AMD

AMD HELIOS
RACK-SCALE SOLUTION

THE OPEN, SCALABLE RACK-SCALE FOUNDATION FOR AI FACTORIES, HYPERSCALE AI, AND SOVEREIGN COMPUTING

AI Demand Is Accelerating Faster Than Today's Infrastructure Can Support.

As organizations scale training and deploy large-scale inference services, legacy systems are becoming a bottleneck—slowing deployment, increasing integration risk, increasing operational cost, and limiting scalability. At the same time, rapid growth in AI users, models, and data is placing unprecedented pressure on compute, power, and networking in hyperscale environments.

This inflection point is driving a shift toward open ecosystems as the foundation for scalable AI. AMD, working with Meta and the Open Compute Project (OCP) community, is meeting this moment with AMD Helios, an open, rackscale solution designed to accelerate deployment of next generation AI infrastructure. Built on OCP standards, the AMD Helios rackscale solution integrates GPUs, CPUs, networking, and open software into a single rack level platform that is scalable, serviceable, and future ready for hyperscale AI environments.

The Open Advantage: Faster Today, Flexible for What's Next

- Open hardware and software standards reduce integration complexity and accelerate deployment.
- AMD ROCm™ software and open standards approach to hardware enable multivendor interoperability and long-term flexibility.
- The AMD open software ecosystem supports leading AI frameworks like PyTorch, ONNX, and JAX, enabling workload portability and Day-0 support for latest AI models.
- Open standards-based networking spans both the scale-up GPU fabric and scale-out cluster interconnect.

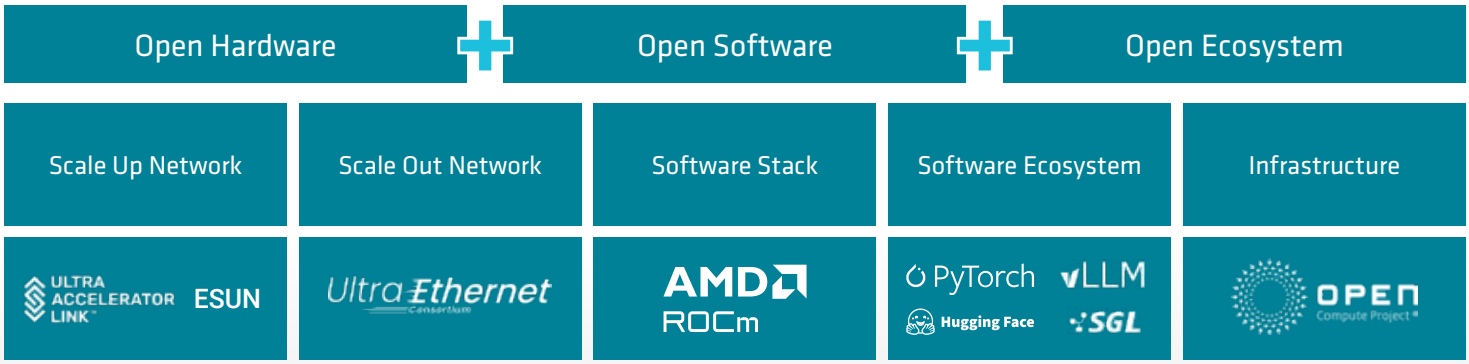
Breakthrough Performance for Next-Generation AI and HPC

- Single rack supports multi-trillion-parameter AI training, distributed inference, sovereign AI deployments, and large-scale HPC environments.
- Extreme compute density with 72 AMD Instinct™ MI455X GPUs delivering up to 2.9 exaFLOPS (OCP MXFP4) and 1.4 exaFLOPS (OCP MXFP8).
- AMD Instinct™ MI455X GPUs deliver up to 31 TB of HBM4 per rack and 19.6 TB/s of memory bandwidth per GPU for large-scale AI training and inference.

Rack-Scale Design Optimized for Efficiency and Operations

- Aligned with Meta's Open Rack Wide (ORW) approach and OCP rack standards for seamless hyperscale integration.
- Rack level integration of compute, GPUs, and networking for power constrained data centers.
- Liquid cooling and a rack ready footprint increase GPU density and performance.
- Unified GPUs, CPUs, networking, and open software streamline deployment and management.

Open Ecosystem Drives AI Innovation



System Architecture Highlights

AMD Instinct™ MI455X GPUs, built on the AMD CDNA™ architecture, combine massive HBM4 capacity, extreme memory bandwidth, and multi-petaflop compute performance to accelerate large-scale inference and AI training across the Helios rack-scale platform. MI455X GPUs enable efficient scaling for large models and high-throughput AI infrastructure deployments.

6th Gen AMD EPYC™ CPUs deliver high core counts, fast CPU-to-GPU connectivity, and high-bandwidth memory access to support AI orchestration, training, and inference workloads at rack scale. Optimized for modern AI infrastructure, EPYC processors help orchestrate networking, storage, and workloads at rack scale.

AMD Pensando™ "Vulcano" 800 AI NIC delivers leadership performance with 800 Gbps Ethernet speeds and is the only AI NIC to offer up to 2.4 Tbps of scale-out bandwidth per GPU, with full hardware and software programmability to keep GPUs and CPUs continuously fed at rack scale.

Scale-Up Fabric built on an open networking approach, leverages UALink over Ethernet (UALoE) to deliver high bandwidth, low-latency GPU connectivity at rack scale. Powered by advanced Ethernet silicon, it enables seamless interconnect up to 72 GPUs per Helios rack, supporting efficient scaling and reduced vendor lock-in.

10X performance increase[1]

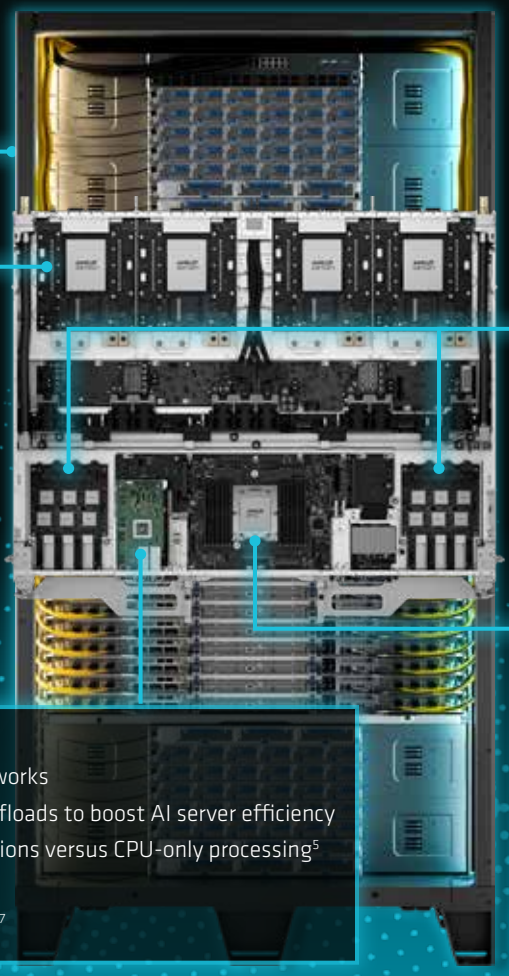
(vs MI355X)

AMD Helios AI Rack specs

- GPU domain: 72
- Scale up bandwidth: 260 TB/s
- FP4 – FP8 FLOPS: 2.9 EF – 1.4EF[2]
- HBM4 memory capacity: up to 31 TB[2]
- Memory bandwidth: up to 19.6 TB/s per GPU[2]
- Scale out bandwidth: up to 43 TB/s[2]

AMD Instinct MI455X GPUs specs

- Up to 432 GB of HBM4 memory
- Up to 19.6 TB/s peak memory bandwidth
- 10x Gen vs Gen performance[4]



AMD Pensando "Vulcano" 800 AI NIC specs

- 800Gbps high-performance Ethernet network throughput
- Up to 8× scale-out bandwidth per GPU
- UAL | PCIe® Gen6 host interface for low-latency GPU communication
- UEC-ready RDMA Ethernet optimized for large-scale AI clusters

6th Gen AMD EPYC™ processor specs

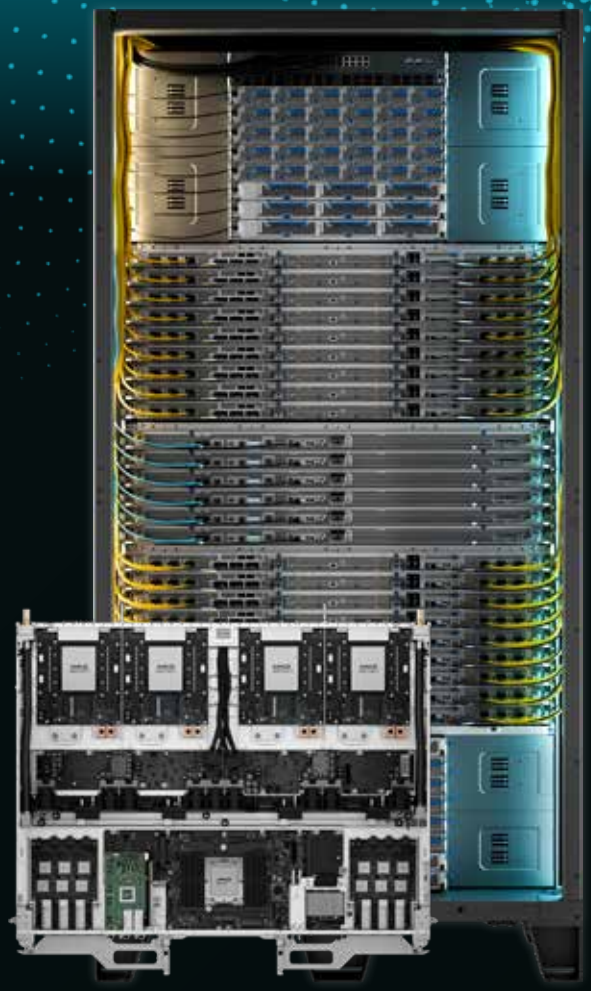
- Up to 256 cores[3]
- 2.0x CPU to GPU bandwidth[3]
- 1.7x Gen vs. gen performance[3]
- 1.6 TB/s memory bandwidth[3]

AMD Pensando DPU specs

- Securely bridge AI servers to enterprise networks
- Accelerate network, security, and storage offloads to boost AI server efficiency
- Up to 40× acceleration for networking functions versus CPU-only processing[5]
- 2× performance versus previous generation[6]
- 1.4× performance versus NVIDIA BlueField-3[7]

AMD AMD INSTINCT AMD EPYC AMD PENSANDO AMD ROCm

BUILDING THE FUTURE OF OPEN AI INFRASTRUCTURE



The AMD Helios rack-scale solution brings openness, performance, and rack scale integration together in a way the industry has never seen, giving hyperscalers and neoclouds a faster, more flexible path to next generation AI.

Learn more at:

amd.com

[1] Performance projection as of 06/05/2025 using engineering estimates based on the design of a future AMD Instinct MI400 Series GPU. For GenAI inference, an MoE model was evaluated for Instinct MI400 Series & MI355X with 2K & 16K prefill with TP8, EP8 and projected inference improvements for MI400 Series vs MI355X shows 10x better performance. Performance projection as of 06/05/2025 using engineering estimates based on the design of a future AMD Instinct MI400 Series GPU. A GenAI training model was evaluated for Instinct MI400 Series vs MI355X and projected improvements for GEMM and Attention Algorithms for MI400 Series. Results may vary when actual products are released to the public. MI350-050

[2] Source: AMD specs based on engineering projections, results subject to change

[3] PCIe Gen comparison based on PCI-SIG published statements, https://pcisig.com/pci-express-6.0-specification. 2P 6th Gen EPYC CPU with 128 lanes of PCIe Gen 6 and 5th Gen EPYC with 128 lanes of PCIe Gen 5 as of 6/3/2025. PCIe is a registered trademark of PCI-SIG Corporation. VEN-003

[4] Based on engineering projections by AMD Performance Labs in September 2025, to estimate the peak theoretical precision performance of seventy-two (72) AMD Instinct™ MI455X GPUs "Helios" AI Rack using MXFP4 dense Matrix datatype vs. an 8xGPU AMD Instinct MI355X platform using the MXFP4 dense Matrix datatype. Results subject to change when products are released in market. MI340-047B

[5] AMD, Driving Next Generation Scalability and Performance for Data Centers, May 11, 2025

[6] PEN-012: Measurements conducted by AMD Performance Labs as of Aug 27, 2024 on the current specification for the AMD Pensando™ Salina DPU accelerator designed with AMD Pensando™ 5nm process technology, projected to result in delivering 400Gb/s line-rate estimated performance. Estimated delivered results calculated for AMD Pensando™ Elba DPU designed with AMD Pensando 7nm process technology resulted in 200Gb/s line-rate performance. Actual results and specifications may vary based on production silicon.
Salina projected performance:
Bandwidth: 400Gbps
Connections per second: 10M
Packets per Second: 100MPPS
Encryption Offloads: 400 Gbps
Storage IOPS: 4 Million

[7] PEN-017: Testing conducted by AMD Performance Labs as of 15th April 2025 on the AMD Pensando Salina DPU, on a test system comprising of 2x Dual socket Xeon Dell power edge XE9680 function; Cisco 64x400G Switch; IXIA 2x400G tester as a traffic generator from Keysight; 2xAMD Pensando™ Salina DPU; 5th gen Xeon 8568 - 48 core CPU with PCIe Gen-5; Operating System Version : Ubuntu 22.04.5 LTS; Kernel Version: 5.15.0-139-generic; BIOS version: 1.3.6 [Mitigation: Off (default); System profile setting: Performance (default); SMT: enabled (default)]. AMD Pensando performs an average of 117 MPPS (millions packets per second) in AMD testing while Nvidia Bluefield-3 published performance is 80 MPPS [https://hc33.hotchips.org/assets/program/conference/day1/HC2021.NVIDIA.IdanBurstein.v08.norecording.pdf Slide 6] for ~1.45x the performance with AMD Pensando Salina DPU. Results may vary based on factors including but not limited to system configuration and software settings.