



AMD Instinct MI355X

Also known as: MI355X OAM · Instinct MI350 Series

The Instinct MI355X is AMD's flagship 4th-gen CDNA accelerator, an OAM module with 288 GB of HBM3E at 8 TB/s, 256 compute units, and a 1,400 W board power rating, launched 12 June 2025. It is the high-power member of the MI350 series, adds MXFP6 and MXFP4 datatypes, and is built for 8-GPU OAM baseboards joined by Infinity Fabric. GPU Smith specs it as the AMD alternative to B200 and B300 class parts for training and large-model inference, where 288 GB per GPU removes much of the parallelism a smaller-memory part needs.

288 GB HBM3E ECC Memory · full-chip ECC	8 TB/s Memory bandwidth · peak	10.1 PFLOPS MXFP4 / MXFP6 matrix · dense, peak theoretical
5 PFLOPS FP8 matrix · OCP-FP8, dense; 10.1 PFLOPS with sparsity	78.6 TFLOPS FP64 · vector and matrix	1,400 W Board power · typical board power (TBP)

COMPUTE		01/05
Architecture	CDNA 4 · TSMC 3nm / 6nm FinFET	
Compute units	256	
Stream processors	16,384	
Matrix cores	1,024	
Peak engine clock	2,400 MHz	
MXFP4 / MXFP6 matrix	10.1 PFLOPS · dense	
MXFP8 / OCP-FP8 matrix	5 PFLOPS · dense; 10.1 PFLOPS with structured sparsity	
FP16 / BF16 matrix	2.5 PFLOPS · dense; 5 PFLOPS with structured sparsity	
INT8 matrix	5 POPS · dense; 10.1 POPS with structured sparsity	
FP32	157.3 TFLOPS · vector and matrix	
FP64	78.6 TFLOPS · vector and matrix	
Transistors	185 billion	

MEMORY02/05	
Capacity	288 GB HBM3E · full-chip ECC
Bandwidth	8 TB/s · peak
Last-level cache	256 MB · Infinity Cache
RAS	Page retirement, page avoidance · SR-IOV supported
INTERCONNECT & FORM FACTOR03/05	
Form factor	OAM module
Host bus	PCIe 5.0 x16
Infinity Fabric links	7 · 153 GB/s peak per link, scale-up
Scale-out link	1 per GPU · 128 GB/s peak
POWER & THERMAL04/05	
Board power (TBP)	1,400 W
Cooling	Passive & Active · as listed by AMD; confirm air vs liquid per OEM platform
SOFTWARE & PLATFORM05/05	
Software stack	ROCm
Launch date	12 June 2025 · per AMD product page
Cloud availability	Vultr, Oracle · reported, as of March 2026
PROCUREMENT	
Typical lead time No lead time published. Reported cloud availability from Vultr and Oracle by March 2026; we found no sourced lead-time figure for OEM server purchases. Pricing AMD has not published a price. Reported cloud rental ranges from about \$2.95 per GPU-hour (TensorWave, July 2026 data) up; Vultr 8-GPU pods were listed at \$18.32 to \$20.72 per hour (March 2026). No sourced purchase price per GPU. ▪ A 1,400 W module drives facility planning: confirm the OEM's cooling approach (air or direct liquid) and per-rack power before ordering. ▪ Sold as 8-GPU OAM platforms; treat the baseboard as the unit of purchase. ▪ Rent before you buy: several clouds offer 8x MI355X on demand, which de-risks the ROCm port.	

FIELD NOTES

- 288 GB per GPU is the headline advantage over smaller-memory parts: by arithmetic, a 70B model in FP16 (about 140 GB of weights) fits on one GPU.
- Peak PFLOPS are theoretical at the 2,400 MHz clock and at the datatype named; sustained throughput at 1,400 W depends on cooling and power capping.
- MXFP4 and MXFP6 need framework and kernel support. Verify your serving stack uses them on CDNA 4 before counting on the 10.1 PFLOPS figure.
- The scale-out link is one 128 GB/s port per GPU, so multi-node fabric design (NIC choice, topology) is a separate decision from the node.
- ROCm maturity is the risk, not the silicon: pilot your exact model and framework versions on rented hardware first.

QUESTIONS WE GET ON THIS PART

How much HBM does the MI355X have?

288 GB of HBM3E at 8 TB/s peak bandwidth, with full-chip ECC.

What is the MI355X power draw?

AMD lists 1,400 W typical board power for the OAM module.

MI355X vs B200?

On AMD's published peaks the MI355X has more memory (288 GB) and bandwidth (8 TB/s). Delivered performance depends on software; AMD's comparison figures are its own calculations, so benchmark your workload.

Can I buy a single MI355X?

It is an OAM module sold on 8-GPU platforms through OEMs; AMD's PCIe option in this generation is the MI350P.

REFERENCES

[1] <https://www.amd.com/en/products/accelerators/instinct/mi350/mi355x.html>
[2] <https://discover.vultr.com/mi355x-now-available>
[3] <https://gpus.io/en/gpus/mi355x>
[4] <https://deploybase.ai/articles/amd-mi355x-cloud-pricing-where-to-rent-and-how-much-it-costs>