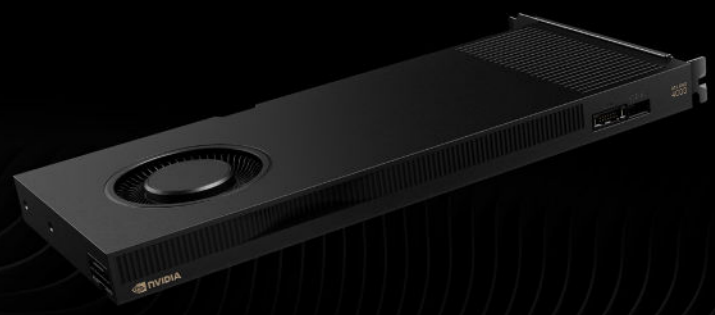
Datasheet

NVIDIA RTX PRO 4000 Blackwell

Powering the next era of AI.



Transform Workflows With the Most Advanced Single-Slot Workstation Solution

As AI continues to advance at an incredible pace, industries face mounting pressure to harness its transformative power and adopt tools capable of handling generative AI, real-time simulation, and hyper-realistic rendering. Enterprises need solutions that combine breakthrough performance, scalability, and versatility to tackle the rise of increasingly complex workloads—from training domain-specific AI models to rendering billion-polygon engineering designs or simulating real-world physics with higher fidelity and precision.

The NVIDIA RTX PRO™ 4000 Blackwell answers this demand with NVIDIA's revolutionary NVIDIA Blackwell architecture, redefining what's possible in a single-slot GPU. Built to accelerate demanding professional workflows, it combines breakthrough AI compute, neural graphics, and power efficiency to deliver unprecedented performance.

Equipped with 24 GB of ultra-fast GDDR7 memory, fifth-generation Tensor Cores, and fourth-generation RT cores, the RTX PRO 4000 tackles large, complex datasets and multi-app workflows effortlessly. Fifth-generation Tensor Cores accelerate generative AI and LLM inference, while fourth-generation RT cores enable cinematic-quality ray tracing, empowering professionals to visualize photorealistic scenes in real time. Its sleek single-slot design ensures seamless integration into compact workstations, high-density rendering nodes, medical imaging systems, and AI research labs, offering unmatched versatility for studios, engineering firms, financial institutions, and healthcare providers.

These capabilities are powered by the NVIDIA Blackwell architecture, a paradigm shift in accelerated computing that merges unprecedented AI, ray tracing, and neural rendering advancements to redefine professional workflows for the next decade.

Key Features

- Enhanced streaming multiprocessors (SMs) built for neural shaders
- Fifth-generation Tensor Cores support FP4 precision, DLSS 4 Multi Frame Generation
- Fourth-generation ray-tracing cores built for detailed geometry
- 24 GB of GDDR7 memory
- 672 GB/s of memory bandwidth
- Ninth-generation NVENC and sixth-generation NVDEC with 4:2:2 support
- PCIe 5.0
- Four DisplayPort 2.1b connectors
- AI management processor

Breakthrough Innovations

The NVIDIA Blackwell architecture combines breakthrough AI, ray tracing, and neural rendering technology with massive performance and memory improvements to drive cutting-edge professional creative, design, and engineering workflows and power the next decade of innovation.

NVIDIA Blackwell Streaming Multiprocessor: The new SM features increased processing throughput and new neural shaders that integrate neural networks inside of programmable shaders to drive the next decade of AI-augmented graphics innovations.

Fifth-Generation Tensor Cores: Deliver up to 3x the performance of the previous generation and support for FP4 precision for faster AI model processing times with reduced memory usage, enabling local fine-tuning of LLMs and generative AI.

Fourth-Generation Ray-Tracing Cores: Double the ray-triangle intersection rate of the previous generation to create photoreal, physically accurate scenes and immersive 3D designs with NVIDIA RTX™ Mega Geometry, which enables up to 100x more ray-traced triangles.

Next-Gen Video Engines: Enhance video conferencing, video production, and streaming workflows with real-time AI processing. Ninth-generation NVENC and sixth-generation NVDEC engines provide support for 4:2:2 encoding and decoding to explore a new realm of high-resolution video workflows.

GDDR7 Memory: New and improved GDDR7 memory significantly boosts bandwidth and capacity, empowering your applications to run faster and work with larger, more complex datasets. With 24 GB of GPU memory and 672 GB/s bandwidth, tackle large 3D and AI projects, fine-tune AI models locally, explore immersive VR environments, and drive multi-app workflows.

DLSS 4: Multi Frame Generation ensures ultra-smooth frame pacing for lifelike simulations. Experience up to 3x faster frame rates and stunning image quality in supported game engines and 3D rendering applications for smoother, more responsive performance.

PCIe 5.0: Support for PCIe 5.0 provides double the bandwidth of PCIe 4.0, improving data-transfer speeds from CPU memory and unlocking faster performance for data-intensive tasks like AI, data science, and 3D modeling.

DisplayPort 2.1b: Achieve unparalleled visual clarity and performance, driving high-resolution displays at up to 8K at 240 Hz and 16K at 60 Hz. Increased bandwidth enables seamless multi-monitor setups, ideal for multitasking and collaboration, while HDR and higher color depth support ensures superior color accuracy for precision work, such as video editing, 3D design, and live broadcasting.

Enterprise Reliability

Designed for professionals who demand the best, NVIDIA RTX PRO solutions deliver unparalleled performance, reliability, and support. Every GPU is rigorously tested for a wide range of design, engineering, and AI workflows and continually optimized through enterprise drivers. With extensive ISV certifications, robust IT management tools, and enterprise-grade support, RTX PRO workstations are the trusted choice for enterprise and mission-critical work.

Technical Specifications	
GPU architecture	NVIDIA Blackwell
NVIDIA® CUDA® cores	8,960
Tensor Cores	Fifth generation
Ray Tracing Cores	Fourth generation
TOPS/TFLOPS	
AI Performance[1]	1178 AI TOPS[2]
Single-Precision performance[1]	37 TFLOPS
RT Core performance[1]	112 TFLOPS
GPU memory	24 GB GDDR7 with ECC
Memory interface	192-bit
Memory bandwidth	672 GB/s
System interface	PCIe 5.0 x16
Display connectors	4x DisplayPort 2.1b
Max simultaneous displays	**>**4x 3840 x 2160 @ 165 Hz **>**2x 7680 x 4320 @ 100 Hz
Video engines	**>**2x NVENC (ninth generation) **>**2x NVDEC (sixth generation)
Power consumption	Total board power: 145 W
Power connector	1x PCIe CEM5 16-pin
Thermal solution	Active
Form factor	4.4” x 9.5” L, single slot, full height
Graphics APIs	DirectX 12, Shader Model 6.6, OpenGL 4.6[3], Vulkan 1.3[3]
Compute APIs	CUDA 12.8, OpenCL 3.0, DirectCompute

Ready to Get Started?

To learn more, visit: nvidia.com/rtx-pro-4000

1 Peak rates are based on GPU boost clock.

2 Theoretical FP4 TOPS using the sparsity feature.

3 Product is based on a published Khronos specification and is expected to pass the Khronos conformation testing process when available. Current conformance status can be found at khronos.org/conformance.

NVIDIA