

# Generative AI SuperCluster

With 256 1U NVIDIA MGX™ GH200 Grace™ Hopper Superchip Systems



## Scalable Compute Unit Built For Large Language Model Inference

- Unified GPU and CPU memory for cloud-scale high volume, low-latency, and high batch size inference
- 1U Air-cooled NVIDIA MGX Systems in 9 Racks, 256 NVIDIA GH200 Grace Hopper Superchips in one scalable unit
- Up to 144GB of HBM3e + 480GB of LPDDR5X, enough capacity to fit a 70B+ parameter model in one node
- 400Gb/s InfiniBand or Ethernet non-blocking networking connected to spine-leaf network fabric
- Customizable AI data pipeline storage fabric with industry leading parallel file system options
- NVIDIA AI Enterprise software ready

## Building Blocks for Cloud-Scale AI Inference Deployment

In the era of AI, 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. The full turn-key data center solution accelerates time-to-delivery for mission-critical enterprise use cases, and eliminates complexity of building a large cluster, that used to be only achievable through intensive design tuning and time-consuming optimization of supercomputing.

### MGX Systems with GH200 Grace Hopper

At the system level, the modular MGX platform features NVIDIA GH200 Grace™ Hopper Superchip which combines the power of an NVIDIA H100 GPU and NVIDIA Grace CPU on a single chip. This integrated solution delivers superior efficiency for a wide range of workloads, including large-scale AI inference.

The NVIDIA GH200 Grace Hopper Superchip equipped system addresses a key bottleneck in training and inference of Generative AI: the GPU memory bandwidth and capacity for running large language (LLM) models. NVIDIA® NVLink® Chip-2-Chip (NVLink®-C2C) provides a coherent CPU-GPU link that is 7x faster than PCIe 5.0. Utilize a unified pool of the 96GB of HBM3 or 144GB of HBM3e and 480GB of LPDDR5X totaling 576GB of memory to accelerate AI and HPC applications.

### Plug-and-Play Cluster

Our SuperCluster design for Supermicro NVIDIA MGX Systems with NVIDIA GH200 comes with 400Gb/s networking fabrics with non-blocking architecture. It enables the 32 nodes (32 GPUs) per rack and 256-node cluster to operate as a unit of compute with a coherent pool of high bandwidth memory which is essential for LLM high batch size, and high volume inference. Whether building a cloud-scale inference infrastructure for LLMs or fitting large models for optimal training performance, the spine and leaf network topology allows to scale from 256 nodes to thousands of nodes. Supermicro’s proven testing processes thoroughly validate the operational effectiveness of the cluster before shipping. Customers receive plug-and-play units at the rack or multi-rack cluster level 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

### Compute and Storage

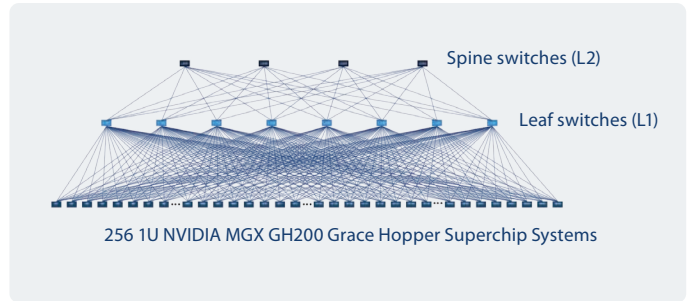
- 32x ARS-111GL-NHR per rack
- 32x NVIDIA GH200 Grace Hopper Superchips per rack (1x Grace CPU and 1x Hopper GPU per system).
- Up to 4.6TB HBM3e fast-access memory plus 15.3TB integrated LPDDR5X per rack.
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage support



## 256-Node LLM Scalable Unit

The spine-leaf network fabric allows 256-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

ARS-111GL-NHR

Overview	1U system with single NVIDIA Grace Hopper Superchip (air-cooled)
CPU	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip
Memory	Up to 480GB of integrated LPDDR5X with ECC (Up to 480GB + 144GB of fast-access memory)
GPU	NVIDIA H100 Tensor Core GPU with 96GB of HBM3 or 144GB of HBM3e (coming soon)
Networking	2x NVIDIA ConnectX®-7 single-port 400Gbps/NDR OSFP NICs, or 1x NVIDIA ConnectX-7 and 1x NVIDIA BlueField®-3
Storage	Up to 8x Hot-swap E1.5 drives and 2x M.2 NVMe drives
Power Supply	2x 2000W Redundant Titanium Level power supplies

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

### 256-Node Scalable Unit

SRS-MGX256-SU-001

Overview	Fully integrated 256-node cluster with 256 Hopper GPUs and Grace CPUs
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	4x SSE-MSN4600-CS2FC 64-port 100GbE QSFP28, 2U switch
Out-of-band Management Switch	8x SSE-G3748R-SMIS, 48-port 1Gbps 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