GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

Cisco Secure AI Factory with Supermicro: October 2026 Buying Gate

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-10-13 · cisco secure ai factory with supermicro · cisco secure ai factory · supermicro gpu server · nvl72 · hgx b300 · ai infrastructure · liquid cooling · ai networking · gpu procurement

Original article: https://gpusmith.com/articles/en/cisco-secure-ai-factory-supermicro-procurement-gate

In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Compute Configuration and the BOM Gate
5. Implementation Considerations and Process Changes
6. Facilities and Cooling Readiness
7. Networking, Security, and Telemetry Gate
8. Support, Lifecycle, and Quote Normalization
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Conclusion

Executive Summary

Cisco Secure AI Factory with Supermicro warrants a conditional shortlist for an October procurement, with release of the purchase order dependent on a reconciled configuration, site survey, priced support scope, and acceptance plan. Cisco's August announcement added liquid- and air-cooled compute to its existing AI infrastructure architecture. (newsroom.cisco.com) The original FAQ describes ordering in **October 2026**, while the current partner page specifies **the end of October 2026** for global full-stack ordering. (www.cisco.com) (www.cisco.com) Supermicro also describes October as the scheduled start through Cisco and its partners. (learn-more.supermicro.com) These statements establish a procurement window, rather than delivery or verified ordering-system status as of this report's October 11 cutoff.

The published compute choices are meaningfully different. Liquid-cooled HGX B300 requires complete rack integration and onsite service in Supermicro's product description. (www.supermicro.com) MGX itself is a modular reference architecture, so its name cannot substitute for an itemized configuration. (www.nvidia.com) For rack-scale GB300, Supermicro documents **132 to 140 kW operating power**, while NVIDIA's enterprise reference hardware guide says **up to 142 kW**. (www.supermicro.com) (docs.nvidia.com) These are differently scoped specifications, not interchangeable site design limits.

The integration proposition is credible enough to investigate, but must be tested against a named build. Cisco describes **Cisco Validated Infrastructure Services (CVIS)** as aligned with **NVIDIA Infrastructure Services (NVIS)** and validating the delivered infrastructure against its intended design. (newsroom.cisco.com) Independent coverage establishes the portfolio expansion, without providing an independent performance measurement. (www.networkworld.com) A commissioning plan should cover compute, power, cooling, and network function together, consistent with the **American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE)** commissioning framework. (www.ashrae.org)

The recommended buying gate therefore separates **architecture eligibility**, **physical readiness**, and **measured acceptance**. Require firmware recovery evidence, demonstrate tenant policy enforcement, and compare throughput only under matched workloads. The **National Institute of Standards and Technology (NIST)** firmware guidance addresses protection, detection, and recovery; Kubernetes warns that policy objects

require an enforcing network plugin; SPEC cautions against comparisons across different benchmark workloads. (csrc.nist.gov) (kubernetes.io) (spec.org) None of these generic requirements demonstrates that a particular Cisco/Supermicro build has already passed. A complete quote and reproducible pilot are the evidence needed to convert the shortlist into an order.

Introduction and Background

The procurement question is whether a combined rack architecture reduces enough integration uncertainty to justify selection. It is not answered by naming the **graphics processing unit (GPU)** generation alone. Network World describes Supermicro's liquid- and air-cooled systems joining Cisco's broader AI infrastructure portfolio. (www.networkworld.com) Supermicro's collaboration description assigns rack integration, testing, and validation to its own delivery capabilities. (learn-more.supermicro.com) The buyer must determine where that factory preparation ends and site-specific commissioning begins.

This report treats **October 11, 2026** as the research cutoff. The ordering wording summarized above should govern scheduling assumptions until the supplying partner provides actual orderability evidence. A public configuration identifier, accepted quotation, released purchase order, factory completion, site delivery, and acceptance are separate milestones.

The engineering perspective is consistent with GPU Smith's stated practice of delivering engagements against written acceptance criteria and an as-built documentation set. (gpusmith.com) That perspective supplies a decision method; it does not establish a partnership with the vendors or a completed deployment of this architecture.

The evidence hierarchy also matters. Vendor documents establish their own advertised configurations and requirements. Standards and software documentation establish generic design principles. Trade coverage can corroborate the announcement, but Network World's account attributes architecture qualification and validation to Cisco and NVIDIA. (www.networkworld.com) This report found no independent end-to-end benchmark in the reviewed sources.

The sections below translate that evidence into a controlled **bill of materials (BOM)**, facilities worksheet, fabric gate, responsibility matrix, and acceptance sheet. Recommendations are identified as buyer requirements. Unknowns mean that the reviewed evidence does not settle a specific purchasing question, rather than that the architecture cannot support the requirement.

Key Changes

Compute expands beyond the original infrastructure proposition

Supermicro's collaboration description names **GB300 NVL72**, **Vera Rubin NVL72**, and expanded liquid-cooled HGX B300 NVL8 and HGX Rubin NVL8 support. (learn-more.supermicro.com) These names define portfolio scope. They do not establish the exact accelerator, **central processing unit (CPU)**, memory, adapter, and cooling combination contained in an individual quotation. NVIDIA describes MGX as an open modular reference architecture, which reinforces why a platform-family name is insufficient purchasing evidence. (www.nvidia.com)

The initial shortlist should therefore record both the announced family and the exact proposed system. Keep alternatives separate instead of constructing a composite rack from specifications belonging to different products. The Supermicro GB300 datasheet and general cooling page illustrate this issue: the former documents a particular rack cooling option, while the latter advertises broader cooling capabilities. (www.supermicro.com) (www.supermicro.com)

Timing becomes a gate, rather than an assumption

The August FAQ's October ordering statement is broader than the current landing page's end-of-October wording. (www.cisco.com) (www.cisco.com) This is a scheduling refinement that deserves explicit treatment in the procurement record. As of the research cutoff, the reviewed public sources do not demonstrate that a particular configuration can already be transacted in a buyer's jurisdiction.

Cisco's liquid-HGX datasheet further states that configured-system Cisco ordering product identifiers are **in progress**. (www.cisco.com) Procurement should request the active Cisco identifier, corresponding Supermicro identifier, revision, geography, and evidence that the supplying channel can submit the configuration. A published datasheet is useful engineering evidence, but does not by itself settle lead time or delivery. Cisco's sovereign-cloud overview also identifies a customer facility questionnaire and compliance site survey as pre-sales requirements for order processing. (www.cisco.com)

Validation becomes an inspectable deliverable

Cisco's FAQ says CVIS customers receive a report containing validated configuration and test results. (www.cisco.com) Make that report a deliverable with an agreed manifest, test coverage, results, exceptions, and sign-off process.

NVIDIA Cloud Partner (NCP) alignment should likewise be tied to a particular reference architecture and revision. NVIDIA's inference-provider requirements distinguish mandatory language from strong recommendations, and some linked facilities and reference documents require access granted by an NVIDIA representative. (docs.nvidia.com) (docs.nvidia.com) A public alignment statement cannot replace the applicable requirements package or prove that the buyer's selected substitutions remain within its scope.

Compute Configuration and the BOM Gate

NVL72, **HGX**, and **MGX** identify NVIDIA platform families, rather than three equivalent server sizes. NVIDIA's enterprise design guidance describes HGX expansion in a scalable unit of four nodes, while NVL72 expansion occurs one rack at a time. (docs.nvidia.com) (docs.nvidia.com) Those are NVIDIA reference-design conventions, not proof that every Cisco proposal must use the same grouping.

Table 1 maps the documented options to procurement evidence. Identifiers below describe published platforms, not independently verified active ordering codes.

Configuration	Documented compute and integration	Evidence required before release
GB300 NVL72 rack	Cisco identifies SRS-GB300-NVL72-M0 without a CDU and M1 with in-rack CDUs and support. (www.cisco.com)	Reconcile the ordered variant, cooling inclusion, revision, active Cisco identifier, populated mass, and site envelope.
Liquid HGX B300	Supermicro SYS-422GS-NB3RT-LCC requires complete rack integration and onsite service; its page lists eight ConnectX-8 SuperNICs up to 800 Gb/s. (www.supermicro.com) (www.supermicro.com)	Obtain the configured BOM, active ordering identifier, and rack services schedule.
Air-cooled HGX B300	Cisco documents SYS-822GS-NB3RT and a net server mass of 119 kg. (www.cisco.com)	Verify exact GPU/CPU/memory/adapter population, air delivery, rack density, acoustic envelope, and electrical schedule.
MGX Peripheral Component Interconnect Express (PCIe) server	Cisco documents eight H200 NVL GPUs in a 4U Supermicro-built server. (www.cisco.com)	Identify GPU bridging, adapter allocation, included software entitlements, and measured workload fit.
Rubin-family options	The collaboration names Vera Rubin NVL72 and HGX Rubin NVL8 support. (learn-more.supermicro.com)	Require the exact dated datasheet, ordering code, release schedule, facilities envelope, and supported software manifest.

The table deliberately preserves unknown cells. A qualification decision should compare **complete proposed configurations**, including CPU architecture, host memory, storage path, GPU interconnect, network adapters, and cooling. It should not assume that an accelerator-family announcement defines all of those choices.

Use the following decision sequence as an analyst recommendation:

- **Workload first:** Document model size, context, concurrency, training or inference mix, and target latency.
- **Scale-up need:** Establish whether the application benefits from a rack-sized communication domain before selecting rack-scale compute.
- **Incremental expansion:** Compare the operational consequences of server-group versus full-rack growth against the reference-design grouping. (docs.nvidia.com) (docs.nvidia.com)
- **Facility constraint:** Select an air-cooled path only after confirming that the site can remove the proposed system heat.
- **Rack integration:** Treat the liquid-HGX complete-rack and onsite-service requirement as a purchasing dependency. (www.supermicro.com)
- **Software entitlement:** Cisco says NVIDIA AI Enterprise is included with each H200 NVL GPU in its MGX configuration; obtain the license term and support entitlement in the quote. (www.cisco.com)

The outcome should be a controlled build sheet. Alternative processors, adapters, cooling assemblies, or management packages should appear as explicit options with separate acceptance assumptions.

Implementation Considerations and Process Changes

Software is part of the configuration

Cisco's FAQ includes high-performance storage and Kubernetes platforms from ecosystem partners. (www.cisco.com) That establishes architectural scope, rather than an exact storage capacity, scheduler release, or license bundle. VAST's earlier validated-design description concerns Cisco EBox running VAST AI OS on Cisco AI PODs. (www.vastdata.com) Its stated initial focus is the EBox cluster. (www.vastdata.com) Do not transfer that validation to a new Supermicro rack configuration without the configuration-specific design.

A software manifest should cover the operating system, firmware, accelerator driver, runtime libraries, orchestrator, storage client, security agent, and observability collectors. Kubernetes requires the vendor GPU drivers and corresponding device plugin, and the plugin exposes a schedulable hardware resource. (kubernetes.io) (kubernetes.io) Device-plugin documentation also covers high-performance network interface cards (NICs), rather than GPUs alone. (kubernetes.io) This makes adapter exposure and resource discovery a concrete deployment check.

For a Slurm deployment, **generic resources (GRES)** require explicit configuration; Slurm says none are enabled by default. (slurm.schedmd.com) Its topology-aware allocation requires matching network topology information. (slurm.schedmd.com) A physically correct cluster can still schedule jobs poorly if the orchestration model does not describe its resources and fabric.

Recommended implementation controls are:

- **Immutable manifest:** Archive component identities, firmware, driver, runtime, switch operating system, and configuration checksums. NIST's firmware guidance supplies the protection and recovery rationale. (csrc.nist.gov)
- **Resource discovery:** Demonstrate that GPU and adapter resources match the physical inventory and scheduler view. (kubernetes.io)
- **Scheduler health:** Test missing-resource behavior; Slurm documents that nodes with fewer detected resources than configured enter DRAIN. (slurm.schedmd.com)
- **Storage path:** Identify protocol, capacity, throughput, protection, tenant allocation, and recovery objectives. Include response time with throughput, consistent with SPECstorage's measurement scope. (www.spec.org)
- **Disconnected operation:** If required, prove local registries, update imports, licensing, time services, and telemetry retention without external dependencies.
- **Upgrade rehearsal:** Demonstrate an update and a documented rollback using the agreed manifest, preserving the ability to detect and recover from unauthorized platform changes. (csrc.nist.gov)

Segmentation and observability need application evidence

Kubernetes documentation says namespace-based isolation requires other resources, network plugins, and security practices. (kubernetes.io) A namespace is therefore an administrative boundary whose effectiveness must be demonstrated. Cilium's documentation similarly distinguishes policy semantics: deny policies take precedence over allow policies. (docs.cilium.io) In Cilium's default enforcement mode, a selecting policy normally enables default-deny for the traffic direction it governs—ingress or egress—unless

`enableDefaultDeny` disables that transition. (docs.cilium.io) The buyer should test intended access and intended denial separately for ingress and egress, including management, storage, and tenant boundaries, and record the effective enforcement mode and `enableDefaultDeny` settings. Confirm namespace isolation alongside the additional controls required by Kubernetes. (kubernetes.io)

Observability needs both infrastructure measurements and application instrumentation. OpenTelemetry explicitly states that an application must be properly instrumented. (opentelemetry.io) Its Collector receives, processes, and exports telemetry to multiple targets. (opentelemetry.io) A purchasing specification should identify those targets and the ownership of retention, access, and alert rules. Trace collection should follow requests through the distributed application. (opentelemetry.io)

Acceptance should correlate a job identifier with accelerator state, scheduler allocation, NIC counters, switch counters, and storage activity. OpenTelemetry permits raw measurements or predefined aggregations, so collection interval and aggregation settings belong in the test record. (opentelemetry.io) Without that detail, dashboards from different pilots may represent materially different observation windows.

Facilities and Cooling Readiness

The physical gate begins with the **selected rack**, not a family-level marketing maximum. Record **information technology (IT)** demand separately from total facility demand. Supermicro's GB300 datasheet specifies operating power of **132 to 140 kW**. (www.supermicro.com) NVIDIA's GB300 enterprise reference guide lists **up to 142 kW**. (docs.nvidia.com) Those values should remain separately attributed until the supplier confirms the electrical and thermal envelope of the ordered build.

CDU, rack, and building boundaries are equally important. OCP treats cold-plate interfaces as a standardization concern. (www.opencompute.org) A **coolant distribution unit (CDU)** can isolate the secondary technology loop from facility chilled water and control temperature, pressure, and flow. (www.vertiv.com) The Open Compute Project (OCP) describes the technology cooling system as extending from cold plate to CDU, including tubing and manifolds. (www.opencompute.org) Procurement should assign responsibility for that complete path, including connectors and commissioning.

Cooling capacity is conditional. Eaton, for example, specifies a CDU capacity of **2 MW at a 3°C approach temperature**; this is an illustration of performance specification practice, not a selected Cisco component. (www.eaton.com) Supermicro's broader cooling page advertises in-row capacity up to **2.2 MW** and liquid-to-air capacity up to **500 kW**, while its GB300-specific materials describe narrower options. (www.supermicro.com) (www.supermicro.com) Do not combine the maxima across these documents into an approved cooling schedule.

The facilities release checklist should require:

- **Electrical envelope:** Continuous operating load, peak and transient demand, redundancy mode, protection, earthing, and derating.
- **Heat split:** Liquid-captured heat and residual air load, with the site heat-rejection path identified.
- **Water conditions:** Supply/return limits, chemistry, pressure, filtration, allowable flow, and connector compatibility.

- **Dew point:** Verify condensation control; Vertiv describes keeping secondary-loop supply above the data center dew point. (www.vertiv.com)
- **CDU location:** Account for rack space and service access; Schneider notes that in-rack units occupy rack space. (blog.se.com)
- **Mechanical envelope:** Populated mass, point loads, transport route, lift capacity, seismic restraint where applicable, and service clearances.
- **Rack standard:** Identify the actual interface; OCP distinguishes EIA-compatible and Open Rack architectures. (www.opencompute.org)
- **Alarm integration:** Demonstrate leak, temperature, pressure, flow, pump, and power alarms through the operations workflow. NVIDIA's reference hardware documents tray- and rack-level leak detection, which should be verified in the selected build. (docs.nvidia.com)

ASHRAE's water-class names indicate upper temperature limits, but a class designation alone does not supply a selected system's flow, chemistry, or pressure requirements. (www.ashrae.org) Likewise, OCP's interface-standardization work should prompt connector and manifold checks, rather than an assumption that all liquid-cooling components are interchangeable. OCP publishes a distinct Open Rack V3 blind-mate manifold specification, illustrating the need to identify the actual component standard. (www.opencompute.org) (www.opencompute.org)

A liquid-to-air CDU still needs a building air-cooling path; Eaton describes that arrangement as dissipating heat through data center air conditioning. (www.eaton.com) Moving liquid within a rack does not eliminate the final heat-rejection obligation.

Networking, Security, and Telemetry Gate

Name the topology and the supported combination

Cisco's at-a-glance document assigns frontend and storage traffic to Silicon One-based N9300 smart switches and lossless backend traffic to Spectrum-X-based N9100 switches. (www.cisco.com) Its GB300 cloud reference architecture separates east-west compute, converged north-south storage/management, and out-of-band management fabrics. (www.cisco.com) These are design roles that should appear explicitly in the quote and port map.

The same documents permit Cisco NX-OS or **SONiC**, Software for Open Networking in the Cloud, on the validated hardware. (www.cisco.com) Buyers should nevertheless require the exact switch model, operating-system release, feature set, adapter firmware, optical modules, cable lengths, and reference-design revision. A family-level operating-system choice does not establish every combination's validated scope.

NCP status can change with a topology substitution. Cisco's reference paper permits an alternative switch configuration that it explicitly labels as lacking NCP compliance. (www.cisco.com) The purchasing record should distinguish **supported architecture**, **reference-architecture compliance**, and **tested deployed configuration**.

The network acceptance requirements should include:

- **Port map:** List every host-facing, inter-switch, storage, and management connection with its endpoint and speed.
- **Rail mapping:** Record adapter-to-GPU and adapter-to-switch-plane relationships.
- **Optics compatibility:** Obtain approved modules, reach, breakout configuration, connector type, and spare policy.
- **Oversubscription:** State downstream and upstream capacity separately for each fabric and each expansion stage.
- **Congestion behavior:** Test collective traffic together with the representative workload and collect relevant counters.
- **Recovery:** Measure behavior after an approved link, adapter, or switch-path interruption.
- **Scheduler locality:** Confirm that placement reflects topology; Slurm supports topology-aware resource allocation. (slurm.schedmd.com)

A worksheet can calculate connections from the accepted endpoint map, but it should not infer port counts from GPU count alone. Dual-plane or breakout designs alter cable, transceiver, and switch-port accounting. The scheduler's topology representation also needs to match the accepted map. (slurm.schedmd.com)

Security controls require enforceable policies

Cisco assigns model validation and runtime guardrails to **AI Defense**, and east-west segmentation to **Hypershield** and **Isovalent**. (www.cisco.com) NIST's zero-trust model withholds implicit trust based on physical or network location alone. (www.nist.gov) Accordingly, the acceptance plan should test identities and policies, rather than infer security from the private location of the rack.

The proposed security tests should cover:

- **Identity:** Administrative roles, service identities, authentication, and management separation, without assuming trust from network location. (www.nist.gov)
- **Tenant isolation:** Allowed and denied paths across compute, storage, and management; namespace boundaries alone need supporting controls. (kubernetes.io)
- **Policy engine:** Verify actual enforcement; Kubernetes says policies are ignored without an implementing network plugin. (kubernetes.io)
- **Runtime guardrails:** Record policy versions, test inputs, expected decisions, and retained evidence.
- **Firmware resilience:** Demonstrate the protection, detection, and recovery processes described by NIST. (csrc.nist.gov)
- **Local operations:** For sovereignty requirements, prove the proposed deployment and its external dependencies; Cisco describes on-premises AI Defense with NeMo Guardrails. (www.cisco.com)

Cilium describes **Hubble** as a networking and security observability platform. (docs.cilium.io) That is useful supporting documentation for a policy-evidence approach, but it does not prove Hubble is included in the quoted stack. OpenTelemetry also describes signals as functioning independently, so a complete dashboard requires deliberate correlation across those signals. (opentelemetry.io)

Support, Lifecycle, and Quote Normalization

The support boundary deserves its own purchasing schedule. Cisco's FAQ assigns initial triage to Cisco while stating that each vendor supports its respective components. (www.cisco.com) The GB300 datasheet applies standard Supermicro warranty terms and Cisco support under the applicable service contract. (www.cisco.com) These descriptions are complementary, but they do not replace a binding escalation map or define every response-time commitment.

Supermicro describes rack-scale integration, testing, and validation. (learn-more.supermicro.com) NVIDIA's **August 2024 NVIS datasheet** describes project planning with a schedule and roles-and-responsibilities plan. (resources.nvidia.com) The buyer should request the actual CVIS statement of work instead of assuming that every element of a separate NVIDIA services description transfers automatically.

Table 2 is a proposed responsibility and acceptance sheet. It assigns owners to be confirmed in the contract, rather than presenting a documented universal RACI, meaning responsible, accountable, consulted, and informed.

Work package	Owner to name in the contract	Acceptance evidence
Configuration integrity	Lead integrator accountable; compute, fabric, and software vendors consulted	Signed BOM, serial inventory, firmware/license manifest, approved deviations, and as-built record; recoverable firmware belongs in this evidence package. (csrc.nist.gov)
Physical commissioning	Site engineering and integrator, with CDU/loop boundaries assigned	Electrical, cooling, leak, and alarm tests; ASHRAE treats commissioning as a combined infrastructure test. (www.ashrae.org)
GPU and fabric validation	Lead integrator with accelerator/network specialists	Correctness and performance tests; NVIDIA distinguishes single-node testing from whole-cluster health checks. (docs.nvidia.com) (docs.nvidia.com)
Application acceptance	Buyer workload owner and integrator	Named model/application, dataset, latency and throughput targets, reproducible run settings, and retained traces; application instrumentation is a prerequisite. (opentelemetry.io)
Operations readiness	Buyer operations accountable, service providers supporting	Training, procedures, maintenance, spares, and escalation contacts; these are also elements of Uptime's operating assessment. (uptimeinstitute.com)
Warranty and escalation	Contractually named first contact and component owners	Coverage, response terms, onsite entitlement, cross-vendor escalation, exclusions, and closure authority.

The matrix's value is boundary clarity. An alarm that crosses cooling, compute, and software cannot be resolved operationally by a list of product names. The contract should identify who gathers evidence, who coordinates the vendors, and who authorizes return to service.

Normalize competing quotations using the following buyer requirements:

- **Hardware:** Match compute population, adapters, switches, optics, storage, rack components, and cooling assemblies.
- **Software:** Separate included entitlements from subscriptions; Cisco's liquid-HGX datasheet requires an Intersight license for Intersight management. (www.cisco.com)

- **Services:** Itemize design, factory integration, shipment, onsite work, CVIS, training, and documentation.
- **Support:** Specify duration, hours, geography, response targets, onsite terms, and renewal basis.
- **Spares:** Identify stocked parts, location, replenishment, ownership, and coverage for liquid components.
- **Commercial milestones:** Separate deposit, shipment, delivery, commissioning, and acceptance payments.
- **Delivery:** Request a dated, configuration-specific commitment and dependencies, rather than treating orderability as delivery.
- **Lifecycle:** Require an approved upgrade path, compatibility ownership, rollback method, and replacement strategy, with operational training and maintenance responsibilities. (uptimeinstitute.com)

The reviewed sources do not establish a public configured-stack price. Treat **Cisco Secure AI Factory Supermicro pricing** as a quote-normalization exercise. No defensible transaction amount follows from adding unrelated server list prices or assigning a generic price to an announced rack family.

Data Analysis and Evidence

Site worksheet and reproducible calculations

Table 3 separates published data from calculated buyer inputs. Server mass and power-supply ratings must not be silently converted into populated rack mass or measured operating load.

Worksheet item	Verified evidence or explicit input	Procurement use
GB300 operating load	Supermicro publishes 132 to 140 kW; NVIDIA reference hardware lists up to 142 kW. (www.supermicro.com) (docs.nvidia.com)	Require the selected build's continuous, peak, and derated electrical envelope.
Liquid HGX server mass	Cisco lists 84.6 kg net and 100.7 kg gross. (www.cisco.com)	Add every installed rack component; shipping mass is not installed mass.
Air HGX server mass	Cisco lists 119 kg net and 137 kg gross. (www.cisco.com)	Obtain populated rack mass and point-load distribution.
Liquid HGX power supplies	Supermicro lists four 6.6-kW supplies in a 2+2 arrangement. (www.supermicro.com)	Ratings describe the supply inventory, not measured consumption.
Facility energy	The **U.S. Department of Energy (DOE)** defines **power usage effectiveness (PUE)** using annual facility energy divided by annual IT energy. (www.energy.gov)	Use a stated planning assumption and later reconcile it to measured energy.
Water and heat	DOE describes **water usage effectiveness (WUE)** in liters per kWh and gives the single-phase heat-flow relationship in its thermodynamics handbook. (www.energy.gov) (www.energy.gov)	Obtain the actual coolant properties, temperatures, flow, and annual water boundary.

The table prevents false precision. A different Supermicro liquid-HGX variant, SYS-422GS-NB3RT-ALC, has a published **101 kg net mass**. (www.supermicro.com) That figure should not overwrite the Cisco LCC variant's documented mass. Product suffixes and revision dates belong in the worksheet.

Hypothetical Example: Assume buyer inputs of **140 kW IT load**, **1.20 planning PUE**, and **8,760 operating hours per year**. The calculated annual IT energy is **1,226,400 kWh**; applying the planning factor gives **1,471,680 kWh** facility energy. This is analyst arithmetic, not a Cisco performance result. DOE's PUE definition uses annual energy, so an instantaneous estimate of **168 kW** facility demand is only a planning approximation. (www.energy.gov)

For a single-phase loop, use **Q = m × Cp × ΔT**, where Q is heat-transfer rate, m is mass-flow rate, Cp is the actual coolant's specific heat, and ΔT is return minus supply temperature. DOE cautions that this temperature-difference form assumes no phase change. (www.energy.gov) With hypothetical buyer inputs of **140 kW liquid heat**, **Cp = 4.2 kJ/kg·K**, and **ΔT = 10 K**, calculated flow is **3.33 kg/s**. The Cp value is an assumed worksheet input, not a verified coolant property or product requirement. Residual air heat, exchanger approach, pump head, and redundancy still need separate engineering.

A floor-loading worksheet should divide installed mass by the relevant footprint, then separately check wheel or foot reactions, transport loads, and raised-floor limits. Do not invent the missing rack mass to complete the calculation.

Fabric counts and pilot measurements

Cisco's GB300 reference design describes each east-west ConnectX-8 adapter operating as **two 400G connections** into separate planes. (www.cisco.com) The connection count should therefore follow the accepted adapter and plane map. For an explicitly hypothetical **32 endpoints per plane**, **two planes**, and **four spare endpoint ports per plane**, the calculated host-facing budget is **72 ports**. This excludes spine uplinks, management, and storage. Transceiver counts depend on cable type and breakout; an active optical cable is not two separately purchased optical modules.

For pilot acceptance, **NVIDIA Collective Communications Library (NCCL)** tests assess correctness and performance. (docs.nvidia.com) The Data Center GPU Manager, or **DCGM**, NCCL plugin is single-node only, so passing it does not validate the inter-node fabric. (docs.nvidia.com) NVIDIA's NCP inference requirements recommend SKU baselines and a **24-hour representative-inference soak**. (docs.nvidia.com) (docs.nvidia.com) Capture software versions, message sizes, topology, power cap, thermal steady state, errors, job throughput, and latency distributions. Archive the telemetry export configuration as well as dashboards: the Collector separates receiving, processing, and exporting. (opentelemetry.io)

Storage testing should report response time as well as throughput; SPECstorage measures both. (www.spec.org) SPEC instructs users to compare results within the same workload. (spec.org) A generic file-server score cannot establish end-to-end model-serving capacity.

Implications and Future Directions

The architecture can justify a shortlist when its proposed integration scope matches the buyer's staffing, facilities, and application requirements. The decisive evidence is a configuration-specific acceptance record, rather than the breadth of the announced portfolio. Uptime Institute's construction certification provides a

useful distinction: it checks whether a facility was built as designed and includes live demonstrations. (ats.uptimeinstitute.com) (ats.uptimeinstitute.com) That analogy does not imply that the Cisco/Supermicro offering has Uptime certification.

GPU Smith's published procurement method describes a controlled build sheet containing part numbers, link budget, lead times, and pricing. (gpusmith.com) Applied here, that approach means giving uncertainty a named owner and an evidence deadline. Commercial review can then distinguish a complete proposal from one that still requires engineering closure.

Operational readiness should carry equal weight with initial bring-up. Uptime's operating guidance includes operating set points and load-management procedures. (uptimeinstitute.com) Schneider's discussion of floor-mounted CDUs also highlights shared-loop service and maintenance consequences. (blog.se.com) Those considerations should inform training, spare strategy, maintenance access, and planned isolation of equipment. Facility operating procedures should incorporate the approved set points and load-management assumptions. (uptimeinstitute.com)

Future compute options should be admitted through the same gate. Reusing a rack or pod concept does not by itself demonstrate that a later accelerator generation has the same electrical load, coolant envelope, adapter mapping, or software dependencies. Obtain the relevant revised documents and repeat only the affected acceptance checks.

Finally, correlate infrastructure telemetry with the actual application path. OpenTelemetry describes distributed tracing as following requests through distributed systems. (opentelemetry.io) That supports investigating whether a serving delay originates in scheduling, storage, networking, or the application itself. It remains a buyer instrumentation requirement, rather than evidence of a measured result for this stack.

Conclusion

Shortlist conditionally; require evidence before committing to a specific build. The reviewed documents establish a rack-scale expansion with Supermicro compute, documented server and rack families, Cisco fabrics, software dependencies, and validation services. They provide enough technical detail to begin configuration and site planning, while leaving configuration-specific commercial and acceptance questions open.

The immediate procurement deliverable should be a reconciled BOM, active ordering identifiers, dated quotation, facilities envelope, reference-architecture scope, software/license manifest, and support schedule. The release decision should then depend on a reproducible pilot and an acceptance report for that same configuration.

Keep the boundaries explicit. Orderability does not establish delivery. Power-supply inventories do not establish operating consumption. A reference-platform maximum does not establish a selected rack's electrical schedule. A policy object does not demonstrate enforcement. Single-node diagnostics do not demonstrate the inter-node fabric. A validation methodology does not establish that the buyer's build has passed.

This discipline accommodates both positive and unknown findings. A proposal that closes the gates can move forward on an inspectable engineering basis. A proposal with unresolved site, version, topology, pricing, or escalation questions should remain conditional until the responsible supplier supplies the missing evidence. The objective is a supported, maintainable AI deployment whose installed behavior can be compared with its ordered specification.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m08/cisco-secure-ai-factory-nvidia-rack-scale.html
2. https://www.cisco.com/c/en/us/solutions/collateral/artificial-intelligence/cisco-secure-ai-factory-nvidia-rackscale-faq.html
3. https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/index.html
4. https://learn-more.supermicro.com/data-center-stories/supermicro-cisco-secure-ai-factory-nvidia
5. https://www.supermicro.com/en/products/system/gpu/4u/sys-422gs-nb3rt-lcc
6. https://www.nvidia.com/en-us/data-center/products/mgx/
7. https://www.supermicro.com/datasheet/datasheet_SuperCluster_GB300_NVL72.pdf
8. https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html
9. https://www.networkworld.com/article/4213820/cisco-bulks-up-its-ai-infrastructure-portfolio-with-supermicros-liquid-cooled-servers.html
10. https://www.ashrae.org/technical-resources/topics-and-initiatives/ai-data-center/introduction-and-purpose
11. https://csrc.nist.gov/pubs/sp/800/193/final
12. https://kubernetes.io/docs/concepts/security/multi-tenancy/
13. https://spec.org/storage2020/results/
14. https://gpusmith.com/
15. https://gpusmith.com/articles/en/nvidia-hgx-b300-price-specs
16. https://www.supermicro.com/en/solutions/liquid-cooling
17. https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/nvidia-hgx-b300-liquid-cooled-ds.html
18. https://www.cisco.com/c/en/us/solutions/collateral/artificial-intelligence/sovereign-ai-factory-nvidia-rackscale-neosov-so.html
19. https://docs.nvidia.com/dsx/ncp/inference-provider-requirements/home
20. https://gpusmith.com/articles/en/dgx-vs-hgx-vs-nvl72-vs-mgx
21. https://docs.nvidia.com/enterprise-reference-architectures/white-paper/latest/design-tenets.html
22. https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/nvidia-gb300-nvl72-ds.html
23. https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/nvidia-hgx-b300-air-cooled-ds.html
24. https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/mgx-pcie-gpu-server-ds.html
25. https://www.vastdata.com/blog/cisco-vast-establishing-data-foundation-secure-ai-factory-nvidia
26. https://kubernetes.io/docs/tasks/manage-gpus/scheduling-gpus/
27. https://kubernetes.io/docs/concepts/extend-kubernetes/compute-storage-net/device-plugins/
28. https://slurm.schedmd.com/gres.html

29. https://slurm.schedmd.com/topology.html
30. https://www.spec.org/storage2020/
31. https://docs.cilium.io/en/stable/security/policy/deny/
32. https://docs.cilium.io/en/stable/security/policy/intro/
33. https://opentelemetry.io/docs/concepts/observability-primer/
34. https://opentelemetry.io/docs/collector/architecture/
35. https://opentelemetry.io/docs/specs/otel/overview/
36. https://www.opencompute.org/projects/cooling-environments/cold-plate/
37. https://www.vertiv.com/en-in/insights/articles/educational-articles/understanding-coolant-distribution-units-cdus-for-liquid-cooling/
38. https://gpusmith.com/articles/en/cold-plate-gpu-rack-commissioning-checklist
39. https://www.eaton.com/us/en-us/catalog/thermal-management-solutions/coolant-distribution-unit-cdu.html
40. https://blog.se.com/datacenter/2026/04/29/in-rack-cdu-floor-mounted-cdu-option-better-data-center/
41. https://www.opencompute.org/projects/rack-and-power/
42. https://gpusmith.com/articles/en/gpu-data-center-cooling-design
43. https://www.cisco.com/c/en/us/solutions/collateral/artificial-intelligence/secure-ai-factory-nvidia-rackscale-aag.html
44. https://www.cisco.com/c/en/us/products/collateral/data-center-networking/n9000-gb300-nvl72-cra.html
45. https://www.cisco.com/c/en/us/solutions/collateral/artificial-intelligence/secure-ai-factory-nvidia-rackscale-enterprise-so.html
46. https://www.nist.gov/publications/zero-trust-architecture
47. https://docs.cilium.io/en/stable/overview/intro/
48. https://resources.nvidia.com/_pfcdn/assets/10412/contents/867639/163243c3-8c8f-4b02-85fc-17d95d392d72.pdf
49. https://docs.nvidia.com/datacenter/dcgm/latest/reference/diagnostics/plugins/nccl-tests.html
50. https://docs.nvidia.com/dgx-superpod/administration-guide-dgx-superpod/latest/system-health.html
51. https://uptimeinstitute.com/tier-certification/operations
52. https://www.energy.gov/cmei/femp/cooling-water-efficiency-opportunities-federal-data-centers
53. https://www.energy.gov/sites/default/files/2026-04/DOE-HDBK-1012-92_VOL1.pdf
54. https://www.supermicro.com/en/products/system/datasheet/sys-422gs-nb3rt-alc
55. https://gpusmith.com/articles/en/gpu-cluster-energy-calculator
56. https://gpusmith.com/articles/en/gpu-cluster-acceptance-testing-runbook
57. https://ats.uptimeinstitute.com/tier-certification/construction
58. https://gpusmith.com/articles/en/gpu-cluster-reliability-budget

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.