October 10, 2024

AMD Unveils Leadership AI Solutions at Advancing AI 2024

— AMD launches 5th Gen AMD EPYC processors, AMD Instinct MI325X accelerators, next gen networking solutions and AMD Ryzen AI PRO processors powering enterprise AI at scale —

— Dell, Google Cloud, HPE, Lenovo, Meta, Microsoft, Oracle Cloud Infrastructure, Supermicro and AI leaders Databricks, Essential AI, Fireworks AI, Luma AI and Reka AI joined AMD to showcase expanding AMD AI solutions for enterprises and end users —

— Technical leaders from Cohere, Google DeepMind, Meta, Microsoft, OpenAI and more discussed how they are using AMD ROCm software to deploy models and applications on AMD Instinct accelerators —

SAN FRANCISCO, Oct. 10, 2024 (GLOBE NEWSWIRE) -- AMD (NASDAQ: AMD) today launched the latest high performance computing solutions defining the AI computing era, including 5th Gen AMD EPYC™ server CPUs, AMD Instinct™ MI325X accelerators, AMD Pensando™ Salina DPUs, AMD Pensando Pollara 400 NICs and AMD Ryzen™ AI PRO 300 series processors for enterprise AI PCs. AMD and its partners also showcased how they are deploying AMD AI solutions at scale, the continued ecosystem growth of AMD ROCm™ open source AI software, and a broad portfolio of new solutions based on AMD Instinct accelerators, EPYC CPUs and Ryzen PRO CPUs.

“The data center and AI represent significant growth opportunities for AMD, and we are building strong momentum for our EPYC and AMD Instinct processors across a growing set of customers,” said AMD Chair and CEO Dr. Lisa Su. “With our new EPYC CPUs, AMD Instinct GPUs and Pensando DPUs we are delivering leadership compute to power our customers’ most important and demanding workloads. Looking ahead, we see the data center AI accelerator market growing to $500 billion by 2028. We are committed to delivering open innovation at scale through our expanded silicon, software, network and cluster-level solutions.”

Defining the Data Center in the AI Era
AMD announced a broad portfolio of data center solutions for AI, enterprise, cloud and mixed workloads:

- New AMD EPYC 9005 Series processors deliver record-breaking performance[1] to enable optimized compute solutions for diverse data center needs. Built on the latest “Zen 5” architecture, the lineup offers up to 192 cores and will be available in a wide range of platforms from leading OEMs and ODMs starting today.
- AMD continues executing its annual cadence of AI accelerators with the launch of AMD Instinct MI325X, delivering leadership performance and memory capabilities for the most demanding AI workloads. AMD also shared new details on next-gen AMD Instinct MI350 series accelerators expected to launch in the second half of 2025, extending

AMD Instinct leadership memory capacity and generative AI performance. AMD has made significant progress developing the AMD Instinct MI400 Series accelerators based on the AMD CDNA Next architecture, planned to be available in 2026.
- AMD has continuously improved its AMD ROCm software stack, doubling AMD Instinct MI300X accelerator inferencing and training performance[2] across a wide range of the most popular AI models. Today, over one million models run seamlessly out of the box on AMD Instinct, triple the number available when MI300X launched, with day-zero support for the most widely used models.
- AMD also expanded its high performance networking portfolio to address evolving system networking requirements for AI infrastructure, maximizing CPU and GPU performance to deliver performance, scalability and efficiency across the entire system. The AMD Pensando Salina DPU delivers a high performance front-end network for AI systems, while the AMD Pensando Pollara 400, the first Ultra Ethernet Consortium ready NIC, reduces the complexity of performance tuning and helps improve time to production.

AMD partners detailed how they leverage AMD data center solutions to drive leadership generative AI capabilities, deliver cloud infrastructure used by millions of people daily and power on-prem and hybrid data centers for leading enterprises:

- Since launching in December 2023, AMD Instinct MI300X accelerators have been deployed at scale by leading cloud, OEM and ODM partners and are serving millions of users daily on popular AI models, including OpenAI's ChatGPT, Meta Llama and over one million open source models on the Hugging Face platform.
- Google highlighted how AMD EPYC processors power a wide range of instances for AI, high performance, general purpose and confidential computing, including their AI Hypercomputer, a supercomputing architecture designed to maximize AI ROI. Google also announced EPYC 9005 Series-based VMs will be available in early 2025.
- Oracle Cloud Infrastructure shared how it leverages AMD EPYC CPUs, AMD Instinct accelerators and Pensando DPUs to deliver fast, energy efficient compute and networking infrastructure for customers like Uber, Red Bull Powertrains, PayPal and Fireworks AI. OCI announced the new E6 compute platform powered by EPYC 9005 processors.
- Databricks highlighted how its models and workflows run seamlessly on AMD Instinct and ROCm and disclosed that their testing shows the large memory capacity and compute capabilities of AMD Instinct MI300X GPUs help deliver an over 50% increase in performance on Llama and Databricks proprietary models.
- Microsoft CEO Satya Nadella highlighted Microsoft's longstanding collaboration and co-innovation with AMD across its product offerings and infrastructure, with MI300X delivering strong performance on Microsoft Azure and GPT workloads. Nadella and Su also discussed the companies' deep partnership on the AMD Instinct roadmap and how Microsoft is planning to leverage future generations of AMD Instinct accelerators including MI350 series and beyond to deliver leadership performance-per-dollar-per-watt for AI applications.
- Meta detailed how AMD EPYC CPUs and AMD Instinct accelerators power its compute infrastructure across AI deployments and services, with MI300X serving all live traffic on Llama 405B. Meta is also partnering with AMD to optimize AI performance from silicon, systems, and networking to software and applications.
- Leading OEMs Dell, HPE, Lenovo and Supermicro are expanding on their highly performant, energy efficient AMD EPYC processor-based lineups with new platforms designed to modernize data centers for the AI era.

Expanding an Open AI Ecosystem
AMD continues to invest in the open AI ecosystem and expand the AMD ROCm open source software stack with new features, tools, optimizations and support to help developers extract the ultimate performance from AMD Instinct accelerators and deliver out-of-the-box support for today's leading AI models. Leaders from Essential AI, Fireworks AI, Luma AI and Reka AI discussed how they are optimizing models across AMD hardware and software.

AMD also hosted a developer event joined by technical leaders from across the AI developer ecosystem, including Microsoft, OpenAI, Meta, Cohere, xAI and more. Luminary presentations hosted by the inventors of popular AI programming languages, models and frameworks critical to the AI transformation taking place, such as Triton, TensorFlow, vLLM and Paged Attention, FastChat and more, shared how developers are unlocking AI performance optimizations through vendor agnostic programming languages, accelerating models on AMD Instinct accelerators, and highlighted the ease of use porting to ROCm software and how the ecosystem is benefiting from an open-source approach.

Enabling Enterprise Productivity with AI PCs
AMD launched AMD Ryzen AI PRO 300 Series processors, powering the first Microsoft Copilot+ laptops enabled for the enterprise[3]. The Ryzen AI PRO 300 Series processor lineup extends AMD leadership in performance and battery life with the addition of enterprise-grade security and manageability features for business users.

- The Ryzen AI PRO 300 Series processors, featuring the new AMD "Zen 5" and AMD XDNA™ 2 architectures, are the world's most advanced commercial processors[4], offering best in class performance for unmatched productivity[5] and an industry leading 55 NPU TOPS[6] of AI performance with the Ryzen AI 9 HX PRO 375 processor to process AI tasks locally on Ryzen AI PRO laptops.
- Microsoft highlighted how Windows 11 Copilot+ and the Ryzen AI PRO 300 lineup are ready for next generation AI experiences, including new productivity and security features.
- OEM partners including HP and Lenovo are expanding their commercial offerings with new PCs powered by Ryzen AI PRO 300 Series processors, with more than 100 platforms expected to come to market through 2025.

Supporting Resources

- Watch the AMD Advancing AI keynote and see the news here
- 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.

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; AMD's expected data center and AI growth opportunities; the ability of AMD to build momentum for AMD EPYC™ and AMD Instinct™ processors across its customers; the ability of AMD to deliver leadership compute to power to its customers workloads; the anticipated growth of the data center AI accelerator market by 2028; and AMD's commitment to delivering open innovation at scale, 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.

---

[1] EPYC-022F: For a complete list of world records see: http://amd.com/worldrecords.

[2] 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.
MI300-62
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.

[3] Based on Microsoft Copilot+ requirements of minimum 40 TOPS using AMD product specifications and competitive products announced as of Oct 2024. Microsoft requirements found here - https://support.microsoft.com/en-us/topic/copilot-pc-hardware-requirements-35782169-6eab-4d63-a5c5-c498c3037364. STXP-05.

[4] Based on a small node size for an x86 platform and cutting-edge, interconnected technologies, as of September 2024. GD-203b

[5] Testing as of Sept 2024 by AMD performance labs using the following systems: HP EliteBook X G1a with AMD Ryzen AI 9 HX PRO 375 processor @40W, Radeon™ 890M graphics, 32GB of RAM, 512GB SSD, VBS=ON, Windows 11 Pro; Lenovo ThinkPad T14s Gen 6 with AMD Ryzen™ AI 7 PRO 360 processor @22W, Radeon™ 880M graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11 Pro; Dell Latitude 7450 with Intel Core Ultra 7 165U processor @15W (vPro enabled), Intel Iris Xe Graphics, VBS=ON, 32GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Professional; Dell Latitude 7450 with Intel Core Ultra 7 165H processor @28W (vPro enabled), Intel Iris Xe Graphics, VBS=ON, 16GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Pro. The following applications were tested in Balanced Mode: Teams + Procyon Office Productivity, Teams + Procyon Office Productivity Excel, Teams + Procyon Office Productivity Outlook, Teams + Procyon Office Productivity Power Point, Teams + Procyon Office Productivity Word, Composite Geomean Score. Each

Microsoft Teams call consists of 9 participants (3X3). Laptop manufactures may vary configurations yielding different results. STXP-10.
Testing as of Sept 2024 by AMD performance labs using the following systems: (1) Lenovo ThinkPad T14s Gen 6 with an AMD Ryzen™ AI 7 PRO 360 processor (@22W), Radeon™ 880M graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11 Pro; (2) Dell Latitude 7450 with Intel Core Ultra 7 165U processor (@15W) (vPro enabled), Intel Iris Xe Graphics, VBS=ON, 32GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Professional; and (3) Dell Latitude 7450 with Intel Core Ultra 7 165H processor (@28W) (vPro enabled), Intel Arc Graphics, VBS=ON, 16GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Pro. Tested applications (in Balanced Mode) include: Procyon Office Productivity, Procyon Office Productivity Excel, Procyon Office Productivity Outlook, Procyon Office Productivity Power Point, Procyon Office Productivity Word, Composite Geomean Score. Laptop manufactures may vary configurations yielding different results. STXP-11.
[6] Trillions of Operations per Second (TOPS) for an AMD Ryzen processor is the maximum number of operations per second that can be executed in an optimal scenario and may not be typical. TOPS may vary based on several factors, including the specific system configuration, AI model, and software version. GD-243.

Media Contacts:
Brandi Martina
AMD Communications
+1 512-705-1720
brandi.martina@amd.com

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

Source: Advanced Micro Devices, Inc.