June 2, 2024

AMD Accelerates Pace of Data Center AI Innovation and Leadership with Expanded AMD Instinct GPU Roadmap

— Updated AMD Instinct accelerator roadmap brings annual cadence of leadership AI performance and memory capabilities —

— New AMD Instinct MI325X accelerator expected to be available in Q4 2024 with up to 288GB of HBM3E memory; new AMD Instinct MI350 series accelerators based on AMD CDNA 4 architecture expected to be available in 2025 with 35x generational increase in AI inference performance —

TAIPEI, Taiwan, June 02, 2024 (GLOBE NEWSWIRE) -- At Computex 2024, AMD (NASDAQ: AMD) showcased the growing momentum of the AMD Instinct™ accelerator family during the opening keynote by Chair and CEO Dr. Lisa Su. AMD unveiled a multiyear, expanded AMD Instinct accelerator roadmap which will bring an annual cadence of leadership AI performance and memory capabilities at every generation.

The updated roadmap starts with the new AMD Instinct MI325X accelerator, which will be available in Q4 2024. Following that, the AMD Instinct MI350 series, powered by the new AMD CDNA™ 4 architecture, is expected to be available in 2025 bringing up to a 35x increase in AI inference performance compared to AMD Instinct MI300 Series with AMD CDNA 3 architecture[1]. Expected to arrive in 2026, the AMD Instinct MI400 series is based on the AMD CDNA "Next" architecture.

"The AMD Instinct MI300X accelerators continue their strong adoption from numerous partners and customers including Microsoft Azure, Meta, Dell Technologies, HPE, Lenovo and others, a direct result of the AMD Instinct MI300X accelerator exceptional performance and value proposition," said Brad McCredie, corporate vice president, Data Center Accelerated Compute, AMD. "With our updated annual cadence of products, we are relentless in our pace of innovation, providing the leadership capabilities and performance the AI industry and our customers expect to drive the next evolution of data center AI training and inference."

AMD AI Software Ecosystem Matures
The AMD ROCm™ 6 open software stack continues to mature, enabling AMD Instinct MI300X accelerators to drive impressive performance for some of the most popular LLMs. On a server using eight AMD Instinct MI300X accelerators and ROCm 6 running Meta Llama-3 70B, customers can get 1.3x better inference performance and token generation compared to the competition[2]. On a single AMD Instinct MI300X accelerator with ROCm 6, customers can get better inference performance and token generation throughput compared to the competition by 1.2x on Mistral-7B[3]. AMD also highlighted that Hugging Face, the largest and most popular repository for AI models, is now testing 700,000 of their most popular models nightly to ensure they work out of box on AMD Instinct MI300X accelerators. In addition, AMD is continuing its upstream work into popular AI frameworks like PyTorch, TensorFlow and JAX.

AMD Previews New Accelerators and Reveals Annual Cadence Roadmap
During the keynote, AMD revealed an updated annual cadence for the AMD Instinct accelerator roadmap to meet the growing demand for more AI compute. This will help ensure that AMD Instinct accelerators propel the development of next-generation frontier AI models. The updated AMD Instinct annual roadmap highlighted:

- The new AMD Instinct MI325X accelerator, which will bring 288GB of HBM3E memory and 6 terabytes per second of memory bandwidth, use the same industry standard Universal Baseboard server design used by the AMD Instinct MI300 series, and be generally available in Q4 2024. The accelerator will have industry leading memory capacity and bandwidth, 2x and 1.3x better than the competition respectively[4], and 1.3x better[5] compute performance than competition.
- The first product in the AMD Instinct MI350 Series, the AMD Instinct MI350X accelerator, is based on the AMD CDNA 4 architecture and is expected to be available in 2025. It will use the same industry standard Universal Baseboard server design as other MI300 Series accelerators and will be built using advanced 3nm process technology, support the FP4 and FP6 AI datatypes and have up to 288 GB of HBM3E memory.
- AMD CDNA "Next" architecture, which will power the AMD Instinct MI400 Series accelerators, is expected to be available in 2026 providing the latest features and capabilities that will help unlock additional performance and efficiency for inference and large-scale AI training.

Finally, AMD highlighted the demand for AMD Instinct MI300X accelerators continues to grow with numerous partners and customers using the accelerators to power their demanding AI workloads, including:

- Microsoft Azure using the accelerators for Azure OpenAI services and the new Azure ND MI300X V5 virtual machines.
- Dell Technologies using MI300X accelerators in the PowerEdge XE9680 for enterprise AI workloads.
- Supermicro providing multiple solutions with AMD Instinct accelerators.
- Lenovo powering Hybrid AI innovation with the ThinkSystem SR685a V3
- HPE is using them to accelerate AI workloads in the HPE Cray XD675.

Read more AMD AI announcements at Computex here and watch a video replay of the keynote on the AMD YouTube page.

Supporting Resources

- Follow AMD on LinkedIn
- Follow AMD on X

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.

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 products including the AMD Instinct™ accelerator family, AMD CDNA™ 4 and AMD CDNA™ "Next", product roadmaps, leadership AI performance and growing momentum as well as partner and customer demand, 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; cyclical nature of the semiconductor industry; market conditions of the industries in which AMD products are sold; loss of a significant customer; competitive markets in which AMD's products are sold; economic and market uncertainty; 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 cyberattacks; 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; long-term impact of climate change on AMD's business; 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; evolving expectations from governments, investors, customers and other stakeholders regarding corporate responsibility matters; issues related to the responsible use of AI; restrictions imposed by agreements governing AMD's notes, the guarantees of Xilinx's notes and the revolving credit facility; 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; 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]MI300-55: Inference performance projections as of May 31, 2024 using engineering estimates based on the design of a future AMD CDNA 4-based Instinct MI350 Series accelerator as proxy for projected AMD CDNA™ 4 performance. A 1.8T GPT MoE model was evaluated assuming a token-to-token latency = 70ms real time, first token latency = 5s, input sequence length = 8k, output sequence length = 256, assuming a 4x 8-mode MI350 series proxy (CDNA4) vs. 8x MI300X per GPU performance comparison. Actual performance will vary based on factors including but not limited to final specifications of production silicon, system configuration and inference model and size used.

[2] MI300-54: Testing completed on 05/28/2024 by AMD performance lab attempting text generated Llama3-70B using batch size 1 and 2048 input tokens and 128 output tokens for each system.

Configurations:
2P AMD EPYC 9534 64-Core Processor based production server with 8x AMD InstinctTM MI300X (192GB, 750W) GPU, Ubuntu® 22.04.1, and ROCm™ 6.1.1
Vs.
2P Intel Xeon Platinum 8468 48-Core Processor based production server with 8x NVIDIA Hopper H100 (80GB, 700W) GPU, Ubuntu 22.04.3, and CUDA® 12.2
8 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.
[3] MI300-53: Testing completed on 05/28/2024 by AMD performance lab attempting text generated throughput measured using Mistral-7B model comparison.
Tests were performed using batch size 56 and 2048 input tokens and 2048 output tokens for Mistral-7B

Configurations:
2P AMD EPYC 9534 64-Core Processor based production server with 8x AMD InstinctTM MI300X (192GB, 750W) GPU, Ubuntu® 22.04.1, and ROCm™ 6.1.1
Vs.
2P Intel Xeon Platinum 8468 48-Core Processor based production server with 8x NVIDIA Hopper H100 (80GB, 700W) GPU, Ubuntu 22.04.3, and CUDA® 12.2

Only 1 GPU 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.

[4]MI300-48 - Calculations conducted by AMD Performance Labs as of May 22nd, 2024, based on current specifications and /or estimation. The AMD Instinct™ MI325X OAM accelerator is projected to have 288GB HBM3e memory capacity and 6 TFLOPS peak theoretical memory bandwidth performance. Actual results based on production silicon may vary.

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 Blackwell HGX B100 (192GB) 700W GPU accelerator resulted in 192GB HBM3e memory capacity and 8 TB/s GPU memory bandwidth performance.

https://resources.nvidia.com/en-us-blackwell-architecture?_gl=1*1r4pme7*_gcl_aw*R0NMLjE3MTM5NjQ3NTAuQ2p3S0NBancyNkt4weknow
QmhCREVpd0F1NktYdDlweXY1dlUtaHNKNmhPdHM4UVdPSlM3dFdQaE40WkI4THZBaWFVajFyTGhYd3hLQmlZQ3pCb0NsVElRQXZEX0J

The highest published results on the NVidia Blackwell HGX B200 (192GB) GPU accelerator resulted in 192GB HBM3e memory capacity and 8 TB/s GPU memory bandwidth performance.

https://resources.nvidia.com/en-us-blackwell-architecture?_gl=1*1r4pme7*_gcl_aw*R0NMLjE3MTM5NjQ3NTAuQ2p3S0NBancyNkt4QmhCREVpd0F1NktYdDlweXY1dlUtaHNKNmhPdHM4UVdPSlM3d

[5]MI300-49: Calculations conducted by AMD Performance Labs as of May 28th, 2024 for the AMD Instinct™ MI325X GPU resulted in 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance. Actual performance will vary based on final specifications and system configuration.

Published results on Nvidia H200 SXM (141GB) GPU: 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance. BFLOAT16 Tensor Core, FP16 Tensor Core, FP8 Tensor Core and INT8 Tensor Core performance were published by Nvidia using sparsity; for the purposes of comparison, AMD converted these numbers to non-sparsity/dense by dividing by 2, and these numbers appear above.

Nvidia H200 source: https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446 and https://www.anandtech.com/show/21136/nvidia-at-sc23-h200-accelerator-with-hbm3e-and-jupiter-supercomputer-for-2024

Note: Nvidia H200 GPUs have the same published FLOPs performance as H100 products https://resources.nvidia.com/en-us-tensor-core/

Contact:
Aaron Grabein
AMD Communications
+1 (737) 256-9518
Aaron.Grabein@amd.com

Suresh Bhaskaran
AMD Investor Relations
+1 (408) 749-2845
Suresh.Bhaskaran@amd.com

Source: Advanced Micro Devices, Inc.