

Cisco Secure AI Factory with NVIDIA: Rack-Scale and dense AI for the Enterprise



The platform you pilot on should be the platform you scale on, and the team you already have should be able to run it. Cisco Secure AI Factory with NVIDIA now spans a single GPU server to rack-scale, liquid-cooled AI factories, managed through Cisco Cloud Control alongside the rest of your data center.

Figure 1. Rack-scale, liquid-cooled AI capacity deploys inside the same Cisco Secure AI Factory with NVIDIA architecture that enterprises already run

Overview

Enterprise AI has left the lab. Agentic workflows, copilots, and retrieval-augmented generation (RAG) are moving into production, and the infrastructure question has changed with them. Gartner forecasts worldwide AI spending of \$2.59 trillion in 2026, with AI infrastructure the largest segment of the market. Yet most corporate estates were never designed for dense, liquid-cooled GPU systems, and the harder problem is operational: a step to this level of scaling usually means new vendors, new tools, a new security model, and a team you do not have.

Cisco Secure AI Factory with NVIDIA, now expanded with rack-scale compute built by Supermicro, gives you one validated path from first pilot to production AI factory. NVIDIA GB300 NVL72, Vera Rubin NVL72, HGX, and MGX systems join the architecture you already deploy: connected by Cisco® fabrics, secured by Cisco AI Defense, Cisco Hypershield™, and Isovalent®, observed through Splunk®, and operated through Cisco Cloud Control. Cisco Validated Infrastructure Services (CVIS) validate that each build matches its intended design. You choose the entry point. The operating model stays the same all the way up.

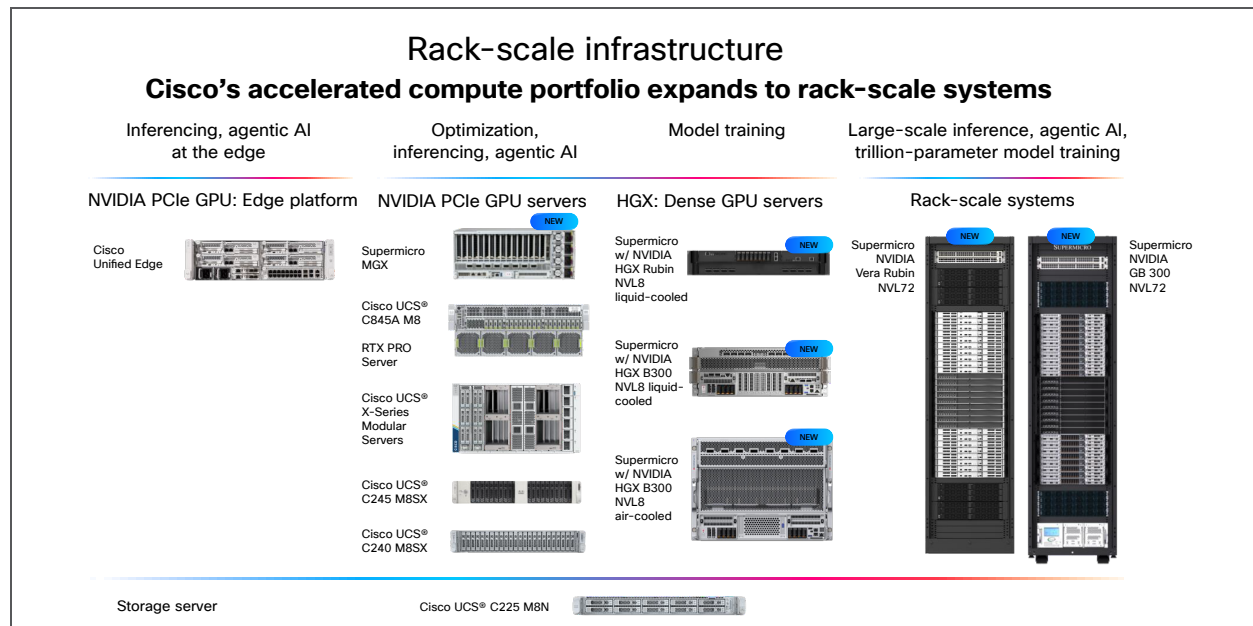


Figure 2. The enterprise portfolio scales from a first GPU server to a rack-scale AI factory on one validated architecture

Benefits

- **Start where you are.** Enter with a Cisco AI Server for NVIDIA MGX for early agentic and RAG workloads, then grow into NVIDIA HGX systems and rack-scale NVL72 factories on the same architecture, with Cisco UCS® and Cisco AI PODs still covering the workloads you run today.
- **Run AI as part of the estate, not beside it.** Cisco Cloud Control brings server lifecycle, liquid cooling and power, and network connectivity and policy into one console, one inventory, and one topology, so AI clusters and non-AI workloads share an operating model.
- **Prove the build before you depend on it.** CVIS qualifies the full-stack design, verifies the deployed configuration against performance and job completion targets, and delivers an evidence report at handover.
- **Protect your inference economics.** Cisco Silicon One® frontend and NVIDIA Spectrum-X Ethernet backend fabrics keep GPUs fed, so cost per token stays a business decision rather than a congestion problem.
- **Secure AI from day one.** Cisco AI Defense, Cisco Hypershield, Isovalent®, and Cisco Secure Firewall protect models, agents, workloads, and infrastructure as part of the design, with Splunk correlating security posture against infrastructure telemetry.

Trends and challenges

Infrastructure is now the AI budget

Gartner forecasts worldwide AI spending of \$2.59 trillion in 2026, a 47 percent increase year over year, and expects AI infrastructure to become the largest segment of that market at over 45 percent of spending. When AI answers customers and runs internal workflows, latency and cost per token become numbers the business tracks, not research metrics.

The estate is not AI-ready

Conventional air-cooled racks are designed for roughly 30 to 50 kW. An NVIDIA NVL72-class rack can exceed 200 kW, which makes direct liquid cooling a physical requirement rather than a preference. Facility decisions now come before hardware orders, and many teams are sequencing that for the first time.

The operating model is the real gate

Performance you can buy. Operations you cannot. Compute from one vendor, fabric from another, and security added afterward means every integration seam extends deployment, and every idle week on NVIDIA GB300-class capacity burns capital that should already be producing results. The scarce thing is not GPUs. It is a full stack that your existing team can secure, observe, and operate.

How it works

One factory, every layer

Cisco Secure AI Factory with NVIDIA is a validated full stack: NVIDIA-accelerated compute built by Supermicro, Cisco frontend and backend fabrics, an AI data platform, security fused from silicon to agents, Splunk observability, and unified operations through Cisco Cloud Control. Deployments align to the NVIDIA Enterprise Reference Architecture for enterprise builds and to NCP-RA where you scale into cloud-style capacity, then are validated as Cisco Validated Designs with Cisco networking, security, and management layered in.

What you buy: four entry points, one architecture

- **On-ramp:** Cisco AI server for MGX with NVIDIA GPUs. Air-cooled PCIe GPU server for RAG, copilots, departmental inference, vGPU, and visual computing, with fast delivery times. This is the pragmatic first step.
- **Facility bridge:** Cisco AI server for NVIDIA HGX B300, air-cooled. An 8RU eight-GPU NVIDIA Blackwell Ultra system for sites without a liquid loop, so the project keeps moving while facility work catches up.
- **Dense AI factory:** Cisco AI server for NVIDIA HGX B300, liquid-cooled, and Cisco AI server for NVIDIA HGX Rubin NVL8. 4RU and 2RU eight-GPU building blocks for training, fine-tuning, agent fleets, simulation, and digital twins where a liquid loop exists or is funded.
- **Rack scale:** Cisco AI rack for NVIDIA GB300 NVL72. 72 NVIDIA Blackwell Ultra GPUs and 36 Grace CPUs in one liquid-cooled NVIDIA NVLink domain for extreme model scale and high token demand. Cisco AI rack for NVIDIA Vera Rubin NVL72 follows on the same pod design, so the facility, fabric, and operating model you build today will carry forward.

Cisco UCS remains the foundation for the workloads you run every day, and Cisco Unified Edge extends the same model to distributed inference. You are matching a tier to a workload, not choosing between portfolios. All platforms above are orderable from Cisco authorized channel partners in October 2026.

A fabric built for AI economics

Cisco N9300 Series Smart Switches built on Cisco Silicon One carry frontend and storage traffic. Cisco N9100 Series Switches built on NVIDIA Spectrum-X Ethernet silicon carry lossless backend traffic, with liquid-cooled Cisco N9000 Series Switch options for the densest rows so the fabric shares the thermal envelope of the compute it serves. Choose Cisco NX-OS for enterprise operational consistency or SONiC where it fits, on the same validated hardware, unified through Cisco Nexus® One.

Operated as one system through Cisco Cloud Control

Cisco Cloud Control delivers unified management and AgenticOps across the stack, from rack-scale core to edge. It brings server lifecycle, liquid cooling and power, and network connectivity and policy into one console, one inventory, and one topology, so operators and agents can correlate an issue across domains, move from alert to evidence-backed action, and apply policy consistently from day 0 onward. Network management is enabled today by Cisco Nexus One. Management of compute systems will be available through Cisco Intersight® and Cisco Nexus One in CY26Q4.

Secured and observed at every layer

Cisco AI Defense validates models and enforces runtime guardrails, integrated with NVIDIA AI Enterprise components including NIM and NeMo. Hypershield and Isovalent segment and protect east/west traffic between GPU nodes, and Cisco Secure Firewall guards the boundary. Splunk observability correlates job health with compute, NICs, optics, and network performance, so when a run degrades your team can tell whether the cause is the GPU, the NIC, an optic, or the fabric instead of triaging across separate tools.

Validated before handover with CVIS

Cisco Validated Infrastructure Services (CVIS) are an end-to-end architecture qualification and validation program with an automation toolkit that reduces deployment and validation timelines from months to weeks. Every deployment starts from a qualified reference design; every deployed configuration is verified against reference architecture metrics such as infrastructure performance, job completion time, and token throughput; and you receive an evidence report covering validated configuration and test results. CVIS is anchored by a dedicated large-scale engineering AI cluster in Cisco's own AI lab, where deployment tooling, performance profiles, and software releases are continuously validated.

Shorten time to first intelligence with Cisco CX

Cisco Customer Experience (Cisco CX) and certified partners cover the full lifecycle: plan, design, implement, validate, knowledge-transfer, optimize, and scale out. Facility and AI readiness assessments settle the liquid-cooling question before any hardware order, and the customer facility questionnaire and compliance site survey are pre-sales activities carried out with your team. CVIS is the named delivery and validation motion for the AI factory itself, delivered by the Cisco CVIS team with channel partners authorized by Cisco, and ending in an evidence report your architects can review.



“Performance you can buy.
Operations you cannot. The
scarce thing is not GPUs, it is
a full stack your existing team
can secure, observe, and
operate.”

“Compliance at the
architecture level is not proof
at the deployment level.
CVIS verifies the cluster you
actually built and hands you
the evidence.”

Prashant Kalika,
Vice President, Product Management
Networking Infrastructure, Cisco Systems, Inc.

Use cases

Use case	What it looks like in practice
Agentic workflows and copilots	Automation across finance, R&D, sales, and support, grounded in enterprise systems and protected by runtime guardrails
RAG and knowledge retrieval	Answers grounded on private corporate data, with the inference path secured, and observable from end to end
Document and content processing	Summarization, extraction, and drafting at production scale with predictable latency and cost per token
Forecasting and analytics	Demand, risk, and operations modeling on dedicated GPU capacity instead of contended, shared infrastructure
Simulation, digital twins, and VDI	Physics, rendering, and virtual desktop workloads on NVIDIA MGX and HGX platforms, sharing the same fabric and operating model

Cisco Capital

Financing to Help You Achieve Your Objectives

Cisco Capital can help you acquire the technology you need to achieve your objectives and stay competitive. We can help you reduce CapEx. Accelerate your growth. Optimize your investment dollars and ROI. Cisco Capital financing gives you flexibility in acquiring hardware, software, services, and complementary third-party equipment. And there’s just one predictable payment. Cisco Capital is available in more than 100 countries. Learn more <http://www.cisco.com/web/ciscocapital/index.html>.

Liquid cooling decides your timeline

Rack-scale AI has crossed the liquid threshold. Conventional air-cooled racks are designed for roughly 30 to 50 kW, while an NVIDIA NVL72 class rack can exceed 200 kW, so direct liquid cooling becomes a system-level requirement rather than a preference. The first question in any enterprise AI-capacity plan is now whether you have a liquid loop today, or a funded timeline to one.

- **No liquid loop yet:** The air-cooled NVIDIA HGX B300 server and NVIDIA MGX PCIe GPU server keep the project moving inside your current power and cooling envelope, and the Cisco AI POD portfolio remains fully air-cooled.
- **Loop in place or funded:** Liquid-cooled NVIDIA HGX systems and NVIDIA GB300 NVL72 racks unlock full density, with liquid-cooled Cisco N9000 Series Switches interoperating directly with the compute.
- **First step:** Get a Cisco CX or partner AI and facility readiness assessment, before any hardware order.

The Cisco Advantage

Cisco is the first NVIDIA technology partner to deliver an NVIDIA Cloud Partner-compliant reference architecture built on partner-developed networking systems, spanning Cisco Silicon One and NVIDIA Spectrum-X Ethernet switch silicon under a single Cisco Nexus One architecture. Add to your enterprise 40 years of networking and security operations trusted by roughly 300,000 organizations, with security fused into every layer, CVIS validation with an evidence report, and one partner to quote, deploy, finance, and support your AI factory.

Move your AI to production without a new operating model.

See how Cisco Secure AI Factory with NVIDIA extends to rack-scale, liquid-cooled compute. Explore the architecture on the [Cisco and Supermicro partnership page](#), read the [rack-scale architecture FAQ](#), or contact your Cisco account team to request an AI and facility readiness assessment.