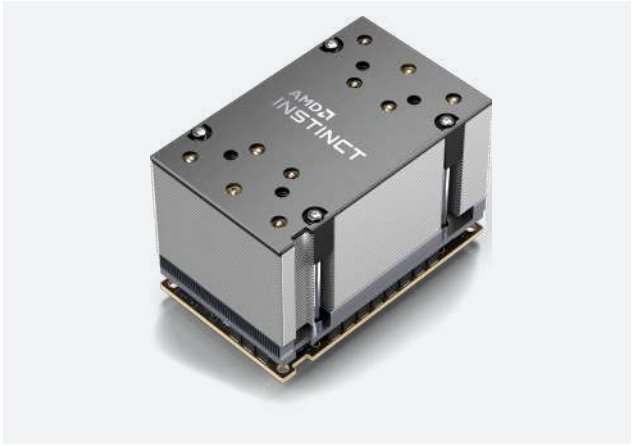
DATA SHEET

AMD INSTINCT™ MI300X ACCELERATOR

Leading-Edge, industry-standard accelerator module
for generative AI, training, and high-performance computing

Leading-Edge Discrete GPU for AI and HPC

The AMD Instinct™ MI300X discrete GPU is based on next-generation AMD CDNA™ 3 architecture, delivering leadership efficiency and performance for the most demanding AI and HPC applications. It is designed with 304 high-throughput compute units, AI-specific functions including new data-type support, photo and video decoding, plus an unprecedented 192 GB of HBM3 memory on a GPU accelerator. Using state-of-the-art die stacking and chiplet technology in a multi-chip package propels generative AI, machine learning, and inferencing, while extending AMD leadership in HPC acceleration. The MI300X offers outstanding performance to our prior generation that is already powering the fastest exaFLOP-class supercomputer[1], offering 13.7x the peak AI/ML workload performance using FP8 with sparsity compared to prior AMD MI250X* accelerators using FP16[MI300-16] and a 3.4x peak advantage for HPC workloads on FP32 calculations.[MI300-11] Do we have your attention? Turn the page to learn more.



AI PEAK THEORETICAL PERFORMANCE		with sparsity
TF32 (TFLOPs)	653.7	1307.4
FP16 (TFLOPs)	1307.4	2614.9
BFLOAT16 (TFLOPs)	1307.4	2614.9
INT8 (TOPS)	2614.9	5229.8
FP8 (TFLOPS)	2614.9	5229.8

HPC PEAK THEORETICAL PERFORMANCE (TFLOPS)	
FP64 vector	81.7
FP32 vector	163.4
FP64 matrix	163.4
FP32 matrix	163.4

DECODERS AND VIRTUALIZATION	
Decoders†	4 groups for HEVC/H.265, AVC/H.264, V1, or AV1
JPEG/MJPEG CODEC	32 cores, 8 cores per group
Virtualization support	SR-IOV, up to 8 partitions

The MI200 Series does not support TF32, FP8, or sparsity

†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	OAM module
Lithography	5nm FinFET
Active interposer dies (AIDs)	6nm FinFET
GPU compute units	304
Matrix cores	1216
Stream processors	19,456
Peak engine clock	2100 MHz
Memory capacity	Up to 192 GB HBM3
Memory bandwidth	5.3 TB/s max. peak theoretical
Memory interface	8192 bits
AMD Infinity Cache™ (last level)	256 MB
Memory clock	Up to 5.2 GT/s
Scale-up Infinity Fabric™ Links	7x 128 GB/s
I/O to host CPU	1 PCIe® Gen 5 x16 (128 GB/s)
Scale-out network bandwidth	PCIe Gen 5 x16 (128 GB/s)
RAS features	Full-chip ECC memory, page retirement, page avoidance
Maximum TBP	750W

Designed to Accelerate Modern Workloads

The increasing demands of generative AI, large-language models, machine learning training, and inferencing puts next-level demands on GPU accelerators. The discrete AMD Instinct MI300X GPU delivers leadership performance with efficiency that can help organizations get more computation done within a similar power envelope compared to last-generation accelerators from AMD.[MI300-23] For HPC workloads, efficiency is essential, and AMD Instinct GPUs have been deployed in some of the most efficient supercomputers on the Green500 supercomputer list[2], these types of systems– and yours–can take now take advantage of a broad range of math precisions to push high-performance computing (HPC) applications to new heights.

Based on 4th Gen Infinity Architecture

The AMD Instinct MI300X is one of the first AMD CDNA 3 architecture-based accelerators with high throughput based on improved AMD Matrix Core technology and highly streamlined compute units. AMD Infinity Fabric™ technology delivers excellent I/O efficiency, scaling, and communication within and between industry-standard accelerator module (OAM) device packages. Each discrete MI300X offers a 16-lane PCIe® Gen 5 host interface and seven AMD Infinity Fabric links for full connectivity between eight GPUs in a ring. The discrete MI300X is sold as an AMD Instinct Platform with eight accelerators interconnected on an AMD Universal Base Board (UBB 2.0) with industry-standard HGX host connectors.

Multi-Chip Architecture

The MI300X 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 OAM module includes:

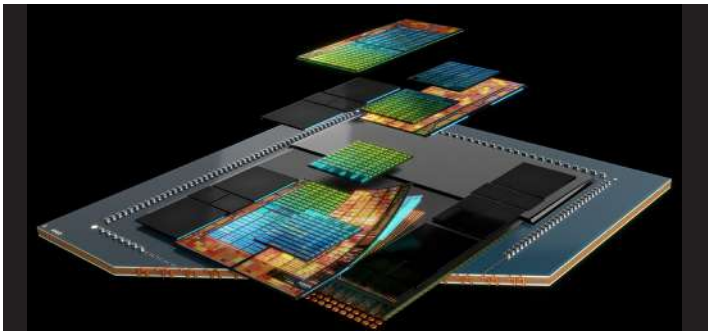
- Eight accelerated compute dies (XCDs) with 38 compute units (CUs), 32 KB of L1 cache per CU, 4 MB shared L2 cache shared across CUs, and 256 MB of AMD Infinity Cache™ shared across 8 XCDs. The compute units support a broad range of precisions for both AI/ML and HPC acceleration, native hardware support for sparsity, and enhanced computational throughput.
- Four supported decoders for HEVC/H.265, AVC/H.264, V1, or AV1, each with an additional 8-core JPEG/MPEG CODEC
- 192 GB of HBM3 memory shared coherently between CPUs and GPUs with 5.3 TB/s on-package peak throughput
- SR-IOV for up to 8 partitions

Coherent Shared Memory and Caches

Machine-learning and large-language models have become highly data intensive, and they need to split jobs across multiple GPUs. AMD Instinct accelerators facilitate large models with shared memory and caches. The large amount of HBM3 memory is supported with 5.3 TB/s of local bandwidth, and direct connectivity of 128 GB/s bidirectional bandwidth between each GPU, accelerating memory-intensive AI, ML, and HPC models.

Learn More

For more information about the AMD Instinct MI300X, the AMD Instinct MI300X Platform, the MI300A APU, and the AMD ROCm™ software platform, 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.

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.

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.

Footnote explanations are available at: https://www.amd.com/en/claims/instinct.

1. Top500, November 2023, https://www.top500.org/lists/top500/2023/11/
2. Top500, The Green500 list, November 2023, https://www.top500.org/lists/green500/2023/11/