June 13, 2023

AMD Expands Leadership Data Center Portfolio with New EPYC CPUs and Shares Details on Next-Generation AMD Instinct Accelerator and Software Enablement for Generative AI

— AMD unleashes the power of specialized compute for the data center with new AMD EPYC processors for cloud native and technical computing —

—AMD reveals details on next-generation AMD Instinct products for generative AI and highlights AI software ecosystem collaborations with Hugging Face and PyTorch —

SANTA CLARA, Calif., June 13, 2023 (GLOBE NEWSWIRE) -- Today, at the "Data Center and AI Technology Premiere," AMD (NASDAQ: AMD) announced the products, strategy and ecosystem partners that will shape the future of computing, highlighting the next phase of data center innovation. AMD was joined on stage with executives from Amazon Web Services (AWS), Citadel, Hugging Face, Meta, Microsoft Azure and PyTorch to showcase the technological partnerships with industry leaders to bring the next generation of high performance CPU and AI accelerator solutions to market.

"Today, we took another significant step forward in our data center strategy as we expanded our 4th Gen EPYC™ processor family with new leadership solutions for cloud and technical computing workloads and announced new public instances and internal deployments with the largest cloud providers," said AMD Chair and CEO Dr. Lisa Su. "AI is the defining technology shaping the next generation of computing and the largest strategic growth opportunity for AMD. We are laser focused on accelerating the deployment of AMD AI platforms at scale in the data center, led by the launch of our Instinct MI300 accelerators planned for later this year and the growing ecosystem of enterprise-ready AI software optimized for our hardware."

Compute Infrastructure Optimized for The Modern Data Center

AMD unveiled a series of updates to its 4th Gen EPYC family, designed to offer customers the workload specialization needed to address businesses' unique needs.

- **Advancing the World's Best Data Center CPU**. AMD highlighted how the 4th Gen AMD EPYC processor continues to drive leadership performance and energy efficiency. AMD was joined by AWS to highlight a preview of the next generation Amazon Elastic Compute Cloud (Amazon EC2) M7a instances, powered by 4th Gen AMD EPYC processors ("Genoa"). Outside of the event, Oracle announced plans to make available new Oracle Computing Infrastructure (OCI) E5 instances with 4th Gen AMD EPYC processors.

- **No Compromise Cloud Native Computing.** AMD introduced the 4th Gen AMD EPYC 97X4 processors, formerly codenamed “Bergamo.” With 128 “Zen 4c” cores per socket, these processors provide the greatest vCPU density[1] and industry leading[2] performance for applications that run in the cloud, and leadership energy efficiency. AMD was joined by Meta who discussed how these processors are well suited for their mainstay applications such as Instagram, WhatsApp and more; how Meta is seeing impressive performance gains with 4th Gen AMD EPYC 97x4 processors compared to 3rd Gen AMD EPYC across various workloads, while offering substantial TCO improvements as well, and how AMD and Meta optimized the EPYC CPUs for Meta’s power-efficiency and compute-density requirements.
- **Enabling Better Products With Technical Computing.** AMD introduced the 4th Gen AMD EPYC processors with AMD 3D V-Cache™ technology, the world’s highest performance x86 server CPU for technical computing[3]. Microsoft announced the general availability of Azure HBv4 and HX instances, powered by 4th Gen AMD EPYC processors with AMD 3D V-Cache technology.

Click here to learn more about the latest 4th Gen AMD EPYC processors and read about what AMD customers have to say, here.

AMD AI Platform – The Pervasive AI Vision

Today, AMD unveiled a series of announcements showcasing its AI Platform strategy, giving customers a cloud, to edge, to endpoint portfolio of hardware products, with deep industry software collaboration, to develop scalable and pervasive AI solutions.

- **Introducing the World’s Most Advanced Accelerator for Generative AI[4].** AMD revealed new details of the AMD Instinct™ MI300 Series accelerator family, including the introduction of the AMD Instinct MI300X accelerator, the world’s most advanced accelerator for generative AI. The MI300X is based on the next-gen AMD CDNA™ 3 accelerator architecture and supports up to 192 GB of HBM3 memory to provide the compute and memory efficiency needed for large language model training and inference for generative AI workloads. With the large memory of AMD Instinct MI300X, customers can now fit large language models such as Falcon-40, a 40B parameter model on a single, MI300X accelerator[5]. AMD also introduced the AMD Instinct™ Platform, which brings together eight MI300X accelerators into an industry-standard design for the ultimate solution for AI inference and training. The MI300X is sampling to key customers starting in Q3. AMD also announced that the AMD Instinct MI300A, the world’s first APU Accelerator for HPC and AI workloads, is now sampling to customers.
- **Bringing an Open, Proven and Ready AI Software Platform to Market.** AMD showcased the ROCm™ software ecosystem for data center accelerators, highlighting the readiness and collaborations with industry leaders to bring together an open AI software ecosystem. PyTorch discussed the work between AMD and the PyTorch Foundation to fully upstream the ROCm software stack, providing immediate “day zero” support for PyTorch 2.0 with ROCm release 5.4.2 on all AMD Instinct accelerators. This integration empowers developers with an extensive array of AI models powered by PyTorch that are compatible and ready to use “out of the box” on AMD accelerators. Hugging Face, the leading open platform for AI builders, announced that it will optimize thousands of Hugging Face models on AMD platforms, from AMD Instinct accelerators to AMD Ryzen™ and AMD EPYC processors, AMD Radeon™ GPUs and Versal™ and Alveo™ adaptive processors.

A Robust Networking Portfolio for the Cloud and Enterprise
AMD showcased a robust networking portfolio including the AMD Pensando™ DPU, AMD Ultra Low Latency NICs and AMD Adaptive NICs. Additionally, AMD Pensando DPUs combine a robust software stack with "zero trust security" and leadership programmable packet processor to create the world's most intelligent and performant DPU. The AMD Pensando DPU is deployed at scale across cloud partners such as IBM Cloud, Microsoft Azure and Oracle Compute Infrastructure. In the enterprise it is deployed in the HPE Aruba CX 10000 Smart Switch, and with customers such as leading IT services company DXC, and as part of VMware vSphere® Distributed Services Engine™, accelerating application performance for customers.

AMD highlighted the next generation of its DPU roadmap, codenamed "Giglio," which aims to bring enhanced performance and power efficiency to customers, compared to current generation products, when it's expected to be available by the end of 2023.

AMD also announced the AMD Pensando Software-in-Silicon Developer Kit (SSDK), giving customers the ability to rapidly develop or migrate services to deploy on the AMD Pensando P4 programmable DPU in coordination with the existing rich set of features already implemented on the AMD Pensando platform. The AMD Pensando SSDK enables customers to put the power of the leadership AMD Pensando DPU to work and tailor network virtualization and security features within their infrastructure, in coordination with the existing rich set of features already implemented on the Pensando platform.

Supporting Resources

- Watch the "Data Center and AI Technology Premiere" Keynote Video
- Read more about 4th Gen AMD EPYC Processors
- Read more about AMD Instinct Accelerators
- Read more about AMD Networking Solutions
- Follow AMD on Twitter
- 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 Twitter pages.

AMD, the AMD Arrow logo, EPYC, AMD Instinct, ROCm, Ryzen, Radeon 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 products including, the AMD 4th Gen EPYC™ processor family, the AMD Instinct™ MI300 Series accelerator family, including AMD Instinct™ MI300X and

AMD Instinct™ MI300A, and the AMD Pensando DPU codenamed “Giglio”, 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; global 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 upgrading and operating AMD’s new 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 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 administration 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 on the combined company’s financial position and results of operation; 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, economic risks and natural disasters; future impairments of goodwill and technology license purchases; AMD’s ability to attract and retain qualified personnel; AMD’s stock price volatility; and worldwide political conditions. 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] EPYC-049: AMD EPYC 9754 is a 128 core dual threaded CPU and in a 2 socket server with 1 thread per vCPU delivers 512 vCPUs per EPYC powered server which is more than any Ampere or 4 socket Intel CPU based server as of 05/23/2023.

[2] SP5-143A: SPECrate®2017_int_base comparison based on performing system published scores from www.spec.org as of 6/13/2013. 2P AMD EPYC 9754 scores 1950 SPECrate®2017_int_base http://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36617.html is higher than all other 2P servers. 1P AMD EPYC 9754 scores 981 SPECrate®2017_int_base score (981.4 score/socket) http://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36613.html is higher per socket than all other servers. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

[3] SP5-165: The EPYC 9684X CPU is the world's highest performance x86 server CPU for technical computing, comparison based on SPEC.org publications as of 6/13/2023 measuring the score, rating or jobs/day for each of SPECrate®2017_fp_base (SP5-009E), Altair AcuSolve (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-altair-acusolve.pdf), Ansys Fluent (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-ansys-fluent.pdf), OpenFOAM (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-openfoam.pdf), Ansys LS-Dyna (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-ansys-ls-dyna.pdf), and Altair Radioss (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-altair-radioss.pdf) application test case simulations average speedup on 2P servers running 96-core EPYC 9684X vs top 2P performance general-purpose 56-core Intel Xeon Platinum 8480+ or top-of-stack 60-core Xeon 8490H based server for technical computing performance leadership. "Technical Computing" or "Technical Computing Workloads" as defined by AMD can include: electronic design automation, computational fluid dynamics, finite element analysis, seismic tomography, weather forecasting, quantum mechanics, climate research, molecular modeling, or similar workloads. Results may vary based on factors including silicon version, hardware and software configuration and driver versions. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

[4] MI300-09 - The AMD Instinct™ MI300X accelerator is based on AMD CDNA™ 3 5nm FinFet process technology with 3D chiplet stacking, utilizes high speed AMD Infinity Fabric technology, has 192 GB HBM3 memory capacity (vs. 80GB for Nvidia Hopper H100) with 5.218 TFLOPS of sustained peak memory bandwidth performance, higher than the highest bandwidth Nvidia Hopper H100 GPU.

[5] MI300-07K: Measurements by internal AMD Performance Labs as of June 2, 2023 on current specifications and/or internal engineering calculations. Large Language Model (LLM) run or calculated with FP16 precision to determine the minimum number of GPUs needed to run the Falcon (40B parameter) models. Tested result configurations: AMD Lab system consisting of 1x EPYC 9654 (96-core) CPU with 1x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator tested at FP16 precision. Server manufacturers may vary configuration offerings yielding different results.

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.