Supermicro Solutions Featuring Arm AGI CPUs

Optimized architectures unlocking breakthrough performance, efficiency, and rack density for 24/7 agentic AI

A New Class of Compute Solutions for AI Orchestration

- New 2U, 5U, and multi-node architectures with liquid and air cooling options
- Rack-scale designs to handle the increasing demands of modern agentic AI
- Deploy faster at scale with Supermicro's complete DCBBS modular infrastructure
- Based on power-efficient, core-dense Arm AGI CPUs with up to 136 cores per CPU and 6GB/s memory bandwidth per core

Designed for AI at Scale

Supermicro solutions featuring AI-centric Arm AGI CPUs are built for the age of massive-scale agentic AI orchestration, delivering performance, efficiency, and density that maximizes the economics of rack-scale deployments. Supermicro's proven first-to-market leadership combines with large-scale DCBBS deployment capabilities and in-house developed thermal management technologies to create a new class of architecture optimized for modern agentic AI.

DCBBS for Complete Agentic AI Solutions

Supermicro's Data Center Building Block Solutions (DCBBS) are the blueprint for modern data centers, providing complete, modular AI infrastructure built from validated components and sub-systems for end-to-end deployment flexibility. By leveraging Arm AGI CPUs, Supermicro solutions can deliver over 2x performance per rack compared to traditional architectures and help enterprises save up to $10 billion in CAPEX per Gigawatt of AI data center capacity. Building on Supermicro's industry-leading rack density and performance-per-watt, these solutions help ensure maximum utilization of data center space and power resources.

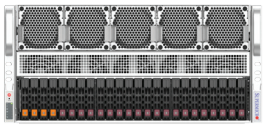
Supermicro Deploys AI Infrastructure Faster

These new and innovative, workload-specific architectures are built to maximize the performance and density of Arm AGI CPUs while also ensuring power and thermal efficiency, lowering TCO and TCE (total cost to the environment). Supermicro's modular building block philosophy, proven in-house liquid cooling technology, and rack integration expertise mean this new range of architectures can be brought to market faster, so that organizations can begin to realize returns on their AI infrastructure investments sooner. Backed by a global manufacturing footprint of up to 6,000 racks per month, Supermicro is poised to deliver fully integrated AI clusters at any scale that are integrated, tested, and validated in-house before shipping, ready to power on from day one.

Compute for a New Age of Agentic AI

Leveraging the high core density and performance-per-watt of Arm AGI CPUs, Supermicro systems deliver up to 2× performance per rack and 2× core density versus traditional platforms within the same power envelope, yielding significant savings in both power consumption and floor space. The 136-core Arm AGI microarchitecture minimizes legacy overhead to deliver higher work per cycle and sustained, unthrottled performance. Additional features include 6GB/s memory bandwidth per core with latency-optimized access for linear scaling, plus expanded memory capacity and flexible I/O to support energy-efficient, distributed agentic AI infrastructure where CPUs orchestrate thousands of parallel tasks.

Rack-Scale Solutions	2U 4-Node ORV3 Rack-Scale Solution Ultra-density multi-node rack-scale liquid cooled architecture	1U 4-Node ORW Rack-Scale Solution Maximum compute density multi-node liquid cooled ORW rack architecture
Key Applications	Large-scale AI Inference, Cloud Compute, Hyperscale	Large-scale AI inference, Agentic AI Orchestration, Cloud Compute, Hyperscale CPU Services, High-density Scale-out Workloads
Rack Configuration	44-0U/48-0U OCP ORV3 Up to 76 nodes per rack	44-0U/48-0U OCP ORW Up to 168 nodes per rack
CPUs/Cores per Rack	Up to 152 CPUs/20,672 cores	Up to 336 CPUs/45,696 cores
Cooling Solution	CPU - Direct-to-Chip In-row or in-rack CDU	CPU - Direct-to-Chip In-row or in-rack CDU
Power Budget per Rack	Max 95kW	Max 210kW



High-Density Rackmounts	5U 8-GPU Flagship Rackmount for a New Generation of Performance and Efficiency	2U Hyper Thermally-optimized Architecture with Flexible GPU Support	2U Hyper-E Single-socket Front I/O Architecture Optimized for Edge Deployments
Model	ARS-522GP-NR	ARS-222H-NR	ARS-212HE-FNR
Key Applications	GPU-accelerated Inference, AI Training, Model Development, HPC, Simulation, Rendering	AI Orchestration, CPU-driven Inference, Memory-intensive Workloads, Cloud Compute, Virtualization, Flexible GPU-accelerated Inference	AI Inference, Edge Computing, Cloud Computing, CDN
CPU	Dual-socket Arm AGI CPU Neoverse V3 with 64/128/136 cores per CPU	Dual-socket Arm AGI CPU Neoverse V3 with 64/128/136 cores per CPU	Single-socket Arm AGI CPU Neoverse V3 with 64/128/136 cores per CPU
Memory	24 DIMM slots; up to 6TB DDR5-8800MT/s in 1 DPC	24 DIMM slots; up to 6TB DDR5-8800MT/s in 1 DPC	12 DIMM slots; up to 6TB DDR5-8800MT/s in 1 DPC
PCIe Expansion	8 PCIe 5.0 x16 FHFL double-width slots 4 PCIe 5.0 x16 FHHL slots 1 PCIe 6.0 x16 FHFL slot 1 PCIe 5.0 x8 AIOM slot (OCP 3.0 compatible)	5 PCIe 6.0 x16 FHHL slots 1 PCIe 6.0 x8 FHHL slot 1 PCIe 6.0 x8 AIOM slot (OCP 3.0 compatible)	Default 3 PCIe 6.0 x16 FHHL slots 1 PCIe 6.0 x8 FHHL slot 1 PCIe 6.0 x16 AIOM slot (OCP 3.0 compatible) Option A 4 PCIe 6.0 x16 FHHL slots 1 PCIe 6.0 x16 AIOM slot (OCP 3.0 compatible) Option B 1 PCIe 6.0 x16 FHHL slot 4 PCIe 6.0 x8 FHHL slots 1 PCIe 6.0 x16 AIOM slot (OCP 3.0 compatible)
GPU Support	Up to 8 double-width	Up to 2 double-width	Up to 2 double-width
Storage	8 front hot-swap 2.5" NVMe drive bays	8 front hot-swap 2.5" NVMe drive bays	Default 6 front hot-swap E1.S NVMe drive bays Option A 4 front hot-swap E1.S NVMe drive bays Option B 6 rear hot-swap 2.5" NVMe drive bays
Power	6x 2700W Redundant (3 + 3) Titanium Level (96%) Hot-plug power supplies	2x 2700W Redundant (1 + 1) Titanium Level (96%) Hot-plug power supplies	2x 2000W Redundant (1 + 1) Titanium Level (96%) power supplies