

# 5 REASONS WHY YOU SHOULD CHOOSE AMD INSTINCT™ MI350X AND MI355X GPUs TO ADVANCE GENERATIVE AI IN THE DATA CENTER

## AT A GLANCE

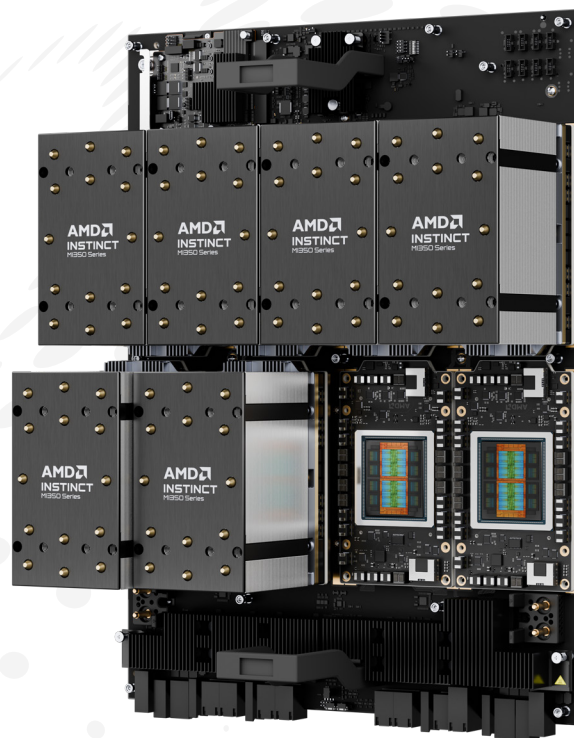
Generative artificial intelligence (AI) has put data centers under tremendous pressure. Because it trains on massive data sets, it forces you to find ways to make your infrastructure more powerful and efficient at the same time. As enterprises continue to explore generative AI development and integrate AI everywhere they can while simultaneously increasing their deployments of resource-intensive high-performance computing (HPC) applications, the demands on the data center will only increase.

What if you could get more GPU performance in a drop-in platform with leadership compute unit counts and advanced high-bandwidth, high-capacity memory? What if you could develop innovative AI and HPC applications on multiple acceleration platforms without worrying about compatibility, operational complexity, specialized design requirements or looming security issues?

## AMD INSTINCT™ MI350 SERIES GPUS: FUTURE-READY YOUR DATA CENTER FOR AI ADVANCES

**AMD Instinct MI350 Series GPUs** set a new standard for generative AI and HPC applications in data centers. Built on the cutting-edge 4th Gen AMD CDNA™ architecture, these GPUs deliver exceptional efficiency and performance for training massive AI models, high-speed inference and complex HPC workloads.

At the same time, the AMD ROCm™ platform, the industry's premier open-source alternative for AI and HPC, allows AMD Instinct Series GPUs to easily integrate with multiple operating environments and platforms, enabling effortless AI and HPC deployment with minimal code changes.



1

**LEADERSHIP PERFORMANCE FOR AI AND HPC*****Discover the possibilities of accelerated performance at scale.***

AMD introduces the new MI350 Series GPUs, consisting of the AMD Instinct MI350X GPU, delivering leading performance for AI inference and training at scale, as well as the AMD Instinct MI355X GPU, designed for high-density, efficient and scalable infrastructure. With enhanced memory capacity, robust datatype support and multiple cooling options, the MI350 Series GPUs handle diverse AI workloads efficiently, from large-scale inferencing to generative AI model training.

---

2

**SEAMLESS SCALABILITY AND DEPLOYMENT*****Ease the configuration of GPU-accelerated workloads such as machine learning and generative AI.***

Air cooled AMD Instinct MI350 Series GPUs integrate seamlessly with prior generation AMD Instinct MI300 Series platforms (MI300X and MI325X) and competitive infrastructures, offering cost-effective performance upgrades for evolving infrastructure demands. They work with the AMD GPU Operator, which simplifies the deployment and management of AMD Instinct GPU accelerators within Kubernetes clusters.

---

3

**TRUSTED BY AI LEADERS*****Join the community of CSPs, OEMs and ODMs driving next-gen AI at scale.***

AMD Instinct Series GPUs power AI, driving adoption across leading cloud service providers (CSPs) and original equipment manufacturer (OEMs) partners. AMD delivers scalable, efficient AI solutions, empowering customers with greater choice and a competitive edge in AI innovation. Their adoption underscores their proven performance, efficiency and scalability and reinforces AMD as a trusted supplier for next-generation AI.

---

4

**PROVEN, OPEN, LIMITLESS AI SOFTWARE*****Enable effortless AI and HPC deployment with minimal code changes and maximum flexibility.***

AMD facilitates the adoption and use of multiple acceleration platforms and cross-platform HPC and AI development with the ROCm 7.0 open-source software ecosystem and programming toolset. A collection of drivers, development tools and APIs, it extracts the best performance from AMD Instinct GPUs while maintaining compatibility with industry software frameworks.

---

5

**BUILT-IN SECURITY FOR AI AND HPC DEPLOYMENTS*****Get advanced security to protect AI models, data and system integrity.***

AMD Instinct MI350 Series GPUs deliver robust security features to protect AI models, data and system integrity. They help ensure only trusted firmware runs, verify hardware authenticity and encrypt high-speed GPU communication. These protections make them ideal for cloud AI, enterprise and mission-critical workloads.

---

## TECHNICAL DEEP DIVE

### #1 DISCOVER THE POSSIBILITIES OF ACCELERATED PERFORMANCE AT SCALE

- AMD Instinct MI350 Series GPUs feature powerful and energy-efficient cores, maximizing performance per watt to drive the next era of AI and HPC innovation. With newly **expanded support for FP6 and FP4 datatypes, a massive 288GB of HBM3E memory and up to 8TB/s bandwidth**, they maximize computational throughput, memory bandwidth utilization and energy.
- AMD Instinct MI350X and MI355X 8xGPU platforms offer up to **36.9 and 40.3 PFLOPs bfloat16 performance**, respectively, with structured sparsity, and with new expanded support for new datatypes, they offer up to **73.8 and 80.5 PFLOPs MXFP6/MXFP4 matrix performance**, respectively. [MI350-010A](#)
- AMD Instinct MI350X and MI355X 8xGPU platforms offer up to **2.05x and 2.2x the peak theoretical MXFP6 matrix PFLOPs performance**, respectively, vs. the NVIDIA HGX Blackwell B200 GPU. [MI350-010A](#)
- The AMD Instinct MI355X GPU offers up to **2x the peak theoretical MXFP6 matrix PFLOPs performance** vs. the NVIDIA Grace Blackwell GB200 GPU. [MI350-008A](#)
- AMD Instinct MI350X and MI355X GPUs offer up to **77% and 93% higher peak theoretical OCP-FP8 matrix performance**, respectively, than the previous-generation AMD Instinct MI300X GPU. [MI350-005](#)
- AMD Instinct MI350X and MI355X GPUs with MXFP4 matrix datatype are projected to deliver up to **7x and 7.7x the performance**, respectively, **on AI and training workloads** compared to the AMD Instinct MI300X GPU using FP16 matrix. [MI350-005](#)
- AMD Instinct MI350 Series GPUs offer up to **2.04x the memory capacity** of the NVIDIA H200 GPU and up to **60% more memory capacity** than the NVIDIA HGX B200 GPU. [MI350-001](#)
- AMD Instinct MI350 Series GPUs offer up to **50% more memory capacity** than previous generation AMD Instinct MI300X GPUs. [MI350-002](#)
- AMD data center solutions support AI and HPC workloads at tremendous scale with enhanced efficiency. AMD powers **5 of the top 10 supercomputer systems** on the Top500 list<sup>1</sup> and more than half of the top 20 supercomputer systems on the Green500 list.<sup>2</sup>

### #2 EASE THE CONFIGURATION OF GPU-ACCELERATED WORKLOADS SUCH AS MACHINE LEARNING AND GENERATIVE AI

- AMD Instinct MI350 Series GPUs optimize efficiency, bandwidth and energy use for fast, power-efficient AI inference and training.

- Enhanced processing with next-gen FP6 and FP4 capabilities enables AMD Instinct MI350 Series GPUs to deliver exceptional performance for generative AI models, pushing the boundaries of AI acceleration.
- Support for matrix sparsity further helps optimize AI training and inference, helping scale AI solutions across data centers.

### #3 JOIN THE COMMUNITY OF CSPs, OEMs AND ODMs DRIVING NEXT-GEN AI AT SCALE

- With support for third-party partners such as Clarifai, ClearML, Fireworks, Flex.ai, Lamini, Rapt.ai, RHEL, OpenShift and UbiOps, AMD provides customers with flexibility, choice and best-in-class tools to accelerate AI innovation.

### #4 SMOOTH YOUR AI AND HPC DEPLOYMENTS WITH MINIMAL CODE CHANGES AND MAXIMUM FLEXIBILITY

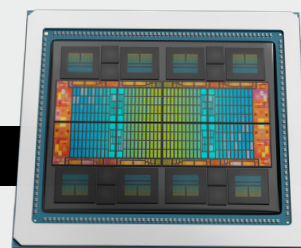
- ROCm 7.0 lets you develop applications in an open-source system so that the software is portable, allowing movement among GPUs from different vendors or among different inter-GPU connectivity architectures—all in a device-independent manner, unlike NVIDIA's proprietary hardware solutions.
- The latest ROCm enhancements optimize AI inference, training and framework compatibility, delivering high throughput and ultra-low latency for workloads such as Natural Language Processing (NLP), computer vision and beyond. ROCm is well-suited to GPU-accelerated HPC, AI, scientific computing and computer-aided design (CAD).
- AMD Instinct MI350 Series GPUs easily integrate with JAX, Kokkos, ONNX, PyTorch, Raja, Runtime, TensorFlow, Triton, vLLM and more, enabling effortless AI and HPC deployment with minimal code changes.
- With Day 0 support, ROCm software helps optimize AI workloads from Hugging Face, Meta, OpenAI and others, delivering high throughput, low latency and peak efficiency for NLP, computer vision and generative AI inference—accelerating AI innovation at scale

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

- AMD Instinct MI350 Series GPUs offer Device Secure Boot and 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.
- Infinity Fabric™ Link Security helps ensure high-speed GPU-to-GPU communication that's safe.

**AMD**  
together we advance AI

LEARN MORE AT [AMD.COM/INSTINCT](https://amd.com/instinct) AND [AMD.COM/ROCM](https://amd.com/rocm).



<sup>1</sup> Top500 list, November 2024, <https://top500.org/lists/top500/2024/11/>.

<sup>2</sup> Green500 list, November 2024, <https://top500.org/lists/green500/list/2024/11/>.

©2025 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD arrow, AMD Instinct, Infinity Fabric, AMD Infinity Platform, AMD CDNA, ROCm and combinations thereof, are trademarks of Advanced Micro Devices, Inc. NVIDIA is a trademark of NVIDIA Corporation in the U.S. and/or other countries. The OpenMP name and the OpenMP logo are registered trademarks of the OpenMP Architecture Review Board. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.