AMD Computex MI325x & MI350 Accelerator Announcements

STEVE MCDOWELL, CHIEF ANALYST
6/4/24

CONTEXT

At the 2024 Computex event in Taiwan, AMD CEO Lisa Su revealed details about AMD's upcoming MI350 and MI325X accelerators, follow-ons to its current MI300x products, highlighting significant advancements in AI performance and memory capacity. The new products are positioned as key components in AMD's strategy to lead the AI accelerator market.

Let's look at what AMD disclosed.

MI325X ACCELERATOR

Set to launch later in the year, the MI325X will feature up to 288 GB of ultrafast HBM3E memory and 6 TB/s of memory bandwidth, continuing the trend of high performance and large memory capacity.

Compared to competing products, the MI325X offers twice the memory and 1.3x faster memory bandwidth, making it capable of running advanced models with up to 1 trillion parameters on a single server:

1. **Enhanced Memory and Bandwidth**:
    - **Memory Capacity**: The MI325X will feature up to 288 GB of HBM3E memory, twice the capacity of its closest competitors.
    - **Bandwidth**: With 6 TB/s of memory bandwidth, the MI325X promises 1.3x faster than competing solutions. This improvement is crucial for handling large AI models and datasets efficiently.

2. **Performance and Compatibility**:
    - **AI Model Support**: The increased memory and bandwidth make the MI325X capable of running advanced AI models up to 1 trillion parameters on a single server, doubling the capacity supported by NVIDIA's H100 server.

- **Ease of Transition**: The MI325X uses the same infrastructure as the MI300, ensuring an easy transition for existing customers looking to upgrade their systems.

MI350 Accelerator

The MI350 series, expected in 2025, will utilize the new CDNA 4 architecture and advanced 3-nanometer process technology. This will mark AMD's largest generational leap in AI performance, offering a 35x increase in AI performance compared to CDNA 3.

The MI350 series will deliver 1.2x more performance overall compared to competitors, further enhancing AMD's leadership in AI accelerator technology:

1. **CDNA 4 Architecture**:
    - **Technological Leap**: The MI350 series will be built on the new CDNA 4 architecture, using advanced 3-nanometer process technology. This upgrade represents AMD's largest generational leap in AI performance, offering a 35x increase in performance compared to the previous CDNA 3 architecture.
    - **Advanced Data Types**: The MI350 series will support new data types, including FP4 and FP6, enhancing its ability to handle complex AI workloads with higher precision and efficiency.

2. **Market Positioning**:
    - **Performance Advantage**: Compared to NVIDIA's B200, the MI350 series will offer 1.5x more memory and 1.2x more overall performance, reinforcing AMD's competitive edge in the AI accelerator market.
    - **Industry Adoption**: AMD's strategic partnerships and the adoption of open standards, such as UALink and Ultra Ethernet, will further enhance the appeal of the MI350 series for enterprise and cloud service providers.

ANALYSIS

AMD's announcements emphasize their commitment to advancing AI hardware capabilities and maintaining a leadership position in the high-performance computing market. The MI325X's superior memory and bandwidth capabilities, combined with the MI350 series' significant performance improvements, illustrate AMD's focus on addressing the growing demands of AI workloads.

These advancements will likely attract a wide range of customers, from cloud service providers to enterprises, looking for efficient and scalable AI solutions. AMD's strategy to offer an annual cadence of new products ensures they remain at the forefront of technological innovation in the AI accelerator space.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.