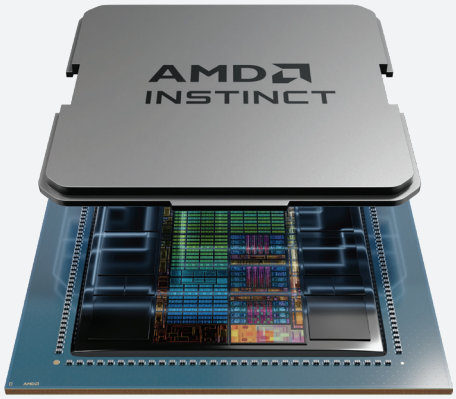
DATA SHEET

AMD INSTINCT™ MI300A APU

Integrated CPU/GPU accelerated processing unit for high-performance computing, generative AI, and ML training

Breakthrough discrete APU for HPC and AI

Based on next-generation AMD CDNA™ 3 architecture, the AMD Instinct™ MI300A accelerated processing unit (APU) is designed to deliver outstanding efficiency and performance for the most-demanding HPC and AI applications. The APU is built from the ground up to overcome the challenges that discrete GPUs present: performance bottlenecks from the narrow interfaces between CPU and GPU, burdensome programming overhead for managing data, and the need to refactor and recompile code for every GPU generation. The AMD Instinct MI300A integrates 24 AMD 'Zen 4' x86 CPU cores with 228 AMD CDNA™ 3 high-throughput GPU compute units, 128 GB of unified HBM3 memory that presents a single shared address space to CPU and GPU, all of which are interconnected into the coherent 4th Gen AMD Infinity architecture. Slated for next-generation supercomputers, this technology is available to enterprise data centers through platforms offered by our solution partners.



HPC PEAK THEORETICAL PERFORMANCE (TFLOPS)

FP64 vector	61.3
FP32 vector	122.6
FP64 matrix	122.6
FP32 matrix	122.6

AI PEAK THEORETICAL PERFORMANCE

		with sparsity
TF32 matrix (TFLOPs)	490.3	980.6
FP16 (TFLOPs)	980.6	1961.2
BFLOAT16 (TFLOPs)	980.6	1961.2
INT8 (TOPS)	1961.2	3922.3
FP8 (TFLOPS)	1961.2	3922.3

DECODERS AND VIRTUALIZATION

Decoders†	3 groups for HEVC/H.265, AVC/H.264, V1, or AV1
JPEG/MJPEG CODEC	24 cores, 8 cores per group
Virtualization support	SR-IOV, up to 3 partitions

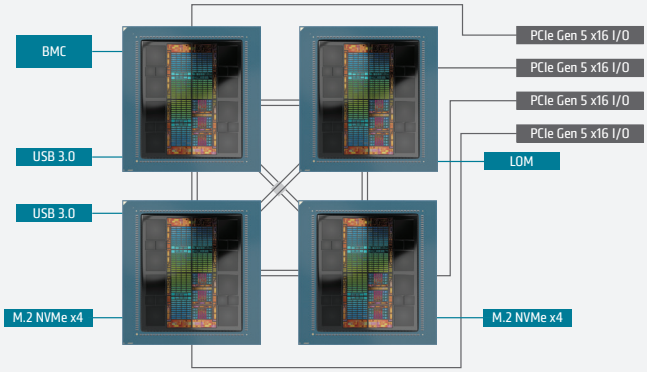
†Video codec acceleration (including at least the HEVC (H.265), H.264, VP9, and AV1 codecs) is subject to and not operable without inclusion/installation of compatible media players. GD-176

SPECIFICATIONS

Form factor	APU SH5 socket
Lithography	5nm FinFET
Active interposer dies (AIDs)	6nm FinFET
AMD 'Zen 4' x86 CPU cores	24
GPU compute units	228
Matrix cores	912
Stream processors	14,592
Peak engine clock	2100 MHz
Memory capacity	128 GB HBM3
Memory bandwidth	5.3 TB/s max. peak theoretical
Memory interface	8192 bits
AMD Infinity Cache™ (last level)	256 MB
Memory clock	5.2 GT/s
Scale-up Infinity Fabric™ Links	4 x16 (128 GB/s)
Scale-out assignable PCIe® Gen 5 or Infinity Fabric Links	4 x16 (128 GB/s)
Scale-out network bandwidth	400 Gbps Ethernet or InfiniBand™
RAS features	Full-chip ECC memory, page retirement, page avoidance
Maximum TDP	550W (air & liquid cooling) 760W (liquid cooling)

Converged Computing and Acceleration

The AMD Instinct MI300A is built to accelerate the convergence of HPC and AI applications at scale. To meet the increasing demands of AI applications, the APU is optimized for widely used data types including FP64, FP32, FP16, BF16, TF32, FP8, and INT8, including native hardware sparsity support for efficiently gathering data from sparse matrices. This helps save power and compute cycles while helping reduce memory use. By integrating 'Zen 4' CPU cores and GPU accelerators, you can achieve high efficiency by eliminating time-consuming data copy operations, transparently managing CPU and GPU caches, offloading tasks easily between GPU and CPU, and efficient synchronization, all supported by the AMD ROCm™ 6 open software platform. Virtualized environments can be supported through SR-IOV to share resources with up to three partitions per APU.



Example server architecture with four interconnected APUs

Multi-Chip Architecture

The APU uses state-of-the-art die stacking and chiplet technology in a multi-chip architecture that enables dense compute and high-bandwidth memory integration. This helps reduce data-movement overhead while enhancing power efficiency. Each device includes:

- Twenty-four x86-architecture 'Zen 4' cores in three chiplets
- Six accelerated compute dies (XCDs) with 38 compute units (CUs), each with 32 KB of L1 cache, 4 MB L2 cache shared across CUs, and 256 MB AMD Infinity Cache™ shared between XCDs and CPUs
- 128 GB of HBM3 memory shared coherently between CPUs and GPUs with 5.3 TB/s on-package peak throughput
- Three decoders for HEVC/H.265, AVC/H.264, V1, or AV1, each with an additional 8-core JPEG/MPEG CODEC
- SR-IOV for up to 3 partitions, each with 24 GB HBM3 memory

Designed for Multi-APU Architectures

Each APU provides 1 TB/s of bidirectional connectivity through eight 128 GB/s AMD Infinity Fabric™ interfaces. Four interfaces are dedicated Infinity Fabric links, while four can be flexibly assigned to deliver either Infinity Fabric or PCIe Gen 5 connectivity.

In a typical 4-APU configuration, six interfaces are dedicated to inter-GPU Infinity Fabric connectivity for a total of 384 GB/s of peer-to-peer connectivity per APU, with one interface assigned to support x16 PCIe® Gen 5 connectivity to external I/O devices. In addition, each MI300A includes two x4 interfaces to storage, such as M.2 boot drives, plus two USB Gen 2 or 3 interfaces.

Exascale-Class Technology in Your Data Center

Designed with the AMD Instinct MI300A APU, the El Capitan system at Lawrence Livermore National Labs is expected to become the next world's fastest supercomputer. You can put the same technology in your data center to accelerate your most challenging HPC and AI workloads.

Learn More

For more information about the AMD Instinct MI300A and AMD ROCm software, visit AMD.com/INSTINCT.

AMD ROCm 6 Open Software Platform for HPC, AI, and ML Workloads

Whatever your workload, AMD ROCm software opens doors to new levels of freedom and accessibility. Proven to scale in some of the world's largest supercomputers, ROCm software provides support for leading programing languages and frameworks for HPC and AI. With mature drivers, compilers and optimized libraries supporting AMD Instinct accelerators, ROCm provides an open environment that is ready to deploy when you are.

Accelerate Your High Performance Computing Workloads

Some of the most popular HPC programing languages and frameworks are part of the ROCm software platform, including those to help parallelize operations across multiple GPUs and servers, handle memory hierarchies, and solve linear systems. Our GPU Accelerated Applications Catalog includes a vast set of platform-compatible HPC applications, including those in astrophysics, climate & weather, computational chemistry, computational fluid dynamics, earth science, genomics, geophysics, molecular dynamics, and physics. Many of these are available through the AMD Infinity Hub, ready to download and run on servers with AMD Instinct accelerators.

Propel Your Generative AI and Machine Learning Applications

Support for the most popular AI & ML frameworks—PyTorch, TensorFlow, ONYX-RT, Triton and JAX—make it easy to adopt ROCm software for AI deployments on AMD Instinct accelerators. The ROCm software environment also enables a broad range of AI support for leading compilers, libraries and models making it fast and easy to deploy AMD based accelerated servers. The AMD ROCm Developer Hub provides easy access point to the latest ROCm drivers and compilers, ROCm documentation, and getting started training webinars, along with access to deployment guides and GPU software containers for AI, Machine Learning and HPC applications and frameworks.