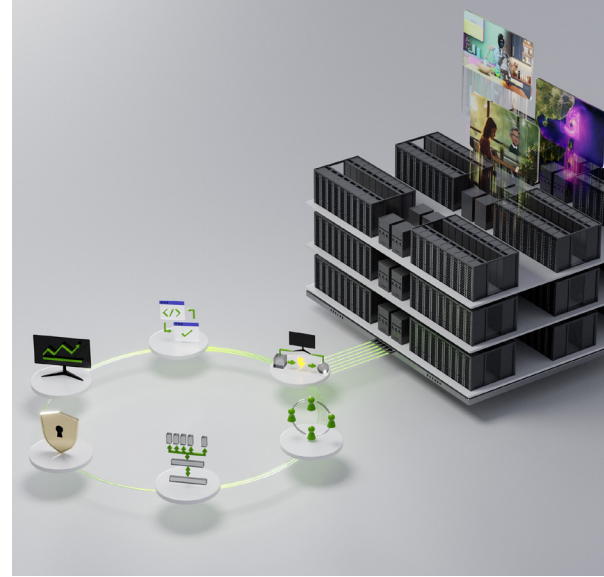




NVIDIA RTX PRO AI Factory

The on-ramp for physical and agentic AI in the enterprise.



Summary

- > The NVIDIA RTX PRO AI Factory is an air-cooled infrastructure design that helps enterprises run AI and accelerated computing workloads in existing data centers.
- > Each [NVIDIA RTX PRO Server](#) supports up to eight [NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs](#) or [NVIDIA RTX PRO 4500 Blackwell Server Edition GPUs](#), giving IT teams a practical unit for scaling inference while managing power, space, GPU use, and cost per token.
- > The full stack combines NVIDIA-Certified Systems, [NVIDIA Spectrum-X](#) networking, [NVIDIA BlueField-3 DPUs](#), [NVIDIA-Certified Storage](#), [NVIDIA AI Enterprise](#), [NVIDIA Omniverse](#), and [NVIDIA Run:ai](#) for secure deployment, policy-based orchestration, monitoring, and production operations.

Recipes for Enterprise AI Factories

NVIDIA provides enterprises with comprehensive guidance for building AI factories from the ground up to deploy at scale. This guidance represents decades of NVIDIA expertise in accelerated computing, networking, and AI software to remove guesswork, improve speed-to-value, optimize performance and TCO, and ensure enterprise-grade security.

The NVIDIA RTX PRO™ AI Factory is designed for enterprise data centers and offers breakthrough AI inference performance across a wide range of enterprise workloads, excelling in agentic AI, physical AI, and visual computing.

This AI factory is based on the infrastructure configurations recommended in [NVIDIA Enterprise Reference Architectures](#), with solutions offered by our global system partners. The NVIDIA RTX PRO AI Factory includes recommendations for configuring NVIDIA Accelerated Computing, NVIDIA networking, NVIDIA-Certified Storage, and NVIDIA software. It leverages an ecosystem of partner hardware rigorously tested and certified by NVIDIA and partner software validated by NVIDIA in enterprise deployments.

NVIDIA RTX PRO AI Factory connects enterprise workloads to accelerated computing, networking, storage, and software.

Key Workloads

- > Enterprise assistants, including deep researchers
- > Secured long-running autonomous agents
- > Video search and summarization
- > Digital twin visualization
- > Synthetic data generation for robotics
- > High-speed rendering
- > Advanced video processing
- > Large-scale data analytics
- > Scientific simulation

Universal Workload Acceleration for Every Enterprise

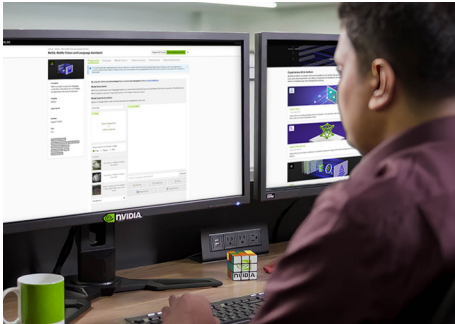
The foundational compute building block for the NVIDIA RTX PRO AI Factory is the [NVIDIA RTX PRO Server](#)—an air-cooled system that requires approximately 7 kW of power, ideal for power-constrained enterprise racks. RTX PRO Servers are x86-based [NVIDIA-Certified Systems™](#) designed for interoperability and compatibility with existing tools and applications commonly used in enterprise data center environments. Each server can support up to eight PCIe-based [NVIDIA RTX PRO 6000 Blackwell Server Edition](#) or [NVIDIA RTX PRO 4500 Blackwell Server Edition](#) GPUs, providing the benefit of configuration flexibility, incremental scaling, and simplified maintenance. [NVIDIA Run:ai](#) simplifies AI workload deployment and GPU orchestration by ensuring optimal GPU utilization and performance.

Powering Agentic and Generative AI at Scale

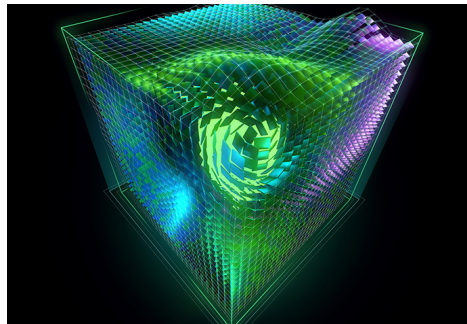
RTX PRO Servers and [NVIDIA AI Enterprise](#) software provide scalable infrastructure for building and deploying advanced AI workloads—including generative models, deep learning, and agentic applications. Powered by NVIDIA Blackwell architecture Tensor Cores and NVIDIA Transformer Engine, these servers deliver major gains in inference throughput, efficiency, and reliability, thus powering the era of useful AI.

From Digital Twins to Autonomous Machines

RTX PRO Servers speed up development and testing for physical AI with NVIDIA Omniverse™ software, enabling large-scale [digital twin modeling](#), video analytics and summarization for industrial operations, and synthetic data generation for robotics and autonomous systems. This boosts perception, decision-making, and task execution for intelligent machines in real-world environments.



Enterprise agentic AI development with NVIDIA software.



Physical AI simulation and digital twin modeling.

Best-in-Class for Visual Computing and Media Processing

RTX PRO Servers offer high-speed rendering, ray tracing, and advanced video processing. This makes them well-suited for professional workflows in simulation, CAD, and media production with demanding graphical requirements.

Furthering Scientific Simulation and Data Analytics

RTX PRO Servers are purpose-built for compute-intensive jobs and scientific simulation workloads via [NVIDIA CUDA-X™](#) library support. With each RTX PRO 6000 Blackwell GPU offering powerful FP32 compute performance, they enable faster results and innovation by accelerating large-scale simulations, data analytics, and predictive modeling for applications in engineering, genomics, fluid dynamics, and research.

NVIDIA RTX PRO AI Factory Bill of Materials

NVIDIA Software

- > [NVIDIA AI Enterprise](#)
- > [NVIDIA Omniverse](#)
- > [NVIDIA Run:ai](#)
- > [NVIDIA NetQ™](#)
- > [NVIDIA RTX™ vWS](#)
- > [vPC](#)

NVIDIA Partner Software

- > Refer to [AI factory design guides](#) for Kubernetes and other ecosystem software components

Optimal Design Points

- > 16 nodes
- > 32 nodes

Accelerated Computing

- > [NVIDIA RTX PRO Servers](#) (nodes)
- > [NVIDIA RTX PRO 6000 Blackwell GPUs](#)

East-West Networking

- > [NVIDIA Spectrum-X™](#)

North-South Networking

- > [NVIDIA BlueField™-3](#)

Cooling

- > Air

Power/GPU

- > 600 W

Storage

- > [NVIDIA-Certified Storage](#)



Photoreal rendering for professional visual computing.

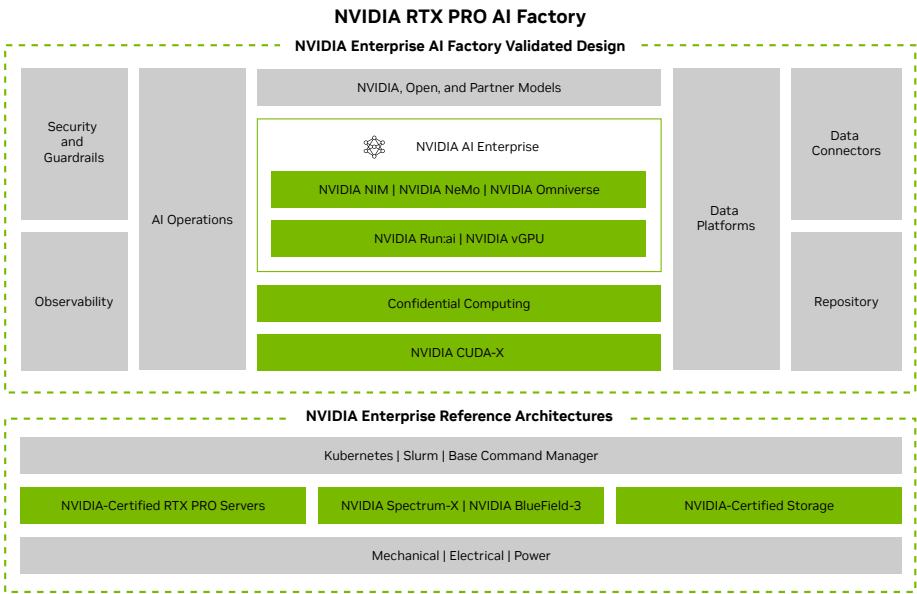


GPU-accelerated engineering and airflow simulation.

A Full-Stack Approach

As part of the NVIDIA AI Factory, [NVIDIA Enterprise Reference Architectures](#) provide recommendations for optimal infrastructure configurations. They are supplemented by a broad ecosystem of partners included in the [NVIDIA Enterprise AI Factory validated design](#) as well as the [NVIDIA AI Factory for Government reference design](#). These software components include tools for AI development and operations, observability, security, and Kubernetes. They enable open source AI models, data, and tools built by NVIDIA, such as the [NVIDIA Nemotron™](#) family for AI reasoning and the NVIDIA Cosmos™ platform for [physical AI](#).

Figure 1: A Full-Stack Approach



NVIDIA RTX PRO AI Factory Full Deck Design

NVIDIA Accelerated Computing

[NVIDIA RTX PRO Servers](#) from global systems partners represent a new category of on-premises enterprise computing, bringing the performance and energy efficiency of NVIDIA Blackwell to a wide range of enterprise use cases. Featuring the latest-generation NVIDIA Blackwell Tensor Cores, fourth-generation RTX capability, and up to 96 GB of GDDR7 memory per GPU, RTX PRO Servers deliver performance for multimodal agentic AI, industrial AI, robotics simulation, design, scientific computing, graphics, and video.

NVIDIA Networking

NVIDIA networking enables high AI performance and scale while ensuring enterprise interoperability and security. The NVIDIA RTX PRO AI Factory includes an East-West AI compute fabric with [NVIDIA Spectrum-X Ethernet switches](#). For the North-South network, [NVIDIA BlueField-3 data processing units](#) optimize data storage access while operating in a zero-trust mode.

NVIDIA-Certified Storage

AI workloads require storage architectures built to handle massive, unstructured, and multimodal datasets. The NVIDIA RTX PRO AI Factory includes [NVIDIA-Certified Storage](#) from leading vendors, validated by NVIDIA for performance and design best practices.

NVIDIA AI and Infrastructure Software

The full capabilities of NVIDIA RTX PRO Servers are unlocked by a comprehensive set of NVIDIA enterprise software for production use. NVIDIA enterprise software streamlines IT operations and drives maximum efficiency for a wide range of enterprise workloads—from agentic AI development and deployment to advanced 3D workflows, industrial digitalization, and GPU-accelerated knowledge worker applications. It includes:

- [NVIDIA AI Enterprise](#): Brings together microservices, frameworks, and libraries for AI development with advanced GPU orchestration and infrastructure management:
 - [NVIDIA Agent Toolkit](#) is an open source development platform that gives developers the building blocks—reasoning models, an agentic harness, tools and CUDA-X skills, and the OpenShell™ secure runtime—to build long-running agents. Security is built into the infrastructure, so enterprises can deploy agents more safely inside the systems where work is done.
 - [NVIDIA Omniverse](#) is a collection of libraries and microservices that enable developers to integrate OpenUSD, NVIDIA RTX rendering, and physical AI simulation into existing industrial and robotics workflows.
 - [NVIDIA Run:ai](#) is an AI workload and GPU orchestration platform that helps enterprises accelerate the AI lifecycle through dynamic resource allocation and policy-driven scheduling to maximize throughput and utilization.
- [NVIDIA RTX Virtual Workstation \(vWS\)](#): Virtualization software that delivers workstation performance for AI-enhanced workloads and compute-intensive applications. [NVIDIA Virtual PC \(vPC\)](#) provides GPU-acceleration for virtual desktop infrastructure.
- [NVIDIA NetQ](#): A highly scalable, modern network operations toolset that provides real-time visibility, troubleshooting, and validation of the fabric, ensuring the AI factory is operating smoothly and at optimal performance.

The NVIDIA RTX PRO AI Factory provides enterprises with a proven, easy on-ramp to realizing the competitive benefits of AI at scale. Designed to help organizations build AI systems from the ground up, it delivers universal workload acceleration across enterprise AI and visual computing workloads. Optimized for modern data centers where space and power are constrained and PCIe-based, air-cooled infrastructure is required, the RTX PRO AI Factory offers performance and flexibility across generative AI, graphics rendering, video search and summarization, and digital twin visualization.

Frequently Asked Questions (FAQ)

> What is the NVIDIA RTX PRO AI Factory?

The NVIDIA RTX PRO AI Factory is an air-cooled, PCIe-based enterprise AI design built on [NVIDIA RTX PRO Servers](#). It supports agentic AI, physical AI, visual computing, analytics, and scientific simulation in existing data centers.

> What hardware does the NVIDIA RTX PRO AI Factory use?

The NVIDIA RTX PRO AI Factory uses x86-based NVIDIA-Certified Systems with up to eight GPUs per server, plus NVIDIA Spectrum-X networking, BlueField-3 DPUs, and NVIDIA-Certified Storage.

> How does the NVIDIA RTX PRO AI Factory help enterprises run autonomous agents securely?

The NVIDIA RTX PRO AI Factory includes the tools needed to run autonomous agents securely at enterprise. The NVIDIA OpenShell secure runtime provides policy-defined sandboxes and controls for agent access to sensitive systems. NVIDIA Confidential Computing protects models and data in use during inference. The [Secure Agent Workspace Reference Design](#) shows enterprises how to deploy autonomous agents securely at enterprise scale.

> How does the NVIDIA RTX PRO AI Factory support token economics?

The NVIDIA RTX PRO AI Factory gives IT teams an air-cooled scale unit with up to eight RTX PRO 6000 Blackwell GPUs and about 7 kW of power per server. Teams can compare token throughput with GPU use, rack power, floor space, and operating cost to manage cost per token.

For interactive agentic AI, IT teams can evaluate time to first token alongside throughput and cost per token.

> Which software is included in the NVIDIA RTX PRO AI Factory stack?

The NVIDIA RTX PRO AI Factory stack includes [NVIDIA AI Enterprise](#), [NVIDIA Agent Toolkit](#), [NVIDIA Omniverse](#), and [NVIDIA Run:ai](#).

Ready to Get Started?

To learn more about the NVIDIA RTX PRO AI Factory, read the [NVIDIA RTX PRO AI Factory Enterprise Reference Architecture](#) and [explore NVIDIA Enterprise AI Factory validated designs and NVIDIA Enterprise Reference Architectures](#).

Or contact your preferred AI factory [systems manufacturer](#) or [specialization partner](#) about AI factory configurations with RTX PRO Servers.

Explore AI Factory configurations on [NVIDIA Marketplace](#).