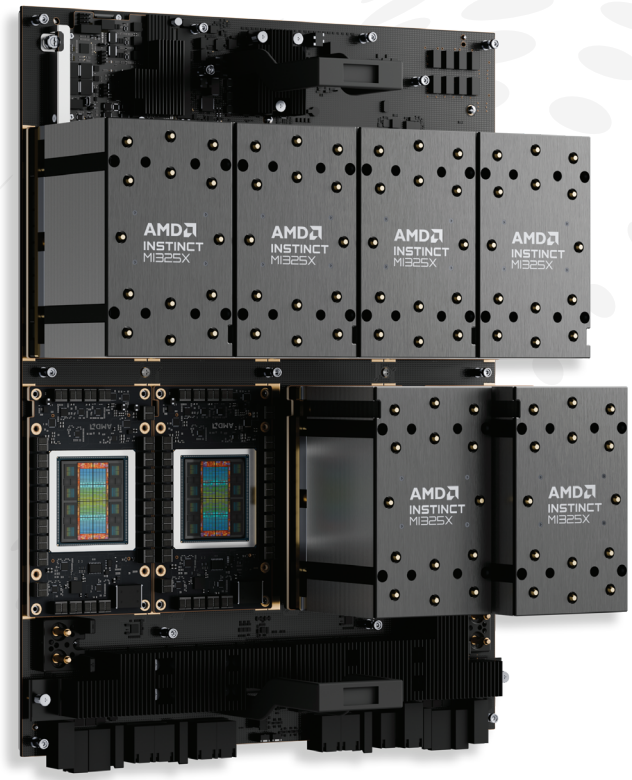
5 REASONS WHY AMD INSTINCT™ MI325X ACCELERATORS ARE THE NEW STANDARD TO ADVANCE GENERATIVE AI AND THE DATA CENTER

AT A GLANCE

Today's great technology opportunity is fueled by acceleration. Generative AI requires training on massive data sets and inference at scale. Exceptional efficiency must be driven in the data center to make room, while exceptional performance is still required for the new class of HPC applications needed for continued progress in healthcare, energy, climate science, transportation, scientific research and more. These needs are only becoming more intense as the enterprise ramps up generative AI development and integrates AI everywhere it can.

What if you could get more acceleration performance in a drop-in platform with leadership compute unit counts and advanced high-bandwidth memory with leadership capacity? Would you like to develop AI and HPC applications quickly and reliably on more than one acceleration platform without having to worry about compatibility, operational complexity or specialized design requirements?



WHERE VAST MEMORY MEETS LEADERSHIP PERFORMANCE.

The **AMD Instinct™ MI325X** accelerator is the quick-to-market, improved performance release in the MI300 family, built on the AMD CDNA™ 3 accelerator microarchitecture. The MI325X data center GPU is designed to deliver even more raw acceleration power for the most demanding generative AI, training and HPC applications while also improving total cost of ownership. Packaging high-throughput GPU compute units (CUs) with leadership 256GB of upgraded high-bandwidth memory (HBM3E), AMD Instinct MI325X is deployed on an industry standard 8× OAI-UBB 2.0 based platform with all GPUs fully connected over high bandwidth, low latency AMD Infinity Fabric™ technology.

Built to address cost, compatibility and power/cooling efficiency challenges inherent in the modern data center, the AMD Instinct MI325X accelerator can help organizations rapidly tap new levels of performance for fast results in AI and HPC.

1

LEADERSHIP PERFORMANCE FOR AI AND HPC

Discover the possibilities of accelerated performance at scale.

The AMD Instinct MI325X accelerator integrates high-throughput CUs and 256GB of stacked HBM3E memory interconnected over a coherent, high-bandwidth fabric with leadership 6TB/s peak theoretical bandwidth, 1.3× the bandwidth of the competition.[MI325-001A] This raw capacity and speed can deliver industry leading acceleration for the latest generative AI models like GPT4, Llama 3.1 and many more, as well as shorter job times for HPC applications.

2

MORE ACCELERATION FOR MASSIVE DATA SETS

Increase compute and memory in each data center system.

The performance demands of large-scale production require packing the most compute processing and memory capacity possible in the systems. The AMD Instinct™ MI325X accelerator uses state-of-the-art die-stacking, chiplet technologies and special processing to enable more compute and high-bandwidth memory on the rack and enhance space efficiency.

3

ENHANCED DATA CENTER POWER USE

Leverage new levels of efficiency for AI and accelerated HPC.

The AMD Instinct MI325X platform packs compute and HBM3E onto the rack on a single OAI UBB 2.0 8-GPU OAM unit per node. Fabrication and design enhance energy usage for high computation per watt. Native hardware support for matrix sparsity helps save power in AI training.

4

LOW TCO

Build out data centers that address budget and sustainability goals.

AMD Instinct MI300 Series allows CU inventory to be partitioned on each accelerator for use by more than one virtual client to reduce capacity waste and increase hardware utilization: get 2, 4 or 8 partitions per MI325X for multi-client access to the same GPU. Floating-point and integer operations can co-execute for additional efficiencies.

5

OPEN, HIGHLY PROGRAMMABLE GPU SOFTWARE PLATFORM

Ease the way to results with an ecosystem designed for accessibility and adaptability.

AMD facilitates the adoption and use of multiple acceleration platforms, and cross-platform AI and HPC development, with the ROCm™ open software ecosystem and programming toolset. The ROCm stack provides an open-source and easy-to-use set of tools that are built around industry standards, the AMD Instinct MI325X accelerator providing numerous features for simplifying programming and assuring consistent runtime performance at scale.

TECHNICAL DEEP DIVE

#1 DISCOVER THE POSSIBILITIES OF ACCELERATED PERFORMANCE AT SCALE

- Offered as a fast-to-deploy, fully connected 8× AMD Instinct MI325X platform, each node offers **2TB of high-bandwidth memory (HBM3E)** for low-latency processing of large ML models. That's at least 1.3× the capacity of the competition.MI325-001A
- Each AMD Instinct MI325X accelerator offers **leadership HBM3E peak theoretical throughput** compared to the Nvidia H200 SXM (141GB) accelerator.MI325-001A
- Get **improved AI processing** with ~1.3× the peak theoretictal half-precision (FP16 Tensor) floating-point performance of the competition.MI325-002
- The AMD Instinct MI325X accelerator provides up to 1TB/s peak aggregate theoretical GPU I/O bandwidth performance.MI325-011
- AMD Instinct accelerators are featured in the world's first exascale supercomputer, Frontier,[1] and will also be featured in the still-under-construction El Capitan exascale supercomputer.

#2 INCREASE COMPUTE AND MEMORY IN EACH SYSTEM

- The AMD Instinct™ MI325X accelerator's die stacking and chiplet architecture delivers compute density and efficiency from the reduction of data-movement overhead.
- Multi-device fully connected designs through AMD Infinity Fabric™ links enable up to 896GB/s of peak theoretical peer-to-peer I/O bandwidth through 128GB/s of bi-directional I/O bandwidth per link in 8× GPU AMD Instinct platforms.MI325-011
- A one-trillion-parameter model is calculated to run on a single MI325X accelerator.MI325-003

#3 LEVERAGE NEW LEVELS OF EFFICIENCY FOR AI AND HPC

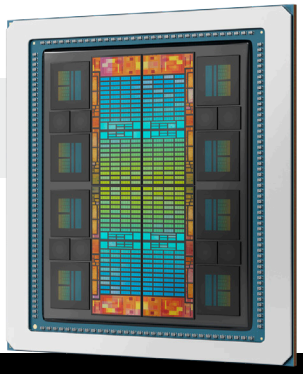
- AMD Instinct MI325X (1,000 watts) offers up to 7.7x the peak theoretical generative AI and training workload performance FP8 performance per watt of the previous generation MI250X (560 watts).MI325-010,2
- AMD Instinct MI325X accelerators deliver 46% more compute density than previous-generation AMD Instinct GPUs.MI325-009
- Native hardware support for sparse matrices on AMD Instinct MI300 Series accelerators helps save power, lower compute cycles and reduce memory use during AI and ML training.

#4 ADDRESS BUDGET AND SUSTAINABILITY GOALS

- The AMD Instinct MI325X accelerator offers high Compute Unit (CU) utilization, supporting up to eight partitions per GPU, enabling a total of up to 64 partitions per server for multi-client access in virtualized deployments and enhanced computational throughput from co-execution of floating-point and integer operations.
- The AMD Instinct MI325X accelerator offers superior capabilities for large language models (LLMs), better than that of the Nvidia H200 SXM (141GB) GPU.MI325-004

#5 EASE THE WAY TO RESULTS WITH AN ECOSYSTEM DESIGNED FOR ACCESSIBILITY AND ADAPTABILITY

- The AMD Instinct MI325X accelerator offers a highly programmable architecture featuring the proven AMD ROCm™ 6.0 open software platform for AI and HPC development and deployment.
- The frictionless software ecosystem includes drop-in support for major AI and HPC frameworks and programming models, a key advantage for your evolving software needs.
- All source code is published on Github, including drivers, tools and libraries. Build environment scripts (CMake) are available to compile source for target devices.
- AMD partners with many AI technical leaders including the Allen Institute for AI, Hugging Face, Lamini, OpenAI and many more LLM initiatives and continues to participate in open-source cross-platform initiatives such as MLIR, OpenMP®, OpenCL™, OpenXLA, PyTorch, TensorFlow and Triton.



AMD INSTINCT™ MI300 SERIES IS READY TO DEPLOY.

AMD
together we advance_data centers

LEARN MORE AT AMD.COM/INSTINCT AND AMD.COM/ROCm

1 Top500 June 2024 list, https://top500.org/lists/top500/list/2024/06/.

2 MI200 series accelerators don't support FP8 and exploit structured sparsity. The MI325X is expected to be able to take advantage of fine-grained structure sparsity providing a 2x improvement in math efficiency. MI325X FP8 performance calculated with sparsity.