September 26, 2024

AMD Instinct MI300X Accelerators Available on Oracle Cloud Infrastructure for Demanding AI Applications

— Customers including Fireworks AI are powering their AI inference and training workloads with new OCI Compute instances --

— OCI Supercluster leads among cloud providers with support for up to 16,384 AMD Instinct MI300X GPUs in a single ultrafast network fabric --

SANTA CLARA, Calif., Sept. 26, 2024 (GLOBE NEWSWIRE) -- AMD (NASDAQ: AMD) today announced that Oracle Cloud Infrastructure (OCI) has chosen AMD Instinct™ MI300X accelerators with ROCm™ open software to power its newest OCI Compute Supercluster instance called BM.GPU.MI300X.8. For AI models that can comprise hundreds of billions of parameters, the OCI Supercluster with AMD MI300X supports up to 16,384 GPUs in a single cluster by harnessing the same ultrafast network fabric technology used by other accelerators on OCI. Designed to run demanding AI workloads including large language model (LLM) inference and training that requires high throughput with leading memory capacity and bandwidth, these OCI bare metal instances have already been adopted by companies including Fireworks AI.

“AMD Instinct MI300X and ROCm open software continue to gain momentum as trusted solutions for powering the most critical OCI AI workloads,” said Andrew Dieckmann, corporate vice president and general manager, Data Center GPU Business, AMD. “As these solutions expand further into growing AI-intensive markets, the combination will benefit OCI customers with high performance, efficiency, and greater system design flexibility.”

“The inference capabilities of AMD Instinct MI300X accelerators add to OCI’s extensive selection of high-performance bare metal instances to remove the overhead of virtualized compute commonly used for AI infrastructure,” said Donald Lu, senior vice president, software development, Oracle Cloud Infrastructure. “We are excited to offer more choice for customers seeking to accelerate AI workloads at a competitive price point.”

Bringing Trusted Performance and Open Choice for AI Training and Inference
The AMD Instinct MI300X underwent extensive testing which was validated by OCI that underscored its AI inferencing and training capabilities for serving latency-optimal use cases, even with larger batch sizes, and the ability to fit the largest LLM models in a single node. These Instinct MI300X performance results have garnered the attention of AI model developers.

Fireworks AI offers a fast platform designed to build and deploy generative AI. With over 100+ models, Fireworks AI is leveraging the benefits of performance found in OCI using AMD Instinct MI300X.

“Fireworks AI helps enterprises build and deploy compound AI systems across a wide range of industries and use cases," said Lin Qiao, CEO of Fireworks AI. "The amount of memory capacity available on the AMD Instinct MI300X and ROCm open software allows us to scale services to our customers as models continue to grow."

Supporting Resources

- Follow AMD on LinkedIn
- Follow AMD on Twitter

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.

Contact:
David Szabados
AMD Investor Relations
+1 408-472-2439
david.szabados@amd.com

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

Source: Advanced Micro Devices, Inc.