Solution Overview

NVIDIA

Experiment, Prototype, and Innovate With AI

NVIDIA RTX-Powered AI Workstations

Meeting the Demands for AI Computing Resources

Generative AI is bringing profound change across industries, accelerating the adoption of AI-infused technologies at an incredible scale. These new AI-powered workflows offer the promise of new levels of creativity and productivity, improving efficiency across industries.

But they also require significantly more computing power than before. The models used for generative AI are very large, taking weeks or months to train on clusters of servers. The results are highly complex models, capable of understanding language, voice, and audio—or creating content such as articles, images, and music, and much more.

The rush to rapidly add AI computing power to data centers and increase the availability of accelerated cloud instances is straining the availability of hardware. This makes it difficult to meet the continually growing demand.

NVIDIA RTX-Powered AI Workstations

NVIDIA RTX™-powered AI workstations offer incredible desktop 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 6000 Ada Generation GPUs per workstation, these AI workstations provide an incredible 5.8 petaflops of combined AI compute performance and 192GB of total system GPU memory. This elegant solution brings data-center-levels of AI computing power to the desktop with ease.

Key Workloads

AI Workstations for Data Science

RTX-powered AI workstations, equipped with up to four GPUs and substantial GPU memory, are well-suited for data science tasks. AI workstations reduce latency, which helps facilitate real-time data preprocessing, exploration, visualization, and evaluation of features and models, saving time and valuable data center and dedicated cloud compute resources. RTX-powered AI workstations are designed to integrate seamlessly with NVIDIA CUDA®-X libraries. This includes RAPIDS, a comprehensive

Key challenges

- **Hardware:** Demand for accelerated AI hardware for data centers and cloud service providers (CSPs) is exceeding supply. Current desktop computing resources may not be suitable for AI-augmented workflows.
- **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 puts additional requirements on current computing solutions.

open-source collection of GPU-accelerated data science and AI libraries that aligns with popular open-source data tools. Specifically, RAPIDS cuDF significantly boosts Pandas' performance by up to 110X (as compared to CPU-only systems), without any code modifications.

AI Workstations for Model Training and Fine-tuning

For businesses or individuals starting with AI or using smaller models, NVIDIA RTX-powered AI workstations offer a robust and cost-effective solution for AI R&D. These workstations serve as an ideal workspace for developing, evaluating and testing models, augmenting 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 RTX™ Professional GPUs for desktops, laptops, data centers, and the cloud to GPU-accelerated AI frameworks, tools, and pretrained AI models. Also included is:

> Easy access to NVIDIA accelerated AI software from NVIDIA® NGC™, an online portal for enterprise services, software, management tools, and support for end-to-end AI workflows.

> NVIDIA AI Workbench enables developers and data scientists to create, collaborate, customize, and scale AI projects on infrastructure of your choice - from fully harnessing the power of AI workstations to scaling up to data center or cloud.

> NVIDIA AI Enterprise is an end-to-end software platform that accelerates data science pipelines and streamlines development and deployment of production-grade co-pilots and other generative AI applications. Easy-to-use microservices provide optimized model performance with enterprise-grade security, support, and stability to ensure a smooth transition from prototype to production.

> Part of NVIDIA AI Enterprise, NVIDIA NIM is a set of easy-to-use inference microservices designed to speed up generative AI deployment in enterprises. NIM supports a wide range of AI models, including NVIDIA AI foundation and custom models and leverages industry-standard APIs, enabling developers to quickly build enterprise-grade AI applications with just a few lines of code.

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.

> Enterprise-grade solutions that are widely available from OEM workstation vendors worldwide.

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

	Good	Better	Best
CPU	W3 or 4th Gen Intel Xeon Silver or AMD Ryzen Threadripper Pro	Intel Xeon w5 Silver or AMD Ryzen Threadripper Pro	Intel Xeon w5 Gold or AMD Ryzen Threadripper Pro
System Memory	128GB ECC DDR5	256GB ECC DDR5	1TB ECC DDR5
Storage	1TB boot + 2B SSD, NVMe	1TB boot + 2-4TB SSD, NVMe - RAID[3]	2TB boot + 2TB SSD, NVMe - RAID[3]
NIC	10 GbE NIC	NVIDIA ConnectX-6Dx (256GbE)	NVIDIA ConnectX-6Dx (256GbE)
OS	Ubuntu/RHEL/SUSE[1]	Ubuntu/RHEL/SUSE[1]	Ubuntu/RHEL/SUSE[1]
GPU	NVIDIA RTX 6000 Ada Generation or NVIDIA A800 40GB Active[2]	2x NVIDIA RTX 6000 Ada Generation or 2x NVIDIA A800 40GB Active[2]	4x NVIDIA RTX 6000 Ada Generation or 3x NVIDIA A800 40GB Active[2]

	Best
CPU	Intel i7 or i9
System Memory	32GB DDR5
Storage	1TB NVMe
OS	Ubtunbtu/RHEL/SUSE[1]
GPU	NVIDIA RTX 5000 Ada Generation laptop GPU

[1] See NVIDIA AI Enterprise **documentation** for Linux release support.

[2] The NVIDIA A800 40GB Active does not come equipped with display ports. Either the NVIDIA RTX 4000 Ada Generation, NVIDIA RTX A1000 is required to support display out capabilities.

[3] For large datasets that need fast/reliable storage.

AI Workstations for Inference

Applications with AI-enabled features—Adobe® Photoshop's® Neural Filters, DaVinci Resolve's face tracking, NVIDIA Broadcast's noise and room echo removal, and image denoising in major rendering software—have been available for several years. Generative AI is bringing new levels of capabilities and efficiency to professionals, demanding more computing power and GPU memory. Professionals often work with high-resolution content and use multiple applications simultaneously, which increases these requirements.

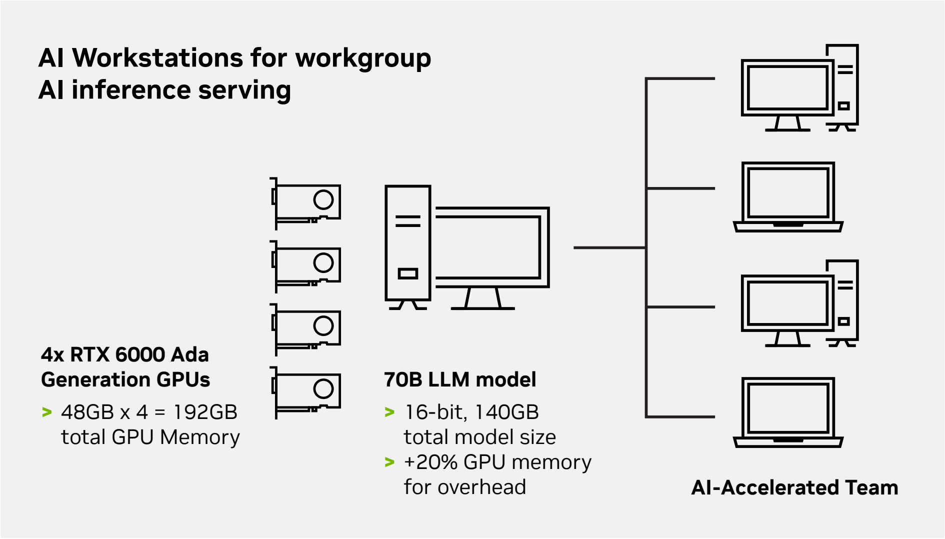
As generative AI tools become integral to professional workflows, the need for powerful GPUs grows. Running large language models (LLMs) like chatbots and code copilots locally on workstations further amplifies these demands. NVIDIA RTX-powered AI workstations are designed to handle these intensive workloads. The NVIDIA RTX 6000 Ada Generation GPU, with 48GB of GPU memory, provides the raw AI computing power and memory necessary to manage high-resolution generative AI content, iterate designs, and integrate with other applications seamlessly, without compromising performance or content fidelity.

AI-Enabled Applications

Latest-generation workstations and laptops plus:

Best	RTX 6000, 5000 Series, Multi-GPU	RTX 5000, 4000 Series
Better	RTX 4000 Series Multi-GPU	RTX 3000, 2000 Series
Good	RTX 2000 Series	RTX 1000, A500 Series

High-end workstation configurations can also meet the inferencing needs of small user groups, such as workgroups or departments, thereby augmenting data center and cloud resources.



AI Workstations for workgroup AI inference serving

4x RTX 6000 Ada Generation GPUs

- 48GB x 4 = 192GB total GPU Memory

70B LLM model

- 16-bit, 140GB total model size
- +20% GPU memory for overhead

AI-Accelerated Team

Enterprise-Ready Solutions

Available from leading OEM workstation manufacturers, NVIDIA AI workstations are designed and built for demanding enterprise deployments. Powered by the latest generation of workstation CPUs and available with NVIDIA ConnectX® high-performance networking solutions, NVIDIA 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, NVIDIA 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 cloud to data center, desktop, laptop, and embedded device, 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-powered AI workstations, visit: **www.nvidia.com/ai-workstations**

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

Partner Logo