Building Telco AI Factories for Sovereign AI

Powered by Supermicro, NVIDIA HGX™ H100/H200/B100, and NVIDIA MGX™

Industry leading Scalable Compute Unit Built For Large Language Models

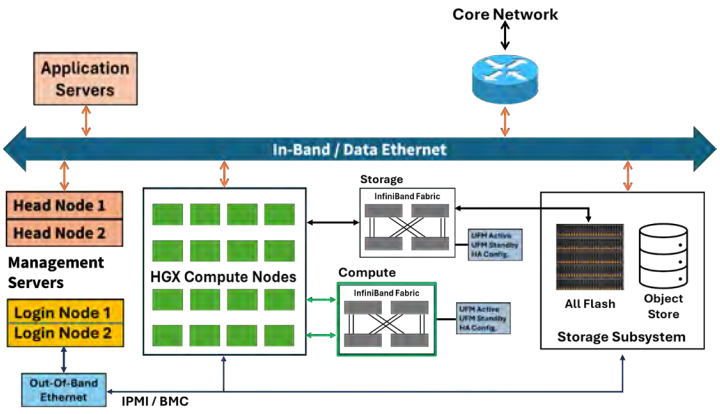
- Proven industry leading architecture for large scale AI infrastructure deployments
- 256 NVIDIA H100/H200 GPUs in one scalable unit
- 20TB of HBM3 with H100 or 36TB of HBM3e with H200 in one scalable unit
- 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage for training large language model with up to trillions of parameters
- Customizable AI data pipeline storage fabric with industry leading parallel file system options
- Supports NVIDIA Quantum-2 InfiniBand and Spectrum™-X Ethernet platform
- Certified for NVIDIA AI Enterprise Platform including NVIDIA NIM microservices

Enabling Telecom Companies as Providers of Sovereign AI

Telecommunication companies, as trusted national technology providers, are looking to provide sovereign AI for regional governments, enterprises, and startups to build, customize, and deploy generative AI applications. By transforming into an "AI factory," telecom companies can leverage accelerated computing infrastructure, software, and services in their existing data center footprint to deliver AI intelligence at a national scale.

The NVIDIA Partner Network (NPN) program enables telecom companies to transform into "AI factories." The program provides ecosystem partners with reference architectures (RAs) certified by NVIDIA and professional services to build optimized and scalable AI infrastructure and cloud services. Supermicro enables telecom companies to build AI factories by providing solutions aligned with NVIDIA RAs for training and inference of generative AI models.

Building Blocks for Highest Density Generative AI Infrastructure Deployment



In the era of generative AI, the data center is the new unit of compute rather than individual servers. Interconnected GPUs, CPUs, memory, storage, and other resources across multiple nodes in racks orchestrate large-scale AI workflows. This infrastructure requires high-speed and low-latency network fabrics, 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 to build, customize, and deploy rapidly evolving generative AI and large language models (LLMs). The turn-key data center solution accelerates time-to-delivery for mission-critical enterprise use cases and eliminates the complexity of building a large compute cluster. This compute infrastructure was previously only achievable through intensive design tuning and time-consuming optimization of supercomputing resources.

© Copyright Super Micro Computer, Inc. All rights reserved

AI Systems for Training Core Components

The core compute component for AI training is Supermicro's proven, industry-leading servers powered by 8-way NVIDIA HGX H100, H200, and B100. Using dedicated PCIe 5.0 slots, each NVIDIA GPU is paired 1:1 with NVIDIA Quantum-2 400Gb/s InfiniBand networking, which includes NVIDIA ConnectX-7 network interface card (NIC) to enable NVIDIA GPUDirect RDMA and storage for direct data flows to GPU memory with the lowest latency possible.

The Supermicro solution powered by NVIDIA HGX H100 and H200 8-way GPUs is ideal for training next-generation, large-scale generative AI models. These systems are available in both air-cooled and liquid-cooled variants. The high-speed interconnected NVIDIA GPUs utilize NVIDIA® NVLink® for high GPU memory bandwidth and capacity to run LLMs cost-effectively. The SuperCluster creates a massive pool of GPU resources acting as one AI supercomputer.

Plug-and-Play for Rapid Generative AI Deployment

The SuperCluster design with 8U air-cooled (shown) or optional 4U liquid-cooled HGX systems comes with 400Gb/s of networking fabrics and non-blocking architecture. These are interconnected into four 8U (or eight 4U) nodes per rack and further into a 32-node cluster that operates as a scalable unit "SU" 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 cloud-scale LLM inference infrastructure, the SuperCluster leaf-spine 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 of compute infrastructure before shipping. Customers receive plug-and-play scalable units for rapid deployment.

32-Node Scalable Unit	SYS-821GE-TNHR / AS -8125GS-TNHR
Overview	8U Air-cooled System with NVIDIA HGX H100/H200
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 NVLink
Networking	8x NVDIA ConnectX®-7 Single-port 400Gbps/NDR OSFP NICs 1x NVIDIA BlueField®-3 DPU (B3220) Dual-port 200Gbps/NDR200
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.

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 -8125GSTNHR 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 leaf-spine network fabric enables a 32-node scalable compute unit to scale to thousands of nodes. With high network performance for GPU-GPU connectivity, the SuperCluster is optimized for high-volume LLM training and high batch size inference. Supermicro provides L11 and L12 validation testing and on-site deployment services for telecom companies to quickly scale up their AI infrastructure.

32-Node Scalable Unit	SRS-48UGPU-AI-ACSU
Overview	Fully integrated liquid-cooled 32-node cluster with 256 H100/H200 GPUs
Compute Fabric Leaf	8x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 InfiniBand 400G NDR, 32 OSFP ports switch
Compute Fabric Spine	4x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 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 and PDU	9x 48U 750mm x 1200mm 34x 208V 60A 3Ph

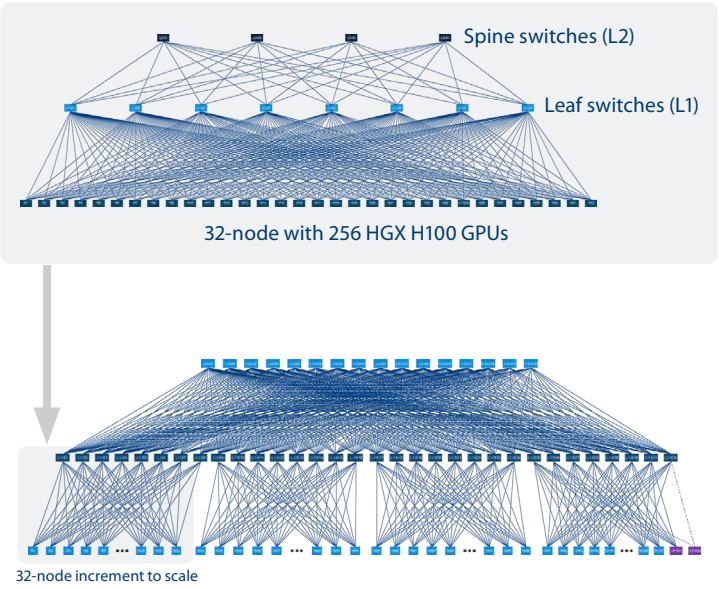
*Recommended configuration, other network switch options and rack layouts are available.
*Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional

Compute Fabric Scaleout

To scale out compute infrastructure, the compute nodes are connected over a rail-optimized leaf-spine network:

- First, the 8-way NVIDIA GPU servers are clustered to form a 32-node LLM Scalable Unit (SU) with 256 GPUs.
- The SU can be optimally connected over a single spine layer into a 128-node SuperCluster with 1000+ GPUs.
- The SuperCluster is multiplied to build a supercomputer-scale "AI factory, "which can expand into multiple thousands of GPUs.

Network Fabrics

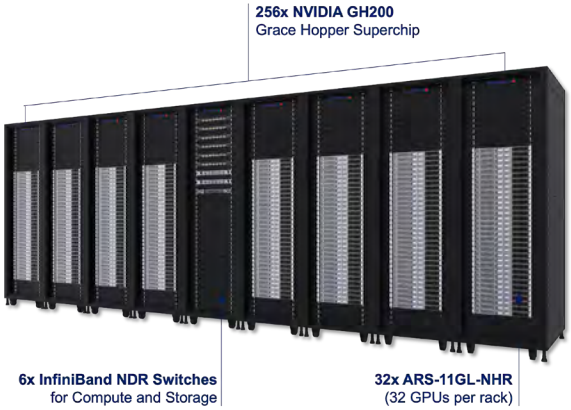


NVIDIA MGX Reference Architecture (AI Inferencing)

Once generative AI models are trained on SuperClusters, AI inference serves the models to users of generative AI applications. This unlocks possibilities for telecom companies to increase revenue opportunities. Each inferencing task is time-bound and often transactional. AI inferencing systems scale-up to manage multiple models and serve user requests in parallel with low latency from the cloud to the edge.

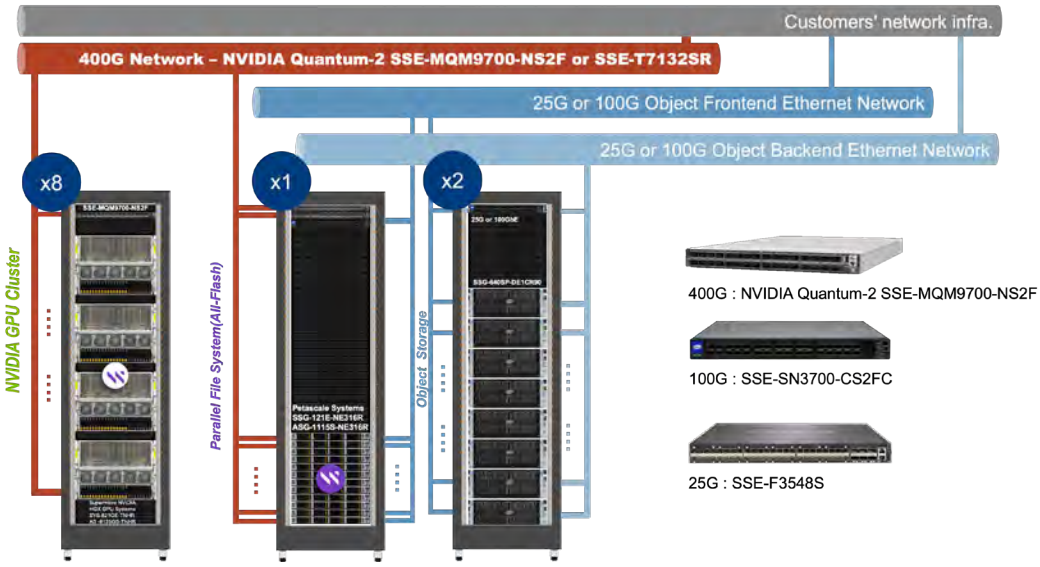
Small-scale inferencing workloads run on the Supermicro ARS-111GL-NHR (air- or liquid-cooled) featuring the NVIDIA MGX architecture with the NVIDIA GH200 Grace Hopper™ Superchip. For large-scale AI inference, the NVIDIA MGX servers with NVIDIA GH200 can be clustered together in a similar manner to the HGX-powered systems based on 560 Supermicro ARS-111GL-DNHR-LCC 1U 2-node systems with 1x NVIDIA Grace Hopper Superchip per node (liquid-cooled).

- System nodes with enough GPU memory capacity to contain pre-trained model
- GPU and CPU performance to handle high volume queries with low-latency response time
- Cloud-optimized for energy-efficiency and TCO



Storage Subsystem for Generative AI

Generative AI models require processing large amounts of stored data, particularly for training. Data stored needs to be accessed with high bandwidth and low latency for the compute infrastructure to maximize performance. The SuperCluster storage system includes GPU-direct working storage, and lower cost ingest, preprocessing, and archival-tiered storage.



Management and Data Ethernet Network

SuperCluster-powered AI infrastructure includes:

- General data network to pair Superclusters to application servers and core networks
- In-band management systems network
- Out-of-band management systems network

These networks run on an Ethernet-based fabric.

The data and in-band management networks can run over the same ethernet network, typically at 100G. The out-of-band ethernet network typically runs at 1G. It may share the data network TOR switches, although the best practice is to keep them physically separate.

Professional Services

Supermicro and NVIDIA offer a full range of professional services to ensure the complete and successful commissioning of AI factories in telecommunications with the shortest time to revenue and full operation.

Complete Integration at Scale	Test, Validate, Deploy with On-site Service	Liquid-Cooling / Air-Cooling	Supply and Inventory Management
Design and build of full racks and clusters with a global manufacturing capacity of up to 4,000 racks per month	Proven L11, L12 testing processes throughly validate the operational effectiveness and efficiency before shipping	Fully integrated liquid-cooling or air-cooling solution with GPU & CPU cold plates, Cooling Distribution Units and Manifolds	One-step-shop to deliver fully integrated racks fast and on-time to reduce time-to-solution for rapid deployment

Get Started Building an AI Factory

Learn more about Supermicro and NVIDIA solutions for AI Factories: https://www.supermicro.com/en/accelerators/nvidia

Test drive Supermicro systems powered by NVIDIA: https://www.supermicro.com/en/jumpstart