

Cisco Secure AI Factory with NVIDIA: Compute Expansion to Rack-Scale



Product overview

Rack-scale AI, validated and operated as one system

AI has moved past single servers and small pods. Enterprises pushing pilots into production, neoclouds selling GPU hours, and sovereign programs building national capacity all need dense, liquid-cooled infrastructure that reaches production quickly. Assembling the infrastructure from separate compute, fabric, and security vendors adds integration seams, and every week a rack that sits unproven is capacity earning nothing.

Cisco Secure AI Factory with NVIDIA now extends to rack-scale, liquid-cooled AI infrastructure. Working with Supermicro, Cisco adds NVIDIA GB300 NVL72, Vera Rubin NVL72, HGX, and MGX systems to the same architecture, connected by Cisco fabrics, secured and observed at every layer, and operated through Cisco Cloud Control. Cisco Validated Infrastructure Services then confirm that the deployed cluster matches the design. All platforms are orderable from Cisco authorized channel partners in October 2026.

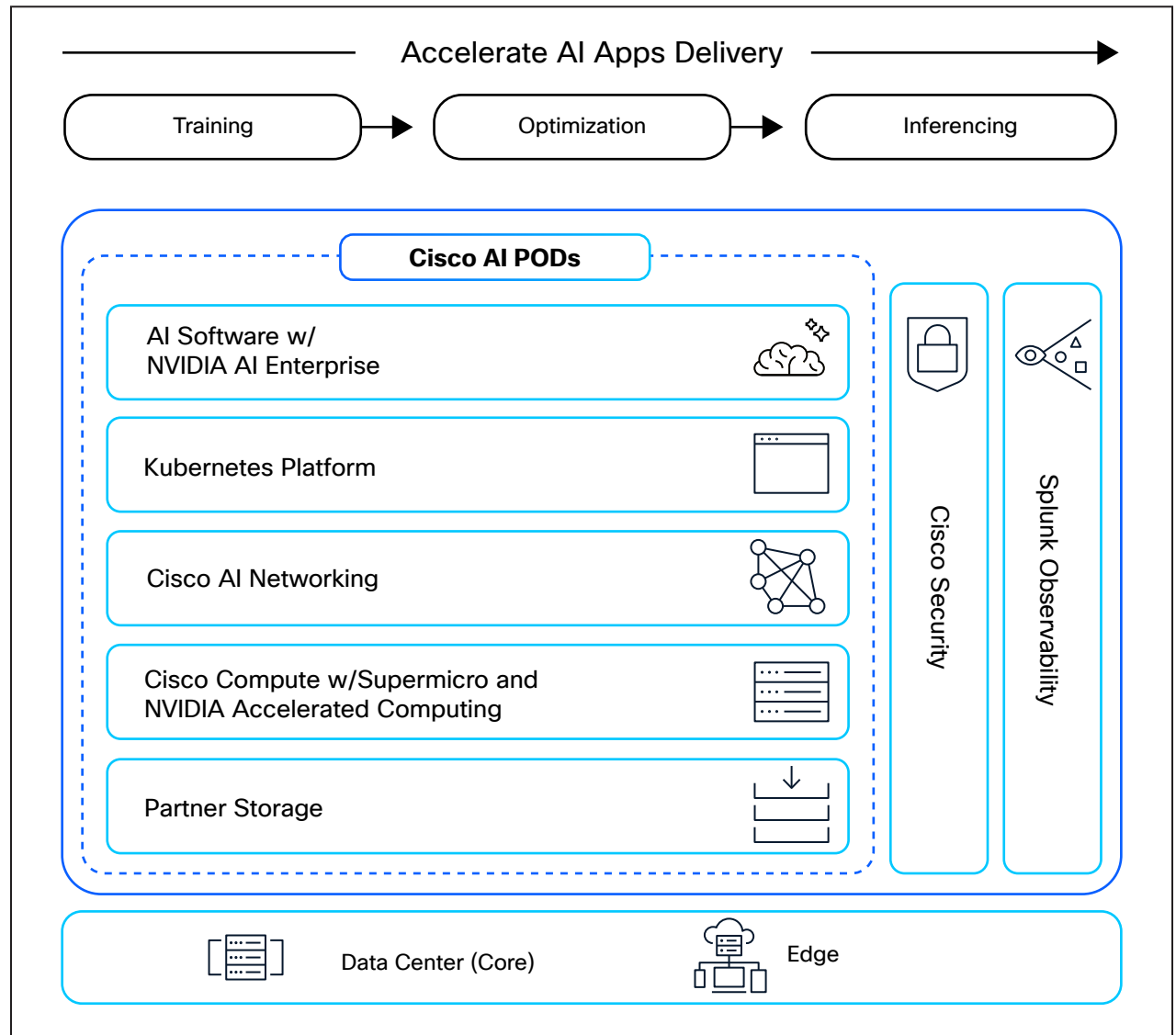


Figure 1. Cisco Secure AI Factory with NVIDIA, extended to rack-scale, liquid-cooled compute built by Supermicro



Figure 2. The compute portfolio: rack-scale NVIDIA NVL72 systems, NVIDIA HGX servers in liquid-cooled and air-cooled form factors, and NVIDIA MGX PCIe GPU servers

“NCP validation gives us the confidence that our infrastructure is optimized from day one, while the platform’s rack-scale capability provides a seamless path to scale our AI operations as our business grows.”

James Manning, Sharon AI, Inc., Co-founder and CEO

Benefits

- **Deploy validated capacity in weeks, not months.** Cisco Validated Infrastructure Services (CVIS) qualify the full-stack design, verify the built cluster against it, and hand you an evidence report.
- **Run an NCP-compliant architecture on Cisco® networking.** Cisco is the first NVIDIA technology partner to deliver an NVIDIA Cloud Partner-compliant reference architecture built on partner-developed networking systems.
- **Scale without redesign** on one reference architecture, from a first NVIDIA NVL72 rack to clusters of tens of thousands of GPUs.
- **Secure and observe every layer** with Cisco AI Defense, Cisco Hypershield™, Isovalent®, and Splunk® observability engineered into the stack rather than added afterward.
- **Protect your delivery date** with Cisco supply chain scale and Supermicro manufacturing scale, at a time when GPU, memory, CPU, and SSD supplies are tight.

One validated full stack, one operating model

The compute-expansion pairs rack-scale and dense NVIDIA compute built by Supermicro with Cisco networking, security, observability, management, and validation that turn racks of GPUs into an operable AI factory. Unlike a self-integrated build, it starts from a qualified design and is proven against that design before handover.

Cisco Secure AI Factory with NVIDIA enables you to:

- **Deploy to a proven design.** Every build starts from an NVIDIA Cloud Partner-aligned or NVIDIA Enterprise Reference Architecture-aligned design, then CVIS verifies the deployed configuration against performance, job completion time, and token throughput targets.
- **Choose the silicon per pod.** Cisco N9300 Series Smart Switches on Cisco Silicon One carry frontend and storage traffic; Cisco N9100 Series Switches on NVIDIA Spectrum-X Ethernet silicon carry lossless backend traffic. Run Cisco NX-OS or SONiC on the same validated hardware.

- **Match cooling to your facility.** Liquid-cooled Cisco N9000 Series Switches interoperate directly with liquid-cooled rack-scale compute, and air-cooled NVIDIA HGX and MGX systems keep projects moving where a liquid loop is not yet in place.
- **Operate AI alongside everything else.** Cisco Cloud Control brings server lifecycle, liquid cooling and power, and network connectivity and policy into one console, one inventory, and one topology.

Cisco is the first NVIDIA technology partner to deliver an NCP-compliant reference architecture built on partner-developed networking systems, with security fused from silicon to agents and one partner to quote, deploy, finance, and support the factory.

Learn more

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](#), or contact your Cisco account team to plan your AI factory.

Data sheets

- [Cisco AI Rack for NVIDIA GB300 NVL72](#)
- [Cisco AI Rack for NVIDIA Vera Rubin NVL72](#)
- [AI Server for NVIDIA HGX Rubin NVL8, Liquid Cooled](#)
- [AI Server for NVIDIA HGX B300, Liquid Cooled](#)
- [AI Server for NVIDIA HGX B300, Air Cooled](#)
- [Cisco AI Server for MGX with NVIDIA GPUs](#)

Partnership: <https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/index.html>.