Solution Overview

NVIDIA

NVIDIA RTX-Powered AI Workstations for AI Training

AI model development and fine-tuning training on the desktop.

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. This rapid deployment has increased demand on computing resources, with data centers and cloud service providers (CSPs) racing to add the hardware required to meet that demand.

NVIDIA data center GPUs are powering these AI workflows around the globe. They provide the compute power required to train trillion+ parameter models and the inferencing necessary to drive the new AI-intensive workflows these models enable.

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

NVIDIA RTX-Powered AI Workstations for AI Training

The latest generation of NVIDIA's AI-focused OEM workstations provide for up to four NVIDIA RTX™ 6000 Ada Generation GPUs per workstation for an incredible 5.8 petaflops of combined compute performance and 192GB of total system GPU memory.

For businesses or individuals just getting started with AI or working with smaller models, NVIDIA RTX-powered AI workstations provide a powerful and cost-effective solution for AI research and development workloads. Equipped with large system memory, storage, and high-performance ConnectX® networking, AI workstations are ideal for training AI models with smaller domain-specific data sets.

Key Challenges for AI Model Development and Training

- **Training:** AI model size continues to grow, taking months to train and taxing already oversubscribed data center and cloud instance resources.
- **Hardware:** Demand for accelerated AI hardware for data centers and CSPs is exceeding supply.
- **Fine-Tuning:** Foundation models don't include data for a specific business, which means models need to be fine-tuned to provide the desired results.

Benefits for AI Model Development and Training

- Provides additional AI computing resources to augment data center and cloud instances for development and R&D tasks
- Ideal solutions for fine-tuning training locally
- Enterprise-grade solutions that are widely available from OEM workstation vendors worldwide

	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 + 2TB 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-7Dx (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

AI Workstations Powered by NVIDIA RTX

Augmenting AI Model Development With NVIDIA RTX-Powered AI Workstations

While the largest generative AI models are created with trillions of parameters and may take weeks or months to train on large clusters of GPU-equipped servers, model development, R&D, and exploration can be done with fewer numbers of parameters and smaller datasets. AI workstations are ideal platforms for AI researchers and developers to experiment and validate work prior to full-scale model training in the data center or on cloud instances. AI workstations can augment data center and cloud compute resources, providing additional AI computing power to maximize the productivity of AI development teams.

NVIDIA provides a full-stack solution for AI development, from NVIDIA RTX professional GPUs for desktop, laptops, data center, and cloud to GPU-accelerated AI frameworks and tools to pretrained AI models. 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, or NVIDIA AI Enterprise, the end-to-end software platform for production AI, lets developers easily access the full power of AI workstations.

As researchers work to create smaller generative AI models with the same accuracy as larger models, AI workstations are an ideal testbed for evaluating these smaller models without putting additional development tasks on data center servers or cloud instances.

[1] Linux: See NVIDIA AI Enterprise Documentation for exact release support

[2] A800: Display out options include T1000 (8GB), A4000, RTX 4000 Ada Gen; see mix-and-match support matrix for more details

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

Fine-Tuning Training With NVIDIA RTX-Powered AI Workstations

As enterprises deploy generative AI models, commercially available models may not provide the results required for business tasks. These models were likely trained on data that didn't include specific company assets, such as product images, datasheets, installation guides, marketing style guides, data that generative AI needs to provide accurate responses, or the look and feel of a particular business's product or services. AI models will likely need to be "tuned"—or receive additional training with company-specific data—to produce the desired results.

AI workstations can help augment data center and cloud instance resources by providing a local computing resource for AI model fine-tuning. As businesses continue to release new products and services and refine the look and feel of their branding, AI models will need to be updated to include the latest data so generated content can stay current. With AI workstations, groups can fine-tune models as required and experiment with various datasets and dataset sizes to optimize results without needing data center or cloud instance compute cycles.

GPT3-40B Fine Tuning Training with 860M Tokens - 15 Hours on AI Workstation with 4X RTX 6000 Ada Generation GPUs.

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 workstation 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.

Ready to Get Started?

To learn more about the NVIDIA RTX-powered AI Workstations, visit:

www.nvidia.com/ai-workstations

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

2958645. OCT23