AI Factory Solutions with NVIDIA HGX™ B300

Full-stack, 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's building block architecture enables rapid adoption of new NVIDIA accelerated compute platforms, helping enterprises bring the latest AI infrastructure online faster to accelerate time-to-revenue and sharpen competitive advantage. With U.S. production capacity exceeding 5,000 racks per month, Supermicro builds, tests, and validates cluster-scale deployments at speed — delivering solutions ready to generate revenue from day one. Supermicro's Data Center Building Block Solutions® (DCBBS) further streamline AI factory build-outs by providing everything needed to develop or modernize a data center, reducing lead times and eliminating multi-vendor coordination complexity.

Flexible, End-to-End AI Solutions Tailored to Your Enterprise, Endorsed by NVIDIA

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. From data preparation to model training and inference, Supermicro's comprehensive storage system portfolio with NVIDIA AI Data Platform integration ensures enterprises can unlock the full value of their data for AI-driven innovation. Supermicro's NVIDIA HGX B300 solution is endorsed by NVIDIA for Infrastructure Configuration, optimized for NVIDIA Spectrum™-X Ethernet, and based on the NVIDIA Enterprise Reference Architecture for NVIDIA HGX™ B300 GPUs.

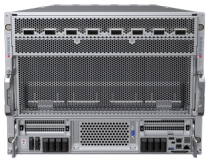
Turnkey solutions with Enterprise-grade support from Supermicro & NVIDIA

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.

Air-Cooled Design with Ultra Performance

The Supermicro NVIDIA HGX platform is the building block of the world's largest AI clusters, delivering the immense computational output required for powering today's transformative AI applications. Supermicro's NVIDIA HGX B300 solutions enable rapid deployment of the industry's highest-performing AI infrastructure to tackle the largest-scale AI training, real-time AI reasoning, agentic AI applications, multimodal AI inference, and physical AI applications. The 8U air-cooled system is designed to maximize performance of eight 1100W TDP NVIDIA HGX B300 GPUs with up to 2.3TB of total HBM3e memory*. The front eight OSFP ports supporting integrated NVIDIA ConnectX®-8 SuperNICs at 800 Gb/s enable plug-and-play deployment with NVIDIA NVIDIA Spectrum-X Ethernet compute fabric.

AI Factory Solutions	SRS-48AC-4N-B300SX	SRS-48AC-8N-B300SX	SRS-48AC-32N-B300SX
Nodes per Cluster	4	8	32
GPUs per Node/Cluster	8/32	8/64	8/256
System SKUs	SYS-822GS-NB3RT		
Networking	NVIDIA Spectrum-X Ethernet	NVIDIA Spectrum-X Ethernet	NVIDIA Spectrum-X Ethernet
Node Pattern (CPU-GPU-NIC-Bandwidth)	2-8-9-800	2-8-9-800	2-8-9-800
Power per Rack (4-node)	60kW	60kW	60kW
Target Deployment Use Cases	High-volume AI inference, foundation AI model training, FP32 HPC, large scale Agentic AI workloads		



NVIDIA-Certified Systems	SYS-822GS-NB3RT
Form Factor	8U
GPU	NVIDIA HGX B300 8-GPU (288GB HBM3e per GPU)
CPU	SYS-822GS-NB3RT: 2x Intel® Xeon® 6767P Processor 64-core 2.40GHz 336MB Cache (350W)
CPU Memory	SYS-822GS-NB3RT: 24x 64GB DDR5 6400MT/s ECC RDIMM
Local Node Storage	8x hot-swap E1.S NVMe drive bays 2x M.2 NVMe slots
Networking	1x NVIDIA BlueField®-3 (B3240) 8x integrated NVIDIA ConnectX®-8 SuperNICs, up to 800Gb/s
Node Max Power Draw (Full Load)	15kW
Node Max Heat (Full Load)	50,000 BTU/h