5 REASONS WHY AMD INSTINCT™ MI300X ACCELERATORS ARE THE RIGHT CHOICE FOR DEMANDING AI AND HPC APPLICATIONS

AT A GLANCE

The technology revolution of today is all about acceleration. Accelerated HPC applications, used for discovery, modeling and prediction, are required for progress in healthcare, energy, climate science, transportation, scientific research and more. AI, in sudden and huge demand across nearly every field and industry, requires training on massive data sets and inference at scale. But creating an HPC or AI system is an involved and complex endeavor, with great needs for accuracy, relevance and reliability. Going forward these needs will only become more intense, and most stakeholders may not realize it.

What if you could get more raw acceleration performance in a standard platform with leadership compute unit counts and high-bandwidth memory 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?

AMD INSTINCT™ MI300 SERIES IS READY TO DEPLOY.

The **AMD Instinct MI300X** data center GPU is the performance release in the MI300 family, built on next-generation AMD CDNA™ 3 accelerator microarchitecture. The MI300X GPU is designed to deliver raw acceleration power for the most demanding generative AI, training and HPC applications while also improving energy efficiency. Packaging high-throughput GPU compute units (CUs) with leadership 192GB of high-bandwidth (HBM3) memory, AMD Instinct MI300X 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™.

Built to address the cost, compatibility and power/cooling efficiency challenges inherent in deploying GPU systems at scale, the AMD Instinct MI300X GPU can help researchers and enterprises tap into breakthrough performance for fast results in AI and HPC applications.

1 LEADERSHIP PERFORMANCE FOR AI AND HPC

Discover the possibilities of accelerated performance at scale.

The AMD Instinct MI300X GPU integrates high-throughput CUs and 192GB of stacked HBM3 memory interconnected over a coherent, high-bandwidth fabric with leadership 5.3TB/s peak theoretical bandwidth, 1.7× the bandwidth of the previous generation.[MI300-13] This raw capacity and speed can enable new levels of acceleration for both longstanding HPC applications and new AI workloads.

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™ MI300X GPU uses state-of-the-art die-stacking, chiplet technologies and special processing such as Matrix Core Technologies in a multi-chip package to enable more compute and high-bandwidth memory on the rack, reduce the need for data movement and enhance space efficiency.

3

ENHANCED DATA CENTER POWER USE

Leverage new levels of efficiency for AI and accelerated HPC.

AMD Instinct MI300X accelerators pack compute and HBM3 onto the rack, offering a single industry standard OAI UBB 2.0 8-GPU OAM module platform per node and featuring fabrication and design to enhance energy usage and produce high computation per watt compared to previous-generation AMD Instinct products.[MI300-15, MI300-22] New native hardware support for matrix sparsity helps save power and compute cycles.

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 MI300X for multi-client access to the same GPU. Realize additional efficiencies from co-execution of floating-point and integer operations.

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 and enable creating well-optimized portable software for AI and HPC. The AMD Instinct MI300X GPU provides numerous features especially helpful for simplifying programming and assuring consistent runtime performance at scale, including a unified memory space and AMD Infinity Cache™. Simplified programmability and portability and high usability accelerate time to results.

TECHNICAL DEEP DIVE

#1 DISCOVER THE POSSIBILITIES OF ACCELERATED PERFORMANCE AT SCALE

- Offered as a fast-to-deploy, fully connected 8× AMD Instinct MI300X GPU solution, the AMD Instinct platform offers **1.5TB of high-bandwidth (HBM3) memory capacity** for low-latency processing of large ML models. That's up to 1.4× the HBM3 capacity of the competition.[MI300-25]
- Each AMD Instinct MI300X GPU offers **leadership HBM3 peak theoretical throughput** compared to the Nvidia H100 SXM3 (80GB) accelerator.[MI300-05A]
- Get **improved AI processing** with ~3.4× the half-precision (FP16) and Bfloat16 total peak theoretical floating-point performance of previous-generation AMD Instinct GPUs.[MI300-11]
- The AMD Instinct MI300X GPU provides **up to 1TB/s peak aggregate theoretical GPU I/O bandwidth performance**.[MI300-06]
- AMD Instinct accelerators are featured in the world's fastest supercomputer, Frontier,[1] and will also be featured in the still-under-construction El Capitan exascale supercomputer.

#2 INCREASE COMPUTE AND MEMORY IN EACH DATA CENTER SYSTEM

- The AMD Instinct™ MI300X GPU'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.MI300-06
- The Hugging Face OPT transformer 66B-parameter large language model (LLM) can run on a single MI300X GPU using FP16.MI300-07
- Matrix Core Technologies increase the amount of matrix processing from a given number of CUs.

#3 LEVERAGE NEW LEVELS OF EFFICIENCY FOR AI AND ACCELERATED HPC

- AMD Instinct MI300X (750 watts) offers up to ~2.5× the peak theoretical generative AI and training workload FP16 performance per watt of the previous generation MI250X (560 watts).MI300-27
- AMD Instinct MI300X accelerators deliver 46% more compute density than previous-generation AMD Instinct GPUs.MI300-15
- For the first time, get native hardware support for sparse matrices on AMD Instinct accelerators, which helps save power, lower compute cycles and reduce memory use during AI and ML training.

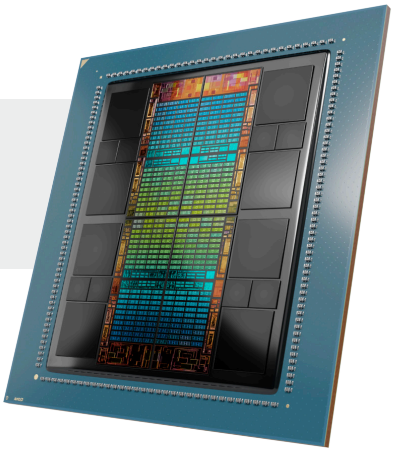
#4 BUILD OUT DATA CENTERS THAT ADDRESS BUDGET AND SUSTAINABILITY GOALS

- The AMD Instinct MI300X GPU offers high CU utilization from up to eight partitions of individual GPUs for multi-client access in virtualized deployments as well as from enhanced computational throughput from co-execution of floating-point and integer operations.
- The AMD Instinct MI300X GPU is expected to deliver superior capabilities for large language models (LLMs), better than that of the Nvidia H100 SXM3 (80GB) GPU.MI300-08
- AMD Instinct powers 7 of the top 10 supercomputer systems on the Green500 list.[2]

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

- The AMD Instinct MI300X GPU offers a highly programmable GPU architecture featuring the AMD ROCm™ 6.0 open software platform, a proven platform for hyper-class HPC and AI deployments.
- 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 and OpenAI and continues to participate in open-source cross-platform initiatives such as MLIR, OpenMP®, OpenCL™, OpenXLA, PyTorch, TensorFlow and Triton.

AMD INSTINCT MI300X ACCELERATORS



AMD
together we advance_data centers

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

1 Top500, The Top500 list, 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/

PID 232405395-H 12/23