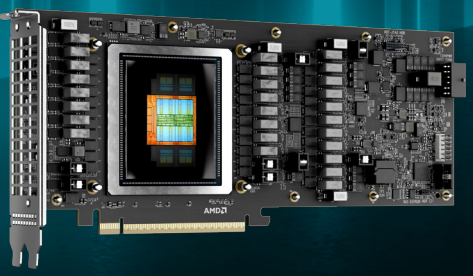
5 REASONS WHY YOU SHOULD CHOOSE *AMD INSTINCT™ MI350P PCIe® CARDS* TO DEVELOP YOUR AI STACK. YOUR WAY, TODAY.



AT A GLANCE

Generative and agentic artificial intelligence (AI) has put enterprises under tremendous pressure. The demand to adopt AI across the organization is ever-increasing, but the supply of AI compute is constrained. While leveraging public clouds is ideal for some workloads, others benefit from keeping compute locally within the enterprise. Integrating GPU capabilities into existing enterprise infrastructure – without costly overhauls or new build-outs – is fast becoming a business necessity. What if you could meet many of the tremendous demands of AI while continuing to drive low total cost of ownership (TCO) with a flexible and adaptive infrastructure? AMD Instinct™ MI350P PCIe® cards make this possible.

AMD INSTINCT MI350P SERIES GPUS: THE AMD INSTINCT AI JOURNEY CONTINUES

The AMD Instinct MI350P PCIe based card enables enterprises to deploy and scale generative AI and agentic AI workloads via a full-height, full-length (10.5 inches) 2-slot PCIe card built for the infrastructure you already own and the AI performance you need. AMD Instinct MI350P PCIe cards deliver exceptional performance, leadership TCO and simplified deployment with an enterprise-ready AI stack. With this PCIe form factor, AMD extends the Instinct GPU series further into mainstream enterprise-optimized AI platforms. Here's how.

1 AI PERFORMANCE WITHIN YOUR EXISTING INFRASTRUCTURE

DEPLOY AI WITHIN EXISTING RACK INFRASTRUCTURE.

The AMD Instinct MI350P PCIe card enables enterprises to deploy and scale generative AI and agentic AI workloads within your current rack infrastructure and power-and-cooling envelope, while maximizing your throughput per server. Get state-of-the-art MXFP6 and MXFP4 precision performance for general enterprise AI workloads.

2 LEADERSHIP TCO FOR ENTERPRISE AI

DON'T JUST SCALE AI. BUILD ROI.

AMD Instinct MI350P GPUs deliver leadership TCO through lower MXFP6 and MXFP4 precision, more HBM3E memory, more raw performance, and an open low and no-cost software ecosystem, for reduced software licensing costs and maximum performance per watt for generative AI and agentic AI workloads.

3 AN OPEN, ENTERPRISE-READY AI STACK

IT'S A FULL AI STACK THAT'S BACKED BY THE BEST.

AMD combines AMD Instinct GPUs, a fully integrated open enterprise AI reference stack and AMD Inference Microservices to deliver a streamlined, enterprise-ready platform that accelerates AI deployment and simplifies operations at scale.

4 A SCALABLE PORTFOLIO WITH A UNIFIED DEVELOPMENT ENVIRONMENT

ENABLE EFFORTLESS AI DEPLOYMENT WITH MINIMAL CODE CHANGES.

AMD facilitates the adoption and use of multiple acceleration platforms and cross-platform AI development with the AMD Enterprise AI reference stack which operates seamlessly with the broad ecosystem of development and programming toolset. You can focus on fostering collaboration, continuous improvement and accelerating development velocity without increasing spend.

5 BUILT-IN SECURITY FEATURES FOR AI DEPLOYMENT

GET ADVANCED SECURITY TO PROTECT YOUR AI MODELS AND DATA.

AMD Instinct MI350P GPUs combined with AMD EPYC™ 9005 Series CPUs change the threat equation by securing applications and data while they are in use inside encrypted, trusted execution environments (TEEs). With confidential computing, sensitive data stays private and encrypted making the combination of AMD Instinct MI350P GPUs and EPYC 9005 Series CPUs ideal for cloud AI, enterprise and mission-critical workloads.

TECHNICAL DEEP DIVE

#1 AI PERFORMANCE WITHIN YOUR EXISTING AI INFRASTRUCTURE

- AMD CDNA™ 4 architecture features 128 compute units and enhanced AI support delivering powerful performance for MXFP8, MXFP6, and MXFP4 calculations and their matrix equivalents.[1]
- Up to four isolated partitions per GPU with 36GB HBM3e each for right-sizing accelerated workloads.
- Get 600W of passive air-cooled Total Board Power (TBP), configurable to 450W TBP for mainstream servers.
- Leverage impressive performance for PCIe-based GPUs with over 4.6PFLOP peak performance at MXFP6 precision and 4.6PFLOP peak performance at MXFP4 precision.

#2 LEADERSHIP TCO FOR ENTERPRISE AI

- Optimized for high-throughput, power-efficient AI inference with little to no infrastructure changes.
- Realize low total cost per token with no need for software licensing fees using the open-source AMD Enterprise AI Suite.
- Get leadership HBM3E memory for PCIe-based GPUs of up to 144GB with up to 4.0 TB/s theoretical peak bandwidth to power large LLM models.

#3 AN OPEN, ENTERPRISE-READY REFERENCE STACK

- Simplify on-prem deployment with the AMD Instinct™ MI350P PCIe card and the AMD enterprise-ready AI software with pre-validated blueprints, to simplify on-prem hosting of LLMs and accelerate AI deployment.
- Use pre-integrated AMD Inference Microservices to help reduce operational overhead, allowing teams to deploy and manage AI applications efficiently.
- Work within a broad ecosystem validated and supported by leading OEMs and ISVs, including Cisco, Cohere, Dell, HPE, Lenovo, Nutanix, Red Hat, Seekr, Supermicro, Uniphore, VMware and more.

#4 A SCALABLE PORTFOLIO WITH A UNIFIED GPU DEVELOPMENT ENVIRONMENT

- Build AI services and applications rapidly, reliably, and affordably with the free ROCm™ software ecosystem and programming toolset that spans all AMD GPU platforms.
- Enable seamless scaling from local development to data center environments to deploy flexible, resource-efficient AI services that grow in alignment with business use and goals.
- Leverage portability between frameworks such as Caffe2, MXNet, PyTorch and TensorFlow to reduce vendor lock-in, lower switching costs and mitigate long-term financial risk.
- Adopt open source to align your technology strategy with cost discipline, achieving sustainable cost advantages at scale.

#5 ADVANCED SECURITY TO PROTECT AI MODELS, DATA AND INTEGRITY

- AMD Instinct MI350P GPUs help ensure trusted firmware, verify hardware integrity, enable secure multi-tenant GPU sharing, and encrypt GPU communication—helping enhance reliability, scalability, and data security for cloud AI and mission-critical workloads.
- Use AMD Instinct MI350P GPU features such as Device Secure Boot, Secure Update and Recovery to help ensure only trusted firmware runs, while Platform-Level DICE Identity and Attestation verify GPU authenticity to prevent unauthorized access.
- Combine AMD Instinct MI350P GPUs with EPYC 9005 Series CPUs to embrace Confidential AI using AMD SEV.
- Work within a broad ecosystem validated and supported by leading OE.

AMD INSTINCT MI350P PCIe® CARD

ENTERPRISE AI, READY WHERE YOU ARE

LEARN MORE AT AMD.COM/INSTINCT