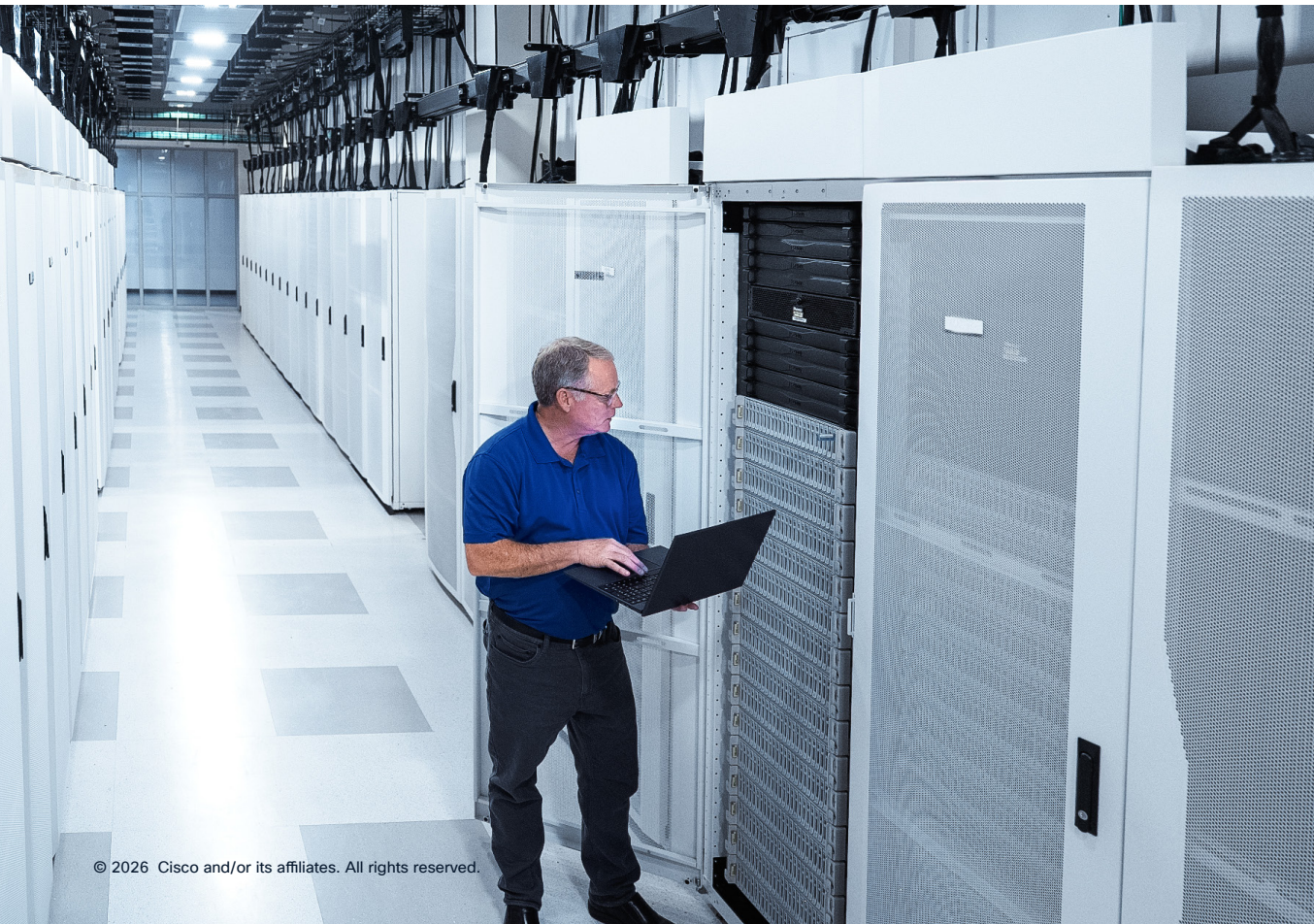


Cisco Secure AI Factory with NVIDIA: Rack-Scale AI for Neo Clouds and Sovereign AI



Whether your capacity is funded by debt or by mandate, the same two things decide the outcome: how fast a rack goes from delivery to trusted production, and how completely you control what runs on it. This architecture is **NVIDIA Cloud Partner Reference Architecture (NCP-RA)**-compliant, isolated per tenant, and validated against its own design before you hand it over.

Overview

Neoclouds and sovereign programs buy the same class of infrastructure, but for different reasons. A neocloud underwrites a fleet against utilization, where every idle week and every congested fabric link is a margin loss that cannot be regained. A sovereign program answers to a mandate, where data residency alone is not enough and control of the hardware, inference path, and operations has to be demonstrable to auditors. Both neoclouds and sovereign programs build multitenant, capital-intensive capacity, and both are exposed to the same failure mode: an assembled stack that nobody can prove was built the way it was designed.

Cisco Secure AI Factory with NVIDIA, expanded with rack-scale compute built by Supermicro, answers both. It is an NCP-RA-compliant design that runs on Cisco® networking with a choice of Cisco NX-OS or SONiC, isolates tenants natively in the fabric, keeps management on premises where that is required, and is verified cluster by cluster through Cisco Validated Infrastructure Services. The published Cisco Cloud Reference Architecture scales to 73,728 GPUs. All platforms are orderable from Cisco authorized channel partners in October 2026.

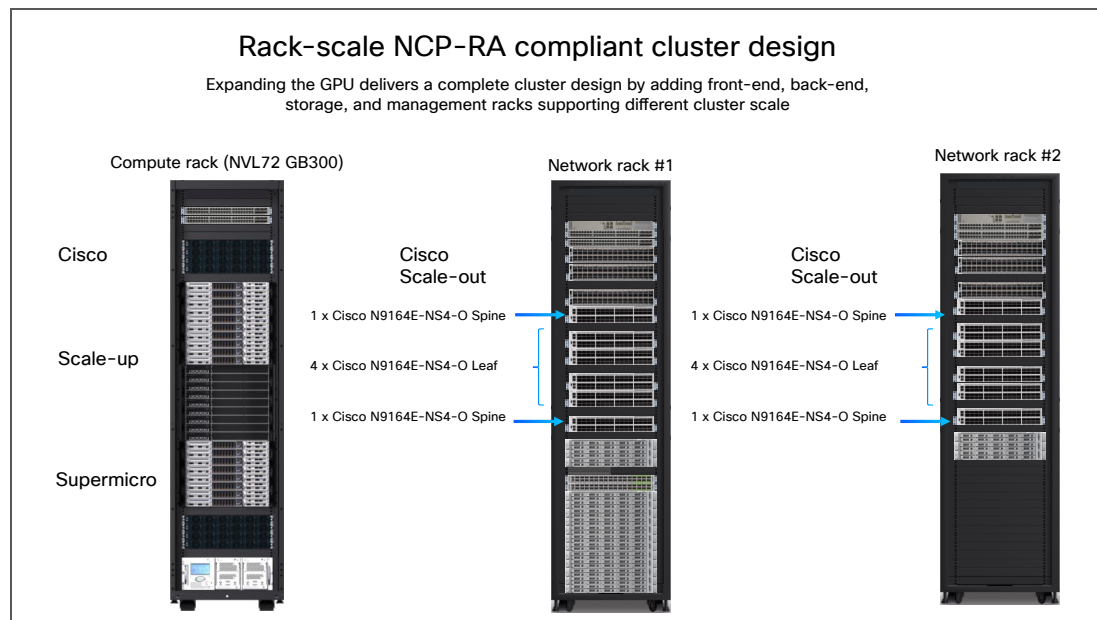


Figure 1. Rack-scale NVL72 capacity, connected by Cisco NCP-RA compliant fabrics and operated as one multi-tenant AI factory

Benefits

- **Deploy on a compliant architecture, on Cisco networking.** Cisco is the first NVIDIA technology partner to deliver an NCP-compliant reference architecture built on partner-developed networking systems, spanning Cisco Silicon One® and NVIDIA Spectrum-X Ethernet switch silicon under one Cisco Nexus® One architecture.
- **Prove the cluster, not just the blueprint.** CVIS qualifies the full-stack design, verifies the deployed configuration against infrastructure performance, job completion time, and token throughput targets, and delivers an evidence report you can put in front of an investment committee or an audit body.
- **Isolate every tenant in the fabric.** VXLAN with BGP EVPN gives each tenant its own backend and frontend VRFs, its own storage tenancy, and its own management nodes, with operator-only VRFs for out-of-band management and internal storage traffic.
- **Keep control where you need it.** SONiC as an open, standards-based option, on-premises fabric management through Cisco Nexus Dashboard, and Cisco AI Defense with NVIDIA NeMo Guardrails deployable fully on premises, with no SaaS hop in the inference path.
- **Grow without redesign.** One reference architecture spans a first NVL72 pod to 73,728 GPUs, with scaling units of two racks so capacity is added incrementally rather than re-cabled.

Trends and challenges

Two funding models, one operational problem

Neocloud fleets are largely debt-financed, which makes utilization the difference between profit and loss and leaves no cushion for idle racks or congested fabrics. Sovereign capacity is funded by policy: Gartner projects worldwide sovereign cloud infrastructure-as-a-service spending of \$80 billion in 2026, up 35.6 percent year over year, and expects 65 percent of governments to introduce technological sovereignty requirements by 2028. Source: Gartner, Feb. 2026; Gartner, Sept. 2025. The funding logic differs, but the operational exposure is identical, because in both cases the asset only earns once it is running in trusted production.

Compliance at the architecture level is not proof at the deployment level

A reference architecture tells you the design is sound; however, it does not tell you that the cluster on your floor was built to it. Between the rack landing and the first trusted workload sits a validation gap that teams usually close by hand, over weeks, with no artifact at the end. For a commercial operator that is unbilled capacity. For a sovereign program it is an audit finding waiting to happen.

Control and cost per token are both network numbers

Mixture-of-experts and reasoning models generate all-to-all traffic at every layer, so NVIDIA NVLink domain size and Ethernet fabric quality set the real cost per token, and a congested fabric idles seven-figure racks. Control follows the same path: if the management plane, the guardrails, or the inference path depend on a service outside your boundary, residency alone does not make the workload sovereign.



Figure 2. One fabric, one operating model, and one validation program across flagship NVL72 pods, HGX clusters, and regional or institutional tiers

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 NCP-RA for cloud and service provider builds and to the NVIDIA Enterprise Reference Architecture for enterprise-scale builds, and CVIS additionally validates against Cisco Enterprise Reference Architecture and Cisco Cloud Reference Architecture.

What you buy: a fleet that you can tier

- **Core: Cisco AI rack for NVIDIA GB300 NVL72.** 72 Blackwell Ultra GPUs and 36 Grace CPUs in one liquid-cooled NVIDIA NVLink domain, for frontier training, reasoning inference, long-context serving, and sovereign foundation models. Cisco AI rack for NVIDIA Vera Rubin NVL72 will follow, on the same pod design, so that today's facility, fabric, and operations will carry forward.

- **Modular: 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 dedicated clusters, managed fine-tuning offers, national labs, and regulated-sector deployments
- **Facility-constrained: Cisco AI server for NVIDIA HGX B300, air-cooled.** An 8RU air-cooled option for sites and regions where liquid facility work is not yet complete
- **Service and institutional tier: Cisco AI server for NVIDIA MGX with NVIDIA GPUs.** Lower-cost inference, visual compute, and vGPU for regional service tiers, government departments, hospitals, and education, with fast delivery times

A fabric that isolates tenants and keeps GPUs fed

Cisco N9100 Series Switches built on NVIDIA Spectrum-X Ethernet silicon deliver the lossless scale-out backend, with dual-plane, rail-optimized topologies and globally aware load balancing rather than static hash-based distribution. Cisco N9300 Series Smart Switches built on Cisco Silicon One carry the converged frontend, storage, and management network,

and liquid-cooled Cisco N9000 Series Switches interoperate directly with liquid-cooled rack-scale compute so the fabric shares the thermal envelope of the systems it serves. The whole fabric runs VXLAN with a BGP EVPN control plane, so every tenant gets isolated backend and frontend VRFs, its own storage account and VLAN, and its own management nodes, while out-of-band management and internal storage traffic stay in operator-only VRFs. Run Cisco NX-OS or SONiC per pod on the same validated hardware, unified through Cisco Nexus One.

Storage and platform, validated at NCP levels

Cisco EBox pairs the VAST Data AI operating system with Cisco UCS C225 M8 Rack Servers and is NVIDIA-Certified high-performance storage at NCP level. Its distributed, everything-shared design scales capacity and throughput by adding servers to a single namespace, with native multitenancy and multiprotocol access over NFS, Amazon S3, and SMB. The architecture also supports other NVIDIA-Certified storage certified at the NCP level, alongside Kubernetes and Slurm orchestration on per-tenant management nodes.

Security and observability that you can operate, or air-gap

Cisco AI Defense enforces runtime guardrails and integrates NVIDIA NeMo Guardrails so policy and enforcement work as one system, deployable fully on premises for sovereignty cases.

Hypershield and Isovalent apply workload-level segmentation across the AI fabric, and Cisco Secure Firewall guards the boundary. Splunk observability correlates job health with compute, NIC, optics, and network performance, so when a run degrades the operator can determine whether the cause is the GPU, the NIC, an optic, or the fabric. At rack scale, where idle GPU capacity is the dominant cost, that is the difference between minutes and days.

Validated with CVIS before handover

Cisco Validated Infrastructure Services (CVIS) are an end-to-end architecture qualification and validation program built on the NVIDIA NVIS methodology, with an automation toolkit that reduces deployment and validation timelines from months to weeks. Every deployment starts from a qualified, NCP-RA- aligned design, every deployed configuration is verified against reference architecture metrics, and delivery is specialist-assisted through the CVIS team with channel partners authorized by Cisco. You receive a fully validated, NCP-RA-compliant

cluster and 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.

Operations you can own

Cisco Cloud Control delivers unified management and AgenticOps from rack-scale core to edge, bringing server lifecycle, liquid cooling and power, and network connectivity

and policy into one console, one inventory, and one topology. Network management is enabled today by Cisco Nexus One, with Cisco Nexus Dashboard available as an on-premises controller when the management plane must stay inside your boundary. Management of compute systems will be available through Cisco Intersight® and Nexus One in CY26Q4. Cisco Customer Experience (Cisco CX) knowledge transfer builds the operating capability alongside the infrastructure, because owning the asset means little if your people cannot run it.





Use Cases

Table 1. Use Cases

Use case	What it looks like in practice
GPU-as-a-service and dedicated clusters	Bare-metal and managed Blackwell and Rubin clusters, provisioned and isolated per tenant with their own VRFs, storage tenancy, and management nodes
Training capacity for AI builders	Frontier and foundation model runs on NVIDIA NVL72 pods, with lossless dual-plane scale-out between pods and rail-optimized scaling units
High-volume inference serving	Reasoning and agentic token serving where fabric quality directly sets cost per token, and observability tells you which layer is throttling it
Sovereign foundation models	Local-language and policy models trained and served under national control on dedicated NVIDIA NVL72 capacity, inside the boundary
Government and citizen services	Eligibility, case triage, and public-sector agents running on premises, with guardrails enforced locally and citizen data never leaving the data center
Defense, intelligence, and classified work	Air-gap-capable analysis on controlled hardware, with on-premises guardrails, workload segmentation, and observability
Regulated-sector and national research AI	Healthcare, finance, energy, and national lab workloads meeting residency, audit, and critical-infrastructure requirements on shared validated clusters
Tiered and regional service offers	MGX-based lower-cost inference and visual compute tiers alongside premium NVIDIA NVL72 capacity, under one 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.](#)

Services

Reach trusted production in weeks, not quarters

Cisco CX and certified partners plan, design, implement, validate, and optimize programs tuned for time-to-first-token, with CVIS as the named validation action and structured knowledge transfer so that your own team owns the fleet. Facility and liquid-cooling readiness assessments settle the power and cooling question before any hardware order, and the customer facility questionnaire and compliance site survey are pre-sales activities required for order processing. For sovereign programs, the same methods produce the documentation that public-sector audit expects.

“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.”

Sharon AI, Inc., Co-founder and CEO

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

What the validation gap actually costs

Between a rack landing on the floor and that rack running securely in production sits work that most teams do by hand: proving that the build matches the design, fine-tuning the fabric, running collective and benchmark tests, and documenting the result. Done manually, it takes months and produces no validation artifact that anyone else can trust.

CVIS closes that gap with qualified full-stack designs, repeatable provisioning, and automated validation against reference architecture metrics, then hands over an evidence report covering configuration and test results.

- **For a commercial operator**, every week of that gap is capacity earning nothing against a financing schedule that does not pause.
- **For a sovereign program**, the evidence report is the difference between asserting control and demonstrating it to an audit body.

The Cisco Advantage

Cisco is the first NVIDIA technology partner to deliver an NCP-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, with SONiC available for operators that require it. Add CVIS validation resulting in a report evidencing conformance of architecture to design, native multitenant isolation, on-premises management options, security enforceable inside your boundary, and Cisco and Supermicro supply chain scale behind your delivery date.

Learn more

Plan for capacity that you can prove meets design

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](#), review the [Cisco N9000 Cloud Reference Architecture with NVIDIA GB300 NVL72](#) and the [rack-scale architecture FAQ](#), or contact your Cisco account team to schedule a capacity, facility, and validation planning session.