AMD INSTINCT™ MI350P PCIe® CARD

BUILT FOR THE INFRASTRUCTURE YOU HAVE.
READY FOR THE AI YOU NEED

AMD INSTINCT

PURPOSE-BUILT FOR ENTERPRISE AI

The AMD Instinct™ MI350P PCIe® card brings generative and agentic AI workloads into existing data centers. Built on a standard PCIe form factor, the AMD Instinct MI350P integrates seamlessly into existing infrastructure without the cost or complexity of purpose-built AI systems. The result is a straightforward path to exceptional AI performance, cost savings, and an enterprise-ready software stack designed to reduce time to value.

UPGRADE YOUR INFRASTRUCTURE, NOT YOUR DATA CENTER

Designed for the PCIe® form factor in a full-height, full-length, dual-slot configuration, the AMD Instinct MI350P PCIe® card drops into mainstream air-cooled servers—no specialized cooling, no rack redesigns, no building from scratch. Just the Gen AI and agentic AI performance business demands, deployed in current data centers.

The efficiency gains are substantial. The AMD Instinct MI350P PCIe® card connects via x16 channel Gen 5 to deliver leadership estimated peak 4600 TFLOPS at the most efficient MXFP6 and MXFP4 precision data formats for enterprise AI workloads as well as leadership 144 GB of HBM3E capacity. These offer outstanding throughput per server to deploy and scale Gen AI and agentic AI workloads within current rack infrastructure and power-and-cooling envelopes. Up to eight AMD Instinct MI350P PCIe® cards can be configured per server, and to meet power or cooling constraints, TDP can be reduced from 600W to 450W.

LEADERSHIP TCO FOR ENTERPRISE AI

Hyperscalers optimize for raw scale. Enterprises optimize for value. The AMD Instinct MI350P PCIe® card is built for the latter—delivering leadership total cost of ownership through every dimension of AI deployment. With the AMD Instinct MI350P, enterprises get exceptional performance in a low overhead form factor that integrates into existing enterprise data center infrastructure, giving finance and infrastructure teams a compelling reason for adoption alongside the outstanding performance numbers.

Unlike solutions that bundle proprietary software at a premium, the AMD Instinct MI350P PCIe® card is backed by the open-source AMD™ Enterprise AI reference stack—alleviating software licensing fees and restoring local control. Combined with exceptional performance,

HPC PEAK PERFORMANCE (ESTIMATED)		W/SPARSITY
FP16 VECTOR (TFLOPS)	72	N/A
FP16 MATRIX (TFLOPS)	1150	2300
BF16 (TFLOPS)	1150	2300
INT8 MATRIX (TOPS)	2300	4600
MXFP8 (TFLOPS)	2300	N/A
OCP-FP8 (TFLOPS)	2300	4600
MXFP6 (TFLOPS)	4600	N/A
MXFP4 (TFLOPS	4600	N/A

HPC PEAK PERFORMANCE (ESTIMATED)	
FP64 VECTOR (TFLOPS)	36
FP32 VECTOR (TFLOPS)	72
FP64 MATRIX (TFLOPS)	36
FP32 MATRIX (TFLOPS)	72

DECODERS AND VIRTUALIZATION	
DECODERS†	2 groups for HEVC/H.265,AVC/H.264, VP9, or AV1
JPEG/MJPEG CODEC	20 cores, 10 cores per group
GPU PHYSICAL PARTITIONS	Up to 4 @ 36GB
MEMORY PARTITIONS	1

SPECIFICATIONS	
FORM FACTOR	FHFL 2-slot PCIe CEM Card
LITHOGRAPHY	TSMC 3nm/6nm FinFET
I/O DIES (IODS)	1
GPU COMPUTE UNITS	128
MATRIX CORES	512
STREAM PROCESSORS	8,192
PEAK ENGINE CLOCK	2.2 GHz
MEMORY CAPACITY	144GB HBM3e
MEMORY BANDWIDTH	4 TB/s
MEMORY INTERFACE	4096 bits
AMD INFINITY CACHE™ (LAST LEVEL)	128 MB
I/O INTERCONNECT	1 PCIe® Gen 5 x16 (128 GB/s)
RAS FEATURES	Full-chip ECC memory, page retirement, page avoidance
MAXIMUM TBP	600W (configurable to 450W)

†Video codec acceleration (including at least the HEVC (H.265), H.264, VP9, and AV1 codecs) is subject to change and not operable without inclusion/installation of compatible media players. GD-176

these accelerators simplify deployment and reduce costs, making enterprise AI not just capable, but economically sustainable at scale.

ENTERPRISE AI, OPEN BY DESIGN

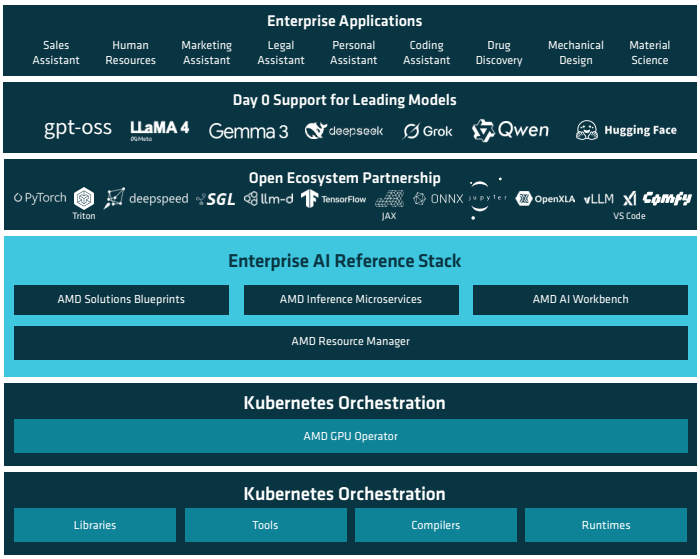
Adopting the AMD Instinct MI350P shouldn't mean rebuilding AI software stacks from scratch. The accelerator pairs with a fully integrated, open, standards-based AI platform designed to make migration seamless—facilitating moves from current GPU infrastructure to the AMD Instinct MI350P without code rewrites, retraining, or rearchitecting deployments. Cross-platform interoperability means existing workloads translate easily while adopting the latest AMD Instinct technology.

A proven on-premises foundation is in place to support moving to the AMD Instinct MI350P. Prevalidated solutions, including AMD Inference Microservices (AIMs) within the AMD Enterprise AI reference stack, simplify the hosting of large language models and accelerate time to production (see figure). Open-source, standards-based software eliminates licensing fees and gives teams full visibility into the code they run—a requirement for enterprises operating under strict compliance and security mandates.

Enterprises don't have to go it alone. The AMD Instinct MI350P is validated and supported by an ecosystem of leading OEMs and ISVs, giving enterprises the confidence that comes with broad industry backing.

LEARN MORE

For more information, visit AMD.com/INSTINCT



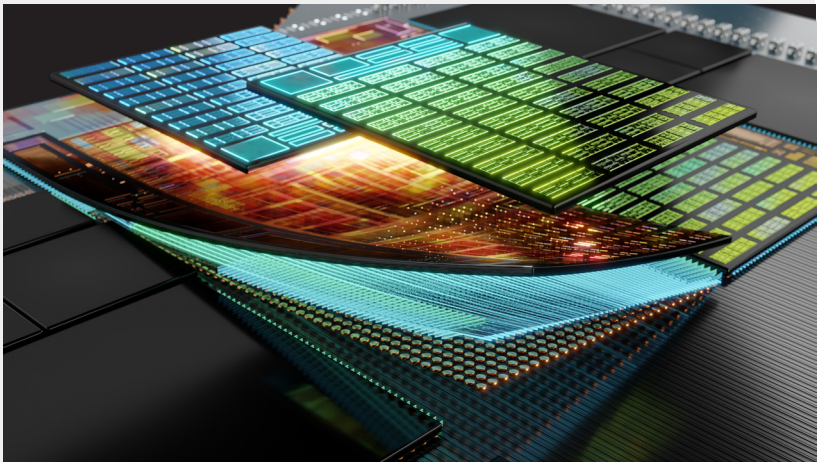
AMD ENTERPRISE AI SOFTWARE

In combination with the AMD GPU Operator, the AMD Enterprise AI reference stack offers a reference architecture for how to build your own custom development stack integrating preferred ecosystem partner software to accelerate the deployment of enterprise AI solutions on AMD based hardware. These tools, designed to address the complexities of modern AI workflows, enable seamless integration, optimized performance and scalable operations.

MULTI-CHIP ARCHITECTURE

The AMD Instinct MI350P PCIe® card is based on AMD CNDA™4 multi-chip architecture with a 3nm process technology to enable dense compute and high-bandwidth memory integration. Each AMD Instinct MI350P PCIe® card includes:

- Four accelerated compute dies (XCDs) with 32 compute units (CUs), 32 KB of L1 cache per CU, 4 MB shared L2 cache shared across CUs, and 128 MB of AMD Infinity Cache™ shared across the four XCDs.
- Designed to enhance AI operations, the accelerator supports not just native BF/FP16, FP8, and INT8 with hardware support for sparsity, but also MXFP8, MXFP6, and MXFP4 data types and enhanced computational throughput.
- Two supported decoders for HEVC/H.265, AVC/H.264, VP9, or AV1, each with an additional 20-core JPEG/MPEG CODEC. The open-source AMD ROCm™ Augmented Library (rocAL™) decodes and augments images and video for deep-learning applications.
- 144 GB of HBM3E memory
- Future SR-IOV support for up to four partitions



LE-93401-00 05/26