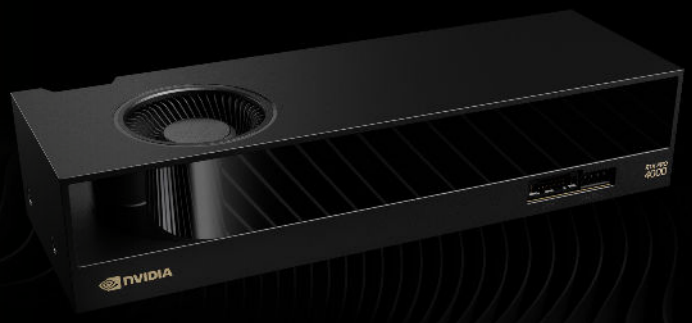
Datasheet

NVIDIA RTX PRO 4000 Blackwell SFF Edition

Powering the next era of AI.



The Power of NVIDIA Blackwell in Compact Workstations

As AI continues to advance at an incredible pace and industries start relying more on instant, compute-heavy tasks—from generative AI to hyper-realistic rendering and physics-based simulation—they face a dual challenge: deploying infrastructure powerful enough to these tools while operating in space-constrained environments. From manufacturing floors to media studios, enterprises demand powerful performance in compact systems—solutions that balance raw computational power with energy efficiency, thermal control, and seamless integration into slim towers, portable edge systems, or dense rack-mounted workstations.

For design studios, engineering labs, and financial hubs, this means accelerating large CAD assemblies, AI-driven simulations, and real-time 3D rendering without sacrificing desk space or scalability. Whether optimizing factory layouts, visualizing smart cities, or powering real-time 3D animation and AI-upscaled visual effects, professionals also need solutions that fit where traditional GPUs cannot—yet still deliver the performance to tackle tomorrow's challenging workloads.

The NVIDIA RTX PRO™4000 Blackwell SFF Edition redefines what's possible in a compact, low-profile GPU, combining NVIDIA's revolutionary Blackwell architecture with 24 GB of ultra-fast GDDR7 memory to provide full size performance in a small form factor. Built to accelerate demanding professional workflows, it offers incredible AI compute, next-generation neural graphics, and unmatched power efficiency for its class.

Equipped with fifth-generation Tensor Cores and fourth-generation RT Cores, the RTX PRO 4000 SFF effortlessly tackles large, complex datasets and multi-app workflows. Its Tensor Cores accelerate generative AI and LLM inference, while RT Cores enable cinematic-quality ray tracing, empowering professionals to visualize photorealistic scenes in real time. Its sleek, low-profile, small form-factor design ensures seamless integration into compact workstations, edge devices, and advanced automation systems. The RTX PRO 4000 SFF is engineered as the most powerful low-profile workstation GPU, offering unmatched versatility for design studios, engineering firms, financial institutions, and industrial applications.

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, all within the constraints of truly small systems.

Key Features

- Enhanced Streaming Multiprocessors (SMs) built for neural shaders
- 5th Gen Tensor Cores support FP4 precision, DLSS 4 Multi Frame Generation
- 4th Gen Ray Tracing Cores built for detailed geometry
- 24 GB of GDDR7 memory
- 432 GB/s of memory bandwidth
- 9th Gen NVENC and 6th Gen NVDEC with 4:2:2 support
- PCIe Gen 5
- Four Mini DisplayPort 2.1b connectors
- AI Management Processor

Breakthrough Innovations

The NVIDIA Blackwell architecture combines 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 is optimized for neural shaders, supporting the integration of neural networks inside programmable shaders to drive the next wave of AI-augmented graphics innovation.

5th Gen Tensor Cores: Deliver up to 3X the performance of the previous generation and support for FP4 precision, enabling faster AI model processing times with reduced memory usage. This supports local fine-tuning of LLMs and generative AI.

4th Gen Ray Tracing Cores: Double the ray-triangle intersection rate of the previous generation to create photoreal, physically accurate scenes and immersive 3D designs with 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 engines support 4:2:2 10-bit encoding/decoding for high-quality color fidelity and smooth multi-stream 4K workflows, powering advanced creative and video production needs.

GDDR7 Memory: The latest generation of GPU memory technology, GDDR7, 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 432 GB/s bandwidth, tackle large 3D and AI projects, fine-tune AI models locally, explore large-scale VR environments, and drive larger 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 Gen 5: Support for PCIe Gen 5 provides double the bandwidth of PCIe Gen 4, enhancing data-transfer speeds from CPU memory and unlocking faster performance for data-intensive tasks such as 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 ensure 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.

Specifications	
GPU architecture	NVIDIA Blackwell
NVIDIA® CUDA® Cores	8,960
Tensor Cores	5th Generation
Ray Tracing Cores	4th Generation
GPU memory	24 GB GDDR7 with ECC
Memory interface	192 bit
Memory bandwidth	432 GB/s
System interface	PCIe 5.0 x8[1]
Display connectors	4x Mini DisplayPort 2.1b
Max simultaneous displays	> 4x 3840x2160 @ 165 Hz > 2x 7680x4320 @ 100 Hz
Video Engines	2x NVENC (9th Gen) 2x NVDEC (6th Gen)
Power consumption	Total board power: 70 W
Thermal solution	Active
Form factor	2.7” x 6.6” L, dual slot, half height
Graphics APIs	DirectX 12, Shader Model 6.7, OpenGL 4.6[2], Vulkan 1.4[2]
Compute APIs	CUDA 12.8, OpenCL 3.0

Ready to Get Started?

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

1. Uses full-length PCIe interface.
2. Product is based on a published Krhonos specification and is expected to pass the Khronos conformation testing process when available. Current conformance status can be found at www.khronos.org/conformance.