October 10, 2024

AMD Delivers Leadership AI Performance with AMD Instinct MI325X Accelerators

— Latest accelerators offer market leading HBM3E memory capacity and are supported by partners and customers including Dell Technologies, HPE, Lenovo, Supermicro and others —

— AMD Pensando Salina DPU offers 2X generational performance and AMD Pensando Pollara 400 is industry's first UEC ready NIC—

SAN FRANCISCO, Oct. 10, 2024 (GLOBE NEWSWIRE) -- Today, AMD (NASDAQ: AMD) announced the latest accelerator and networking solutions that will power the next generation of AI infrastructure at scale: AMD Instinct™ MI325X accelerators, the AMD Pensando™ Pollara 400 NIC and the AMD Pensando Salina DPU. AMD Instinct MI325X accelerators set a new standard in performance for Gen AI models and data centers.

Built on the AMD CDNA™ 3 architecture, AMD Instinct MI325X accelerators are designed for exceptional performance and efficiency for demanding AI tasks spanning foundation model training, fine-tuning and inferencing. Together, these products enable AMD customers and partners to create highly performant and optimized AI solutions at the system, rack and data center level.

"AMD continues to deliver on our roadmap, offering customers the performance they need and the choice they want, to bring AI infrastructure, at scale, to market faster," said Forrest Norrod, executive vice president and general manager, Data Center Solutions Business Group, AMD. "With the new AMD Instinct accelerators, EPYC processors and AMD Pensando networking engines, the continued growth of our open software ecosystem, and the ability to tie this all together into optimized AI infrastructure, AMD underscores the critical expertise to build and deploy world class AI solutions."

AMD Instinct MI325X Extends Leading AI Performance
AMD Instinct MI325X accelerators deliver industry-leading memory capacity and bandwidth, with 256GB of HBM3E supporting 6.0TB/s offering 1.8X more capacity and 1.3x more bandwidth than the H200[1]. The AMD Instinct MI325X also offers 1.3X greater peak theoretical FP16 and FP8 compute performance compared to H200[1].

This leadership memory and compute can provide up to 1.3X the inference performance on Mistral 7B at FP16[2], 1.2X the inference performance on Llama 3.1 70B at FP8[3] and 1.4X the inference performance on Mixtral 8x7B at FP16 of the H200[4].

AMD Instinct MI325X accelerators are currently on track for production shipments in Q4 2024 and are expected to have widespread system availability from a broad set of platform providers, including Dell Technologies, Eviden, Gigabyte, Hewlett Packard Enterprise, Lenovo, Supermicro and others starting in Q1 2025.

Continuing its commitment to an annual roadmap cadence, AMD previewed the next-generation AMD Instinct MI350 series accelerators. Based on AMD CDNA 4 architecture, AMD Instinct MI350 series accelerators are designed to deliver a 35x improvement in inference performance compared to AMD CDNA 3-based accelerators[5].

The AMD Instinct MI350 series will continue to drive memory capacity leadership with up to 288GB of HBM3E memory per accelerator. The AMD Instinct MI350 series accelerators are on track to be available during the second half of 2025.

AMD Next-Gen AI Networking
AMD is leveraging the most widely deployed programmable DPU for hyperscalers to power next-gen AI networking. Split into two parts: the front-end, which delivers data and information to an AI cluster, and the backend, which manages data transfer between accelerators and clusters, AI networking is critical to ensuring CPUs and accelerators are utilized efficiently in AI infrastructure.

To effectively manage these two networks and drive high performance, scalability and efficiency across the entire system, AMD introduced the AMD Pensando™ Salina DPU for the front-end and the AMD Pensando™ Pollara 400, the industry's first Ultra Ethernet Consortium (UEC) ready AI NIC, for the back-end.

The AMD Pensando Salina DPU is the third generation of the world's most performant and programmable DPU, bringing up to 2X the performance, bandwidth and scale compared to the previous generation. Supporting 400G throughput for fast data transfer rates, the AMD Pensando Salina DPU is a critical component in AI front-end network clusters, optimizing performance, efficiency, security and scalability for data-driven AI applications.

The UEC-ready AMD Pensando Pollara 400, powered by the AMD P4 Programmable engine, is the industry's first UEC-ready AI NIC. It supports the next-gen RDMA software and is backed by an open ecosystem of networking. The AMD Pensando Pollara 400 is critical for providing leadership performance, scalability and efficiency of accelerator-to-accelerator communication in back-end networks.

Both the AMD Pensando Salina DPU and AMD Pensando Pollara 400 are sampling with customers in Q4'24 and are on track for availability in the first half of 2025.

AMD AI Software Delivering New Capabilities for Generative AI
AMD continues its investment in driving software capabilities and the open ecosystem to deliver powerful new features and capabilities in the AMD ROCm™ open software stack.

Within the open software community, AMD is driving support for AMD compute engines in the most widely used AI frameworks, libraries and models including PyTorch, Triton, Hugging Face and many others. This work translates to out-of-the-box performance and support with AMD Instinct accelerators on popular generative AI models like Stable Diffusion 3, Meta Llama 3, 3.1 and 3.2 and more than one million models at Hugging Face.

Beyond the community, AMD continues to advance its ROCm open software stack, bringing the latest features to support leading training and inference on Generative AI workloads. ROCm 6.2 now includes support for critical AI features like FP8 datatype, Flash Attention 3, Kernel Fusion and more. With these new additions, ROCm 6.2, compared to ROCm 6.0, provides up to a 2.4X performance improvement on inference[6] and 1.8X on training for a variety of LLMs[7].

Supporting Resources

- Follow AMD on LinkedIn
- Follow AMD on Twitter
- Read more about AMD Next Generation AI Networking here
- Read more about AMD Instinct Accelerators here
- Visit the AMD Advancing AI: 2024 event page

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™ MI325X accelerators; AMD Pensando™ Salina DPU; AMD Pensando Pollara 400; continued growth of AMD's open software ecosystem; AMD Instinct MI350 series accelerators, 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; Nvidia's dominance in the graphics processing unit market and its aggressive business practices; the 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; 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 agreement; 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.

AMD, the AMD Arrow logo, AMD CDNA, AMD Instinct, Pensando, ROCm, 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.

---

[1]MI325-002 -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/.

[2] Based on testing completed on 9/28/2024 by AMD performance lab measuring overall latency for Mistral-7B model using FP16 datatype. Test was performed using input length of

128 tokens and an output length of 128 tokens for the following configurations of AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator.

1x MI325X at 1000W with vLLM performance: 0.637 sec (latency in seconds)
Vs.
1x H200 at 700W with TensorRT-LLM: 0.811 sec (latency in seconds)

Configurations:
AMD Instinct™ MI325X reference platform:
1x AMD Ryzen™ 9 7950X 16-Core Processor CPU, 1x AMD Instinct MI325X (256GiB, 1000W) GPU, Ubuntu® 22.04, and ROCm™ 6.3 pre-release
Vs
NVIDIA H200 HGX platform:
Supermicro SuperServer with 2x Intel Xeon® Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs [only 1 GPU was used in this test], Ubuntu 22.04), CUDA 12.6 Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-005

[3] MI325-006: Based on testing completed on 9/28/2024 by AMD performance lab measuring overall latency for LLaMA 3.1-70B model using FP8 datatype. Test was performed using input length of 2048 tokens and an output length of 2048 tokens for the following configurations of AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator.

1x MI325X at 1000W with vLLM performance: 48.025 sec (latency in seconds)
Vs.
1x H200 at 700W with TensorRT-LLM: 62.688 sec (latency in seconds)

Configurations:
AMD Instinct™ MI325X reference platform:
1x AMD Ryzen™ 9 7950X 16-Core Processor CPU, 1x AMD Instinct MI325X (256GiB, 1000W) GPU, Ubuntu® 22.04, and ROCm™ 6.3 pre-release
Vs
NVIDIA H200 HGX platform:
Supermicro SuperServer with 2x Intel Xeon® Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs, Ubuntu 22.04), CUDA 12.6

Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

[4] MI325-004: Based on testing completed on 9/28/2024 by AMD performance lab measuring text generated throughput for Mixtral-8x7B model using FP16 datatype. Test was performed using input length of 128 tokens and an output length of 4096 tokens for the following configurations of AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator.

1x MI325X at 1000W with vLLM performance: 4598 (Output tokens / sec)
Vs.
1x H200 at 700W with TensorRT-LLM: 2700.7 (Output tokens / sec)

Configurations:
AMD Instinct™ MI325X reference platform:

1x AMD Ryzen™ 9 7950X CPU, 1x AMD Instinct MI325X (256GiB, 1000W) GPU, Ubuntu® 22.04, and ROCm™ 6.3 pre-release
Vs
NVIDIA H200 HGX platform:
Supermicro SuperServer with 2x Intel Xeon® Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs [only 1 GPU was used in this test], Ubuntu 22.04) CUDA® 12.6

Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

[5] CDNA4-03: 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.

[6] MI300-62: Testing conducted by internal AMD Performance Labs as of September 29, 2024 inference performance comparison between ROCm 6.2 software and ROCm 6.0 software on the systems with 8 AMD Instinct™ MI300X GPUs coupled with Llama 3.1-8B, Llama 3.1-70B, Mixtral-8x7B, Mixtral-8x22B, and Qwen 72B models.

ROCm 6.2 with vLLM 0.5.5 performance was measured against the performance with ROCm 6.0 with vLLM 0.3.3, and tests were performed across batch sizes of 1 to 256 and sequence lengths of 128 to 2048.

Configurations:
1P AMD EPYC™ 9534 CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA per socket), 1.5 TiB (24 DIMMs, 4800 mts memory, 64 GiB/DIMM), 4x 3.49TB Micron 7450 storage, BIOS version: 1.8, , ROCm 6.2.0-00, vLLM 0.5.5, PyTorch 2.4.0, Ubuntu® 22.04 LTS with Linux kernel 5.15.0-119-generic.
vs.
1P AMD EPYC 9534 CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA per socket), 1.5TiB 24 DIMMs, 4800 mts memory, 64 GiB/DIMM), 4x 3.49TB Micron 7450 storage, BIOS version: 1.8, ROCm 6.0.0-00, vLLM 0.3.3, PyTorch 2.1.1, Ubuntu 22.04 LTS with Linux kernel 5.15.0-119-generic.

Server manufacturers may vary configurations, yielding different results. Performance may vary based on factors including but not limited to different versions of configurations, vLLM, and drivers.

[7] MI300-61: Measurements conducted by AMD AI Product Management team on AMD Instinct™ MI300X GPU for comparing large language model (LLM) performance with optimization methodologies enabled and disabled as of 9/28/2024 on Llama 3.1-70B and Llama 3.1-405B and vLLM 0.5.5.

System Configurations:
- AMD EPYC 9654 96-Core Processor, 8 x AMD MI300X, ROCm™ 6.1, Linux® 7ee7e017abe3 5.15.0-116-generic #126-Ubuntu® SMP Mon Jul 1 10:14:24 UTC 2024 x86_64 x86_64 x86_64 GNU/Linux, Frequency boost: enabled.

Performance may vary on factors including but not limited to different versions of configurations, vLLM, and drivers.

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

Mitch Haws
AMD Investor Relations
+1 512-944-0790
mitch.haws@amd.com



Source: Advanced Micro Devices, Inc.