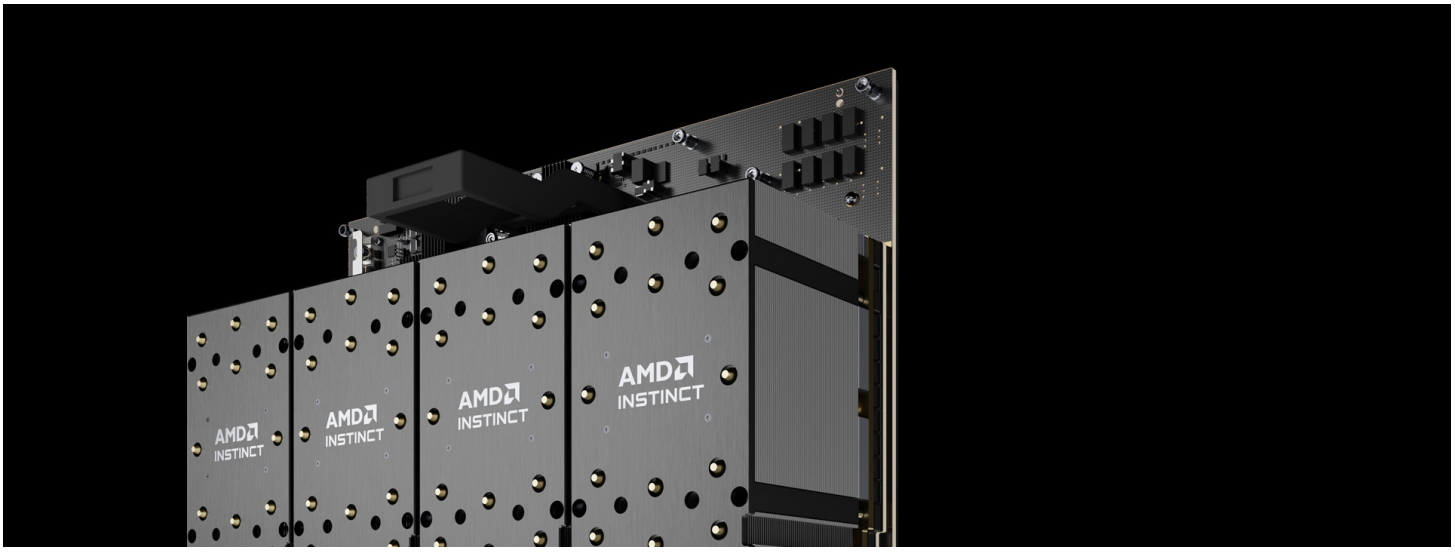
IBM Newsroom

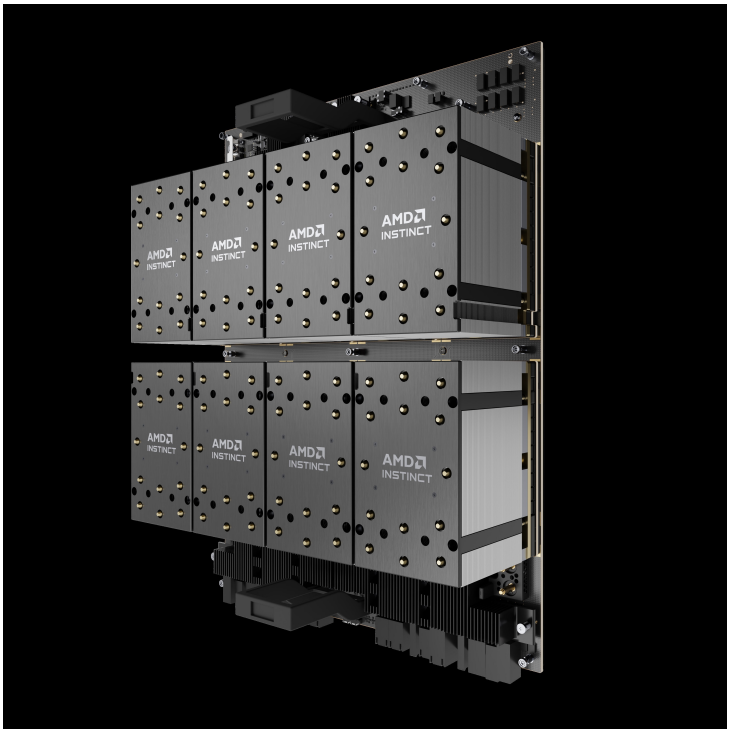
IBM Expands its AI Accelerator Offerings; Announces Collaboration with AMD

IBM Cloud to Deploy AMD Instinct™ MI300X Accelerators to Support Performance for Generative AI Workloads and HPC Applications



November 18, 2024 - Armonk, NY - IBM (NYSE:IBM) and AMD have announced a collaboration to deploy AMD Instinct MI300X accelerators as a service on IBM Cloud. This offering, which is expected to be available in the first half of 2025, aims to enhance performance and power efficiency for Gen AI models such as and high-performance computing (HPC) applications for enterprise clients. This collaboration will also enable support for AMD Instinct MI300X accelerators within IBM's watsonx AI and data platform, as well as Red Hat® Enterprise Linux® AI inferencing support.

“As enterprises continue adopting larger AI models and datasets, it is critical that the accelerators within the system can process compute-intensive workloads with high performance and flexibility to scale,” said Philip Guido, executive vice president and chief commercial officer, AMD. “AMD Instinct accelerators combined with AMD ROCm software offer wide support including IBM watsonx AI, Red Hat Enterprise Linux AI and Red Hat OpenShift AI platforms to build leading frameworks using these powerful open ecosystem tools. Our collaboration with IBM Cloud will aim to allow customers to execute and scale Gen AI inferencing without hindering cost, performance or efficiency.”



(Credit: AMD)

“AMD and IBM Cloud share the same vision around bringing AI to enterprises. We’re committed to bringing the power of AI to enterprise clients, helping them prioritize their outcomes and ensuring they have the power of choice when it comes to their AI deployments,” said Alan Peacock, General Manager of IBM Cloud. “Leveraging AMD’s accelerators on IBM Cloud will give our enterprise clients another option to scale to meet their enterprise AI needs, while also aiming to help them optimize cost and performance.”

IBM and AMD are collaborating to deliver MI300X accelerators as a service on IBM Cloud to support enterprise clients leveraging AI. To help enterprise clients across industries, including those that are heavily regulated, IBM and AMD intend to leverage IBM Cloud’s security and compliance capabilities.

- **Support for Large Model Inferencing:** Equipped with 192GB of high-bandwidth memory (HBM3), AMD Instinct MI300X accelerators offer support for the largest model inferencing and fine tuning. The large memory capacity can also help customers run larger models with fewer GPUs, potentially lowering costs for inferencing.
- **Enhanced Performance and Security:** Offering AMD Instinct MI300X accelerators as a service on IBM Cloud Virtual Servers for VPC, as well as through container support with IBM Cloud Kubernetes Service and IBM Red Hat OpenShift on IBM Cloud, can help optimize performance for enterprises running AI applications.

For generative AI inferencing workloads, IBM plans to enable support for AMD Instinct MI300X accelerators within IBM's watsonx AI and data platform, providing watsonx clients with additional AI infrastructure resources for scaling their AI workloads across hybrid cloud environments. Additionally, Red Hat Enterprise Linux AI and Red Hat OpenShift AI platforms can run Granite family large language models (LLMs) with alignment tooling using InstructLab on MI300X accelerators.

IBM Cloud with AMD Instinct MI300X accelerators are expected to be generally available in the first half of 2025. Stay tuned for more updates from AMD and IBM in the coming months.

To learn more about IBM’s GPU and Accelerator offerings, visit: https://www.ibm.com/cloud/gpu

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

Media Contact:

Kate Connors
IBM
Kate.Connors@ibm.com

---

https://stage.mediaroom.com/ibmnewsroom/2024-11-18-ibm-expands-its-ai-accelerator-offerings-announces-collaboration-with-amd