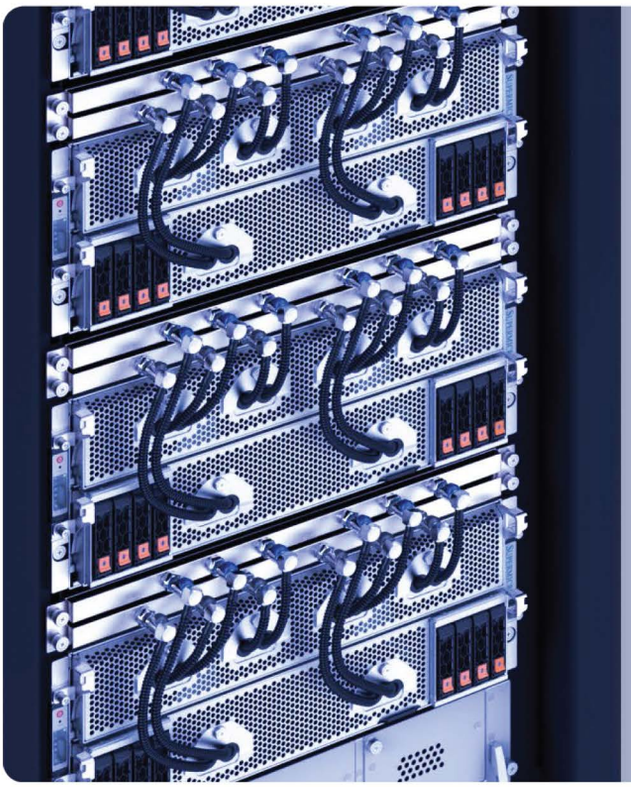
Rack Scale Liquid Cooling Solutions

Superior Cooling, Density, and Sustainability



UP TO 40% REDUCTION in electricity costs for entire data center

UP TO 55% REDUCTION in data center server noise

UP TO 89% REDUCTION in electricity costs for server cooling infrastructure

Supermicro liquid cooling solutions can reduce OPEX by up to 40%, and allow data centers to run more efficiently with lower PUE. Supermicro has proven liquid cooling deployments at scale and enables data centers operators to deploy the latest and most performance CPUs and GPUs.

Liquid Cooling Rack Sample Configurations

SRS-48UGPU-SKU1-L1-SMCI
Up to 8 GPU servers (4U, 8 NVIDIA H100/ H200 GPU per sever) per 48U rack (64 NVIDIA H100/200 GPUs per rack

SRS-48UBTW-SKU1-L1-SMC
Up to 76 server nodes / 19 servers in a 48U rack

SRS-48UBLD-SKU1-L1-SMCI
Up to 80 server blades / 4 enclosures in a 48U rack

Supermicro Liquid Cooling Solutions

Supermicro Direct-to-Chip Cooling Solutions

- Broad range of modular cold plate designs
- Unique liquid cooled server designs to double GPU density at server and rack level
- Rack scale validation of customer applications and environments to ensure the highest quality and satisfaction
- Plug-and-play data center level integration readiness
- Single-vendor total IT solution from design to delivery

Liquid Cooling Key Components

Coolant Distribution Unit (CDU)

Contains the pumping system that circulates coolant to the cold plates, which carry heat away from CPUs, GPUs, and DIMMs.

CPU/GPU Cold Plate

Modular design cold plates for exceptional thermal performance and minimal pressure drop

Coolant Distribution Manifold (CDM)

CDM are the distribution pipes that supply coolant to each server and collect the hotter coolant back to the CDU.

Supermicro Liquid Cooling Servers

Universal GPU Server: Flexible support for NVIDIA, AMD, and Intel GPU

SYS-821GE-TNHR

8U 8-GPU, NVIDIA HGX H100 SXM5

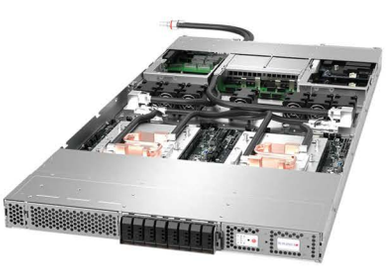
SYS-421GE-TNHR2-LCC

4U 8-GPU, NVIDIA HGX H100

SYS-221GE-TNHT-LCC

2U 4-GPU, NVIDIA HGX H100

NVIDIA MGX™ Server



ARS-111GL-DNHR-LCC

1U 2-Node, NVIDIA GH200

AMD APU Server

AS-2145GH-TNMR

2U 4-APU, AMD Instinct™ MI300A

BigTwin® Server

SYS-221BT-HNTR

2U, 4-Node, 16 DIMMs

Supermicro Intel Servers

Supermicro Gaudi®2 AI Training Server

Gaudi2 AI Training Server prioritizes 2 keys real-world considerations: integrating multi AI training system to analyze diverse AI models faster, while simultaneously multiple scalability function and price advantages. Gaudi2 enables up to 40% better price/performance for deep learning training than traditional AI solutions, for advanced training and inference performance.

SYS-820GH-TNR2
8U, 8 Gaudi®2 HL-225H

Supermicro X14 Servers Support Intel Xeon 6 Processors

The latest generation of proven platforms designed for maximum performance, efficiency, and flexibility for AI, Cloud, Storage, and 5G/Edge workloads

- Industry's broadest portfolio of systems based on Intel® Xeon® 6 processors
- Supermicro liquid cooling including CPU/GPU cold plates, Cooling Distribution Unit and Cooling Distribution Manifolds for a complete integrated solution
- Support for the latest industry technologies including PCIe 5.0, DDR5, CXL 2.0, Open Compute Project (OCP) DC-MHS and OCP 3.0, as well as EDSFF E3.S and E1.S storage form factors

X14 Universal GPU Server
8U air-cooled and 4U liquid-cooled servers with the most advanced GPU

X14 8U SuperBlade
Blade server with up to 10-Node in 6U or 20-Node in 8U

X14 Multi-Node Server
2U 2-Node and 2U 4-Node optimized for compute or storage density

X14 Hyper Server
1U and 2U architectures with flexible I/O options

X14 Hyper-E Server
2U with front I/O and optional DC power for edge data centers

X14 CloudDC with DC-MHS
for Cloud data centers designed to OCP DC-MHS specifications

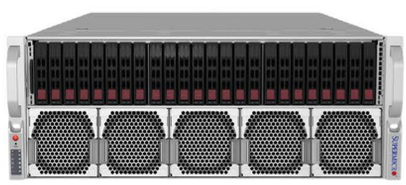
Supermicro AMD Accelerated Servers

Supermicro Servers with AMD Instinct™ MI300 Accelerators

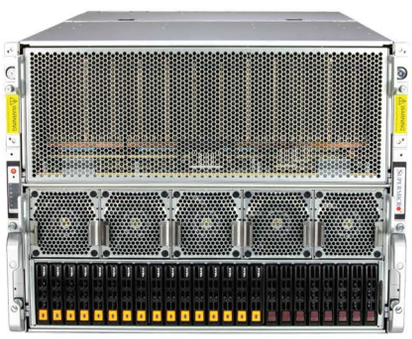
Supermicro servers incorporating the AMD Instinct MI300A or MI300X accelerators, are a leap forward in system design for demanding AI and HPC workloads. The 8U Universal GPU server, which includes 8 AMD Instinct MI300X accelerators and Dual AMD EPYC 9004 series processors, offers exceptionally high performance for HPC and AI workloads, significantly improving over previous generations of AMD Instinct accelerators.

In addition, the new 2U and 4U servers with Quad AMD Instinct MI300A accelerators, which combine CPU and GPU, leverage Supermicro's expertise in multiprocessor system architecture and cooling design, finely tuned to tackle the convergence of AI and HPC.

AS -2145GH-TNMR
2U, 4-APU, AMD Instinct MI300A Accelerators (liquid-cooled)



AS -4145GH-TNMR
4U, 4-APU, AMD Instinct MI300A Accelerators (air-cooled)



AS -8125GS-TNMR2
8U, 8-GPU, AMD Instinct MI300X Accelerators (air-cooled)

Supermicro H13 Servers Support EPYC 9004 Processors and AMD 3D Technology

The H13 AMD-based systems with the AMD EPYC™ 9004 & 8004 series processors featuring the new "Zen 4c" architecture and AMD 3D V-Cache™ Technology delivers unprecedented rack density and scalable performance with energy efficiency for a wide range of compute-intensive workloads.

H13 8U Universal GPU Server
Machine learning server with 8x Nvidia H100 GPU

H13 4U PCIe GPU Server
Maximum acceleration for AI/deep learning and HPC

H13 Flash Storage Server
Purpose-built All-Flash E3.S petascale storage solution

H13 GrandTwin Server
Leading Multi-Node architecture with front or rear I/O

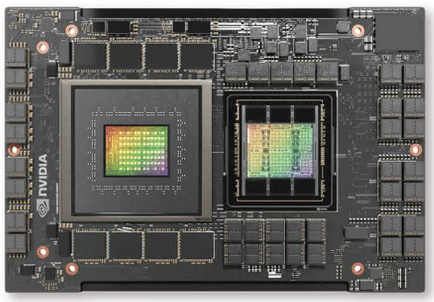
H13 Hyper Server
Industry leading IOPS server with energy efficiency and flexibility

H13 CloudDC Server
All-in-One servers with flexible I/O for cloud scale data centers

Supermicro NVIDIA MGX™ Servers

1U NVIDIA GH200 Grace Hopper™ Superchip Systems

NVIDIA MGX™ Systems enables new possibilities in system design and bleeding-edge technologies, including support for NVIDIA GH200 Grace Hopper™ Superchip which combines the power of an NVIDIA H100 GPU and NVIDIA Grace CPU on a single chip. In a mere 1- rack unit form factor, Supermicro NVIDIA MGX™ Systems can be equipped with up to 2 Grace Hopper Superchips and deliver the highest accelerated computing density in this compact form factor.



Grace Hopper Superchip

H100 GPU+ Grace CPU on one Superchip

The Grace Hopper Superchip addresses a key bottleneck in training and inference of AI models: access to plenty of high-bandwidth memory. NVLink® Chip-2-Chip (NVLink-C2C) provides a coherent CPU-GPU link that is 7x faster than PCIe 5.0.

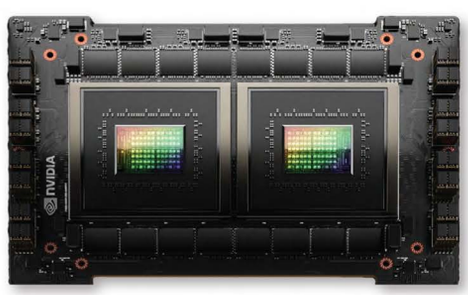
ARS-111GL-NHR
1U, NVIDIA GH200

ARS-111GL-NHR-LCC
1U, NVIDIA GH200, Liquid-Cooled

ARS-111GL-DNHR-LCC
1U 2N, NVIDIA GH200, Liquid-Cooled

1U/2U NVIDIA Grace™ CPU Superchip and x86 Intel® Xeon® Systems

NVIDIA MGX™ Systems are designed to standardize AI infrastructure and accelerated computing in 1U and 2U form factors while providing ultimate flexibility for current and future GPUs, DPUs, and CPUs. Featuring NVIDIA's new Arm-based Grace ™ CPU Superchip as well as x86 processors in the same form factors, these systems support up to 4 doublewidth GPUs such as the NVIDIA H100 and L40S to enable accelerated computing for hyperscalers, edge, HPC, and cloud.



NVIDIA Grace CPU Superchip

2x Grace CPUs on one Superchip

The NVIDIA Grace CPU Superchip uses NVLink® Chip-to-Chip (NVLink-C2C) technology to deliver 144 cores (2x 72) and 1TB/s of memory bandwidth with 480GB LPDDR5X on the integrated board. NVIDIA MGX Systems feature up to 2 Grace CPU Superchips, providing up to 288 CPU cores in a system.

ARS-121L-DNR
1U, 2-Node, NVIDIA Grace CPU

ARS-221GL-NR
2U, NVIDIA Grace CPU

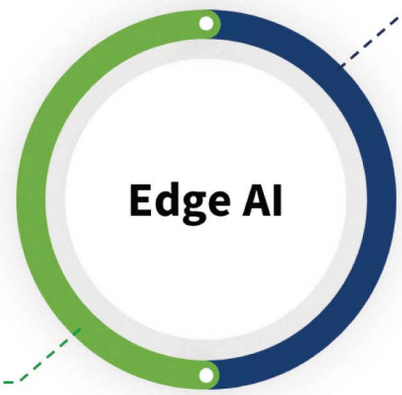
SYS-221GE-NR
2U, x86 Intel® Xeon® CPU

Supermicro Edge AI Systems

Accelerating Adoption of Edge AI

The edge refers to computational processes occurring close to the data sources at these locations – be they sensors, devices, or end-users – rather than relying on centralized systems that are common in traditional AI processing. Supermicro delivers Edge AI solutions that streamline deployment through pre-integrated components with optimized hardware and software for real-time processing.

The Two Halves of Edge AI



GENERATIVE

Focuses on capturing patterns and trends in historical data and creating **new data** based on that

PREDICTIVE

Focuses on understanding historical data and making **predictions** based on that understanding

Supermicro Edge Platforms for Predictive + Generative AI

Small	Medium	Large	Extra Large
			
SYS-E300-13AD Single A2 or T100	SYS-110D-20C-FRAN8TP Single L4 or RTX 4000 SFF Ada	SYS-E403-13E-FRN2T Single L40S, L40, or RTX 6000 Ada	SYS-221HE-TNR(D) Multiple L40S, L40, or RTX 6000 Ada
Key Features for Predictive			
AI computer vision for up to 8 streams	AI computer vision for up to 16 streams	AI computer vision for up to 32 streams	AI computer vision for up to 48 streams per GPU
Up to ~2000 automatic speech recognition (ASR) samples per second	Up to ~6000 automatic speech recognition (ASR) samples per second	Up to ~22,000 automatic speech recognition (ASR) samples per second	Up to ~32,000 automatic speech recognition (ASR) samples per second per GPU
Key Features for Generactive			
LLM up to 8 billion parameters	LLM up to 24 billion parameters	LLM up to 48 billion parameters	LLM up to 80 billion parameters
Image and video generation with Stable Diffusion at ~1 image every 4-5 seconds	Image and video generation with Stable Diffusion at ~1 image every 2 seconds	Image and video generation with Stable Diffusion at ~1-2 images per second	Image and video generation with Stable Diffusion at ~3 images per second
Dimensions (mm) H × W × D			
43.0 × 264.8 × 225.8	43.0 × 437.0 × 399.0	117.3 × 266.7 × 406.4	88.9 × 436.9 × 574.0