NAND RESEARCH

RESEARCH NOTE

AMD MI300A & MI300X AI/HPC ACCELERATORS

STEVE MCDOWELL, CHIEF ANALYST
DECEMBER 10, 2023

CONTEXT

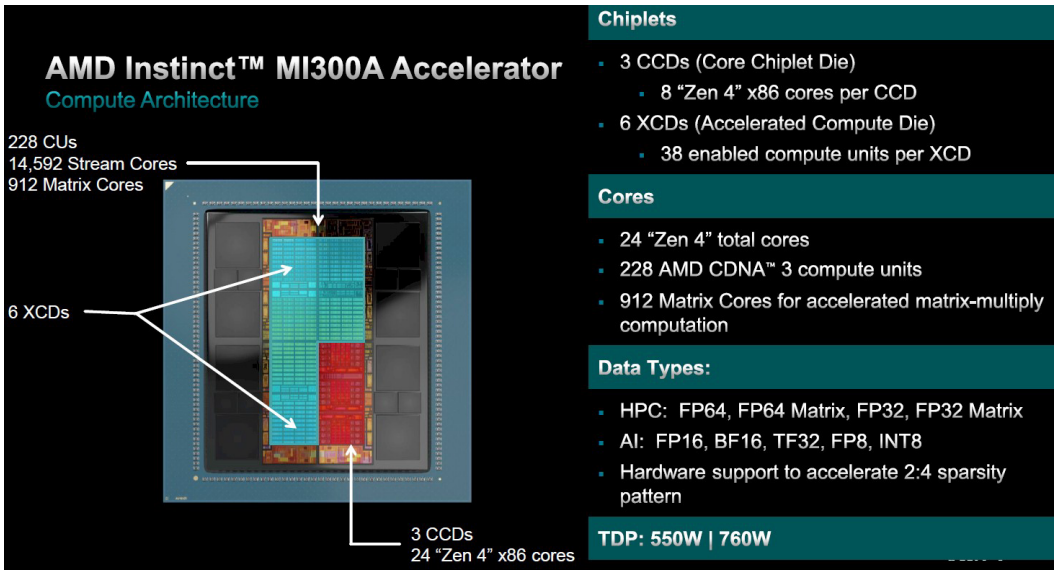
AMD unveiled its new AMD Instinct™ MI300X and MI300A accelerators at its recent AI-focused event in San Jose.

The new MI300X offers leading memory bandwidth for generative AI and top performance for training and inferencing large language models (LLMs).

The company also introduced the AMD Instinct™ MI300A accelerated processing unit (APU), which combines AMD CDNA™ 3 architecture and “Zen 4” CPUs for high-performance computing (HPC) and AI workloads.

NEW: AMD INSTINCT MI300X

The AMD Instinct MI300X is a high-performance GPU accelerator, part of AMD's "Antares" Instinct MI300 family, specifically designed to address the demands of the generative AI market.



SOURCE: AMD

The MI300X can be summarized as follows:

- **Architecture and Design**: The MI300X is built on the advanced AMD CDNA 3 architecture, which is a significant step forward from its predecessors in terms of performance and efficiency. This architecture enables the GPU to handle complex AI and high-performance computing (HPC) workloads effectively.

- **Memory and Bandwidth**: The MI300X is its large memory capacity and high memory bandwidth. It has 192 GB of HBM (High Bandwidth Memory), offering a peak memory bandwidth of 5.3 TB/s. Its vast memory and bandwidth are crucial for handling large, complex datasets typical in AI applications and LLM (Large Language Model) training.

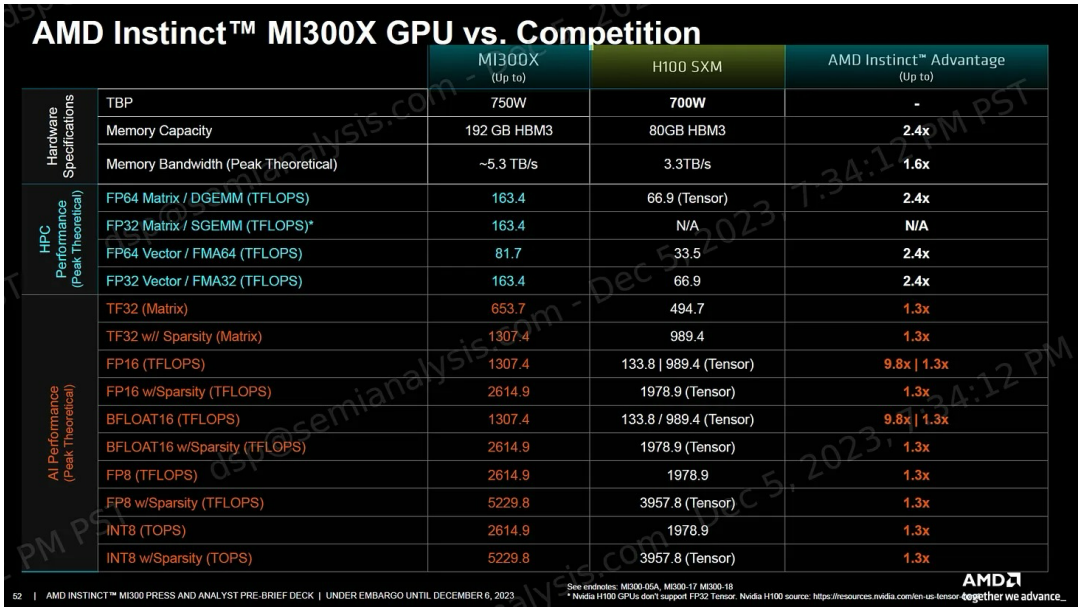
- **Compute Performance**: The MI300X delivers impressive compute performance. It has numerous compute units and stream processors designed for vector and matrix operations, crucial for AI and machine learning tasks. The GPU supports various data formats and can handle sparse matrices efficiently, doubling their throughput, which is particularly beneficial for AI workloads.

- **Energy Efficiency**: Despite its high performance, the MI300X is designed to be energy efficient. This efficiency is increasingly important in data centers where power consumption and heat generation are significant concerns.

- **Compatibility and Integration**: The MI300X is compatible with existing platforms, simplifying its deployment and integration into current systems. This compatibility is crucial for organizations upgrading their computing capabilities without overhauling their entire infrastructure.

PERFORMANCE

AMD shared a large number of relevant benchmarks for the new accelerator, showing that the MI300X impressively outperforms NVIDIA's H100 in several vital areas, boasting 30% more FP8 FLOPS, 60% higher memory bandwidth, and over double the memory capacity.



AMD Instinct™ MI300X GPU vs. Competition

		MI300X (Up to)	H100 SXM	AMD Instinct™ Advantage (Up to)
Hardware Specifications	TBP	750W	700W	-
	Memory Capacity	192 GB HBM3	80GB HBM3	2.4x
	Memory Bandwidth (Peak Theoretical)	~5.3 TB/s	3.3TB/s	1.6x
HPC Performance (Peak Theoretical)	FP64 Matrix / DGEMM (TFLOPS)	163.4	66.9 (Tensor)	2.4x
	FP32 Matrix / SGEMM (TFLOPS)*	163.4	N/A	N/A
	FP64 Vector / FMA64 (TFLOPS)	81.7	33.5	2.4x
	FP32 Vector / FMA32 (TFLOPS)	163.4	66.9	2.4x
AI Performance (Peak Theoretical)	TF32 (Matrix)	653.7	494.7	1.3x
	TF32 w// Sparsity (Matrix)	1307.4	989.4	1.3x
	FP16 (TFLOPS)	1307.4	133.8 | 989.4 (Tensor)	9.8x | 1.3x
	FP16 w/Sparsity (TFLOPS)	2614.9	1978.9 (Tensor)	1.3x
	BFLOAT16 (TFLOPS)	1307.4	133.8 / 989.4 (Tensor)	9.8x | 1.3x
	BFLOAT16 w/Sparsity (TFLOPS)	2614.9	1978.9 (Tensor)	1.3x
	FP8 (TFLOPS)	2614.9	1978.9	1.3x
	FP8 w/Sparsity (TFLOPS)	5229.8	3957.8 (Tensor)	1.3x
	INT8 (TOPS)	2614.9	1978.9	1.3x
	INT8 w/Sparsity (TOPS)	5229.8	3957.8 (Tensor)	1.3x

SOURCE: AMD

The AMD MI300X, as described in the content, is a significant new entrant in the GPU market, particularly in HPC and AI. Here are its key features and performance aspects based on the provided information:

- **Performance Specs**: The MI300X stands out with its impressive specifications, outperforming NVIDIA's H100 in several critical areas. It offers 30% more FP8 FLOPS and 60% more memory bandwidth, with over double the memory capacity. However, it falls slightly short of its initial memory bandwidth target, reaching 5.3TB/s instead of the projected 5.6TB/s.

- **Benchmark Performance**: While the MI300X boasts superior raw specifications, its actual performance in benchmarks shows it underperforming against its theoretical peak. In tests like FlashAttention2 and LLAMA2-70B, which focus on forward pass and compute-bound workloads, the MI300X demonstrates a 10% to 20% performance advantage, notably less than what might be expected from its raw specs.

- **Inference Capabilities**: The MI300X shows more significant advantages in inference benchmarks. For example, the Bloom benchmark leverages its larger memory capacity for higher throughput. The LLAMA 2-70B inference

benchmark exhibits a 40% latency advantage over the H100, a benefit attributed to its superior bandwidth.

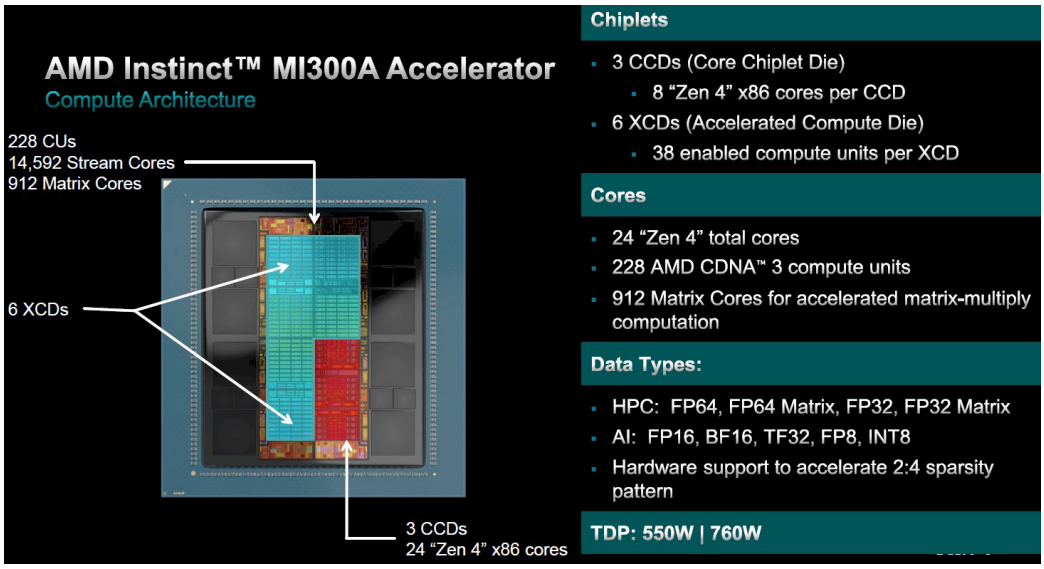
- **Training Performance**: AMD's MI300X reveals some limitations in training due to its software stack, achieving less than 30% of the theoretical FLOPS. This is where NVIDIA often outperforms AMD, frequently reaching around 40% of theoretical FLOPS.

- **Software Optimizations and Future Prospects**: Despite current shortcomings in training performance, there is an expectation of improvement as AMD's software optimization progresses. The GPU is expected to become more competitive over time, especially with the increasing support from software frameworks like OpenAI Triton and PyTorch 2.0.

The AMD MI300X is a powerful GPU that shines in raw specs and inference tasks, with room for growth in training performance. Its ongoing software improvements and growing ecosystem support suggest a promising future in the competitive landscape of GPUs.

NEW: AMD INSTINCT MI300A

The AMD Instinct MI300A is AMD's new APU, combining a CPU and GPU into a single package. The new offering fits into AMD's "Antares" Instinct MI300 family of GPUs, specifically designed for HPC and AI applications.



SOURCE: AMD

The MI300A can be summarized as follows:

- **Hybrid Design**: The MI300A is a unique hybrid device that combines AMD EPYC CPUs and Instinct GPUs in a single package. This integration allows for efficient

data processing, as both CPU and GPU can access the same HBM space, reducing the need for data transfer between the two and enhancing overall performance.

- **Memory and Bandwidth**: The MI300A features 128 GB of next-generation HBM3 memory, offering a bandwidth of 5.3 TB/s. This high bandwidth and memory capacity are essential for handling the large and complex data sets typical in HPC and AI tasks.

- **Compute Performance**: Equipped with many compute units and stream processors, the MI300A is designed to handle vector and matrix computations efficiently. This makes it well-suited for a range of various applications, from AI model training to complex scientific computations.

- **Energy Efficiency**: Like the MI300X, the MI300A emphasizes energy efficiency. It integrates both CPU and GPU cores on a single package, enhancing performance and improving power efficiency – a critical aspect for large-scale computing environments where energy consumption is a significant concern.

- **Software and Training Performance**: While the MI300X shows some limitations in training performance due to its software stack, achieving less than 30% of the theoretical FLOPS, AMD did not provide the MI300A's performance. However, similar software optimizations and improvements will likely benefit the MI300A as well.

- **Packaging and Technology**: The MI300A utilizes advanced packaging techniques, including 3D stacking of compute tiles and a mix of 2.5D and 3D chip interconnects. This sophisticated packaging allows for a high level of integration and performance in a compact form factor.

The AMD Instinct MI300A is an innovative hybrid GPU accelerator offering significant computational power and efficiency. Its integrated design, high memory capacity, and advanced technology make it a competitive choice for HPC and AI applications, including use in some of the world's most powerful supercomputers.

COMPETITIVE POSITION VS NVIDIA

The AMD MI300 is a significant entrant in the GPU market, particularly in comparison with Nvidia's products, notably the H100 and H200 GPUs. Here's a detailed comparison based on various aspects:

1. **Performance Metrics**:

    - **FP8 FLOPS (Floating Point Operations per Second)**: The MI300X outperforms Nvidia's H100 by 30% in FP8 FLOPS, indicating a higher computational capacity for specific tasks.

- **Memory Bandwidth**: The MI300X offers 60% more memory bandwidth compared to the H100. This translates to faster data transfer rates within the GPU, beneficial for memory-intensive tasks. However, against the H200, the gap in memory bandwidth narrows significantly, falling into the single-digit range.

- **Memory Capacity**: The MI300X has more than twice the memory capacity of the H100. This is advantageous for handling larger datasets and complex computational tasks. The gap narrows with the H200, but the MI300X still maintains a lead of less than 40%.

2. **Benchmarks and Real-World Performance**:

    - Benchmarks show that while the MI300X has impressive raw specifications, its peak performance doesn't fully align with its theoretical capabilities.

    - Inference benchmarks such as FlashAttention2 and LLAMA2-70B indicate a performance advantage of 10-20% for the MI300X, which is less than what raw specs might suggest.

    - For more realistic inference scenarios, like the LLAMA 2-70B, AMD's MI300X demonstrates a significant 40% latency advantage over the H100, aligning with its bandwidth advantage.

    - In a broader benchmark (LLAMA 2-13B), the MI300X shows a 20% performance improvement and is also noted to be more cost-effective.

3. **Software and Ecosystem**:

    - AMD's software stack for the MI300 shows some weaknesses in training performance, achieving less than 30% of the theoretical FLOPS. Nvidia often reaches around 40%.

    - OpenAI's collaboration with AMD for supporting the MI300 in the Triton distribution is significant. It suggests a growing ecosystem around AMD GPUs, potentially challenging Nvidia's dominance in certain areas of machine learning and inference.

4. **Market Impact and Adoption**:

    - The MI300 is gaining significant traction among major companies and cloud providers, which indicates confidence in its capabilities and potential for widespread adoption.

    - AMD's strategic partnerships and openness to collaborate, as shown in their alliance with Broadcom and opening up of their infinity fabric network, position them as a strong competitor to Nvidia in the GPU market.

The AMD MI300X stands out for its superior memory bandwidth, capacity, and FP8 FLOPS compared to Nvidia's H100, positioning it as a strong competitor, especially in specific scenarios like high-throughput inference tasks.

However, when compared to Nvidia's H200, the performance gap narrows, particularly in memory bandwidth and capacity. The MI300X's real-world performance, while impressive, doesn't fully match its theoretical potential, a gap that AMD is likely to narrow with future optimizations and software updates.

ANALYSIS

The release of AMD's MI300X and MI300A signifies a pivotal moment in the GPU market, particularly for applications in high-performance computing and artificial intelligence.

The MI300X, with its advanced AMD CDNA 3 architecture and a substantial memory capacity of 192 GB HBM3, coupled with a remarkable 5.3 TB/s memory bandwidth, represents an impressive leap forward in GPU-only accelerators. While its performance overshadows competitors like Nvidia's H100, its actual performance, notably in training, reveals areas for improvement.

The MI300A, on the other hand, emerges as the world's first data center APU for HPC and AI, blending high-performance AMD CDNA 3 GPU cores with the latest "Zen 4" CPU cores. Its hybrid design, featuring 128GB of HBM3 memory and a shared memory architecture, underscores AMD's innovative approach in creating a highly efficient platform. This APU is positioned to accelerate AI model training and set new benchmarks in energy efficiency, resonating with AMD's 30x25 energy efficiency goal.

The strategic release of these accelerators demonstrates AMD's intent to mount a substantial challenge to NVIDIA's dominance in the generative AI space. With cloud providers and enterprises increasingly eager to diversify their technological portfolios, AMD's new offerings are well-positioned to disrupt the current market dynamics.

Market response to AMD's announcement has been telling. The adoption of the new offerings by major cloud providers like Microsoft and Oracle, along with enterprises such as Databricks (MosaicML), points to a broader industry trend of seeking out competitive alternatives to NVIDIA's solutions.

AMD's approach to design and performance, coupled with the potential for a more competitive pricing structure, could entice a significant segment of the market to reconsider their hardware allegiances.

It’s also worth noting the strategic alliances being formed as part of AMD’s broader market strategy. The collaboration with Broadcom to support infinity fabric on their PCIe switches is particularly noteworthy. This alliance may well be seen as a direct challenge to NVIDIA, setting the stage for a more diverse and competitive high-performance networking landscape.

While AMD's MI300X and MI300A are technically impressive, the true test will be in their deployment and the real-world gains they can deliver to the end users. As they stand, these accelerators are a testament to AMD's innovative spirit and its commitment to pushing the boundaries of what's possible in AI and HPC.

The industry will be watching closely to see if these products can fulfill AMD's promise and help reshape the future of computing. Whether these advances will translate into a significant market share gain for AMD remains to be seen, but the potential for disruption is unmistakable. With the groundwork laid for a significant shift, AMD's next moves will be crucial in determining the future dynamics of the AI hardware space.

RESEARCH NOTE

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