



AMD Instinct MI300X

Also known as: MI300X OAM · Instinct MI300 Series

The Instinct MI300X is AMD's CDNA 3 accelerator, an OAM module with 192 GB of HBM3 at 5.3 TB/s, 304 compute units and a 750 W peak board power, launched 6 December 2023. AMD positions it against the H100 on memory capacity. Eight modules join over Infinity Fabric on a standard OAM baseboard. GPU Smith specs it for inference-heavy deployments where 192 GB per GPU and available capacity matter more than newest-generation datatypes; it has no FP4 or FP6.

192 GB HBM3 ECC Memory · 8192-bit	5.3 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	750 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	192 GB HBM3 · full-chip ECC
Interface	8192-bit · 5.2 GHz memory clock
Bandwidth	5.3 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)	750 W peak
Power delivery	54V UBB
Cooling	Passive OAM · air-cooled in the host chassis
SOFTWARE & PLATFORM05/05	
Software stack	ROCm
Launch date	6 December 2023 · 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: we found no sourced purchase price in this review; obtain OEM quotes or cloud rate cards.	
▪ Sold as OAM modules on 8-GPU baseboards; the baseboard is the unit of purchase. ▪ Check ROCm release notes for the exact version your stack needs on CDNA 3.	

FIELD NOTES

- 192 GB per GPU lets a 70B model in FP16 (about 140 GB of weights, by arithmetic) fit on one GPU with room for KV cache, which was the point of the part against 80 GB H100s.
- FP8 is the lowest precision with matrix hardware; there is no FP6 or FP4, so quantised throughput gains stop at 8-bit versus MI350 series.
- FP64 matrix at 163.4 TFLOPS is high for the class, which matters if the same cluster also runs HPC.
- AMD's comparison charts against H100 use AMD's own calculations; benchmark your model rather than relying on ratios.
- At 750 W peak it is the lowest-power OAM part among the AMD Instinct sheets here, which eases chassis cooling.

QUESTIONS WE GET ON THIS PART

How much memory does the MI300X have?

192 GB of HBM3 at 5.3 TB/s peak bandwidth with full-chip ECC.

MI300X vs MI325X?

Same CDNA 3 compute (304 CUs, 2.61 PFLOPS FP8); the MI325X raises memory to 256 GB HBM3E at 6 TB/s and board power to 1,000 W.

What is the MI300X power draw?

AMD lists 750 W peak typical board power, delivered over 54V UBB.

Does the MI300X support FP4?

No. FP8 is its lowest-precision floating-point matrix datatype; MXFP4 and MXFP6 arrive with the MI350 series.

REFERENCES

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

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