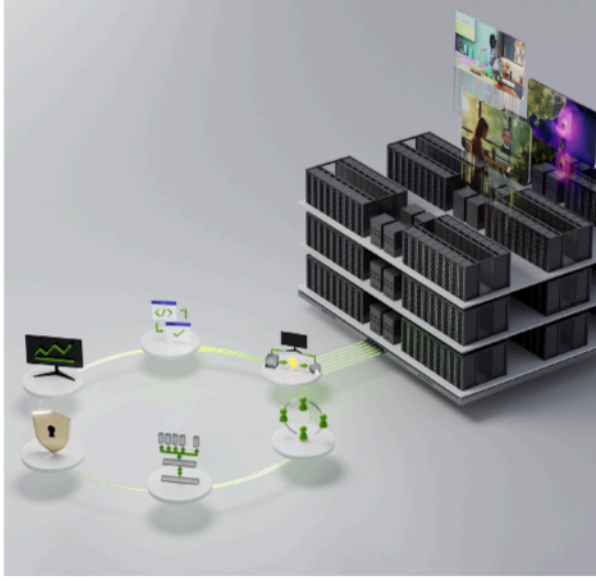
Solution Overview

NVIDIA HGX AI Factory

Breakthrough efficiency and performance for enterprise AI builders.



Key Takeaways

- The NVIDIA HGX™ AI Factory delivers scalable performance for model training, fine-tuning, and enterprise inference.
- The NVIDIA HGX B300 powers copilots, reasoning systems, and autonomous agents at scale.
- The NVIDIA HGX B300 provides high memory and processing power, enabling organizations to run complex AI models directly without needing to manage memory offloading.
- The NVIDIA HGX B300 trains LLMs 2.6x faster and provides up to 15x higher token throughput than the prior generation, helping enterprises scale inference efficiently as token demand grows.

Key Workloads

- Copilots
- Reasoning systems
- Autonomous agents
- Enterprise context (RAG)
- Analytics and decisioning
- Industry AI solutions
- HPC and simulations

Recipes for Enterprise AI Factories

NVIDIA provides enterprises with comprehensive guidance to build AI factories from the ground up and 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 total cost of ownership, and ensure enterprise-grade security.

The NVIDIA HGX AI Factory is engineered for enterprises that need scalable performance across medium- to large-scale model training, foundation model fine-tuning, and enterprise-grade inference for copilots, reasoning systems, and autonomous agents. As inference becomes the dominant enterprise AI workload and agentic AI drives token demand higher, HGX helps organizations deliver AI at scale.

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 HGX 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.

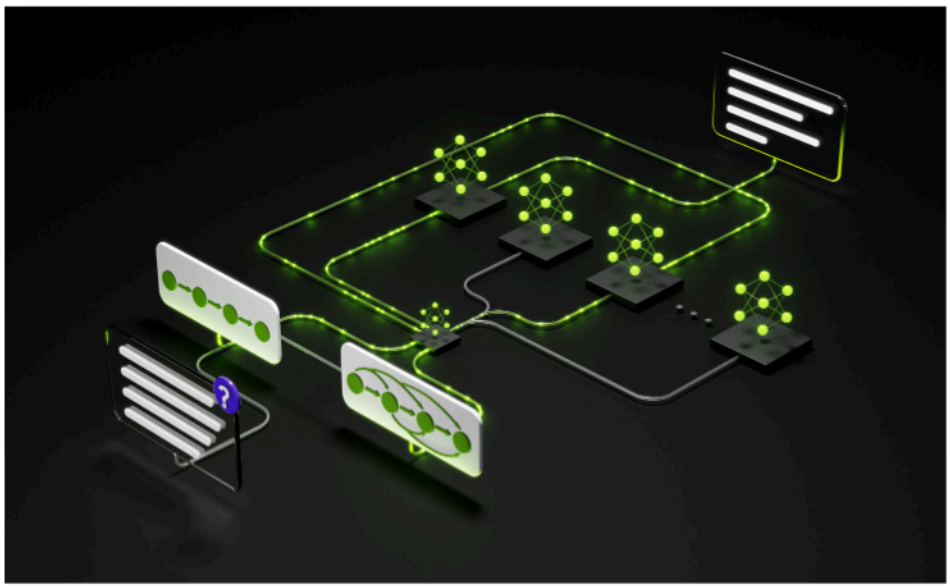
Breakthrough Performance for Enterprise AI

Tailored to Enterprise AI Builders

The NVIDIA HGX AI Factory supports training and fine-tuning large language models (LLMs) for enterprises that want to maximize AI's competitive advantage by leveraging their proprietary data. The accelerated computing relies on a second-generation Transformer Engine, which enables a remarkable 2.6x faster training for very large language models, such as DeepSeek-R1. At the orchestration layer, NVIDIA Run:ai helps schedule distributed GPU clusters and manage compute resources for large model training jobs. At the low-level programming layer, NVIDIA cuDNN™ libraries provide hand-optimized GPU kernels for compute-heavy operations. At the AI operations layer, the NVIDIA NeMo™ suite of software tools supports the whole AI agent lifecycle, including data curation, model customization, and the development of agentic AI applications.

Ultra-Fast Inference for Enterprise Reasoning

As organizations operationalize copilots, reasoning systems, and autonomous agents, inference is becoming the dominant enterprise AI workload in today's data centers. NVIDIA HGX AI Factory delivers unparalleled inference performance to meet this growing demand. It offers x86 compatibility and is available in an air-cooled configuration, allowing for interoperability and easy scalability from existing infrastructure. The NVIDIA HGX B300 achieves a 15x relative speed-up as measured in tokens per second in 1 MW of power, helping enterprises maximize token throughput and improve inference economics. The system's 2.3 TB of HBM3E memory per node allows it to run 500B+ parameter models, such as the NVIDIA Nemotron™ 3 Ultra, without memory offloading.



Optimized for Large-Scale Data Analytics

As autonomous agents draw on increasingly complex enterprise context, quickly processing and analyzing large volumes of data has become a critical part of the AI life cycle. The HGX B300 is designed to handle massive datasets, facilitate extract, transform, and load (ETL) processes, and deliver GPU-accelerated big data analytics.

NVIDIA HGX AI Factory

Agentic AI Software

- NVIDIA AI Enterprise
- NVIDIA CUDA-X + software solutions from NVIDIA partners

Orchestration Software

- NVIDIA Run:ai
- NVIDIA Net-Q™
- Kubernetes
- Slurm

Optimal Design Points

- 32 nodes
- 64 nodes
- 128 nodes

Accelerated Computing

- 8x NVIDIA Blackwell Ultra GPUs per node in NVIDIA-Certified Servers

East-West Networking

- 8x NVIDIA ConnectX™-8 SuperNICs™ per node

North-South Networking

- 1x BlueField™-3 DPU per node

Cooling

- Air or liquid

Power/GPU

- Up to 1,100 W (configurable)

Storage

- NVIDIA-Certified Storage

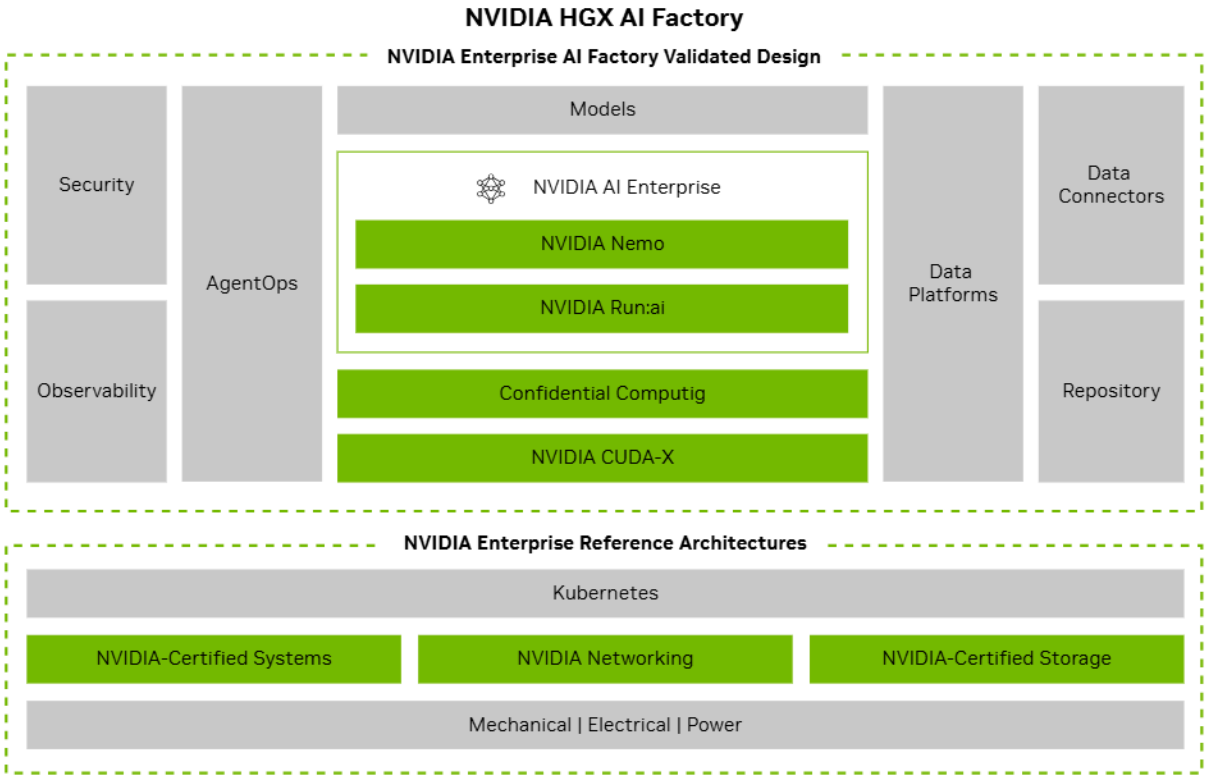
Projected performance subject to change. Inference in FP4 for B300, inference in FP8 for H100, factory inference with DeepSeek-R1, ISL:32K, OSL: 8K, multi-node disaggregated inference with NVIDIA Dynamo. H100 with 400 Gbps CX-7 NIC (Ethernet), B300 with 800 Gbps CX-8 NIC (Ethernet).

Ultra-fast interconnects are enabled between GPU and CPU with the PCIe Gen5 x16, and GPU to GPU by the fifth-generation NVIDIA NVLink™, thus limiting data transfer bottlenecks. Combined with NVIDIA CUDA-X™ libraries, this brings acceleration to a broad range of analytical applications, including decision optimization with NVIDIA cuOpt™ and dense linear algebra with NVIDIA cuSOLVER™.

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. Best-practice configurations are available for 32, 64, and 128 nodes, in both single-plane and dual-plane networking topologies.

Figure 1: A Full-Stack Approach



NVIDIA HGX AI Factory full-stack design

NVIDIA Accelerated Computing

NVIDIA HGX B300 servers from global system providers are built to usher in the age of AI reasoning. Each HGX B300 server features eight NVIDIA Blackwell Ultra GPUs. NVIDIA Blackwell Ultra offers 7x more AI compute than NVIDIA Hopper™ platforms and delivers up to 20 petaFLOPS of FP4 sparse inference performance, providing efficiency for large-scale deployments. With 288 GB of HBM3E memory, it supports expansive KV caching and long-context inference without offloading.

NVIDIA Networking

NVIDIA NVLink accelerates the multi-GPU communication required for reasoning models and autonomous agents, helping deliver higher token throughput and faster responses at scale.

The NVIDIA Spectrum-X™ Ethernet platform, featuring NVIDIA Spectrum SN5610 switches and NVIDIA ConnectX-8 SuperNICs, is the world's first end-to-end, AI-optimized Ethernet platform for multi-tenant generative AI factories.

The networking configuration includes an East-West AI compute fabric with NVIDIA Spectrum-X and NVIDIA ConnectX-8 SuperNICs, which offer up to 800 Gb/s of throughput in a dual plane topology, doubling interconnect bandwidth compared to previous generations to enable seamless scaling across data centers. Additionally, fifth-generation NVLink interconnects provide 1,800 GB/s of GPU-to-GPU connection, reducing data transfer times.

For the North-South network, the NVIDIA BlueField-3 data processing unit offers automated provisioning and elasticity and optimizes data storage access (400 Gb/s) 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 HGX 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 HGX B300 are unlocked by a comprehensive set of NVIDIA software to streamline IT operations, drive maximum efficiency, and guarantee enterprise-grade security. The software includes:

- NVIDIA AI Enterprise: a suite of software tools, libraries and frameworks, including:
  - NVIDIA NIM™ microservices and NeMo toolkit that accelerate and simplify the development, deployment, and scaling of AI applications. The suite also includes world-class NVIDIA Nemotron models delivered as easy-to-use NIM microservices.
  - NVIDIA Run:ai, an AI workload and GPU orchestration platform, helps enterprises accelerate the AI lifecycle through dynamic resource allocation and policy-driven scheduling to maximize throughput and utilization.
- NVIDIA CUDA-X libraries, built on NVIDIA CUDA™, deliver dramatically higher performance across application domains, including AI and HPC. Through domain-specific optimizations, they enable breakthroughs in engineering, research, and scientific discovery. By leveraging the parallel processing power of GPUs, CUDA-X libraries accelerate computations significantly faster than traditional methods, with continuous NVIDIA optimization further boosting performance.
- NVIDIA NetQ, a highly scalable, modern network operations toolset that provides visibility, troubleshooting, and validation of the fabric in real time, ensuring the AI factory is operating smoothly without interruption and delivering optimal performance.

NVIDIA AI Factories offer enterprises an easy on-ramp to reap the benefits of AI by allowing them to build an AI system from the ground up and at scale. The NVIDIA HGX AI Factory offers breakthrough performance for AI power users and builders. It enables training and fine-tuning of large language models while dramatically accelerating inference for enterprise AI applications. As agentic AI increases token demand, HGX B300 helps organizations maximize the value created from every token while improving the efficiency of producing them.

Frequently Asked Questions (FAQ)

- **What is the NVIDIA HGX AI Factory?** The NVIDIA HGX AI Factory is an enterprise full-stack guidance built on NVIDIA Enterprise Reference Architectures and offered through global system partners, engineered for scalable medium-to-large model training, foundation model fine-tuning, and enterprise-grade inference.
- **What accelerated compute powers the NVIDIA HGX AI Factory?** The NVIDIA HGX AI Factory runs on the NVIDIA HGX B300 server with 8 NVIDIA Blackwell Ultra GPUs per node. Each NVIDIA Blackwell Ultra GPU provides 288 GB of HBM3E for 2.3 TB per node and can run 500B+ parameter models, such as NVIDIA Nemotron 3 Ultra, without offloading.
- **What networking and storage does the NVIDIA HGX AI Factory use?** The NVIDIA HGX AI Factory uses NVIDIA Spectrum-X Ethernet with SN5610 switches and NVIDIA ConnectX-8 SuperNICs, supporting up to 800 Gb/s in a dual-plane design. North-South traffic runs over the BlueField-3 DPU at 400 Gb/s with zero-trust security, paired with NVIDIA-Certified Storage.
- **How much faster is the NVIDIA HGX AI Factory for training and inference?** The NVIDIA HGX AI Factory trains very large LLMs such as DeepSeek-R1 2.6x faster using the second-generation Transformer Engine. For inference, the NVIDIA HGX B300 achieves a 15x relative speed-up in tokens per second per megawatt and 15x higher token throughput than the HGX H100 through FP6 and FP4 precision.
- **What software stack runs on the NVIDIA HGX AI Factory?** The NVIDIA HGX AI Factory runs NVIDIA AI Enterprise, including NIM, NeMo, and Nemotron, alongside NVIDIA Run:ai, CUDA-X, and NetQ. For orchestration, the NVIDIA HGX AI Factory supports both Kubernetes and Slurm.

Ready to Get Started?

To learn more about the NVIDIA HGX AI Factory, read the NVIDIA HGX AI Factory Enterprise Reference Architecture Whitepaper, the NVIDIA Enterprise AI Factory Validated Design, and the AI Factory for Government Reference Design.

Or contact your preferred systems manufacturer or channel partner about their NVIDIA-endorsed AI factory solution including the NVIDIA HGX B300.

Partner Logo

NVIDIA