AMD HELIOS RACKSCALE SOLUTION

AN OPEN BLUEPRINT FOR BUILDING AI FACTORIES FOR AGENTIC AI AND BEYOND

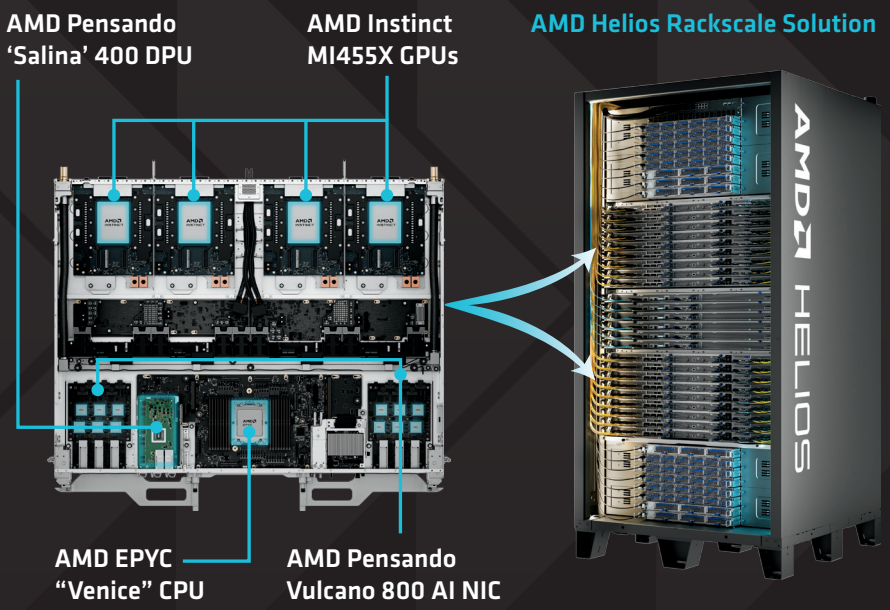
AMD HELIOS

INFRASTRUCTURE FOR THE NEXT ERA OF AI

Artificial Intelligence is moving from single-model interactions to agentic systems: AI agents that reason, retrieve knowledge, use tools, collaborate with other agents, and act across complex workflows. Agentic AI requires high-throughput inference, large memory capacity, fast GPU-to-GPU communication, secure multi-tenancy, scalable orchestration, and always-on operational reliability. The AMD Helios rackscale solution is designed to help hyperscalers, sovereign AI providers, OEMs, neoclouds, and data center teams build the infrastructure foundation for this next era of AI.

AMD Helios is an open, validated rack-scale AI reference architecture for building open AI factory infrastructure. The solution combines the latest AMD Instinct™ GPUs, AMD EPYC™ server CPUs, AMD Pensando™ networking, and open industry standards to enable large-scale inference, frontier model training, and fine tuning. It organizes 72 AMD Instinct MI455X GPUs into a single rack using 18 4-GPU compute trays, UALink™ over Ethernet for scale-up connectivity, standards-based Ethernet for scale-out networking, and the AMD ROCm™ software stack that simplifies workload migration and optimizes performance.

As a reference architecture, AMD Helios gives OEMs and ODMs a clear integration target. It helps hyperscalers and neoclouds scale from one rack to many racks, and provides infrastructure teams with a practical view of deployment requirements before data center decisions are locked.



AT A GLANCE	
SYSTEM TYPE	72-GPU rackscale AI reference rack / rackscale solution
RACK DESIGN	18× 1OU 1P:4G compute trays; 6 AMD Helios switch trays
GPU	72× AMD Instinct MI455X GPUs
CPU	18× AMD EPYC "Venice" server CPUs, one per compute tray
SCALE-UP	UALoE over Ethernet with single-switch-hop, multi-plane fabric, blind-mate rear cable cartridge
SCALE-OUT	Ethernet/UEC-aligned scale-out via AMD Pensando 'Vulcano' AI NICs; up to 3×800 Gb/s per GPU, or up to 43 TB/s rack bidirectional bandwidth
PERFORMANCE	Up to 2.9 EFLOPS OCP MXFP4; up to 1.4 EFLOPS OCP MXFP6/MXFP8
MEMORY	Up to 31 TB HBM4 per rack
MEMORY BANDWIDTH	Up to 1.67 PB/s aggregate per rack
COOLING	Liquid Cooled
SOFTWARE	AMD ROCm open software stack for rack-to-cluster AI deployment

SCALE FROM A SINGLE RACK TO AN AI FACTORY

SINGLE RACK	MULTI-RACK CLUSTER	AI FACTORY
A 72-GPU scale-up pod for dense AI compute, inference, training, fine-tuning, and agentic AI workloads.	Standards-based Ethernet scale-out links Helios racks into larger AI clusters for distributed inference, multi-agent systems, training, and fine-tuning.	A repeatable rack-scale architecture for hyperscalers, neoclouds, sovereign AI providers, and HPC organizations building for agentic AI and beyond.
BUILD	**CUSTOMIZE**	**BUY**
OEMs and ODMs can use the blueprint to create differentiated Helios-based offerings.	Hyperscalers can adapt the blueprint to specific data center environments while preserving common rack-level specifications.	Customers can deploy Helios-based systems through AMD ecosystem partners.

WHERE AMD HELIOS FITS

AGENTIC AI PLATFORMS	LARGE-SCALE TRAINING AND FINE TUNING
For organizations building AI agents that reason, plan, retrieve, call tools, coordinate with other agents, and operate across complex enterprise workflows.	For hyperscale AI, domain-specific models, pipelines are needed to support specialized AI agents.
HIGH-THROUGHPUT INFERENCE CLOUDS	**HPC + AI CONVERGENCE**
For neoclouds and AI cloud providers delivering multitenant GPU services, secure allocation, serviceability, high utilization, and open standards-based scale-out.	For environments that need dense GPU compute, high memory bandwidth, multi-GPU communication, and a software foundation spanning AI and HPC workloads.
SOVEREIGN AI FACTORIES	**CONTINUOUS FINE TUNING**
For on-premises or in-country AI infrastructure requiring control over data, operations, security posture, supply chain decisions, and deployment geography.	Adapt models across domains, languages and tenants on a continuous basis.

FIVE REASONS FOR BUILDING THE AMD HELIOS RACKSCALE SOLUTION

1	**LEADERSHIP COMPUTE PERFORMANCE**	Integrates 72 AMD Instinct MI455X GPUs into a unified in-rack compute domain for inference, training, fine-tuning, and multi-agent workloads, delivering delivering the scale and flexibility needed for inference, training, fine-tuning, and multi-agent workloads. Delivers up to 2.9 EFLOPS of OCP MXFP4 and up to 1.4 EFLOPS of OCP MXFP6/MXFP8 performance, up to 15% higher peak theoretical performance than NVIDIA® Vera Rubin NVL72 when comparing OCP MXFP4 to NVIDIA NVFP4. MI400-005
2	**SCALE FROM RACK TO AI FACTORY**	UALink over Ethernet connects GPUs through the in-rack switch fabric; standards-based Ethernet scales across racks and clusters.
3	**INDUSTRY'S BEST MEMORY**	Large HBM memory capacity, implemented as a single load-store domain in the rack, helps support high-concurrency needs of agentic AI workloads. With 5th Gen AMD CDNA architecture, AMD Instinct MI455X GPUs are designed to deliver exceptional memory scale and throughput, with up to 50% more memory capacity and up to 6% higher peak memory bandwidth than the NVIDIA Vera Rubin GPU. MI400-004
4	**FLEXIBILITY THAT COMES WITH AN OPEN BLUEPRINT**	The AMD Helios rackscale solution gives you a path to build, customize, or buy through an AMD multivendor ecosystem spanning OEMs, clouds, neoclouds, software partners, and infrastructure builders. Modify implementations for your specific workload needs.
5	**SOFTWARE AND LIFECYCLE SUPPORT**	AMD ROCm software and lifecycle tools help support orchestration, observability, distributed execution, and fleet-scale deployment.

BUILD YOUR AI FACTORY FOR AGENTIC AI AND BEYOND WITH AMD HELIOS

The future requires dense compute, high-bandwidth memory, fast scale-up, open scale-out, secure multi-tenancy, software orchestration, and operational control. AMD Helios is the open rackscale blueprint to power your AI factory.

Footnote explanations are available at: https://www.amd.com/en/legal/claims/instinct.html

LE-93205-00 0726 PID # 5158301