Generative AI SuperCluster

With 256 NVIDIA HGX™ H100/H200 GPUs, 32 8U Air-cooled Systems

Industry leading Scalable Compute Unit Built For Large Language Models

- Proven industry leading architecture for large scale AI infrastructure deployments
- 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.

8U 8-GPU System

Supermicro's proven industry-leading 8U system is powering NVIDIA HGX H100/H200 8-GPU at its full potential. 8 of PCIe 5.0 slots are dedicated to 1:1 400Gb/s networking for GPUs. Each GPU is paired with 400Gb/s networking such as NVIDIA ConnectX-7 to enable NVIDIA GPUDirect RDMA and Storage so that the data flows directly to the GPU memory with the lowest latency possible.

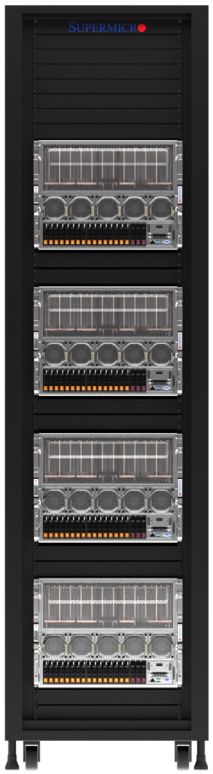
The NVIDIA HGX H100/H200 8-GPU equipped system is ideal for training Generative AI. The high-speed interconnected GPUs through NVIDIA® NVLink®, high GPU memory bandwidth and capacity are the keys to running large language (LLM) models cost-effectively. The SuperCluster creates a massive pool of GPU resources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

Plug-and-Play, Reduced Lead-time

The SuperCluster design with the 8U air-cooled systems comes with 400Gb/s networking fabrics and non-blocking architecture. The 4 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. Supermicro's proven testing processes thoroughly validate the operational effectiveness and efficiency before shipping. Customers receive plug-and-play scalable units for rapid deployment.

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
- Leaf switches in the dedicated networking rack or in the individual compute racks

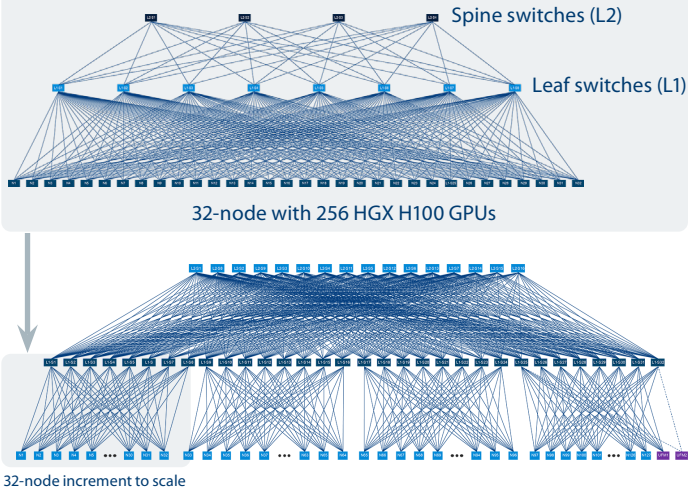
Compute and Storage

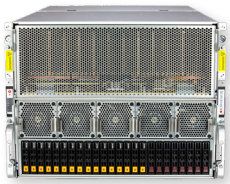
- 4x SYS-821GE-TNHR or AS -8125GS-TNHR per rack
- 4x NVIDIA HGX H100/H200 8-GPU per rack
- 32x NVIDIA H100/H200 Tensor Core GPUs
- 5TB of HBM3 or 9TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage support

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





Node Configuration — SYS-821GE-TNHR / AS -8125GS-TNHR

Overview	8U Air-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	6x 3000W Redundant Titanium Level power supplies

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

32-Node Scalable Unit — SRS-48UGPU-AI-ACSU

Overview	Fully integrated air-cooled 32-node cluster with 256 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	2x 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	9x 48U 750mm x 1200mm
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