NVIDIA GB200 NVL72 SuperCluster

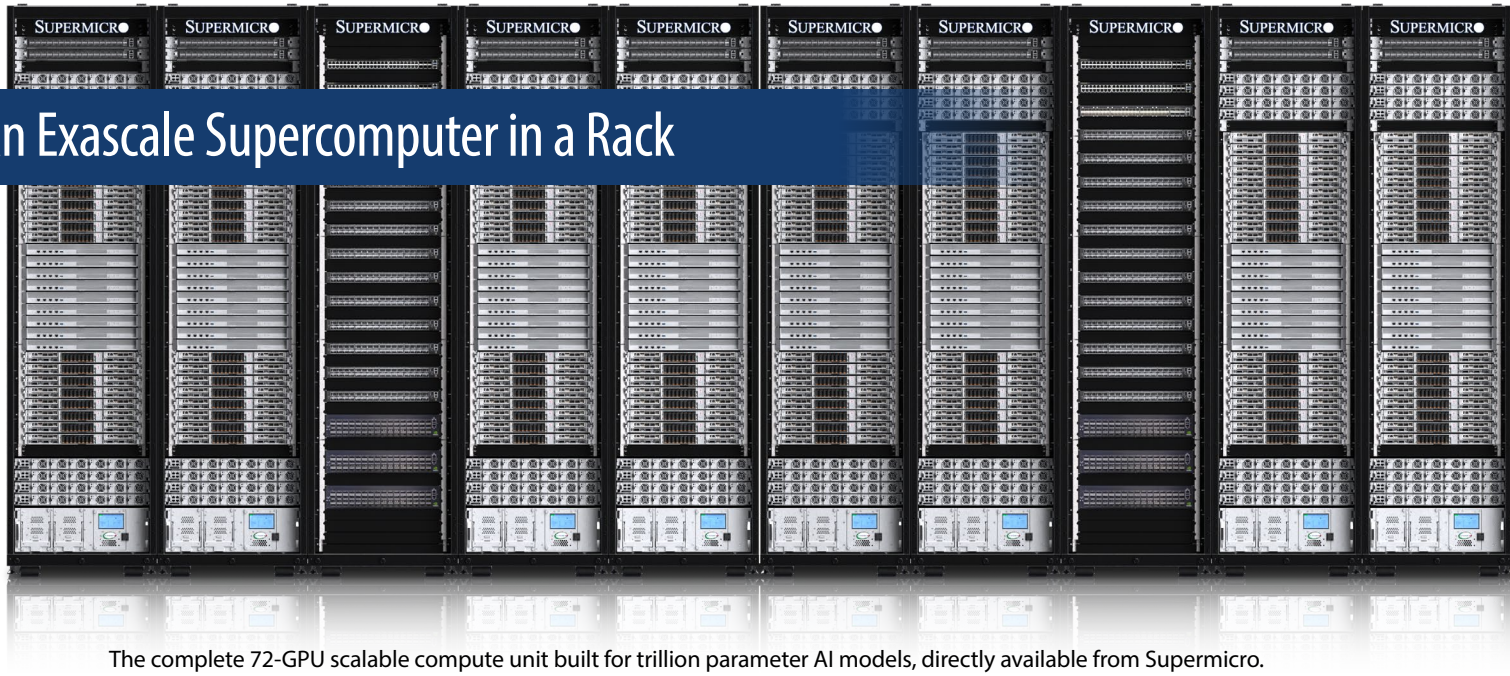
AI Data Center End-to-End Liquid-Cooling Solutions

- ✓ 5,000+ Racks Per Month Capacity with Worldwide Manufacturing
- ✓ Onsite Deployment Services and Management Software
- ✓ Unmatched Time-to-Deployment with Extensive Experience

October 2024

NVIDIA GB200 NVL72 SuperCluster

An Exascale Supercomputer in a Rack



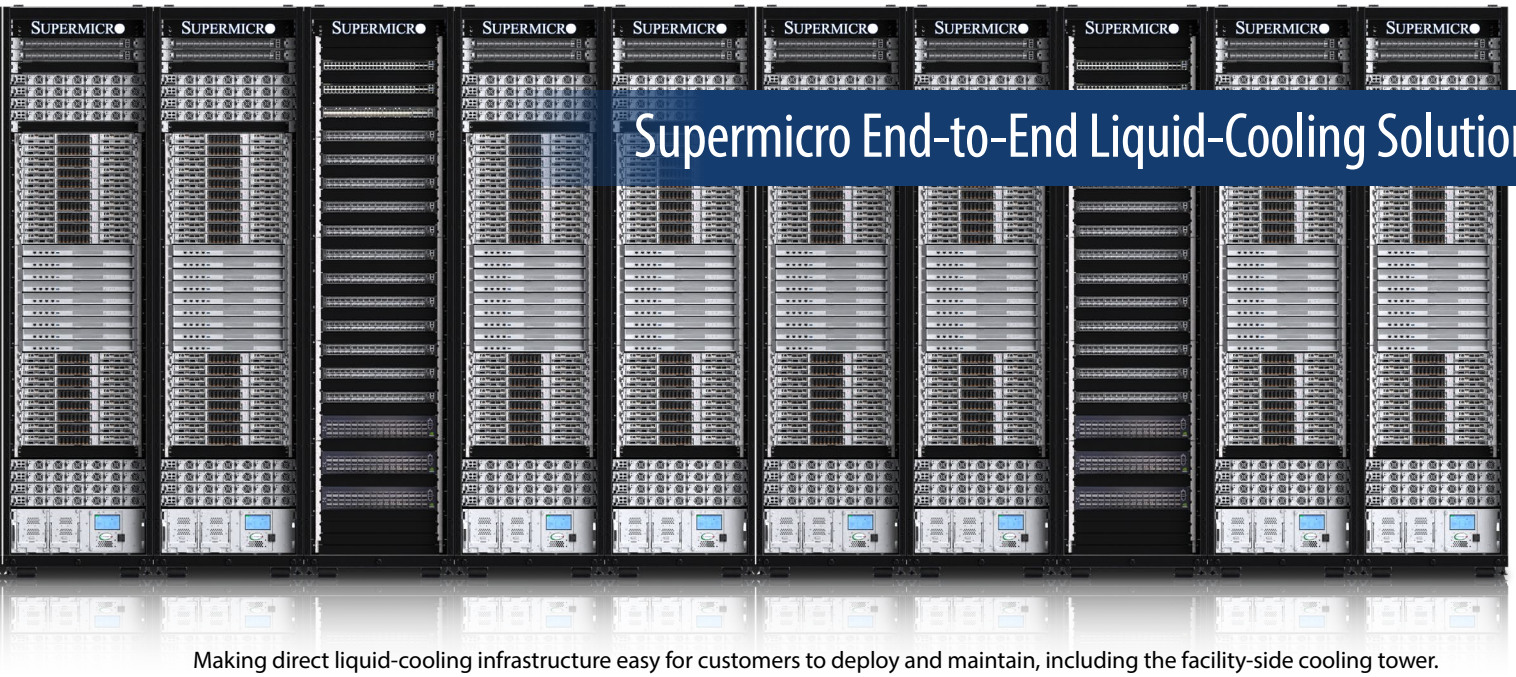
The complete 72-GPU scalable compute unit built for trillion parameter AI models, directly available from Supermicro.

The Most Powerful and Efficient NVIDIA Blackwell Platform

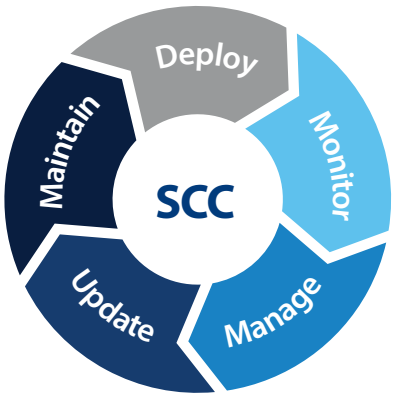
Supermicro accelerates the industry's transition to liquid-cooled data centers with NVIDIA Blackwell to deliver a new paradigm of energy-efficiency for the rapidly heightened energy demand of AI infrastructure. With extensive experience deploying large scale direct-to-chip (DLC) liquid-cooled AI systems, Supermicro's leading Liquid-Cooling technology advancement powers NVIDIA GB200 NVL72, an exascale computing in a single rack, providing up to 25x more performance at the same power than the previous generation could offer.

- **72 NVIDIA Blackwell GPUs**: acting as one GPU with a massive pool of HBM3e memory to deliver the most efficient exascale computing in a rack
- **Pioneers in Liquid-Cooling**: end-to-end Liquid-Cooling solution with up to 40% reduction in electricity cost for data center
- **Unmatched Manufacturing Scale**: with the largest Liquid-Cooling rack-level manufacturing capacity, Supermicro ensures timely and high-quality deployment of the GB200 NVL72, supported by production facilities in San Jose, CA, Europe, and Asia
- **Comprehensive Service Offering**: from proof of concept to full-scale deployment, Supermicro is one-stop shop, providing all necessary parts, networking solutions, and on-site installation services
- **Advanced Networking Ready**: Supermicro is at the forefront of adopting NVIDIA BlueField®-3 SuperNIC, Spectrum™-X, Quantum-2, and next generation 800 Gb/s networking platforms

Supermicro End-to-End Liquid-Cooling Solution



Making direct liquid-cooling infrastructure easy for customers to deploy and maintain, including the facility-side cooling tower.

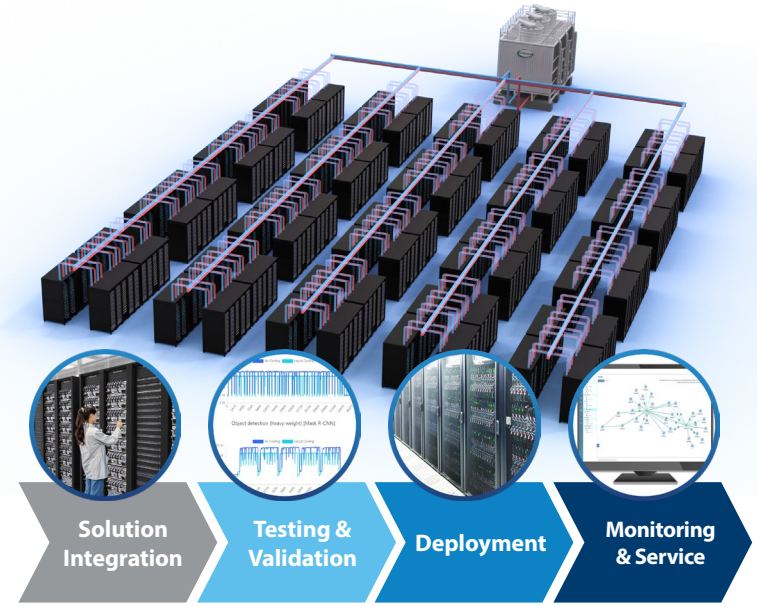


SuperCloud Composer (SCC) for Liquid-Cooled Data Center Management

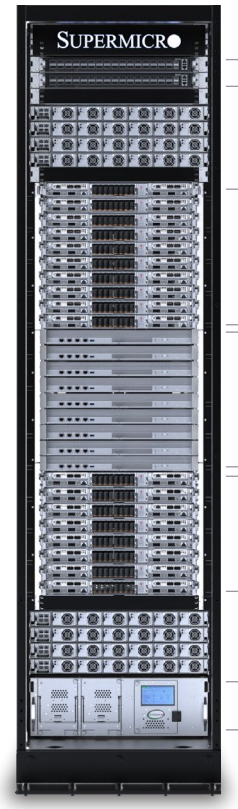
Supermicro's comprehensive datacenter management platform, SuperCloud Composer software, provides powerful tools to monitor vital information on liquid-cooled systems and racks, coolant distribution units, and cooling towers, including pressure, humidity, pump and valve conditions, and more. SuperCloud Composer's Liquid-Cooling Consult Module (LCCM) optimizes the operational cost and manages the integrity of liquid-cooled data centers.

Powered by Supermicro End-to-End Liquid-Cooling Solution

Supermicro NVIDIA GB200 NVL72 SuperCluster features the new advanced in-rack coolant distribution unit (CDU) and custom coldplates designed for the compute trays housing the NVIDIA GB200 Grace™ Blackwell Superchips. The NVIDIA GB200 NVL72 delivers exascale computing capabilities in a single rack with fully integrated Liquid-Cooling. It incorporates 72 NVIDIA Blackwell GPUs and 36 Grace CPUs interconnected by NVIDIA's largest NVLink™ network to date. The NVLink Switch System facilitates 130 terabytes per second (TB/s) of total GPU communications with low latency, enhancing performance for AI and high-performance computing (HPC) workloads.



NVIDIA GB200 NVL72 Rack-Scale Configuration



Management Networking
- In-band management switch
- Out-of-band management switch

10 Compute Trays
- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Compute Interconnect
- 9x NVLink Switches
- 72 GPUs and 36 CPUs interconnected at 1.8TB/s

8 Compute Trays
- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Liquid-Cooling Options
- Supermicro 250kW capacity coolant distribution unit (CDU) with redundant PSU and dual hot-swap pumps
- 240kW or 180kW capacity Liquid-to-air solution (no facility water required)

End-to-End Onsite Deployment Services

From proof-of-concept (PoC) to full-scale deployment, Supermicro is a one-stop shop, providing all necessary parts, Liquid-Cooling, networking solutions, management software, and onsite installation services. As a one-stop shop, Supermicro delivers a comprehensive, in-house Liquid-Cooling ecosystem, encompassing custom-designed cold plates optimized for various GPUs, CPUs, and memory modules, along with multiple coolant distribution unit form factors and capacity, manifolds, hoses, connectors, cooling towers, and monitoring and management software. This end-to-end solution seamlessly integrates into rack-level configurations, significantly boosting system efficiency, mitigating thermal throttling, and simultaneously reducing both the Total Cost of Ownership (TCO) and environmental impact of data center operations for the era of AI.

72-GPU Scalable Unit	SRS-GB200-NVL72-M1
GPUs	72x NVIDIA Blackwell B200 GPUs
CPUs	36x NVIDIA 72-core Grace Arm Neoverse V2
Compute Trays	18x 1U ARS-121GL-NBO
NVLink Switch Trays	9x NVLink Switch, 4-ports per compute tray connecting 72 GPUs to provide 1.8TB/s GPU-to-GPU interconnect
Power Shelves	8x 1U 33kW (6x 5.5kW PSUs), total power 132kW
Rack Dimensions (mm)	W 600 x D 1068 x H 2236
Liquid Cooing Options	• 1x in-rack Supermicro 250kW capacity CDU with redundant PSU and dual hot-swap pumps • 1.3MW capacity in-row CDU • 180kW/240kW capacity liquid-to-air solutions for facilities without cooling tower and water supply

Subject to change

Compute Tray	ARS-121GL-NBO
Overview	1U Liquid-cooled System with 2x NVIDIA GB200 Grace Blackwell Superchips
CPU and GPU	• 2x 72-core NVIDIA Grace Arm Neoverse V2 CPU • 4x NVIDIA Blackwell B200 per Superchip
GPU Memory	Up to 384GB HBM3e per Superchip (768GB per tray)
CPU Memory	Up to 480GB LPDDR5X per Superchip (960GB per tray)
Networking	4x NVIDIA NVLink Switch ports
Storage	8x E1.S PCIe 5.0 drives
Power Supply	Shared power through 4+4 rack power shelves

Subject to change

AI Data Center End-to-End Liquid-Cooling

Total Liquid-Cooling Offerings for a Wide Range of AI Data Center Environments

Coolant Distribution Unit

Liquid-to-Air Cooling Rack
(no facility water required)

Cooling Tower

NVIDIA HGX B200 8-GPU SuperClusters

New 4U Liquid-Cooled and 10U Air-Cooled Systems for NVIDIA HGX B200 8-GPU

Supermicro's leading 4U liquid-cooled systems and the new 10U air-cooled systems now support NVIDIA HGX™ B200 8-GPU and ready for production. The newly developed cold plates and the 250kW capacity in-rack coolant distribution unit maximize the performance and efficiency of the 8x systems, providing 64x 1000W Blackwell GPUs and 16x 500W CPUs in a 48U rack. The new 10U air-cooled systems support up to 4x systems fully integrated in a rack, the same density as the previous generation while providing up to 15x inference and 3x training performance.

4U Liquid-Cooled Rack Configuration

Networking

- In-band management switch
- Out-of-band IPMI management switch
- Non-blocking network
- Leaf switches in the dedicated networking rack or in the individual compute racks

Compute and Storage

- 8x SYS-422GA-NBRT-LCC or AS -4126GS-NBR-LCC per rack
- 8x NVIDIA HGX B200 8-GPU per rack
- 64x NVIDIA B200 Tensor Core GPUs
- 8x 1440GB HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA support

Liquid-Cooling Options

- Supermicro 250kW capacity Coolant Distribution Unit with redundant PSU and dual hot-swap pumps
- Supermicro Coolant Distribution Manifolds (horizontal or vertical)
- 240kW or 180kW capacity Liquid-to-air solution (no facility water rquired)

Next-Generation Scalable Unit

Scaling the infrastructure for multi-trillion parameter AI models, Supermicro is at the forefront of adopting networking innovations for both InfiniBand and Ethernet, including NVIDIA BlueField®-3 SuperNIC and ConnectX-7 at 400Gb/s, as well as ConnectX-8, Spectrum-4, and Quantum-3 to enable 800Gb/s networking for the NVIDIA Blackwell Platform. The NVIDIA Quantum-2 InfiniBand and Spectrum-X Ethernet with Supermicro's 4U liquid-cooled and 8U air-cooled NVIDIA HGX H100 and H200 system clusters now power one of the largest AI deployments to date.

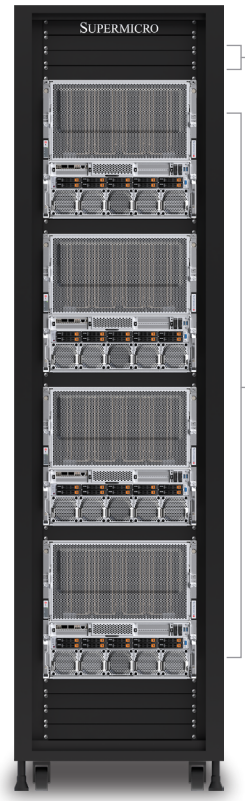
64-GPU Scalable Unit	
GPUs	8x NVIDIA HGX B200 8-GPU (64 GPUs)
CPUs	16x Intel® Xeon® 6 or AMD EPYC™ 9005 Series Processors
GPU Systems	8x SYS-422GA-NBRT-LCC / AS -4126GS-NBR-LCC
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking*	• NVIDIA Quantum-2 InfiniBand 400G NDR • NVIDIA Spectrum-X Ethernet 400Gb/s • 2x Ethernet ToR management switch
Rack Dimension*	48U 800mm x 1200mm
Liquid Cooing Options*	• 1x in-rack Supermicro 250kW capacity CDU with redundant PSU and dual hot-swap pumps • 1.3MW capacity in-row CDU • 180kW/240kW capacity liquid-to-air solutions for facilities without cooling tower and water supply

*Subject to change

4U 8-GPU System	SYS-422GA-NBRT-LCC / AS-4126GS-NBR-LCC
Overview	4U Liquid-cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6 or AMD EPYC™ 9005 Series Processors
Memory	• Up to 6TB DDR5-6400 or MRDIMM 8800MT/s (Intel) • Up to 9TB DDR5-6000 (AMD)
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
Networking*	• 8x NVIDIA ConnectX®-7 or BlueField®-3 SuperNICs • 2x NVIDIA ConnectX®-7 Dual-port 200Gbps/NDR200 QSFP112 NICs • 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage
Storage	8 front hot-swap 2.5" PCIe 5.0 x4 NVMe drive bays
Power Supply	4x 6.6kW redundant Titanium Level power supplies

*Subject to change.

10U Air-Cooled Rack Configuration



Networking

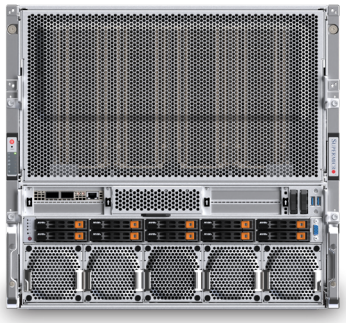
- In-band management switch
- Out-of-band IPMI management switch
- Non-blocking network
- Leaf switches in the dedicated networking rack or in the individual compute racks

Compute and Storage

- 4x SYS-A22GA-NBRT or AS -A126GS-TNBR per rack
- 4x NVIDIA HGX B200 8-GPU per rack
- 32x NVIDIA B200 Tensor Core GPUs
- 4x 1440GB HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA Support

Air or Liquid-Cooling

Multiple system configurations and form factors are available depending on the specific cooling requirements and infrastructure of data centers. An all-new 10U form factor enables support for NVIDIA HGX B200 8-GPU in air-cooled environments, with a modular GPU tray capable of accommodating enlarged heatsinks for enhanced thermal performance and easy maintenance. For maximum GPU density, a liquid-cooled 4U architecture can be integrated using Supermicro's end-to-end liquid-cooling solution, allowing up to 8 systems in a standard 48U rack for a total of 64 GPUs.



32-GPU Scalable Unit	
GPUs	4x NVIDIA HGX B200 8-GPU (32 GPUs)
CPUs	8x Intel® Xeon® 6 or AMD EPYC™ 9005 Series Processors
GPU Systems	SYS-A22GA-NBRT or AS -A126GS-TNBR
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking*	• NVIDIA Quantum-2 InfiniBand 400G NDR • NVIDIA Spectrum-X Ethernet 400Gb/s • 2x Ethernet ToR management switch
Rack Dimension	48U 750mm x 1200mm

*Subject to change

10U 8-GPU System	SYS-A22GA-NBRT / AS-A126GS-TNBR
Overview	10U Air-cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6 or AMD EPYC™ 9005 Series Processors
Memory	• Up to 6TB DDR5-6400 or MRDIMM 8800MT/s (Intel) • Up to 9TB DDR5-6000 (AMD)
GPU	• NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) • 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
Networking*	• 8x NVIDIA ConnectX®-7 or BlueField®-3 SuperNICs • 2x NVIDIA ConnectX®-7 Dual-port 200Gbps/NDR200 QSFP112 NICs • 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage
Storage	10 front hot-swap 2.5" PCIe 5.0 x4 NVMe drive bays
Power Supply	6x 5250W redundant Titanium Level power supplies

*Subject to change

Supermicro AI Data Center End-to-End Solutions

Worldwide Headquarters

Super Micro Computer, Inc.
980 Rock Ave.
San Jose, CA 95131, USA
Tel: +1-408-503-8000

EMEA Headquarters

Super Micro Computer, B.V.
Het Sterrenbeeld 12, 5215 ML,
's-Hertogenbosch, The Netherlands
Tel: +31-73-640-0390

APAC Headquarters

Super Micro Computer, Taiwan Inc.
3F, No. 150, Jian 1st Rd., Zhonghe Dist.,
New Taipei City 235, Taiwan
Tel: +886-2-8226-3990

www.supermicro.com

MKT-0002_2024-GB200-NVL72_R08

Please Recycle