AMD's new GPU lineup aims to rival Nvidia in data centre AI market

capacitymedia.com/article/amd-new-data-centre-gpus

Ben Wodecki

October 10, 2024

AMD has come out swinging in the GPU battle to take on Nvidia unveiling new hardware and plans to launch new units yearly to keep pace with its rival’s annual product cycle.

At the company’s Advancing AI event in San Francisco, AMD provided greater clarity on its data centre hardware, including the MI325X and MI350X, unveiled earlier this year.

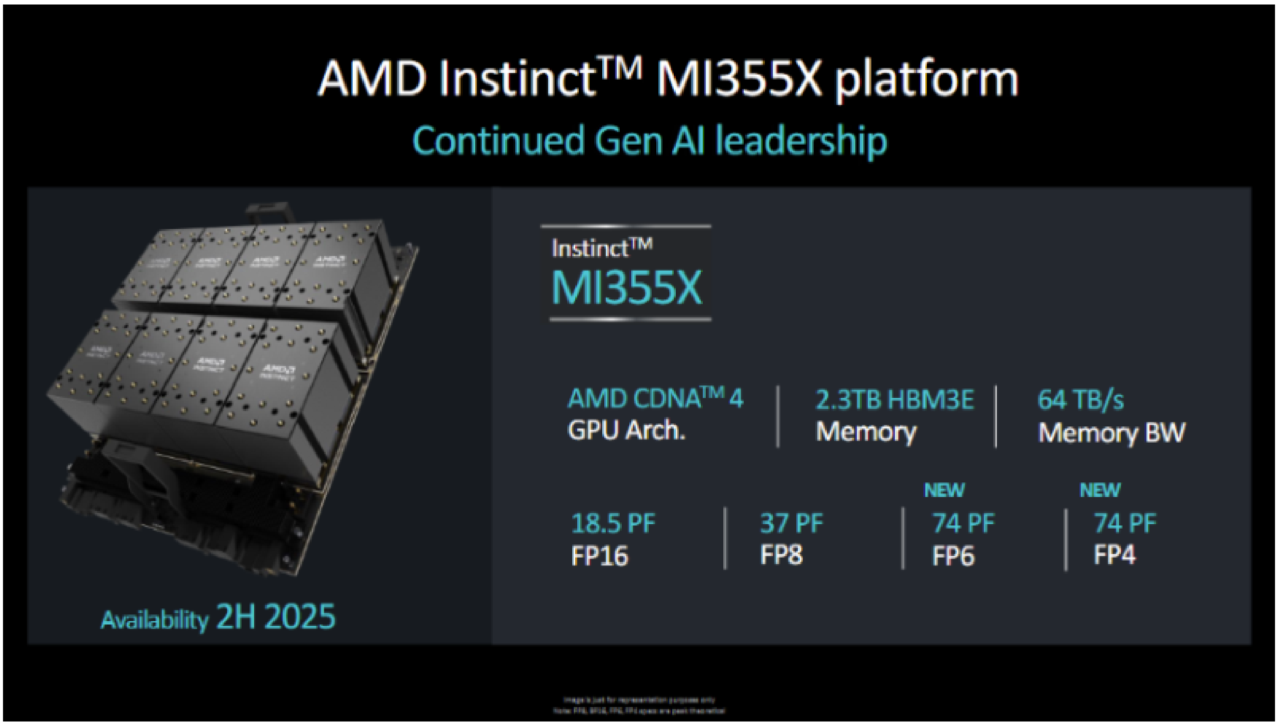
AMD also revealed the MI355X and teased a complete architectural overhaul coming to its GPUs in 2026.

The headline GPU update was the reveal of the MI355X — the next in AMD’s MI350 series and a potential challenger to Nvidia’s Blackwell GPUs.

Andrew Dieckmann, corporate VP and general manager for data centre GPUs at AMD, told a press gathering that the MI355X represents a “complete bottoms-up redesign” of AMD’s Instinct hardware.

The MI355X marks a move to a 3nm process with the chips housing a total of 10 compute elements and eight integrated memory stacks, enabling it to support up to 288GB of HBM3E.

AMD has also added support for new data types, including FP4 (4-bit floating point) and FP6 (6-bit floating point). These are lower-precision data types that can boost running AI and machine learning applications by increasing computational throughput, enhancing energy efficiency, and minimising data movement overhead, making them ideal for large-scale and power-constrained environments.



AMD Instinct™ MI355X platform

Continued Gen AI leadership

Instinct™
MI355X

AMD CDNA™ 4
GPU Arch.

2.3TB HBM3E
Memory

64 TB/s
Memory BW

18.5 PF
FP16

37 PF
FP8

NEW
74 PF
FP6

NEW
74 PF
FP4

Availability 2H 2025

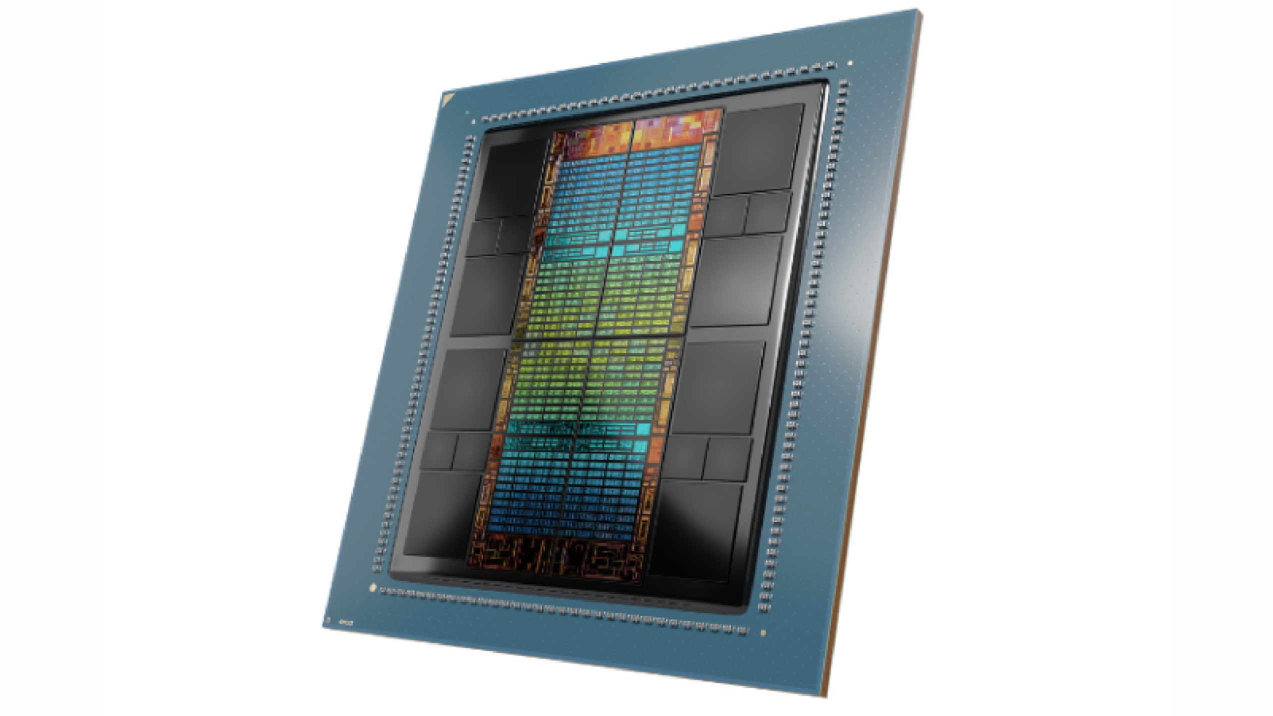
Image is just for representation purposes only
Note: FP8, BF16, FP6, FP4 specs are peak theoretical

The MI355X provides substantial increases in the total compute capabilities of AMD's data centre GPUs, notably a 50% increase over the current-gen MI300X, supporting eight terabytes per second per GPU and 288 gigabytes of total memory capacity.

"When we look at the comparison of what the extra compute means, as well as the memory bandwidth, the support of the lower precision data formats, along with the increased memory footprint, allows us to support up to six times the parameters into a single GPU platform. We're pretty excited about those capabilities and bringing that to the market," Dieckmann said.

It'll be some time though before the MI355X finds its way into data centres, with the hardware expected to start shipping in the second half of 2025.

MI325X updates



While infrastructure providers will have to wait for MI355X, they can get their hands on another AMD GPU a bit sooner: The MI325X, which AMD confirmed will start shipping towards the end of 2024.

Brad McCredie, corporate VP at AMD, said that the MI325X stands apart from the competition thanks to its boosted performance and memory capabilities.

Billed as the next generation of the MI300X, the 325 offers 30% more floating-point operations per second (FLOPS) compared to its predecessor and boasts increased memory for both bandwidth and capacity.

Compared to Nvidia's H200 GPU, AMD's MI325X offers 1.2 to 1.4 times improved performance powering open source AI workloads for models like Meta's Llama 3, Qwen from Alibaba Cloud, and Mixtral from Mistral AI.

"[The MI325X] is going to, we believe, serve our clients very well as they continue to grow," McCredie said during a press briefing.

The MI325X will be made available in the same form factor as the MI300X — with eight GPUs connected on one pod, the UBD8.

McCredie suggested the interconnected eight MI325Xs will provide "the largest memory and the largest amount of memory bandwidth in a pod for the entire industry."

Compared to the Nvidia equivalent, the DGX H200, the AMD pod offers 30% more memory bandwidth and FLOPs with 1.8 times the memory capacity.

McCredie also confirmed during the press briefing that the MI325X can support up to 1,000 watts of power.

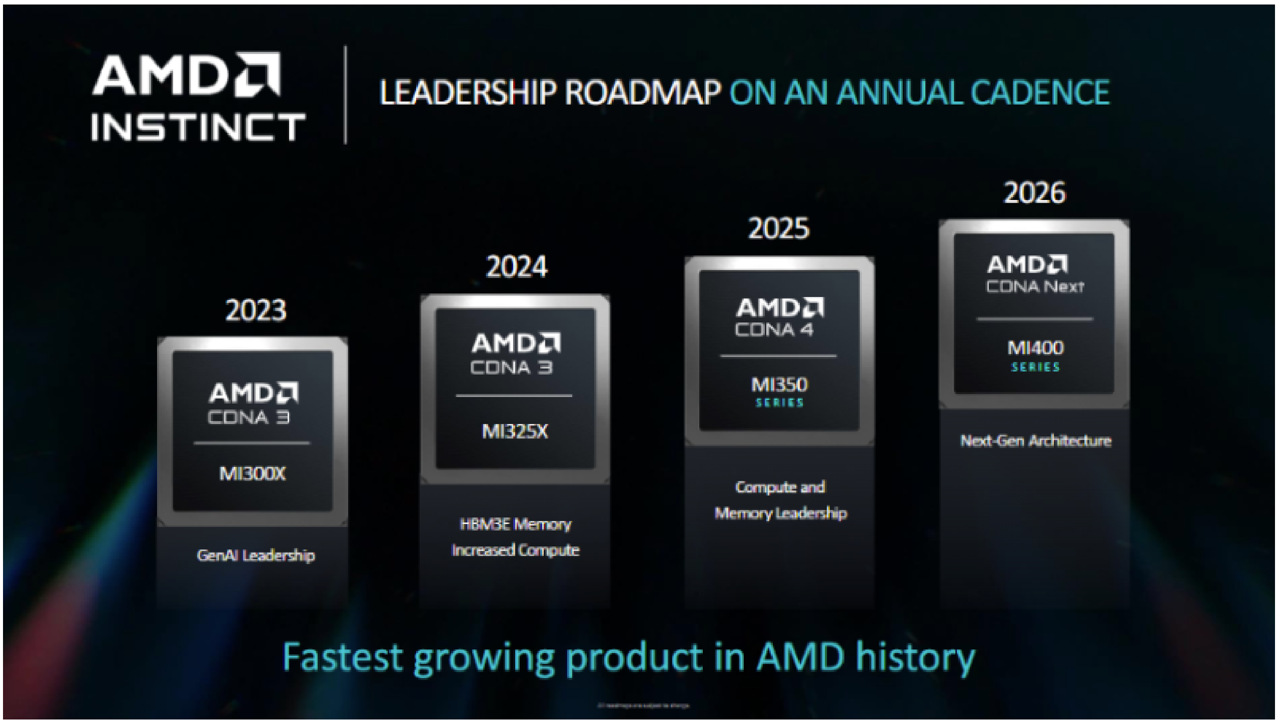
No details were provided on the power consumption capabilities of the MI355X, however, McCredie said it'll be expected to follow industry trends around the time of release.

No word was provided on pricing for the GPUs, either.

Annual cadence: Keeping up with Nvidia's rhythm

Earlier this year at Computex, Nvidia CEO Jensen Huang unveiled the company's plan to launch a powerful new GPU annually, which he described as the "one-year rhythm."

AMD appears to be following suit with what it describes as its "annual cadence" for its Instinct GPUs.



AMD INSTINCT | LEADERSHIP ROADMAP ON AN ANNUAL CADENCE

2023
AMD CDNA 3
MI300X
GenAI Leadership

2024
AMD CDNA 3
MI325X
HBM3E Memory
Increased Compute

2025
AMD CDNA 4
MI350 SERIES
Compute and
Memory Leadership

2026
AMD CDNA Next
MI400 SERIES
Next-Gen Architecture

Fastest growing product in AMD history

Briefly outlining the company's upcoming hardware roadmap, Dieckmann said that the MI325X will go into volume production in the current quarter and will ship next year. The MI355X will also ship in 2025.

Also teased for 2026 is the MI400X — a next-gen GPU with little information known about it.

Dieckmann suggested the MI400X would be a complete architectural overhaul of its Instinct product.

Nothing concrete was revealed about what the MI400X's architecture would look like, but at AMD's Computex update in June, the company said the architecture would be called 'Next.'

“We're going to be adding many new features and capabilities in [the MI400X],” Dieckmann added. “Generally, we'll talk more about that next year as well.”

RELATED STORIES

AMD unleashes next-gen AI processors to power enterprise PCs

AMD launches 5th gen EPYC CPUs to supercharge enterprise AI, cloud workloads

AMD launches new DPUs to boost AI efficiency and network performance in data centres