PRODUCT BRIEF

SUPERMICRO AND AMD DELIVER RACK SCALE AI AND HPC SOLUTIONS WITH THE NEW AMD INSTINCT™ MI300 SERIES ACCELERATORS

Supermicro Delivers Innovative Servers Incorporating the New AMD Instinct MI300X and AMD Instinct MI300A Accelerators

TABLE OF CONTENTS

Executive Summary

AI is quickly becoming a technology that cannot be ignored as more and more enterprises and organizations take advantage of the latest acceleration technology. The AMD Instinct™ MI300A and AMD Instinct™ MI300X are designed to accelerate both AI and HPC workloads, with results showing significant leaps in performance over previous generations. Supermicro has developed application optimized servers incorporating the AMD Instinct MI300A and AMD Instinct MI300X.

Workloads Demand Higher Performance

A number of workloads constantly push the performance envelope and demand the increasing power of GPUs/Accelerators. These include researchers who are solving the most significant problems in physics, astrophysics, and material science, to name a few. Academics and governments are using the latest computer hardware and algorithms to understand and improve the

lives of many, including modeling communities and climate change to offer better outcomes for all. Enterprises require accelerators to create more optimized products, invent new vaccines, and simulate new energy sources.

Overall System Architecture

The new technology that addresses the latest workload requirements relies on both the CPUs, GPUs, and server architecture. The server that houses the latest generations of CPUs and GPUs must be designed to connect to peripherals, deal with thermal (TDP) issues, and contain security safeguards. A leadership class optimized and productive server must include the following elements:

- Compute – Workload optimized compute architecture with a wide range of data format support for different workloads.
- Memory – High memory capacity and bandwidth to deliver data to the CPUs
- Networking – Advanced network bandwidth with industry standard and customer technologies
- Software – Software ecosystem with drop-in support for leading programming models and AI frameworks.

Supermicro Servers with AMD Instinct MI300 Technology

Supermicro has a broad line of servers that address a wide range of workloads. For advanced AI and HPC workloads that contain the AMD Instinct MI300X or the AMD Instinct MI300A GPUs.

Supermicro expands its rack-scale GPU solutions with new accelerated AI and HPC optimized servers powered by AMD Instinct™ MI300 series accelerators, including additions to the universal 8-GPU family as well as new 2U and 4U 4-Way Application Processing Unit (APU) systems that combine GPU, CPU, and high-bandwidth memory (HBM3) on a single chip. Both product families are powered with AMD's MI300 series accelerators, the 8U 8GPU featuring the AMD Instinct MI300X targeted for AI workloads such as Large Language Models (LLM), generative AI training, and the 2U liquid cooled, and 4U air cooled 4-Way systems with the AMD Instinct MI300A, which is designed for high-performance computing workloads such as CFD simulations and data analytics with optimized liquid and air cooling options, unparalleled performance, and efficiency at scale.

Supermicro GPU server with AMD Instinct MI300X - AS -8125GS-TNMR2

Supermicro is extending the universal 8-GPU family with AMD Instinct™ MI300X accelerators to deliver leading edge rack scale AI infrastructure for large scale AI training applications. Dual 4th Gen AMD EPYC™ CPUs power the system with up to 256 cores and eight AMD Instinct MI300X accelerators. The 8U 8GPU system is designed with robust capabilities that optimize performance, density, and efficiency for advanced AI models supporting industry-standard OCP Accelerated Module (OAM) form factors. This system has a balanced 1:1 networking to GPU, which reduces networking bottlenecks when AI training runs require performance that a single server cannot deliver and needs to connect to other servers. Each rack can support over 1000 CPU cores and up to 24TB of high-bandwidth memory (HBM3). The Supermicro proven 8U 8GPU systems also provide best in class compute throughput on double precision FP64 for HPC, delivering 1728 TFLOPS per rack.

Key Details

- Accelerator Support: 8x AMD Instinct MI300X OAM Accelerators
- Processor Support: Dual AMD EPYC 9004 Series processors, up to 256 cores/512 threads per server
- Memory Capacity: 24x DIMM slots, Up to 6TB, ECC DDR5 with support up to 4800MHz
- 8 high-speed 400G networking cards for 1:1 paring with GPUs
- Chassis: 8U

Figure 1 - 8U 8-GPU Server with AMD Instinct MI300X Accelerators

Supermicro GPU servers with AMD Instinct MI300A Series Accelerators - AS -2145GH-TNMR and AS -4145GH-TNMR

Supermicro's 2U liquid and 4U air cooled APU servers with a 4-way integration of AMD Instinct™ MI300A accelerators combining the highest-performing AMD CPU, GPU, and HBM3 on a single chip. Each server contains 96 "Zen4" cores and 512GB of HBM3 memory with a full rack (48U) solution of 21 2U systems containing 2016 cores and over 10TB of HBM3 memory. Both systems feature dual AIOMs with 400G Ethernet support and expanded networking options. This design improves space, scalability, and efficiency for high-performance computing. The 2U direct-to-chip liquid-cooled system also delivers excellent TCO with over 35% energy consumption savings based on 21 2U systems in a rack solution producing 61,780 watts per rack over 95,256 watts air cooled rack, as well as a 70% reduction in fans compared to the air cooled system.

Supermicro GPU server with AMD Instinct MI300A Accelerator - AS -2145GH-TNMR

- CPU/GPU: Quad AMD Instinct MI300A accelerators
- 96 "Zen 4" cores
- Memory Capacity: 128GB HBM3 (per APU), Total Per Server: 512GB HBM3
- Dual AIOM, supporting 400 G
- Liquid Cooling Only enables up to 760W TDP per APU
- Chassis: 2U, for maximum density

Figure 2 - AS -2145GH-TNMR Rear View

Figure 3 - AS -2145GH-TNMR Front View

Supermicro GPU server for AMD MI300A - AS -4145GH-TNMR

- CPU/GPU: Quad AMD Instinct MI300A accelerators
- 96 "Zen 4" cores
- Memory Capacity: 128GB HBM3 (per APU), Total Per Server: 512GB HBM3
- 8 NVMe or 24 SAS/SATA storage devices
- Dual AIOM supporting 400G networking
- Air Cooled
- Chassis: 4U

Figure 4 - AS -4145GH-TNMR Front View

Summary of Supermicro GPU Servers with AMD Instinct MI300 Accelerators

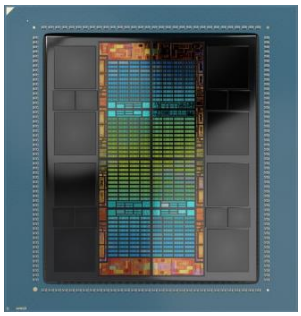
Configuration	AS -8125GS-TNMR2 (8U)	AS -2145GH-TNMR (2U)	AS -4145GH-TNMR (4U)
Form Factor	8U 8-GPU System with AMD Instinct MI300X Accelerators (air-cooled)	2U 4-GPU System with AMD Instinct MI300A Accelerators (liquid-cooled)	4U 4-GPU System with AMD Instinct MI300A Accelerators (air-cooled)
CPU	Dual AMD EPYC 9004 Series Processors with up to 128 cores/256 threads per socket	Quad AMD Instinct MI300A APUs with total of 96 CPU cores (4x 24 AMD "Zen4" cores)	Quad AMD Instinct MI300A APUs with total of 96 CPU cores (4x 24 AMD "Zen4" cores)
GPU	8x AMD Instinct MI300X Accelerators with 192GB HBM3 memory per GPU interconnected on AMD Universal Base Board (UBB 2.0)	4x AMD Instinct MI300A with 228 AMD CDNA 3 GPU compute units per APU	4x AMD Instinct MI300A with 228 AMD CDNA 3 GPU compute units per APU
Memory	Up to 6TB (24x 256GB DRAM) 4800MT/s ECC DDR5 RDIMM/LRDIMM	512 GB unified HBM3 memory with up to 4.3 TB/s bandwidth	512 GB unified HBM3 memory with up to 4.3 TB/s bandwidth
Drives	16x hot-swap PCIe 5.0 U.2 NVMe, 1x onboard M.2 NVMe, 2x 2.5" SATA	2x onboard 2280 or 22110 M.2 NVMe and 8x hot-swap 2.5" U.2 NVMe	2x onboard 2280 or 22110 M.2 NVMe and 8x hot-swap 2.5" U.2 NVMe or 24x 2.5" SAS/SATA
Networking	8x PCIe 5.0 high-performance networking cards, up to 400G with Ethernet or InfiniBand	2x AIOM (OCP 3.0) with up to 400G and additional 4x PCIe 5.0 (x8) slots	2x AIOM (OCP 3.0) with up to 400G and additional 4x PCIe 5.0 (x8) slots
Interconnect	AMD Infinity Fabric™ Links (7x 128GB/s)	AMD Infinity Fabric™ Links (4x 128GB/s)	AMD Infinity Fabric™ Links (4x 128GB/s)
Power	6x or 8x 3000W redundant Titanium Level power supplies	4x 1600W redundant Titanium Level power supplies	4x 1600W redundant Titanium Level power supplies
Cooling	Air Cooling	Liquid Cooling	Air Cooling

Rack Scale Solutions

Supermicro, with its worldwide manufacturing facilities, is able to deliver thousands of racks per month. Supermicro designs complete solutions with servers, switches, and liquid cooling as customers need as part of a total solution delivery.

AMD Instinct MI300 Series Accelerator Details

MI300X

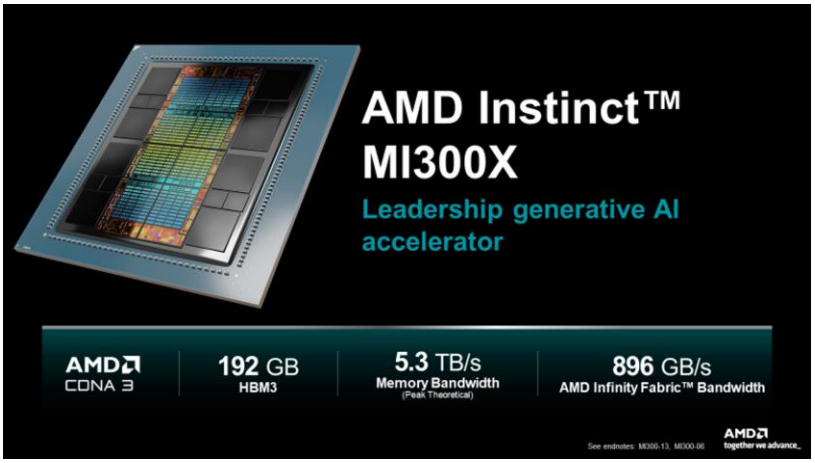


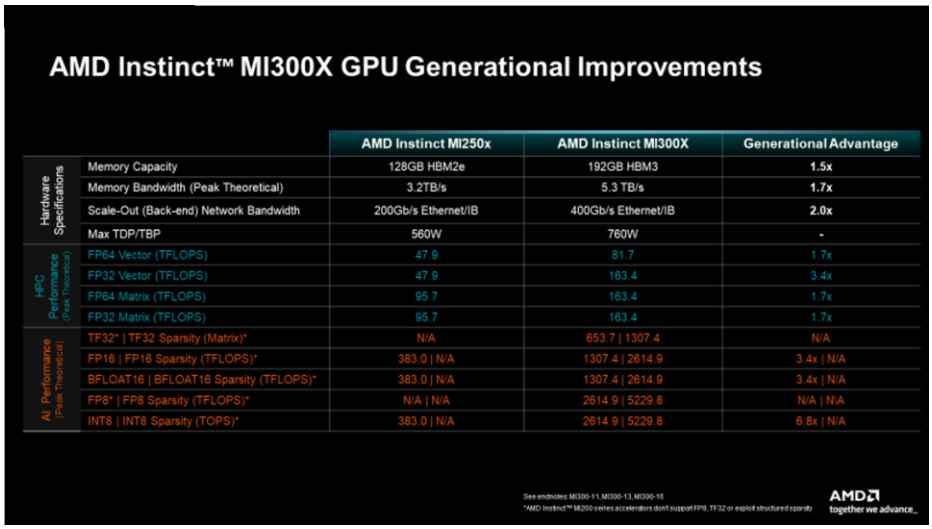
- Discrete GPU (needs a host CPU)
- 304 GPU compute units for accelerated matrix-multiply computation
- 192GB HBM3, Bandwidth – 5.3TB/s
- Ideal for AI large model training/inference, High Performance Computing (HPC)
- Eight fully meshed OAMs with Instinct™ Platform
- OAM-UBB design
- Data types include FP16, BF16, INT8 and FP8

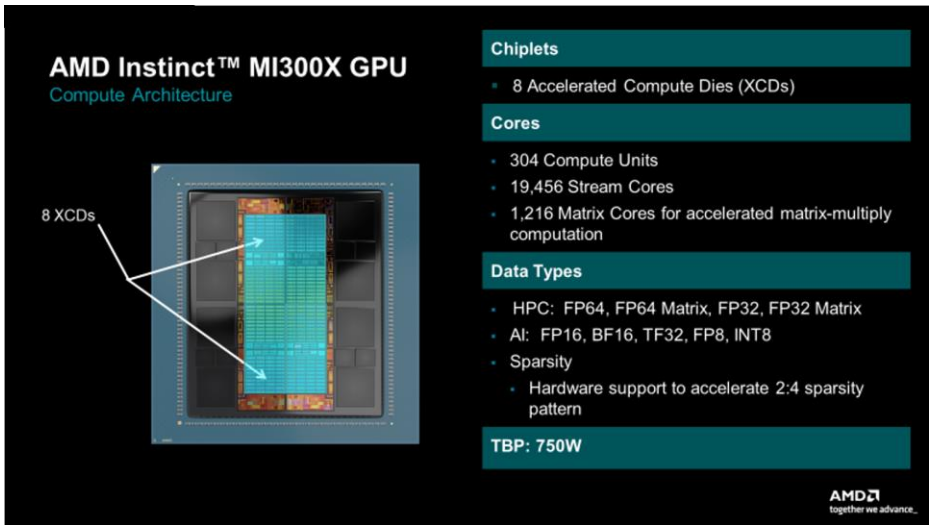
8x MI300X on a Universal BaseBoard

Summary of AMD Instinct MI300X Accelerator

Below is the relative performance per watt of INT8, FP16, and Bfloat16 over the previous generations. These numerical representations are commonly used in AI and ML applications.



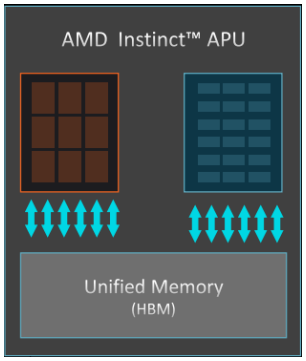


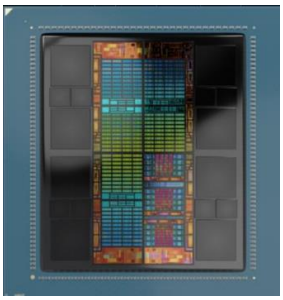


MI300A

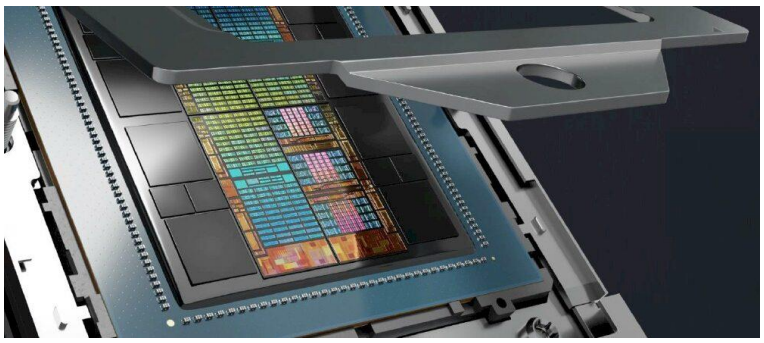
- Accelerated Processing Unit (APU)
- 228 CDNA compute units
- 24 AMD EPYC ZEN 4 CPU cores
- 128GB HBM3, Bandwidth @ 5.2 TB/s
- Ideal for high end HPC and AI
- Unified memory across CPU & GPU (256MB) Eliminates memory copies
- APU-APU coherent interconnect
- TDP: 760W

- Data types supported (Different data types for different applications)
    - HPC: FP64, FP32
    - AI: FP16, BF16, TF32, FP8, INT8
    - Sparsity support for matrix operations
    - Significant performance improvement in HPC and AI workloads over previous AMD MI generations. (See AMD for more details)



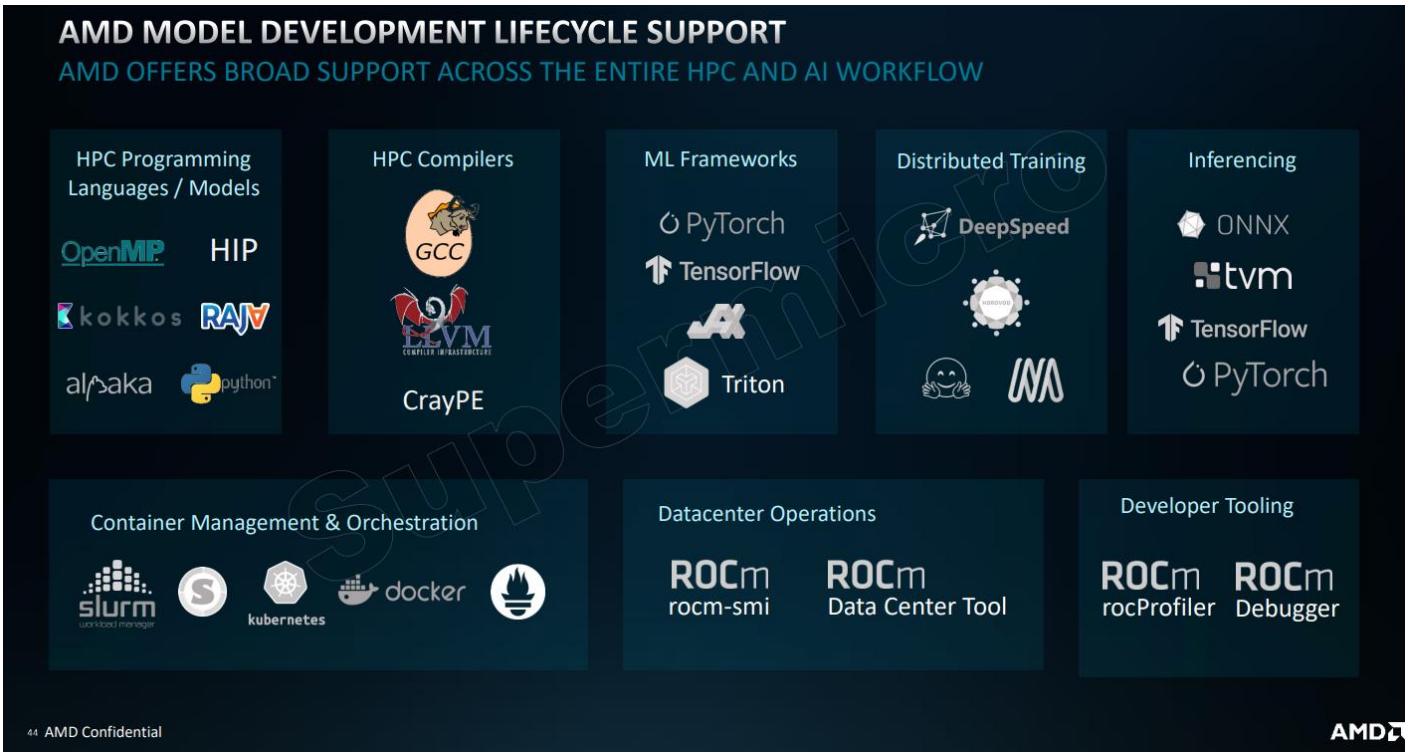


The MI300A shows significant improvement over the previous generation of MI250. (Chart courtesy of AMD)



AMD Software

A software environment is critical to the productivity of developers and users and to take advantage of the underlying hardware.



Summary

The new 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 eight 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 reduce the system's complexity and offer developers and customers an innovative server with tremendous performance and significantly lower power consumption.

For More Information:

Supermicro AMD Accelerated Systems – https://www.supermicro.com/accelerators/amd

Supermicro H13 Aplus Servers – www.supermicro.com/aplus

8U 8GPU Server - https://www.supermicro.com/en/products/system/gpu/8u/as-8125gs-tnmr2

2U 4-way Server – Liquid Cooled https://www.supermicro.com/en/products/system/gpu/2u/as-2145GH-tnmr

4U 4-way Server – Air Cooled https://www.supermicro.com/en/products/system/gpu/4u/as-4145gh-tnmr