



AMD Instinct MI325X

Also known as: MI325X OAM · Instinct MI300 Series

The Instinct MI325X is the memory-upgraded refresh of the MI300X: the same CDNA 3 compute (304 compute units) with 256 GB of HBM3E at 6 TB/s and a 1,000 W peak board power, launched 10 October 2024. It stays on the OAM form factor with Infinity Fabric, passively cooled, so it slots into MI300-generation platforms where the board supports the higher power. GPU Smith specs it where memory capacity per GPU is the constraint but a move to the MI350 generation (and its ROCm/datatype changes) is not yet justified.

256 GB HBM3E ECC Memory · 8192-bit	6 TB/s Memory bandwidth · peak	2.61 PFLOPS FP8 matrix · dense; 5.22 PFLOPS with sparsity
1.3 PFLOPS FP16 / BF16 matrix · dense; 2.61 PFLOPS with sparsity	163.4 TFLOPS FP64 matrix · vector 81.7 TFLOPS	1,000 W peak Board power · OAM, 54V UBB

COMPUTE		01/05
Architecture	CDNA 3 · TSMC 5nm / 6nm FinFET	
Compute units	304	
Stream processors	19,456	
Matrix cores	1,216	
Peak engine clock	2,100 MHz	
FP8 matrix	2.61 PFLOPS · dense; 5.22 with structured sparsity	
FP16 / BF16 matrix	1.3 PFLOPS · dense; 2.61 with structured sparsity	
TF32 matrix	653.7 TFLOPS · 1.3 PFLOPS with sparsity	
INT8	2.6 POPS · 5.22 with structured sparsity	
FP32	163.4 TFLOPS · matrix and vector	
FP64	81.7 TFLOPS vector · 163.4 TFLOPS matrix	
Transistors	153 billion	

MEMORY02/05	
Capacity	256 GB HBM3E · full-chip ECC
Interface	8192-bit · 6 GHz memory clock
Bandwidth	6 TB/s · peak
Last-level cache	256 MB · Infinity Cache
INTERCONNECT & FORM FACTOR03/05	
Form factor	OAM module · passive
Host bus	PCIe 5.0 x16
Infinity Fabric links	8 · 128 GB/s peak per link
POWER & THERMAL04/05	
Board power (TBP)	1,000 W peak
Power delivery	54V UBB
Cooling	Passive OAM
SOFTWARE & PLATFORM05/05	
Software stack	ROCm
Launch date	10 October 2024 · per AMD product page
RAS	Page retirement, page avoidance · SR-IOV supported
PROCUREMENT	
Typical lead time No lead time published by AMD; we found no sourced lead-time figure. Not verified. Pricing AMD has not published a price. Not verified: no sourced purchase price found; obtain OEM quotes. Sold as OAM modules on 8-GPU platforms; the baseboard is the unit of purchase. 1,000 W per GPU is 250 W above the MI300X; confirm the host platform's power and cooling are rated for it. Its compute equals the MI300X; the buying reason is memory, so price per GB is the comparison that matters.	

FIELD NOTES

- Compute is identical to the MI300X on paper; the gain is 256 GB at 6 TB/s (33 percent more capacity, 13 percent more bandwidth by arithmetic from AMD's figures). Decode-bound inference benefits; compute-bound work does not.
- 256 GB lets a 70B model in FP16 (about 140 GB of weights) leave over 100 GB for KV cache on a single GPU.
- No FP6 or FP4 hardware; if quantised throughput is the plan, the MI350 series is the generation that adds MXFP4.
- The 1,000 W figure is a peak board rating; validate sustained clocks in your chassis.

QUESTIONS WE GET ON THIS PART

How much memory does the MI325X have?

256 GB of HBM3E at 6 TB/s peak bandwidth with full-chip ECC.

MI325X vs MI300X?

Same CDNA 3 compute. The MI325X has 256 GB HBM3E at 6 TB/s versus 192 GB HBM3 at 5.3 TB/s, and a 1,000 W versus 750 W peak board power.

What is the MI325X power draw?

AMD lists 1,000 W peak board power over 54V UBB.

Is the MI325X replaced by the MI350 series?

The MI350 series (MI355X, MI350P) is the newer CDNA 4 generation with MXFP4 and MXFP6; the MI325X remains on AMD's product list.

REFERENCES

[1] <https://www.amd.com/en/products/accelerators/instinct/mi300/mi325x.html>
[2] <https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html>