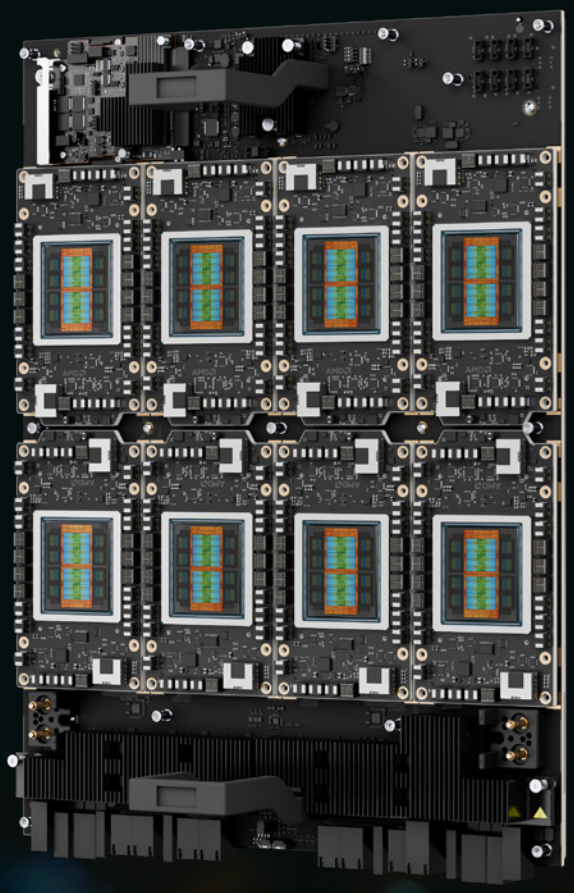
AMD

OPEN. POWERFUL. *READY.*



AMD INSTINCT™ GPUs AND AMD ROCm™ ECOSYSTEM: A COMPLETE AI PLATFORM BUILT ON OPEN-SOURCE INNOVATION

THE FOUNDATION: AMD INSTINCT™ GPUs AND AMD ROCm™ ECOSYSTEM

Purpose-built hardware and an open software platform, working as two halves of one complete solution.

AMD INSTINCT GPUs

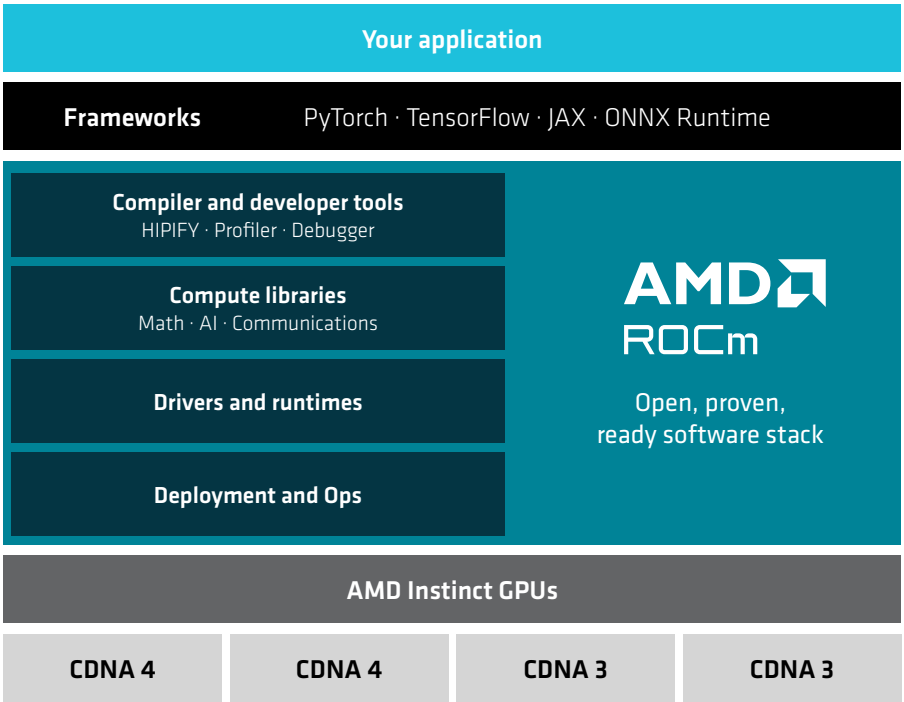
PURPOSE-BUILT FOR AI AT SCALE

AMD Instinct MI350 Series GPUs are engineered specifically for the demands of generative AI, LLMs, and HPC workloads. Built on 4th Gen AMD CDNA™ architecture, they deliver massive memory, extreme bandwidth, and support for the precision formats modern AI and HPC require.

AMD ROCm 7 SOFTWARE

THE OPEN SOFTWARE ENGINE

AMD ROCm **is a no cost, open-source software** platform with no licensing fees and no proprietary black boxes. It provides everything from low-level GPU drivers up to the AI frameworks your teams are using. ROCm 7 software consists of a rich collection of drivers, libraries, and developer tools that enable GPU programming from low-level kernels up to end-user applications, and it's customizable to meet specific needs.



Unleash the full performance of AMD Instinct™ GPUs with open-source AMD ROCm™ software.

Visit amd.com/rocm to learn more.

AMD INSTINCT MI350 SERIES GPUs

288 GB
HBM3E per GPU[1]

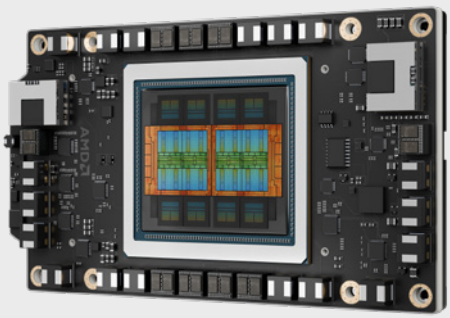
8 TB/S
memory bandwidth[1]

MXFP6, MXFP4
support

FP64 PERFORMANCE
for HPC

UP TO 4.2X BETTER
inference performance on an 8x GPU platform using Llama 3.1 405B[2]

UP TO 3.5X BETTER
training performance on an 8x GPU platform using Llama 3 70B[3]



UP AND RUNNING FASTER WITH A CLEAR PATH FORWARD

OPEN ECOSYSTEM DEPLOYMENT SIMPLIFIED

Built on AMD ROCm™ software, the AMD Enterprise AI Suite enables organizations to deploy AI faster with less complexity. Optimized for AMD Instinct™ GPUs, it provides an enterprise-ready platform with **Day-0 framework support, optimized libraries, and standards-based runtimes,** reducing integration effort and accelerating both training and inferencing.

The suite includes tools that streamline the full AI lifecycle:

- **AMD Inference Microservices (AIMs)** deliver pre-optimized models to accelerate application deployment.
- **AMD AI Workbench provides** a unified environment for model development, fine-tuning, and scalable inference.
- **AMD Resource Manager** enables efficient orchestration and GPU utilization across cluster environments.

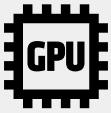
Together, these capabilities extend the **open, high-performance AMD ROCm foundation—combining portability, scalability, and operational control** to help organizations move from development to production with confidence.

ACCELERATE AI WORKLOADS

Speed time to insight with pre-built containers and deployment guides for AI and HPC, available through the AMD Infinity Hub. These resources enable rapid deployment while maintaining consistency across environments.

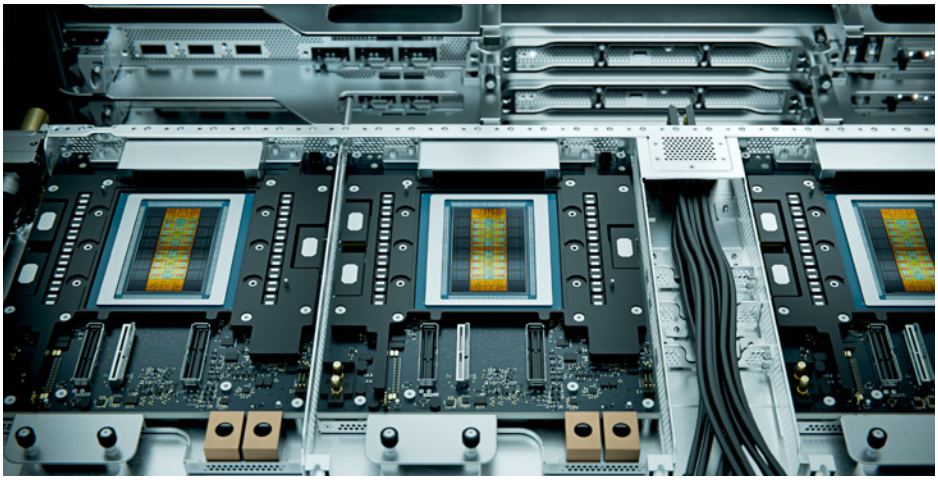
Drive performance optimization strategies with **AI Tensor Engine for ROCm (AITER)**, a robust AI operators repository of pre-optimized kernels purpose-built for AMD Instinct GPUs.

Accelerate AI performance with pre-optimized kernels for training and inference workloads.



Maximize GPU utilization with optimized GEMM and communication operations.

Experience high throughput and low latency across end-to-end AI workflows.



ALREADY ON CUDA? NO PROBLEM.

With **HIPIFY**, ROCm simplifies migration by enabling developers to convert CUDA code to HIP, preserving existing workflows while unlocking portability across AMD GPU environments.

A COMPLETE OPEN ECOSYSTEM, READY TO GO

AMD Instinct™ GPUs, powered by AMD ROCm™ software, are supported by a broad and growing open ecosystem spanning leading AI frameworks, inference engines, cloud platforms, and enterprise software.

Additionally, ROCm 7 supports all leading AI frameworks—including PyTorch, TensorFlow, JAX, and ONNX Runtime—and a broad range of tools such as DeepSpeed and CuPy, with comprehensive model compatibility through the Hugging Face Hub.

GO FROM ZERO TO RUNNING, FASTER THAN YOU THINK

Getting started with AMD Instinct™ GPUs and the AMD ROCm™ software stack is fast and straightforward. With streamlined setup, pre-built containers, and Day-0 framework support, teams can quickly move from installation to **production-ready AI workloads on an open, scalable platform.**

1. **Get the hardware**
   AMD Instinct GPUs

2. **Install ROCm**
   Download the ROCm runfile installer

3. **Run your workloads**
   Accelerate with AITER, fine tune with HIPify

2M+
pre-trained models
via Hugging Face

DAY-0 SUPPORT
for Llama, Gemma 3

Open platform.
No licensing fees.
Enterprise-ready.

Learn more at
amd.com/instinct

AMD ROCm™ blogs

AMD Developer Cloud

AMD INSTINCT

AMD ROCm

1. Based on calculations by AMD Performance Labs as of April 17, 2025, on the published memory specifications of the AMD Instinct MI350X / MI355X GPU (288GB) vs MI300X (192GB) vs MI325X (256GB). Calculations performed with FP16 precision datatype at (2) bytes per parameter, to determine the minimum number of GPUs (based on memory size) required to run the following LLMs: OPT (130B parameters), GPT-3 (175B parameters), BLOOM (176B parameters), Gopher (280B parameters), PaLM 1 (340B parameters), Generic LM (420B, 500B, 520B, 1.047T parameters), Megatron-LM (530B parameters), Llama (405B parameters) and Samba (1T parameters). Results based on GPU memory size versus memory required by the model at defined parameters plus 10% overhead. Server manufacturers may vary configurations, yielding different results. Results may vary based on GPU memory configuration, LLM size, and potential variance in GPU memory access or the server operating environment. MI350-012. *All data based on FP16 datatype. For FP8 = X2. For FP4 = X4.

2. Based on AMD measurements as of 6/6/2025 of the text-generated offline inference throughput for Llama 3.1-405B chat model on an 8x GPU AMD Instinct MI355X platform running 8x TP1 (8 copies of model on 1 GPU) with (FP4) compared to an 8x GPU AMD Instinct MI300X platform running 2x TP4 (2 copies of model on 4 GPUs) with (FP8). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input/output tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on the use of the latest drivers and optimizations. (MI350-042)

3. Based on AMD internal testing as of 6/4/2025, using an (8) GPU AMD Instinct™ MI355X Platform for overall GPU-normalized training throughput (processed tokens per second) for text generation using the Llama3-70B and Llama3-8B chat models running Torchtitan or Megatron-LM (FP8 and BF16) as applicable, using a maximum sequence length of 8192 tokens, compared to an (8) GPU AMD Instinct™ MI300X Platform using Megatron-LM (FP8 and BF16). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI350-034