



A BLUEPRINT FOR LLM AND GENERATIVE AI INFRASTRUCTURE AT SCALE



Table of Contents

Introduction - Overcoming the Biggest Challenges in AI Computing	2
AI Infrastructure: A Calculated Investment	2
Part 1: GPU System Building Blocks of the AI SuperCluster	3
System Architecture	4
System Topology	Error! Bookmark not defined.
Part 2: Populating the Racks of an AI SuperCluster	5
Organizing the Rack for Smart Cabling, Management, and Thermals	5
Doubling Computing Density Per Rack with Liquid Cooling	6
Part 3: High-Performance Networking: The Fabric to Weave System Nodes into One SuperCluster	7
Optimizing Network Efficiency with a Non-Blocking Fat-Tree Rail-Optimized Topology	8
Part 4: SuperCluster Solution Design and Deployment	11
Evaluating Data Center Power and Other Resources	12
End-to-End Software Platform for Generative AI with NVIDIA AI Enterprise	12
Conclusion - Accelerate Time-to-Deployment	13
Further Information	13
Appendix - SuperCluster Configurations	14
SuperCluster Current and Future Offerings	15

Scaling out Supermicro SuperCluster Architecture with NVIDIA Quantum InfiniBand

The Supermicro SuperCluster is the Gold Standard purpose-built data center infrastructure - meticulously crafted to tackle the computational needs of modern AI and HPC workloads. Supermicro designs, manufactures, and deploys data center solutions built for massively parallel computing capabilities to drive modern AI applications, including LLM training and real-time Generative AI Inference. As one of the industry's leaders in deploying infrastructure in some of the world's largest AI data centers, Supermicro offers a unique perspective on data center solutions.

Supermicro's SuperCluster reference architecture is designed to solve the challenges of planning and deploying highly complex scale-out AI infrastructure. SuperCluster vastly simplifies infrastructure projects by providing a base package of interoperable components, known as a "scalable unit (SU)". Featuring 32 of Supermicro's incredibly powerful GPU systems, utilizing NVIDIA's ground-breaking H100/H200 GPUs along with InfiniBand compute fabric - Supermicro SuperCluster SU is the ultimate building block towards building the largest AI training infrastructures of the world. As demands grow, this uniquely designed SU scales out effortlessly utilizing the power of NVIDIA Quantum InfiniBand to expand infrastructures - ensuring that customers will always have the capacity required to meet evolving

computing demands. Capable of scaling out to thousands of GPU nodes, the Supermicro SuperCluster has the capability to harness the maximum computational power from NVIDIA's most powerful GPUs, making it the ultimate powerhouse of Generative AI infrastructure.

This white paper reveals blueprints of a Generative AI rack cluster with NVIDIA HGX™ H100/H200 GPUs. Based on NVIDIA's validated reference architectures and Supermicro's hand-in-hand collaboration with NVIDIA, this paper delves into the design of SuperCluster's individual system nodes, component selection, rack layout, network topology, and deployment steps.

Introduction - Overcoming the Biggest Challenges in AI Computing

Artificial Intelligence applications operate on the same physical processes involved in all forms of computing: manipulating the flow of electricity in a circuit to perform operations. But training today's AI large language models requires computing performance at unprecedented magnitudes that can amount to hundreds of millions of quadrillions of training operations.

The heightened computing demands of AI applications present unique challenges for data centers. Solutions for large-scale AI training and inference should be designed with these factors in mind:

- **Parallel computing capacity:** GPU system nodes must be highly effective at splitting workloads and executing a vast number of operations in tandem to complete AI workloads in a timely manner.
- **Networking scalability:** the cluster topology needs to aggregate the computing capacity of individual system nodes into a single powerful supercomputer with a shared memory system without introducing major network bottlenecks.
- **Deployment complexity:** to ensure high uptime, high performance, and interoperability of the individual parts, key aspects of the data center deployment must be carefully planned, including data center power, floor plan, rack layout, and thermal management.

AI Infrastructure: A Calculated Investment

In addition to these challenges, there are also practical business considerations when evaluating AI infrastructure investments. Assuming a company is reasonably well-positioned, AI infrastructure investments will likely deliver positive ROI if deployed at the correct scale for the appropriate applications. Therefore, a starting point is to properly size the project based on the desired business objective.

To better illustrate this, imagine a rapidly growing AI software company that is training custom large language models for other enterprises, roughly equivalent to GPT-3 in complexity. The company already has major revenue-generating clients and is starting to think long-term. Its GPU cloud billing costs keep increasing, so it's decided that investing in hardware on-premises or via co-location is the next logical step.

Training today's AI foundation models requires thousands of GPUs. However, GPUs are of little use without systems to provide the power delivery and cooling needed to operate effectively. Furthermore, the system and network architecture need to deliver a reliable pipeline of training data to the GPUs to ensure adequate utilization rates. The systems must be interconnected via high-speed networking that enables fast GPU-to-GPU communication with a shared memory pool.

To understand how to scale effectively to the cluster level, Supermicro will profile the GPU systems that compose Supermicro's SuperCluster solution before building up to the rack and cluster level.



Figure 1 - Scaling from System Nodes to an Interconnected Cluster

Part 1: GPU System Building Blocks of the AI SuperCluster

GPU systems do the heavy lifting for AI workloads and can be referred to as the "compute nodes" within the SuperCluster network. The rack-scale characteristics of the cluster, such as the network topology, are defined by patterns established in the system architecture of these individual compute nodes.

SuperCluster's base package provides 32 interconnected 9kW GPU systems, each containing 8 GPUs. For each 32 node scalable unit, the GPU systems (8U, 8GPU) are populated in a total of 8 rack enclosures for the air-cooled version or four rack enclosures for the double-density liquid-cooled version (4U, 8 GPU). An additional rack enclosure hosts the networking components.



Figure 2 - Supermicro 8U NVIDIA HGX H100/H200 8-GPU Server (Air Cooled or Liquid Cooled)

SYS-821GE-TNHR / AS-8125GS-TNHR



Supermicro 4U NVIDIA HGX H100/H200 8-GPU Server (Liquid Cooled Only)

SYS-421GE-TNHR2-LCC / AS-4125GS-TNHR2-LCC

Although the Supermicro 8-GPU NVIDIA HGX Systems are powerful on their own, they feature a system architecture and topology intended for scalability. The reasons for this will become clear as we build up to rack-scale, but let's first briefly cover some of the system's key components:

- Dual Socket E (LGA-4677) 4th/5th Gen Intel® Xeon® Scalable processors or AMD EPYC™ 9004 Series Processors
- NVIDIA HGX™ H100/H200 GPUs, totaling 8 GPUs per system
- Memory Capacity: Up to 8TB DDR5-5600 via up to 32 DIMM slots
- Up to 19x 2.5" hot-swap NVMe/SATA drive bays, 2x NVMe M.2 boot drives
- 8 PCIe 5.0 x16 LP slots, 4 PCIe 5.0 x16 FHHL slots
- InfiniBand Compute Network: 8x NVIDIA ConnectX-7 Single-Port 400Gbps/NDR OSFP NICs
- Storage & In-Band Management Network: 1 x NVIDIA BlueField-3 DPU (Dual-Port 200GbE) per system.
 - *Note: These systems are also compatible with NVIDIA Spectrum-X, if ethernet is preferred.*

Eight GPUs and two CPUs occupy each of the systems. The 8-to-2 ratio of GPUs to CPUs is suitable for AI applications since their parallelizable workloads rely primarily on GPU computing. Both AMD EPYC CPUs and Intel Xeon CPUs are available as options.

The NVIDIA H100/H200 GPU has become nearly synonymous with AI. The Hopper architecture's powerful parallel computing capabilities are explicitly made for AI applications, featuring re-designed streaming multiprocessors and a high-bandwidth memory system. In addition, NVIDIA introduced new lightweight data types specifically optimized to allow Hopper's cores to perform AI arithmetic at unprecedented speeds.

This system utilizes NVIDIA SXM version of GPUs, specifically the NVIDIA HGX 8-GPU H100 or H200. Each H100/H200 GPU is interconnected via NVLink to 4 NVSwitches, delivering 900GB/s bi-directional bandwidth for GPU-to-GPU communication between any of the GPUs in the local group of 8. The HGX H200 system's 1,128GB HBM3e GPU memory is enough to fully contain a large AI model. This combined pool of coherent memory enables a single system to act as a powerful real-time inference engine, even without extending over a network.

System Architecture

A significant amount of R&D work goes into refining key aspects of a system that may not be apparent from its list of technical specifications. Supermicro develops its own system architecture through a multi-stage design process that includes the chassis, motherboard, and electromechanical hardware (such as fans and connectors).

Up to 8x 3000W Redundant Titanium Level PSUs provide ample power to the 8 GPUs and other system components with headroom to spare. 8 AC input connections on the rear of the system ensure reliable power delivery to the PSUs.

Running 8x 700W TDP GPUs inevitably leads to heat as a byproduct. The 8U chassis provides a high level of mechanical airflow to ensure thermal stability at max load within an AI data center. The Motherboard Air Shroud and GPU Air Blocker boost cooling efficiency by concentrating airflow.

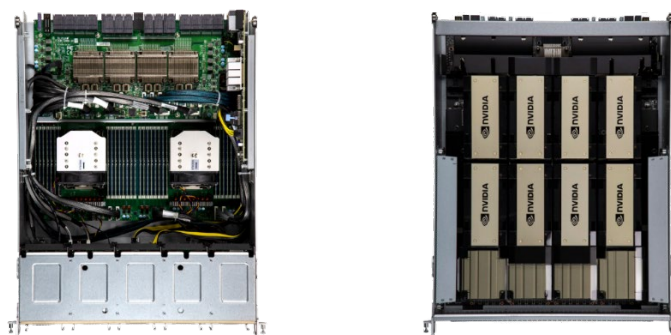


Figure 3 - Motherboard Tray (Left), GPU Tray (Right)

The system is composed of 2 trays that can be accessed independently, as shown in Figure 3:

- Motherboard tray, positioned on the bottom of the chassis
- GPU tray, positioned on the top of the chassis

Isolating the GPU tray and CPU tray reduces heat transfer between the components. Each tray has a full row of 5 heavy-duty fans spanning the width of the chassis. Fan speed control is supported by Thermal Management using the BMC 2.0 interface. The GPU tray hosts the NVIDIA HGX H100/H200 8-GPU baseboard. Its 4U height accommodates the tall heatsinks attached to the GPUs. On the other hand, the 4U 8GPU system is completely liquid-cooled (both CPUs & GPUs) utilizing Supermicro's Direct to Chip Liquid Cooling (D2C) solution. Supporting highest densities and highest TDP CPUs and GPUs with up to 100kW power and cooling per rack, the 4U liquid cooled cluster is the ultimate resource saving data center solution.

The HGX GPU baseboard combines H100/H200 GPUs with high-speed interconnects – utilizing and harnessing the power of NVIDIA ConnectX-7 NICs and BlueField3 DPUs. ConnectX-7 is available in 1, 2, or 4-port configurations and delivers up to 400Gb/s (InfiniBand/Ethernet) of bandwidth, enabling highest utilization of GPU compute. BlueField-3 DPUs can reach up to 400Gb/s that enables offloading, accelerating, and isolating software-defined networking, storage, security, and management functions, significantly enhancing data center performance. With features such as NVIDIA's Accelerated Switching and Packet Processing, advanced RoCE, NVIDIA GPUDirect Storage, and inline hardware acceleration for TLS/IPsec/MACsec encryption/decryption, NVIDIA ConnectX-7 and BlueField-3 DPU along with Supermicro's most powerful GPU systems empower agile and high-performance solutions from edge to core data centers to clouds, all while enhancing network security and reducing the total cost of ownership.

PCIe switches enable GPU-to-GPU network traffic to bypass the host CPU via GPU Remote Direct Memory Access (RDMA). Each of the 8 GPUs is paired with NVIDIA ConnectX-7 via the 4 PCIe switches, providing 400Gb/s bandwidth of InfiniBand or Ethernet connectivity. There are two additional PCIe switches for further expansion. These two PCIe x16 slots are typically used with a pair of NICs to connect to a high-performance storage cluster over a network fabric.

Part 2: Populating the Racks of an AI SuperCluster

Organizing the Rack for Smart Cabling, Management, and Thermals

Going beyond system nodes, the rack-level can be considered as the next tier of organization for the cluster. It's important to note that the rack layout is independent of the cable endpoints between components. Theoretically, two clusters with

identical components and connections could use different rack layouts. However, the rack layout should still be optimized to best suit the cluster's topology.

The rack layout can aid in the deployment, management, and servicing of the cluster. Optimized rack layouts offer additional benefits, such as allowing for shorter cable lengths, which can improve performance and reduce airflow blockage. Supermicro will explain the rationale behind our rack layout design choices, with the caveat that there is some flexibility for customers to adjust as needed.

Optimizing the rack layout is based on factors including:

- To reduce cable length and to improve cable organization
- To simplify the physical deployment and service
- To improve thermal performance
- To maximize use of available space (such as improving density)



Figure 4 – 8U SuperCluster (Available in both air-cooled & liquid-cooled)

Let's examine the rack layout, starting at the center of the rack cluster. The middle rack is for housing the networking switches, including NVIDIA Quantum InfiniBand switches to harness maximum compute performance, and additional ethernet switches for optimum storage and management. On both sides, 8 identical "compute racks" contain four 8U 8-GPU systems. This rack layout, as opposed to placing more switches in a top-of-rack style, is optimized for the cross-rack cabling required for the topology, which streamlines GPU-to-GPU communication by reducing network hops. In this air-cooled configuration, blanking panels occupy the remaining rack units, allowing for more thermal headroom to avoid throttling.

Doubling Computing Density Per Rack with Liquid Cooling

Liquid cooling presents an opportunity for substantial energy savings of up to 40% for the entire data center and substantial space savings. For customers seeking to maximize computing capacity within their available data center footprint, Supermicro offers a SuperCluster option with direct-to-chip liquid cooling. In the liquid-cooled version, the Supermicro 4U 8-GPU NVIDIA HGX System plays the role of the cluster's compute nodes. Both the CPUs and GPUs are liquid-cooled with cold plates that efficiently move heat away from the chips.



Figure 5 - Liquid-Cooled SuperCluster

The increased cooling efficiency allows 8 4U systems to reside in a 48U rack. The total solution consisting of 32 system nodes and 256 GPUs only occupies a total of 5 racks (four compute racks and one switching rack).

A 4U Cooling Distribution Unit (CDU) is positioned at the bottom of each compute rack, moving the hot liquid away from the systems to the facility-side where it is dissipated via an external cooling tower. Each 4U system is paired with a 1U manifold that handles the distribution of liquid to and from the systems. Aside from the increased density and the in-rack CDU, the liquid-cooled version shares most other similarities with the air-cooled SuperCluster, such as the same network topology.

Supermicro chose to utilize an in-rack CDU with direct-to-chip liquid cooling due to its effectiveness and ease of deployment as a complete integrated liquid-cooling solution. Supermicro develops custom in-house liquid cooling components, including cold plates which are tailored to each type of socket. The in-rack CDU offers additional benefits over other approaches (such as in-row CDU) by providing rack-level intelligent flow adjustment and monitoring. Lastly, the in-rack CDU streamlines deployment by allowing much of the closed-loop liquid cooling setup to be configured off-site. For customers interested in deploying infrastructure for modern high-density data centers with liquid cooling, Supermicro can evaluate its suitability, and ease in the deployment process.

Part 3: High-Performance Networking: The Fabric to Weave System Nodes into One SuperCluster

Imagine a scenario where a company has just purchased 32 GPU systems, each with 8 GPUs but nothing else. DevOps loads the AI training application designed to train a 175B parameter AI model and a massive training dataset. With these systems alone, unfortunately, they would only be able to utilize a single system for the training application. Without a fast and reliable network fabric, the compute nodes are blind to the activity of other nodes.

3.1: Compute Fabric

Networking is a way to solve the biggest challenges of AI computing at scale, including maintaining a coherent, high-capacity, high-bandwidth memory system. The NVIDIA Quantum-2 InfiniBand platform is an end-to-end network solution featuring high-bandwidth, ultra-low-latency NVIDIA Quantum-2 switches and ConnectX®-7 adapters for accelerated performance. Leveraging NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ technology for In-Network Computing, NVIDIA Quantum-2 gives AI developers the fastest networking performance available to take on the world's most challenging AI applications.

For the compute fabric, NVIDIA Networking Platforms serve as the nervous system for compute infrastructure.



Figure 6 - NVIDIA Quantum-2 InfiniBand Switch

Optimizing Network Efficiency with a Non-Blocking Fat-Tree Rail-Optimized Topology

AI clusters should employ a topology that allows the GPUs to communicate through the most optimal network path. We describe SuperCluster's networking as a "Fat-Tree Spine-Leaf Non-blocking Rail-Optimized Topology". Here's a quick breakdown of these terms:

- **Fat-Tree:** a tree data structure is an efficient means to connect the nodes in a cluster. In a spine-leaf topology, spine switches branch out into connections with a greater number of leaf switches, branching out further to an even greater number of nodes. The term "fat tree" means that the bandwidth at the top of the tree is greater than the bandwidth of connections at the bottom to ensure balanced bandwidth between all levels.
- **Non-blocking:** in a network, if the amount of traffic sent through a switch exceeds its capacity, network bottlenecks will occur (oversubscription). A non-blocking network uses a 1:1 balance of downlink and uplink bandwidth to maximize throughput.
- **Rail-optimized:** An essential aspect of optimizing network performance is to reduce the number of network "hops" when traffic gets sent from GPU to GPU. Rail-optimization is a strategy to accomplish this by grouping GPUs that are connected to the same leaf switch.

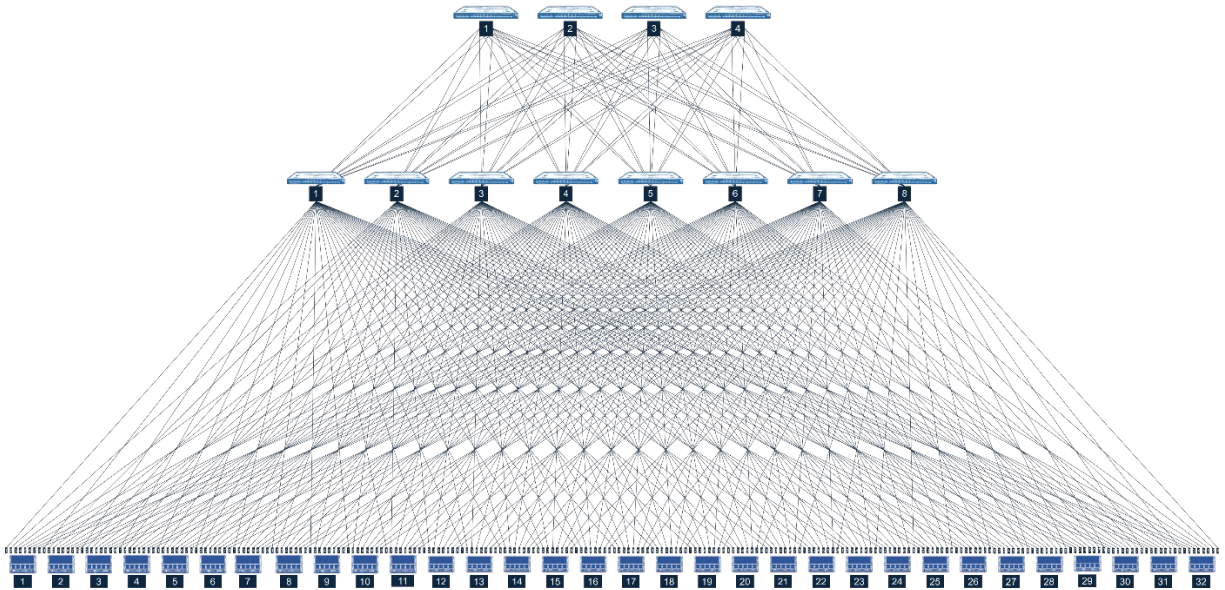


Figure 7 - SuperCluster Scalable Unit (32 Nodes) Network Topology

GPUs sharing the same rank are classified as a single "rail group". This approach streamlines communication by prioritizing traffic routing through members of the same rail group to avoid unnecessary "hops" when traversing the tree. The spine switches are still important to allow GPUs to communicate outside of their rail group.

Altogether, SuperCluster's base package (32 node Scalable Unit) contains a total of 12 NVIDIA Quantum-2 InfiniBand switches (QM9700), delivering speeds up to 400Gb/s, to unify its 256 GPUs.

- **4x NVIDIA InfiniBand (QM9700) as Spine switches** (Layer 2) with NVIDIA LinkX Quantum-2 InfiniBand Twin-port OSFP Transceivers (MMA4Z00-NS)
- **8x NVIDIA InfiniBand QM9700 as Leaf switches** (Layer 1) with NVIDIA LinkX Quantum-2 InfiniBand Twin-port OSFP NDR Transceivers (MMA4Z00-NS).
- **8x NVIDIA ConnectX®-7 per System Node** with NVIDIA LinkX Quantum-2 InfiniBand OSFP Transceivers (MMA4Z00-NS400)
- NVIDIA MPO-12/APC Passive Fiber Cables

Example – Aggregating 4 Scalable Units (Scaling out to 128 Nodes)

Below is an example of how effortlessly the Infiniband fabric can be scaled out using rail-optimized topology to create a 128 nodes SuperCluster. As mentioned before, each NVIDIA HGX H100 node has 8x back-end NVIDIA ConnectX-7 adapters to create 1 leaf switch group, and 4 leaf-switch groups are sufficient to connect 128 nodes for maximum GPU-GPU connectivity. Comprising of 32 Leaf and 16 Spine switches, the example below demonstrates the easy scalability of this meticulously designed AI SuperCluster. Note, adding an NVIDIA Unified Fabric Manager (UFM) node requires to remove 1 compute node, hence, the following topology shows a 127 compute node cluster.

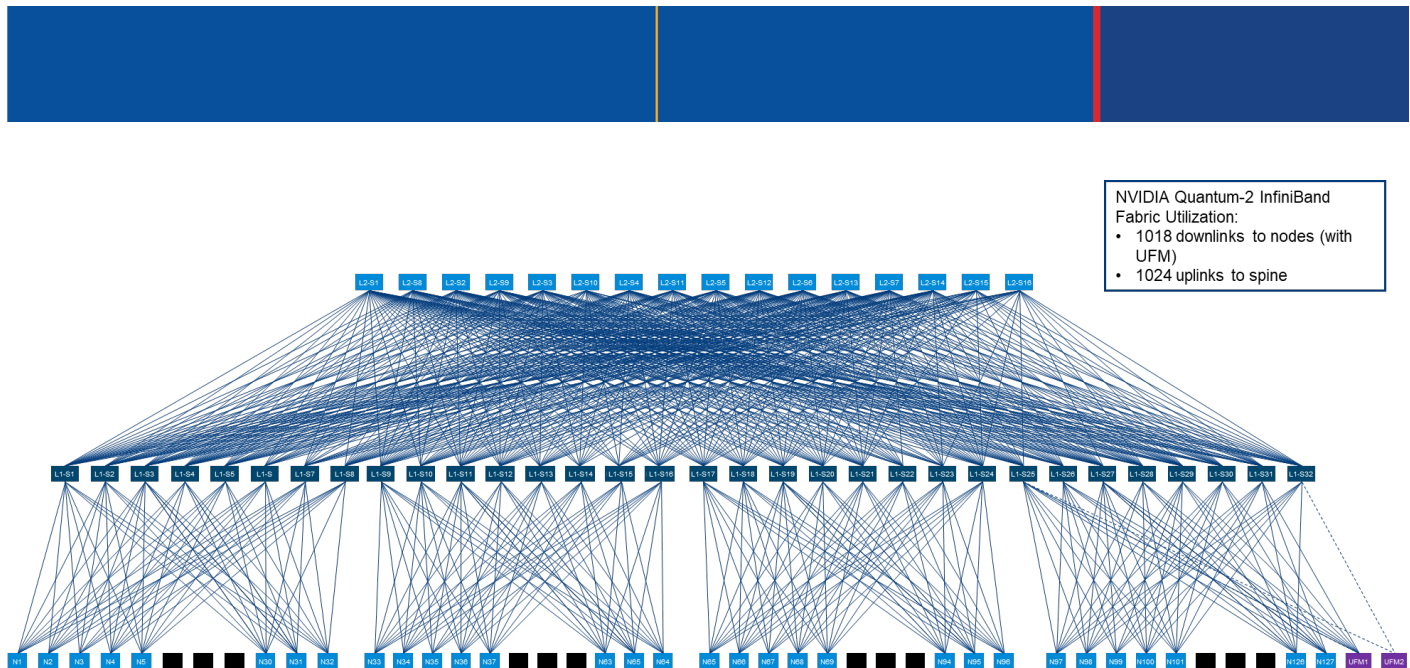


Figure 8 - SuperCluster 128 Nodes (4 x SU) Network Topology

To ensure maximum performance for AI and LLM workloads, GPU-to-GPU connectivity, link utilization and critical events happening across the fabric path needs to be monitored. Especially, while aggregating multiple scalable units, it is recommended to include UFM nodes which add enhanced network monitoring, management, workload optimizations and periodic configuration checks.

3.2: Converged Network (Storage Fabric & In-Band Management)

The converged network comprises of the storage network fabric and in-band management fabric. This provides flexible storage allocation and simplified network management while monitoring in-band. The converged network also provides high bandwidth to HPS and connects to the data center network. It's independent of the compute fabric to maximize both storage and application performance.

There are the variety of workloads, datasets, and need for training locally and directly from the high-speed storage system (external storage – provided by additional storage servers) . The Storage Fabric provides high bandwidth to shared storage, independent of the compute fabric to maximize performance of both storage and application performance. The in-band Management fabric maximizes stability, performance, and ease of management by using underlay network, as well as providing some redundancy. It also provides connectivity for in-cluster services such as Slurm, and to other services outside of the cluster such as the NGC registry, code repositories, and data sources. Basically, it connects all the services that manage the cluster. It is recommended that storage & in-band management is provided utilizing NVIDIA Spectrum-4 Ethernet switches (SN5600) and BlueField-3 DPU (B3220, 2 x NDR200/200GbE connections per node).

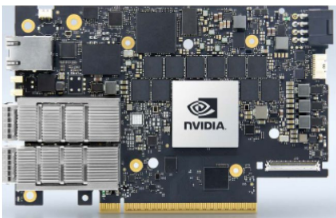


Figure 9 - SuperCluster Converged Network (Storage and In-Band Management)

3.3: Out of Band Management

The out-of-band Ethernet network is used for system management via the BMC, IPMI and provides connectivity to manage all networking equipment. Out-of-band management is critical to the operation of the cluster by providing low usage paths that ensure management traffic does not conflict with other cluster services. NVIDIA Spectrum SN2201 is ideal as an out-of-band (OOB) management switch - connecting up to 48 x 1G Base-T host ports with non-blocking 100 GbE spine uplinks. Featuring highly advanced hardware and software along with ASIC-level telemetry and a 16MB fully shared buffer, the Spectrum SN2201 delivers unique and innovative features to 1G switching.

*Figure 10 - NVIDIA Spectrum SN2201 (1GbE Switch)*

Part 4: SuperCluster Solution Design and Deployment

Designing and deploying AI infrastructure requires a multi-staged process, starting with the solution design. For rack-scale projects, it is often difficult to finalize the bill of materials (BOM), which can contain over 10,000 components. SuperCluster speeds up the process by maintaining a pre-validated list of interoperable GPU systems, rack enclosures, rack rail kits, blanking panels, PDUs, NVIDIA Quantum InfiniBand and Spectrum Ethernet switches, LinkX cables, transceivers, and more.

That doesn't mean SuperCluster is an "off-the-shelf" solution: it can be tailored to fit the customer's exact requirements. Supermicro's Solution Design team ensures that the proposal addresses the needs of the customer's application, existing software and hardware infrastructure, and the data center deployment environment. Supermicro can adjust SuperCluster's BOM accordingly and will create a proposal for the customer's approval.

Supermicro uses a 6-step process to ensure project success from start to finish:

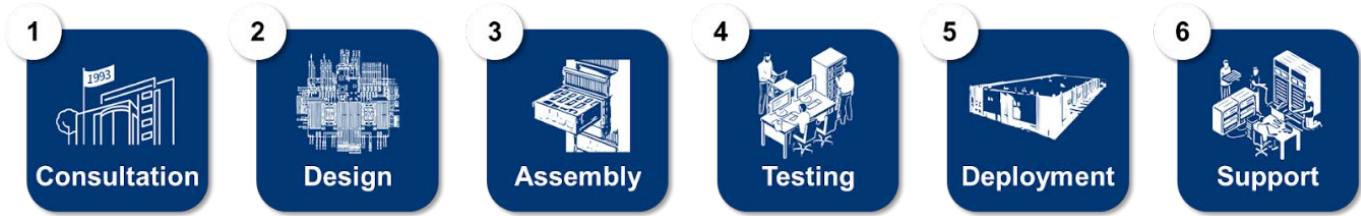


Figure 9 - Supermicro Solution & Integration Process

Evaluating Data Center Power and Other Resources

Many details are out of the scope of this paper, but we will briefly cover SuperCluster's power requirements to illustrate an important piece of the consultation and design process. Data center power specifications will vary depending on things like its geographical location. For example, the type of AC power input will vary by region. In any case, our team will determine a suitable solution.

Total power capacity is one of the critical metrics to classify modern data centers, ranging from local Edge data centers rated in kilowatts to hyperscalers rated up to hundreds of megawatts. Supermicro's team will evaluate if the data center power available covers the power draw of the cluster. Each system in an air-cooled 8U 8-GPU system-based SuperCluster draws about 9kW. A SuperCluster Scalable Unit with 32 nodes (256 GPUs) consumes about 288kW of power, plus roughly 25kW from the networking components.

Our reference SuperCluster architecture for 8U 8-GPU systems utilizes 34 208V 60A 3-phase PDUs. To calculate total power, multiply volts by amps by the number of PDUs ($208 \times 60 \times 34$), which equals 424kW of power. This power capacity is enough to drive the SuperCluster with headroom to spare. Note that the power requirements of other SuperCluster configurations will vary.

In the consultation and design phase, Supermicro also includes the data center floor plan and rack layout in the proposal. The goal is to create a plug-and-play data center deployment experience, with Supermicro overseeing the delivery, cabling, configuration, testing, and support with a team of on-site engineers.

End-to-End Software Platform for Generative AI with NVIDIA AI Enterprise

SuperCluster is specialized for the computing needs of AI but is intended to deliver performance across all types of AI applications. The need for AI infrastructure goes beyond application-specific limitations to support the expanding ways that companies are integrating AI into their businesses. As state-of-the-art open-source generative AI models become increasingly accessible, enterprises across all industries are experimenting with new use cases for generative AI.

Supermicro collaborates closely with NVIDIA to ensure a seamless and flexible transition from experimentation and piloting AI applications to production deployment and large-scale data center AI. This high level of integration is achieved through rack and cluster-level optimization with the NVIDIA AI Enterprise Software platform, enabling a smooth journey from initial exploration to scalable AI implementation.

Managed services compromise infrastructure choices, data sharing, and generative AI strategy control. NVIDIA NIM, part of the NVIDIA AI Enterprise platform, offers managed generative AI and open-source deployment benefits without drawbacks. Its versatile inference runtime with microservices accelerates generative AI deployment across a wide range of models, from

open source to NVIDIA's foundation models. NVIDIA NeMo enables custom model development with data curation, advanced customization, and retrieval-augmented generation (RAG) for enterprise-ready solutions.

Supermicro's SuperCluster is validated to run the NVIDIA AI Enterprise software platform, providing a unified hardware and software solution for AI infrastructure.

Conclusion - Accelerate Time-to-Deployment

The computing requirements of today's AI applications have led to a rethinking of data centers. Due to the fast growth of AI and GPU computing, many of the conventional IT practices are no longer a blueprint for success. The key concepts discussed in this whitepaper represent a proven approach to AI infrastructure that will carry forward for the foreseeable future (For example, future SuperClusters that leverage NVIDIA Blackwell architecture will have a similar network topology to the one described here).

Supermicro's SuperCluster is a validated solution that balances cost, performance, and flexibility for a variety of AI workloads. We have a vertically integrated global supply chain underpinning our US-based final assembly facilities, with a manufacturing capacity of up to 5,000 racks per month. Supermicro offers complete 'white-glove service' to ensure customers have the complete satisfaction of our plug-n-play deployment. Supermicro believes this has allowed us to deliver complex AI infrastructure projects with reduced lead times and better value to our customers. For those interested in a quotation for SuperCluster or other solutions, please contact a Supermicro sales rep.

Further Information

Supermicro Generative AI SuperCluster: <https://www.supermicro.com/ai-supercluster>



Supermicro AI Infrastructure: <https://www.supermicro.com/ai>

Supermicro NVIDIA Solutions: <https://www.supermicro.com/accelerators/nvidia>



Supermicro GPU Systems: <https://supermicro.com/products/gpu>

Supermicro Liquid Cooling Solutions: <https://www.supermicro.com/liquid-cooling>

Appendix - SuperCluster Configurations

System Nodes	 4U 8-GPU SuperCluster Nodes	 8U 8-GPU SuperCluster Nodes
Overview	Liquid-cooled 32-node with 256 H100/H200 GPUs	Air-cooled 32-node with 256 H100/H200 GPUs
Part Number	SYS-421GE-TNHR2-LCC / AS-4125GS-TNHR2-LCC	SYS-821GE-TNHR / AS-8125GS-TNHR
CPU	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC™ 9004 Series	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC™ 9004 Series
Memory	2TB DDR5 (recommended)	2TB DDR5 (recommended)
GPU	NVIDIA® HGX™ H100/H200 8-GPU	NVIDIA® HGX™ H100/H200 8-GPU
Networking	8x NVIDIA ConnectX®-7 400Gbps/NDR OSFP 1 x BlueField®-3 DPU (B3220) NDR200 QSFP112, dual port	8x NVIDIA ConnectX®-7 400Gbps/NDR OSFP 1 x BlueField®-3 DPU (B3220) NDR200 QSFP112, dual-port
Storage	30.4TB NVMe (4x 7.6TB U.3) 3.8TB NVMe (2x 1.9TB U.3, Boot) [Optional M.2 available]	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	6x 3000W Redundant Titanium Level power supplies



Scalable Unit	 4U 8-GPU Liquid-Cooled SuperCluster	 8U 8-GPU Air-Cooled SuperCluster
Overview	Liquid-cooled 32-node, 256 H100/H200 GPUs	Air-cooled 32-node, 256 H100/H200 GPUs
Compute Leaf	8x NVIDIA Quantum-2 SSE-MQM9700-NS2F, 64-port 400G InfiniBand	8x NVIDIA Quantum-2 SSE-MQM9700-NS2F, 64-port 400G InfiniBand
Compute Spine	4x NVIDIA Quantum-2 SSE-MQM9700-NS2F, 64-port 400G InfiniBand	4x NVIDIA Quantum-2 SSE-MQM9700-NS2F, 64-port 400G InfiniBand
Converged (Storage & In-Band Management)	NVIDIA Spectrum-4 SSE-MSN5600-CS2R, 64 OSFP 800GbE	NVIDIA Spectrum-4 SSE-MSN5600-CS2R, 64 OSFP 800GbE
Out-of-Band	3x NVIDIA Spectrum SSE-MSN2201-CB2FC 48-port 1G, TOR	3x NVIDIA Spectrum SSE-MSN2201-CB2FC 48-port 1G, TOR
Rack / PDU	Rack: 5x 48U 750mmx1200mm. PDU: 18x 415V 60A 3Ph	Rack: 9x 48U 750mmx1200mm. PDU: 34x 208V 60A 3Ph

SuperCluster Current and Future Offerings

Supermicro's current Generative AI SuperCluster offerings include:

- Supermicro NVIDIA HGX H100/H200 SuperCluster with 256 H100/H200 GPUs as a scalable unit of compute in 9 racks
- Liquid-cooled Supermicro NVIDIA HGX H100/H200 SuperCluster with 256 H100/H200 GPUs as a scalable unit of compute in 5 racks (with one dedicated networking rack)
- Supermicro NVIDIA MGX™ GH200 SuperCluster with 256 GH200 Grace™ Hopper Superchips as a scalable unit of compute in 9 racks (with one dedicated networking rack)

SuperClusters are NVIDIA AI Enterprise ready with NVIDIA NIM microservices and NVIDIA NeMo platform for end-to-end generative AI customization and optimized for NVIDIA Quantum-2 400Gb/s InfiniBand and NVIDIA Spectrum-X Ethernet. Supermicro will offer AI rack solutions designed for NVIDIA's upcoming Blackwell Platform.

Supermicro's future SuperCluster offerings include:

- Supermicro NVIDIA GB200 NVL72 or NVL36 SuperCluster, liquid-cooled
- Supermicro NVIDIA HGX B100/B200 SuperCluster, air-cooled
- Supermicro NVIDIA HGX B200 SuperCluster, liquid-cooled

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions based upon your requirements.

Learn more at www.supermicro.com