AMD INSTINCT

DATA SHEET

AMD INSTINCT™ MI325X PLATFORM

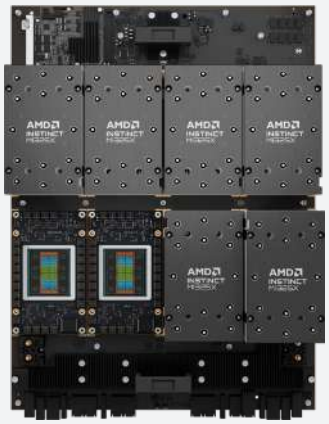
Advanced accelerator solution for AI, inference, and training

Powerful Industry-Standard 8-GPU Solution

Today's generative AI and large-language models need three elements to speed responses: fast acceleration across multiple data types, large memory and bandwidth to handle huge data sets and models, and extreme I/O bandwidth.

You get all three with the AMD Instinct™ MI325X Platform with 3rd Gen AMD CDNA™ architecture-based GPUs: 42 petaFLOPs of peak theoretical FP8 with sparsity precision performance for generative AI and ML training. Our industry-standard-based universal baseboard (UBB 2.0) platform hosts 8 AMD Instinct™ MI325X accelerators and 2 TB of HBM3E memory to help process the most demanding AI models With eight x16 PCIe® Gen 5 host I/O connections, you don't have to worry about data bottlenecks.

With drop-in compatibility with the AMD Instinct MI300X Platform, you can update your technology with the MI325X Platform, where vast memory meets leadership performance. With this platform you can deploy fast, train fast, and optimize your total cost of ownership (TCO).



AI PEAK THEORETICAL PERFORMANCE		with sparsity
TF32	5.2 PFLOPs	10.5 PFLOPs
FP16	10.5 PFLOPs	20.9 PFLOPs
BFLOAT16	10.5 PFLOPs	20.9 PFLOPs
INT8	20.9 POPs	41.8 POPs
FP8	20.9 PFLOPs	41.8 PFLOPs

HPC PEAK THEORETICAL PERFORMANCE	
FP64 vector	653.6 TFLOPs
FP32 vector	1.3 PFLOPs
FP64 matrix	1.3 PFLOPs
FP32 matrix	1.3 PFLOPs

DECODERS AND VIRTUALIZATION	
Decoders*	32 groups for HEVC/H.265, AVC/H.264, VP9, or AV1
JPEG/MJPEG CODEC	256 cores, 8 cores per group
Virtualization support	SR-IOV, up to 64 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	Universal baseboard (UBB) module with 8 Instinct MI325X OAM GPUs
Lithography	5nm FinFET
Active interposer dies (AIDs)	6nm FinFET
GPU compute units	2432
Matrix cores	9728
Stream processors	155,648
Peak engine clock	2100 MHz
Memory capacity	2.048 TB HBM3E
Memory bandwidth	6 TB/s max. peak theoretical
Memory interface	8192 bits per GPU
AMD Infinity Cache™ (last level)	256 MB per GPU
Memory clock	Up to 6.0 GT/s
Scale-up Infinity Fabric™ Links	7x 128 GB/s per GPU
Ring of 8 aggregate bandwidth	896 GB/s
Scale-out network bandwidth	8 PCIe® Gen 5 x16 (128 GB/s) per GPU
RAS features	Full-chip ECC memory, page retirement, page avoidance
Maximum TBP	1000W per GPU

The Challenges of Diverse Data Requirements

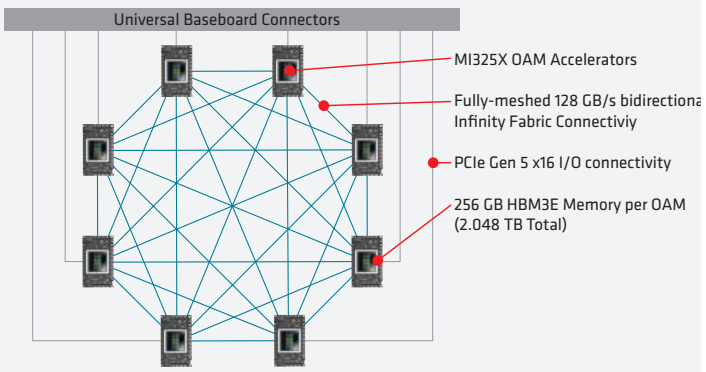
Emerging generative AI and large-language models have a voracious appetite for data. Support for a wide range of data types, compute density, and large memory capacities prepare the AMD Instinct MI325X Platform to tackle these diverse workloads. We bring low-precision data types such as FP8, INT8, FP16, and BF16 with hardware-based sparsity to propel scale-out generative AI and machine-learning models. With the introduction of sparsity, AI models lacking dense data structures can be accelerated with even greater memory efficiency. Today, an Instinct MI325X Platform is projected to handle trillion-parameter models on a single platform. [MI325-003]

AMD Instinct MI325X Platform

To offer the power of the AMD Instinct MI325X accelerator through industry-standard servers, we have designed a platform to combine the power of eight accelerators on an industry-standard universal baseboard (UBB 2.0). The eight Open Compute Project (OCP) Accelerator Modules (OAMs) are connected with an AMD Infinity Fabric™ mesh that provides direct connectivity between each of the GPUs over 128 GB/s bidirectional links. Each MI325X connects with its peers through seven links, plus one PCIe® Gen 5 x16 connection per OAM device for upstream server and/or I/O connectivity. Remote DMA I/O transfers can stream data to each GPU where it is needed and where it can be processed in each module's large 256 GB HBM3E memory.

Based on 3rd Gen AMD CDNA Architecture

The AMD Instinct MI325X accelerator is based on the AMD CDNA 3 architecture offering high throughput based on generationally improved AMD Matrix Core technology and streamlined compute units. The AMD Instinct MI325X GPU also supports PCIe Gen 5 with AMD Infinity Fabric™ technology helping to improve I/O performance, efficiency, and scaling within and between each OAM device on the universal baseboard.



AMD Infinity Platform Architecture

High-Speed GPU Interconnects

Generative AI, machine learning, and large-language models have become highly data intensive, and they often need to split jobs across multiple GPUs. The AMD Instinct MI325X Platform facilitates large models through a 2 TB coherent shared memory with 6.0 TB/s of peak bandwidth within each GPU accelerator and 128 GB/s of bidirectional Infinity Fabric bandwidth between each GPU for a peak aggregate bandwidth of 896 GB/s. Cache coherency is supported by a shared 256 MB Infinity Cache™ that supports all compute units in a single GPU.

Learn More

The AMD Instinct MI325X Platform solution is available through AMD solution partners. Please contact your preferred solution partner to find out when their AMD Instinct MI325X Platform-based solutions will be available. Standard form factors such as UBB-based solutions facilitate adoption into enterprise servers so that you can use the same power in your data center, or in the cloud from the leading superscalars. Learn more at AMD.com/INSTINCT.

Proven, Open, Limitless Software Ecosystem

The Instinct MI325X accelerator leverages latest AMD ROCm™ 6 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.

AMD ROCm

Propel Generative AI and Machine Learning Applications

Support for the most popular AI & ML frameworks–PyTorch™, TensorFlow™, ONYX-RT, Triton, 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..