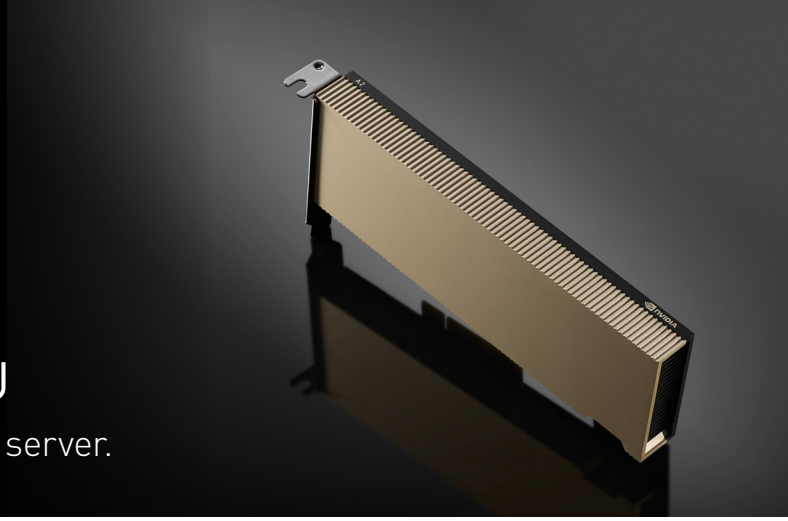
DATASHEET



NVIDIA A2 TENSOR CORE GPU

Entry-level GPU that brings NVIDIA AI to any server.

Versatile Entry-Level Inference

The NVIDIA A2 Tensor Core GPU provides entry-level inference with low power, a small footprint, and high performance for NVIDIA AI at the edge. Featuring a low-profile PCIe Gen4 card and a low 40-60 watt (W) configurable thermal design power (TDP) capability, the A2 brings adaptable inference acceleration to any server.

A2's versatility, compact size, and low power exceed the demands for edge deployments at scale, instantly upgrading existing entry-level CPU servers to handle inference. Servers accelerated with A2 GPUs deliver higher inference performance versus CPUs and more efficient intelligent video analytics (IVA) deployments than previous GPU generations—all at an entry-level price point.

NVIDIA-Certified Systems™ featuring A2 GPUs and NVIDIA AI, including the NVIDIA Triton™ Inference Server, deliver breakthrough inference performance across edge, data center, and cloud. They ensure that AI-enabled applications deploy with fewer servers and less power, resulting in easier deployments, faster insights, and significantly lower costs.

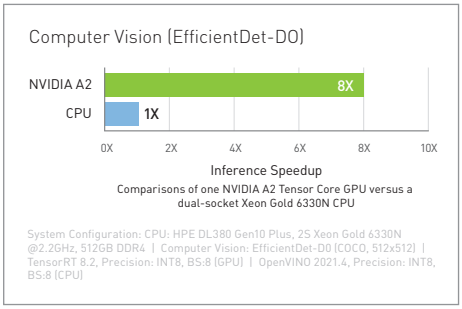
Up to 20X More Inference Performance

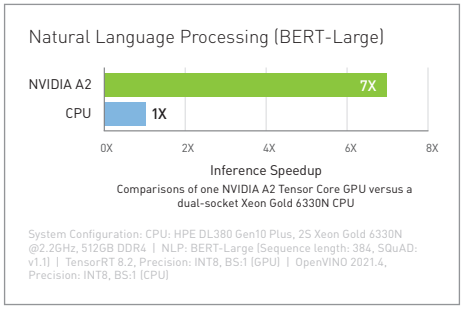
AI inference is deployed to make consumer lives more convenient through real-time experiences, and enables them to gain insights on trillions of end-point sensors and cameras. Compared to CPU-only servers, the servers built with NVIDIA A2 Tensor Core GPU offer up to 20X more inference performance, instantly upgrading any server to handle modern AI.

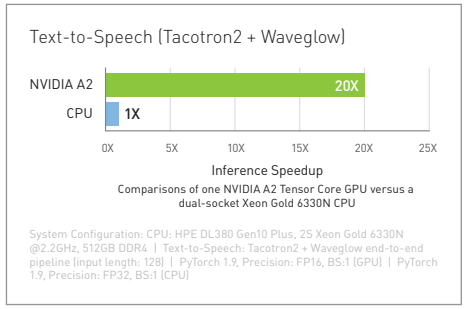
SYSTEM SPECIFICATIONS

Peak FP32	**4.5 TF**
TF32 Tensor Core	**9 TF | 18 TF[1]**
BFLOAT16 Tensor Core	**18 TF | 36 TF[1]**
Peak FP16 Tensor Core	**18 TF | 36 TF[1]**
Peak INT8 Tensor Core	**36 TOPS | 72 TOPS[1]**
Peak INT4 Tensor Core	**72 TOPS | 144 TOPS[1]**
RT Cores	**10**
Media engines	**1 video encoder** **2 video decoders (includes AV1 decode)**
GPU memory	**16GB GDDR6**
GPU memory bandwidth	**200GB/s**
Interconnect	**PCIe Gen4 x8**
Form factor	**1-slot, Low-Profile PCIe**
Max thermal design power (TDP)	**40-60W (Configurable)**
vGPU software support[2]	**NVIDIA Virtual PC (vPC), NVIDIA Virtual Applications (vApps), NVIDIA RTX Virtual Workstation (vWS), NVIDIA AI Enterprise, NVIDIA Virtual Compute Server (vCS)**

[1] With sparsity
[2] Supported in future vGPU release

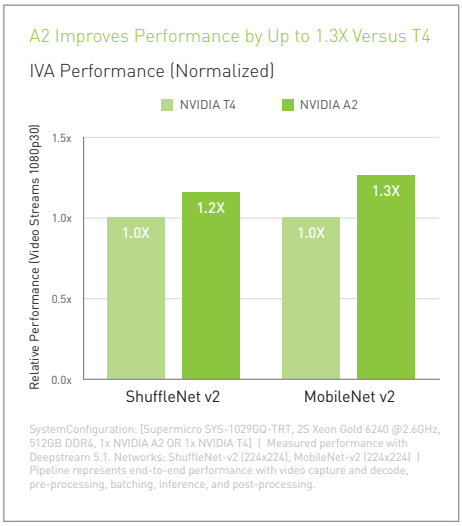




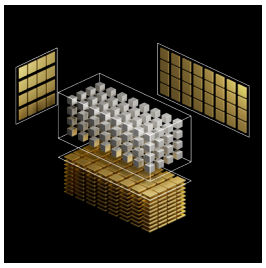


Higher IVA Performance for Intelligent Edge

Servers equipped with A2 offer up to 1.3X more performance in intelligent edge use cases, including smart cities, manufacturing, and retail. NVIDIA A2 GPUs running IVA workloads result in more efficient deployments with up to 1.6X better price-performance and ten percent better energy efficiency than previous GPU generations.

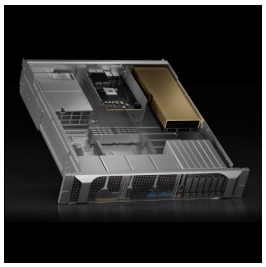


NVIDIA A2 Brings Breakthrough NVIDIA Ampere Architecture Innovations



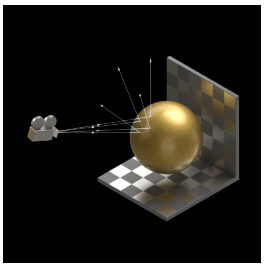
THIRD-GENERATION TENSOR CORES

The third-generation Tensor Cores in A2 support integer math, down to INT4, and floating point math, up to FP32, to deliver high AI training and inference performance. The NVIDIA Ampere architecture also supports TF32 and NVIDIA's automatic mixed precision (AMP) capabilities.



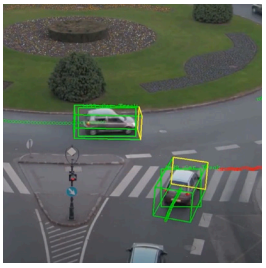
ROOT OF TRUST SECURITY

Providing security in edge deployments and end-points is critical for enterprise business operations. A2 optionally supports secure boot through trusted code authentication and hardened rollback protections to protect against malicious malware attacks.



SECOND-GENERATION RT CORES

A2 includes dedicated RT Cores for ray tracing that enable groundbreaking technologies at breakthrough speed.
With up to 2X the throughput over the previous generation and the ability to concurrently run ray tracing with either shading or denoising capabilities.

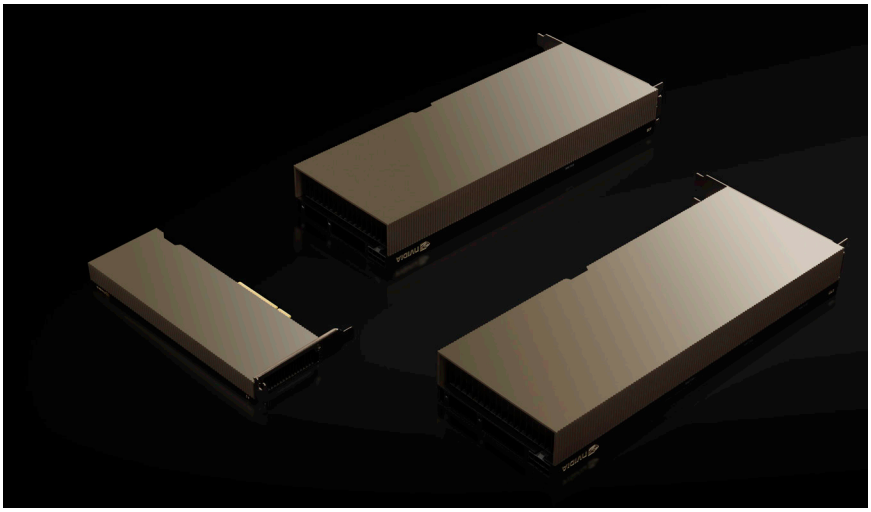


HARDWARE TRANSCODING PERFORMANCE

Exponential growth in video applications demand real-time scalable performance, requiring the latest in hardware encode and decode capabilities. A2 GPUs use dedicated hardware to fully accelerate video decoding and encoding for the most popular codecs, including H.265, H.264, VP9, and AV1 decode.

Complete Inference Portfolio

NVIDIA offers a complete portfolio of NVIDIA-Certified Systems featuring Ampere Tensor Core GPUs as the inference engine powering NVIDIA AI. A2 Tensor Core GPUs add entry-level inference in a low-profile form factor to the NVIDIA AI portfolio that already includes A100 and A30 Tensor Core GPUs. A100 features the highest inference performance at every scale and A30 brings optimal inference performance for mainstream servers. NVIDIA A2, NVIDIA A30, and NVIDIA A100 Tensor Core GPUs deliver leading inference performance across edge, data center, and cloud.



NVIDIA-CERTIFIED SYSTEMS

NVIDIA-Certified Systems with NVIDIA A2 combine compute acceleration and high-speed, secure networking to systems from leading NVIDIA partners in configurations validated for optimum performance, reliability, and scale. With NVIDIA-Certified Systems, enterprises can confidently choose performance-optimized hardware solutions to power accelerated computing workloads—from the desktop to the data center to the edge.

NVIDIA CERTIFIED

Optimized Software and Services for Enterprise

NVIDIA AI Enterprise

NVIDIA AI Enterprise, an end-to-end cloud-native suite of AI and data analytics software, is certified to run on A2 in hypervisor-based virtual infrastructure with VMware vSphere. This enables management and scaling of AI and inference workloads in a hybrid cloud environment.

NVIDIA