



Cisco AI Server for MGX with NVIDIA GPUs

(Supermicro SYS-422GL-FNR2)

Contents

Product overview	2
Features and benefits.....	3
Prominent feature.....	4
Platform support	5
Licensing.....	5
Product sustainability	6
Product specifications.....	6
System requirements	8
Ordering information	8
Warranty information.....	9
Cisco and Partner Services	9
Cisco Capital.....	9
Learn more.....	9

Eight NVIDIA H200 NVL GPUs in a 4U Supermicro-built PCIe server, sold and supported by Cisco and managed through Cisco Intersight, for enterprise AI inference without rack-scale complexity

Product overview

The AI Server SYS-422GL-FNR2 is a 4RU air cooled MGX PCIe GPU server for large model inference and fine tuning. Configured with eight NVIDIA H200 NVL Tensor Core GPUs, it provides 1,128 GB of HBM3e memory and 38.4 TB/s of aggregate GPU memory bandwidth in a single node.

Memory capacity is the reason to choose this platform. Each H200 NVL carries 141 GB of HBM3e. With optional NVLink bridging, four GPUs form a single 564 GB memory domain, large enough to hold and fine tune model classes that would otherwise be sharded across multiple servers. Sharding costs interconnect bandwidth, scheduling complexity, and engineering time. Keeping a model resident in one domain avoids all of it.

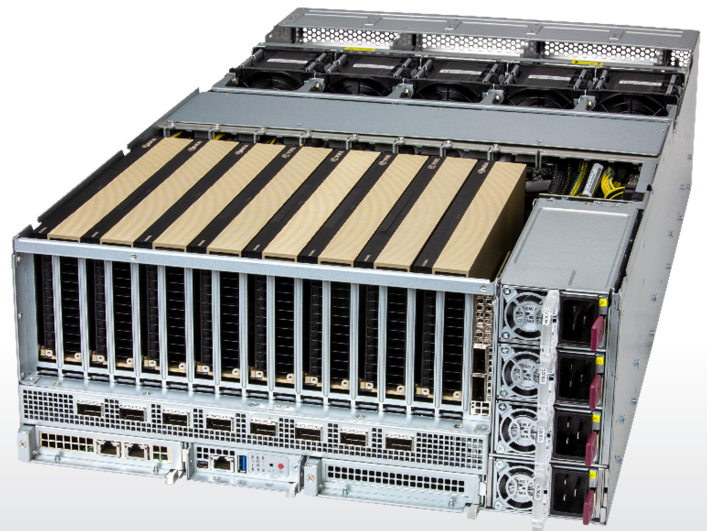


Figure 1. AI Server for MGX with NVIDIA GPUs (Supermicro SYS-422GL-FNR2)

H200 NVL is a PCIe dual slot, air cooled accelerator. Organizations can use existing air-cooling for simplified deployment without the requirements of liquid cooling infrastructure. The Supermicro SYS-422GL-FNR2 Networking is handled by an NVIDIA ConnectX-8 PCIe switch board, which combines a 400/800 Gb/s SuperNIC with an integrated PCIe Gen6 switch. Collapsing those two functions into a single component removes a hop between GPU and fabric and frees up to eight PCIe 6.0 x16 double-width slots for accelerators.

Two Intel Xeon processors, up to 3 TB of DDR5 memory, and a dual root PCIe 5.0 switch topology keep the accelerators supplied. NVIDIA ConnectX-8 SuperNICs handle east-west traffic and NVIDIA BlueField-3 handles storage and north-south connectivity. Cisco Intersight provides lifecycle management across the fleet, and CX Services provides support.



Features and benefits

Table 1 summarizes the primary features and benefits of the platform.

Table 1. Features and benefits

Feature	Benefit
Eight NVIDIA H200 NVL GPUs, 141 GB HBM3e each	1,128 GB of GPU memory per node, enough to serve or fine tune large models without multi-node sharding
4.8 TB/s memory bandwidth per GPU, 38.4 TB/s aggregate	Sustains token throughput on memory bound inference, where bandwidth rather than compute sets the ceiling
4-way NVLink bridging at 900 GB/s per GPU	Forms 564 GB unified memory domains for models that exceed single GPU capacity
Multi-Instance GPU, up to 7 instances per GPU	Partitions the system into as many as 56 hardware isolated instances for multi-tenant serving
NVIDIA Confidential Computing on Hopper	Keeps data and models encrypted in use, isolated from the hypervisor and from co-resident tenants
Dual Intel Xeon Series, 128 cores each	Supplies data preprocessing and orchestration without starving the accelerators
NVIDIA ConnectX-8 SuperNICs and BlueField-3	Provides high bandwidth cluster networking with offloaded storage and security services
Dual root PCIe 5.0 switch topology	Balances GPU and NIC bandwidth across both CPU roots and reduces cross socket traffic
Up to 3 TB DDR5-6400 system memory	Holds large datasets and embedding tables in host memory alongside GPU resident models
NVIDIA AI Enterprise included	Adds supported inference runtimes, NIM microservices, and framework support with the GPUs
Cisco Intersight management	One operations model across the Cisco compute estate
4x 3200 W Titanium (3+1) power supplies	96% efficiency and redundancy; survives a PSU failure under full GPU load
Cisco CX Services support	Dedicated support contracts from Cisco and Supermicro for a multi-vendor system

Prominent feature

Dual-root PCIe architecture with NVIDIA ConnectX-8

Deployment topologies

The dual root PCIe 5.0 switch supports two configurations, selected at order time by whether the NVLink bridge kit is included.

Without bridging, the system presents eight independent GPUs. Each is an addressable inference endpoint with 141 GB of HBM3e. This configuration maximizes tenant density and replica count, and suits high concurrency serving of models that fit within a single GPU.

With 4-way bridging, the system presents two 564 GB memory domains. This configuration suits fine tuning and inference of models that exceed single GPU memory, and it keeps those models inside one server rather than distributing them across a cluster.

On network sizing, the platform carries four ConnectX-8 SuperNICs across eight GPUs. This ratio is designed for inference serving and single node fine tuning. Customers with multi-node distributed training requirements should evaluate the Cisco Supermicro HGX B300 platforms, which provide a one to one GPU to SuperNIC ratio for scale out fabrics.

Multi-tenancy and workload isolation

Multi-Instance GPU divides each H200 NVL into as many as seven instances, each with dedicated memory, cache, and streaming multiprocessors. A fully populated server therefore supports up to 56 concurrent isolated workloads. Because the partitioning is enforced in hardware, a fault or a resource spike in one instance does not affect the others, which makes quality of service predictable in shared environments.

NVIDIA Confidential Computing is supported on the Hopper architecture. Data and model weights remain

encrypted while in use, protected from the hypervisor, the host operating system, and other tenants on the same physical server. For regulated industries, sovereign deployments, and service providers who must demonstrate isolation rather than assert it, this is often the requirement that determines platform selection.

Software and management

NVIDIA AI Enterprise is included with each H200 NVL GPU. It provides supported inference runtimes, NVIDIA NIM microservices for deploying models as API endpoints, and validated builds of common training and inference frameworks.

Cisco Intersight manages the platform through its lifecycle, covering server profiles, firmware, configuration policy, hardware telemetry, and health monitoring. Servers can be managed individually or as a pool, and Intersight applies consistent policy across select Supermicro server deployments.

Most PCIe GPU servers bottleneck at the host bridge. This system uses a dual-root design: each Xeon CPU owns a PCIe 5.0 x16 path into the NVIDIA ConnectX-8 switch board, which fans out to the GPU slots and integrated QSFP networking at up to 400 Gb/s. GPU-to-GPU traffic stays on the switch board instead of crossing the CPU interconnect, and optional NVLink bridges pair adjacent GPUs for workloads that need more than PCIe bandwidth.

The practical result: eight GPUs behave like a coherent inference pool, not eight isolated cards. For scale-out, four ConnectX-8 SuperNICs carry east-west traffic and up to two NVIDIA BlueField-3 B3220 Crypto controllers offload storage and north-south networking from the host CPUs.



Platform support

Cisco offers this Supermicro platform in a defined configuration, listed in Table 2.

Table 2. Cisco selected configuration, SYS-422GL-FNR2 (MGX H200)

Item	Selected configuration
PID	SYS-422GL-FNR2 (Supermicro SKU)
Platform	MGX PCIe GPU server
GPU	Up to 8x NVIDIA H200 NVL, 141 GB HBM3e, PCIe Gen5, dual slot, air cooled
4-way NVLink bridging at 900 GB/s per GPU	Forms 564 GB unified memory domains for models that exceed single GPU capacity
CPU	2x Intel Xeon 6980P (128 cores, 2.0 GHz, 504 MB cache, 500 W)
Memory	16x or 24x 128 GB DDR5-6400 2Rx4 ECC RDIMM (6.0 TB max)
M.2 boot drive	2x 1.9 TB NVMe PCIe Gen4 M.2 22x110, 1 DWPD, SED
NVMe storage	Up to 8x 3.84 TB or 7.68 TB NVMe PCIe 5.0 x4 E1.S 9.5 mm, 1 DWPD, SED
East-west NICs	4x NVIDIA ConnectX-8 SuperNICs
North-south NICs	Up to 2x NVIDIA BlueField-3 B3220SH E-Series storage controllers
Management	Cisco Intersight (SaaS and on-premises) and HyperFabric (SaaS only)
Support	Cisco CX Services
Availability	October 2026

Licensing

This is a hardware product; no platform license is required. Cisco Intersight management requires an Intersight license, listed in Table 3.

Table 3. Licensing

License	Description
Cisco Intersight	License tier appropriate to the deployment see Cisco Sales specialist



Product sustainability

Information about Cisco’s Environmental, Social and Governance (ESG) initiatives and performance is provided in Cisco’s CSR and sustainability [reporting](#).

Table 4. Cisco Environmental Sustainability Information

Sustainability Topic		Reference
General	Information on product-material-content laws and regulations	Materials
	Information on electronic waste laws and regulations, including our products, batteries and packaging	WEEE Compliance
	Information on product takeback and resuse program	Cisco Takeback and Reuse Program
	Sustainability Inquiries	Contact: csr_inquiries@cisco.com
Material	Product packaging weight and materials	Contact: environment@cisco.com

Product specifications

Table 5 reflects the Supermicro data sheet for the SYS-422GL-FNR2. Product specifications may change without notice.

Table 5. Product specifications, Supermicro GPU SuperServer SYS-422GL-FNR2

Category	Specification
Form factor	4U rackmount; enclosure 438.4 x 176 x 800 mm (17.26 x 6.93 x 31.5 in)
Processor	Dual Socket BR (LGA-7529), Intel Xeon 6900 series with P-cores, up to 128C/144T, up to 500 W air-cooled
GPU support	Up to 8 double-width PCIe GPUs; NVIDIA H200 NVL (141 GB)
GPU interconnect	CPU-GPU: PCIe 5.0 x16 switch, dual-root; GPU-GPU: optional NVIDIA NVLink bridge



Category	Specification
System memory	24 DIMM slots; up to 6.0 TB DDR5-6400 ECC RDIMM or 6.0 TB DDR5-8800 ECC MRDIMM (1DPC)
Drive bays	8 rear hot-swap E1.S NVMe bays; 2x M.2 PCIe 4.0 x4 NVMe (M-key 2280 default)
Expansion slots	8x PCIe 6.0 x16 FH double-width; 2x PCIe 5.0 x16 FHHL
Networking	NVIDIA ConnectX-8 PCIe switch board with 8 integrated QSFP ports up to 400 Gb/s; 1x RJ45 1GbE BMC port (ASPEED AST2600)
I/O	1x USB 3.0 Type-A (rear), 1x mini-DP, 1x TPM header
Cooling	10x 80 mm fans
Power	4x 3200 W redundant (3+1) Titanium Level (96%)
BIOS	AMI 32 MB UEFI
System management	Cisco Intersight and Cisco Nexus One (Hyperfabric)
Operating environment	10 to 35 C operating; 8% to 90% relative humidity, non-condensing
Weight	Net 22.45 kg (49.5 lb); gross 43.75 kg (96.45 lb)
Motherboard/chassis	Super X14DBG-MAP/CSE-MG401TS-R0BNDFP



System requirements

Table 6 lists what must be in place before installation.

Table 6. System requirements

Requirement	Description
Rack	Standard 19-inch rack, 4U, 800 mm chassis depth; verify rail and airflow clearance
Power	4x 3200 W inputs; provision high-line AC feeds sized for full GPU population
Operating temperature	10 to 35 C (50 to 95 F)
Operating humidity	8% to 90% non-condensing
Management	Cisco Intersight account and claimed device connectivity

Ordering information

The base platform is a Supermicro SKU offered through Cisco. Cisco ordering PIDs for the configured system are in progress. To order, contact your Cisco account representative or visit the [Cisco Ordering Home Page](#).

Table 7. Ordering information

Part number	Product description
SYS-422GL-FNR2-1	Supermicro 4U MGX Dual-Root PCIe 8-GPU Server, 16 DIMM
SYS-422GL-FNR2-2	Supermicro 4U MGX Dual-Root PCIe 8-GPU Server, 24 DIMM
SVC-NVSTDSUP-3Y	NVIDIA Enterprise Support Entitlement, 3Yr

Warranty information

Warranty terms for this Cisco-offered Supermicro platform are standard Supermicro warranty terms. Support is delivered through Cisco CX Services under the applicable service contract, contact your Cisco sales specialist for more details.

Cisco and Partner Services

Cisco CX and Supermicro Services cover this platform from planning through deployment and ongoing operations, with dedicated support contracts. For more information, visit <https://www.cisco.com/go/services>.

Cisco Capital

Flexible payment solutions to help you achieve your objectives

Cisco Capital makes it easier to get the right technology to achieve your objectives, enable business transformation and help you stay competitive. We can help you reduce the total cost of ownership, conserve capital, and accelerate growth. In more than 100 countries, our flexible payment solutions can help you acquire hardware, software, services and complementary third-party equipment in easy, predictable payments. [Learn more](#).

Learn more

For more on Cisco AI infrastructure, visit <https://www.cisco.com/go/ai>. Full platform specifications are available on the Supermicro product page for the SYS-422GL-FNR2.

Learn more about Cisco and Supermicro partnership:

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