5 REASONS WHY AMD INSTINCT™ MI300A ACCELERATORS ARE A GAME-CHANGER IN HPC AND AI

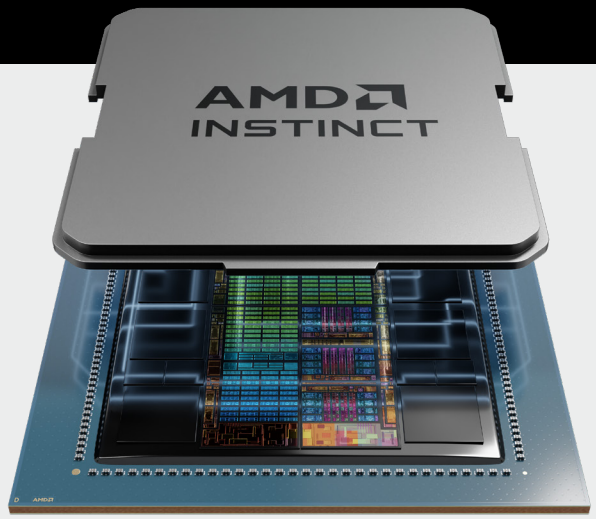
AT A GLANCE

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 accelerated performance from the same amount of space and power in the data center? Would you like to develop HPC and AI 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 MI300A** is a new packaging approach, the accelerated processing unit (APU), for the advancement of HPC and AI. Built on the next-generation AMD CDNA™ 3 accelerator microarchitecture, the MI300A APU is designed with state-of-the-art die stacking and chiplet technologies, putting GPU, CPU and high-bandwidth memory (HBM3) together in a single package for each server socket. **With the MI300A, AMD delivers leadership HPC and AI performance with breakthrough density and efficiency,** helping researchers and enterprises speed their way to results.



1 ACCELERATED HPC AND AI ON A STREAMLINED PLATFORM

Discover the possibilities of accelerated performance at scale.

AMD Instinct MI300 Series accelerators deliver new levels of leadership performance, both for longstanding accelerated HPC applications and for the newly exploding demand for generative AI. The AMD Instinct MI300A APU integrates high-throughput AMD CDNA 3 based GPU compute units (CUs) and high-performance AMD "Zen 4" x86-based CPU cores with 128GB of unified HBM3 memory on a coherent, high-bandwidth fabric. The cores and CUs also share a large unified AMD Infinity Cache™ that provides additional uplift for memory-bound workloads.

2

MORE ACCELERATION FOR MASSIVE DATA SETS

Reduce data movement for streamlined processing flows.

The performance demands of accelerated systems require a large amount of processing and memory as well as very fast throughput in order to ensure optimal speed of computation and the fewest possible bottlenecks. The AMD Instinct™ MI300A multi-chip package enables dense compute and high-bandwidth memory integration, reducing data-movement overheads and enhancing space efficiency.

3

ENHANCED DATA CENTER POWER USE

Leverage new levels of efficiency for AI and accelerated HPC.

The AMD Instinct MI300A APU architecture confers the power advantages of packaging CPU, GPU and memory together in the same unit. Additionally, with the MI300 Series, AMD introduces native hardware support for sparsity at AI precisions, efficiently compressing sparse matrices before processing to help save power and compute cycles and reduce memory use.

4

LOW TCO

Build out data centers that address budget and sustainability goals.

The AMD Instinct MI300A APU allows for granular thermal design power (TDP) settings down to ~25% below maximum, letting users fine-tune power usage against workload requirements to help avoid energy waste and lessen the impact on data center budget costs and related emissions. CU inventory can be partitioned on each APU for use by more than one virtual client in virtualized environments to increase capacity utilization.

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 HPC and AI development with the ROCm™ open software ecosystem and programming toolset. The AMD Instinct MI300A APU provides numerous features especially helpful for simplifying programming and enabling consistent runtime performance at scale, including a unified memory space, cache coherence and efficient inherent GPU-CPU synchronization, for CPU-like coding of a GPU-accelerated platform. Simplified programmability, portability and high usability accelerate time to results.

TECHNICAL DEEP DIVE

#1 DISCOVER THE POSSIBILITIES OF ACCELERATED PERFORMANCE AT SCALE

- By packaging the latest CPU, GPU and high-bandwidth memory all in one unit, the AMD Instinct MI300A APU shorten data pathways and contain them within the socket, helping to solve large compute challenges faster and better.
- The AMD Instinct MI300A APU offers **2.6× the peak theoretical AI performance** (TFLOPS) on FP16 and Bfloat16 of previous-generation AMD MI250X accelerators.[MI300-10]
- Wide range of use cases thanks to **superior or comparable performance** in FP64 and FP32 vector/matrix calculations for HPC; and FP16, Bfloat16, TF32, FP8 and INT8 for AI compared to the Nvidia H100 SXM5 (80GB).[MI300-20]
- Experience fewer latency-inducing data copy operations thanks to a leadership capacity and resource sharing of 128GB of on-APU HBM3 memory.
- Get improved AI capabilities from hardware support for E5M2 and E4M3 (FP8).

#2 REDUCE DATA MOVEMENT FOR STREAMLINED PROCESSING FLOWS

- The AMD Instinct™ MI300A APU encompasses the 4th Gen AMD Infinity architecture and offers a single coherent fabric for GPU, CPU and unified HBM3 along with 256MB shared L3 AMD Infinity Cache™. It packs 228 AMD CDNA™ 3 CUs and 24 AMD EPYC™ "Zen 4" x86-based cores with 128GB of HBM3 onto a single package.
- 2- and 4- socket fully connected nodes with 128GB/s bi-directional APU-to-APU I/O bandwidth per link through AMD Infinity Fabric™ offers accelerated computing performance within a smaller physical space.MI300-24

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

- Matrix Core Technologies increase the amount of matrix math processing from a given number of CUs.
- Improved HPC job efficiencies with up to a ~89% FP32 performance-per-watt gain over previous-generation AMD Instinct GPUs.MI300-23
- For the first time, get native hardware support for sparse matrices on AMD Instinct accelerators, designed to save power, lower compute cycles and reduce memory use for AI workloads.

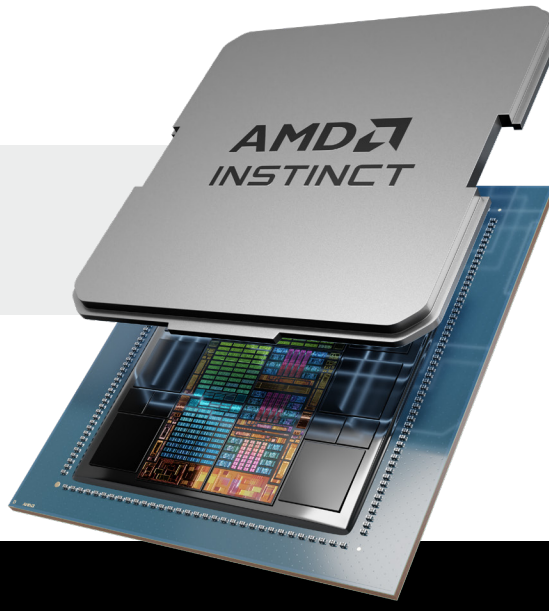
#4 BUILD OUT DATA CENTERS THAT ADDRESS BUDGET AND SUSTAINABILITY GOALS

- The AMD Instinct MI300A APU offers high CU utilization from partitioning of individual GPUs for multi-client access in virtualized environments and from the enhanced computational throughput of co-execution of floating-point and integer operations.
- The AMD Instinct MI300A APU has a lower TDP option of as low as 550W for applications that do not need the full 760W envelope. The TPD can be adjusted granularly as needed for optimal power use.
- AMD Instinct powers 7 of the top 10 supercomputer systems on the Green500 list.[1]

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

- The AMD Instinct MI300A APU offers a highly programmable GPU architecture, including unified memory space across GPU and CPU for simpler programming and transparent management of CPU and GPU caches via cache coherence.
- The AMD ROCm™ 6.0 open software platform is a proven platform for hyper-class HPC and AI deployments. The frictionless software ecosystem includes drop-in support for major AI frameworks 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 MI300A ACCELERATORS



AMD
together we advance_data centers

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

1 Top500, The Green500 list, November 2023, https://www.top500.org/lists/green500/2023/11/

PID 232405395-E 12/23