PRODUCT BRIEF

SUPERMICRO'S ARM-BASED RACK-SCALE SOLUTIONS OFFER HIGH-END PERFORMANCE TO SUPPORT THE RAPID GROWTH OF AGENTIC AI WORKLOADS

New Systems Address the Growing Need for Energy-Efficient Servers for Token-Intensive Always-on Agents

Supermicro Rack-Scale Solution with Arm AGI CPUs

Executive

TABLE OF CONTENTS

Summary

The rapid proliferation of AI agents is creating a need for energy-efficient systems that deliver the required results within the expected timeframe. While AI training has dominated recent discussions, the rising demand for inference and agentic workloads calls for energy-optimized systems that operate across diverse environments. The rise of agentic AI will require CPUs capable of orchestrating and managing continuously running agents. This inflection point will require significantly

more CPUs at rack-scale density, while keeping energy consumption within the power envelopes of current enterprise data centers.

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, combined with large-scale Data Center Building Block Solutions® (DCBBS) deployment capabilities and in-house-developed thermal management technologies, creates a new class of architecture optimized for modern agentic AI.

Target Workloads

Agentic workloads, which are systems that can act on behalf of the user, encompass a wide range of actions. These include, but are not limited to, customer support automation, computer coding, and supply chain management. Other "agents" may include, across the enterprise, human resources and onboarding activities, as well as financial workloads. An innovative agentic system will also learn over time, fix errors, and communicate with other agents.

Supermicro Systems

New systems from Supermicro incorporate the latest Arm technology, featuring the Arm AGI CPU in two systems, custom-designed for agentic AI. These systems both include:

- **Performance:** Up to 136 Arm Neoverse® V3 cores per CPU, delivering leading performance per core, SoC, blade, and rack, with 6 GB/s memory bandwidth per core at sub-100ns latency. Supermicro systems are available in dual-socket configurations.

- **Scale:** Base TDP of 300 watts with a dedicated core per program thread enables deterministic performance under sustained load, eliminating throttling and idle threads.

- **Efficiency:** Supports high-density 2U server air-cooled chassis that support air-cooled deployments with up to 6528 cores per rack, and liquid-cooled systems delivering 26,111 cores per rack.

The new Supermicro hyper system is designed for environments that require very high-density computing for AI inference and agentic AI. This system can be ordered as a single system or in a rack-scale configuration. Specifically, the system components:

System SKU: ARS-222H-NR - General compute including Agentic AI & Edge Computing, Cloud, and Memory-Intensive Workloads

- 2U Height
- # of CPUs: 2
- Arm AGI CPU Neoverse V3 with 64, 128, or 136 cores
- Memory:
  - Slot Count: 24 DIMM slots
  - Max Memory (1DPC): Up to: 6TB ECC DDR5-8800 MT/s RDIMM
- One OCP 3.0 Compatible AIOM
- Up to 8 front hot-swap 2.5” NVMe bays
- Redundant 2700W Titanium Level power supplies
- Up to 2 GPUs

System SKU: ARS-522GP-NR - Agentic AI Inferencing and AI Training

- 5U Height
- # of CPUs: 2
- Arm AGI CPU Neoverse V3 with 64, 128, or 136 cores
- Memory:
  - Slot Count: 24 DIMM slots
  - Max Memory (1DPC): Up to: 6TB ECC DDR5-8800 MT/s RDIMM
- One OCP 3.0 Compatible AIOM
- Up to 8 front hot-swap 2.5” NVMe bays
- Up to 6x N+N redundant 2700W AC hot-swappable power supplies
- Up to 8 Double-Width GPUs

System SKU: ARS-212HE-FNR

- 2U Height
- # of CPUs: 1
- Arm AGI CPU Neoverse V3 with 64, 128, or 136 cores
- Memory:
  - Slot Count: 12 DIMM Slots
  - Max Memory (1DPC): Up to: 3TB ECC DDR5-8800 MT/s RDIMMs
- One OCP 3.0 Compatible AIOM
- Up to 6 front hot-swap NVMe bays
- 1+1 Redundant 2000W Titanium Level power supplies
- Up to 2 GPUs

System SKU: ARS-242TP-QNR-LCC

- 2-OU Height
- 4 Independent nodes
- # of CPUs per node: 2
- Arm AGI CPU Neoverse V3 with 64, 128, or 136 cores
- Memory Per Node:
    - Slot Count: 24 DIMM Slots
    - Max Memory (1DPC): Up to: 6TB ECC DDR5-8800 MT/s RDIMMs
- Two OCP 3.0 Compatible AIOMs per node
- Two front hot-swap NVMe bays per node
- Liquid Cooled

System SKU: ARS-142TP-QNR-LCC

- 1OU Height
- # of CPUs per node: 2
- Arm AGI CPU Neoverse V3 with 64, 128, or 136 cores
- Memory Per Node:
    - Slot Count: 24 DIMM Slots
    - Max Memory (1DPC): Up to: 6TB ECC DDR5-8800 MT/s RDIMMs
- OCP Compatible AIOM
- Two front hot-swap NVMe bays per node
- Open Rack DC -48V with PDB to each node
- Liquid Cooled

Rack Scale Configurations

While the performance of a single server with the Arm AGI CPU is impressive, at rack scale, it is truly amazing. Depending on the rack configuration, over 26,000 Neoverse V3 cores are available to customers in a single rack.

Rack Configuration 1: (48U rack, external networking)

- 24 X 2U servers
- 48 Arm AGI CPUs
- 6,528 cores per rack

Rack Configuration 2: (48U rack, external networking)

- 9 X 5U servers
- 18 Arm AGI CPUs
- 2,448 cores per rack
- Up to 72 double-width GPUs

Rack Configuration 3: (48U rack, external networking)

- 16x 2U 4N (nodes)
- Dual Arm AGI CPUs per node, 128 Arm AGI CPUs total
- 26, 112 cores per rack
- Liquid cooling

Rack Configuration 4: (ORW - 48U rack, external networking)

- 168 Servers per rack
- 336 Arm AGI CPUs per rack (42x 1U 4N))
- 45,696 cores per rack
- Liquid cooling

Arm AGI CPU

The Arm AGI CPU is manufactured using an advanced 3nm process node. The processor supports up to 136 Neoverse V3 cores, 12-channel DDR5-8800 memory, and 96 lanes of PCIe Gen 6 and CXL 3.0. The max CPU frequency is up to 3.5 GHz for the 136C SKU and 3.7 GHz for the 64C SKU, with 2 MB of dedicated L2 cache per core. Dual 128-bit SVE2 (Scalable Vector Extension 2) units per core, which support bfloat16 and INT8 MMLA instructions.

The evolution of AI and the rise of agentic AI are changing the way organizations deploy their AI infrastructure. The inherent limitations of data center power supplies, the rising cost of powering large-scale AI data centers, and the physical space required

to house growing AI clusters necessitate a balance among raw compute power, compute density, and efficiency. This balance becomes ever more important as AI workloads scale.

Supermicro's system portfolio spans a range of form factors from 1U compute servers to rack-scale AI clusters. By leveraging the enhanced core density and performance-per-watt of new Arm AGI CPUs, Supermicro can provide up to 2x performance per rack and 2x core density compared to a traditional rack with the same power envelope. This can result in significant data center cost savings in terms of both ongoing power consumption and physical floor space.

- The dense 136-core microarchitecture of Arm's AGI CPUs is purpose-built for performance, minimizing legacy overhead and completing more work per cycle for sustained, unthrottled performance.
- Low per-core base TDP combined with Supermicro's industry-leading air and liquid cooling thermal management facilitates energy efficiency and system density for up to 2x cores per rack compared to traditional systems.
- 6GB/s memory bandwidth per core and latency-optimized memory access to support linear scaling.
- Expanded memory capacity and flexible I/O for energy-efficient, scalable agentic AI infrastructure where CPUs orchestrate thousands of parallel tasks across distributed infrastructure.

For More Information

Supermicro Arm AGI CPUs: https://www.supermicro.com/en/solutions/arm-agi
Supermicro Arm AGI CPU-based server 2U: https://www.supermicro.com/en/products/system/hyper/2u/ars-222h-nr
Supermicro Arm AGI CPU-based server 5U: https://www.supermicro.com/en/products/system/gpu/5u/ars-522gp-nr
Supermicro Arm AGI CPU-based server, 2U, Single Socket: https://www.supermicro.com/en/products/system/hyper/2u/ars-212he-fnr
Supermicro Arm SGI CPU-based server, 2U, 4Node: https://www.supermicro.com/en/products/system/hyper/2-ou/ars-242tp-qnr-lcc

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com

ARM

Arm develops the compute platform that is powering the global AI revolution. Our high-performance, energy-efficient CPU products are trusted by the world's leading semiconductor companies and deployed in more than 350 billion chips to date—including over 99% of smartphones.

Visit www.arm.com