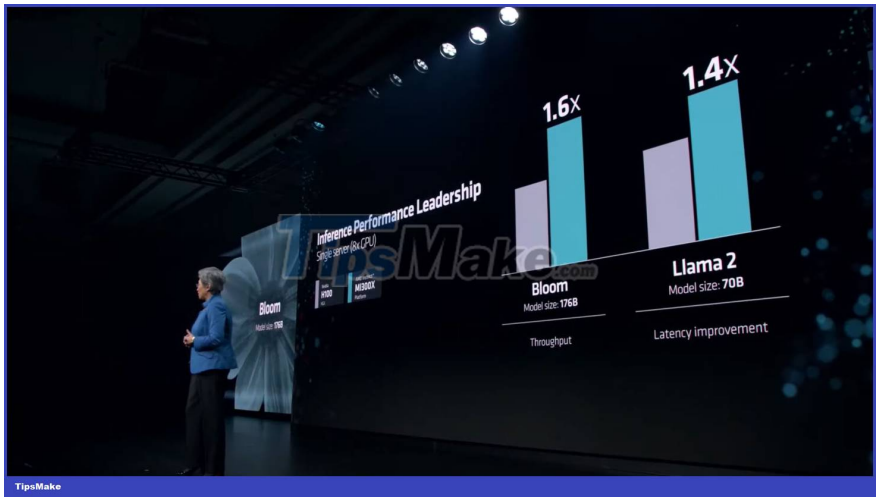
AMD launches AI Instinct MI300X GPU, up to 60% faster than NVIDIA H100

AMD has just officially launched the flagship MI300X AI GPU, which is said to be able to deliver up to 60% better performance than NVIDIA's H100 platform.

AMD has just officially launched the flagship MI300X AI GPU, which is said to be able to deliver up to 60% better performance than NVIDIA's H100 platform, and is mainly targeted at the data center market, serving processing. HPC and AI. Compared to its predecessor MI300A, Instinct MI300X focuses on taking advantage of the entire CDNA 3 architecture GPU processing core tile, instead of combining CPU and GPU as before.

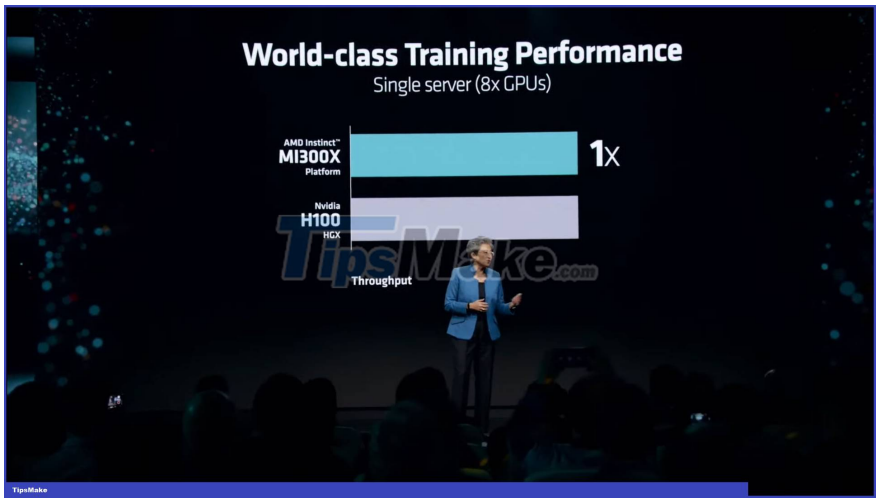
With the launch of the Instinct MI300X, AMD has for the first time used general specifications for comparison to highlight the performance of the advanced CDNA 3 accelerator (versus NVIDIA H100):

1. Memory capacity is 2.4 times higher
2. 1.6x higher memory bandwidth
3. 1.3 times FP8 TFLOPS
4. 1.3 times FP16 TFLOPS
5. Up to 20% faster than H100 (Llama 2 70B) in 1v1 comparison
6. Up to 20% faster than H100 (FlashAttention 2) in 1v1 comparison
7. Up to 40% faster than H100 (Llama 2 70B) in 8v8 Server
8. Up to 60% faster than H100 (Bloom 176B) in 8v8 server



Technically, the presence of LLM kernel TFLOP helps MI300X deliver up to 20% higher performance in FlashAttention-2 and Llama 2 70B. Looking from a platform perspective comparing the 8x MI300X solution

with the 8X H100 solution, a much larger 40% increase can be seen for the Llama 2 70B and even up to 60% for the Bloom 176B.

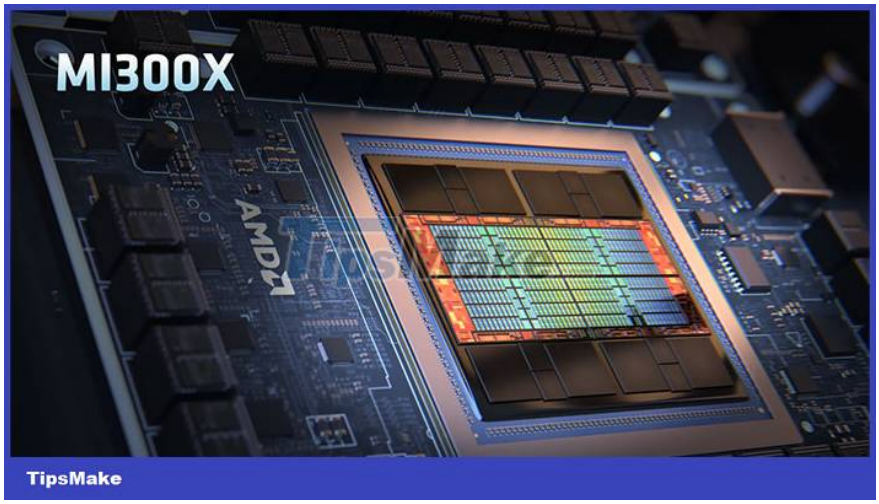


AMD mentions that in terms of training performance, the MI300X is on par with the competition (H100) and offers extremely competitive price/performance, while 'shining' in workloads involving inference computing .

One of the factors that helps AMD feel truly confident with its new MI300 platform lies in ROCm 6.0. The software stack has been updated to the latest version with powerful new features, including support for a variety of AI workloads such as creative AI and large language models.

The new software package supports the latest computing formats such as FP16, Bf16 and FP8 (including Sparsity). Optimizations combine to deliver up to 2.6x speedup in vLLM via optimized inference libraries, 1.4x speedup in HIP Graph via runtime optimized, and a 1.3x Flash Attention improvement through optimized Kernels. It will be interesting to compare ROCm 6 with NVIDIA's latest version of CUDA, which is a real competitor.

The AMD Instinct MI300X is the chip that will get the most attention as it targets NVIDIA's Hopper and Intel's Gaudi accelerators in the AI ??segment. This chip is exclusively designed on the CDNA 3 architecture with many notable improvements. The MI300X will host a mix of 5nm and 6nm IP, all combining to deliver up to 153 billion transistors. Combined with HBM3 VRAM memory with a capacity of up to 192GB, MI300X is capable of operating the largest scale machine learning models today.



In terms of design, the main interposer is arranged with a passive die containing an interconnect layer based on the 4th generation Infinity Fabric solution. The Interposer includes a total of 28 dies including 8 HBM3 packages, 16 dummy dies between HBM package and 4 active dies, and each active die contains two calculation dies.

Each GCD based on CDNA 3 GPU architecture has a total of 40 compute units equivalent to 2560 cores. There are eight compute dies (GCD) in total, thus providing a total of 320 compute units and 20,480 core units. To gain performance, AMD will scale back a fraction of these cores, and we will see a total of 304 compute units (38 CUs per GPU chiplet), enabled for a total of 19,456 processors. stream. stream.

Going deeper into memory, MI300X has 50% higher HBM3 capacity than its predecessor MI250X (128 GB). To achieve a total memory of 19 GB, AMD equips the MI300X with 8 HBM3 stacks and each stack is 12-Hi, and incorporates 1Gb ICs providing a capacity of 2GB per IC or 24GB per stack.

The storage will offer up to 5.3TB/s bandwidth and 896GB/s Infinity Fabric bandwidth. For comparison, NVIDIA's upcoming AI H200 GPU offers 141GB capacity, while Intel's Gaudi 3 will offer 144GB capacity. This is a very important factor in LLM-related workloads that rely heavily on memory, and AMD can demonstrate superior AI power on its products in this aspect.

1. Instinct MI300X - 192 GB HBM3
2. Gaudi 3 - 144 GB HBM3
3. H200 - 141 GB HBM3e
4. MI300A - 128 GB HBM3
5. MI250X - 128 GB HBM2e
6. H100 - 96 GB HBM3
7. Gaudi 2 - 96 GB HBM2e

In terms of power consumption, the AMD Instinct MI300X has a TDP of 750W, an increase of 50% compared to the 500W of the Instinct MI250X and 50W more than the NVIDIA H200.

Currently, AMD understands that its competitors are also making every effort to gain a solid foothold in the AI ??fever. NVIDIA has posted impressive numbers for the 2024 Hopper H200 and Blackwell B100 GPUs, while Intel is also preparing to launch the Guadi 3 and Falcon Shores GPUs in 2024. The same holds true for the coming years. Companies such as Oracle, Dell, META, and OpenAI have announced support for AMD's Instinct MI300 AI chip in their ecosystems.

With the launch of AMD Instinct MI300X, this manufacturer is expected to break Nvidia's monopoly in the AI ??research and operation chip market. The launch of MI300 will open up opportunities for AMD in this market. MI300X will also make an important contribution to AMD's annual financial report this year.

You finished reading the article "**AMD launches AI Instinct MI300X GPU, up to 60% faster than NVIDIA H100**" edited by the **TipsMake** team. We hope this article has provided you with many useful tech tips and tricks. You can search for similar articles on tips and guides. Thank you for reading and for following us regularly.