

AMD Instinct MI350P

Also known as: MI350P PCIe · Instinct MI350P

The Instinct MI350P is AMD's CDNA 4 accelerator in a dual-slot, full-height, passively cooled PCIe card, announced in May 2026. It carries 144 GB of HBM3E at 4 TB/s and 128 compute units, in a 600 W envelope that can be configured down to 450 W, so it fits standard air-cooled 19-inch servers instead of requiring an OAM baseboard. Cards talk to each other over PCIe 5.0 only; AMD configurations run from one to eight cards per node, reported. It is the AMD answer to the PCIe H200 NVL and RTX PRO 6000 Server Edition class of part, and GPU Smith specs it for single-node private inference where HBM capacity and ROCm support matter more than a scale-up fabric.

144 GB HBM3E ECC Memory · 4096-bit interface	4 TB/s Memory bandwidth · peak	4.6 PFLOPS MXFP4 / MXFP6 matrix · dense, peak theoretical
2.3 PFLOPS FP8 matrix · OCP-FP8, dense; 4.6 PFLOPS with sparsity	600 W max Board power · 450 W configurable	Dual-slot FHFL, passive Form factor · PCIe 5.0 x16, 12V-2x6

COMPUTE		01/05
Architecture	CDNA 4 · TSMC 3nm / 6nm FinFET	
Compute units	128	
Stream processors	8,192	
Matrix cores	512	
Peak engine clock	2,200 MHz	
MXFP4 / MXFP6 matrix	4.6 PFLOPS · dense	
MXFP8 / OCP-FP8 matrix	2.3 PFLOPS · dense; 4.6 PFLOPS with structured sparsity	
FP16 / BF16 matrix	1.15 PFLOPS · dense; 2.3 PFLOPS with structured sparsity	
INT8 matrix	2.3 POPS · dense; 4.6 POPS with structured sparsity	
FP32	72 TFLOPS · vector and matrix	
FP64	36 TFLOPS · vector and matrix	
Transistors	73 billion	

PROCUREMENT

Typical lead time

Reported OEM availability from July 2026, starting with Dell PowerEdge XE7745 and R7725. No published lead time; quotes come through OEMs and distributors (reported: Dell, HPE, Supermicro, Gigabyte).

Pricing

AMD has not published a price or MSRP. Third-party estimates of \$30,000 to \$40,000 per card circulate from a secondary source and are not confirmed by AMD or the trade press we checked. Treat any figure as unverified until you hold an OEM quote.

- Sold as part of qualified OEM servers rather than as a boxed retail card; confirm chassis, PSU and airflow qualification before ordering.
- The card uses a 12V-2x6 connector, so server power distribution must supply it; check the OEM's cabling kit.
- Because there is no published price, budget by quote and ask for the 450 W configuration if rack power is the constraint.

FIELD NOTES

- No high-speed card-to-card link: tensor-parallel splits across cards ride PCIe 5.0 through the host. Run one model per card, or pipeline-parallel across cards, rather than wide tensor parallelism in one node.
- 144 GB on a single card is the point. Many 70B-class models in FP8, and some larger ones in MXFP4, fit without any parallelism, which simplifies serving.
- Peak figures are theoretical matrix throughput at the engine clock. AMD's own page does not publish delivered numbers; compare with measured results on your model, not the FP4 headline.
- ROCm is the software decision, not the silicon. Check that your serving stack (vLLM, SGLang, your kernels) is supported on CDNA 4 before committing.
- The 450 W configuration trades some peak for rack power headroom; confirm the throughput hit on your workload before standardising on it.

QUESTIONS WE GET ON THIS PART

How much memory does the MI350P have?

144 GB of HBM3E on a 4096-bit interface with 4 TB/s peak bandwidth, with full-chip ECC.

How much does the MI350P cost?

AMD has not published a price. Unconfirmed third-party estimates run \$30,000 to \$40,000 per card. Get an OEM quote.

MI350P vs MI355X?

The MI355X is the OAM module: 288 GB HBM3E, 8 TB/s, 256 compute units and 1,400 W, built for 8-GPU baseboards with Infinity Fabric. The MI350P is a half-scale PCIe card (144 GB, 4 TB/s, 128 CUs, 600 W) for ordinary servers with no scale-up fabric.

Does it need special server hardware?

Yes. It is passively cooled and draws up to 600 W through a 12V-2x6 connector, so it belongs in an OEM-qualified chassis with enough airflow and power delivery.

REFERENCES

- [1] <https://www.amd.com/en/products/accelerators/instinct/mi350/mi350p.html>
- [2] <https://www.theregister.com/a/5231481>
- [3] <https://tweaktown.com/news/111491/amd-launches-the-instinct-mi350p-gpu-with-144gb-of-hbm3e-and-a-600w-tbp/index.html>