May 21, 2024

AMD Instinct MI300X Accelerators Power Microsoft Azure OpenAI Service Workloads and New Azure ND MI300X V5 VMs

— The new Azure ND MI300X V5 instances are now generally available, with Hugging Face as the first customer —

— Microsoft is using VMs powered by AMD Instinct MI300X and ROCm software to achieve leading price/performance for GPT workloads —

SANTA CLARA, Calif., May 21, 2024 (GLOBE NEWSWIRE) -- Today at Microsoft Build, AMD (NASDAQ: AMD) showcased its latest end-to-end compute and software capabilities for Microsoft customers and developers. By using AMD solutions such as AMD Instinct™ MI300X accelerators, ROCm™ open software, Ryzen™ AI processors and software, and Alveo™ MA35D media accelerators, Microsoft is able to provide a powerful suite of tools for AI-based deployments across numerous markets. The new Microsoft Azure ND MI300X virtual machines (VMs) are now generally available, giving customers like Hugging Face, access to impressive performance and efficiency for their most demanding AI workloads.

“The AMD Instinct MI300X and ROCm software stack is powering the Azure OpenAI Chat GPT 3.5 and 4 services, which are some of the world’s most demanding AI workloads,” said Victor Peng, president, AMD. “With the general availability of the new VMs from Azure, AI customers have broader access to MI300X to deliver high-performance and efficient solutions for AI applications.”

“Microsoft and AMD have a rich history of partnering across multiple computing platforms: first the PC, then custom silicon for Xbox, HPC and now AI,” said Kevin Scott, chief technology officer and executive vice president of AI, Microsoft. “Over the more recent past, we’ve recognized the importance of coupling powerful compute hardware with the system and software optimization needed to deliver amazing AI performance and value. Together with AMD, we’ve done so through our use of ROCm and MI300X, empowering Microsoft AI customers and developers to achieve excellent price-performance results for the most advanced and compute-intense frontier models. We’re committed to our collaboration with AMD to continue pushing AI progress forward.”

Advancing AI at Microsoft

Previously announced in preview in November 2023, the Azure ND MI300x v5 VM series are now available in the Canada Central region for customers to run their AI workloads. Offering industry-leading performance, these VMs provide impressive HBM capacity and memory bandwidth, enabling customers to fit larger models in GPU memory and/or use less GPUs, ultimately helping save power, cost, and time to solution.

These VMs and the ROCm™ software that powers them, are also being used for Azure AI Production workloads, including Azure OpenAI Service, providing customers with access to GPT-3.5 and GPT-4 models. With AMD Instinct MI300X and the proven and ready ROCm open software stack, Microsoft is able to achieve leading price/performance on GPT inference workloads.

Beyond Azure AI production workloads, one of the first customers to use these VMs is Hugging Face. Porting their models to the ND MI300X VMs in just one-month, Hugging Face was able to achieve impressive performance and price/performance for their models. As part of this, ND MI300X VM customers can bring Hugging Face models to the VMs to create and deploy NLP applications with ease and efficiency.

“The deep collaboration between Microsoft, AMD and Hugging Face on the ROCm open software ecosystem will enable Hugging Face users to run hundreds of thousands of AI models available on the Hugging Face Hub on Azure with AMD Instinct GPUs without code changes, making it easier for Azure customers to build AI with open models and open source,” said Julien Simon, chief evangelist officer, Hugging Face.

Additionally, developers are able to use AMD Ryzen AI software to optimize and deploy AI inference on AMD Ryzen AI powered PCs[1]. Ryzen AI software enables applications to run on the neural processing unit (NPU) built on AMD XDNA™ architecture, the first dedicated AI processing silicon on a Windows x86 processor[2]. While running AI models on a CPU or GPU alone can drain the battery fast, with a Ryzen AI powered-laptop, AI models operate on the embedded NPU, freeing-up CPU and GPU resources for other compute tasks. This helps significantly increase battery life and allows developers to run on-device LLM AI workloads and concurrent applications efficiently and locally.

Advancing Video Services and Enterprise Compute
Microsoft has selected the AMD Alveo™ MA35D media accelerator to power its vast live streaming video workloads, including Microsoft Teams, SharePoint video, and others. Purpose-built to power live interactive streaming services at scale, the Alveo MA35D will help Microsoft ensure high-quality video experience by streamlining video processing workloads, including video transcoding, decoding, encoding, and adaptive bitrate (ABR) streaming. Using the Alveo MA35D accelerator in servers powered by 4th Gen AMD EPYC™ processors, Microsoft is getting:

- **Ability to consolidate servers and cloud Infrastructure** - harnessing the high channel density, energy-efficient and ultra-low latency video processing capabilities of the Alveo MA35D, Microsoft can significantly reduce the number of servers required to support its high-volume live interactive streaming applications.
- **Impressive Performance** - the Alveo MA35D features ASIC-based video processing units supporting the AV1 compression standard and AI-enabled video quality optimizations that help ensure smooth and seamless video experiences.
- **Future-Ready AV1 Technology** - with an upgrade path to support emerging standards like AV1, the Alveo MA35D provides Microsoft with a solution that can adapt to evolving video processing requirements.

4th Gen AMD EPYC™ processors today power numerous general purpose, memory-intensive, compute-optimized, and accelerated compute VMs at Azure. These VMs showcase the growth and demand for AMD EPYC processors in the cloud and can provide up to 20% better performance for general purpose and memory-intensive VMs with

better price/performance, and up to 2x the CPU performance for compute-optimized VMs versus the previous generation of AMD EPYC processor-powered VMs at Azure. Now in preview, the Dalsv6, Dasv6, Easv6, Falsv6 and Famsv6 VM-series will become generally available in the coming months.

Supporting Resources

- Read more about AMD and Microsoft Collaboration
- Read the Microsoft Blog and News Hub
- 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.

©2024 Advanced Micro Devices, Inc. All rights reserved. AMD, Alveo, AMD Instinct, AMD XDNA, EPYC, ROCm, Ryzen, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names used herein are for informational purposes only and may be trademarks of their respective owners.

[1] Ryzen™ AI is defined as the combination of a dedicated AI engine, AMD Radeon™ graphics engine, and Ryzen processor cores that enable AI capabilities. OEM and ISV enablement is required, and certain AI features may not yet be optimized for Ryzen AI processors. Ryzen AI is compatible with: (a) AMD Ryzen 7040 and 8040 Series processors except Ryzen 5 7540U, Ryzen 5 8540U, Ryzen 3 7440U, and Ryzen 3 8440U processors; and (b) All AMD Ryzen 8000G Series desktop processors except the Ryzen 5 8500 G/GE and Ryzen 3 8300 G/GE. Please check with your system manufacturer for feature availability prior to purchase. GD-220b

[2] As of May 2023, AMD has the first and only available dedicated AI engine on an x86 Windows processor, where 'dedicated AI engine' is defined as an AI engine that has no function other than to process AI inference models and is part of the x86 processor die. For detailed information, please check: https://www.amd.com/en/products/ryzen-ai. PHX-3.

Contact:
Aaron Grabein
AMD Communications
(512) 602-8950
Aaron.Grabein@amd.com

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



Source: Advanced Micro Devices, Inc.