H13 Quad APU Systems

Powered by AMD Instinct™ MI300A Accelerators

2U AS -2145GH-TNMR (Liquid-cooled)

4U AS -4145GH-TNMR (Air-cooled)

Integrated CPU/GPU Accelerated Processing Unit (APU) Systems for HPC and AI Workloads

Targeting accelerated HPC workloads, 4U air-cooled and 2U liquid-cooled systems integrate 4 AMD Instinct™ MI300A APU accelerators

- Four APUs combine high-performance AMD CPU, GPU and HBM3 memory for a total of 912 AMD CDNA™ 3 GPU compute units and 96 "Zen 4" cores, and 512GB of unified HBM3 memory in one system.
- PCIe 5.0 expansion slots for high-speed networking including RDMA to APU memory
- 2 PCIe 5.0 Open Compute Project (OCP) 3.0 AIOM slots
- Fully optimized for the most popular AI & ML frameworks—PyTorch, TensorFlow, JAX, ONYX-RT, Triton

The challenges of generative artificial intelligence (AI), large-language models (LLMs) and high-performance computing (HPC) have put extreme pressure on the traditional separation between CPUs and GPUs in a server. For example, the need for higher speed connections between CPU and GPU memory to simplify the overhead of managing data pipelines and eliminating software refactoring for GPU acceleration have propelled the development of accelerated processing units such as the new AMD Instinct™ MI300A APU. Combined with Supermicro's expertise in multiprocessor system architecture, these quad APU systems push the limits of HPC that optimizes serial (CPU) and parallel (GPU) processing to EXASCALE level.

Quad-Socket MI300A APU-Based Systems

Leveraging a long history of building multiprocessor systems, Supermicro offers two quad-socket APU-based servers that give flexibility in cooling models. The 2U AS -2145GH-TNMR is a dense, liquid-cooled system that delivers exceptional TCO with over 51% data center energy cost savings. Furthermore, there is a 70% reduction in fans compared to air-cooled solutions.

The 4U AS -4145GH-TNMR provides more storage and 8-16 extra PCIe 5.0 acceleration cards. Each server supports two compact AIOMs and offers PCIe 5.0 x16 slots for 400G Ethernet InfiniBand networks to develop a supercomputing cluster while speeding the flow of data directly to APU memory. With 2+2 redundant 1600W Titanium-Level power supplies, these systems can accelerate your workloads at peak performance while maintaining thermal parameters.

Integration Means High Performance

Each APU combines high-performance AMD CPU, GPU and HBM3 memory for a total of 912 AMD CDNA™ 3 GPU compute units and 96 "Zen 4" cores, and 512GB of unified HBM3 memory in one system. The large, unified 128GB HBM3 memory per APU is more than three times faster than traditional DDR5 memory speeds, helping increase both CPU and GPU performance. Cache coherence is maintained through a 256 MB last-level AMD Infinity Cache™ that mediates memory requests from both

CPU and GPU cores. With the quad APU configuration, six Infinity Fabric™ links are dedicated to inter-GPU connectivity for a total of 384 GB/s of peer-to-peer bandwidth per APU.

Shorten Time to Value with the AMD ROCm Platform

Whether you are deploying HPC or AI applications, AMD ROCm™ software opens doors to new levels of freedom. With mature drivers, compilers, and optimized libraries supporting AMD Instinct accelerators, ROCm is open and ready to deploy. Proven in some of the world's largest supercomputers, ROCm software provides support for leading programming languages and frameworks for AI, including PyTorch, TensorFlow, ONYX-RT, Triton, and JAX. If HPC is your workload, you can use frameworks that help parallelize operations across multiple GPUs and solve linear systems. The AMD Infinity Hub includes platform-compatible HPC applications, including those for astrophysics, weather and climate, computational chemistry, computational fluid dynamics, earth sciences, genomics, geophysics, molecular dynamics, and physics.

Open Management

Our approach to management enables you to deliver the scale your organization requires. Supermicro® SuperCloud Composer with open-source Redfish® compliant software helps you configure, maintain, and monitor all of your systems using single-pane-of-glass management. If your DevOps teams prefer to use their own tools, our accessible Redfish-compliant APIs provide access to higher-level tools and scripting languages. More traditional management approaches, including IPMI 2.0, are available as well. Regardless of your data center needs, our open management APIs and tools are ready to support you.

H13 Generation	AS -2145GH-TNMR Server	AS -4145GH-TNMR Server
Form Factor	• 2U rackmount, liquid cooled	• 4U rackmount, air cooled
APU Support	• 4x AMD Instinct MI300A APUs in SH5 sockets	
APU Features	• Total of 96 'Zen 4' cores per server • Total of 912 compute units per server • Total of 3648 Matrix Cores per server • Total of 12 decoder groups for HEVC/H.264, AVC/H.264, V1, or AV1 (requires inclusion/installation of compatible media players • Total of 96 cores JPEG/MJPEG CODECs • Virtualization support for up to 12 partitions with SR-IOV • HPC data types: FP64 and FP32 vector, FP64 and FP32 matrix • AI/ML data types: TF32 matrix, FP16, BFLOAT16, INT8, FP8 w/sparsity support	
Memory Capacity	• 128 GB HBM3 memory (5.3 TB/s) per APU, 512 GB total per system • Last-level 256 MB AMD Infinity Cache™ shared between CPU and GPU cores	
On-Board Devices	• System on chip with integrated GPU • IPMI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics	
Expansion Slots	• 6 PCIe 5.0 x16 (or x8) slots • 2 Optional PCIe 5.0 (x8 or x16) slots • 2 OCP PCIe 5.0 x16 AIOM slots	• 8 PCIe 5.0 x16 (or x8) slots or 16 Optional PCIe 5.0 x8 slots • 2 OCP PCIe 5.0 x16 AIOM slots
Storage	• 8 optional 2.5" U.2 NVMe hot-swap drives • 2 M.2 NVMe boot drives	• 8 x2.5" U.2 NVMe hot-swap drives or 24x 2.5" SAS/SATA drives (optional) • 2 M.2 NVMe boot drives
I/O Ports	• 1 RJ45 dedicated management LAN port • 2 rear USB 3.0 ports • 1 VGA Connector • 1 COM port • 1 Display port	
BIOS	• AMI Code Base 256 Mb (32 MB) SPI EEPROM	
System Management	• Built-in server management tool (IPMI 2.0, KVM/media over LAN) with dedicated LAN port • Supermicro SuperCloud Composer • SuperDoctor® 5 • Supermicro Server Manager (SSM) • Supermicro Update Manager (SUM) • Redfish APIs	
System Cooling	• Liquid cooling for APUs • 3x heavy-duty 80mm hot-swappable fans	• 5 external and 5 internal 80mm hot-swappable fans
Power Supply	• 2+2 redundant 1600W hot-swappable Titanium-Level power supplies	

† Certain high TDP CPUs may be supported only under specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization.