



together we advance

THE AMD VISION FOR FULL-STACK AI INFRASTRUCTURE

AMD is shaping the future of data centers with full-stack AI infrastructure built for scale, efficiency, and adaptability.

With AI's undeniable momentum, AMD gives enterprises the flexibility and performance to manage increasingly complex AI workloads through a **three-pillar** approach:

<p>1</p> <p>LEADERSHIP COMPUTE ENGINES</p> <p>Designed to power high-performance computing workloads</p>	<p>2</p> <p>AN OPEN ECOSYSTEM</p> <p>Enables multi-vendor interoperability and freedom of choice</p>	<p>3</p> <p>FULL-STACK SOLUTIONS</p> <p>Maximize efficiency from silicon to systems</p>
--	--	---

\$500B+ AI market by 2028, growing at **60% CAGR¹**

Inference will make up two-thirds of AI workloads, growing at **80% CAGR²**

AI-READY LEADERSHIP COMPUTE ENGINES

Modern enterprises need advanced compute engines (GPUs, CPUs, DPUs, and adaptive SoCs) within an open hardware and software ecosystem. AMD is the only provider that delivers all four.

This approach gives developers, service providers, and enterprises the performance and interoperability needed to scale AI across domains and deployment models.

<p>35x</p> <p>increase in inference performance from AMD Instinct™ MI300 to MI350 GPU³</p>	<p>Up to 40%</p> <p>more tokens per dollar for LLM inference with Instinct MI355X GPU vs. NVIDIA's B200⁴</p>	<p>Up to 520B</p> <p>parameter model support with FP4/FP6 precision⁵</p>
--	--	--

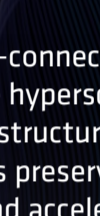
Made possible with AMD

Providers using AMD Instinct MI355X GPUs report meaningful cost and latency improvements thanks to higher tokens-per-dollar efficiency and larger on-GPU memory helping them run larger models with fewer GPUs and faster response times.

AI-first hardware	
AMD Instinct™ MI350 Series GPUs	Up to 4x Gen-on-Gen AI compute, up to 35x inference performance for generative AI and LLM workloads ⁶
AMD Instinct MI355X GPUs	Delivers 40% more tokens per dollar vs. NVIDIA's B200, optimized FP4/FP6 precision for inference ⁷
AMD Instinct MI400 GPUs	Projected 10x inference over MI355X GPUs, tuned for agentic AI and Mixture of Experts models ⁸
6th Gen AMD EPYC™ CPUs	High memory bandwidth, ideal for data preprocessing and orchestration in AI pipelines
AMD Pensando™ 800G AI NIC	800G network interface card, ultra-low latency for distributed AI workloads

EMBRACING AN OPEN ECOSYSTEM

Using tools such as open hardware standards (via OCP), AMD ROCm™ software, and UALink/UEC interconnects, AMD accelerates AI through openness and a customer-first approach to collaboration.

<p>IDC projects that by 2026</p> <p>80% +</p> <p>of enterprises will prioritize infrastructure interoperability to support multi-vendor AI strategies²</p>	 <p>UALink-connected AMD GPUs will enable hyperscalers to integrate into existing infrastructure without replacing everything. This preserves investments, reduces costs, and accelerates deployment</p>
--	---

Made possible with AMD

In addition to Llama Inference, Meta's AI Recommendation Inference & Training models run on AMD Instinct MI300X GPUs.

Open Ecosystem Enablers	
AMD ROCm™ 7 Software Stack	Open software stack with drivers, tools, and APIs that support GPU programming. Optimized for Generative AI and HPC workloads, making it easy to migrate existing x86 code
AMD Developer Cloud	A free-to-use AMD ROCm™ software-based environment to help lower adoption barriers for startups, researchers, and open-source communities
Networking support backing the UALink Consortium	Vendor-neutral GPU interconnect, offers an alternative to NVIDIA's proprietary, NVLink for high-speed AI interconnects, enabling greater deployment flexibility
Adoption Across Oracle Cloud Zettascale AI clusters	OCL is scaling the full-stack portfolio of AMD GPUs and CPUs across zettascale AI clusters including +131,000+ AMD GPUs in open, multi-vendor cloud architecture

“AMD is positioning itself as the "open alternative" in a landscape dominated by vertically integrated, proprietary AI stacks”²
– IDC

FULL-STACK SOLUTIONS

Through a vertically integrated strategy and key acquisitions like ZT Systems, AMD solves for today's enterprise needs with optimized, integrated platforms from silicon to systems.

<p>Silicon</p> 	<p>Software</p> 	<p>Systems</p>  <p>AMD AI Rack</p>
---	---	---

Made possible with AMD

Financial services providers can look forward to benefits such as strengthening fraud detection, harnessing larger LLMs, improving accuracy, and advancing AI-driven security with the planned AMD AI rack design, featuring AMD Instinct MI400 GPUs, AMD EPYC CPUs, and the AMD ROCm™ software stack.

A full-stack approach	
AMD AI Rack	Designed for large-scale AI training and inference deployments, it merges CPUs, GPUs, networking, and software into one unified system, optimized for agentic AI and MoE workloads
ZT Systems Integration	Expands system integration capabilities that unify AMD CPUs, GPUs, networking, and software (like AMD ROCm stack) at the rack level
AMD ROCm™ software + Developer Cloud	Enables accelerated innovation, seamless scaling from development to production, and the flexibility of open standards, all in one integrated ecosystem
Adaptive Computing	AMD Xilinx products add adaptive computing to the stack, complementing CPUs and GPUs for heterogeneous AI

“AMD is no longer just catching up—it is forging a differentiated path with openness, developer-first engagement, and full-stack integration.”²
– IDC

BUILDING AN AI-READY DATA CENTER

Over the next decade, leadership will belong to data centers that are AI-first, open, and full-stack. They will have to power advanced workloads with efficiency and flexibility. The AMD portfolio is built for this future, uniting high-performance compute, open standards, and full-stack solutions to meet tomorrow's demands today.

From large-scale AI clusters to high-performance enterprise environments, AMD helps data centers:

 <p>Scale AI workloads with high performance and efficiency</p>	 <p>Experience the openness and cost savings that come with freedom of choice</p>	 <p>Accelerate innovation with ready-to-run integrated systems</p>	 <p>Support sustainability goals while optimizing system capabilities</p>
--	--	---	--

Download the full report to read IDC's takeaways from Advancing AI 2025.