Air-Cooled AI SuperCluster

With 256 NVIDIA HGX™ B200 GPUs, 32 10U Air-Cooled Systems

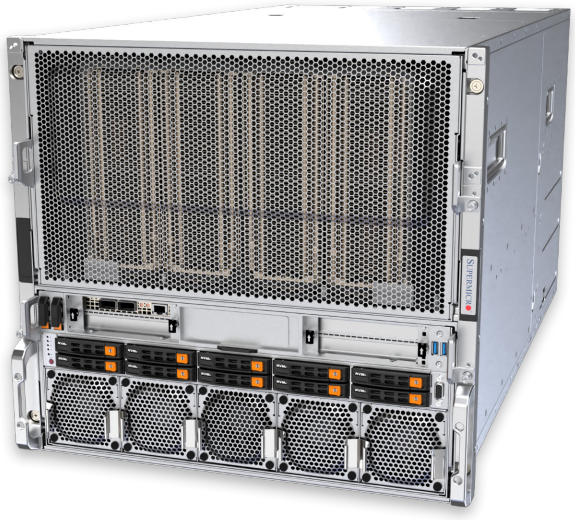
Industry Leading AI Performance with Advanced Air-Cooling Technology

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 AI cluster with a plug-and-play deployment experience.

Supermicro's new air-cooled SuperCluster is composed of the new Supermicro NVIDIA HGX B200 8-GPU systems. Featuring a redesigned 10U chassis to accommodate the thermals of its leading-edge AI compute performance, it is designed to tackle heavy AI workloads of all types, from training to fine-tuning to inference. NVIDIA Quantum InfiniBand or NVIDIA Spectrum™ networking in a centralized rack enables a non-blocking, 256-GPU scalable unit in nine racks.

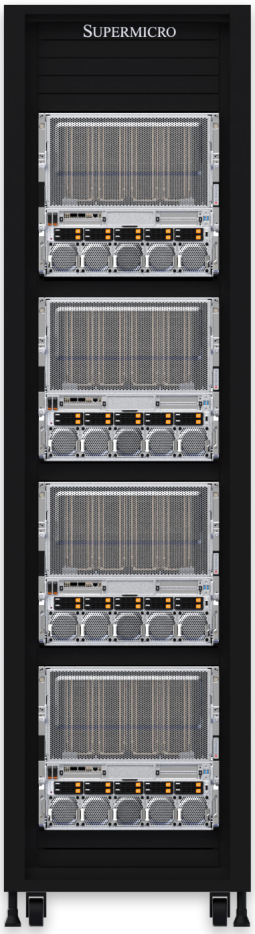
Supermicro NVIDIA HGX B200 8-GPU Systems, Air-Cooled

Supermicro NVIDIA HGX Systems are the building blocks of the world's largest AI data centers. The new 10U air-cooled NVIDIA B200 8-GPU system features an upgraded mechanical design to improve airflow over key components, including the GPU heatsink and high-speed NICs. The system features 8 NVIDIA Blackwell GPUs, each with 180GB of 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 GPU-to-NIC ratio supporting NVIDIA ConnectX®-7 NICs or BlueField®-3 SuperNICs for scaling across high-performance compute fabric.

Rack Scale Design Close-up



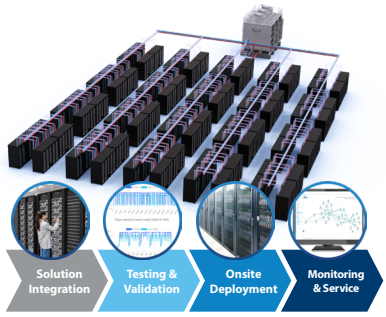
Networking

- 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

- 4x SYS-A22GA-NBRT / AS -A126GS-TNBR / SYS-A21GE-NBRT per rack
- 32x NVIDIA B200 GPUs per rack
- 5.76TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage or RoCE support

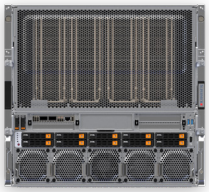
Software and Services



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.



Node configuration — SYS-A22GA-NBRT / AS -A126GS-TNBR / SYS-A21GE-NBRT

Overview	10U Air-cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6900 series processors with P-cores (SYS-A22GA-NBRT) Dual AMD EPYC™ 9005/9004 Series Processors (AS -A126GS-TNBR) Dual 5th/4th Gen Intel® Xeon® Scalable processors (SYS-A21GE-NBRT)
Memory	24 DIMMs, up to DDR5-6400 (SYS-A22GA-NBRT) 24 DIMMs, up to DDR5-6000 (AS -A126GS-TNBR) 32 DIMMs, up to DDR5-5600 (SYS-A21GE-NBRT)
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	10 front hot-swap 2.5" NVMe drive bays 2 M.2 NVMe slots
Power Supply	6x 5250W Redundant Titanium Level power supplies

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

32-Node Scalable Unit — SRS-48UAC-10U4N-A1

Overview	Fully integrated air-cooled 32-node cluster with 256 NVIDIA B200 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	9x 48U x 750mm x 1295mm
PDU	34x 208V 60A 3Ph

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