



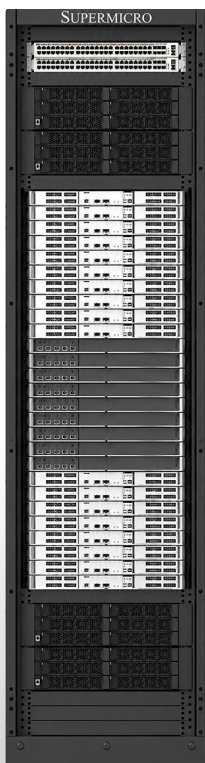
Cisco AI Rack for NVIDIA Vera Rubin NVL72

(Supermicro SRS-VR200-NVL72-M0)

Contents

Product overview	2
Features and benefits.....	3
Prominent feature.....	4
Platform support	4
Product sustainability	5
Product specifications.....	6
System requirements	7
Ordering information	7
Warranty information.....	8
Cisco and Partner Services	8
Cisco Capital.....	8
Learn more.....	8

A complete, factory-integrated NVIDIA Vera Rubin NVL72 rack with 72 Rubin GPUs and 36 Vera CPUs in a single NVLink 6 domain, sold and supported by Cisco and Supermicro and managed through Cisco Nexus One, delivered without a coolant distribution unit and paired with in-row CDU infrastructure scoped through Cisco sales.



Product overview

This is the Rubin-generation successor to the GB300 NVL72 rack in the Cisco AI compute portfolio, ordered as one product. The underlying hardware is the Supermicro NVIDIA Vera Rubin NVL72 system, built on the third-generation NVIDIA MGX rack architecture: 18 liquid-cooled compute trays, 9 NVLink switch trays, and 4 power shelves, integrated and tested as a complete rack before delivery. The rack ships without a coolant distribution unit; it is designed for in-row CDU deployments, and all liquid cooling infrastructure is scoped separately. For information on any liquid cooling components, or for a quote on liquid cooling products specifically, contact your Cisco sales specialist. Cisco handles the selling motion, support through Cisco CX Services, and lifecycle management through Cisco software. Supermicro builds and integrates the hardware.

Figure 1. Cisco AI Rack for NVIDIA Vera Rubin NVL72 (Supermicro SRS-VR200-NVL72-M)

Each compute tray carries four NVIDIA Rubin GPUs and two NVIDIA Vera CPUs (88 custom Arm cores each), the rack totals 72 Rubin GPUs and 36 Vera CPUs connected through sixth-generation NVLink at 3.6 TB/s per GPU, 260 TB/s of aggregate scale-up bandwidth. GPU memory totals 20.7 TB of HBM4; with 54 TB of CPU-attached LPDDR5X, the rack presents roughly 75 TB of fast memory. Rack-level compute reaches up to 3.6 exaflops of NVFP4 inference performance. Target workloads are frontier-scale training and agentic and reasoning-model inference at the highest density NVIDIA offers.

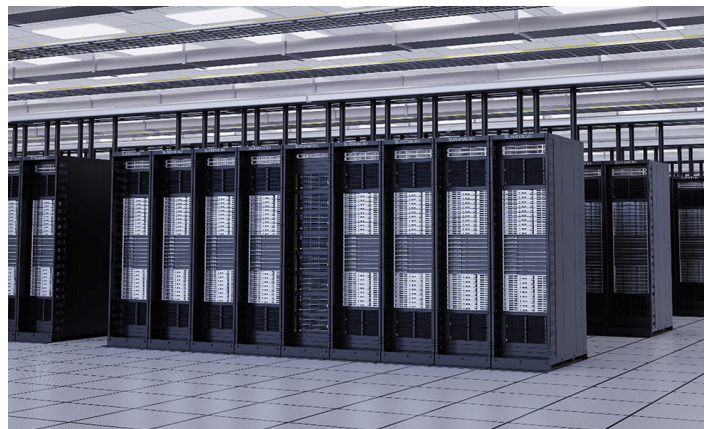


Figure 2. Cisco AI Cluster for NVIDIA Vera Rubin NVL72



Features and benefits

Table 1 summarizes the primary features and benefits of the platform

Table 1. Features and benefits

Feature	Benefit
72x NVIDIA Rubin GPUs in one NVLink 6 domain	The whole rack behaves as a single accelerator, with double the per-GPU interconnect bandwidth of the Blackwell generation
36x NVIDIA Vera CPUs (88 custom Arm cores each)	CPU and GPU share coherent memory over NVLink-C2C with a purpose-built CPU for the Rubin generation
Approx. 75 TB fast memory (20.7 TB HBM4 plus 54 TB LPDDR5X)	Roughly double the memory footprint of GB300 NVL72; frontier models and KV caches stay in one rack
Up to 3.6 exaflops NVFP4 inference per rack	Rack-level throughput for reasoning and agentic AI workloads that previously required multiple racks
9x NVLink 6 switch trays, 260 TB/s aggregate scale-up bandwidth	Non-blocking all-to-all GPU communication without external switching
NVIDIA ConnectX-9 SuperNICs at up to 1.6 Tb/s per GPU	Double the scale-out bandwidth of the ConnectX-8 generation for multi-rack growth
NVIDIA BlueField-4 DPUs	Offloads north-south networking, storage, and security with a 64-core Grace-based processor
Factory-integrated, direct liquid-cooled rack for in-row CDU deployments	Arrives built and tested; warm-water cooling cuts energy and water consumption
Cisco Hyperfabric management	SaaS Management - Hyperfabric for compute and networking On-prem management - Cisco Software or 3 rd party software for compute and Cisco Software for networking
Cisco CX Services and Supermicro support	Dedicated support contracts from both Supermicro and Cisco CX Services for each rack-scale system



Prominent feature

The generational leap, same rack discipline: NVLink 6, HBM4, and 75 TB of fast memory.

Vera Rubin NVL72 is what the rack-as-computer idea looks like one generation later. Against the GB300 NVL72 it replaces, the same 72-GPU, 36-CPU footprint moves to NVLink 6 at twice the per-GPU bandwidth, HBM4 at roughly 1.4 PB/s of aggregate memory bandwidth, a purpose-built Vera CPU instead of Grace, ConnectX-9 scale-out at up to 1.6 Tb/s per GPU, and BlueField-4 DPUs. Total fast memory roughly doubles to 75 TB. Nothing about the operating model changes: one Cisco order, dedicated Cisco and Supermicro support contracts, one management plane (for SaaS deployments) through Cisco Nexus One (Cisco Hyperfabric).

The cooling boundary is explicit. This rack is offered without a coolant distribution unit and is designed for in-row CDU infrastructure that serves multiple racks. It can also be deployed with Liquid-to-Air Sidecar CDUs. For information on any liquid cooling components, or for a quote on liquid cooling products specifically, contact your Cisco sales specialist.

Platform support

Cisco offers this Supermicro platform in a defined rack configuration, listed in Table 2. Final specifications are yet to be defined.

Table 2. Cisco selected configuration, NVIDIA Vera Rubin NVL72 rack

Item	Selected configuration
PID	SRS-VR200-NVL72-M0 (Supermicro-built Vera Rubin NVL72 rack, without coolant distribution unit)
Rack base	1x NVIDIA Vera Rubin NVL72 rack: 18x compute trays, 9x NVLink 6 switch trays, 4x 110 kW power shelves
GPU/CPU	72x NVIDIA Rubin GPUs and 36x NVIDIA Vera CPUs (4x Rubin GPUs and 2x Vera per tray)
Memory	Approx. 75 TB fast memory: 20.7 TB HBM4 plus 54 TB LPDDR5X
M.2 boot drives	To be defined
NVMe storage	To be defined
East-west networking	NVIDIA ConnectX-9 SuperNICs at up to 1.6 Tb/s per GPU (configuration to be defined)
North-south networking	NVIDIA BlueField-4 DPUs (configuration to be defined)



Item	Selected configuration
Liquid cooling	No CDU included; designed for in-row CDU deployments. Contact your Cisco sales specialist for liquid cooling component information and quotes
Software management	Cisco Hyperfabric
Support	Cisco CX Services
Availability	Quotable ahead of hardware availability; availability dates to be announced

Product sustainability

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

Table 3. 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 4 reflects NVIDIA platform specifications and Supermicro announcements for the Vera Rubin NVL72 rack. Supermicro has not yet published the full data sheet for this SKU; product specifications may change without notice.

Table 4. Specifications, Supermicro NVIDIA Vera Rubin NVL72 rack

Category	Specification
Form factor	Rack-scale system on the third-generation NVIDIA MGX architecture; 48U and 52U rack enclosure options
Compute trays	18x liquid-cooled trays; each with 4x NVIDIA Rubin GPUs and 2x NVIDIA Vera CPUs
GPU	72x NVIDIA Rubin; 288 GB HBM4 per GPU, TSMC 3 nm class
CPU	36x NVIDIA Vera; 88 custom Arm cores per CPU, up to 1.5 TB LPDDR5X each
Memory	20.7 TB HBM4 at approx. 1.4 PB/s aggregate bandwidth; 54 TB LPDDR5X; approx. 75 TB total fast memory
Rack compute	Up to 3.6 exaflops NVFP4 inference
GPU interconnect	NVIDIA NVLink 6 via 9x switch trays; 3.6 TB/s per GPU, 260 TB/s aggregate scale-up bandwidth
Scale-out networking	NVIDIA ConnectX-9 SuperNICs, up to 1.6 Tb/s per GPU; NVIDIA Quantum-X800 InfiniBand or Spectrum-X Ethernet fabrics
North-south/storage	NVIDIA BlueField-4 DPUs (64-core Grace-based, up to 800 Gb/s)
Storage	Per-tray M.2 boot and NVMe configuration to be defined
Cooling	Direct-to-chip liquid cooling across compute and switch trays; no CDU in-rack, designed for in-row CDU with warm-water operation
Power	4x 110 kW power shelves with redundant 18.3 kW power supply units
System management	Cisco Nexus One (Hyperfabric)



System requirements

Table 5 lists what must be in place before installation. This is a rack-scale liquid-cooled system delivered without a CDU; for information on any liquid cooling components, or for a quote on liquid cooling products, contact your Cisco sales specialist.

Table 5. System requirements

Requirement	Description
Facility space	Full-rack footprint, fixed enclosure; verify delivery path, floor loading, and service clearance for an integrated rack
Power	Facility feeds sized for the configured rack, delivered through 4x 110 kW power shelves; final operating power is confirmed at configuration time
Liquid cooling	In-row coolant distribution unit and facility water loop, provided separately from this PID; components and quotes scoped with your Cisco sales specialist
Scale-out fabric	Front end fabrics on Cisco Silicon One® and Frontend/Backend fabrics on NVIDIA Spectrum-X Ethernet switch silicon on Cisco N9100 series for multi-rack deployments
Management	Cisco Nexus One (Cisco Hyperfabric management)

Ordering information

The rack is a Supermicro-built system offered through Cisco under the part number below, orderability starting October 2026. No coolant distribution unit is included. Liquid cooling components (in-row CDUs, manifolds, and related infrastructure) are quoted separately: contact your Cisco sales specialist for information on any liquid cooling components or for a quote on liquid cooling products specifically. To order, contact your Cisco account representative or visit the Cisco Ordering Home Page.

Table 6. Ordering information

Part number	Product description
SRS-VRRBN-NVL72-M0	Supermicro 48U liquid-cooled NVIDIA Vera Rubin NVL72 rack, no In-Rack CDU
LCS-SCDU-1K3LR001	1.8MW In-Row CDU, L2L, N+1 Pumps
LCM-SMLCPM3-3Y	Liquid Cooling Maintenance Service, SMLCPM3, 3-Year
LCS-SCDU-200AR001	200kW L2A Sidecar CDU, N+1 Pumps

Part number	Product description
SFT-ODLCRSUPSTD-5Y	BMC Software Support, Standard, 5Yr
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 and Supermicro under the applicable service contracts, 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 will be available on the Supermicro Vera Rubin solutions page once published.

Learn more about Cisco and Supermicro partnership:

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