H13 GPU Solutions w/ AMD Instinct™ MI300 Series Accelerators

Elevator Pitch

Three powerful new systems build upon Supermicro's Building Block and Multiprocessor architectures with platform-specific enhancements to scale performance with the MI300-Series and optimize for the differentiated needs of AI and HPC.

Augment AI infrastructure with Supermicro's new 8U 8-GPU system with MI300X: it's equipped with 1.5TB HBM3 onboard and supports 8 400G NICs to extend handling AI models of any size while providing up to 3.4x FP32 and FP16 performance versus MI250X[1].

Empower advancement in HPC with Supermicro's 4U and liquid-cooled 2U 4-Way system equipped with quad MI300As: APUs with on-die integration of a Zen 4 CPU, a MI300-level GPU, and HBM3 memory for strength in both general-purpose and accelerated workloads.

All provide industry-leading performance in both single-precision (FP32) and double-precision (FP64).

4 Key Facts:

1

Capacity + Scalability

Start with a large pool of onboard HBM3, with up to 192GB per GPU, and scale to massive clusters with 400G networking support.

2

Speed + Precision

Up to 3.4x performance in half-precision (FP16) compared to MI250X[1], and industry-leading performance in single/double precision operations.

3

Density + Efficiency

Improve TCO and maximize data center real estate with the power of 8 MI300s in 8U or 4 MI300s in 2U (liquid-cooled).

4

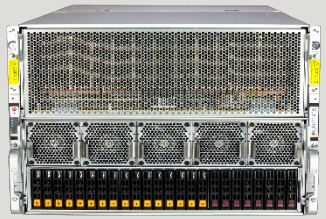
Systems with Purpose

8U system for Generative AI, 4U system for HPC/data science and liquid-cooled 2U for supercomputing clusters.

[1]Theoretical performance estimate comparing MI300X to MI250X

Purpose-Built Solutions with MI300X & MI300A

AI



8U GPU w/ MI300X
AS -8125GS-TNMR2

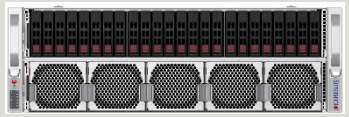
Supermicro Value

- Proven 8U high-performance fabric 8-GPU on open standard OAM form factor in 1 node
- Dual AMD EPYC™ 9004 processor
- Supports dedicated IO and storage per GPU, full performance GPU, CPU, Memory and high-speed networking
- 8 high-speed 400G networking cards for 1:1 pairing with GPU

MI300X Value

- Substantial performance gains for applications that use FP32 or higher precision data types compared to other options
- Provides the highest pool of high-bandwidth memory in a system

HPC



4U 4-Way w/ MI300A
AS -4145GH-TNMR

Supermicro Value

- Quad AMD® MI300A Processors (96 zen4 core, 912 GPU cores, 512GB HBM3) in 1 node
- Advanced air-cooled design with more storage (8 NVMe or 24 SAS/SATA) and more networking expansion.
- Dual AIOM supporting 400G networking
- Enterprise Toolless Design

MI300A Value

- Performance and air-cooled flexibility for converged HPC-AI
- Quad Zen 4 CPUs for strong performance across non-parallelizable workloads
- Highly efficient but still delivers 75% accelerated performance vs. MI300X.

Liquid-Cooled HPC

2U 4-way LC w/ MI300A
AS -2145GH-TNMR

Supermicro Value

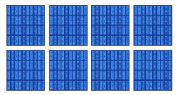
- Quad AMD® MI300A Processors (96 zen4 core, 912 GPU cores, 512GB HBM3) in 1 node
- Liquid cooling enables 760W TDP per APU in 2U, maximizing density
- Up to 51% datacenter-level energy cost savings and 70% reduction in fans.
- Dual AIOM supporting 400G
- Enterprise Toolless Design

MI300A Value

- Strong in both accelerated performance and general-purpose performance
- Highly-suitable for FP64 operations.
- On-package integration enables high-density clusters with 4 APUs for every 2 rack units.

Battle Card: Supermicro H13 with AMD INSTINCT™ MI300 Series Accelerators

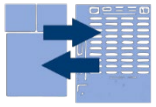
Gen-over-gen GPU Improvements[1]



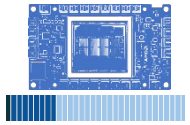
304 Compute Units
+38%

19,456 Stream Processors
+38%

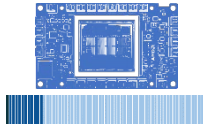
128 GB/s per GPU Infinity Fabric Link
+28%



5.3 TB/s Max Memory Bandwidth
+62%

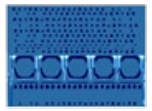


163 TFLOPS per GPU FP32 Performance
+240%

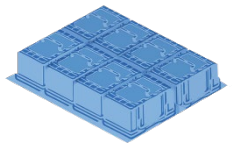


82 TFLOPS per GPU FP64 Performance
+70%

Supermicro MI300X System Benefits

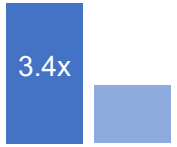


Proven Architecture for Rack Scale

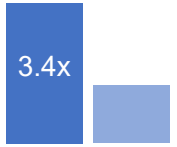


8 OAM GPUs in 1 node, 1.5TB HBM3

8x 400G Networking Support



FP16 per GPU Performance vs MI250X

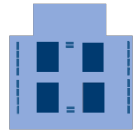


FP32 per GPU Performance vs MI250X

FP64 per GPU Performance vs MI250X

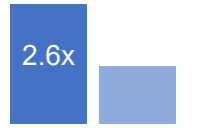
Supermicro MI300A System Benefits

2U Liquid-Cooled and 4U Air-Cooled Options

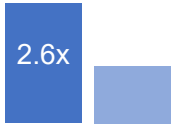


4 APUs in 1 node: CPU, GPU, 512GB total HBM3

Dual AIOM Networking



FP16 per GPU Performance vs MI250X



FP32 per GPU Performance vs MI250X

FP64 per GPU Performance vs MI250X

GPU/APU Key Information

Product name: AMD Instinct™ MI300 Series, comprising the MI300A and MI300X accelerators
Supermicro MI300X System: AS -8125GS-TNMR2
MI300A Systems: AS -2145GH-TNMR, AS -4145GH-TNMR
GPU Architecture: AMD CDNA™ 3
Announcement: 12/6/23

Process node: 5nm FinFET
Onboard Memory (per GPU): 192GB HBM3 (MI300X), 128GB HBM3 (MI300A)
Memory bandwidth: 5.2TB/s (MI300X), 5.3TB/s (MI300A)
Compute Units: 304 (MI300X), 228 (MI300A)

Stream Processors: 19,456 (MI300X), 14,592 (MI300A)
Max TDP: 750W (MI300X), 760W (MI300A)
AMD Infinity Fabric™ Links: Up to 8 GPUs with 8 links (MI300X). Up to 4 GPUs with 6 links (MI300A).
Link speed: 128 GB/s bidirectional bandwidth per GPU

[1]MI300X versus MI250X unless otherwise noted – Theoretical performance estimates based on AMD's claims.