IDC Link
Real-Time IDC Research® Opinion on Industry News, Trends, and Events

Advancing AI 2025: AMD's Full-Stack Infrastructure Vision

June 16, 2025

By: Matthew Eastwood, Jim Hines, Brandon Hoff, Mario Morales, Kuba Stolarski

This IDC Link also includes contributions from Nina Turner, Mohamed Hefny, and Andy Buss.

IDC'S QUICK TAKE

At its 2025 Advancing AI event, AMD unveiled a comprehensive vision to lead in open, scalable AI infrastructure. The company introduced the Instinct MI350 Series accelerators, previewed its next-gen Helios rack, reviewed the road map for its next-generation MI400 and MI450 GPUs as well as its EPYC "Venice" CPUs for the datacenter, and launched ROCm 7 alongside a new developer cloud. AMD emphasized its commitment to openness through its participation in and adoption of the Universal Accelerator Link (UALink) specification, Ultra Ethernet Consortium (UEC), open software stacks, and Open Compute Project (OCP) reference designs, positioning itself as the leading open alternative to proprietary AI platforms. With broad ecosystem support — from Meta to Oracle — AMD is building a full-stack, energy-efficient AI infrastructure platform designed for training, inference, and agentic execution at a global scale. On-premises infrastructure was also an area of emphasis as AMD begins to push into enterprises to address the growth for AI inference with a full stack of solutions with partners.

EVENT HIGHLIGHTS

At its 2025 Advancing AI event, AMD laid out a comprehensive vision to become a foundational leader in global AI infrastructure. As AI workloads diversify across training, inference, and agentic execution — from hyperscale datacenters to the edge — AMD is positioning itself as the only vendor offering leadership compute engines (GPUs, CPUs, DPUs, and adaptive SOCs) alongside a fully open hardware and software ecosystem. This approach is intended to give developers, service providers, and enterprises the flexibility, performance, and interoperability needed to scale AI across domains and deployment models.

By doubling down on openness through ROCm, UALink, UEC, and standard rack architectures — and committing to an annual cadence of silicon and systems innovation — AMD is building an alternative AI stack focused on performance, energy efficiency, and choice. The message is clear: The next era of AI infrastructure will be open, collaborative, and globally distributed — and AMD is working hard to lead it.

AMD's strategy is anchored in three pillars:

- **Leadership compute engines:** Driving performance and efficiency with continuous GPU, CPU, and networking innovation, led by the Instinct MI300, MI350 Series, and Pollara AI NICs
- **Open ecosystem:** Embracing open hardware (via OCP), software (via ROCm), and interconnects (UALink, UEC) to ensure flexibility and innovation across enterprises and hyperscalers
- **Full-stack solutions:** Integrating silicon, software, and systems — including through acquisitions like ZT Systems — culminating in the forthcoming Helios AI rack

Announcements

At its Advancing AI 2025 event, AMD unveiled a comprehensive set of announcements across silicon, software, and systems — reinforcing its strategy to lead in open, scalable AI infrastructure. With the launch of the AMD Instinct MI350 Series, a preview of its next-gen Helios rack, the debut of ROCm 7, and expanded global partnerships, AMD positioned itself as a credible and differentiated force in AI compute. The announcements reaffirm AMD's commitment to an open ecosystem that prioritizes performance, developer accessibility, and energy efficiency at scale:

- **AMD Instinct MI350 Series launch:** The new MI350X and MI355X GPUs offer up to four times more AI compute and a 35 times jump in inference performance. The MI355X reportedly delivers up to 40% more tokens per dollar than NVIDIA's B200, supporting models up to 520 billion parameters with FP4 and FP6 precision. Full availability ramps in the second half of 2025.
- **Preview of MI400 and Helios AI rack:** AMD showcased its upcoming Helios AI rack, powered by next-gen MI400 GPUs, Zen 6–based EPYC "Venice" CPUs, HBM4 memory, and the Pensando "Vulcano" 800G AI NIC. Expected to deliver 10x inference performance over the MI355X, it is optimized for large-scale agentic and Mixture of Experts workloads, with more details coming in 2026.
- **ROCm 7 software stack debuts:** ROCm 7 brings up to 3.5x performance gains; adds support for frontier models, distributed inference, and enterprise features; and expands to Windows and consumer GPUs — showcasing AMD's fast-paced, open source AI development.
- **Developer enablement via AMD Developer Cloud:** AMD launched its Developer Cloud with a credit program offering complimentary access to ROCm-based environments and pay-as-you-go pricing — aimed at lowering adoption barriers for start-ups, researchers, and open source communities.
- **Open deployments with Oracle Cloud:** Oracle Cloud Infrastructure will scale AMD's full open AI stack — including MI355X GPUs and EPYC CPUs — across

zettascale AI clusters with plans for over 131,000 GPUs, enabling massive-scale training and deployment.

- **Support for UALink and open standards:** AMD strengthened its push for openness by backing the UALink Consortium — offering an alternative to NVIDIA's NVLink for high-speed AI interconnects and enabling greater deployment flexibility.
- **Ecosystem momentum:** Meta, OpenAI, Microsoft, Cohere, and others confirmed AMD Instinct GPUs are in use for production workloads, validating the platform's credibility for both training and inference at scale.
- **HUMAIN partnership for AI infrastructure:** AMD and HUMAIN announced a $10 billion initiative to build resilient, open AI infrastructure across the United States and the Middle East. Over five years, it will deploy 500MW of AI compute, fully powered by AMD.
- **Sustainability milestones:** AMD exceeded its 30x25 energy goal and set a new 20x target by 2030 — aiming to reduce the power required to train today's largest models by 95%, down to a single rack.
- **Annual platform update cadence:** AMD reaffirmed its annual platform update strategy with a brief preview of 2027 plans, including EPYC Verano, Instinct MI500 Series, and Vulcano integrated into a next-gen AI rack.
- **Strategic acquisitions:** In recent months, AMD acquired teams from Untether AI, Brium (AI software/tooling), Enosemi (chip IP for optics), and Lamini (AI talent), totaling over 25 acquisitions across the AI stack in the past year.
- **OpenAI collaboration signals shift:** AMD shared the stage with OpenAI's Sam Altman to spotlight early design collaboration on the MI450, indicating a shift as OpenAI — once reliant on NVIDIA — expands its partners. The MI400 series will offer higher memory density and bandwidth than NVIDIA's Vera Rubin, cementing AMD's growing role with OpenAI and Microsoft Azure.

IDC'S POINT OF VIEW

At its Advancing AI 2025 event, AMD highlighted clear progress toward becoming a foundational player in the AI infrastructure ecosystem. Spanning compute, connectivity, software, and systems, AMD's announcements reflected a cohesive strategy aligned with IDC's view of enterprise AI: open, composable, and scalable hybrid infrastructure that supports diverse use cases — from training and inference to agentic execution.

AMD reaffirmed its projection of a $500+ billion AI market by 2028, growing at 60% CAGR, with inference accounting for two-thirds and growing at over 80% CAGR.

The Instinct MI350 Series launch and MI400-based Helios rack preview mark major steps in addressing performance, efficiency, and scale demands — especially for hyperscalers. AMD's emphasis on FP4/FP6 precision and next-gen memory (HBM3E, HBM4) signals a clear focus on advanced inference and Mixture of Experts models moving into production.

The ZT Systems acquisition appears to be bearing fruit, with Helios demonstrating AMD's expanded systems integration capabilities. Combined with Pensando interconnects and AMD's EPYC and Instinct platforms, the company now offers a robust, end-to-end hardware portfolio. While Helios will follow open hardware principles, AMD has not clarified whether it will be a branded system (like NVIDIA DGX), an OCP reference design, or both. Based on AMD's open approach, a hybrid model — branded systems for select partners and reference designs for OEMs — seems likely.

ROCm 7 continues to anchor AMD's software stack, with improvements in developer experience, distributed inference, and framework compatibility — key for enterprise adoption. The launch of AMD Developer Cloud marks a shift from component vendor to full-stack platform provider, aligning with IDC's view of what is needed for long-term AI relevance. ROCm 7 also expands to Windows 11 and supports more devices, including Ryzen AI mobile CPUs and RDNA 4–based Radeon GPUs.

This enables powerful local AI capabilities at the edge — supporting up to four GPUs (eight forthcoming) on platforms like Ryzen Threadripper PRO 9000. Developers can now train and deploy large models locally, enhancing privacy, performance, and cost efficiency while reducing cloud dependency.

Crucially, AMD is positioning itself as the "open alternative" in a landscape dominated by vertically integrated, proprietary AI stacks. Its support for standards like UALink and UEC, alongside adoption by Meta, OpenAI, Microsoft, and Oracle, boosts its credibility with buyers that value vendor-neutral platforms offering flexibility and transparency. AMD is betting that open ecosystems backed by large developer communities will out-innovate closed systems over time. Its challenge: accelerate this ecosystem quickly enough to close the gap with NVIDIA.

AMD's focus on openness — ROCm, UALink, and open rack-scale systems — directly contrasts NVIDIA's CUDA/NVLink stack and Intel's oneAPI. If enterprise demand for flexibility continues to grow, this could pressure rivals to open their platforms further. As top-tier players validate AMD's stack in production, the perception of AI compute being "NVIDIA only" is beginning to erode.

Maintaining its annual cadence of GPU and system updates also pressures competitors to keep pace. ROCm 7's performance improvements and broader framework support are reducing migration friction, but challenges remain. Many enterprise and research workflows are deeply tied to CUDA, and not all frameworks or custom libraries perform equally well on ROCm yet. Despite growing momentum, NVIDIA still dominates in brand, community, and ecosystem inertia. AMD must continue investing in developer engagement, customer success stories, and high-visibility deployments to build confidence among risk-averse buyers.

From IDC's viewpoint, AMD is no longer just catching up — it is forging a differentiated path with openness, developer-first engagement, and full-stack integration. As AI adoption matures from experimentation to scaled deployment, AMD's platform approach offers strong economic and operational value. Challenges persist, but the accelerating pace of innovation may shorten buyers' switching timelines, creating an opportunity for AMD to gain ground.

Subscriptions Covered:

Canadian Datacenter Infrastructure and Cloud Services, Cloud Adoption Trends and Strategies, Cloud to Edge Datacenter Trends, Computing Systems, Platforms and Technologies, Digital Infrastructure Strategies, Enabling Technologies: Artificial Intelligence Edge Processor Architectures, Enabling Technologies: DRAM and NAND Memory, Enabling Technologies: Processors and Architectures, HPC and AI Infrastructure Stacks and Deployments, Storage and Computing Infrastructure Software Platforms, Storage Systems