Supermicro AI Networking SuperSwitch Portfolio

Pioneering Green Solutions for a Sustainable AI Infrastructure

The Supermicro AI Networking Solution

The ever-growing demands of Artificial Intelligence (AI) require a robust and scalable network infrastructure. Supermicro empowers your AI deployments with industry-leading networking solutions. Our comprehensive portfolio of high-performance data center switches is designed to connect and manage your AI infrastructure seamlessly.

Unparalleled Performance

Our Supermicro network switches deliver exceptional performance with blazing-fast speeds, low latency, and high fabric capacity. This ensures smooth data flow, even for demanding AI workloads like deep learning and machine learning training, achieving faster training times and quicker inference cycles with unmatched performance.

Scalability for Growth

Seamlessly adapt your network fabric to accommodate the ever-growing demands of your AI training data center with our solutions that cater to deployments of all sizes. Our AI networking Switch portfolio combines Open Networking SONiC NOS distribution with advanced layer 3 capabilities and hardware equipped with advanced Switch ASIC, allowing you to scale your network fabric as your AI infrastructure demands evolve.

Advanced Features for Optimized AI Environments

Quality of Service (QoS): Prioritize critical AI traffic for real-time performance with advanced ROCEv2 feature set support for end-to-end congestion control (PFC, ECN) & traffic management (DLB).

In-band Telemetry: Gain real-time insights into network performance for proactive management and troubleshooting.

Network Analytics: Analyze network traffic patterns to optimize resource allocation and identify potential bottlenecks.

Simplified Management and Control

All Supermicro network switches come equipped with user-friendly management tools, including industry standard CLI, SNMP, gNMI, and RESTful API, which offer deeper control and automation capabilities.

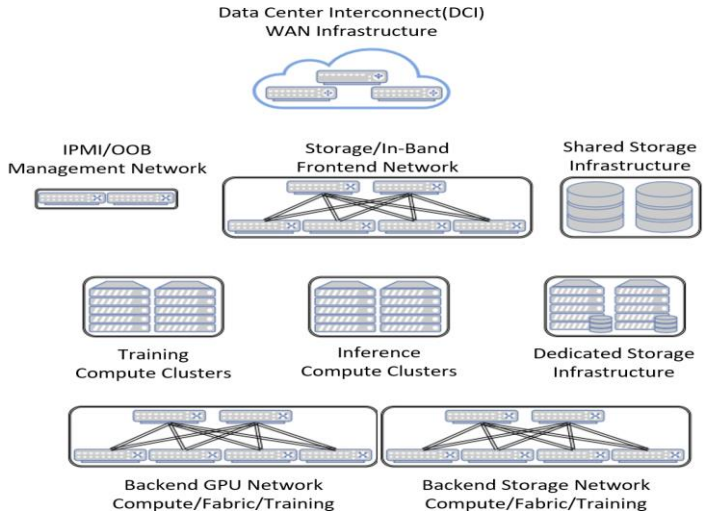

Figure 1: Typical AI/ML Training & Inference DC Infrastructure

Optimal TCO for AI Network Infrastructure

Our solutions offer pre-validated designs qualified with all the leading GPU, NIC, and storage vendors, providing turn-key deployments with short lead times. Our switches enable exceptional cost savings with no additional software licensing cost involved, as well as certified standard cables and transceivers with no vendor lock-ins. Supermicro's exceptional on-site support and services team for rack-scale designs is backed by this.

Supermicro Ethernet switches provide high-speed, high-density, and high-compatibility switches for modern AI/ML pipelines. Our switches provide the foundation for a highly performant and scalable AI infrastructure. Our solutions empower you to unlock the full potential of your AI applications and accelerate innovation.

Switch Type	Product SKUs	Hardware Details	Software
64x 400G ToR/Leaf/Spine	SSE-T8032S 	**Ideal for 400G Leaf/Spine nodes in large AI clusters** • 1RU 25.6T Switch Fabric with 64x 400G in 32 OSFP cages • 2x10G SFP+ and 1x1G RJ45 dedicated Management Port • Scales with supporting multiple Port Breakout Options • Redundant hot swap 1+1 power & N+1 fans option • Supports up to 2048 GPUs in a backend training network per pod/scaled unit(SU)	• Broadcom Advanced Enterprise SONiC OS • Optimized, Cable/TRX assurance • Distributed & Supported by Supermicro • No additional software licensing cost
32x 400G ToR/Leaf/Spine	SSE-T7132S 	**Ideal for high-density 100G/200G storage networking** • 1RU 12.8T Switch Fabric with 32x 400G QSFP-DD • 2x10G SFP+ and 1x1G RJ45 dedicated Management Port • Scales with supporting multiple Port Breakout Options • Redundant hot swap 1+1 power & N+1 fans option	• Supermicro Enterprise SONiC OS • Optimized, Cable/TRX assurance • Distributed & Supported by Supermicro • No additional software licensing cost
32x 100G ToR/Leaf	SSE-C4632SB 	**Ideal for high-density 10G/25G ToR network aggregation** • 1RU 3.2T Switch Fabric with 32x 100G QSFP28 • 1x10G SFP+ and 1x1G RJ45 dedicated Management Port • Scales with supporting multiple Port Breakout Options • Redundant hot swap 1+1 power & N+1 fans option	• Broadcom Advanced Enterprise SONiC OS • Optimized, Cable/TRX assurance • Distributed & Supported by Supermicro • No additional software licensing cost
48x 1G Management	SSE-G3748-SONIC 	**Ideal for dedicated management networks in Datacenters** • 1RU 396G Switch Fabric with 48x 1G RJ45 and additional uplink ports 6x 10G/25G QSFP+ • 1x1G RJ45 dedicated Management Port • Redundant hot swap 1+1 power & N+1 fans option	• Supermicro Enterprise SONiC OS • Optimized, Cable/TRX assurance • Distributed & Supported by Supermicro • No additional software licensing cost

Figure 2: Supermicro AI Networking Switch Portfolio

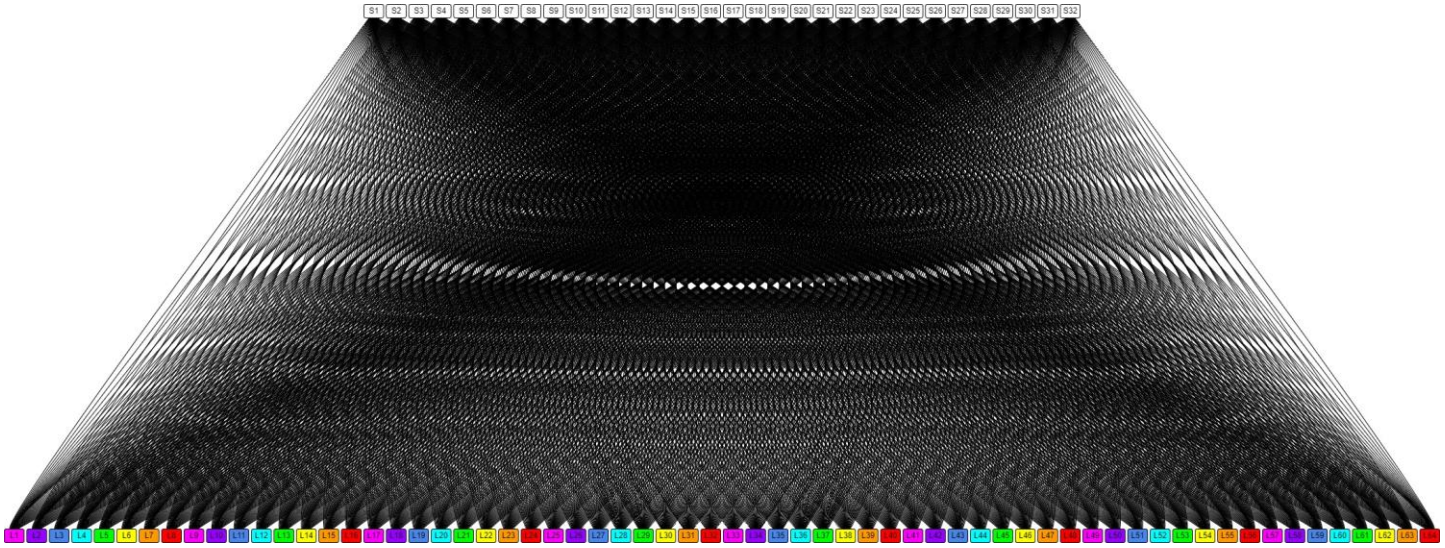

Figure 3: Rail-optimized cluster for 2048 GPUs using SSE-T8032

Details of Rail-optimized cluster for 2048 GPUs using SSE-T8032:

- Cluster size scales from 256 GPUs per scaled unit (SU) to 2048 GPUs, reaching a maximum of 8 SUs in a 2-tier fabric
- Comprises of a 2-tier non-blocking CLOS fabric with 64-colored Leafs (L1 - L64) (each color/rail identifies a specific GPU# - NIC# connection) inTier-1 & 32-uncolored Spines (S1-S32) in Tier-2
- The total number of 400G interconnects between leafs-spines and leaf-compute nodes scales from 256 to 2048
- The scale of each compute with 8-GPUs & 8-NICs connected to the colored Leafs goes up from 32 per SU to 256 spread across several racks
- In an equivalent non-rail optimized cluster, all 8x GPU#-NIC# connections from each compute terminate on 2-uncolored leaf pair per SU in Tier-1 scaling from 64 GPUs per scaled unit (SU) to 2048 GPUs reaching a maximum of 32 SUs, computes scale from 8 per SU to 256 across racks
- Higher scale clusters beyond 2048 can be realized by extending it to a 3-tier non-blocking CLOS fabric with Super spine nodes

Ideal Applications:

- **SSE-G3748:** Perfect for secure and reliable 1G network for out-of-band (OOB) remote IT infrastructure management.
- **SSE-C4632:** Well-suited for high-density 10G/25G ToR network aggregation with the 100G port breakout capability, providing efficient low-latency connectivity storage and compute nodes.
- **SSE-T7132:** Enables high-density 100G/200G, low-latency & high-throughput storage connectivity solutions with the 400G port breakout option in AI clusters.
- **SSE-T8032:** Suited for large-scale AI training and inference deployments with optimal performance and enhanced features requiring high port density and scalability with dual 400G OSFP transceivers.

Why Choose Supermicro?

- **Total Solution:** Achieve the most optimal AI workload performance with a complete, pre-validated solution, including fine-tuned network configurations and templated designs.
- **Unmatched Performance:** Deliver industry-leading throughput and forwarding rates for demanding AI workloads.
- **Scalable Solutions:** Seamlessly scale your network fabric as your AI infrastructure grows.
- **Advanced Features:** Leverage cutting-edge functionalities to optimize your AI deployments.
- **Unified Management:** Simplified management tooling to automate & orchestrate for templatized AI cluster network fabric deployment for Day-0 to Day-2+ configuration, audit, monitoring, performance, and alerting across the fabric nodes.

Next Steps:

- Visit the SMC website for detailed product information: https://www.supermicro.com/en/products/networking/switches
- Contact your SMC sales representative to discuss your specific AI networking needs.

Discover how SMC network switches can power your next-generation AI infrastructure.

Disclaimer: Specifications and features may vary by model. Please refer to the SMC website for detailed product information.

#SMC #AI #Networking #HighPerformance

https://www.supermicro.com/en/products/networking/switches