

7 ARCHITECTURAL PRINCIPLES OF SCALABLE *AI NETWORKING*



TABLE OF CONTENTS

Key Takeaways/Executive Summary	3
The Network Is the Strategic Foundation for AI Systems	4
Architecture Must Scale in Both Bandwidth and Behavior	9
Determinism and Predictability Are First-Order Requirements	11
Programmability Enables Workload-Aware Optimization	12
Operational Excellence	14
Openness Preserves Architectural Optionality	16
Architected for AI Growth	19
Defining the Next Era of AI Networking	21
Sources	22

KEY TAKEAWAYS / EXECUTIVE SUMMARY

- **AI workloads evolve faster than infrastructure, with AI compute demand historically doubling every six months.**
Making AI infrastructure investments future-ready requires a shift from a static hardware-centric mindset to a flexible, software-defined ecosystem that prioritizes data quality, interconnection, and scalable power.
- **Peak performance benchmarking for traditional networking is generally inadequate for AI networking.**
While traditional data center benchmarks prioritize maximum data transfer speeds under ideal conditions, AI workloads are bottlenecked by tail latency and communication synchronization, rather than peak network throughput.
- **Standards-based Ethernet fabrics are one of the only architectures designed to meet the demands of AI at scale.**
A strategy centered on open, standards-based Ethernet lays the foundation for high-performance, flexible, and vendor-agnostic scaling from racks to fabrics across multiple data centers.
- **Most AI networking today ignores existing infrastructure realities, offering disruption or invasive infrastructure instead of credible transition paths.**
Preserving architectural optionality helps ensure AI infrastructures can evolve as quickly as the AI technologies they support, fostering innovation without costly rip-and-replace cycles.
- **Operational excellence in networking is no longer optional.**
As AI operates at production scale, compute performance isn't enough. Technology leaders and networking teams need to ensure networks remain performant, consistent, observable, and upgradable without impacting operations.

As AI workloads expand from training to distributed inferencing and agentic AI, networking has become a defining pillar, shaping economics, interoperability, and vendor freedom far beyond initial performance decisions. Architectures optimized for early gains often harden over time into structural dependencies that become costly and difficult to unwind.

AI networking should not be a technical afterthought but a strategic foundation that governs long-term flexibility, cost behavior, and customer choice and control in the data center. Open, standards-based Ethernet fabrics are among the few architectures designed to meet the demands of AI at scale—whether that means scaling up the GPU memory space, scaling out to a cluster, or scaling the cluster across data centers.

Choosing the right networking infrastructure is one of the most consequential strategic decisions a technology leader can make. These are the seven core AI networking principles essential for the modern AI era.

THE NETWORK IS THE STRATEGIC FOUNDATION FOR AI SYSTEMS

As a strategic foundation, the network orchestrates data movement, job scheduling, and resource coordination across compute, storage, and accelerators when scaling up within a single cluster or scaling out or across distributed infrastructure. As AI systems shift from episodic training to continuous inference and agentic execution, the network increasingly governs responsiveness, utilization, and system-level behavior.

HOW ARE GENAI AND AGENTIC AI RESHAPING AI NETWORKING

These trends are reshaping AI networking through key structural shifts. For example, GenAI architectures have made split-network topologies standard in modern data centers¹: a front end that leverages standard Ethernet to manage standard network traffic, and a back-end AI fabric providing an isolated, high-bandwidth, ultra-low-latency highway dedicated to GPU-to-GPU communication.

While standard GenAI is typically a linear “prompt-in, response-out” transaction, Agentic AI sets off a chain reaction of independent, multi-step actions comprising complex, multi-turn traffic loops that place a significantly heavier burden on the network’s front-end infrastructure and API layers.

Both GenAI and Agentic AI workloads rely heavily on the same underlying LLMs, meaning they share a fundamental need for ultra-low tail latency, massive bandwidth, and lossless data delivery. That means they also share a heavy focus on minimizing packet delays by leveraging technologies like packet spraying and advanced congestion telemetry to throttle or reroute traffic instantly before data drops occur.

AGENTIC AI: THE NEXT WAVE OF TRANSFORMATION

Agentic AI is an architectural evolution built directly on top of GenAI. Representing the next wave of AI transformation, it consists of intelligent agents that not only automate tasks but also drive complex workflows, innovation, and decision-making across industries. These agents rely on a network of interconnected models, applications, and data, demanding more than just GPU horsepower. They need high-performance CPUs to orchestrate tasks, secure data flows across the network, and manage AI workloads seamlessly.

GENERATIVE AI: NOVELTY TO UTILITY

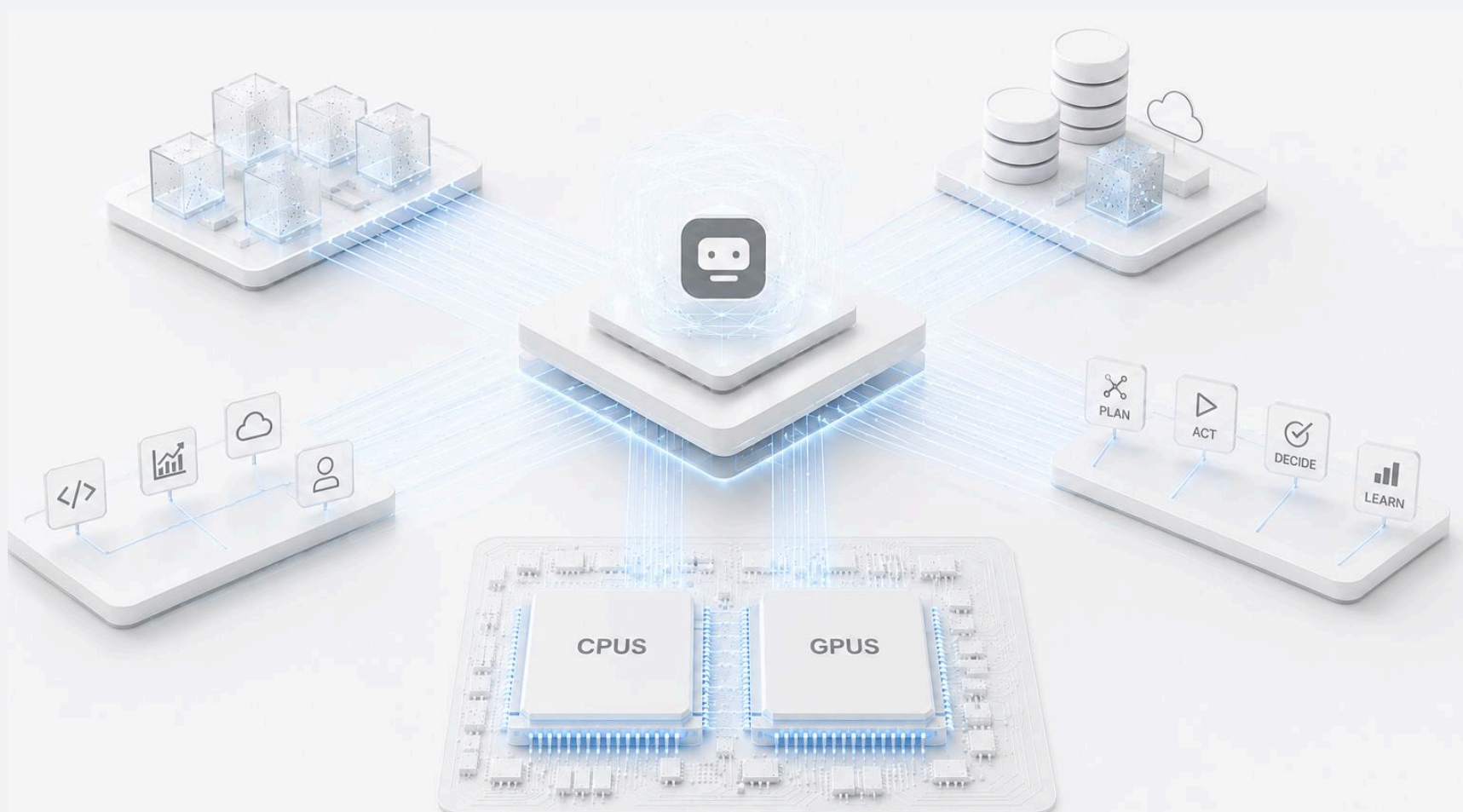
Since the launch of OpenAI's ChatGPT in 2022, Generative AI (GenAI) has shifted from an experimental novelty to a critical utility that delivers massive gains in productivity, creativity, and automation. Unlike older predictive AI models that merely categorize data, GenAI creates contextually rich text, code, imagery, and video from simple prompts.

This shift is reshaping the foundational priorities of data centers, forcing a transition from traditional compute networks to highly specialized AI back-end fabrics. Because training and running massive large language models (LLMs) require continuous, heavy parallel synchronization across thousands of GPUs, the underlying network has evolved from a basic utility into a primary factor governing overall application performance.



A CRITICAL NERVOUS SYSTEM

The network acts as the critical nervous system for training, deploying, scaling, and utilizing AI models. During model training, the network synchronizes terabytes of data across thousands of nodes, ensuring that “all-to-all” communication doesn't become a bottleneck and that expensive hardware doesn't idle when not fully utilized. When deploying, scaling, and utilizing models, the network's programmability and intelligence allow it to dynamically route real-time inference traffic and handle the bursty demands of real-world use so models remain responsive and accessible even as user volume grows.



**RIISING TRENDS:
GENERATIVE AI AND AGENTIC AI**

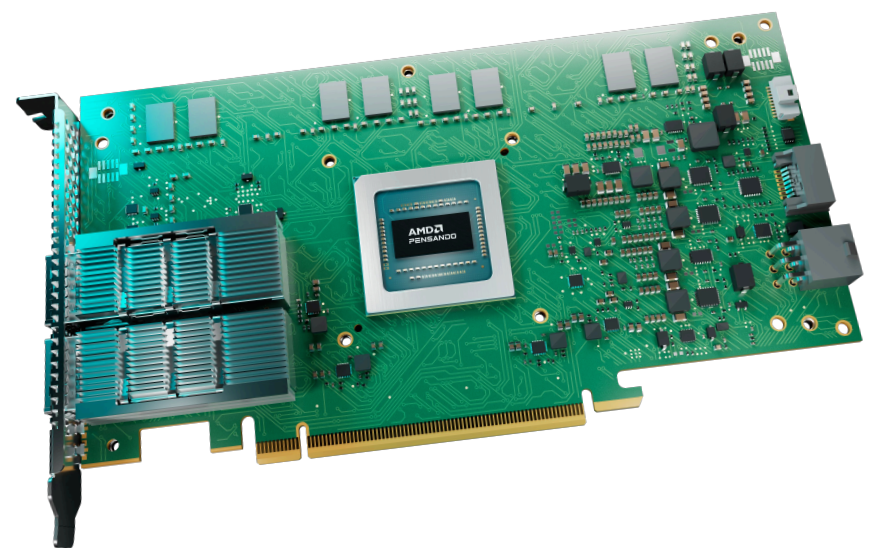
The next frontier of computing demands a network architecture that can keep pace with the evolving AI landscape.

HOW KV CACHE WORKS ON THE FRONT END

Large language models reuse previously computed attention states, known as the KV cache, to avoid recomputing them on every token. As agentic workflows grow longer context windows and reference large, shared knowledge stores, the volume and movement of KV cache data becomes one of the dominant traffic patterns on the front end. Moving KV cache acceleration onto the front-end infrastructure layer, for example onto a programmable DPU such as the **AMD Pensando™ Salina DPU**, keeps this traffic off the CPU, raises tokens per second, and shortens query response times for interactive and agentic applications.

The architectural principle is the same one that applies to the back-end fabric: separate the domains, but compose them into one coherent, programmable system.

Front-end DPUs handle data ingestion, storage offload, security enforcement, and KV cache acceleration, while back-end AI NICs handle high-bandwidth GPU-to-GPU collective communication. Together, they optimize the full path from storage to GPU to GPU, which is what determines real-world tokens per second and job completion time for agentic workloads.

**THE FRONT-END NETWORK BECOMES THE
INFERENCE DATA PATH**

With GenAI, the front-end network primarily ingests data, connects to storage, and enforces security so the back-end fabric can keep GPUs fed. In the agentic era, the front end takes on a much larger role. It now sits directly on the inference data path, managing key-value (KV) cache traffic, dynamically balancing load across GPU resources, and enforcing multi-tenant security at line rate, all without burdening the CPU.

This shift is driven by a changing system balance. The CPU-to-GPU ratio is moving from roughly one CPU per four GPUs closer to one CPU per GPU, because the CPU is constantly feeding accelerators, coordinating inference, and moving data across the system. That simultaneously drives demand on the front-end network, the scale-up interconnect, and the scale-out fabric. If the front end can't keep pace, GPUs sit idle waiting for context, regardless of how much back-end bandwidth is available.



SCALING INFRASTRUCTURE: UP, OUT, AND ACROSS

When workload demand increases, AI infrastructure must scale. There are three different methods of modifying AI infrastructure.

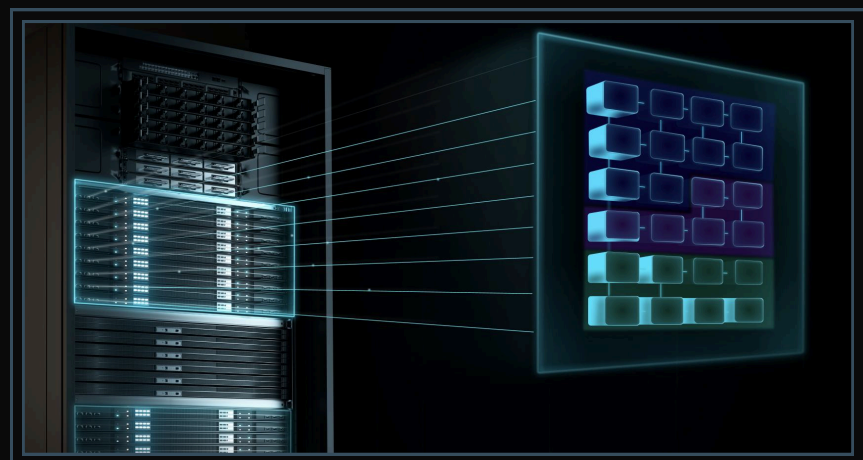
Technology leaders must carefully evaluate each approach to ensure the best performance to meet the demands of AI/ML workloads, along with cost-efficiency and flexibility.

A strategy centered on open standards might use high-performance UALink (Ultra Accelerator Link) for scaling up within the rack, and open, optimized Ultra Ethernet for scaling out across racks, managed by the ROCm (Radeon Open Compute) 7 open software ecosystem. The result is high-performance, flexible, and vendor-agnostic scaling from racks to multiple data centers.

SCALE UP (VERTICAL SCALING)

Scale up enhances the power of one machine or node by adding more processors, memory, or GPUs and connecting them through a high-bandwidth interconnect. As the primary approach for handling models that exceed the memory capacity of a single GPU, it allows multiple GPUs to share memory and operate as a single large logical GPU.

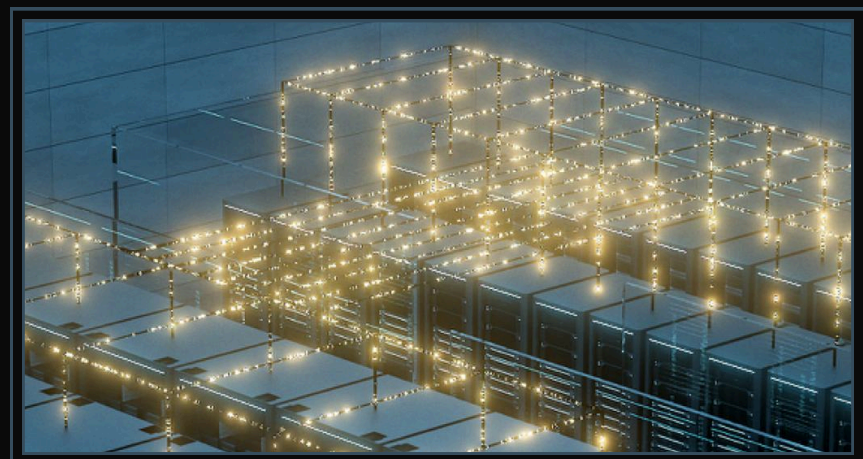
Use cases: Training extremely large models (via tensor parallelism) where GPUs need direct, memory-level access to High Bandwidth Memory (HBM), and high-frequency data sharing.



SCALE OUT (HORIZONTAL SCALING)

When a single server hits its physical performance ceiling, the only way to increase total compute capacity is to add more nodes or clusters. Scale out achieves this by connecting more server nodes through a high-speed network to form a large-scale compute cluster.

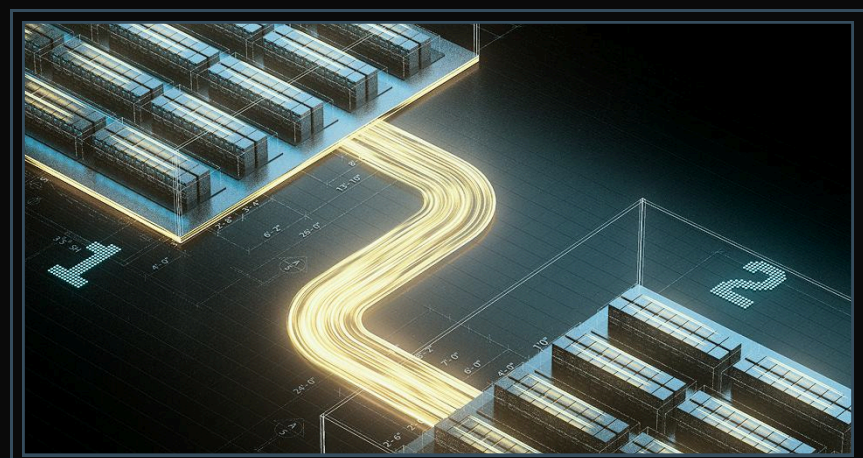
Use cases: Data parallelism, pipeline parallelism, and distributed inference across thousands of GPUs.



SCALE ACROSS (GEOGRAPHIC/MULTI-SITE SCALING)

As a significant evolution in scaling strategy, scale across is an emerging approach to extending computing resources beyond the boundaries of a single data center across multiple, geographically dispersed data centers to function as a single logical supercomputer (e.g., using high-speed optical Data Center Interconnect, or DCI).

Use cases: Very large-scale training, geo-distributed inference, and increasing capacity when a single data center reaches physical limits in power or cooling.



ARCHITECTURE MUST SCALE IN BOTH BANDWIDTH AND BEHAVIOR

To deliver peak performance and reliability, AI networks must scale in three ways: scale **up** for tightly coupled accelerator domains, **out** across growing clusters, and **across** data centers or workload types while maintaining predictable latency, congestion control, and observability. This includes supporting bursty training collectives, iterative GenAI pipelines, and latency-sensitive inference flows, often concurrently, without compromising predictability.

BUILDING HIGH-PERFORMING AI INFRASTRUCTURE

Cloud service providers, hyperscalers, and organizations in general are looking to maximize the performance and resilience of their AI clusters. Advanced AI/ML models have unique networking requirements, including tightly coupled parallel processing, rapid data transfers, and low-latency communication. Mission-critical AI applications rely on the highly efficient synchronization of data in AI GPU clusters, which demands continuous cluster uptime.

However, the network is often cited as a primary bottleneck in GPU utilization. Data transfer speed only matters if the network is properly optimized to capitalize on it, and performance benchmarking for traditional networks is generally inadequate for AI networks.

WHY DOES PEAK PERFORMANCE BENCHMARKING FAIL FOR AI NETWORKING DECISIONS?

Peak performance benchmarking often fails to guide AI networking decisions because AI workloads are bottlenecked by tail latency and communication synchronization, not peak network throughput. While traditional data center benchmarks prioritize maximum data-transfer speeds under ideal conditions, AI training and inference operate as highly synchronized, multi-node collective operations. If a single network packet is delayed, the entire GPU cluster stalls.

NOT ALL AI TRAFFIC IS CREATED EQUAL

Handling the sheer volume of AI traffic is just part of the challenge. AI-generated traffic also presents a variety of new data behaviors that can pose their own networking challenges.

AI traffic runs the gamut, from bursty mice flows used in AI inference to massive elephant flows that can saturate network links for hours, days, or weeks at a time.

BURSTINESS

Unlike traditional smooth traffic where data flows at a relatively constant rate—such as streaming music or video—certain types of AI processes and applications produce sudden, intense data spikes, followed by idle periods. This "burstiness" is common in distributed training and a wide variety of applications and frameworks.

Examples include:

➤ **Agentic AI virtual assistants or AI-powered customer service agents:**

A single user inquiry can trigger multiple internal steps, including searching databases, consulting documents, and calling other AI agents before producing a final answer, then returning to low activity. These real-time inference systems leverage Retrieval-Augmented Generation (RAG), an AI framework that enhances Large Language Models (LLMs) by fetching data from external sources.

➤ **AI-enabled network security tools:**

Agents that analyze network traffic for anomalies in real time may experience high network usage bursts when a security incident triggers a thorough, large-scale inspection.

➤ **Distributed AI training and synchronization:**

Large AI models that are “sharded” (split) across multiple nodes require frequent, massive synchronized transfers of data, causing high spikes in traffic within data centers and across networks.

➤ **Multi-modal generative AI workloads:**

Tools that process and produce complex media (images, video, text) generate large bursts of uplink traffic when submitting complex prompts, followed by high-bandwidth download spikes to display the generated content.

Bursty traffic requires specialized handling to prevent the sudden, irregular surges of data transmission from overwhelming network links and causing high packet latency, bottlenecks, and buffer saturation. The network architecture, therefore, needs to support bandwidth peaks that are much higher than the average data rate, handle the unpredictability of triggering events, and demonstrate high latency sensitivity through priority-based queuing for latency-sensitive traffic, so that bursts of data are absorbed and delivered with minimal delay for rapid analysis.

ELEPHANT FLOWS

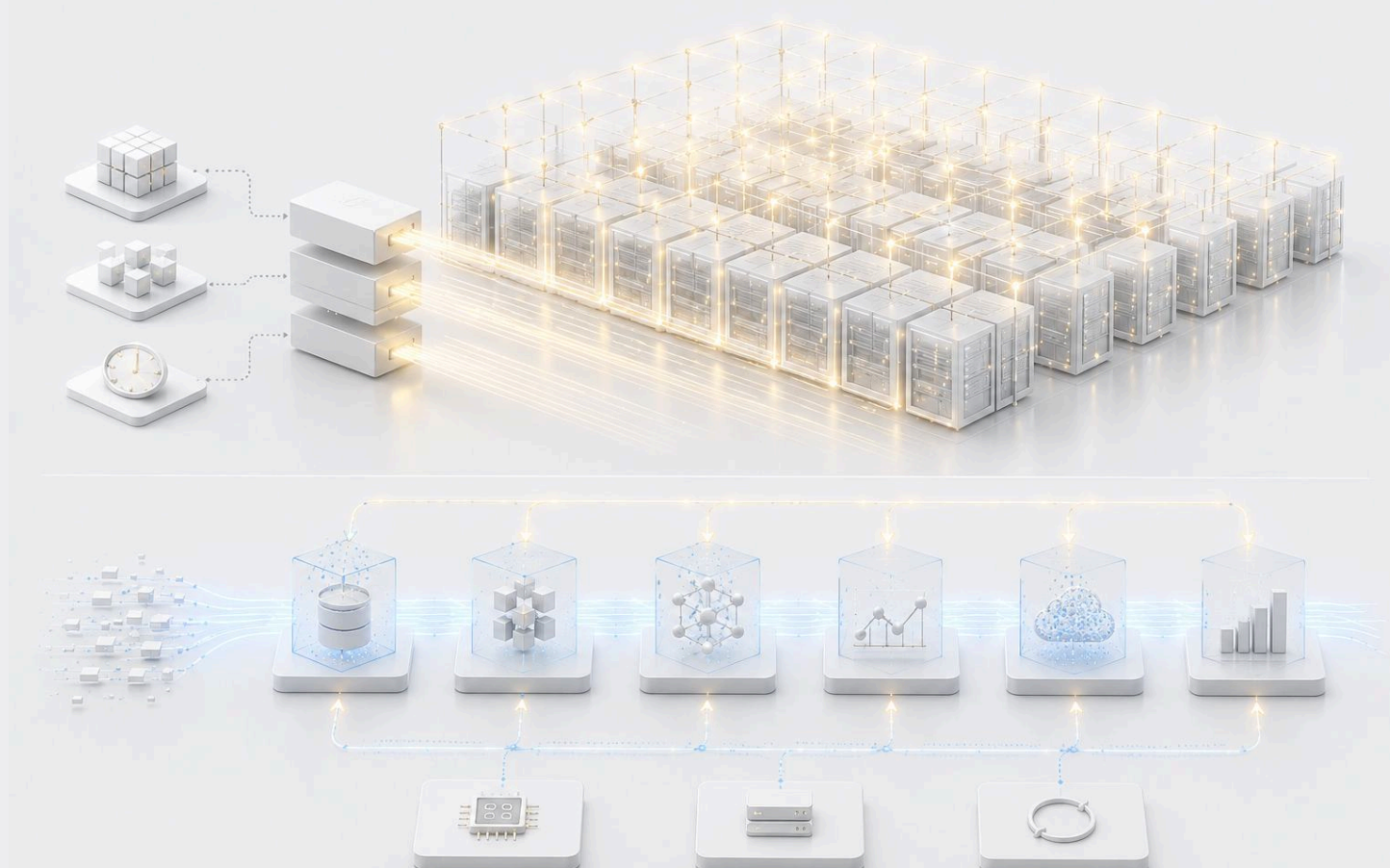
Elephant flows consist of large, continuous streams of data, often ranging from terabytes to petabytes, and are seen in processes such as AI batching. In batch-driven training, traffic often hits the network simultaneously from multiple senders. A distinguishing feature of AI batching is intense “east-west” traffic—high-volume data flows occurring horizontally between GPUs within a data center.

Batch processing in AI networks demands high throughput and parallelization. Unlike real-time inference, which may leave hardware idle between requests, batching ensures high GPU utilization. And because a batch job is not complete until every packet is received, the network must minimize tail latency or risk stalling the entire multi-million dollar cluster.

ITERATIVE GENAI PIPELINES

While batch processing focuses on high-throughput, scheduled, and cost-effective processing of large, static datasets, iterative GenAI pipelines involve continuous refinement, immediate feedback, and deployment of generative models for complex, evolving tasks. Unlike traditional linear pipelines, these end-to-end automated systems are designed to function as cyclical workflows where each stage—from data ingestion to model serving—feeds back into the next for rapid, repeatable improvement. Iterative GenAI pipelines place unique pressures on network infrastructure, requiring GPU-centric architectures with significant back-end bandwidth to handle massive data flows during frequent retraining cycles. Modern pipelines often include agentic AI that executes multi-step strategies, demanding deterministic performance and deep telemetry from the underlying network. While real-time responsiveness is important, these pipelines prioritize consistency over speed, with the primary goal being training-serving consistency—ensuring the model in production matches the quality and data standards of the training environment.

AI networking architectures need to scale across varied traffic patterns, models, and training methods while remaining predictable and delivering adaptive throughput.



DETERMINISM AND PREDICTABILITY ARE FIRST-ORDER REQUIREMENTS

AI networking must deliver predictable performance across diverse AI workloads. Training efficiency, GenAI responsiveness, and inference consistency all depend on controlled latency, effective congestion management, and workload isolation as workload scale and utilization increase.

WHY DETERMINISM AND PREDICTABILITY MATTER

Determinism and predictability in AI networking are critical because they ensure that data flows consistently through AI training and inference pipelines, enabling optimized, predictable performance at scale, as well as supporting security, reliability, and trust in automated systems.

Together, they offer the foundation for trustworthy, high-performance AI networking, balancing the need for creative AI insights with robust, reliable infrastructure. Both exert their own specific AI networking demands, including reliable packet delivery, minimal congestion loss, and predictable, stable response times through bounded latency.

DETERMINISTIC AI ENSURES PREDICTABILITY

Deterministic AI is one of three basic types of AI (along with probabilistic AI and generative AI), each of which operates differently, excels in specific applications, and presents unique opportunities for innovation.

Deterministic AI operates on predefined rules and logic, ensuring predictable, consistent, and identical outputs for the same inputs. Because it's used for rule-based tasks requiring high accuracy, deterministic AI systems are preferred when the cost of error is high and the need for explainability (e.g., auditability) is paramount.

Key use cases include high-stakes, regulated environments like financial compliance, healthcare scheduling, automated legal disclosures, IT security, and robotic process automation (RPA), where repeatability is mandatory.

HYBRID AI WORKFLOWS

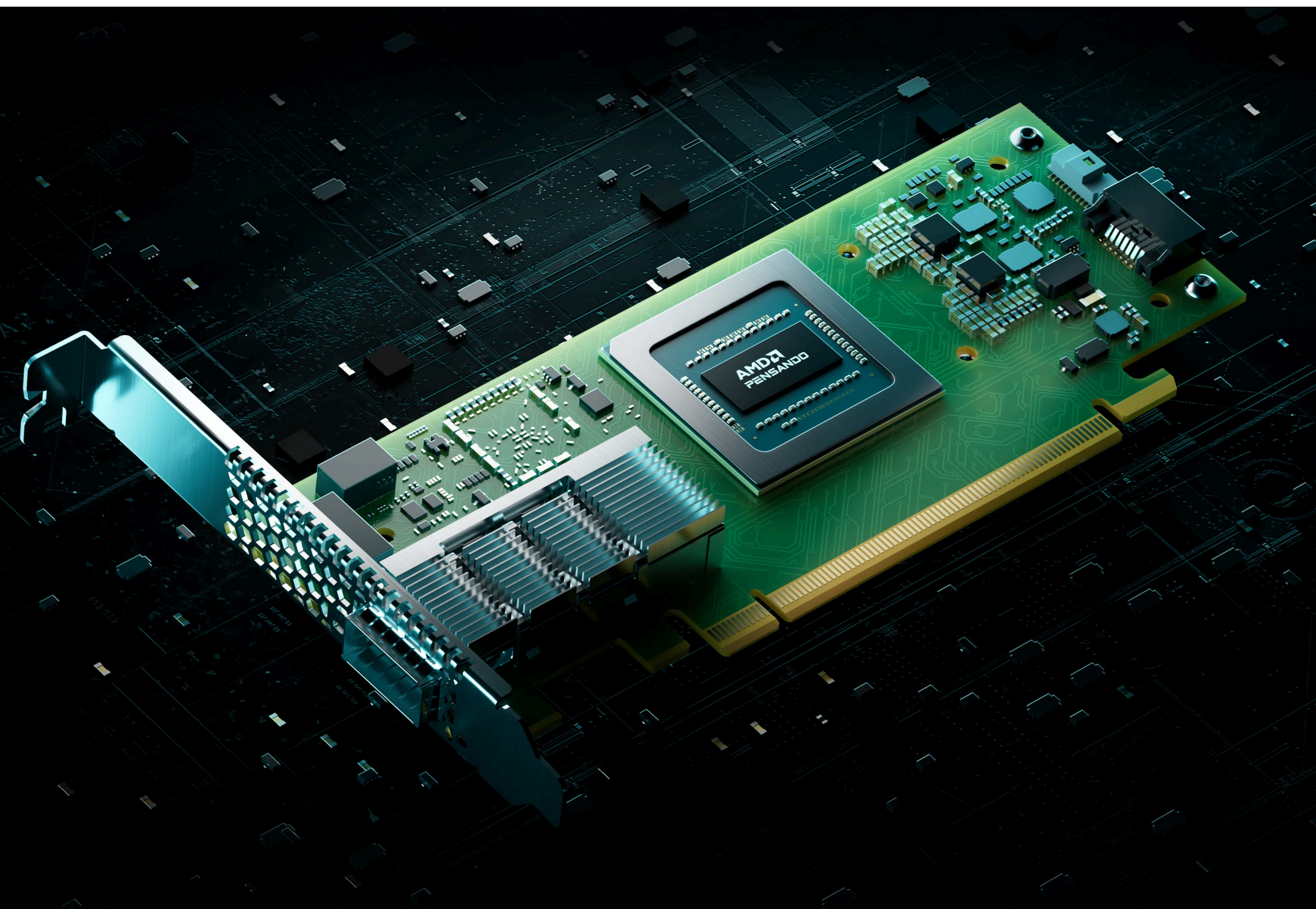
Predictability and clear outcomes matter just as much as generative capability. For this reason, a deterministic layer is often paired with probabilistic AI/GenAI, allowing deterministic networking and logic to “anchor” AI models.

These hybrid workflows rely on probabilistic AI/GenAI’s ability to interpret language, classify data, or make judgments, then execute those decisions in a predictable way. However, even when deterministic AI systems use AI models (such as LLMs) that are probabilistic under the hood, the overall system is deterministic.

Hybrid AI workflows require a hybrid architecture. In these systems, probabilistic AI is deployed inside deterministic workflows, and the AI models may be configured to behave predictably—for example, by fixing model settings or tightly constraining how outputs are used. The system is designed so that the same input leads to the same result under defined constraints.

KEYS TO PREDICTABLE NETWORK PERFORMANCE

Networking features like Intelligent Packet Spray are critical for managing the high bandwidth and low latency required by large AI models, while Path-Aware Congestion Control optimizes network performance through intelligent load balancing, automatically avoiding congested paths and helping sustain near-wire-rate performance during transient congestion. Capabilities like these are found in deep buffer switches, but also, more affordably, in the best network accelerators. The **AMD Pensando™ Pollara 400 AI NIC** and **AMD Pensando™ Vulcano 800 AI NIC**, for example, offer high throughput across all available links, along with congestion avoidance, reduced tail latency, scalable performance, and fast job completion times.



PROGRAMMABILITY ENABLES WORKLOAD-AWARE OPTIMIZATION

Programmability keeps open architectures viable, allowing networks to be optimized for collectives, congestion control, and emerging AI workloads as they scale up in complexity and scale out in volume. Programmable networks can adapt to the distinct demands of training, GenAI pipelines, and inference—optimizing traffic patterns, prioritization, and resource usage as workloads evolve.

WHY IS PROGRAMMABILITY IMPORTANT IN AI NETWORKING?

Programmability is critical in AI networking because it allows the infrastructure to adapt to the massive, dynamic workloads that modern AI requires. Unlike traditional “fixed-function” networks that use hardwired logic, programmable networks use software to manage how data moves and adapts to congestion, enabling the agility needed to support large-scale GPU clusters and high-throughput AI training.

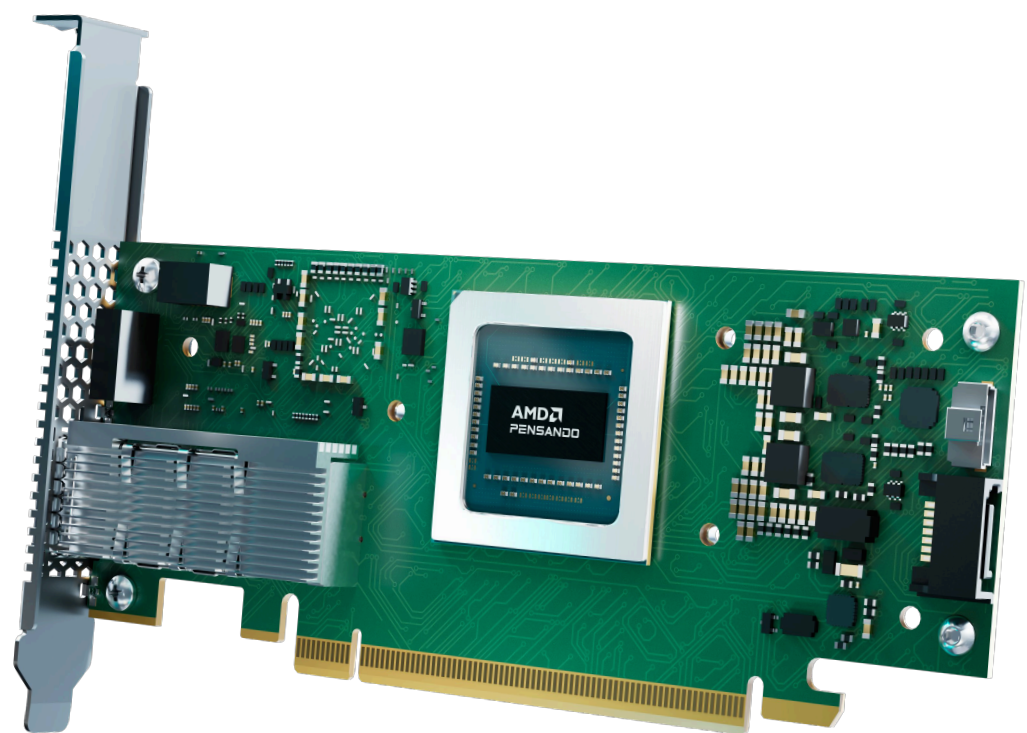
PROGRAMMABILITY FOR TUNABILITY AND FUTURE FLEXIBILITY

While standards are a vital starting point, each organization’s AI infrastructure has unique characteristics based on the specific infrastructure components in use and the software used to train models, manage data, and execute inferencing. Every AI infrastructure instance must therefore be tunable to adjust operations in ways that can reduce training times, accelerate inferencing, and eliminate wasted capacity.

In AI networking, programmability uses software, APIs, and AI-driven, intent-based systems to dynamically configure, manage, and optimize network behavior, moving away from manual, device-by-device configuration.

PROGRAMMABILITY FOR TRANSPORT FLEXIBILITY

In AI networking, programmability is the direct enabler of transport flexibility. It allows the network infrastructure to dynamically adapt to the volatile, high-throughput demands of distributed AI training and inference using standards-based Ethernet protocols such as **UEC-ready RDMA** (Ultra Ethernet Consortium-Ready Remote Direct Memory Access) and now MRC (Multipath Reliable Connection). Released by OpenAI in May 2026 through the **Open Compute Project** (OCP) and co-developed with industry leaders, including AMD,² MRC is a novel protocol that improves GPU networking performance and resilience in large training clusters. MRC is supported by both the AMD Pensando™ Pollara 400 AI NIC and the AMD Pensando™ Vulcano 800 AI NIC.



KEY BENEFITS OF PROGRAMMABILITY

DYNAMIC WORKLOAD ADAPTATION

Programmable silicon allows the network to adapt its transport behavior and congestion control in real time, preventing performance bottlenecks when AI models, especially LLMs, create rapidly fluctuating east-west (server-to-server) traffic.

FUTURE READINESS

AI protocols and communication patterns evolve faster than hardware cycles. A programmable network can be updated with new software-defined protocols (like specialized versions of RDMA) without requiring the physical replacement of expensive switches.

PROACTIVE MANAGEMENT AND SELF-HEALING

While traditional networks are often reactive, programmable networks can think ahead to maintain uptime. In addition to enabling predictive maintenance, this allows networks to self-heal and adapt to AI traffic demands by separating control and data planes.

SCALABILITY AND EFFICIENCY

Programmable networks can burst to handle data spikes and provide bandwidth on demand—essential for distributed AI workloads.

IN-NETWORK PROCESSING

Technologies like P4 architecture allow AI-specific tasks, such as anomaly detection or even simple machine-learning routines, to run directly on AI NICs. This reduces latency by processing data at the edge before it even reaches a CPU or GPU.

INFRASTRUCTURE AS CODE (IAC)

Programmability enables teams to treat network configuration like software. By using APIs instead of manual command-line entries (CLI), entire AI environments can be deployed and scaled with automated troubleshooting.

IMPROVED SECURITY AND COMPLIANCE

Programmability enables real-time threat detection by leveraging machine learning to analyze network behavior and detect anomalies, such as sudden data exfiltration, instantly quarantining affected segments. Centralized control also allows consistent enforcement of security policies across the entire network, reducing the risk of human error during manual configuration.

The best NIC solutions also leverage technologies like P4 architecture, allowing teams to customize network behavior, implement bespoke RDMA transports, and optimize performance for specific AI workloads—all while maintaining compatibility with future industry standards, such as those set by the UEC and OCP. Programmable infrastructure also ensures future and backward compatibility across generations, so organizations can innovate without the cost or disruption of rip-and-replace cycles.

OPERATIONAL EXCELLENCE

As AI operates at production scale, operational rigor is critical to support long-running training jobs and persistent, real-time AI workloads. Operational excellence ensures networks stay performant, consistent, resilient, observable, and upgradable while minimizing impact.

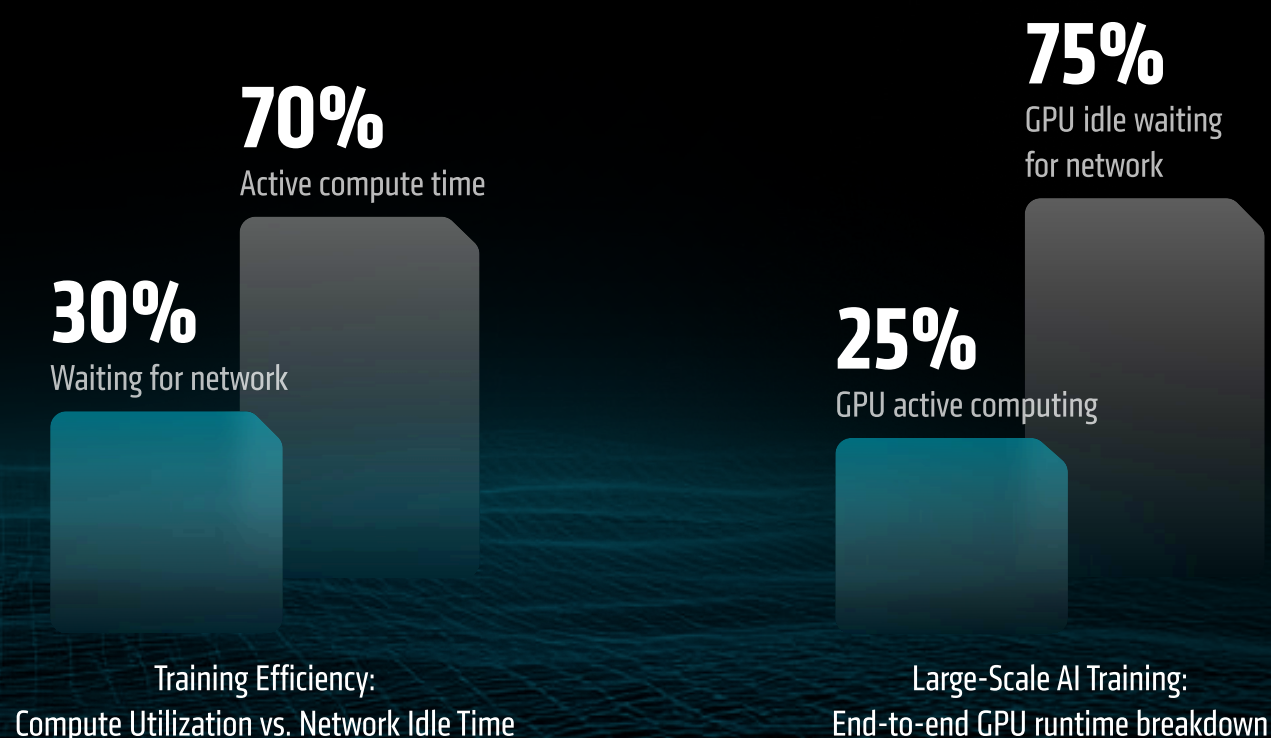
WHY NETWORK OPERATIONAL EXCELLENCE MATTERS FOR AI

The network is no longer just a way to connect devices—it's the backplane of the computer. When thousands of GPUs are working on a single problem, the network's performance directly dictates how fast and how cost-effectively that model learns.

Networking is now a first-order determinant of AI system performance, cost, and reliability. Operational excellence addresses two key realities of AI infrastructure.

COMMUNICATION OVERHEAD

Modern AI workloads, especially large-scale LLM training and distributed inference, are often dominated by communication overhead—time and computing resources spent moving data between nodes instead of performing actual calculations.



THE “WAIT-STATE” PROBLEM

Because distributed AI training depends on GPU synchronization, if the network is poorly managed and a single packet is delayed by a messy burst, GPUs in the same cluster as the data transfer can sit idle waiting for that piece of data. That means paying for world-class GPUs to do nothing while waiting on the network, leading to serious issues with reliability and failures.

Any downtime or degraded performance is exponentially more expensive than in traditional data centers, with even minutes of disruption costing thousands of dollars per hour as GPUs sit idle, training jobs stall, and SLAs slip. Worse, an AI network glitch can cause a checkpoint failure, potentially losing hours or days of training progress.

REDEFINING OPERATIONAL EXCELLENCE

These operational challenges can be addressed across four key pillars.

And because AI clusters scale horizontally (tens of thousands of GPUs, multiple fabrics, and containerized workflows), network operational challenges multiply, including:

- Limited observability into RDMA traffic, path usage, congestion, and link health.
- Configuration drift across hosts, NICs, GPU nodes, and software builds.
- Risky network changes (firmware updates, reconfigs) that can stall running jobs.
- Difficulty correlating degraded jobs to flows, queues, or entropy paths.

MAXIMIZE CLUSTER UPTIME WITH NONDISRUPTIVE UPDATES

Traditionally, network upgrades have required a maintenance window where traffic is stopped or rerouted, causing downtime. Nondisruptive alternatives include:

- Hitless upgrades (no downtime): This software deployment method, enabled by programmable hardware, allows network devices like switches and routers to update their firmware or OS without dropping a single data packet.
- Graceful upgrades (less than one-minute downtime): This cuts maintenance windows from tens of minutes to seconds-scale.

Both types of upgrades result in greater GPU utilization, fewer maintenance windows, and higher cluster uptime and ROI.

CONSISTENCY VALIDATION: PREVENTING FAILURES BEFORE THEY HAPPEN

Cluster-wide configuration assurance ensures consistency across the network, hosts, and GPUs. By eliminating silent errors and configuration drift, this automated approach prevents performance degradation and significantly reduces operational risk and support overhead.

ENHANCED OBSERVABILITY: DEEP, HIGH-RESOLUTION TELEMETRY

Visibility into AI networking with multi-layered counters, precision telemetry, low-CPU-overhead streaming, and programmable RDMA traffic sniffing provides:

- Faster mean time to detect (MTTD).
- Faster mean time to repair (MTTR).
- Ability to root-cause congestion, drops, path imbalance, and protocol faults.

FASTER TIME TO PRODUCTION

By streamlining Kubernetes integration for AI NICs—including automated driver updates, resource labeling, and built-in consistency checks—nodes can become AI-ready the moment they join a cluster. This minimizes manual configuration and significantly accelerates both cluster setup and workload deployment.

As AI operates at production scale, compute performance isn't enough. Technology leaders and networking teams need to ensure networks remain performant, consistent, observable, and upgradable without impact. Operational excellence in networking is no longer optional—it's infrastructure-critical.

OPENNESS PRESERVES ARCHITECTURAL OPTIONALITY

As AI systems evolve—spanning new accelerators, model architectures, and deployment patterns—openness enables change without architectural resets. Openness prevents proprietary lock-in and enables operators to adopt best-of-breed silicon, software, and network fabrics without re-platforming the entire stack.



COST EFFICIENCY

Reduced capital expenditures through competitive component pricing



SUPPLIER CHOICE

Freedom to select best-in-class components



DIGITAL SOVEREIGNTY

Enterprises and nations gain more control over data location

WHY AN OPEN ECOSYSTEM?

In today's dynamically evolving AI landscape, freedom from the constraints of proprietary technologies and vendor lock-in can give organizations competitive edge. An open ecosystem allows hardware and software from different vendors to work together using shared standards, with benefits such as:

- Increased cost efficiency.
- Freedom to choose best-in-class components from multiple suppliers.
- Accelerated innovation.
- Greater digital sovereignty as organizations and nations gain more control over their data location and critical infrastructure.

WHAT CAUSES AI INFRASTRUCTURE LOCK-IN?

AI infrastructure lock-in stems from a combination of high-cost physical dependencies, proprietary software frameworks, and deep operational integration, making switching providers technically and financially prohibitive. Freedom from infrastructure lock-in is a tenet of the Ultra Ethernet Consortium (UEC) mission: “the development of an Ethernet-based, open, interoperable, high-performance, full communications stack architecture to meet the growing network demands of AI and HPC (High-Performance Computing) at scale.”³ AMD spearheaded the founding of the UEC in 2023.

Similarly—though not a standards body—the Open Compute Project (OCP) networking standards promote open, disaggregated, and vendor-neutral hardware and software for data centers.

WHY ETHERNET?

Ethernet has been used in networking for over 50 years and is still the networking technology of choice in AI clusters today, largely due to its ubiquity, familiarity among networking teams, and cost-effectiveness.

Despite its many advantages, however, conventional Ethernet does not prioritize the high bandwidth, low latency, and efficient communication patterns demanded by modern AI workloads. Open, standards-based Ethernet provides the predictable performance and massive throughput necessary to manage high-density AI workloads at scale.



LOSSY VS. LOSSLESS ETHERNET

AI networks are notoriously unforgiving. They have none of the tolerance for dropped or lost data packets that characterizes the “lossy” Ethernet protocol. Scale-out networks for AI infrastructure typically employ one of two networking technologies:

- InfiniBand is designed for low-latency, high-throughput communication in large AI clusters.
- “Lossless” Ethernet eliminates packet loss under heavy traffic, thanks to the underlying routing protocol—the RoCEv2 Underlay (RDMA over Converged Ethernet).

While InfiniBand is an open-standard networking technology, it's significantly more expensive than standard networking gear, and the InfiniBand ecosystem is tightly controlled by a few players. To counter this closed ecosystem, one of the UEC's goals has been to establish lossless Ethernet as a viable alternative to InfiniBand. And indeed, Ethernet-based solutions are emerging as the standard for on-premises AI infrastructure buildouts.⁴



MORE INNOVATION, LESS RIP-AND-REPLACE

Most AI networking narratives today ignore existing infrastructure realities, offering disruption through rip-and-replace instead of credible transition paths. An added benefit of preserving architectural optionality is that it helps ensure AI infrastructures can evolve as quickly as the AI technologies they support, fostering innovation without the cost or disruption of rip-and-replace cycles.



BRIDGING ETHERNET AND AI WORKLOADS

One of the benefits of AMD AI NICs is that they serve as a bridge between Ethernet's broad compatibility and the specialized demands of AI workloads, allowing customers to choose their preferred Ethernet vendor and implementation type, whether lossless or lossy. What's more, eliminating the need for a lossless network vastly reduces the complexities of configuring and operating networks, thereby enhancing AI workload performance and network reliability.



GENAI AND AGENTIC AI DRIVING SHIFT TO SCALE-OUT ETHERNET

In addition, because of the massive scale they require, the GenAI and Agentic AI trends are driving an industry-wide move toward highly scalable Ethernet solutions. This massive shift is powered by open initiatives such as the UEC and OCP, allowing companies to build interconnected clusters of over 100,000 GPUs using next-generation Ethernet protocols (such as UEC-ready RDMA) without getting locked into a single hardware vendor.



ARCHITECTED FOR AI GROWTH

Infrastructure scales linearly, ensuring costs remain aligned with cluster size and model complexity, whether organizations are scaling up capacity or scaling out footprint. Future-ready architectures support forward-compatible hardware, evolving interconnects, and new AI workloads without forcing disruptive redesigns.

WHY DO AI NETWORKING COSTS INCREASE AT SCALE?

As AI networks scale, costs rise due to the exponential increase in data transfer requirements between thousands of GPUs, leading to high-bandwidth, low-latency infrastructure demands. As models grow, they require massive data movement across parallel systems, driving up power consumption, specialized hardware expenses, and inter-data center communication costs.

Aligning cost with cluster growth, utilization, and model complexity over time is complicated by the fact that AI workloads evolve faster than infrastructure, making an already moving target even more elusive.

CHOOSING NETWORK INFRASTRUCTURE

Choosing networking infrastructure is one of the most consequential strategic decisions a technology leader can make, emphasizing the need for cost-efficiency, long-term flexibility, and adaptability in the face of future AI growth and changing AI workloads. Amid these future uncertainties, this constant remains: Costs will increase as AI networking scales.



SLASHING CAPEX

Capital expenditures on network infrastructure can be reduced by up to 58% by leveraging capabilities typically found in expensive deep buffer switches, but now available in top NIC solutions, thereby drastically reducing the number of switches needed. These capabilities include Adaptive Packet Spray and Out-of-Order and In-Order message delivery, which help avoid bottlenecks and ensure full use of available bandwidth.

Solutions like the AMD AI NICs also support multiplane fabric architecture—which conveys transport traffic along multiple, independent, and parallel network planes or fabrics—enabling high-availability failover with better fault isolation. This significantly improves networking RAS (reliability, availability, and serviceability) and supports the cost-effective scale out of AI networks.

HOW DO AI WORKLOADS EVOLVE FASTER THAN INFRASTRUCTURE?

AI workloads evolve faster than infrastructure due to an exponential mismatch. While hardware capabilities follow steady, predictable cycles like Moore's Law (doubling roughly every two years), AI compute demand has historically doubled every six months.⁵ Making AI networking investments future-ready, therefore, becomes paramount.

FUTURE READINESS IN AI NETWORKING INVESTMENTS

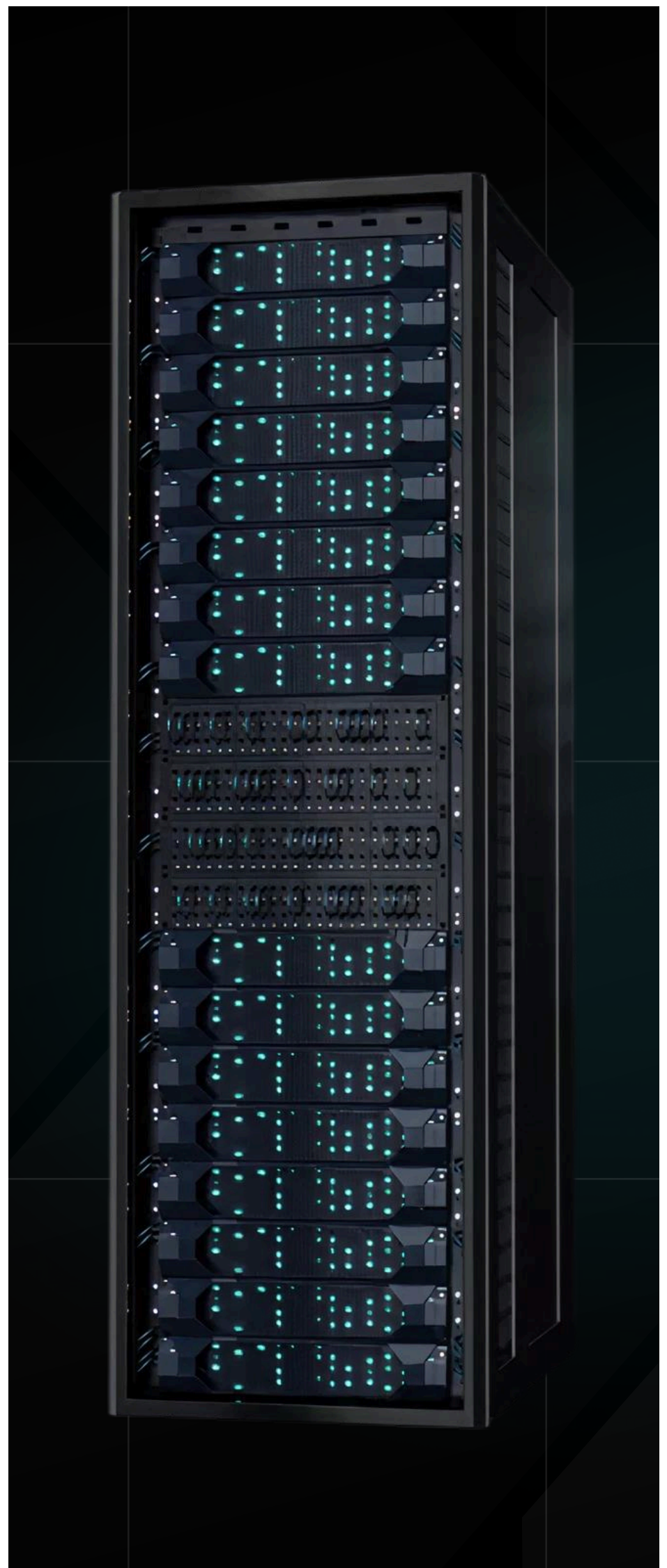
Building future readiness into AI infrastructure investments requires a shift from a static hardware-centric mindset to a flexible, software-defined ecosystem that prioritizes data quality, interconnection, and scalable power.

Instead of buying larger individual boxes, technology leaders should adopt scale-out interconnected architecture. This distributes workloads across many interconnected nodes, allowing organizations to increase capacity linearly as AI model sizes grow.

DATA BEFORE ENGINE

Since AI performance is often limited by the slowest component in the system, prioritizing data and infrastructure readiness is key to aligning cost with growth while staying future-ready. Following the “data before engine” maxim, leaders and networking teams should ensure they have clean AI-ready data pipelines before over-investing in massive GPU clusters.

Following emerging standards like those published by the UEC and OCP to prevent vendor lock-in, and leveraging programmable network infrastructure that ensures future compatibility across generations, enables innovation without the cost or disruption of architectural resets.

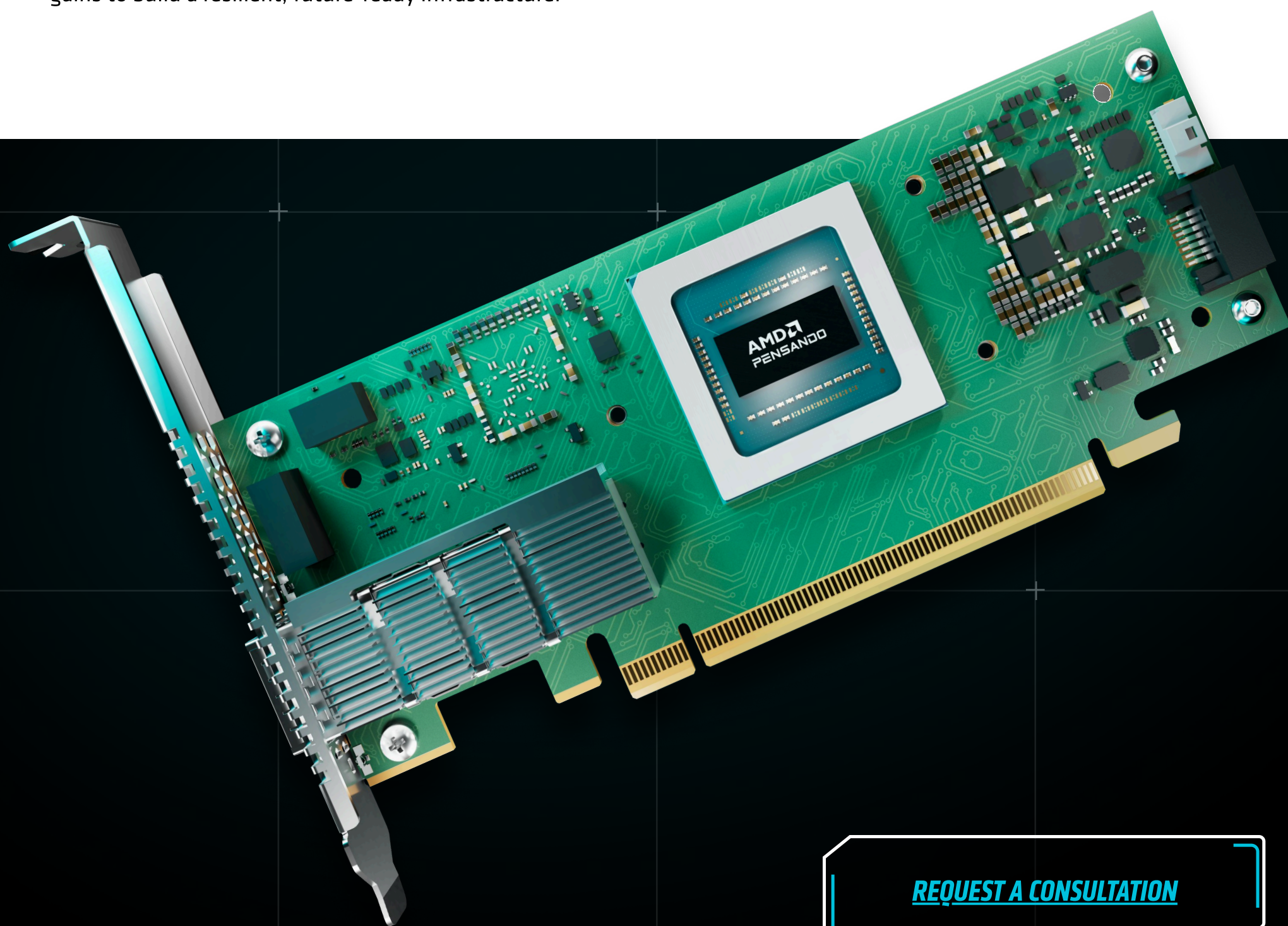


DEFINING THE NEXT ERA OF AI NETWORKING

As AI transitions from experimentation to production, networking can no longer be a technical afterthought—it's the defining strategic foundation for the modern data center. The architectural choices made today dictate long-term economics, vendor flexibility, and scalability.

To avoid the structural dependencies of closed systems, leaders must prioritize open, standards-based Ethernet fabrics. By embracing these networking principles, organizations ensure the performance and control necessary to meet the demands of AI at scale, moving beyond early gains to build a resilient, future-ready infrastructure.

Don't let yesterday's architecture bottleneck tomorrow's AI. Transition to an open Ethernet foundation now and claim your competitive advantage with **AMD AI networking solutions.**



REQUEST A CONSULTATION

Sources:

1. Gartner Insights: Gartner, “2026 Trends in Networking for GenAI,” Naresh Singh, Rithvik Kumar, Menglin Cao, October 2025.
<https://www.gartner.com/en/documents/7120630>
2. “Supercomputer networking to accelerate large scale AI training.” OpenAI, 5 May 2026,
<https://openai.com/index/mrc-supercomputer-networking/>
3. Ultra Ethernet Consortium
<https://ultraethernet.org>
4. Enterprise Strategy Group Research Report, **The Critical Role of NIC Programmability in Scaling Out Data Center Networks for AI**, June 2025.
5. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos, "Compute Trends Across Three Eras of Machine Learning," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9891914