



AMD ENTERPRISE-READY *AI PLATFORMS*

Leadership performance
in hardware and software

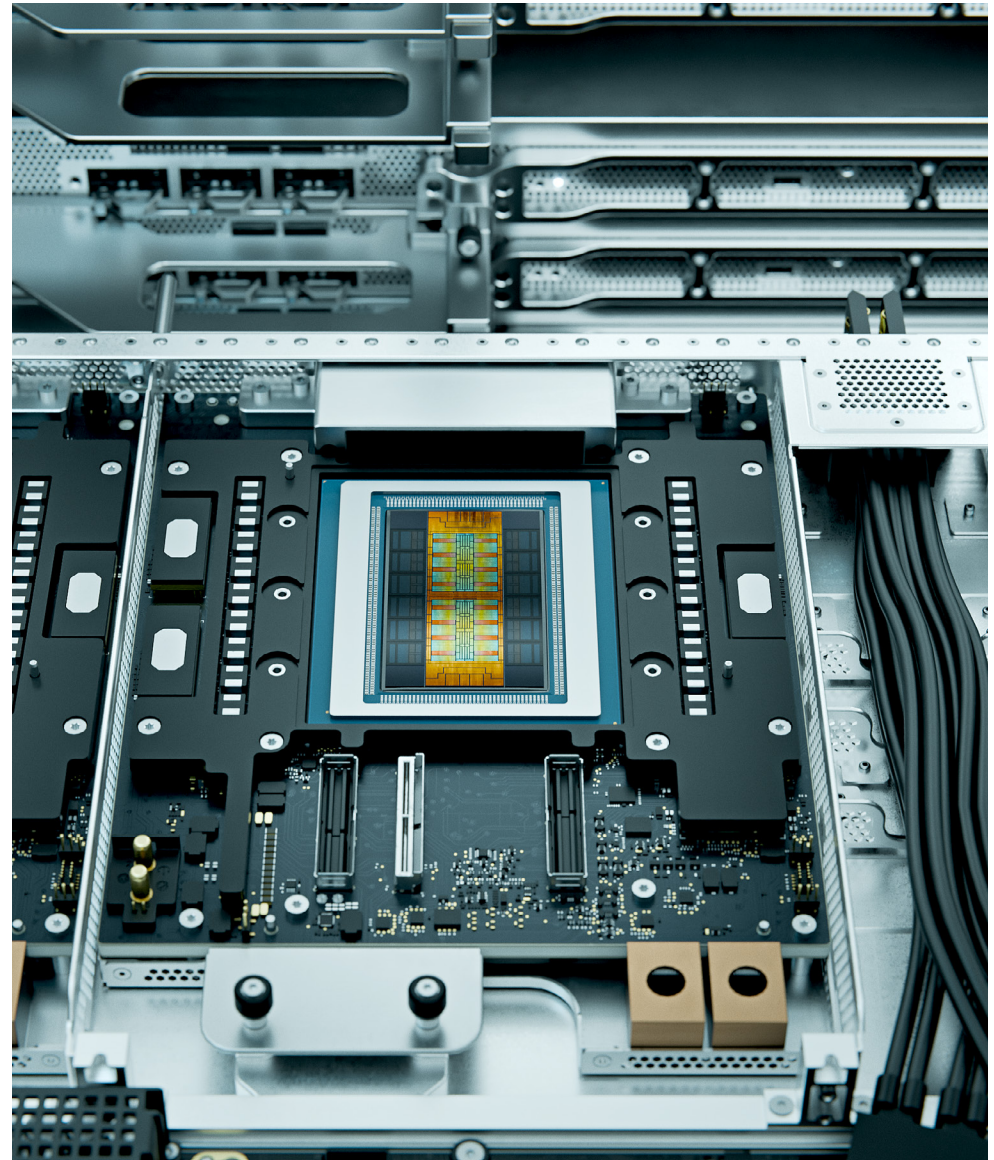
AMD 
together we advance_

THE COMPLETE STACK. FULL CHOICE. NO COMPROMISE.

AI adoption is accelerating across every industry. But most enterprises hit the same wall: the infrastructure is expensive, the tools are fragmented and the path from a promising model to a running production service is longer and harder than it should be.

Fortunately, AMD is changing that equation with AMD Instinct™ GPUs, AMD ROCm™ open-source ecosystem and the AMD enterprise AI reference stack.

AMD delivers everything needed to **build, deploy and scale AI** from beginning to end, in an open ecosystem, and with no licensing fees.



A COMPLETE ECOSYSTEM. ANY STACK, ANY WAY.

Most enterprises don't lack AI infrastructure options; they're overwhelmed by them. And today, most of that market gravitates toward closed, vertically integrated stacks that bundle hardware, software, and orchestration from a single vendor. It's a familiar path, but it comes at a cost: in the form of limited flexibility and a roadmap that's difficult to control.

AMD is built on a different premise: customers should have real choice, and choice shouldn't mean complexity.

The AMD open ecosystem—spanning cloud providers, neoclouds, OEM/ODM partners, and a broad ISV community—is designed so that when these pieces are brought together, customers get a full, enterprise-ready AI solution as easy to deploy and manage as any closed alternative.



Organizations don't have to adopt everything at once. AMD ROCm and AMD enterprise AI reference stack components are modular by design, enabling organizations to **integrate what they need, when they need it, and build toward a complete platform at their own pace**. With AMD Inference Microservices (AIMs) and example-ready blueprint designs, for example, organizations can speed adoption or use as a known-good reference starting point.

AMD INSTINCT GPUs: LEADERSHIP AI AND HPC ACCELERATION

AMD Instinct™ MI350 Series GPUs, are engineered for generative AI, training and inferencing, and HPC on 4th Gen AMD CDNA™ architecture. They deliver the memory, bandwidth, and precision support that all your AI workloads demand.

YOUR FOUNDATION MATTERS

AMD INSTINCT™ MI350X/MI355X OAM

Handle the largest models without multi-node complexity.

- 288 GB HBM3E per GPU; 2.3 TB across an 8-GPU platform
- Supports MXFP4, MXFP6, and OCP-FP8 for higher throughput per watt
- Serves more users at lower latency while reducing total cost of ownership

AMD INSTINCT™ MI350P PCIe® CARD

Deploy and scale generative and agentic AI within your existing infrastructure.

- CDNA™ 4 architecture on advanced 3nm process
- 144GB HBM3E memory at 4.0 TB/s peak bandwidth
- Up to 4.6 PFLOPs peak performance with OCP MXFPformat support
- GPU partitioning into up to 4 independent instances
- Air-cooled PCIe form factor for drop-in server integration
- Open, enterprise-ready ROCm AI software stack

THERMAL FLEXIBILITY

Available in air-cooled and direct liquid-cooled configurations, the MI350 Series supports standard and high-density data center designs. This dual-configuration approach gives customers the flexibility to optimize for their existing infrastructure investments or design purpose-built, next-generation facilities from the ground up.

TRUSTED AT THE LARGEST SCALE

AMD Instinct GPUs are trusted by some of the world's most advanced AI organizations, including Microsoft, Meta, Oracle, TensorWave, Digital Ocean, Vultr, Crusoe, and IBM, to power their most demanding AI workloads at scale. That same hardware is available for enterprise deployment today.

CDNA 4
architecture

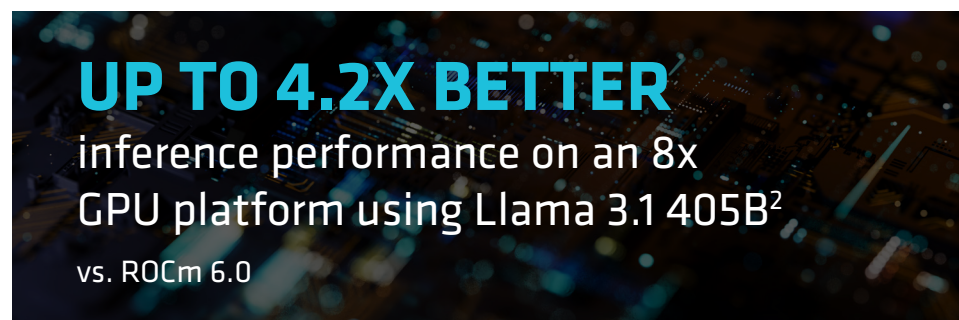
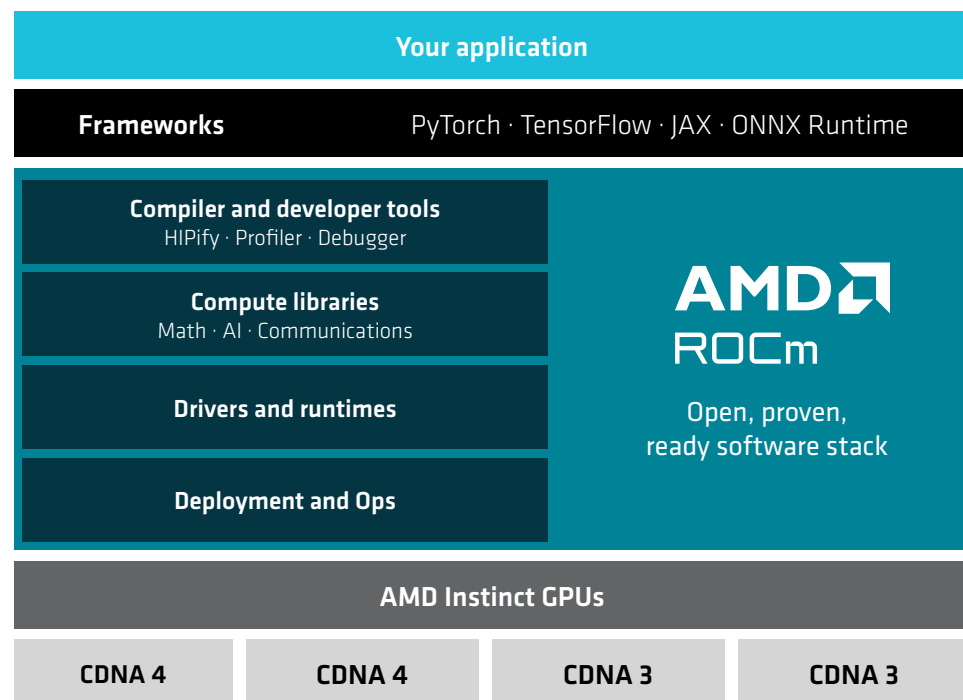
DAY 0
support for leading
AI frameworks
and models

**MXFP4 /
MXFP6 /
OCP-FP8**
low-precision support

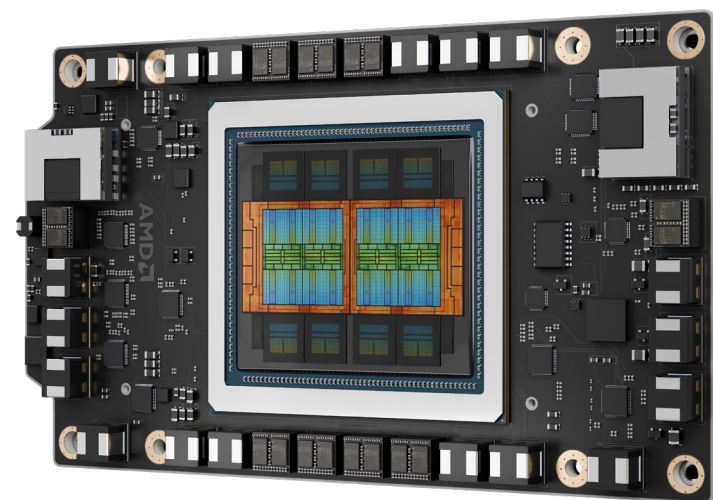
The MI355X platform offers **up to 4X generational uplift** in training and inference throughput vs. the MI300X platform¹

AMD ROCm™ 7: THE OPEN SOFTWARE ENGINE

AMD ROCm is an open-source software ecosystem. No licensing fees. No proprietary black boxes. AMD ROCm 7 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. It provides a complete stack from low-level GPU drivers all the way up to the AI frameworks teams already rely on, including fully upstreamed support for popular frameworks such as PyTorch and JAX.



Because ROCm is built upstream-first and developed in the open, innovations flow back into the community, and teams can pull key components directly from public GitHub repositories.



AMD ENTERPRISE AI REFERENCE STACK: FROM GPUS TO PRODUCTION SERVICES

Having powerful GPUs and an open software stack is necessary, but it is not sufficient. Enterprise AI teams still face the operational challenge of turning raw GPU infrastructure into reliable, governed, production-grade AI services. AMD recognizes that many organizations do not have an abundance of AI experts. That's why we created the AMD enterprise AI reference stack—to ease the implementation of AI solutions and agents for the broad enterprise market.



The AMD enterprise AI reference stack is a **unified, open, Kubernetes-native platform** that manages the entire AI lifecycle—from model development and fine-tuning through deployment and production serving—on AMD Instinct infrastructure.

Enterprise Applications

Sales Assistant

Human Resources

Marketing Assistant

Legal Assistant

Personal Assistant

Coding Assistant

Drug Discovery

Mechanical Design

Material Science

Open Ecosystem Partnership



PyTorch



Triton



DeepSpeed



SGLang



LLM-d



TensorFlow



JAX



ONNX



Jupyter



OpenXLA



vLLM



DeepSeek



ComfyUI



Hugging Face

AMD enterprise AI reference stack

AMD Solutions Blueprints

AMD Inference Microservices

AMD AI Workbench

AMD Resource Manager

AMD ROCm

High-Performance AMD Hardware

AMD Instinct GPUs

AMD EPYC™ CPUs

AMD Pensando™ DPUs

AMD Pensando™ NICs

Storage

AMD SOLUTION BLUEPRINTS: START WITH THE ANSWER, NOT THE QUESTION

Finding the right AI talent is hard. Knowing how to assemble all the pieces—models, infrastructure, software, orchestration—is harder. That’s the real barrier to enterprise AI adoption, and it’s where most teams get stuck.

Solution Blueprints remove that barrier. They are enterprise-ready recipes that combine validated AI models with proven reference architectures, giving teams a clear, tested starting point for enterprise-grade AI without needing to figure it out from scratch.

Available blueprints:



CODING ASSISTANT

for developers and engineering teams



FINANCIAL STOCK INTELLIGENCE

for financial analysts and investment teams



LLM CHAT SANDBOX

for AI teams and business stakeholders



DOCUMENT SUMMARIZATION

covering a range of document types



AGENTIC MACHINE TRANSLATION

for Enterprise applications and global teams



AUTOGEN STUDIO

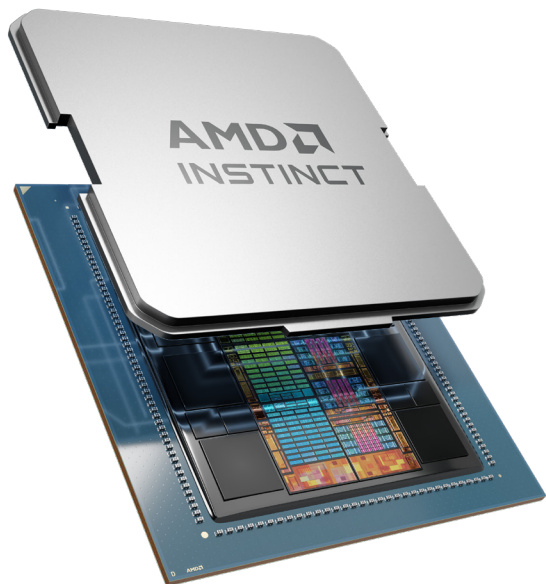
for creating, configuring, and managing multiagent AI conversations

Blueprints give enterprise teams a **validated starting point** with architectures that can be customized to the enterprises’ needs.

For the full list of blueprints, [visit Solution Blueprints Overview—AMD Enterprise AI](#).

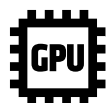
AMD will continue to release new blueprints across industries and use cases.

Solution Blueprints are not the ceiling; they are the floor. Each blueprint is a starting point that teams can extend, customize, and evolve for their specific use cases.



AMD INFERENCE MICROSERVICES (AIMs): PRODUCTION-READY MODEL SERVING

Getting a model to run in a notebook is straightforward. Getting it to serve thousands of requests per day—reliably, securely, with observability and compliance—is an entirely different problem. AIMs is the production serving layer that bridges that gap.



AIMs is a platform that makes it easy to run AI models on AMD Instinct GPUs. It handles all the technical setup by providing **prebuilt containers that bundle model, engine, and tuned configurations** for AMD hardware on industry-standard APIs.

Capabilities include:



INFERENCE MICROSERVICES

Enterprises can run popular open-source models with one click in their own infrastructure.



OPTIMIZED INFERENCE PROFILES

AIM containers automatically detect available compute resources and apply optimized inference parameters.



INFERENCE ORCHESTRATION

AIMs can automatically scale to demand and offer OpenAI-compatible APIs with security and built-in telemetry.

AMD AI WORKBENCH: THE COMPLETE AI DEVELOPMENT TOOLCHAIN

AI developers need a place to work—a consistent environment where they can discover models, run experiments, fine-tune for their specific use case and deploy to production without waiting on infrastructure teams for every step. AMD AI Workbench is that environment.

Core capabilities:

1

INFERENCE MANAGEMENT

Open model catalog for AIMS, that includes the most common open-source models with built-in orchestration for scalable inference.

2

MODEL FINE-TUNING

Recipes and pipelines for model fine-tuning and optimized inference support.

3

BUILT-IN SECURITY AND ANALYTICS

Access control and resource allocation limits configuration for each deployment, with detailed inference metrics per AMD enterprise AI reference stack deployment.

AMD AI Workbench is a unified developer interface that **simplifies AIMS deployments, accelerates model building, and streamlines AI workflow** orchestrations so teams can innovate faster and operate more efficiently.

AMD RESOURCE MANAGER: INTELLIGENT GPU COMPUTE MANAGEMENT

In a shared AI infrastructure environment, GPU compute is the most valuable and most contested resource. Without intelligent allocation, some teams sit idle while others are bottlenecked—and no one has clear visibility into utilization across the fleet. AMD Resource Manager solves this.

AMD Resource Manager is a central tool that **helps administrators manage who can use GPU resources**, how they're organized and what rules govern their use—all from a single location.



POLICY-BASED

Intelligent resource
scheduling



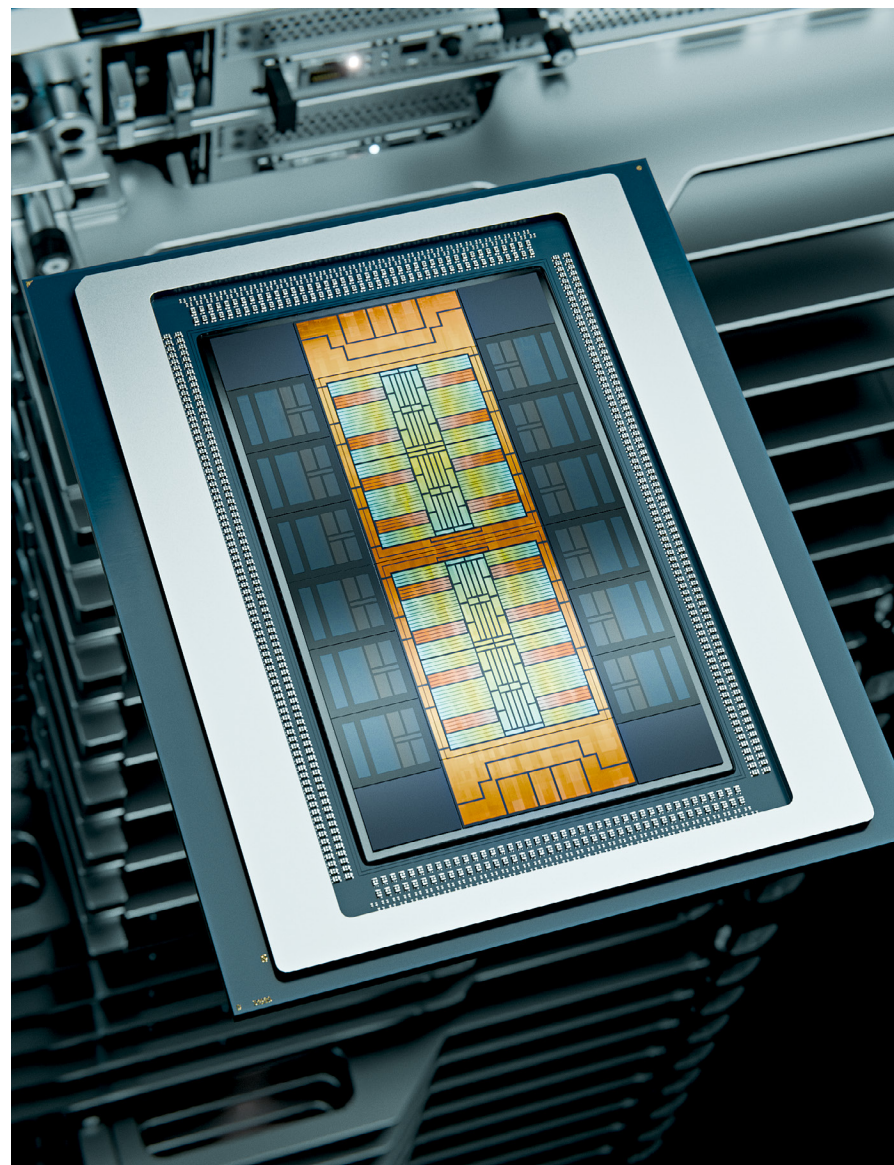
DYNAMIC

Quota scaling
across teams



REAL-TIME

Telemetry
and analytics



UNLEASH INTELLIGENT BUSINESS SOLUTIONS WITH AMD



Get started with serious AI development on AMD compute.



FASTER TIME TO PRODUCTION

Accelerate revenue-generating AI by compressing the path from development to deployment, so your teams ship faster and compete harder.



OPEN-SOURCE MODULAR ARCHITECTURE

Avoid vendor lock-in and future-proof your AI investments with a fully open platform that gives your organization the flexibility to adapt as the technology evolves.



OPTIMIZED COMPUTE UTILIZATION

Maximize return on infrastructure spend with intelligent resource allocation that ensures every GPU dollar works harder across your entire organization.

THE ENTERPRISE AI PLATFORM BUILT FOR WHAT'S NEXT

AMD delivers a complete, open, and production-ready AI platform—from the silicon up. Whether organizations are scaling LLM inference, fine-tuning models for industry or building intelligent applications from scratch, the AMD Reference Architecture provides teams with everything they need to move fast without compromise.

No Lock-In. No Licensing Fees. No Dead Ends.

AMD
INSTINCT

AMD
ROCm

¹ Based on calculations by AMD Performance Labs in May 2025, to determine the peak theoretical precision performance of eight (8) AMD Instinct™ MI355X and MI350X GPUs (Platform) and eight (8) AMD Instinct MI325X, MI300X, MI250X and MI100 GPUs (Platform) using the FP16, OCP-FP8, MXFP6 and MXFP4 datatypes with Matrix. Server manufacturers may vary configurations, yielding different results. Results may vary based on use of the latest drivers and optimizations. (MI350-004)

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) And testing by AMD as of May 15, 2025, measuring the inference performance in tokens per second (TPS) of AMD ROCm 6.x software, vLLM 0.3.3 vs. AMD ROCm 7.0 preview version SW, vLLM 0.8.5 on a system with (8) AMD Instinct MI300X GPUs running Llama 3.1-70B (TP2), Qwen2 72B (TP2), and Deepseek-R1 (FP16) models with batch sizes of 1-256 and sequence lengths of 128-204. Stated performance uplift is expressed as the average TPS over the (3) LLMs tested. Results may vary. (MI300-080)

DISCLAIMER: The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

© 2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, CDNA, EPYC, Instinct, Infinity Fabric, ROCm, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

READY TO TEST DRIVE?

Try out AMD Reference Architecture with [Dev Cloud](#).

Visit [AMD enterprise AI reference stack](#) to learn more.

Additional Resources:

- [Whitepaper](#)
- [Video](#)

AMD
together we advance_