Solution Overview

NVIDIA RTX-Powered AI Workstations for AI Inference

The power of AI-augmented workflows on the desktop.

The Challenges of AI-Augmented Workflows

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.

Data center and cloud resources need to be expanded to take on new AI inferencing workloads. Increasing data center capacity or acquiring additional cloud instances can be prohibitive with respect to cost and hardware availability. To harness the power of generative AI on the desktop, businesses are discovering that their traditional desktop computing solutions are inadequate for these new AI-powered tools and applications.

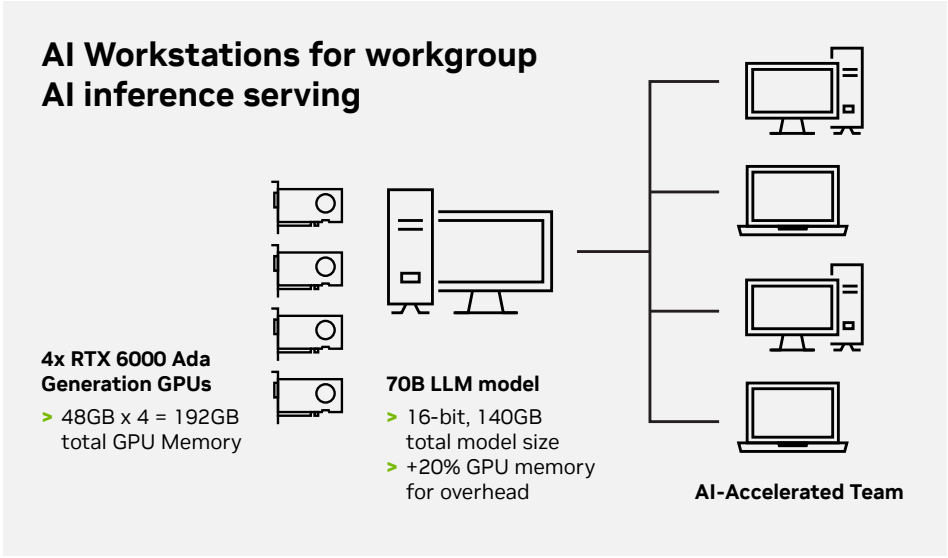
Key Challenges for AI Inference

- **Hardware:** Demand for AI acceleration hardware for data centers and cloud service providers (CSPs) is exceeding supply. Current desktop computing resources may not be suitable for AI-augmented workflows.
- **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.
- **Scaling:** Enhancing workloads with AI-augmented tools and applications requires the latest hardware to take advantage of the latest technology.

AI Workstations for Inference

Large generative AI models require substantial data center and cloud resources to provide inferencing to large numbers of users with acceptable response times. Generative AI models are not only applicable to large data center and cloud deployments but also desktop workstation systems.

NVIDIA RTX™-powered AI workstations can run smaller model inferencing workloads. The latest generation of OEM workstations provide for up to four RTX 6000 GPUs per workstation for an incredible 5.8 petaflops of combined compute performance and 192 gigabytes (GB) of total system GPU memory for local inferencing. High-end workstation configurations should be able to meet the inferencing needs of small numbers of users, such as a workgroup or department.



Benefits for AI Inference

- Additional hardware augments data center and cloud resources and is readily available worldwide from major OEM workstation vendors.
- Large GPU memory configurations enable AI-augmented, multi-application workflows that maximize productivity. Availability of workstation solutions lets businesses increase system capabilities as their workflows expand.
- Enterprise-grade hardware maximizes uptime with enterprise-level performance, reliability, and support.

NVIDIA RTX-Powered AI Workstations for Generative AI

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.

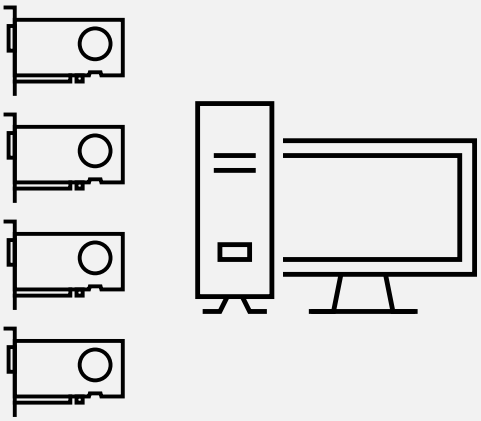
Generative AI brings new levels of capabilities and efficiency to professionals but requires more computing power and GPU memory. Professional users work with high-resolution content, using workflows that require the simultaneous use of multiple professional applications. NVIDIA RTX-powered AI workstations are built for these demanding workloads. The NVIDIA RTX 6000 Ada Generation GPU, with 48GB of GPU memory, has the raw AI computing power and memory necessary to work with high-resolution generative AI content, iterate, and pass content on to other design or creative applications without needing to shut down other applications or reduce content fidelity. As professional workflows include generative AI tools to help with concept development and creation, these compute- and memory-intensive applications will put additional demands on the GPU. Larger, more powerful GPUs will be needed to drive these new AI-augmented workflows.

Enterprise-Class Solutions

NVIDIA RTX-powered AI workstations are based on the latest generation of OEM workstation platforms and are readily available from worldwide OEM workstation partners. AI workstations provide the enterprise-class performance, reliability, and support required for mission-critical enterprise deployments.

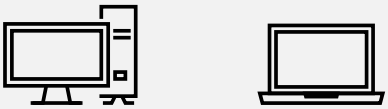
With a full stack of enterprise-level deployment, support, and optimization tools, AI workstations easily fit into existing IT infrastructure, providing drop-in solutions for AI inferencing on the desktop. The NVIDIA GPU architecture scales from cloud to data center, desktop, laptops, and embedded devices,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

Local Inference Serving



Latest-generation workstations plus
1-4 NVIDIA RTX 6000 Ada Generation GPUs

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

Ready to Get Started?

To learn more about NVIDIA RTX-powered AI Workstations, visit: **www.nvidia.com/ai-workstations/**

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

NVIDIA