



A BLUEPRINT FOR LLM AND GENERATIVE AI INFRASTRUCTURE AT SCALE

Supermicro SuperCluster Architecture



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	5
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	10
Evaluating Data Center Power and Other Resources	10
End-to-End Software Platform for Generative AI with NVIDIA AI Enterprise	11
Conclusion - Accelerate Time-to-Deployment	11
Further Information	12
Appendix - SuperCluster Configurations	13
SuperCluster Current and Future Offerings	14

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". This pre-validated bill of materials includes the rack enclosures, 32 Supermicro GPU systems, 256 NVIDIA H100/H200 GPUs, 12 400Gb/s InfiniBand switches, cables, PDUs, and more. SuperCluster is deployed as a complete rack-level solution, ready to tackle a broad range of AI applications. SuperCluster can be further expanded by adding additional scalable units to handle AI models of any size.

This white paper reveals blueprints of a Generative AI rack cluster with NVIDIA HGX™ H100/H200 GPUs. It 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.

To determine the amount of computing capacity required for the business, the math roughly breaks down like this:

- An AI model with 175B parameters takes about 314 million quadrillion floating-point operations to train.
- A single NVIDIA H100 GPU provides approximately 2 PFLOPS (floating point operations per second) of theoretical performance using the FP16 data type.
- 86,400 seconds in a day multiplied by 2 PFLOPS equals 172,800 quadrillion floating-point operations total, assuming a hypothetical 100% utilization rate.
- Therefore, it would take 1,817 days of training time to train a GPT-3 equivalent AI model with a single NVIDIA H100 GPU in this hypothetical example.

Most readers would likely agree that it is not practical to wait this long. But what if an enterprise harnessed the power of a 4.6MW cluster with 256 GPU system nodes, providing a total of 2,048 NVIDIA H100 GPUs? At a 40% utilization rate, it would take a little over two days to complete a training run, a far more realistic timeline to ship an AI model to production.

Calculating the number of required GPUs is a good starting point. 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. The GPU systems are populated in a total of 8 rack enclosures (48U) for the air-cooled version or four rack enclosures (48U) for the double-density liquid-cooled version. An additional rack enclosure hosts the networking components.



Figure 2 - Supermicro 8U HGX H100/H200 8-GPU Server

Although the Supermicro 8U 8-GPU NVIDIA HGX System is powerful on its own, it features 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

- 8x NVIDIA HGX 8-GPU H100 or H200, up to 141GB HBM3e per GPU
- 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

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 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.

There are a few different versions of Hopper-based AI GPUs, including the H100 PCIe, H100 NVL, and H100 SXM (a specialized socket for the GPU module). This system uses an SXM version, 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 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.

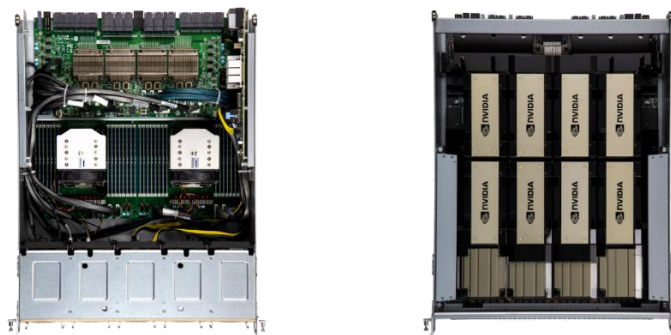



Figure 3 - Motherboard Tray (Left), GPU Tray (Right)



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 8-GPU baseboard. Its 4U height accommodates the tall heatsinks attached to the GPUs.

Effective system architecture is sometimes overlooked, but it's critical to ensuring high uptime, serviceability, and overall performance in a data center operating environment. The system architecture includes things like the chassis design and cooling. Next, we will explain the system topology to examine the data movement patterns within the system.

System Topology

The internal topology is laid out to ensure optimal bandwidth in two aspects: 1) communication between system components and 2) network connections outside the system. The GPU is the most critical component. In highly performant AI systems or clusters, the AI application ideally runs entirely within the high-bandwidth GPU memory. Therefore, the flow of application data should avoid movement through the CPU's PCIe bus.

The dual CPUs provide several PCIe lanes distributed across various PCIe switches. By design, there is no direct PCIe link between the CPUs and GPUs. Instead, four PCIe switches are placed in between them.

These "middleman" 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 additional expansion. These two additional PCIe x16 slots are often used with two additional NICs to interface with an attached high-performance storage cluster over a network fabric.

Although Supermicro NVIDIA HGX Systems are designed with a system architecture and topology intended for rack-scale, the importance of individual systems shouldn't be understated. Systems are a way to consolidate localized groups of computing resources, ensuring high uptime and high performance. An analogy is that a system is like a "squad" whereas a rack of systems can be envisioned as a "platoon". The following section reveals how these systems are organized into data center racks.

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 - Air-Cooled SuperCluster

Let's examine the rack layout, starting at the center of the rack cluster. The middle rack is for housing the networking switches. 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

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.

Liquid cooling presents an opportunity for substantial energy savings of up to 40% for the entire data center and substantial space savings. 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 shared coherent memory system, the compute nodes are blind to the activity of other nodes.

Networking is a way to solve the biggest challenges of AI computing at scale, including maintaining a coherent, high-capacity, high-bandwidth memory system. SuperCluster uses Infiniband NDR and NVIDIA InfiniBand QM9700 switches to connect GPU computing nodes across the network at 400Gb/s.



Figure 6 - NVIDIA InfiniBand QM9700 Switch

Parallelization strategies in LLM training and inference prioritize traffic to be sent between the local group of 8 GPUs within a system node, interconnected at 900GB/s via NVLink. In addition, any GPU can communicate with any other GPU in the cluster over the network fabric.

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 buzzwords:

- **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.

SuperCluster's network topology naturally follows the pattern established within the GPU systems themselves, like how the overall geometry of a city is defined by the layout of individual blocks. Each system node has 8 GPU-NIC pairs. These pairs can be labeled based on their position within the system by numbers 1 through 8; this number is referred to as their "rank". GPUs of the same rank are all routed to the same leaf switch. For example, the first GPU system node's "rank 1" NIC is connected to leaf switch #1. In addition, the last GPU system node's "rank 1" NIC is connected to the same leaf switch #1. Therefore, within a single system node, each NIC is connected to a different switch.

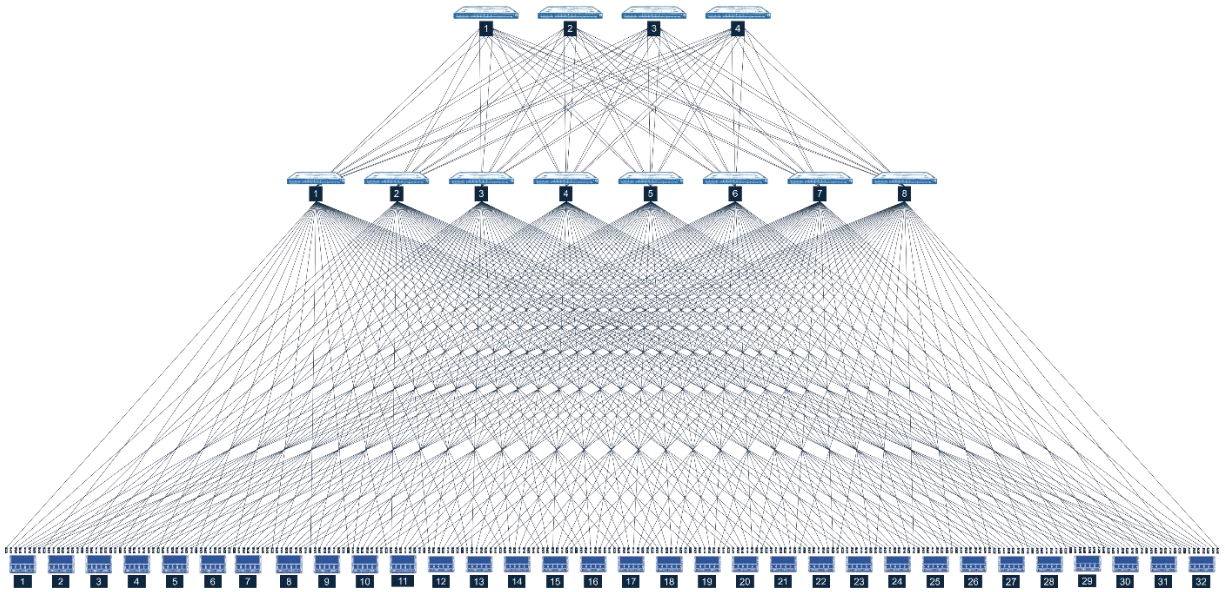


Figure 7 - SuperCluster Network Topology

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 contains a total of 12 NVIDIA Infiniband NDR QM9700 switches to unify its 256 GPUs:

- **4x NVIDIA InfiniBand QM9700 as Spine switches** (Layer 2) with NVIDIA InfiniBand MMA4Z00-NS 2x400Gb/s Twin-port OSFP Transceivers
- **8x NVIDIA InfiniBand QM9700 as Leaf switches** (Layer 1) with NVIDIA InfiniBand MMA4Z00-NS 2x400Gb/s Twin-port OSFP Transceivers
- **8x NVIDIA ConnectX®-7 per System Node** with NVIDIA InfiniBand MMA4Z00-NS400 400Gb/s OSFP Transceivers
- NVIDIA MPO-12/APC Passive Fiber Cables

The cluster is managed via two out-of-band management switches, two in-band management switches, and one IPMI switch per SuperCluster scalable unit. For large SuperClusters with multiple scalable units, one of the GPU system nodes should be switched out for a Unified Fabric Manager appliance node.

A choice must be made between the Ethernet protocol and the Infiniband protocol. By default, SuperCluster utilizes NDR Infiniband for its low latency, high scalability, and features such as UFM and RDMA. However, it is possible for customers to customize their SuperCluster package to use 400GbE with RDMA over Converged Ethernet (RoCE) as an alternative.

SuperCluster's topology delivers strong performance scaling as node count increases. It provides a single coherent memory system and allows for efficient communication between system nodes over a non-blocking network. This approach makes it suitable for large-scale AI training and real-time inference leveraging multiple AI models (MoE).

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, InfiniBand and Ethernet switches, 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:

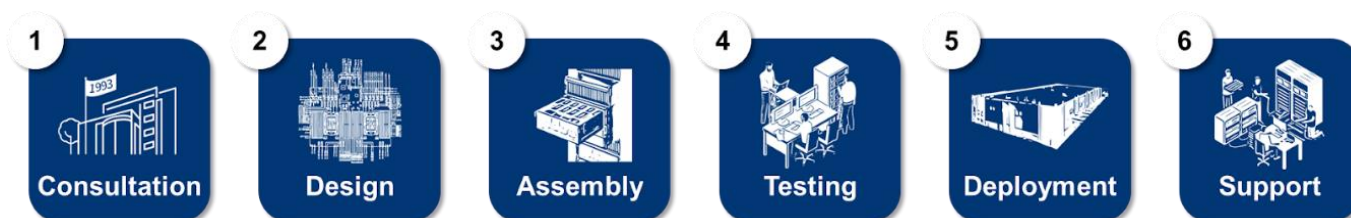


Figure 8 - Supermicro Solution & Integration Process

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 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	 1U MGX GH200 SuperCluster Nodes
Overview	Liquid-cooled 32-node with 256 H100/H200 GPUs	Air-cooled 32-node with 256 H100/H200 GPUs	256-node with 256 Hopper GPUs + Grace CPUs
Part Number	SYS-421GE-TNHR2-LCC / AS-4125GS-TNHR2-LCC	SYS-821GE-TNHR / AS-8125GS-TNHR	ARS-111GL-NHR
CPU	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC™ 9004 Series	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC™ 9004 Series	Integrated 72-core Grace Arm Neoverse V2 CPU
Memory	2TB DDR5 (recommended)	2TB DDR5 (recommended)	Up to 480GB of integrated LPDDR5X with ECC (CPU)
GPU	NVIDIA® HGX™ H100/H200 8-GPU	NVIDIA® HGX™ H100/H200 8-GPU	Integrated NVIDIA H100 Tensor Core GPU
Networking	8x NVIDIA ConnectX®-7 400Gbps/NDR OSFP 2x NVIDIA ConnectX®-7 200Gbps/NDR200 QSFP112	8x NVIDIA ConnectX®-7 400Gbps/NDR OSFP 2x NVIDIA ConnectX®-7 200Gbps/NDR200 QSFP112	2x NVIDIA ConnectX®-7 400Gbps/NDR OSFP, or 1x NVIDIA ConnectX-7 and 1x NVIDIA BlueField®-3
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]	Up to 8x Hot-swap E1.S drives and 2x M.2 NVMe drives
Power Supply	4x 5250W Redundant Titanium Level power supplies	6x 3000W Redundant Titanium Level power supplies	2x 2000W Redundant Titanium Level power supplies
↓			
Scalable Unit	 4U 8-GPU Liquid-Cooled SuperCluster	 8U 8-GPU Air-Cooled SuperCluster	 1U MGX GH200 SuperCluster
Overview	Liquid-cooled 32-node, 256 H100/H200 GPUs	Air-cooled 32-node, 256 H100/H200 GPUs	256-node, 256 Hopper GPUs + Grace CPUs
Compute Leaf	8x SSE-MQM9700-NS2F, 64-port 400G NDR	8x SSE-MQM9700-NS2F, 64-port 400G NDR	8x SSE-MQM9700-NS2F, 64-port 400G NDR
Compute Spine	4x SSE-MQM9700-NS2F, 64-port 400G NDR	4x SSE-MQM9700-NS2F, 64-port 400G NDR	4x SSE-MQM9700-NS2F, 64-port 400G NDR
In-Band	3x SSE-MSN4600-CS2FC 64-port 100GbE	2x SSE-MSN4600-CS2FC 64-port 100GbE	4x SSE-MSN4600-CS2FC 64-port 100GbE
Out-of-Band	2x SSE-G3748R-SMIS, 48-port 1Gbps ToR 1x SSE-F3548SR, 48-port 10Gbps ToR	2x SSE-G3748R-SMIS, 48-port 1Gbps ToR 1x SSE-F3548SR, 48-port 10Gbps ToR	8x SSE-G3748R-SMIS, 48-port 1Gbps ToR
Rack / PDU	Rack: 5x 48U 750mmx1200mm. PDU: 18x 415V 60A 3Ph	Rack: 9x 48U 750mmx1200mm. PDU: 34x 208V 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 400Gb/s NVIDIA Quantum-2 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