



NVIDIA NVL72 AI Factory

The best of NVIDIA Blackwell performance at frontier scale.



Summary

- > The NVIDIA NVL72 AI Factory is a liquid-cooled, rack-scale reference architecture built on NVIDIA GB300 NVL72 for serving inference, data analytics, and HPC workloads at massive scale and training trillion-parameter models.
- > It delivers up to a 50x leap in AI factory output, 10x faster user responsiveness, and 5x higher throughput per megawatt versus the prior NVIDIA Hopper™ generation.
- > Each rack connects 36 NVIDIA Grace™ CPUs and 72 NVIDIA Blackwell Ultra GPUs in a 72-GPU NVLink™ domain, paired with NVIDIA Spectrum-X™ networking, NVIDIA Dynamo, and NVIDIA Mission Control™ software.

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 NVL72 AI Factory offers the best token efficiency for enterprises training and deploying AI at massive scale. As inference becomes the dominant enterprise, AI workload and agentic AI drives exponential growth in token demand, NVL72 helps organizations maximize the value created from tokens while optimizing the cost and efficiency of producing them. 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 NVL72 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.

Key Workloads

- > Training for trillion-parameter models
- > Large-scale real-time inference
- > Autonomous agents
- > Reasoning systems
- > Retrieval-augmented generation (RAG)
- > High-performance data analytics
- > Scientific simulation and HPC workloads

Exceptional Performance at Frontier Scale

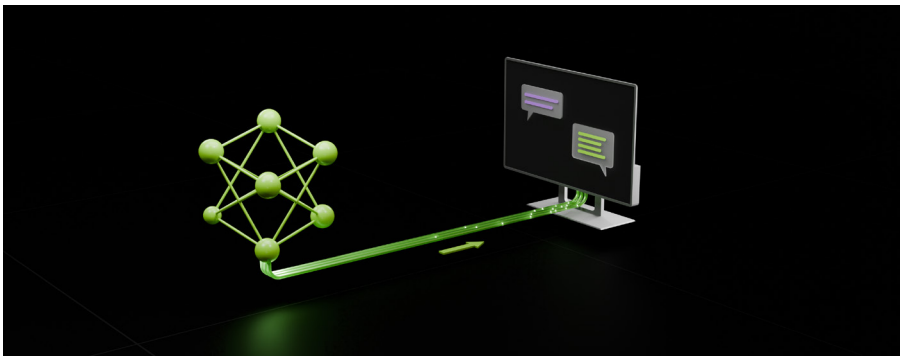
The NVIDIA NVL72 AI Factory includes NVIDIA accelerated computing, NVIDIA Networking, NVIDIA-Certified Storage, and NVIDIA orchestration and observability software. It leverages a trusted ecosystem of partner solutions along the AI stack, all vetted and tested by NVIDIA. These capabilities provide the foundation for enterprise AI applications ranging from large-scale training to inference-intensive agentic AI workloads that require rapid reasoning, retrieval, and multi-step execution.

Best Token Efficiency Performance

The NVIDIA NVL72 AI Factory reduces carbon footprint and energy consumption for large-scale data center deployments. Liquid cooling increases compute density and reduces floor space used. Large [NVLink](#) domain architectures facilitate high-bandwidth, low-latency GPU communication. With [NVIDIA GB300 NVL72](#), this AI factory configuration delivers a 10x boost in user responsiveness (TPS per user) and a 5x improvement in throughput (TPS per megawatt), translating into a 50x leap in overall AI factory output and dramatically reducing total cost of ownership, particularly for 24/7 model training and inference environments supporting copilots and autonomous agents.

Training Trillion-Parameter Models

The NVIDIA NVL72 AI Factory supports the training and inference for trillion-parameter models, which stretch the computational resources of conventional hardware. Trillion-parameter models maintain coherence across long context windows or conversations, digest enormous datasets in which they detect subtle relationships, and perform cross-domain analysis. Very large mixture-of-experts (MoE) LLMs can address queries with large knowledge breadth, long-tail factual knowledge, and multimodal reasoning. Graph neural networks solve problems in manufacturing optimization, financial risk analysis, or customer behavior analysis. The training of very large language models is further supported by [NVIDIA cuDNN™](#) libraries.



Real-Time Inference

Real-time inference is rapidly becoming the dominant enterprise AI workload as organizations operationalize copilots, reasoning systems, retrieval pipelines, and autonomous agents. The NVIDIA NVL72 AI Factory enables real-time LLM inference with NVFP4 precision, which radically reduces memory bandwidth and compute without sacrificing the model quality. Moreover, the full-stack design leverages the TensorRT™ library, which achieves sub-millisecond latency for batched inference workloads with NVIDIA GB300 NVL72. NVIDIA Dynamo further improves system-level inference performance through disaggregated serving.

NVIDIA NVL72 AI Factory Bill of Materials

NVIDIA Software

- > [NVIDIA Mission Control](#)
- > [NVIDIA Dynamo](#)
- > [NVIDIA CUDA-X™](#)

Optimal Design Points

- > 2, 4, 8 racks

Accelerated Computing

- > [72x NVIDIA Blackwell Ultra GPUs per rack in NVIDIA-Certified Servers](#)

East-West Networking

- > [72x NVIDIA ConnectX™](#)
- > 8 SuperNICs™ per rack adapters

North-South Networking

- > [18x NVIDIA BlueField™-3 DPU per rack](#)

Cooling

- > Liquid

Power/GPU

- > Up to 1,400 W (configurable)

Storage

- > [NVIDIA-Certified Storage](#)

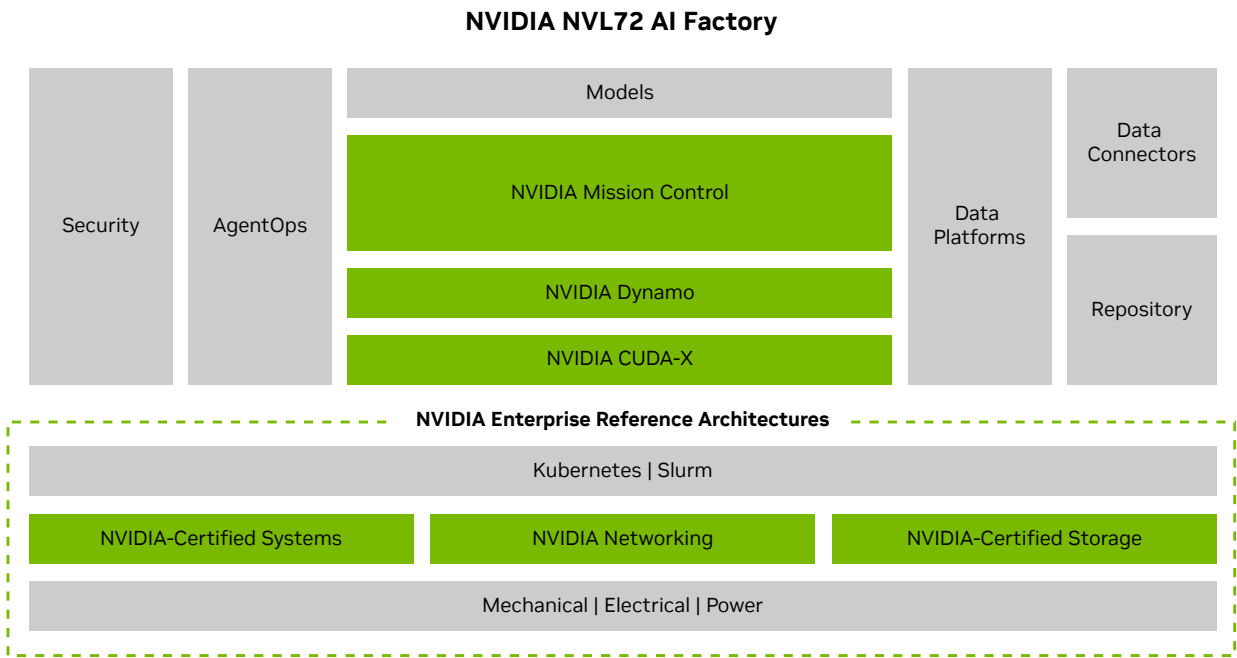
Ultra-Fast Data Processing

Handling, processing, and analyzing large volumes of data is a critical step of the AI lifecycle for any organization. Agentic AI workloads depend on fast access to enterprise context, making data processing and retrieval increasingly important contributors to both inference performance and token efficiency. NVIDIA NVL72 AI Factory speeds up database queries by 6x compared to NVIDIA HGX™ H100, and by 18x compared to x86 CPUs. These gains are enabled by a hardware decompression engine that natively decompresses data at scale and is enabled by the [NVIDIA nvCOMP™](#) library. Database queries are further accelerated by the Grace CPU high-speed [NVLink Chip-to-Chip \(C2C\) interconnect](#), along with 8 TB/s of high-memory bandwidth.

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 NVIDIA NVL72 AI Factory configurations are available for 2-, 4-, and 8-rack versions, with a dual-plane networking topology.

Figure 1: A Full-Stack Approach



NVIDIA NVL72 AI Factory full-stack design

NVIDIA Accelerated Computing

NVIDIA GB300 NVL72 connects 36 Grace CPUs and 72 NVIDIA Blackwell GPUs in a rack-scale, liquid-cooled design. It boasts a 72-GPU NVIDIA NVLink domain that delivers dramatically faster LLM inference. The NVIDIA GB300 Grace Blackwell Superchip is a key component of the system, coherently connecting two high-performance [NVIDIA Blackwell GPUs](#) and an [NVIDIA Grace CPU](#) using the NVLink-C2C interconnect to the two NVIDIA Blackwell GPUs. [NVLink-C2C interconnect](#) delivers up to 7x bandwidth compared to PCIe Gen 5.

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 SN5000 switches and [SuperNICs](#), is the world's first end-to-end, AI-optimized Ethernet platform for multi-tenant generative AI factories.

The NVIDIA NVL72 AI Factory includes an East-West AI compute fabric with NVIDIA [Spectrum-X Ethernet switches](#) and NVIDIA [ConnectX-8 NICs](#), which offer up to 800 Gb/s of throughput.

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 NVL72 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 GB300 NVL72 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 CUDA-X](#) libraries, built on CUDA™, are a collection of libraries that deliver dramatically higher performance across application domains, including AI and HPC. CUDA-X libraries enable breakthroughs in engineering, research, and scientific discovery through domain-specific optimizations. By leveraging the parallel processing power of GPUs, CUDA-X libraries accelerate computations significantly faster than traditional methods. NVIDIA's continuous optimization of these libraries boosts performance.
- [NVIDIA Dynamo](#), the successor to NVIDIA Triton Inference Server™, is an open source AI inference-serving software designed to maximize token revenue generation for AI factories deploying reasoning AI models. It orchestrates and accelerates inference communication across thousands of GPUs, and uses disaggregated serving to separate the processing and generation phases of large language models (LLMs) on different GPUs.
- [NVIDIA Mission Control](#) software simplifies how AI factories are deployed and operated throughout the entire cluster lifecycle, including bring-up, provisioning, and integration with building management systems. It also provides seamless workload orchestration leveraging NVIDIA Run:ai technology and optimizes AI factory infrastructure resiliency with extensive observability dashboards, continuous health checks, and built-in autonomous recovery capabilities at both the hardware and application layer. Enterprises accelerate the AI lifecycle through dynamic resource allocation and policy-driven scheduling to maximize throughput and utilization.

Frequently Asked Questions (FAQ)

- > **What is the NVIDIA NVL72 AI Factory?** A liquid-cooled, rack-scale reference architecture for training and deploying AI at massive scale, enabling real-time applications and trillion-parameter models.
- > **What hardware does the NVIDIA NVL72 AI Factory use?** It is built on NVIDIA GB300 NVL72, connecting 36 Grace CPUs and 72 NVIDIA Blackwell Ultra GPUs per rack in a 72-GPU NVLink domain.
- > **What performance gains does the NVIDIA NVL72 AI Factory deliver?** It delivers up to a 50x leap in AI factory output, 10x faster user responsiveness, and 5x higher throughput per megawatt versus H100; database queries run 6x faster than HGX H100 and 18x faster than x86 CPUs.
- > **What workloads is it built for?** Inference at massive scale, trillion-parameter model training, real-time inference, ultra-fast data analytics, conversational AI, and simulation.
- > **What software comes with it?** It includes NVIDIA CUDA-X libraries, NVIDIA Dynamo, and NVIDIA Mission Control.

Ready to Get Started?

To learn more about the NVIDIA NVL72 AI Factory, read the [NVIDIA NVL72 AI Factory Enterprise Reference Architecture](#), the [NVIDIA Enterprise AI Factory validated design](#), and the [NVIDIA AI Factory for Government reference design](#). Or contact your preferred systems manufacturer or channel partner about their NVIDIA-endorsed AI factory solutions including NVIDIA GB300 NVL72.

