Solution Overview

Experiment, Prototype, and Innovate With AI

NVIDIA RTX PRO AI Workstations

Meeting the Demands for AI Computing Resources

AI is bringing profound change across industries, accelerating the adoption of advanced technologies like Agentic AI, which utilizes sophisticated reasoning and iterative planning to autonomously solve complex, multi-step problems. The next frontier, Physical AI, integrates intelligence into physical systems such as robots and autonomous vehicles, enabling them to perceive, reason, and interact directly with the real world. This progression marks a shift towards proactive workflows in both digital and physical domains, promising unprecedented creativity, productivity, and efficiency that transforms operations and drives innovation across sectors.

However, harnessing the capabilities of these increasingly sophisticated AI systems demands a significant surge in computing power. Unlike previous generations, these models think through complex problems by generating numerous intermediate steps or internal reasoning tokens before producing a final answer. This internal thinking process can generate significantly more tokens per query—sometimes hundreds or thousands—compared to earlier AI models, driving up the computational requirements and costs needed to generate a response.

The evolving sophistication of AI models coupled with demands for real-time performance is pushing legacy systems beyond their operational limits - requiring more robust workstation solutions capable of supporting rapid prototyping and experimentation while enabling easy scaling to data center and cloud resources for building larger, more complex models.

NVIDIA RTX PRO AI Workstations

NVIDIA RTX PRO™ AI Workstations offer incredible desktop and laptop computing performance, perfectly suited for AI training, inference, and data science workflows. These robust workstations are optimized for processing smaller AI models locally with exceptional efficiency, while also providing seamless integration with data center and cloud resources for the development of larger, more complex models. With up to four NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs per workstation, these AI workstations provide an incredible 14.2 petaflops of combined AI compute performance and 384 GB of total system GPU memory, bringing data center AI computing power to the desktop. As a companion for running and prototyping smaller AI models, professional laptops provide up to 24 GB of GPU memory to take AI inference on the go.

Key challenges

- **Hardware:** Current desktop systems often lack the processing power needed for AI workflows, leading to greater reliance on already strained cloud and data center resources.
- **Training and fine-tuning:** AI model size continues to grow, taking months to train and taxing already oversubscribed data center and cloud instance resources. Off-the-shelf models trained on public data require fine tuning with domain-specific data to provide content relevant to business purposes.
- **Workflow complexity:** Modern professional workflows require running multiple applications simultaneously to maximize productivity. Adding AI-augmented tools and applications put additional requirements on current computing solutions.

Fast-Track All Workflows

Data Science

RTX PRO AI Workstations, equipped with up to four GPUs and substantial GPU memory, are well-suited for data science tasks. These workstations reduce latency, facilitating real-time data preprocessing, exploration, visualization, and evaluation of features and models, while saving time and valuable data center and dedicated cloud compute resources. RTX PRO AI Workstations are designed to integrate seamlessly with NVIDIA CUDA®-X Data Science libraries. a comprehensive open-source collection of GPU-accelerated data science and AI libraries that align with popular open-source data tools. Specifically, RAPIDS cuDF significantly accelerates DataFrame operations in libraries like pandas and polars, compared to CPU-only systems, while cuML accelerates machine learning algorithms for libraries such as scikit-learn, UMAP, and HDBSCAN, without requiring any code modifications

AI Model Training and Fine-Tuning

For businesses or individuals starting with AI or using smaller models, NVIDIA RTX PRO AI Workstations offer a robust and cost-effective solution for AI research and development. These workstations serve as an ideal testbed for evaluating smaller generative and reasoning AI models, with similar accuracy as larger models, without overloading data center servers or cloud resources. With large system memory, storage, and high-performance ConnectX® networking, these workstations are well-suited for training AI models on specific, smaller data sets. Dev teams can fine-tune models as needed and experiment with various datasets and data sizes to optimize results, all while conserving data center or cloud computing resources.

NVIDIA provides a full-stack solution for AI development, from NVIDIA professional GPUs for desktop, laptops, data center, and cloud to GPU-accelerated AI frameworks and tools to pretrained AI models. Easily access NVIDIA accelerated AI software from NVIDIA® NGC™, an online portal for enterprise services, software, management tools, and support for end-to-end AI workflows, as well as NVIDIA AI Workbench, a unified, easy-to-use developer toolkit to create, test, and customize pretrained AI models, empowers developers to fully harness the power of AI workstations.

AI Agents

As AI evolves, the shift from generative AI, which creates new content, to agentic AI, which empowers AI systems to make decisions and act autonomously, is becoming more pronounced. Agentic AI demands the ability to reason, plan, and execute, placing even greater demands on computing infrastructure. RTX PRO AI Workstations, with up to 14.2 petaflops of combined AI compute performance and 384 GB of total system GPU memory, provides developers with the ideal sandbox experience to iteratively experiment, test and customize AI agents for optimal performance before large-scale deployment.

To help get started, NVIDIA AI Blueprints provides enterprise developers with reference AI workflows, pre-trained for specific use cases, that can be customized with their own data and run locally on RTX PRO AI Workstations. Each AI Blueprint includes reference code for constructing the workflow using the NVIDIA AI Enterprise catalog of software, including NVIDIA NIM microservices and NVIDIA NeMo™, along with tools and documentation for deployment and customization with proprietary data. With these blueprints, developers can build AI agents for canonical use cases, such as virtual assistants, retrieval-augmented generation, video search and summarization, and more.

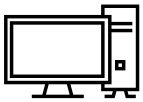
AI Workstations benefits

- Provides additional AI computing resources to augment data center and cloud instances for development and R&D tasks.
- Large GPU memory configurations enable AI-augmented, multi-application workflows that maximize productivity.
- Integration with NVIDIA AI Enterprise to accelerate data science pipelines and streamline the development and deployment of production-grade AI.
- Enterprise-grade solutions that are widely available from OEM workstation vendors worldwide.

Physical AI

Physical AI represents the next evolution where intelligent systems perceive, reason about, and interact directly with the physical world. This requires understanding physical dynamics and often involves real-time decision-making for applications like robotics and autonomous vehicles, demanding significant computing power for tasks like training AI models on physics-based simulations and processing multi-modal sensor data. NVIDIA RTX PRO AI Workstations, powered by high-performance NVIDIA RTX PRO 6000 Blackwell Workstation Edition series GPUs featuring 96 GB of GDDR7 memory per GPU provide the necessary compute and AI acceleration directly on the desktop. This local power enables developers to work with sophisticated tools like NVIDIA Cosmos world foundation models, which are designed to advance physical AI development. Developers can download Cosmos models from platforms like NVIDIA NGC or Hugging Face and leverage the capabilities of RTX PRO Workstations to fine-tune or post-train these models locally using frameworks like NVIDIA NeMo or PyTorch.

Recommended Configurations for Desktop Workstations and Laptops for Data Science and AI Training



Desktop Workstations

	Good	Better	Best
CPU	Intel Core Ultra 9 Series	AMD Threadripper Pro or Intel Xeon w5	AMD Threadripper Pro or Intel Xeon w9
System Memory	192 GB ECC DDR5	384 GB ECC DDR5 (or more)	768 GB ECC DDR5 (or more)
Storage	1 TB boot + 2 TB NVMe SSDs	1 TB boot + 2-4 TB NVMe SSDs with RAID[1]	2 TB boot + 2-8 TB NVMe SSDs with RAID[1]
NIC	10 GbE	NVIDIA ConnectX-6Dx (25 GbE)	NVIDIA ConnectX-6Dx (25 GbE)
OS	Ubuntu / RHEL / SUSE[2] / Windows 11 / WSL2[3]	Ubuntu / RHEL / SUSE[2] / Windows 11 / WSL2[3]	Ubuntu / RHEL / SUSE[2] / Windows 11 / WSL2[3]
GPU	NVIDIA RTX PRO 6000 Blackwell Workstation Edition or NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition	2x NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition	4x NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition

Laptop Workstations

	Best
CPU	Intel Core Ultra 9 Series
System Memory	64 GB DDR5
Storage	1 TB
OS	Ubuntu / RHEL / SUSE[2] / Windows 11 / WSL2[3]
GPU	NVIDIA RTX PRO 5000 Blackwell Generation

AI workstations powered by NVIDIA RTX PRO

1. For large datasets that need fast/reliable storage.
2. See NVIDIA AI Enterprise documentation for Linux release support.
3. NVIDIA AI Enterprise is not supported on Windows 11 or WSL2.

Inference

Applications with AI-enabled features—such as Adobe® Photoshop's® Neural Filters, DaVinci Resolve's face tracking, NVIDIA Broadcast's noise and room echo removal, and image denoising in every major rendering application software—have been available for several years. Workstations equipped with RTX professional GPUs have been the platform of choice for modern AI-powered workflows.

As AI brings new levels of capabilities and efficiency, professionals but requires more computing power and GPU memory. Professional users work with high-resolution content, using workflows that require the simultaneous use of multiple professional applications. As professional workflows include AI-augmented tools to help with concept development and creation, these compute- and memory-intensive applications will put additional demands on the GPU. Running large language models (LLMs) like chatbots and code copilots locally on a workstation further amplifies these requirements, making powerful GPUs essential to ensure smooth performance and efficient multitasking. NVIDIA RTX PRO-powered AI Workstations are built for these demanding workloads. The NVIDIA RTX PRO 6000 Blackwell Workstation Edition and NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition, each with 96 GB of GPU memory, have the raw AI computing power and memory necessary to work with high-resolution AI content, iterate, and pass content on to other design or creative applications without needing to shut down other applications or reduce content fidelity.

NVIDIA RTX PRO AI Workstations, supporting up to four NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs, can provide the AI computing power necessary for local inferencing deployments, enabling AI solutions for workgroups or departments.

AI-Enabled Applications

	Desktop Configurations	Laptop Configurations
Best	RTX PRO 6000 Blackwell Workstation Edition, RTX PRO 6000 Blackwell Max-Q Workstation Edition, Multi-GPU	NVIDIA RTX PRO 5000 Blackwell Generation
Better	RTX PRO 5000 Blackwell, RTX PRO 4500 Blackwell	NVIDIA RTX PRO 4000, 3000 Blackwell Generation
Good	RTX PRO 4000 Blackwell	NVIDIA RTX PRO 2000, 1000, 500 Blackwell Generation

Enterprise-Ready Solutions

Available from leading OEM workstation manufacturers, AI workstations are designed and built for demanding enterprise deployments. Powered by the latest generation of NVIDIA RTX PRO GPUs for workstations, CPUs and available with NVIDIA ConnectX® high-performance networking solutions, AI workstations are ready to tackle demanding AI development workflows. The latest generation of OEM desktop and mobile workstations are available now and ready to ship.

With a full stack of enterprise-level deployment, support, and optimization tools, AI workstations easily fit into existing IT infrastructure, providing drop-in solutions for AI training, development and inferencing on the desktop. The NVIDIA GPU architecture scales from embedded devices, laptop, desktops, data center, and cloud, supporting the same software stack across devices, which enables users to move workloads seamlessly between them.

Ready to Get Started?

To learn more about the NVIDIA RTX PRO AI Workstations, visit nvidia.com/ai-workstations.

Contact sales at nvidia.com/en-us/contact/sales.

NVIDIA