WHITE PAPER

A BLUEPRINT TO BUILD THE WORLD'S LARGEST LIQUID-COOLED GPU CLUSTER

Scaling out Supermicro SuperCluster with NVIDIA Spectrum™-X Ethernet

Contents

Executive Summary

As an industry leader in deploying infrastructure in some of the world's largest AI data centers, Supermicro offers a unique perspective on data center solutions. This whitepaper demonstrates the optimal data center infrastructure for building generative AI clouds and factories of any scale, powered by Supermicro NVIDIA HGX™ GPU servers and the NVIDIA Spectrum-X Ethernet networking platform. Supermicro's liquid-cooled SuperCluster powers some of the world's largest liquid-cooled AI clusters, achieving massive scale using the NVIDIA Spectrum-X Ethernet platform to connect up to more than 100,000 GPUs. This white paper outlines the blueprints of a Generative AI cluster featuring NVIDIA HGX H100/H200/B200 GPUs. It delves into the design of SuperCluster's individual system nodes, component selection, rack layout, network topology, and deployment steps.

Supermicro's SuperCluster reference architecture, paired with NVIDIA Spectrum-X, is designed to maximize performance for the most advanced model training and scale to exaFLOPS of performance. 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 B200/H200/H100 GPUs along with a NVIDIA Spectrum-X compute fabric, a SuperCluster scalable unit is the ultimate building block towards building the largest AI clusters in the world. As demands grow, this uniquely designed SU scales out effortlessly, utilizing the power of NVIDIA Spectrum-X to expand infrastructures, ensuring that customers will always have the capacity required to meet evolving computing demands.

Introduction - Overcoming the Biggest Challenges in AI Factories

In today's rapidly evolving AI landscape, staying competitive means optimizing your infrastructure for speed, security, and scalability. The rise of AI factories and AI clouds presents significant challenges to traditional data center infrastructures, especially in terms of networking performance and scalability. Supermicro's SuperCluster and NVIDIA Spectrum-X are purpose-built to address these obstacles by providing the highest-performing Ethernet fabric, explicitly designed to accelerate AI at the largest scales.

The heightened computing demands of AI applications present unique challenges for data centers. Solutions for large-scale AI infrastructure are a complex, multi-faceted effort with three major obstacles:

- **Parallel Compute 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.
- **Network Scalability:** The cluster topology needs to aggregate the compute 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 and high performance, key aspects of the data center deployment must be carefully planned, including data center power, floor plan, rack layout, and thermal management.

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

To understand how to scale effectively to the cluster level, here are the major GPU systems that comprise Supermicro's SuperCluster solution, which builds upon the server and rack levels.

Part 1: GPU System Building Blocks of the AI SuperCluster

Supermicro's extremely powerful 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 GPU systems, each containing 8 GPUs via an NVIDIA HGX baseboard. For each 32-node scalable unit (SU), the GPU systems 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.

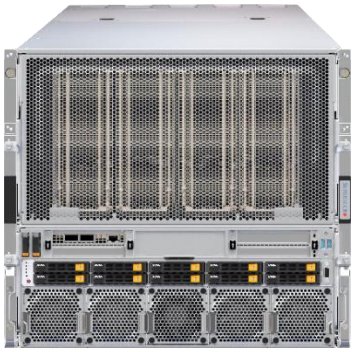

Figure 1 – Air-Cooled Supermicro NVIDIA HGX System

B200: SYS-A22GA-NBRT / AS -A126GS-TNBR
H100/ H200: SYS-821GE-TNHR / AS-8125GS-TNHR

Figure 2 – Liquid-Cooled Supermicro NVIDIA HGX System

B200: SYS-422GA-NBRT-LCC / AS -4126GS-NBR-LCC / SYS-421GE-NBRT-LCC
H100/ H200: SYS-421GE-TNHR2-LCC / AS-4125GS-TNHR2-LC

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 us first briefly cover some of the system's key components:

- Up to Dual Socket Intel® Xeon® 6 Processors or AMD EPYC™ 9004/9005 Series Processors
- NVIDIA HGX B200/H200/H100 8-GPU
- Memory Capacity:
  - 24 DIMMs, up to DDR5-4800 (AS -8125GS-TNHR)
  - 24 DIMMs, up to DDR5-6000 (AS -4126GS-NBR-LCC, AS -A126GS-TNBR)
  - 24 DIMMs, up to DDR5-6400 (SYS-422GA-NBRT-LCC, SYS-A22GA-NBRT)
  - 32 DIMMs, up to DDR5-5600 (SYS-421GE-NBRT-LCC, SYS-421GE-TNHR2-LCC)
- 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
- Spectrum-X Ethernet Compute Fabric: 8x NVIDIA BlueField®-3 SuperNICs, Single-Port 400GbE
- Storage/ In-Band Management Fabric (Converged Network): NVIDIA BlueField-3 DPU Dual-Port 200GbE
  Note: These systems are also compatible with NVIDIA Quantum-2 InfiniBand.

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 are heavily GPU-bound. Both AMD EPYC™ CPUs and Intel® Xeon® CPUs are available as options.

The NVIDIA B200/H200/H100 GPU has become synonymous with AI. The NVIDIA Hopper architecture's powerful parallel computing capabilities are specifically designed for AI applications, featuring redesigned streaming multiprocessors and a high-bandwidth memory system. The NVIDIA HGX platform features extremely fast local GPU-to-GPU interconnects: each of the system's 8 GPUs is connected by NVIDIA NVLink™. It creates a combined pool of coherent memory, enabling a single system to act as a powerful real-time inference engine, even without extending over a network. NVIDIA Blackwell doubles the effective bandwidth of the previous generation to 1.8TB/s of bidirectional throughput per GPU—over 14x the bandwidth of PCIe Gen5, ensuring high-speed communication for today's most complex large models.

System Architecture

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. There are eight AC input connections on the rear of the system to ensure reliable power delivery to the PSUs.

Running 8x GPUs with TDPs up to 1000W each 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. Air-cooled Supermicro Blackwell systems further enhance airflow by utilizing a 10U chassis to accommodate a larger GPU heatsink.

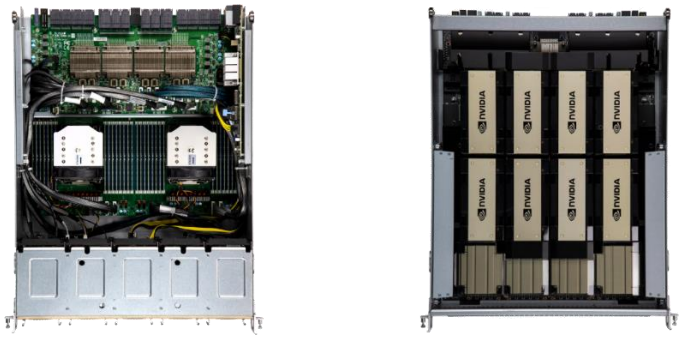

Figure 3 - Motherboard Tray (Left), GPU Tray (Right)

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, thereby improving thermal performance. Fan speed control is supported by thermal management using the BMC 2.0 interface. The GPU tray hosts the NVIDIA HGX B200/H200/H100 8-GPU baseboard. Its 4U height accommodates the tall heatsinks attached to the GPUs. On the other hand, the 4U 8-GPU system is completely liquid-cooled (both CPUs & GPUs) utilizing Supermicro's Direct-to-Chip Liquid Cooling (DLC) solution. Supporting the highest densities and the highest TDP CPUs and GPUs with up to 250kW cooling capacity per rack with Supermicro's in-rack Cooling Distribution Unit (CDU), the 4U liquid-cooled cluster is the ultimate resource-saving data center solution.

Supermicro NVIDIA HGX systems harness the power of NVIDIA BlueField-3 SuperNICs and BlueField-3 DPUs to scale out across the compute fabric and in-band management network. The NVIDIA BlueField-3 SuperNICs deliver up to 400Gb/s of Ethernet bandwidth, enabling AI networking, composable storage, zero-trust security, and GPU compute elasticity in hyperscale AI environments. This significantly enhances data center performance. Supermicro's most powerful GPU systems, along with NVIDIA BlueField-3 SuperNICs and BlueField3 DPU, enable agile and high-performance solutions that span edge-to-core data centers to clouds, all while enhancing network security and reducing the total cost of ownership.

Part 2: High-Performance Networking

Supermicro's deployment of Spectrum-X for AI compute fabric connectivity and BlueField-3 DPUs for in-band management and storage networking stands out due to our deep integration and expertise in high-performance computing and data center solutions. While many vendors offer networking hardware, Supermicro's close collaboration with NVIDIA ensures an optimized full-stack solution. The demand for AI workloads is growing at an unprecedented rate, while the adoption of generative AI is surging. Supermicro and NVIDIA are building some of the world's largest AI factories together — GPU-based data centers that combine NVIDIA NVLink with either InfiniBand or Spectrum-X Ethernet to achieve the highest levels of performance. Some AI factories are focused on training the largest-scale models and require a single tenant, while others are built for a wide range of AI workloads and require multi-tenancy. These multi-tenant AI factories often leverage cloud management and operations models, integrating into the Ethernet service and management network environment.

The NVIDIA Spectrum-X networking platform is the first Ethernet platform designed specifically to improve the performance and efficiency of Ethernet-based AI factories. Traditional Ethernet solutions struggle to meet the networking demands of modern AI models, leading to bottlenecks and inefficiencies. Spectrum-X overcomes this by delivering up to 1.6 times better AI networking performance, ensuring low latency and high effective bandwidth across massive AI clusters. Optimized from host to switch, Spectrum-X is designed to manage the needs of large-scale transformer-based generative AI models. NVIDIA BlueField-3 DPU complements this by offloading tasks such as software-defined networking, storage, and security, thereby dramatically increasing data center efficiency while reducing operational costs. Together, Spectrum-X and BlueField-3 enable seamless, secure, high-performance AI factories that can easily support the most demanding applications, setting a new standard for AI-driven data center networking.

2.1: Compute Fabric

The compute network fabric (also referred to as East-West, E-W) is designed for a multi-job and multi-tenancy AI cloud. Spectrum-X technology innovations include load balancing, congestion control, Quality of Service (QoS), and virtualization, enabling multiple tenants to run multiple jobs on the same infrastructure while maintaining complete security, isolation, and independence from each other, all while achieving the highest levels of performance.

The NVIDIA Spectrum-4 Ethernet switch offers a total of 128 x 400GbE or 64 x 800GbE ports for the GPU-GPU compute fabric:

Figure 4 - NVIDIA Spectrum-4 Switch

The NVIDIA BlueField-3 SuperNIC is utilized in the E/W compute fabric, optimized explicitly for accelerating AI workloads, featuring one 400GbE port, 8 ARM cores, and 16GB of onboard memory. Key features of BlueField-3 SuperNIC include accelerated networking for AI computing, best-in-class RoCE networking, high-speed out-of-order packet reordering, and

advanced end-user programmable congestion control, which makes it the most optimized NIC for E-W in GPU-accelerated systems.

Figure 5 - NVIDIA BlueField-3 SuperNIC

Network Topology

Supermicro SuperCluster deployments based on Spectrum-X are built from SUs, consisting of up to 32 HGX nodes, each with 8 GPUs. Each GPU on a given rail out of the eight rails of the HGX is one hop away from the respective GPUs on the other HGX systems within a given SU. The network is "rail-optimized," specifically configured to maximize performance and efficiency. By grouping the GPUs (and corresponding BlueField-3 SuperNICs) into "rails" and optimizing connectivity and communication within and across these rails, the network bandwidth is maximized, and costs are reduced.

When building a compute fabric based on Spectrum-X, scale is the primary consideration. This 32-node SU makes deployments straightforward to scale to many SUs, built with up to three tiers of cloud-based network with a non-blocking topology:

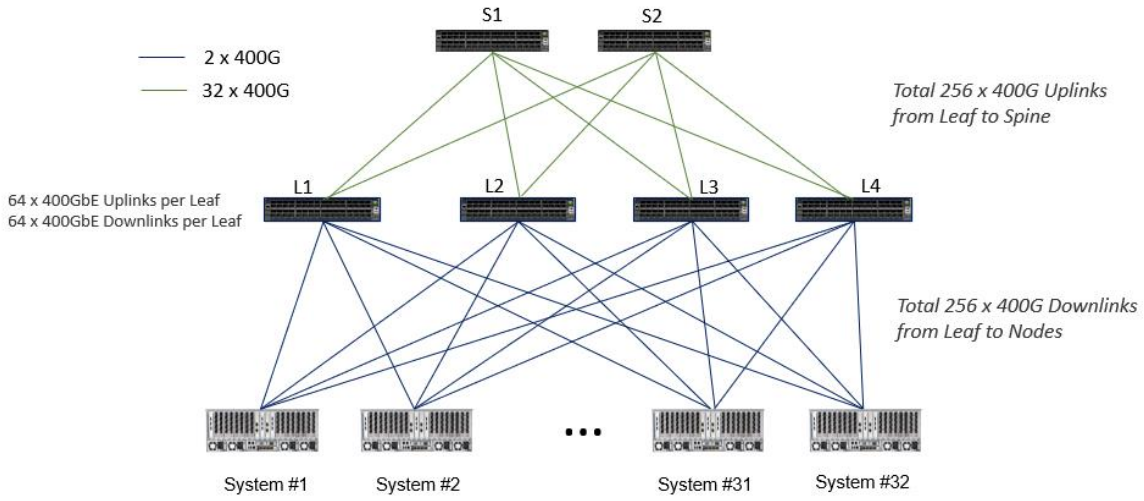

Figure 6- SuperCluster Scalable Unit (32 Nodes) Compute Network Topology – Spectrum X Ethernet

Scalable Unit Highlights:

- 32x GPU nodes, 256 B200/H200/H100 GPUs
- Compute Networking Fabric (E-W): 8x NICs (400GbE) for GPU compute network (BlueField-3 SuperNIC), non-blocking
- NVIDIA Spectrum-4 Switch: 4 Leaf, 2 Spine

Two-Tier Topology

In a two-tier topology, the maximum cluster size is 8K GPUs/ 1024 HGX nodes. The NVIDIA HGX nodes within each SU are connected to four leafs with rail-optimized connectivity where each leaf connects a separate rail group. Below is an example of a 128 node (four SU) cluster topology:

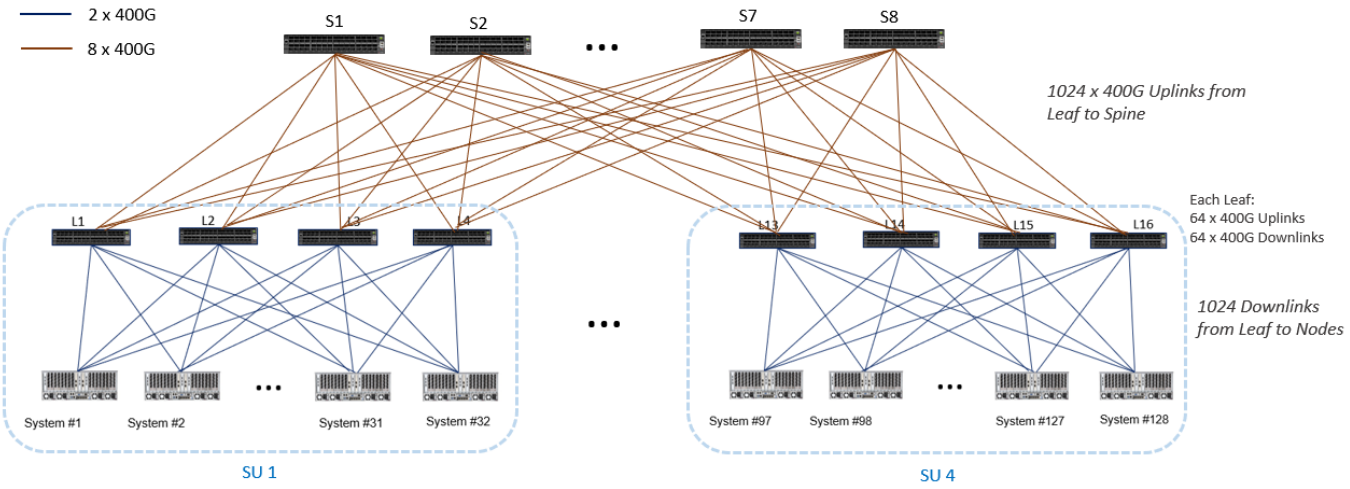

Figure 7- 128 Node (4 x SU) Compute Network Topology – Spectrum X Ethernet (2-tier)

128 Node Topology Highlights:

- 128 x GPU nodes, 1024 x B200/H200/H100 GPUs
- NVIDIA Spectrum-4 Switch: 16 Leaf, 8 Spine, NON-BLOCKING

Three-Tier Topology

When deploying a scale-out three-tier topology with more than 8K GPUs, the SUs are combined into PODs, each consisting of up to 64 SUs. Within each POD, rail-optimized connectivity is applied within the SU (NVIDIA HGX leaf) and between the leaf and spine layers. The leaf-spine rail-optimized connectivity forms rail blocks, each connecting a single rail group (out of 4) between the SUs in the POD. To provide non-blocking connectivity within the POD, each rail block consists of an equal number of leafs and spines. Inter-rail/ POD traffic is connected with the super-spine layer. Below is an example of a 32K GPU cluster:

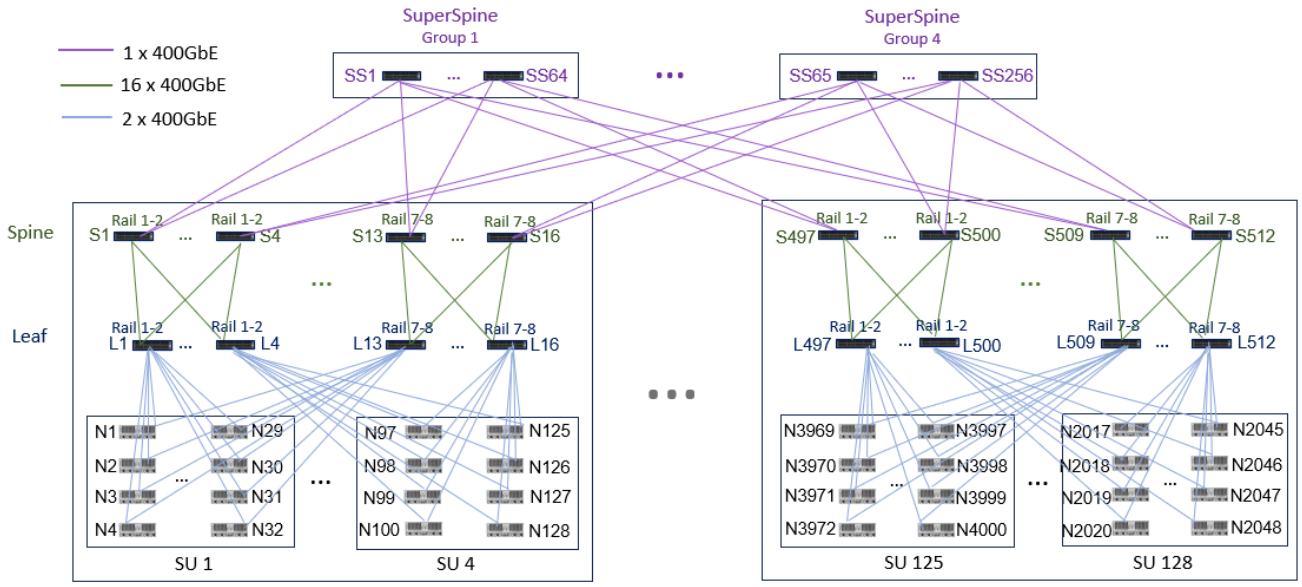

Figure 8 - SuperCluster 4096 Nodes (128 SUs, 32 PODs of 4 SUs each) Network Topology - Spectrum X Ethernet (3-tier)

4096 Node Topology Highlights:

- 4096 x GPU nodes, 32768 x B200/H200/H100 GPUs, Rail- optimized
- NVIDIA Spectrum-4 Switch: 512 Leaf, 512 Spine, 256 SuperSpine
- 128 x Scalable Unit (32 Nodes Each)
- 4 x SuperSpine Groups, each with 64 Switches
- 64 x 400GbE Uplinks, 64 x 400GbE Downlinks per Leaf & Spine Switch

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

The converged network comprises the storage network fabric and the in-band management fabric. This provides flexible storage allocation and simplified network management, while also enabling in-band monitoring. The converged network also provides high bandwidth to the High-Performance Storage (HPS) tier 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 both storage and application performance. The in-band management fabric maximizes stability, performance, and ease of management by utilizing an underlay network, while also providing redundancy. It also provides connectivity for in-cluster services, such as Slurm, as well as other services outside the cluster, including 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 be provided utilizing NVIDIA Spectrum-4 Ethernet switches via the NVIDIA BlueField-3 DPU (2 x 200G connections per node). Key features of the NVIDIA BlueField-3 DPU include offloading, accelerating, and isolating data center infrastructure, secure zero-trust management, data storage acceleration, NVIDIA DOCA SDK, and a cloud Infrastructure processor, which makes it the most optimized solution for N-S in GPU-accelerated systems.

NVIDIA Spectrum-4 Ethernet Switch

NVIDIA BlueField-3 DPU

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

2.3: Out-of-Band Management

The out-of-band Ethernet network is used for system management via the BMC and IPMI, providing connectivity to manage all networking equipment. Out-of-band management is crucial to the cluster's operation, providing low-usage paths that ensure management traffic does not interfere with other cluster services. The NVIDIA 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 NVIDIA SN2201 delivers unique and innovative features to 1G switching.

Figure 10 - NVIDIA Spectrum SN2201 (1GbE Switch)

Part 3: 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 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 customers have some flexibility to adjust as needed.

Optimizing the rack layout is based on several 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)

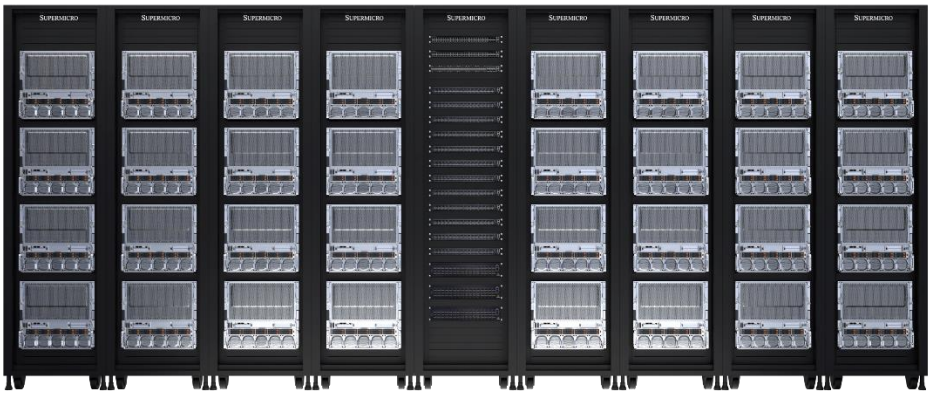

Figure 11 – Air-cooled SuperCluster with NVIDIA HGX B200

Let's examine the rack layout, starting at the center of the rack cluster. The middle rack houses the networking switches, including NVIDIA Spectrum-X Ethernet switches, which create the most optimized AI networking fabric, as well as additional Ethernet switches for ideal storage and management. On both sides, eight identical "compute racks" contain four 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, providing more thermal headroom to prevent throttling.

Doubling Computing Density Per Rack with Liquid Cooling

Liquid cooling offers the potential for substantial energy savings of up to 40% across the entire data center, as well as significant 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 is the cluster's compute node. Both the CPUs and GPUs are liquid-cooled with cold plates that efficiently move heat away from the chips.

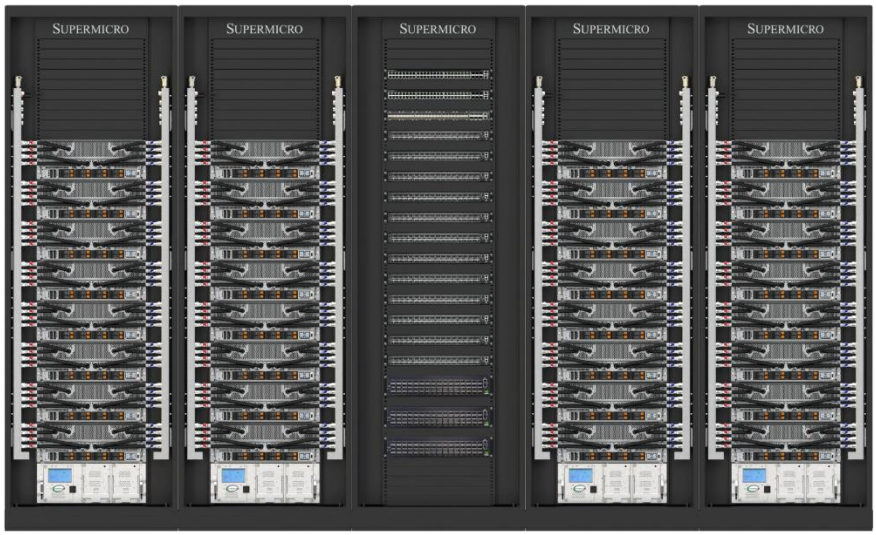

Figure 12 - Liquid-Cooled SuperCluster

The increased cooling efficiency enables 8 x 4U systems to be accommodated 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 Coolant Distribution Unit (CDU) is located 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. Supermicro NVIDIA HGX H200/H100 4U systems are paired with 1U manifolds that handle the distribution of liquid to and from the systems. Supermicro NVIDIA HGX B200 systems utilize a vertical CDM to further improve rack density by freeing up rack units previously occupied by the horizontal CDMs. 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 utilized 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 tailored to each socket type and baseboard. 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 the ease of the deployment process.

Part 4: SuperCluster Solution Design and Deployment

Designing and deploying AI infrastructure involves a multi-staged process, beginning with 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 customer's application needs, existing software

and hardware infrastructure, and the data center deployment environment. Supermicro can adjust the 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: [the images below are non-standard for WPs; not sure if they add much]

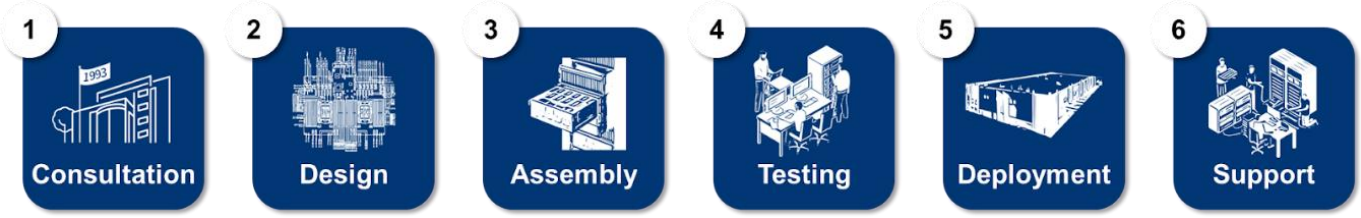

Figure 13 - Supermicro Solution & Integration Process

Evaluating Data Center Power and Other Resources

Not all data center configurations can be addressed in this paper; however, we will briefly cover SuperCluster's power requirements to illustrate an essential aspect of the consultation and design process. Data center power specifications will vary depending on factors such as 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 a critical metric for classifying 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 available power for the data center covers the power draw of the cluster. Each system in an air-cooled 8U 8-GPU system-based SuperCluster consumes approximately 9 kW of power. A SuperCluster Scalable Unit (SU) with 32 nodes (256 GPUs) consumes about 288kW of power, plus roughly 25kW from the networking components.

One of our reference air-cooled SuperCluster architectures utilizes 34 208V 60A 3-phase PDUs to power nine racks, whereas liquid-cooled SuperClusters may utilize 18x 415V 60A 3-phase PDUs to power five racks. To calculate total power, multiply volts by amps by the number of PDUs (208 x 60 x 34 for air-cooled / 18 x 415 x 60A for liquid-cooled), which equals 424kW / 448kW of power. This power capacity is enough to drive the SuperCluster with headroom to spare. Please note that the power requirements of other SuperCluster configurations may 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 extends beyond application-specific limitations to support the growing ways 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 control over infrastructure choices, data sharing, and strategic use of generative AI. 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 the deployment of generative AI 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.

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 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 be carried 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, targeted to build the most optimized AI factory, 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 monthly manufacturing capacity of up to 5,000 racks. Supermicro offers a complete ”white-glove service” to ensure customer satisfaction with our plug-and-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/en/accelerators/nvidia
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

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