Generative AI SuperCluster

With 256 NVIDIA HGX™ H100/H200 GPUs, 32 4U Liquid-cooled Systems

Scalable Compute Unit Built For Large Language Models

- Doubling compute density through Supermicro's custom liquid-cooling solution with up to 40% reduction in electricity cost for data center
- 256 NVIDIA H100/H200 GPUs in one scalable unit
- 20TB of HBM3 with H100 or 36TB of HBM3e with H200 in one scalable unit
- 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage for training large language model with up to trillions of parameters
- Customizable AI data pipeline storage fabric with industry leading parallel file system options
- NVIDIA AI enterprise software ready

Building Blocks for Highest Density Generative AI Infrastructure Deployment

In the era of AI, a unit of compute is no longer measured by just the number of servers. Interconnected GPUs, CPUs, memory, storage, and these resources across multiple nodes in racks construct today's artificial Intelligence. The infrastructure requires high-speed and low-latency network fabrics, and carefully designed cooling technologies and power delivery to sustain optimal performance and efficiency for each data center environment. Supermicro's SuperCluster solution provides foundational building blocks for rapidly evolving Generative AI and Large Language Models (LLMs). The full turn-key data center solution accelerates time-to-delivery for mission-critical enterprise use cases, and eliminates the complexity of building a large cluster, that used to be only achievable through intensive design tuning and time-consuming optimization of supercomputing.

4U 8-GPU System, Liquid-cooled

Supermicro 4U liquid-cooled system with NVIDIA HGX H100/H200 8-GPU doubles the density of the 8U air-cooled system. Our custom direct-to-chip (D2C) cold plates keep both GPUs and CPUs at optimal temperature for sustained maximum performance. Supermicro cooling distribution unit (CDU) and manifold (CDM) are the main arteries for distributing cooled liquid to the cold plates, enabling up to 40% reduction in electricity costs for the entire data center, reducing server noise, and saving data center space.

The NVIDIA HGX H100/H200 8-GPU equipped system is ideal for training Generative AI. The high-speed interconnected GPUs through NVIDIA® NVLink®, high GPU memory bandwidth and capacity are the key for running large language (LLM) models cost effectively. The SuperCluster creates a massive pool of GPU resources acting as one AI supercomputer.

Plug-and-Play, Reduce Lead-time

The SuperCluster design with the 4U liquid-cooled systems comes with 400Gb/s networking fabrics and non-blocking architecture. The 8 nodes per rack and 32-node cluster operate as a scalable unit of compute providing a foundational building block for Generative AI Infrastructure.

Whether fitting an enormous foundation model trained on a dataset with trillions of tokens from scratch, or building a cloud-scale LLM inference infrastructure, the spine and leaf network topology allows it to scale from 32 nodes to thousands of nodes seamlessly. With fully integrated liquid-cooling out of the box, Supermicro's proven testing processes thoroughly validate the operational effectiveness and efficiency before shipping. Customers receive plug-and-play scalable units for rapid deployment.

SUPERMICRO

## Rack Scale Design Close-up

Networking

- 400G InfiniBand NDR leaf switches dedicated for compute and storage
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch Non-blocking network

Compute and Storage

- 8x SYS-421GE-TNHR2-LCC or AS -4125GS-TNHR2-LCC per rack
- 8x NVIDIA HGX H100/H200 8-GPU per rack
- 64x NVIDIA H100/H200 Tensor Core GPUs
- 5TB of HBM3 or 9TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and storage support

CDU and CDM

- Supermicro 100kW capacity Cooling Distribution Unit with redundant PSU and dual hot-swap pumps
- 8x 1U Supermicro Cooling Distribution Manifold
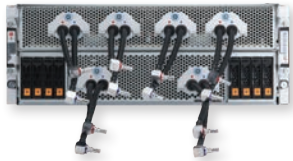
32-Node LLM Scalable Unit

The spine-leaf network fabric allows 32-node compute unit as a increment to scale to thousands of nodes. With highest network performance achievable for GPU-GPU connectivity, the SuperCluster is optimized for LLM training and high volume, high batch size inference. Plus, our L11 and L12 validation testing, and on-site deployment service provides seamless experience.

Network Fabrics

Spine switches (L2)

Leaf switches (L1)

32-node with 256 HGX H100 GPUs

32-node increment to scale
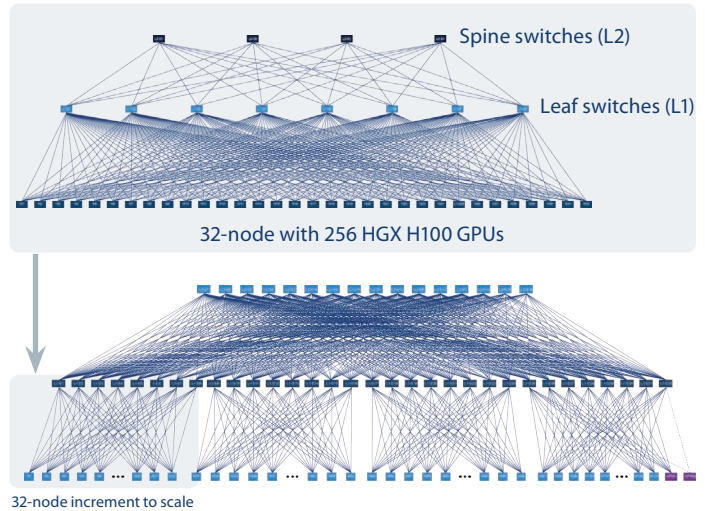
Node Configuration	SYS-421GE-TNHR2-LCC / AS-4125GS-TNHR2-LCC
Overview	4U Liquid-cooled System with NVIDIA HGX H100/H200 8-GPU
CPU	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC™ 9004 Series Processors
Memory	2TB DDR5 (recommended)
GPU	NVIDIA HGX H100/H200 8-GPU (80GB HBM3 or 141GB HBM3e per GPU) 900GB/s NVLink GPU–GPU interconnect with NVSwitch
Networking	8x NVIDIA ConnectX®-7 Single-port 400Gbps/NDR OSFP NICs 2x NVIDIA ConnectX®-7 Dual-port 200Gbps/NDR200 QSFP112 NICs 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage
Storage	30.4TB NVMe (4x 7.6TB U.3) 3.8TB NVMe (2x 1.9TB U.3, Boot) [Optional M.2 available]
Power Supply	4x 5250W Redundant Titanium Level power supplies

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

32-Node Scalable Unit	SRS-48UGPU-AI-LCSU
Overview	Fully integrated liquid-cooled 32-node cluster with 256 NVIDIA H100/H200 GPUs
Compute Fabric Leaf	8x SSE-MQM9700-NS2F, 64-port InfiniBand 400G NDR, 32 OSFP ports switch
Compute Fabric Spine	4x SSE-MQM9700-NS2F, 64-port InfiniBand 400G NDR, 32 OSFP ports switch
In-band Management Switch	3x SSE-MSN4600-CS2FC 64-port 100GbE QSFP28, 2U 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 750mm x 1200mg PDU: 18x 415V 60A 3Ph
Liquid Cooing	4x Supermicro 80kW capacity CDU with redundant PSU and dual hot-swap pumps

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

SUPERMICRO