Supermicro NVIDIA Vera Rubin NVL72

Rack-Scale AI with 72 Liquid-Cooled NVIDIA Rubin GPUs and 36 NVIDIA Vera CPUs

The Highest Level of AI Performance and Scalability

- Rack-scale solution providing 72 NVIDIA Rubin GPUs and 36 NVIDIA Vera CPUs per rack, operating as a single machine
- NVIDIA Rubin GPUs feature up to 288GB HBM4 memory per GPU, delivering 20.7 TB of HBM4 and 1.6 PB/s of aggregate HBM4 bandwidth per rack
- Sixth-generation NVIDIA NVLink™ provides 216 TB/s of scale-up bandwidth per rack
- Supermicro DLC-2 Direct Liquid Cooling total solution stack, purpose-built for the NVIDIA Vera Rubin platform
- Up to 1.6 Tb/s NVIDIA Quantum-X800 InfiniBand or NVIDIA Spectrum™-X Ethernet with integrated NVIDIA ConnectX®-9 SuperNICs

An AI Factory in a Rack, Built for the Agentic Era

Supermicro's NVIDIA Vera Rubin NVL72 is designed to deliver the token throughput that today's agentic AI workloads demand, targeting up to 10x the per-watt throughput of NVIDIA Blackwell. As frontier models grow in parameter count, the NVIDIA Vera Rubin platform relieves bottlenecks in GPU memory, scale-up interconnect, and scale-out networking simultaneously, with up to 2.8x HBM bandwidth, 2x NVLink bandwidth, and 2x per-GPU network bandwidth versus Blackwell. One rack integrates 72 NVIDIA Rubin GPUs and 36 NVIDIA Vera CPUs across eighteen 1U liquid-cooled compute trays, with up to 288GB of HBM4 per GPU for 20.7 TB per rack. Supermicro's Data Center Building Block Solutions® delivers a total liquid-cooled solution to accelerate time-to-online.

Designed to Push the Limits of Scalability

Each rack provides 72 Rubin GPUs interconnected with liquid-cooled NVIDIA Sixth-generation NVLink switches, doubling interconnect speeds to 3 TB/s bi-directional bandwidth per GPU and 216 TB/s of aggregate scale-up bandwidth per rack. The rack operates as a single coherent NVLink domain rather than 72 separate GPUs, making it suitable for frontier AI model training and large-scale agentic AI inference.

Scale out beyond the 72-GPU NVLink domain to build clusters of any size. Each compute tray has four NVIDIA ConnectX®-9 SuperNICs providing up to 1.6 Tb/s of bandwidth per GPU over NVIDIA Spectrum-X Ethernet or NVIDIA Quantum-X800 InfiniBand, plus an NVIDIA BlueField®-4 DPU for up to 800 Gb/s north-south throughput..

Ready for Plug-and-Play Deployment

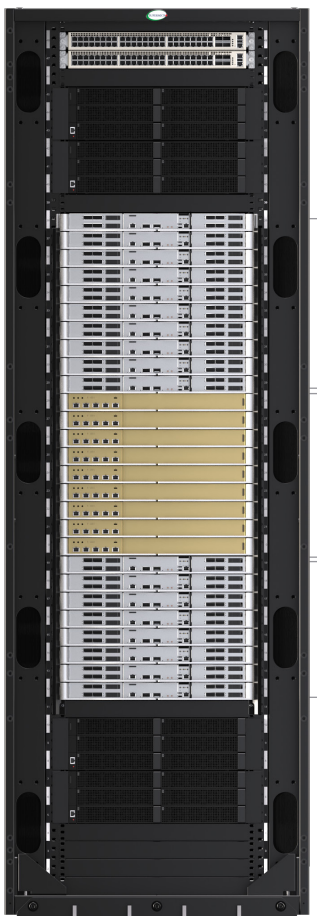
Supermicro's Data Center Building Block Solutions® accelerate time-to-market and time-to-online by offering a total solution with all critical computing and cooling infrastructure, along with on-site services and support. From individual GPUs to full racks and facility-side site infrastructure, Supermicro enables end-to-end deployment with ultimate flexibility.

Supermicro Data Center Building Block Solutions®:

- **Speed up Time-to-Market and Time-to-Online**
- **One-Stop Shop with Optimized Quality and Performance**
- **Reduce Power, OPEX, and CAPEX**

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Rack Scale Design Close-up



Networking

- NVIDIA Spectrum-X Ethernet or NVIDIA Quantum-X800 InfiniBand for up to 1.6 Tb/s compute fabric
- Converged fabric with up to 800 Gb/s north-south bandwidth
- 1G/10G ToR management switches

10 Compute Trays

- 4x Rubin GPUs and 2 Vera CPUs each

Compute Interconnect

- 9x Sixth-generation NVIDIA NVLink™ switches per rack providing 3 TB/s bandwidth per GPU

8 Compute Trays

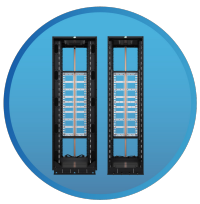
- 4x Rubin GPUs and 2 Vera CPUs each

Liquid-Cooling Options

- Supermicro In-Row CDU: up to 5 MW cooling capacity CDU with redundant pumps
- Optional: Supermicro Rear Door Heat Exchanger
- Optional: Supermicro L2A Sidecar CDU

Direct Liquid Cooling and Supermicro Data Center Building Block Solutions® (DCBBS)

Supermicro's DCBBS covers all critical data center components and management software for liquid cooling suitable for a wide range of data center environments.



MGX 1.2 Racks

Purpose-built 48U/52U racks engineered exclusively for the NVIDIA Vera Rubin NVL72 to streamline deployment and operation

In-Row CDU

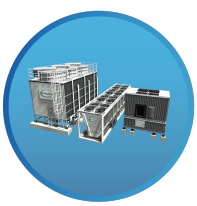
Up to 5 MW cooling capacity in-row CDU enables a single CDU to cool multiple racks of systems. Ensures uptime with dual redundant pumps.

L2A Sidecar CDU

200 kW/500 kW sidecar CDU dissipates heat from liquid to air, allowing for simple deployment when full liquid cooling is not possible.

Rear Door Heat Exchanger

Mounts to rack to remove heat at the source reducing data center air conditioning requirements



Cooling Infrastructure

Water cooling towers in scalable configurations and dry coolers that eliminate water consumption

Management Software Suite

Unify infrastructure control, automate deployment pipelines, and optimize multi-tenant GPU cloud operations

Rack Specifications

SRS-VR-NVL72

Overview	48U rack with 72x Rubin GPUs and 36x Vera CPUs
Rack elevation (from top-to-bottom)	2x ToR switches, 2x 3U Power Shelves, 10x Compute Trays, 9x NVLink Switches, 8x Compute Trays, 2x 3U Power Shelves
Total GPU Memory	20.7 TB HBM4 total, 1.6 PB/s aggregate bandwidth
Total CPU Memory	Up to 54 TB LPDDR5X (SOCAMM) total
NVLink Fabric	72-GPU NVLink via 9 NVLink 6 switch trays, 216 TB/s scale-up bandwidth
Power	4x 3U 110kW power shelves
Cooling	DCBBS direct liquid-cooling stack with up to 5 MW cooling capacity in-row CDU

Compute Tray Specifications

Overview	4x Rubin GPUs and 2x Vera CPUs per tray
GPU	4x NVIDIA Rubin GPUs, each with up to 288 GB HBM4 memory
CPU	2x NVIDIA Vera CPUs, each with up to 1.5TB LPDDR5X memory
NVLink	Sixth-generation NVLink, 3 TB/s bandwidth per GPU
Scale-out Networking	4x NVIDIA ConnectX-9 SuperNICs per compute tray (8 ports), 1.6 Tb/s scale-out bandwidth per GPU
Converged Networking	1x NVIDIA BlueField-4 DPU per compute tray, 800Gb/s north-south bandwidth
Storage	Up to 4× EDSFF E1.S NVMe drive bays 1x E1.S NVMe boot drive

© 2026 Copyright Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are the property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.