Supermicro NVIDIA GB200 NVL72

Liquid-cooled Exascale Compute in a Rack with 72 NVIDIA Blackwell GPUs

Scalable Compute Unit Built For Trillion Parameter AI Models

- **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**: total 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, Spectrum™-X, Quantum-2, and next generation 800 Gb/s networking platforms

An Exascale Supercomputer in a Rack

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 25 times more energy efficiency than the previous generation.

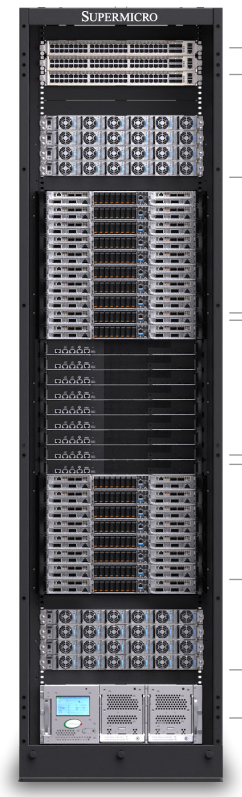
Powered by Supermicro End-to-End Liquid-cooling Solution

Supermicro NVIDIA GB200 NVL72 SuperCluster features the new advanced in-rack or in-row cooling distribution unit (CDU) and 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.

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 distiribution 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.

Rack Scale Design Close-up



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

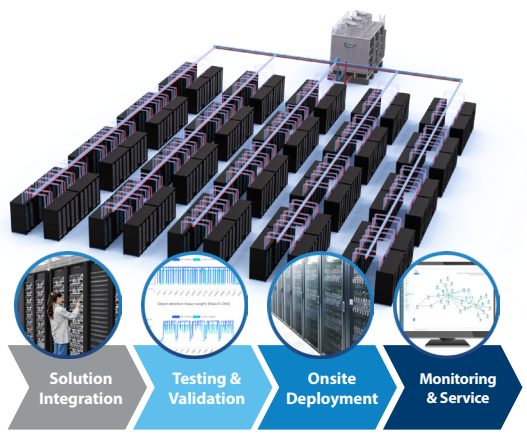
Compute Interconnect
- 9x NVLink Switches
- 72 GPUs and 36 CPUs inteconnected at 1.8TB/s

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)

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.



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)	2236mm x 600 mm x 1068mm
Liquid Cooling Options	1x in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps 1.3MW capacity in-row CDU Optional 180kW/240kW capacity liquid-to-air solutions for facilities without cooling tower and water supply

Compute Tray	ARS-121GL-NBO
Overview	1U Liquid-cooled System with 2 NVIDIA GB200 Grace Blackwell Superchips
CPU and GPU	2 72-core NVIDIA Grace Arm Neoverse V2 CPUs 4 NVIDIA Blackwell Tensor Core GPUs
GPU Memory	Up to 384GB HBM3e per Superchip (768GB per tray)
CPU Memory	Up to 480GB LPDDR5X per Superchip (960GB per tray)
Networking	4 NVIDIA NVLink Switch ports (rear) 4 single-port NVIDIA ConnectX®-7 NICs (front) Up to 2 NVIDIA BlueField®-3 DPUs (front)
Storage	Up to 8 E1.S PCIe 5.0 drives
Power Supply	Shared power through 4+4 rack power shelves