Liquid-Cooled AI SuperCluster

With 256 NVIDIA HGX™ B200 GPUs, 32 4U Liquid-cooled Systems

Unprecedented Density and Efficiency with Next-Generation Liquid Cooling

Supermicro's SuperCluster, accelerated by the NVIDIA Blackwell Platform, empowers the next stage of AI, defined by new breakthroughs, including the evolution of scaling laws and the rise of reasoning models. The SuperCluster provides the core infrastructure elements necessary to scale the NVIDIA Blackwell Platform and deploy the pinnacle of AI training and inference performance. SuperCluster simplifies the complexities of AI infrastructure by providing a fully validated liquid-cooled AI cluster with a plug-and-play deployment experience.

These new SuperCluster offerings powered by the NVIDIA Blackwell Platform are available in 42U, 48U, or 52U configurations. The upgraded cold plates and 250kW coolant distribution unit (CDU) more than double the cooling capacity of the previous generation. The new vertical coolant distribution manifold (CDM) means that horizontal manifolds no longer occupy valuable rack space. NVIDIA Quantum InfiniBand or NVIDIA Spectrum™ networking in a centralized rack enables a non-blocking, 256-GPU scalable unit in five racks, or an extended 768-GPU scalable unit in nine racks.

Supermicro NVIDIA HGX B200 8-GPU Systems, Liquid-Cooled

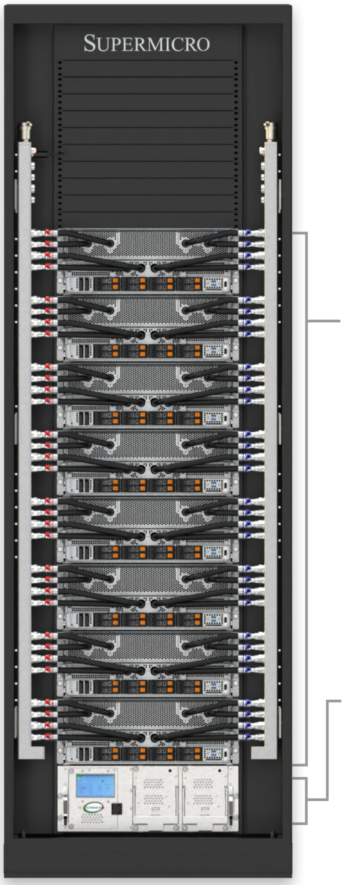
Supermicro NVIDIA HGX Systems power the world's largest liquid-cooled AI data centers. The new 4U NVIDIA HGX B200 8-GPU system features new cold plates and tubing design that further enhances efficiency and serviceability over its predecessor. The system features 8 NVIDIA Blackwell GPUs, each with 180GB HBM3e memory. The GPUs are interconnected at 1.8TB/s through the latest NVIDIA NVLink, with 1.4TB of GPU memory capacity per system.

The SuperCluster creates a massive pool of GPU resources, acting as one AI supercomputer, featuring 1:1 networking to GPU with 8x 400GB/s NVIDIA ConnectX®-7 adapters or BlueField®-3 SuperNICs, as well as 2 NVIDIA BlueField®-3 DPUs per system.

Node Configuration	SYS-422GA-NBRT-LCC / AS -4126GS-NBR-LCC / SYS-421GE-NBRT-LCC
Overview	4U Liquid-cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6900 series processors with P-cores (SYS-422GA-NBRT-LCC) Dual AMD EPYC™ 9005/9004 Series Processors (AS -4126GS-NBR-LCC) Dual 5th/4th Gen Intel® Xeon® Scalable processors (SYS-421GE-NBRT-LCC)
Memory	24 DIMMs, up to DDR5-6400 (SYS-422GA-NBRT-LCC) 24 DIMMs, up to DDR5-6000 (AS -4126GS-NBR-LCC) 32 DIMMs, up to DDR5-5600 (SYS-421GE-NBRT-LCC)
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVIDIA NVLink bandwidth with NVSwitch
Networking	8 single-port NVIDIA ConnectX®-7 NICs, or NVIDIA BlueField®-3 SuperNICs, up to 400Gbps 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 front hot-swap 2.5" NVMe drive bays 2 M.2 NVMe slots
Power Supply	4x 6600W Redundant Titanium Level power supplies

*Recommended configuration, other system memory, networking, storage options are available.

Rack Scale Design Close-up



Networking

- NVIDIA Quantum-2 400G InfiniBand switches or NVIDIA Spectrum-4 400GbE Ethernet switches dedicated for compute and storage
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

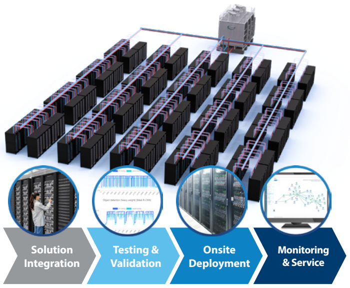
Compute

- 8x SYS-422GA-NBRT-LCC / AS -4126GS-NBR-LCC / SYS-421GA-NBRT-LCC per rack
- 64x NVIDIA HGX B200 GPUs per rack
- 11.5TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Liquid-Cooling

- Supermicro 250kW capacity Cooling Distribution Unit (CDU) with redundant PSU and dual hot-swap pumps
- Vertical Cooling Distribution Manifold (CDM)

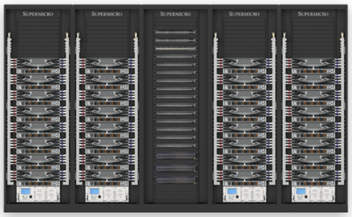
Software and Services



Software: Supermicro's SuperCloud Composer software provides management tools for monitoring and optimizing air or liquid-cooled infrastructure, delivering a complete solution from proof of concept to full-scale deployment. Manage all data center racks, including compute, storage, networking in one unified dashboard.

SuperCluster natively supports NVIDIA AI Enterprise software to accelerate time to online for production AI. NVIDIA NIM microservices allow organizations to easily access and deploy latest AI models and AI agents, fully optimized for the new NVIDIA Blackwell Platforms.

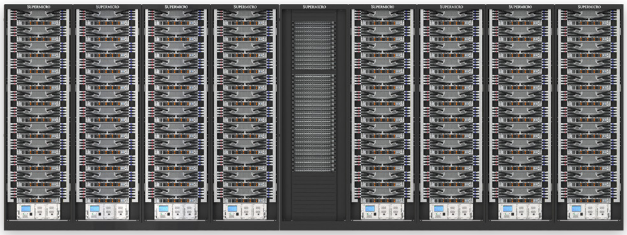
Services: Supermicro's on-site rack deployment helps enterprises build a data center from the ground up, including the planning, design, power-up, validation, testing, installation, and configuration of racks, servers, switches, and other networking equipment to meet the organization's specific needs.



32-Node Scalable Unit SRS-48UDLC-4U8N-L1

Overview	Fully integrated liquid-cooled 32-node cluster with 256 NVIDIA B200v GPUs
Compute Fabric Leaf	8x NVIDIA Quantum-2 400G InfiniBand Switch or 8x NVIDIA Spectrum-4 400GbE Ethernet Switch
Compute Fabric Spine	4x NVIDIA Quantum-2 400G InfiniBand Switch or 4x NVIDIA Spectrum-4 400GbE Ethernet Switch
In-band Management Switch	3x NVIDIA Spectrum SN4600 100GbE Ethernet Switch
Out-of-band Management Switch	2x SSE-G3748R-SMIS, 48-port 1Gbps Ethernet ToR management switch 1x SSE-F3548SR, 48-port 10Gbps Ethernet ToR management switch
Rack and PDU	5x 48U x 800mm x 1400mm PDU: 18x 415V 60A/100A 3Ph
Liquid Cooling	4x Supermicro 250kW capacity CDU with redundant PSU and dual hot-swap pumps

*Recommended configuration. Other network switch options and rack dimension and layouts are available.
*Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.



96-Node Scalable Unit SRS-52UDLC-4U12N-L1

Overview	Fully integrated liquid-cooled 96-node cluster with 768 NVIDIA B200 GPUs
Compute Fabric Leaf	24x NVIDIA Quantum-2 400G InfiniBand switch or 24x NVIDIA Spectrum-4 400GbE Ethernet switch
Compute Fabric Spine	12x NVIDIA Quantum-2 400G InfiniBand switch or 12x NVIDIA Spectrum-4 400GbE Ethernet switch
In-band Management Switch	9x NVIDIA Spectrum SN4600 100GbE Ethernet switch
Out-of-band Management Switch	6x SSE-G3748R-SMIS, 48-port 1Gbps Ethernet ToR management switch 3x SSE-F3548SR, 48-port 10Gbps Ethernet ToR management switch
Rack and PDU	9x 52U x 800mm x 1400mm PDU: 50x 415V 60A/100A 3Ph
Liquid Cooling	8x Supermicro 250kW capacity CDU with redundant PSU and dual hot-swap pumps

*Recommended configuration, other network switch options and rack dimensions and layouts are available.
*Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.