High-Performance AI Infrastructure Cisco and AMD

Overview

As demand for Artificial Intelligence and Machine Learning (AI/ML) workloads grow, organizations need to make their data centers equipped with the best components to meet the needs of their customers and outpace the competition. However, as AI/ML workloads scale out to distributed GPU clusters; maintaining consistent performance and efficiency becomes increasingly challenging. As clusters grow in size and complexity, coordination and data transfer across compute nodes become increasingly complex, introducing inefficiencies that can leave GPUs underutilized and degrade overall Job Completion Time (JCT). Organizations must address these infrastructure challenges to power their next generation AI/ML workloads.

The unified Cisco and AMD AI infrastructure components

Modern AI workloads demand infrastructure that can scale efficiently while delivering consistent performance — a challenge the joint solution from Cisco and AMD is designed to address holistically. It is built on the Cisco Unified Computing System™ (Cisco UCS®) and centered on AMD EPYC™ CPUs, AMD Instinct™ GPUs, AMD Pensando™ Pollara 400 AI NICs, and Cisco Nexus® One for AI Networking — including Cisco N9000 Series Switches powered by the Cisco® Silicon One® G200 switching processor. By integrating balanced compute, high-memory capacity, and an open, Ethernet-based AI networking layer, this solution enables customers to scale AI workloads predictably, maximize GPU productivity, and sustain performance from pilot deployments to massive, multinode clusters.

Cisco UCS® C885A M8 Rack Server

Cisco UCS® C885A M8 Rack Server is Cisco's dense GPU rack server purpose-built for large-scale AI training, fine-tuning, and inference workloads. It brings the Cisco and AMD architecture together in a single validated platform that combines 5th Gen AMD EPYC™ CPUs, AMD Instinct™ MI350X GPUs, and AMD Pensando™ Pollara 400 AI NICs for the east/west GPU interconnected fabric. Customers get a pre-validated AI node that shortens qualification cycles and accelerates time to production. Lifecycle and telemetry are unified through Cisco Intersight®, giving operators a consistent view from a single node to full multirack clusters.

Figure 1. Cisco UCS® C885A M8 Rack Server

Product highlights:

- Form factor: 8U dense GPU rack server
- GPUs: 8x AMD Instinct™ MI350X OAM accelerators
- CPUs: dual 5th Gen AMD EPYC™ processors with 12-channel DDR5 memory
- AI networking: 8x 400G AMD Pensando™ Pollara 400 AI NICs for east/west GPU interconnectivity
- Management: Cisco Intersight

AMD EPYC™ Server CPUs (5th Generation)

Figure 2. 5th Generation AMD EPYC™ Processor

AMD EPYC™ 9005 Series processors are the newest generation of the powerful and efficient AMD EPYC™ processor family for servers, which have set hundreds of performance and efficiency world records. Advancements in the EPYC 9005 processor family are enabled by AMD's breakthrough high-performance, highly efficient "Zen 5" processor core architecture and advanced microprocessor process technologies to better meet the needs of the modern AI-enabled data center. The complete line of processor offerings include a wide range of core counts (up to 192 cores, 384 threads per CPU), frequencies (up to 5 GHz), cache capacities, energy efficiency levels, and competitive cost points – all complemented by the familiar x86 software compatibility that allow AMD EPYC™ 9005–based servers to readily support almost any business need, while allowing customers to free up space and power to accommodate AI.

Product highlights:

- High core density: up to 192 cores per socket
- High memory capacity: up to 6 TB of DDR5 memory per socket
- High memory throughput: 12-channel DDR5 memory architecture
- Advanced I/O: up to 128 lanes of PCIe Gen5

AMD Instinct™ MI350 Series GPU platform

Figure 3. AMD Instinct™ MI350X Series GPU platform

AMD Instinct™ MI350 Series GPUs are the newest generation of AMD accelerators designed to advance generative AI and high-performance computing across modern data centers. Built on the 4th Gen AMD CDNA™ architecture, the MI350 Series introduces architectural and efficiency advancements to better support large-scale AI training, inference, and data-intensive workloads. The MI350 Series portfolio is designed for broad deployment flexibility across air- and liquid-cooled platforms and integrates seamlessly with industry-standard server infrastructure, networking, and the AMD ROCm™ software ecosystem – enabling customers to deploy scalable AI solutions while preserving software compatibility and operational consistency across the AI lifecycle.

Product highlights:

- GPU architecture: AMD CDNA 4
- Memory capacity: up to 288 GB HBM3E
- Memory bandwidth: up to 8 TB/s

Leading AI networking solution: AMD adapters and Cisco networking

The AMD Pensando™ Pollara 400 AI NIC and Cisco Silicon One®–based Cisco N9000 Series Switches deliver the next generation of performant UEC-ready networking for AI workloads.

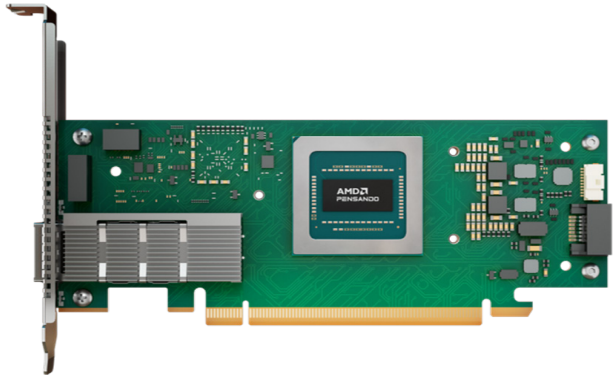

Figure 4. AMD Pensando™ Pollara 400 AI NIC

AMD Pensando™ Pollara 400 AI NIC

As AI clusters scale, network communication becomes a primary limiter of performance. Collective operations generate bursty, all-to-all traffic patterns that can overwhelm traditional Ethernet fabrics, leading to congestion, uneven traffic distribution, packet loss, and variability in message completion time. These network effects leave GPUs underutilized and directly increase overall job completion time.

The AMD Pensando™ Pollara 400 AI NIC is purpose-built to improve performance and efficiency in large-scale AI clusters by accelerating GPU-to-GPU network communication. Designed for high-bandwidth, communication-intensive AI workloads, the AMD Pensando™ Pollara 400 AI NIC enhances Ethernet-based networking for distributed training and inferencing by improving message completion time, traffic efficiency, and reliability for collective communication using UEC-ready RDMA.

UEC-ready RDMA brings intelligent load balancing and path-aware congestion control to dynamically steer traffic away from congested or impaired paths, helping to maintain consistent performance as clusters scale. Built-in support for out-of-order packet handling with in-order message delivery, selective retransmission, and fast failure recovery reduces stalls caused by packet loss or transient network issues.

Together, these capabilities keep GPUs highly productive, minimize idle time, and enable AI clusters to scale efficiently across hundreds or thousands of nodes—delivering faster collective communication and more predictable training performance for workloads such as large language models, recommendation systems, and other data-intensive AI applications.

Cisco Nexus® One for AI networking

Cisco Nexus® One for AI networking delivers a comprehensive, integrated stack comprising of silicon, switches, optics, and software – all managed through a unified operating model. This architecture offers the flexibility to choose between different silicon, software, and operating model, all of which use high-performance Cisco N9000 Series Switches. Key features include advanced congestion-control mechanisms, robust integrated security, and centralized operations through the on-premises Cisco Nexus Dashboard or the cloud-managed Cisco Nexus Hyperfabric™.

Cisco further strengthens this foundation with validated reference designs developed in collaboration with trusted partners such as AMD, enabling organizations to deploy high-performance AI infrastructure confidently.

Cisco offers an Ultra-Ethernet-ready infrastructure to maximize GPU utilization and ROI while maintaining a flexible, open architecture. Innovations include intelligent packet flow, a policy-driven load-balancing engine, including Dynamic Load Balancing (DLB) that dynamically adapts to diverse traffic patterns. By supporting multiple schemes, including flowlet, per-packet, and Equal-Cost Multipath (ECMP), this solution optimizes performance across converged storage networks.

The synergy between Cisco's intelligent packet flow solution and the AMD Pensando™ Pollara 400 AI NIC is a critical differentiator. By leveraging the AMD Pensando™ Pollara 400 AI NIC's out-of-order packet handling, this solution effectively mitigates congestion caused by "elephant flows" impacting "mice flows." This proactive approach improves not only AI training but also increases the efficiency of inference jobs.

Designed for high-speed, low-latency AI fabrics, Cisco N9364E-SG2 switches (Figure 5) are powered by Cisco Silicon One® G200 switching processors. This 2-Rack Unit (RU) platform supports 51.2 Tbps throughput and a wide range of interface options, including 64 x 800 GbE, 128 x 400 GbE, 256 x 200 GbE, or up to 512 x 100 GbE ports.

Figure 5. Cisco N9364E-SG2 switch

For additional deployment flexibility, Cisco offers Cisco N9K-C9364D-GX2A and Cisco N9K-C9332D-GX2B switches, both powered by Cisco CloudScale Silicon. The Cisco N9K-C9364D-GX2A (Figure 6) is a 2-Rack Unit (RU) platform supporting high-density configurations of 64 x 400 GbE, 128 x 200 GbE, or 256 x 100 GbE. For space-constrained environments, the Cisco N9K-C9332D-GX2B provides similar high-performance capabilities in a compact 1-Rack Unit (RU) form factor, supporting 32 x 400 GbE, 64 x 200 GbE, or 128 x 100 GbE interfaces.

Figure 6. Cisco N9K-C9364D-GX2A switch

Cisco Nexus Dashboard (Figure 7) delivers consistent, simplified management across all networks in an AI cluster, including storage, frontend, management, and inter-GPU backend networks – while also offering unified visibility and actionable analytics to optimize performance and streamline operations.

Figure 7. Cisco Nexus Dashboard

The Cisco and AMD advantage for AI infrastructure

Cisco and AMD deliver more than just full-stack AI infrastructure; they provide the critical innovations required to maximize GPU efficiency and minimize Job Completion Time (JCT). Together, they transform AI clusters into high-performance engines for business innovation.

Key benefits of the unified Cisco and AMD AI infrastructure include increased ROI due to maximum GPU utilization, improved cost per token due to operational simplicity, increased uptime due to full-stack observability, integrated security with multitenancy, and an open architecture based on Ultra Ethernet, as described below.

Increased ROI due to maximum GPU utilization

Cisco N9000 Series Switches provide high-speed (up to 800 Gbps per port) connectivity not only for the inter-GPU compute networks, but also the frontend and the storage networks of an AI cluster. This high-speed connectivity together with robust congestion control improves RDMA performance as GPUs sync their states or read and write data to the centralized storage. Cisco Silicon One® G200 switching processor, the powerhouse in the Cisco N9364E-SG2 switches, offers 256 MB of on-die packet buffers, fully shared among all the ports. This large-size buffer helps in absorbing congestion instead of spreading it. The switches can detect network congestion and then notify AMD NICs, enabling the NICs to adjust the congestion control mechanism. Cisco's innovative intelligent packet flow takes control from reactive to proactive, allowing GPUs to communicate without network congestion, leading to improved GPU utilization and return on investments.

AMD focuses on enabling efficient scale out by balancing compute, memory, and networking so distributed AI workloads operate as a cohesive system. By reducing coordination overhead and communication delays, GPUs spend more time performing useful work, improving overall system efficiency and return on infrastructure investment.

Cisco Nexus Dashboard provides consistent and simplified operations of all networks of an AI cluster, not only the storage network, but also the frontend network, management network, and the compute network.

Improved cost per token due to operational simplicity

Key benefits include:

- **Faster provisioning of RDMA fabrics:** built-in AI fabric types and templates with fine-tuned thresholds for ECN and Quality of Service (QoS) of RoCEv2 traffic
- **Consistent operational experience:** support for all widely used fabric technologies, such as VXLAN, routed fabrics (see Figure 7), and even the inter-fabric connectivity, thereby delivering a consistent operational experience for all networks of an AI cluster and even the peripheral networks
- **Congestion analytics:** real-time congestion scoring and statistics such as ECN, drops, and microburst detection
- **AI job observability:** integration with Simple Linux Utility for Resource Management (SLURM) provides visibility into AI workloads, networks, NICs, and GPUs
- **Anomaly detection:** proactive identification of performance bottlenecks with suggested remediation after an automatic correlation of distributed AI job and health of the network paths
- **Sustainability insights:** energy consumption monitoring and optimization recommendations

AMD emphasizes open, standards-based infrastructure to simplify deployment and operations across the AI lifecycle. By aligning with industry ecosystems and familiar software environments, organizations can scale AI workloads with less friction, improving efficiency, and lowering the operational cost per token.

By simplifying operations, Cisco and AMD offerings lower the cost of operating an AI cluster, thereby improving the cost per token.

Increased uptime due to full-stack observability

Cisco provides end-to-end visibility of AI applications, which helps in detecting real-time performance and compliance risks. For example, Splunk® Observability ensures digital resilience of your AI applications, infrastructure, and business processes with real-time visibility into performance issues, root causes, and business impact. Cisco Nexus Dashboard provides network congestion analytics with real-time congestion scoring, statistics such as ECN, drops, and microburst detection. Cisco Nexus Dashboard also provides AI job observability through integration with Simple Linux Utility for Resource Management (SLURM). The built-in anomaly detection feature allows proactive identification of performance bottlenecks with suggested remediation after an automatic correlation of distributed AI jobs to the health of the network paths.

AMD supports predictable AI performance by enabling greater visibility and resilience across the data path. Intelligent handling of congestion and transient failures helps minimize variability and reduce disruptions that can impact distributed job execution at scale.

This level of full-stack observability leads to increased uptime of AI clusters by proactive detection and resolution of issues.

Integrated security with multitenancy

Cisco N9000 Series Switches support hardware-based VXLAN tunnel endpoint (VTEP) functionality for strict isolation of clients, workloads, and storage appliances. This helps with multi-tenancy of different workloads using the shared infrastructure while maintaining the security domains.

AMD enables secure, multitenant AI deployments by supporting isolation and policy enforcement as part of an open infrastructure approach. This allows shared environments to scale while maintaining performance, control, and operational consistency.

Open architecture based on Ultra Ethernet

Both Cisco and AMD are the founding members of the Ultra Ethernet Consortium (UEC) to build a complete Ethernet-based communication stack architecture for high-performance networking for AI and HPC workloads. The Ultra Ethernet 1.0 specification required features for the network of an AI cluster are already supported by Cisco and AMD Nexus® One, and both are committed to not only develop, but also jointly validate the advancements. A key advantage of Cisco Silicon One® is its programmable architecture, which allows newer features to be deployed without a hardware refresh.

AI clusters built using the open architecture of Cisco and AMD solutions benefit from increased affordability, better talent availability, broader ecosystem adoption, and faster innovation through community contributions.

AMD advocates for open connectivity as a foundation for scalable AI infrastructure. By advancing Ethernet for AI workloads and supporting open standards, AMD gives customers flexibility to scale performance while preserving interoperability and long-term choice.

Powering AI innovation: the Cisco and AMD advantage

As AI workloads scale across distributed GPU clusters, performance increasingly depends on efficient coordination and predictable data movement. Cisco and AMD align on an open, Ethernet-based approach that reduces inefficiencies, improves operational confidence, and sustains performance as AI environments grow.

To summarize, Cisco and AMD solution for AI infrastructure has the following advantages:

- Enables efficient scale-out of distributed AI workloads
- Keeps GPUs productive by reducing coordination overhead
- Simplifies operations through open, standards-based design
- Improves stability and uptime with end-to-end visibility
- Preserves flexibility and choice as AI infrastructure evolves

C22-6109251-00 05/26