FIVE REASONS AMD HELIOS IS THE FOUNDATION FOR HYPERSCALE AND SOVEREIGN AI

AMD Helios Rackscale Solution Powered by AMD Instinct™ MI455X GPUs



AT A GLANCE

Generative and agentic AI place immense pressure on data centers to scale computing power and efficiency simultaneously. Building AI infrastructure at this scale requires more than raw performance; it demands an open, modular rack design that evolves across product generations, combining leadership compute engines with high-speed networking to unify GPUs and servers into a single, coherent system.

AMD is accelerating AI infrastructure development with unmatched momentum, adopting open standards and deepening its presence across a broad network of hardware and software collaborators. AMD Helios, powered by AMD Instinct™ MI455X GPUs, is the result of a complete, open, rack-scale AI infrastructure platform purpose-built for hyperscale and sovereign AI deployments.

AMD HELIOS POWERED BY AMD INSTINCT MI455X GPUS: A BLUEPRINT FOR OPEN, SCALABLE AI INFRASTRUCTURE

Deploy rack-scale AI with a consistent, open software foundation and an integrated rack-scale solution for large-scale inference, training, and fine-tuning. Turn AI data center designs into deployable production infrastructure at any scale.

1

ACHIEVE INDUSTRY-LEADING AI PERFORMANCE

AMD Helios transforms a rack into a unified AI compute engine.

The AMD Instinct MI455X GPU delivers a generational leap in low-precision AI performance. Combined with AMD EPYC™ "Venice" server CPUs, AMD Pensando™ Vulcano AI NICs, and the AMD ROCm™ software stack, AMD Helios sets a new benchmark for hyperscale AI, delivering compute density, memory capacity, and bandwidth to support everything from high-volume inference and frontier model training to enterprise fine-tuning and emerging agentic AI.

2

SCALE FROM RACK TO MULTI-DATA CENTER CLUSTERS

Scale up inside the rack. Scale out across racks and clusters without proprietary fabrics.

AMD Helios is designed for linear, predictable scaling at every level. Inside the rack, all 72 AMD Instinct MI455X GPUs are unified into a single high-speed pod via UALink over Ethernet (UALoE), enabling large-model inference and training within a single coherent domain. Across racks and into multi-data center clusters, standards-based Ethernet aligned with the Ultra Ethernet Consortium (UEC) enables high-performance scale-out without proprietary interconnects. Virtualization and partitioning capabilities allow workload allocations to be rightsized from a single GPU partition up to a full 72-GPU pod, yielding seamless, predictable growth from rack to cluster.

3

IMPROVE EFFICIENCY AND SERVICEABILITY

Turn dense AI infrastructure into workable, production-ready deployments.

AMD Helios is engineered to convert dense AI infrastructure into deployable production systems with superior operational efficiency. Flexible workload partitioning, from pod-level Virtual Pods down to sub-GPU compute and memory partitions, drives higher GPU utilization, supports shared and multi-tenant environments, and lowers cost per token at hyperscale. Rack-scale power and cooling optimization and modular serviceability reduce operational overhead, enabling efficient and scalable deployments across hyperscale and sovereign AI environments. AMD is committed to long-term efficiency leadership, which means the platform is built not just for today's workloads, but for the demands of the decade ahead.

4

EMBRACE OPEN STANDARDS FOR HYPERSCALE AND SOVEREIGN AI

Join the next generation of AI leaders on an open, standards-based foundation.

AMD Helios is built on open standards at every layer, from ORW-aligned rack design and UEC-compliant Ethernet networking to OCP MX-standard data formats, qualified multi-vendor NIC flexibility, and the open AMD ROCm software stack. This openness preserves architectural flexibility across product generations and eliminates dependence on proprietary fabrics. Adoption validates the approach: eight of the world's ten largest AI organizations use AMD Instinct GPUs, and leading hyperscalers and AI developers are scaling AMD Helios with multi-year commitments. Advanced security, including device-level attestation, memory encryption, universal link encryption, and secure multi-GPU scaling, protects data and model confidentiality and supports data sovereignty for both hyperscale and sovereign AI environments.

5

DEPLOY WITH A PRODUCTION-READY OPEN SOFTWARE STACK

Take advantage of the industry's premier open-source AI software ecosystem.

AMD ROCm software is the industry's premier open-source AI stack, validated in exascale-class environments and built with a developer-first approach to enablement. AMD ROCm powers the full AMD Helios software foundation, converting GPU innovations into reliable production results for inference, training, and fine-tuning. Day-0 model support and native integration with the largest open-source AI projects help ensure the latest models and frameworks run on AMD hardware from day one. AMD AI Academy and the AMD Developer Cloud empower an expanded developer community, broadening ecosystem partnerships and accelerating the pace of innovation across the AI industry.

TECHNICAL DEEP DIVE

The AMD Helios rackscale solution includes 72 AMD Instinct™ MI455X GPUs per rack alongside AMD EPYC™ "Venice" server CPUs, AMD Pensando™ Vulcano 800 AI NICs, AMD Pensando™ Salina 400 DPUs, and the ROCm™ software stack. Together, these building blocks form a single open, rack-scale system that drives industry-leading performance, seamless scale from rack to cluster, greater efficiency and serviceability, open standards at every layer, and a production-ready open software foundation for hyperscale and sovereign AI.

#1 ACHIEVE INDUSTRY-LEADING AI PERFORMANCE

- Delivers up to 2.9 exaFLOPS peak theoretical four-bit precision performance (OCP MXFP4) and up to 260 TB/s peak theoretical scale-up bandwidth per rack.
- Up to 31 TB HBM4 memory per rack, up to 1.67 PB/s memory bandwidth, and up to 43 TB/s aggregate scale-out bandwidth per rack.
- The AMD Instinct™ MI455X GPU delivers more than 4x the peak theoretical low-precision performance of the previous-generation MI355X GPU at four-bit, eight-bit, and 16-bit bfloat precision, with MXFP4 performance adhering to the OCP MX standard for cross-platform hardware-software interoperability.[MI400-006]
- Powered by the 5th Gen AMD CDNA™ architecture, the AMD Instinct™ MI455X GPU offers up to 2.91X or 191% higher peak memory bandwidth compared to AMD Instinct™ MI355X GPU.[MI400-008]
- Powered by the 5th Gen AMD CDNA architecture, the AMD Instinct MI455X GPU with peak Matrix FP16/BF16 offers up to 1.26X or 26% better peak theoretical precision performance compared to the NVIDIA Vera Rubin GPU with FP16/BF16 datatype.[MI400-003]
- Powered by the 5th Gen AMD CDNA™ architecture, an AMD Instinct™ MI455X GPU with peak Open Compute Project MXFP4, MXFP6, MXFP8 and FP8 datatype offers up to 1.15X or 15% better peak theoretical precision performance compared to the NVIDIA Vera Rubin GPU with NVFP4, FP6/FP8 datatypes.[MI400-003]
- Powered by the 5th Gen AMD CDNA architecture, the AMD Instinct MI455X GPU offers up to 1.5X more memory capacity, or 50% more memory compared to the NVIDIA Vera Rubin GPU.[MI400-004]

#2 SCALE FROM RACK TO MULTI-DATA CENTER CLUSTERS

- AMD Helios uses UALink over Ethernet (UALoE) for in-rack scale-up and UEC-aligned Ethernet for scale-out across racks and clusters.
- All 72 AMD Instinct MI455X GPUs form a unified in-rack domain with consistent bandwidth and rack-wide GPU memory access, sharing a single load/store domain and automatic failover for link resilience.
- The AMD Pensando™ Vulcano 800 AI NIC, fully UEC 1.0 compliant, connects via PCIe® to the host and delivers lossless, high-throughput Ethernet-based scale-out with programmable pipelines and RCCL proxy acceleration, contributing to up to 43 TB/s of aggregate scale-out bandwidth per rack.
- Virtualization and partitioning enable rightsizing from a single GPU partition up to a full 72-GPU pod, supporting seamless growth from rack to cluster with predictable scaling economics.

#3 IMPROVE EFFICIENCY AND SERVICEABILITY

- The AMD Instinct MI455X GPU optimizes efficiency and bandwidth for fast, power-efficient AI inference and training, with a maximum TBP of 2500W in liquid-cooled EAM form factor.
- AMD Helios supports shared and multi-tenant workloads via pod partitioning (Virtual Pods spanning 4 GPUs to full 72-GPU pods), GPU-level compute partitioning, and spatial memory partitioning with NUMA-mode isolation. Compute Nodes within a Virtual Pod do not need to be physically adjacent, providing flexible allocation across the rack. NUMA mode is configured at boot time and requires a reboot to reconfigure.
- Rack-scale power and cooling optimization reduces operational overhead and enables efficient, scalable deployments across hyperscale and sovereign AI environments.
- AMD has a 2030 energy efficiency goal: a 20x increase in rack-scale energy efficiency from a 2024 baseline, enabling a model that today requires more than 275 racks to train in fewer than one fully utilized rack, using 95% less electricity.

#4 EMBRACE OPEN STANDARDS FOR HYPERSCALE AND SOVEREIGN AI

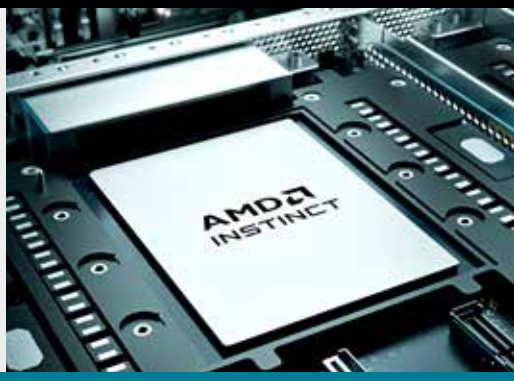
- ORW-aligned rack design and UEC-compliant Ethernet networking preserve openness at every layer, eliminating proprietary fabric dependencies across product generations.
- OCP MX-standard data formats help ensure hardware-software interoperability across modern AI accelerators.
- Security features include device-level DICE identity and attestation, runtime attestation, memory encryption, universal link encryption, and secure multi-GPU scaling, supporting data and model confidentiality for hyperscale and sovereign AI deployments.

#5 DEPLOY WITH A PRODUCTION-READY OPEN SOFTWARE STACK

- AMD ROCm software transforms Helios from powerful hardware into deployable, scalable AI infrastructure, supporting inference, training, and fine-tuning from rack to cluster while leveraging open frameworks and interoperable networking layers for flexibility and consistent performance.
- Day-0 model support and native integration span the largest open-source AI frameworks and runtimes, including PyTorch, JAX, vLLM, SGLang, ONNX Runtime and TensorFlow, plus day-0 enablement for leading open models such as DeepSeek.
- AMD ROCm software enables predictable distributed execution with RCCL communication primitives, KV-cache optimizations, and compute/communication overlap tuned for exaFLOP-class scale.
- AMD AI Academy and the AMD Developer Cloud empower an expanded developer community, broaden ecosystem partnerships, and accelerate the pace of innovation across the AI industry.

AMD
together we advance_AI

POWER HYPERSCALE AND SOVEREIGN AI WITH THE AMD HELIOS RACKSCALE SOLUTION POWERED BY AMD INSTINCT™ MI455X GPUs



LEARN MORE AT AMD.COM/INSTINCT AND AMD.COM/ROCM

©2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD arrow, AMD Instinct, Infinity Fabric, AMD Infinity Platform, 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. PCIE® design mark is a registered trademark and/or service mark of PCI-SIG. Ultra Accelerator Link" and "UALink" are trademarks of the Ultra Accelerator Link Consortium, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. PID # 5158664