

DATA SHEET

AMD INSTINCT™ MI300X PLATFORM™

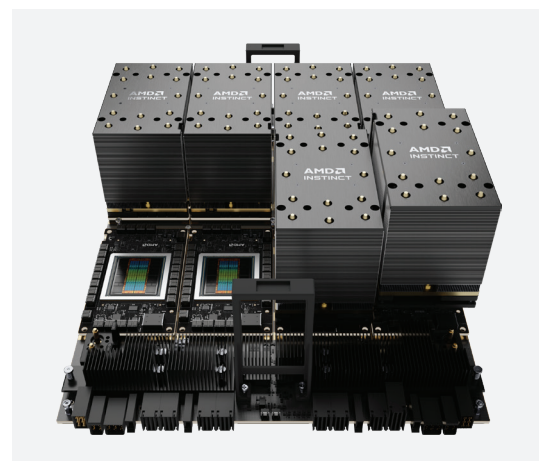
Advanced accelerator solution for generative AI and compute

Powerful Industry-Standard 8-GPU Solution

Today's large-scale AI/ML training sets and HPC data need three elements to accelerate workloads: fast acceleration across multiple data types, large memory and bandwidth to handle huge data, and extreme I/O bandwidth.

You get all three with the AMD Instinct™ MI300X 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 and 1.3 petaFLOPs peak theoretical FP32 precision for the most challenging HPC codes. Our industry-standard-based universal baseboard (UBB 2.0) platform hosts 8 AMD Instinct™ MI300X accelerators and 1.5 TB of HBM3 memory to help process the most demanding AI models and HPC workloads. With eight x16 PCIe® Gen 5 host I/O connections, you don't have to worry about data bottlenecks.

The bottom line is a platform that's based on open standards that incorporate proven AMD Instinct™ technology that is expected to drive some of the world's fastest supercomputers, and an open software platform that is ready to support you.



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	653.7 TFLOPs
FP32 vector	1307.4 TFLOPs
FP64 matrix	1307.4 TFLOPs
FP32 matrix	1307.4 TFLOPs

DECODERS AND VIRTUALIZATION

Decoders*	32 groups for HEVC/H.265, AVC/H.264, V1, 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 MI300X 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	1.5 TB HBM3
Memory bandwidth	5.3 TB/s max. peak theoretical
Memory interface	8192 bits per GPU
AMD Infinity Cache™ (last level)	256 MB per GPU
Memory clock	Up to 5.2 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	750W per GPU

The Challenges of Diverse Data Requirements

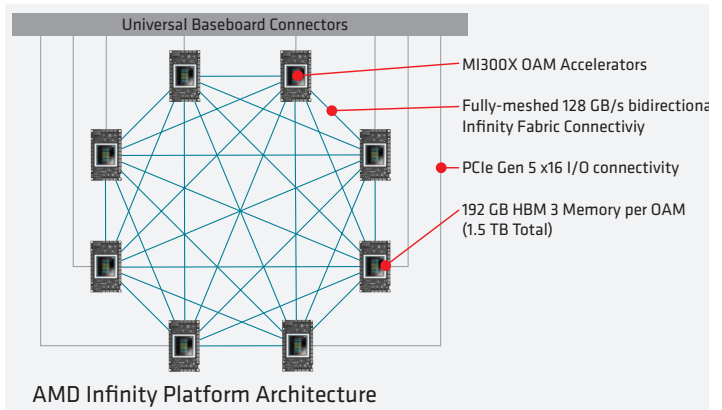
Emerging generative AI and ML training and HPC compute codes have a voracious appetite for data. Support for a wide range of data types, compute density, and large memory capacities prepare the AMD Instinct MI300X Platform to tackle these diverse workloads. We bring low-precision data types such as FP8, INT8, and FP16/BF16 with hardware-based sparsity to propel scale-out generative AI and machine-learning models. FP64 and FP32 operations help drive the most demanding HPC compute codes. With the introduction of sparsity, AI models lacking dense data structures can be accelerated with even greater efficiency.

AMD Instinct MI300X Platform

To offer the power of the AMD Instinct MI300X 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 MI300X connects with its peers through seven links, plus one PCIe® Gen 5 x16 connection per OAM device provides 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 192 GB HBM3 memory.

Based on 4th Gen Infinity Architecture

The AMD Instinct MI300X accelerator is based on the 4th Gen Infinity architecture and the AMD CDNA™ 3 architecture offers high throughput based on generationally improved AMD Matrix Core technology and streamlined compute units. The AMD Instinct MI300X 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.



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. AMD Instinct MI300X accelerators facilitate large models through a large 192 GB HBM3 memory capacity supported by 5.3 TB/s of peak local bandwidth and a large internally shared 256 MB AMD Infinity Cache™ between compute units. In addition, the platform connects all GPUs in the 8-node ring with a peak aggregate bidirectional Infinity Fabric bandwidth of 896 GB/s.

Learn More

The AMD Instinct MI300X Platform solution is expected to be available through AMD solution partners. Please contact your preferred solution partner to find out when their AMD Instinct MI300X 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.

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 programming 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, ONNX-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 programming 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.

AMD
ROCm

© 2023 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD Instinct, Infinity Cache, Infinity Fabric, ROCm, and combinations thereof are trademarks of Advanced Micro Devices, Inc. PCIe is a registered trademark of PCI-SIG Corporation. PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc. TensorFlow, the TensorFlow logo, and any related marks are trademarks of Google Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Use of third party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied GD-83
PID # 232405395-D