AMD INSTINCT

DATA SHEET

AMD INSTINCT™ MI325X ACCELERATOR

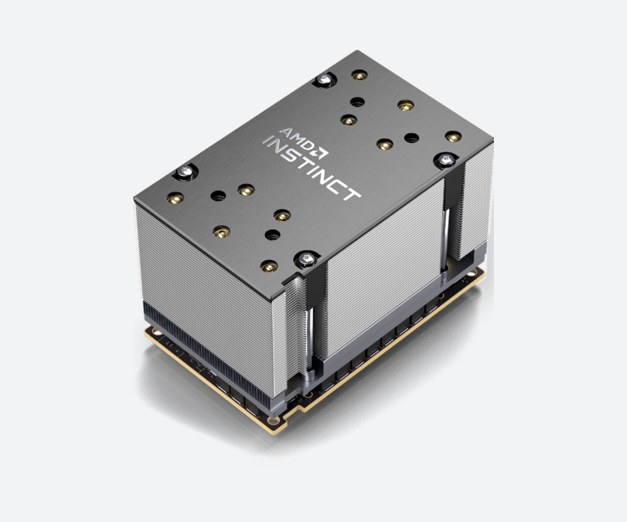
Leading-Edge, industry-standard accelerator module
for generative AI, inference, training, and high performance computing

Designed to Accelerate Modern Workloads

The increasing demands of generative AI, large-language models, inference, and machine learning training puts next-level demands on GPU accelerators. The discrete AMD Instinct MI325X GPU delivers superior performance on a broad set of data types needed for AI software, including FP16, BF16, FP8, and INT8 used in both high-precision inference and training.[MI325-002] An industry-leading 256 GB of HBM3E memory [MI325-001A] and 6 TB/s bandwidth enables a single accelerator to contain and process a one-trillion parameter model while reducing total cost of ownership for select large-language models.[MI325-003]

Support for matrix sparsity further economizes memory use and boosts computational speed, helping enable sustainable scaling of AI solutions across data centers, speeding time to market and enhancing performance.

Integrated with AMD ROCm™ software, the accelerator supports key AI and HPC frameworks, simplifying deployment. Seamless, drop-in compatibility with the AMD Instinct MI300X Platform comes through strong support from our OEM partners, industry-leading frameworks, and thousands of large-language models.



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, VP9, or AV1
JPEG/MJPEG CODEC	32 cores, 8 cores per group
Virtualization support	SR-IOV, up to 8 partitions

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

SPECIFICATIONS

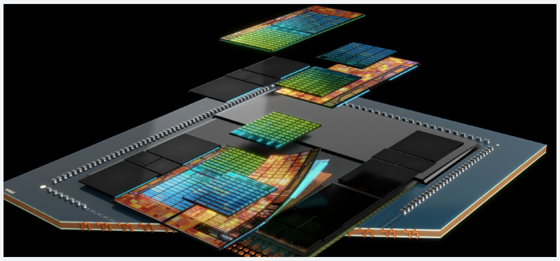
Form factor	OAM module
Lithography	5nm FinFET
Active interposer dies (AIDs)	6nm FinFET
GPU compute units	304
Matrix cores	1216
Stream processors	19,456
Peak engine clock	2100 MHz
Memory capacity	Up to 256 GB HBM3E
Memory bandwidth	6 TB/s max. peak theoretical
Memory interface	8192 bits
AMD Infinity Cache™ (last level)	256 MB
Memory clock	Up to 6.0 GT/s
Scale-up AMD 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	1000W

Introducing the AMD Instinct MI325X Accelerator and Platform

3RD GEN AMD CDNA™ CORE ARCHITECTURE: Utilizes advanced die stacking and chiplet technology. This cutting-edge design is purpose-built to drive the most demanding AI workloads, delivering incredible performance, efficiency and scalability.

MEMORY CAPACITY: Featuring huge HMB3E memory, these accelerators effortlessly handle large datasets and complex computations, making them ideal for AI inferencing, AI training, and data analytics. A platform powered by 8 AMD Instinct™ MI325X GPU accelerators boasts a mind-boggling two terabytes of memory with reduced latency compared to the competition.[MI325-015] The MI325X Platform offers seamless scalability as a drop-in replacement for the Instinct MI300X Platform. Boost multitasking efficiency and optimize resource utilization while supporting multiple virtual machines and extensive AI models.

MEMORY BANDWIDTH: Experience rapid data transfer, enhanced throughput, and minimized latency, enabling quick access to large datasets, accelerating AI training processes. Improve scalability and support complex, high-resolution data with ease, optimizing GPU core utilization for efficient, real-time AI inferencing. Transform your AI capabilities with exceptional data processing speed and efficiency, driving more powerful AI solutions.



Based on 3rd Gen AMD CDNA Architecture

The AMD Instinct MI325X is an AMD CDNA 3 architecture-based accelerator 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 MI325X 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 MI325X is available in an AMD Instinct MI325X Platform with eight accelerators interconnected by an AMD Universal Base Board (UBB 2.0) with HGX host connectors.

Multi-Chip Architecture

The MI325X uses a multi-chip architecture that enables dense compute and high-bandwidth memory integration. 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, VP9, or AV1, each with an additional 8-core JPEG/MPEG CODEC
- 256 GB of HBM3E memory with 6 TB/s on-package peak throughput
- SR-IOV for up to 8 partitions

Coherent Shared Memory

AMD Instinct accelerators facilitate large models with coherent, shared memory between all eight accelerators on a UBB with 128 GB/s bidirectional bandwidth between each GPU to accelerate memory-intensive AI, ML, and HPC models.

Learn More

For more information, visit AMD.com/INSTINCT.

Proven, Open, Limitless Software Ecosystem

AMD ROCm

The Instinct MI325X accelerator leverages latest AMD ROCm™ open software platform, designed to accelerate AI inference and training. Experience extraordinary performance, scalability, and developer productivity with comprehensive tools, compilers, libraries, and APIs that optimize accelerator utilization and streamline AI development.

Propel Generative AI and Machine Learning Applications

Support for the most popular AI & ML frameworks—PyTorch™, TensorFlow™, and Jax, along with LLMs including Hugging Face, Databricks, Lamini, and JAX—make it easy to adopt ROCm software for AI deployments on AMD Instinct accelerators. 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 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/legal/claims/instinct.html