



AMD Instinct MI355X Platform

8x Instinct MI355X CDNA 4 baseboard 2025

OVERVIEW

40.3 FP8 PFLOPS Dense	2.3 TB Aggregate Memory	64.0 TB/s Aggregate Bandwidth	— Aggregate Power
--------------------------	----------------------------	----------------------------------	----------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	0.6	78.6	—
FP32	1.3	157.3	—
FP16	20.1	2516.6	—
BF16	20.1	2516.6	—
FP8	40.3	5033.2	—
FP4	80.5	10066.3	—
INT8	40.3	5033.2	—
FP6	80.5	10066.3	—

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	Instinct MI355X OAM
Count	8
Memory per GPU	288 GB
Bandwidth per GPU	8.0 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	1400 W
FP8 Efficiency	—
Aggregate Bandwidth	64.0 TB/s

SYSTEM DETAILS

Vendor: AMD	Family: Instinct Platform	Architecture: CDNA 4
Form Factor: baseboard	Release Year: 2025	Process Node: 3nm/6nm

INTERCONNECT

4th Gen AMD Infinity Fabric, eight fully connected OAM modules on an OCP UBB design

NOTES

AMD's eight-GPU building block, integrating eight fully connected MI355X OAM modules on an industry-standard OCP universal baseboard over fourth-generation Infinity Fabric. AMD publishes the platform totals of 2.3 TB of HBM3E and 64 TB/s of aggregate memory bandwidth, both of which are

exactly eight times the per-GPU figures, which is why the precision values here are also eight times AMD's published per-GPU specification rather than a separate platform table, since AMD provides none. Sparse figures appear only for the four datatypes where AMD publishes a structured-sparsity number, which are FP8, FP16, BF16 and INT8; AMD gives no sparsity figure for MXFP4 or MXFP6, so neither does this row. Supported in both air-cooled UBB servers and direct liquid cooled platforms. AMD publishes no platform power figure.