Datasheet

NVIDIA.

NVIDIA RTX 5000
Ada Generation

Performance for endless possibilities.
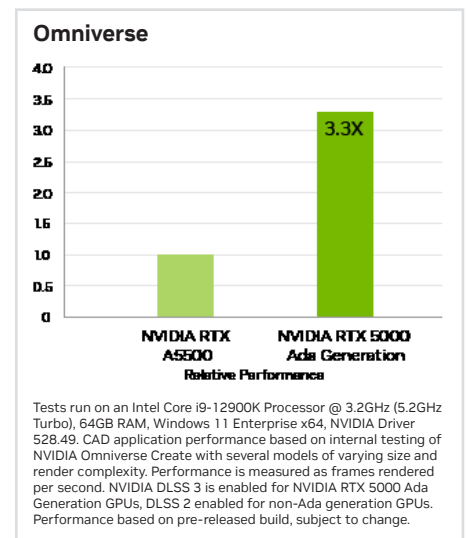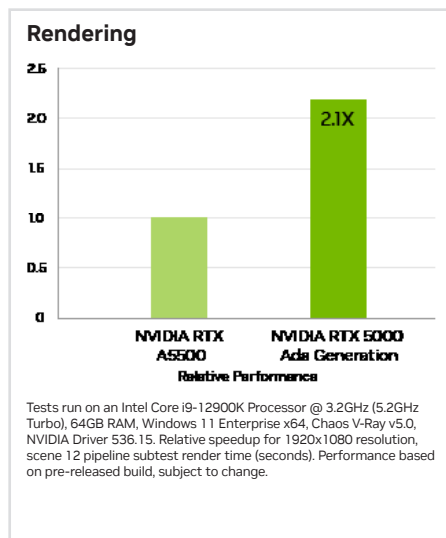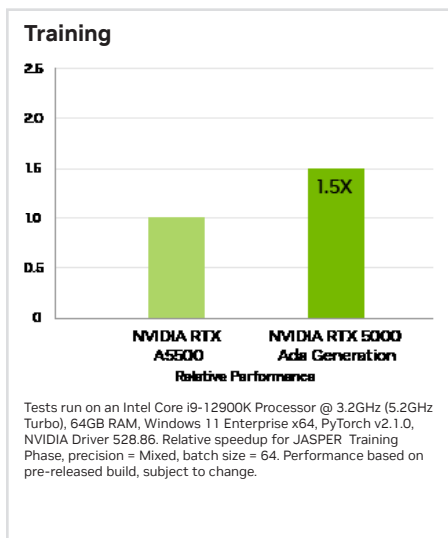
Powering the Next Era of Innovation

Industries are embracing accelerated computing and AI to tackle powerful dynamics and unlock transformative possibilities. Generative AI is reshaping the way professionals create and innovate across various domains, from design and engineering to entertainment and healthcare. The NVIDIA RTX™ 5000 Ada Generation GPU, with third-generation RTX technology, unlocks breakthroughs in generative AI, revolutionizing productivity and offering unprecedented creative possibilities.

The NVIDIA RTX 5000 Ada Generation GPU is purpose-built for today's professional workflows. Built on the NVIDIA Ada Lovelace architecture, it combines 100 third-generation RT Cores, 400 fourth-generation Tensor Cores, and 12,800 CUDA® cores with 32 gigabytes (GB) of graphics memory to deliver the next generation of AI graphics and petaFLOPS inferencing performance, accelerating rendering, AI, graphics, and compute workloads. RTX 5000-powered workstations equip you for success in today's demanding business landscape.
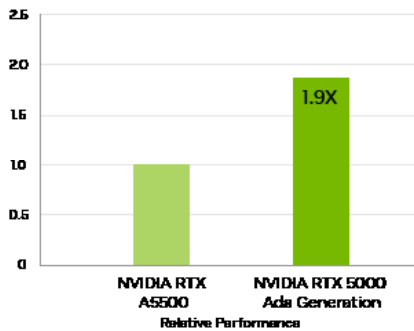
NVIDIA RTX professional graphics cards are certified for a broad range of professional applications, tested by leading independent software vendors (ISVs) and workstation manufacturers, and backed by a global team of support specialists. Get the peace of mind to focus on what matters with the premier visual computing solution for mission-critical business.

Key Features

> PCIe Gen4

> Four DisplayPort 1.4a connectors

> AV1 encode and decode support

> DisplayPort with audio

> 3D stereo support with stereo connector

> NVIDIA® GPUDirect® for Video support

> NVIDIA GPUDirect remote direct memory access (RDMA) support

> NVIDIA Quadro® Sync II[1] compatibility

> NVIDIA RTX Experience

> NVIDIA RTX Desktop Manager software

> NVIDIA RTX IO support

> HDCP 2.2 support

> NVIDIA Mosaic[2] technology

Training

Tests run on an Intel Core i9-12900K Processor @ 3.2GHz (5.2GHz Turbo), 64GB RAM, Windows 11 Enterprise x64, PyTorch v2.1.0, NVIDIA Driver 528.86. Relative speedup for JASPER Training Phase, precision = Mixed, batch size = 64. Performance based on pre-released build, subject to change.

Rendering

Tests run on an Intel Core i9-12900K Processor @ 3.2GHz (5.2GHz Turbo), 64GB RAM, Windows 11 Enterprise x64, Chaos V-Ray v5.0, NVIDIA Driver 536.15. Relative speedup for 1920x1080 resolution, scene 12 pipeline subtest render time (seconds). Performance based on pre-released build, subject to change.

Omniverse

Tests run on an Intel Core i9-12900K Processor @ 3.2GHz (5.2GHz Turbo), 64GB RAM, Windows 11 Enterprise x64, NVIDIA Driver 528.49. CAD application performance based on internal testing of NVIDIA Omniverse Create with several models of varying size and render complexity. Performance is measured as frames rendered per second. NVIDIA DLSS 3 is enabled for NVIDIA RTX 5000 Ada Generation GPUs, DLSS 2 enabled for non-Ada generation GPUs. Performance based on pre-released build, subject to change.
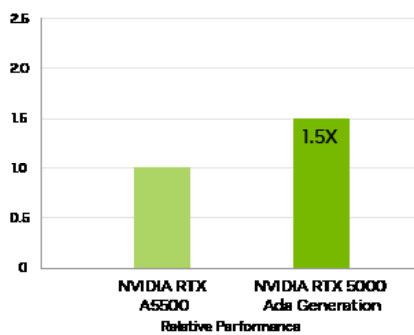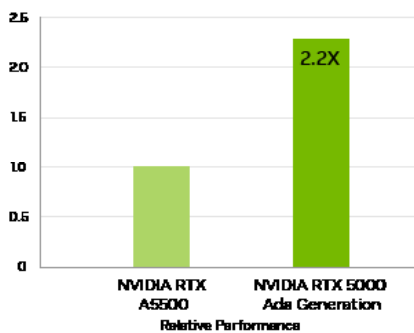
Graphics

Tests run on an Intel Core i9-12900K Processor @ 3.2GHz (5.2GHz Turbo), 64GB RAM, Windows 11 Enterprise x64, SPECviewperf 2020, NVIDIA Driver 528.49. Relative speedup for 4K Siemens NX composite score. Performance based on pre-released build, subject to change.

Generative AI

Tests run on an Intel Core i9-12900K Processor @ 3.2GHz (5.2GHz Turbo), 64GB RAM, Windows 11 Enterprise x64, Stable Diffusion WebUI v1.3.1, NVIDIA Driver 536.15. Relative speedup for 512x512 image generation. Performance based on pre-released build, subject to change.

HPC

Tests run on an Intel Core i9-12900K Processor @ 3.2GHz (5.2GHz Turbo), 64GB RAM, Windows 11 Enterprise x64, CUDA 11.8 (cuBLAS performance), NVIDIA Driver 525.85. Relative speedup for GFLOPS, precision = INT8, input = zero. Performance based on pre-released build, subject to change.

Specifications	
GPU Memory	32GB GDDR6
Memory Interface	256 bit
Memory Bandwidth	576GB/s
Error Correcting Code (ECC)	Yes
NVIDIA Ada Lovelace Architecture-Based CUDA Cores	12,800
NVIDIA Fourth-Generation Tensor Cores	400
NVIDIA Third-Generation RT Cores	100
Single-Precision Performance	65.3 TFLOPS[3]
RT Core Performance	151.0 TFLOPS[3]
Tensor Performance	1044.4 TFLOPS[4]
System Interface	PCIe 4.0 x16
Power Consumption	Total board power: 250W
Thermal Solution	Active
Form Factor	4.4" H x 10.5" L, single slot
Display Connectors	4x DisplayPort 1.4a[5]
Max Simultaneous Displays	4x 4096 x 2160 @ 120Hz 4x 5120 x 2880 @ 60Hz 2x 7680 x 4320 @ 60Hz
Encode/Decode Engines	2x encode, 2x decode (+AV1 encode and decode)
VR Ready	Yes
vGPU Software Support[6]	> NVIDIA vPC/vApps > NVIDIA RTX Virtual Workstation
vGPU Profiles Supported	See the Virtual GPU licensing guide.
Graphics APIs	DirectX 12, Shader Model 6.7, OpenGL 4.6[7], Vulkan 1.3[7]
Compute APIs	CUDA 12.2, OpenCL 3.0, DirectCompute
NVIDIA NVLink ®	No

1 Quadro Sync II card sold separately. I 2 Windows 10 and Linux. I 3 Peak rates based on GPU boost clock. I 4 Effective FP8 teraFLOPS (TFLOPS) using sparsity. I 5 Display ports are on by default for RTX 5000. Display ports aren't active when using vGPU software. | 6 Virtualization support for the RTX 5000 Ada Generation GPU will be available in an upcoming NVIDIA vGPU release, anticipated in Q3, 2023. | 7 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.khronos.org/conformance**

Ready to Get Started?

To learn more about NVIDIA RTX 5000, visit:
www.nvidia.com/rtx-5000