December 6, 2023

AMD Delivers Leadership Portfolio of Data Center AI Solutions with AMD Instinct MI300 Series

— Dell Technologies, Hewlett Packard Enterprise, Lenovo, Meta, Microsoft, Oracle, Supermicro and others showcase AMD hardware for high performance computing and generative AI —

— ROCm 6 open software ecosystem combines next-gen hardware and software to deliver ~8x generational performance increase, power advancements in generative AI and simplify deployment of AMD AI solutions —

SANTA CLARA, Calif., Dec. 06, 2023 (GLOBE NEWSWIRE) -- Today, AMD (NASDAQ: AMD) announced the availability of the AMD Instinct™ MI300X accelerators – with industry leading memory bandwidth for generative AI[1] and leadership performance for large language model (LLM) training and inferencing – as well as the AMD Instinct™ MI300A accelerated processing unit (APU) – combining the latest AMD CDNA™ 3 architecture and "Zen 4" CPUs to deliver breakthrough performance for HPC and AI workloads.

"AMD Instinct MI300 Series accelerators are designed with our most advanced technologies, delivering leadership performance, and will be in large scale cloud and enterprise deployments," said Victor Peng, president, AMD. "By leveraging our leadership hardware, software and open ecosystem approach, cloud providers, OEMs and ODMs are bringing to market technologies that empower enterprises to adopt and deploy AI-powered solutions."

Customers leveraging the latest AMD Instinct accelerator portfolio include Microsoft, which recently announced the new Azure ND MI300x v5 Virtual Machine (VM) series, optimized for AI workloads and powered by AMD Instinct MI300X accelerators. Additionally, El Capitan – a supercomputer powered by AMD Instinct MI300A APUs and housed at Lawrence Livermore National Laboratory – is expected to be the second exascale-class supercomputer powered by AMD and expected to deliver more than two exaflops of double precision performance when fully deployed. Oracle Cloud Infrastructure plans to add AMD Instinct MI300X-based bare metal instances to the company's high-performance accelerated computing instances for AI. MI300X-based instances are planned to support OCI Supercluster with ultrafast RDMA networking.

Several major OEMs also showcased accelerated computing systems, in tandem with the AMD Advancing AI event. Dell showcased the Dell PowerEdge XE9680 server featuring eight AMD Instinct MI300 Series accelerators and the new Dell Validated Design for Generative AI with AMD ROCm-powered AI frameworks. HPE recently announced the HPE Cray Supercomputing EX255a, the first supercomputing accelerator blade powered by AMD Instinct MI300A APUs, which will become available in early 2024. Lenovo announced its design support for the new AMD Instinct MI300 Series accelerators with planned availability in the first half of 2024. Supermicro announced new additions to its H13 generation of

accelerated servers powered by 4th Gen AMD EPYC™ CPUs and AMD Instinct MI300 Series accelerators.

AMD Instinct MI300X

AMD Instinct MI300X accelerators are powered by the new AMD CDNA 3 architecture. When compared to previous generation AMD Instinct MI250X accelerators, MI300X delivers nearly 40% more compute units[2], 1.5x more memory capacity, 1.7x more peak theoretical memory bandwidth[3] as well as support for new math formats such as FP8 and sparsity; all geared towards AI and HPC workloads.

Today's LLMs continue to increase in size and complexity, requiring massive amounts of memory and compute. AMD Instinct MI300X accelerators feature a best-in-class 192 GB of HBM3 memory capacity as well as 5.3 TB/s peak memory bandwidth[2] to deliver the performance needed for increasingly demanding AI workloads. The AMD Instinct Platform is a leadership generative AI platform built on an industry standard OCP design with eight MI300X accelerators to offer an industry leading 1.5TB of HBM3 memory capacity. The AMD Instinct Platform's industry standard design allows OEM partners to design-in MI300X accelerators into existing AI offerings and simplify deployment and accelerate adoption of AMD Instinct accelerator-based servers.

Compared to the Nvidia H100 HGX, the AMD Instinct Platform can offer a throughput increase of up to 1.6x when running inference on LLMs like BLOOM 176B[4] and is the only option on the market capable of running inference for a 70B parameter model, like Llama2, on a single MI300X accelerator; simplifying enterprise-class LLM deployments and enabling outstanding TCO.

AMD Instinct MI300A

The AMD Instinct MI300A APUs, the world's first data center APU for HPC and AI, leverage 3D packaging and the 4th Gen AMD Infinity Architecture to deliver leadership performance on critical workloads sitting at the convergence of HPC and AI. MI300A APUs combine high-performance AMD CDNA 3 GPU cores, the latest AMD "Zen 4" x86-based CPU cores and 128GB of next-generation HBM3 memory, to deliver ~1.9x the performance-per-watt on FP32 HPC and AI workloads, compared to previous gen AMD Instinct MI250X[5].

Energy efficiency is of utmost importance for the HPC and AI communities, however these workloads are extremely data- and resource-intensive. AMD Instinct MI300A APUs benefit from integrating CPU and GPU cores on a single package delivering a highly efficient platform while also providing the compute performance to accelerate training the latest AI models. AMD is setting the pace of innovation in energy efficiency with the company's 30x25 goal, aiming to deliver a 30x energy efficiency improvement in server processors and accelerators for AI-training and HPC from 2020-2025[6].

The APU advantage means that AMD Instinct MI300A APUs feature unified memory and cache resources giving customers an easily programmable GPU platform, highly performant compute, fast AI training and impressive energy efficiency to power the most demanding HPC and AI workloads.

ROCm Software and Ecosystem Partners

AMD announced the latest AMD ROCm™ 6 open software platform as well as the company's commitment to contribute state-of-the-art libraries to the open-source community, furthering the company's vision of open-source AI software development. ROCm 6 software represents a significant leap forward for AMD software tools, increasing AI acceleration performance by ~8x when running on MI300 Series accelerators in Llama 2 text generation compared to previous generation hardware and software[7]. Additionally, ROCm 6 adds support for several new key features for generative AI including FlashAttention, HIPGraph and vLLM, among others. As such, AMD is uniquely positioned to leverage the most broadly used open-source AI software models, algorithms and frameworks – such as Hugging Face, PyTorch, TensorFlow and others – driving innovation, simplifying the deployment of AMD AI solutions and unlocking the true potential of generative AI.

AMD also continues to invest in software capabilities through the acquisitions of Nod.AI and Mipsology as well as through strategic ecosystem partnerships such as Lamini – running LLMs for enterprise customers – and MosaicML – leveraging AMD ROCm to enable LLM training on AMD Instinct accelerators with zero code changes.

Product Specifications

AMD Instinct™	Architecture	GPU CUs	CPU Cores	Memory	Memory Bandwidth (Peak theoretical)	Process Node	3D Packaging w/ 4th Gen AMD Infinity Architecture
MI300A	AMD CDNA™ 3	228	24 "Zen 4"	128GB HBM3	5.3 TB/s	5nm / 6nm	Yes
MI300X	AMD CDNA™ 3	304	N/A	192GB HBM3	5.3 TB/s	5nm / 6nm	Yes
Platform	AMD CDNA™ 3	2,432	N/A	1.5 TB HMB3	5.3 TB/s per OAM	5nm / 6nm	Yes

Supporting Resources

- Watch the full AMD Advancing AI Keynote
- Learn more about AMD Instinct Accelerators
- Follow AMD on X
- Connect with AMD on LinkedIn

About AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies. Billions of people, leading Fortune 500 businesses and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn and X pages.

AMD, the AMD Arrow logo, AMD Instinct, ROCm, EPYC and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names are for informational purposes only and may be trademarks of their respective owners.

CAUTIONARY STATEMENT

This press release contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD Instinct™ MI300X accelerators; AMD Instinct™ MI300A APUs; EI

Capitan, a super-computer powered by AMD Instinct™ MI300 accelerators; AMD's 30x from 2020-2025 energy efficiency goal; AMD Instinct™ platform; AMD Instinct MI300X based-bare metal instances; ROCm™ open software platform, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this press release are based on current beliefs, assumptions and expectations, speak only as of the date of this press release and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Material factors that could cause actual results to differ materially from current expectations include, without limitation, the following: Intel Corporation's dominance of the microprocessor market and its aggressive business practices; economic uncertainty; cyclical nature of the semiconductor industry; market conditions of the industries in which AMD products are sold; loss of a significant customer; impact of the COVID-19 pandemic on AMD's business, financial condition and results of operations; competitive markets in which AMD's products are sold; quarterly and seasonal sales patterns; AMD's ability to adequately protect its technology or other intellectual property; unfavorable currency exchange rate fluctuations; ability of third party manufacturers to manufacture AMD's products on a timely basis in sufficient quantities and using competitive technologies; availability of essential equipment, materials, substrates or manufacturing processes; ability to achieve expected manufacturing yields for AMD's products; AMD's ability to introduce products on a timely basis with expected features and performance levels; AMD's ability to generate revenue from its semi-custom SoC products; potential security vulnerabilities; potential security incidents including IT outages, data loss, data breaches and cyber-attacks; potential difficulties in operating AMD's newly upgraded enterprise resource planning system; uncertainties involving the ordering and shipment of AMD's products; AMD's reliance on third-party intellectual property to design and introduce new products in a timely manner; AMD's reliance on third-party companies for design, manufacture and supply of motherboards, software, memory and other computer platform components; AMD's reliance on Microsoft and other software vendors' support to design and develop software to run on AMD's products; AMD's reliance on third-party distributors and add-in-board partners; impact of modification or interruption of AMD's internal business processes and information systems; compatibility of AMD's products with some or all industry-standard software and hardware; costs related to defective products; efficiency of AMD's supply chain; AMD's ability to rely on third party supply-chain logistics functions; AMD's ability to effectively control sales of its products on the gray market; impact of government actions and regulations such as export regulations, tariffs and trade protection measures; AMD's ability to realize its deferred tax assets; potential tax liabilities; current and future claims and litigation; impact of environmental laws, conflict minerals-related provisions and other laws or regulations; impact of acquisitions, joint ventures and/or investments on AMD's business and AMD's ability to integrate acquired businesses; impact of any impairment of the combined company's assets; restrictions imposed by agreements governing AMD's notes, the guarantees of Xilinx's notes and the revolving credit facility; AMD's indebtedness; AMD's ability to generate sufficient cash to meet its working capital requirements or generate sufficient revenue and operating cash flow to make all of its planned R&D or strategic investments; political, legal and economic risks and natural disasters; future impairments of technology license purchases; AMD's ability to attract and retain qualified personnel; and AMD's stock price volatility. Investors are urged to

review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

---

[1] MI300-05A: Calculations conducted by AMD Performance Labs as of November 17, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8).

The highest published results on the NVidia Hopper H200 (141GB) SXM GPU accelerator resulted in 141GB HBM3e memory capacity and 4.8 TB/s GPU memory bandwidth performance.
https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446

The highest published results on the NVidia Hopper H100 (80GB) SXM5 GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance.

https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

[2] MI300-15: The AMD Instinct™ MI300X (750W) accelerator has 304 compute units (CUs), 19,456 stream cores, and 1,216 Matrix cores.
The AMD Instinct™ MI250 (560W) accelerators have 208 compute units (CUs), 13,312 stream cores, and 832 Matrix cores.
The AMD Instinct™ MI250X (500W/560W) accelerators have 220 compute units (CUs), 14,080 stream cores, and 880 Matrix cores.

[3] MI300-13: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8).

The AMD Instinct™ MI250 (500W) / MI250X (560W) OAM accelerators (128 GB HBM2e) designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 3.277 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 8,192 (4,096 bits times 2 die) and memory data rate is 3.20 Gbps for total memory bandwidth of 3.277 TB/s ((3.20 Gbps*(4,096 bits*2))/8).

[4] MI300-34: Token generation throughput using DeepSpeed Inference with the Bloom-176b model with an input sequence length of 1948 tokens, and output sequence length of 100 tokens, and a batch size tuned to yield the highest throughput on each system comparison based on AMD internal testing using custom docker container for each system as of 11/17/2023.
Configurations:
2P Intel Xeon Platinum 8480C CPU powered server with 8x AMD Instinct™ MI300X 192GB 750W GPUs, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2.

Vs.
An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 80GB 700W GPUs, CUDA 12.0, Ubuntu 22.04.3.
8 GPUs on each system were used in this test.
Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

[5] MI300-23: Calculations conducted by AMD Performance Labs as of Nov 16, 2023, for the AMD Instinct™ MI300X (192GB HBM3 OAM Module) 750W accelerator designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.43 TFLOPS peak theoretical single precision (FP32) floating-point performance.
The AMD Instinct™ MI300A (128GB HBM3 APU) 760W accelerator designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 122.573 TFLOPS peak theoretical single precision (FP32) floating-point performance.
The AMD Instinct™ MI250X (128GB HBM2e OAM module) 560W accelerator designed with AMD CDNA™ 2 6nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in 47.9 TFLOPS peak theoretical single precision (FP32) floating-point performance.

[6] Includes AMD high-performance CPU and GPU accelerators used for AI training and high-performance computing in a 4-Accelerator, CPU-hosted configuration. Goal calculations are based on performance scores as measured by standard performance metrics (HPC: Linpack DGEMM kernel FLOPS with 4k matrix size. AI training: lower precision training-focused floating-point math GEMM kernels such as FP16 or BF16 FLOPS operating on 4k matrices) divided by the rated power consumption of a representative accelerated compute node, including the CPU host + memory and 4 GPU accelerators.
[7] MI300-33: Text generated with Llama2-70b chat using input sequence length of 4096 and 32 output token comparison using custom docker container for each system based on AMD internal testing as of 11/17/2023.
Configurations:
2P Intel Xeon Platinum CPU server using 4x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu® 22.04.2.
Vs.
2P AMD EPYC 7763 CPU server using 4x AMD Instinct™ MI250 (128 GB HBM2e, 560W) GPUs, ROCm® 5.4.3, PyTorch 2.0.0., HuggingFace Transformers 4.35.0, Ubuntu 22.04.6.
4 GPUs on each system was used in this test.
Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

A photo accompanying this announcement is available at https://www.globenewswire.com/NewsRoom/AttachmentNg/793bf8fd-eec4-4460-bec2-cc6620bdd6c7

Contact:
Aaron Grabein
AMD Communications
(512) 602-8950

Aaron.grabein@amd.com

Suresh Bhaskaran
AMD Investor Relations
(408) 749-2845
Suresh.Bhaskaran@amd.com

Source: Advanced Micro Devices, Inc.

AMD Instinct MI300 Series



AMD Instinct MI300X