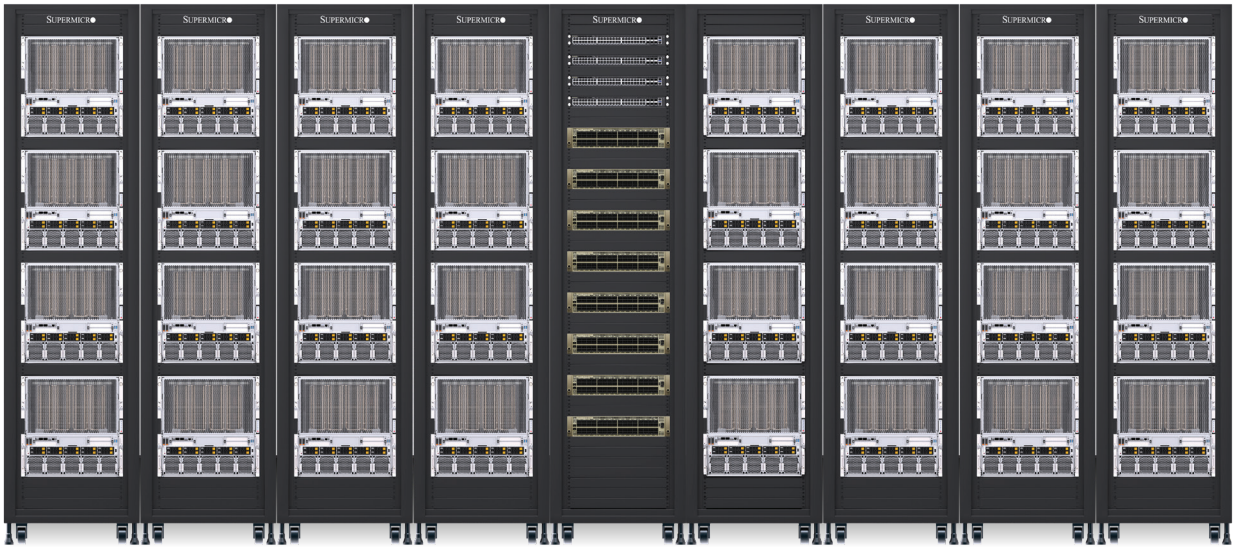
AI Factory Solutions with NVIDIA HGX™ B200

Flexible, proven end-to-end solutions to accelerate at-scale AI factory deployments



Why Choose Supermicro & NVIDIA for AI Factories?

AI factories from Supermicro and NVIDIA are complete, turnkey solutions simplifying the deployment of AI at scale for faster time-to-online and time-to-revenue, with full-stack solutions including compute, software, networking, and storage. Supermicro delivers AI infrastructure optimized for performance and efficiency, with fully-integrated solutions based on NVIDIA Enterprise Reference Architecture Designs and NVIDIA-Certified Systems™ for guaranteed full-stack performance and compatibility. Supermicro's rack-level testing and validation goes beyond industry standards, ensuring quality and seamless plug-and-play deployment for complete AI confidence.

Industry-leading Time-to-Online for the Latest AI Technologies

Supermicro has a proven track record of rapidly bringing new acceleration technologies to market, with a flexible building block approach enabling faster adoption cycles for new NVIDIA accelerated compute platforms. Supermicro can help enterprises bring latest-generation AI infrastructure online faster, accelerating time-to-revenue and maximizing AI-powered competitive advantage. With production capacity in the USA of over 5,000 racks per month, Supermicro is able to build, test, and validate cluster-scale deployments faster, ensuring solutions are delivered ready to begin generating revenue from day one. Additionally, Supermicro's Data Center Building Block Solutions® (DCBBS) can further facilitate the build-out of enterprise AI factories, providing everything needed to develop or refurbish a data center to become an AI factory, reducing lead times and eliminating coordination between multiple vendors.

Flexible, End-to-End AI Solutions Tailored to Your Enterprise

Supermicro offers a broad portfolio of accelerated systems supporting NVIDIA HGX GPUs, enabling customers to create AI solutions that are optimized for maximum AI performance. At scale, Supermicro provides complete AI cluster solutions, backed by deep expertise in networking, topology design, deployment, and cabling. A comprehensive storage portfolio is also available, supporting every stage of the AI data pipeline and integrating the NVIDIA AI Data Platform to simplify data workflows and accelerate innovation. Supermicro AI factory solutions are endorsed by NVIDIA for Infrastructure Configuration, NVIDIA Spectrum™-X Ethernet, and Software Reference Stack and based on the NVIDIA Enterprise Reference Architecture for HGX™ B200.

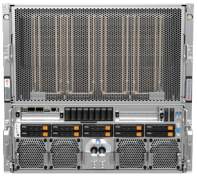
Turnkey solutions with Enterprise-grade support from Supermicro & NVIDIA

Working in close cooperation, Supermicro and NVIDIA ensure performance-optimized AI hardware integrates easily into full-stack AI solutions. Supermicro's range of NVIDIA-Certified Systems™ is fully tested and validated for performance, reliability, and compatibility with the NVIDIA software stack (NVIDIA AI Enterprise and NVIDIA Run:ai), NVIDIA Spectrum-X Ethernet networking, and forms the building blocks for scaling AI factories seamlessly. As a single-vendor provider, Supermicro supplies everything for a complete AI factory while controlling quality, integrity, and compatibility across the supply chain. Complete L12 system and cluster-level validation before shipment ensure seamless plug-and-play deployment at any scale.

Maximize Performance with NVIDIA HGX

The NVIDIA HGX platform brings together the full power of NVIDIA GPUs, NVIDIA NVLink®, NVIDIA networking, and fully optimized AI and high-performance computing (HPC) software stacks to provide the highest application performance and drive the fastest time to insights for every data center. The NVIDIA HGX B200 integrates eight NVIDIA Blackwell GPUs with high-speed fifth-generation NVLink® for up to 1.8TB/s of GPU-GPU interconnect to accelerate AI performance at scale. The NVIDIA HGX B200 propels the data center into a new era of accelerating computing and generative AI to accelerate AI performance at scale for the most demanding AI, data analytics, and HPC workloads.

AI Factory Solutions	SRS-48AC-4N-B200SX	SRS-48AC-8N-B200SX	SRS-48AC-32N-B200SX
Nodes per Cluster	4	8	32
GPUs per Node/Cluster	8/32	8/64	8/256
System SKUs	SYS-A22GA-NBRT-G1	SYS-A22GA-NBRT-G1	SYS-A22GA-NBRT-G1
Networking	NVIDIA Spectrum-X Ethernet	NVIDIA Spectrum-X Ethernet	NVIDIA Spectrum-X Ethernet
Node Pattern (CPU-GPU-NIC-Bandwidth)	2-8-9-400	2-8-9-400	2-8-9-400
Power per Rack (4-node)	53.6kW	53.6kW	53.6kW
Target Deployment Use Cases	High-volume AI inference, foundation AI model training, HPC		



NVIDIA-Certified Systems	SYS-A22GA-NBRT-G1
Form Factor	10U
GPU	NVIDIA HGX B200 8-GPU
CPU	2x Intel® Xeon® 6960P Processor 72-Core 2.70GHz 432MB Cache (500W)
Memory	24x 96GB DDR5 6400MHz ECC RDIMM
Local Node Storage	4x PCIe5.0 NVMe U.2 2x NVMe M.2
Networking	1x NVIDIA BlueField®-3 (B3220) 8x NVIDIA BlueField-3 (B3140H)
Node Max Power Draw (Full Load)	13.4kW
Node Max Heat (Full Load)	47,390 BTU/h