AI Security

Congestion Control and Telemetry Security in AI Fabrics: Architectural Synthesis

Congestion Control and Telemetry Security in AI Fabrics: Architectural Synthesis

● **Author:** Scott Thornton, perfecXion.ai ● **Published:** January 25, 2026 ● **Read Time:** 10 minutes

https://perfecxion.ai

Table of Contents

Executive Summary

Networks matter more than ever. Your multi-million-dollar AI training cluster faces a critical bottleneck—not compute, but the network fabric itself.

Performance vs Security Trade-off

AI fabric congestion control mechanisms optimized for performance can become attack vectors. Balancing throughput with security monitoring is critical for production deployments.

Traditional data centers excel with diverse microservices traffic—web requests, database queries, file transfers, all smoothed by statistical multiplexing. But AI workloads shatter these assumptions completely.

Think about it. AI training doesn't follow web patterns. Thousands of GPUs synchronize in massive, coordinated bursts, creating avalanches that overwhelm conventional networks instantly. The math is unforgiving.

Critical Reality: This paradigm shift changed everything, forcing the industry to build something entirely new: specialized AI fabrics running high-performance RDMA technologies like InfiniBand, which delivers proven results, and RoCEv2, which promises Ethernet scale with microsecond latencies and zero packet loss.

But promises come with prices. Complex new challenges emerge overnight, creating attack surfaces worth billions.

Consider this. Criminals stole $25.6 million from Arup using deepfakes in early 2024, and similar attacks now target telemetry systems controlling AI fabrics, with latest threat intelligence projecting $25-40 billion in annual losses by 2027. Your network infrastructure sits in the crosshairs.

This synthesis examines modern AI fabric architecture—why traditional networking fails spectacularly for AI workloads, how congestion control and telemetry interdependence creates critical security vulnerabilities, and what evolution from crude link-layer controls to sophisticated AI-driven algorithms means for your infrastructure.

The analysis reveals fundamental architectural tension. RDMA protocols demand lossless networks, yet Ethernet is inherently lossy by design. Bridging this gap requires complex engineering that creates cascading vulnerabilities.

Telemetry poisoning attacks emerge as the silent killer. Performance gets manipulated without dropping packets. Billion-dollar training runs get crippled by invisible attackers. The stakes couldn't be higher.

Industry Divergence: Major industry players chose dramatically different paths—Google and AWS built proprietary custom transport protocols, while NVIDIA, Broadcom, and Meta push open standards to their limits as the Ultra Ethernet Consortium creates multi-vendor alternatives.

This report provides forward-looking perspectives on a fundamental question: will vertically integrated proprietary solutions dominate, or will open ecosystems prevail? The answer shapes AI infrastructure's next decade and determines whether your networks become assets or attack surfaces.

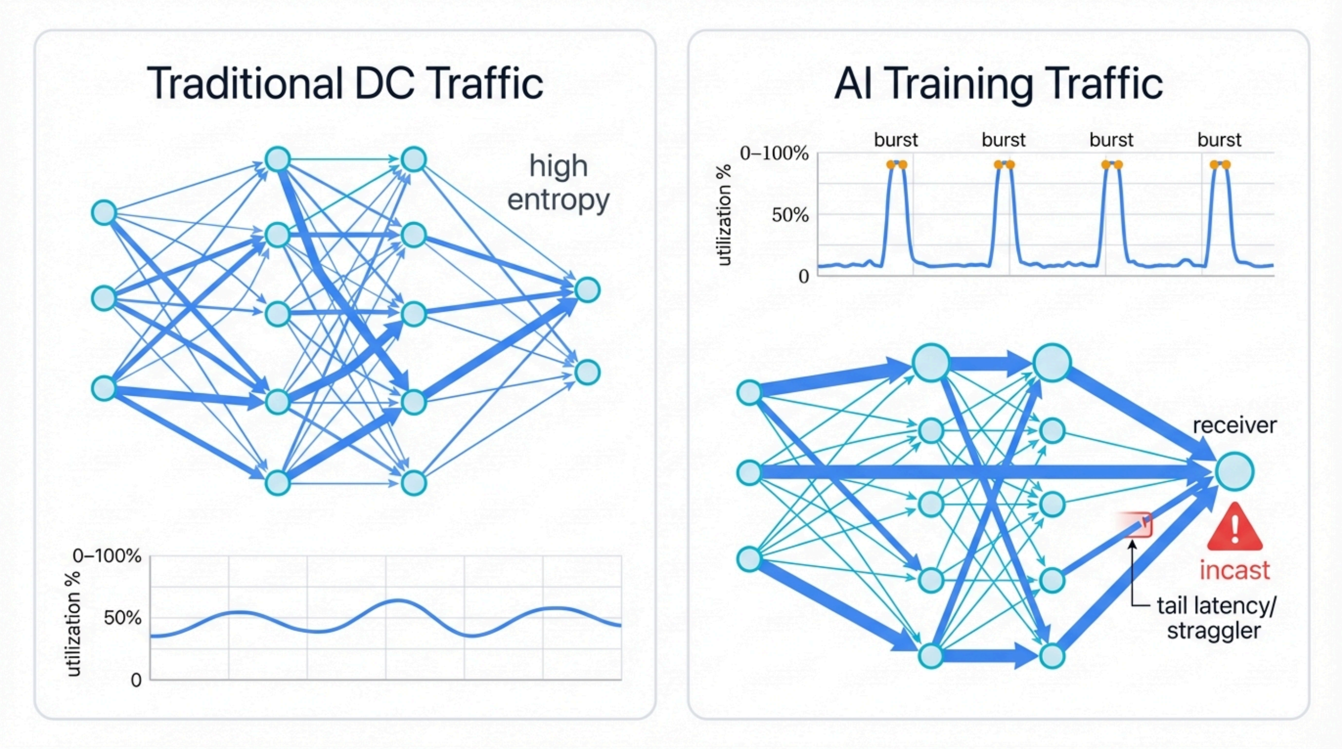
Part I: The Foundational Imperative for Specialized AI Networks

Everything changed with AI. Not evolution—revolution.

Large-scale distributed AI training breaks traditional networks completely, with communication patterns that differ radically from conventional workloads and fundamental networking assumptions that crumble under pressure. The math doesn't lie.

The Unique Traffic Profile of AI Workloads

Investment scale matters here. Your distributed AI cluster represents hundreds of millions in hardware, and performance depends entirely on network efficiency that follows patterns unlike anything traditional data centers handle.



AI Workload Traffic Profile vs Traditional

Analysis of AI Traffic Patterns

Forget typical data center traffic patterns. Traditional networks handle diverse flow mixes beautifully—short "mice" flows coexist with longer "elephant" flows, statistical multiplexing smooths the rough edges, and load balancing works because flows are random and varied.

AI training destroys this playbook completely.

Low-Entropy, High-Bandwidth Flows: AI cluster communication focuses on specific patterns—relatively few, long-lived flows carry massive bandwidth between GPU groups as distributed training demands synchronized exchanges where AllReduce operations require model parameters and gradients from thousands of accelerators.

These flows differ from statistical assumptions by design. They're deterministic. Modern 400/800 Gbps links saturate easily for sustained periods.

Low-entropy flows share similar source-destination patterns, and when traditional load balancing algorithms expect high-entropy header variations, they fail spectacularly.

Periodic, Bursty Nature: Traffic timing creates unique challenges through intense periodic bursts that align with training algorithm cycles—GPUs perform local computation first, then enter communication phases for network-wide gradient exchange and aggregation.

Physics Problem: This creates severe "incast" scenarios instantly where multitudes of senders simultaneously transmit large data volumes to small sets of receivers or aggregation nodes that get overwhelmed because the physics can't be ignored.

Synchronized burstiness drives instant network congestion. Switch buffer capacity gets overwhelmed in microseconds. Traditional congestion control assumes gradual buildup and can't react fast enough.

Job Completion Time as Key Metric: AI workloads differ dramatically from web services here because tightly coupled synchronous training jobs create dependencies where entire clusters wait for the slowest communication links before proceeding to next computation steps.

The "straggler" problem dominates performance. Your cluster's speed gets dictated by network tail latency —average performance doesn't matter when single flows experiencing delay, jitter, or packet loss create cascading effects that leave thousands of expensive GPUs sitting idle while the network catches up.

The Inadequacy of Traditional TCP/IP

Traditional approaches collapse under AI pressure. Standard TCP/IP stacks fail spectacularly with AI traffic profiles because RFC 5681 standard TCP congestion control algorithms assumed different conditions, designed for the lossy, best-effort internet rather than AI fabric microsecond-scale dynamics that demand faster responses.

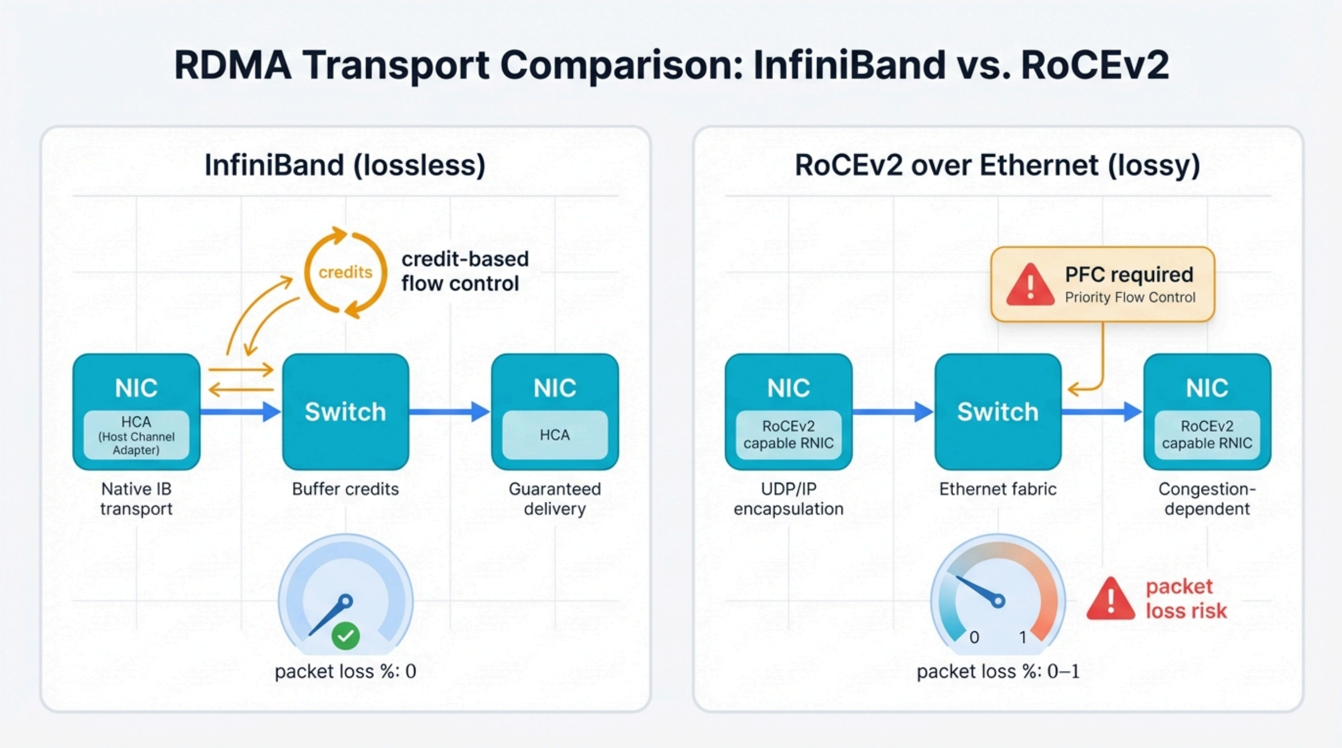
ECMP routing creates bigger problems. Equal Cost Multi-Path routing proves highly ineffective for AI workloads because ECMP uses stateless packet header hashes for distribution where source and destination IPs and ports determine path selection—a technique that works beautifully for high-entropy web services where every flow looks different to the hash function.

AI training generates different patterns. Low-entropy flows have limited header variability, causing hash algorithms to map many large, high-bandwidth flows onto identical physical paths. This creates "hash polarization" or "flow collision" phenomenon that results in severe link oversubscription while parallel fabric paths remain underutilized.

Cascading Impact: Load imbalance creates cascading problems where increased tail latency follows directly, GPU cycles get wasted, and your entire cluster efficiency suffers because the network can't distribute load properly.

The Rise of RDMA-based Interconnects

AI workloads demand extreme performance. Stringent latency and bandwidth requirements forced network designers to circumvent traditional bottlenecks where host operating system kernels and CPUs became the enemy, driving widespread Remote Direct Memory Access adoption as foundational transport technology for AI fabrics.



RDMA over InfiniBand vs RoCEv2 Losslessness

The Need for Kernel Bypass

Traditional network stacks create unacceptable overhead. Conventional approaches require host CPU processing for every packet, adding delays through kernel networking layers, consuming time with multiple memory copies, and compounding the problem with context switches that result in tens of microseconds latency—an eternity for tightly-coupled AI applications.

RDMA solves this through "kernel bypass" innovation where one machine's Network Interface Card directly reads from remote machine memory, writes directly to remote memory without involving remote CPUs, and bypasses operating systems completely. This direct memory-to-memory transfer reduces end-to-end latency to microseconds, which is critical for efficient distributed training.

The Lossless Requirement

RDMA has non-negotiable demands. Performance models assume the underlying network never drops packets because while RDMA protocols include recovery mechanisms by design, they're designed for exceptional errors—not routine Ethernet congestion-based packet loss.

Packet loss creates cascading problems. A single dropped packet triggers slow timeout-based recovery that stalls communication for milliseconds, and these events devastate synchronous AI job performance where every microsecond matters.

Fundamental Requirement: This fundamental RDMA characteristic creates strict requirements where underlying networks must be "lossless fabrics" with no exceptions allowed, and this single requirement drives modern AI networking's immense complexity.

Comparative Analysis of Dominant RDMA Technologies

Two primary technologies emerged for high-performance RDMA delivery, each with distinct architectural philosophies that shape how you'll deploy and manage your AI fabric.

InfiniBand: InfiniBand represents complete architectural thinking—end-to-end network design governed by the InfiniBand Trade Association, designed from ground up for high-performance computing, and inherently lossless by design.

InfiniBand achieves losslessness through credit-based link-layer flow control where devices won't transmit packets without downstream buffer confirmation, ensuring they know downstream devices have available receive buffer space first. This proactive approach prevents buffer overruns by design, eliminates packet drops, and makes reactive loss-prevention mechanisms unnecessary.

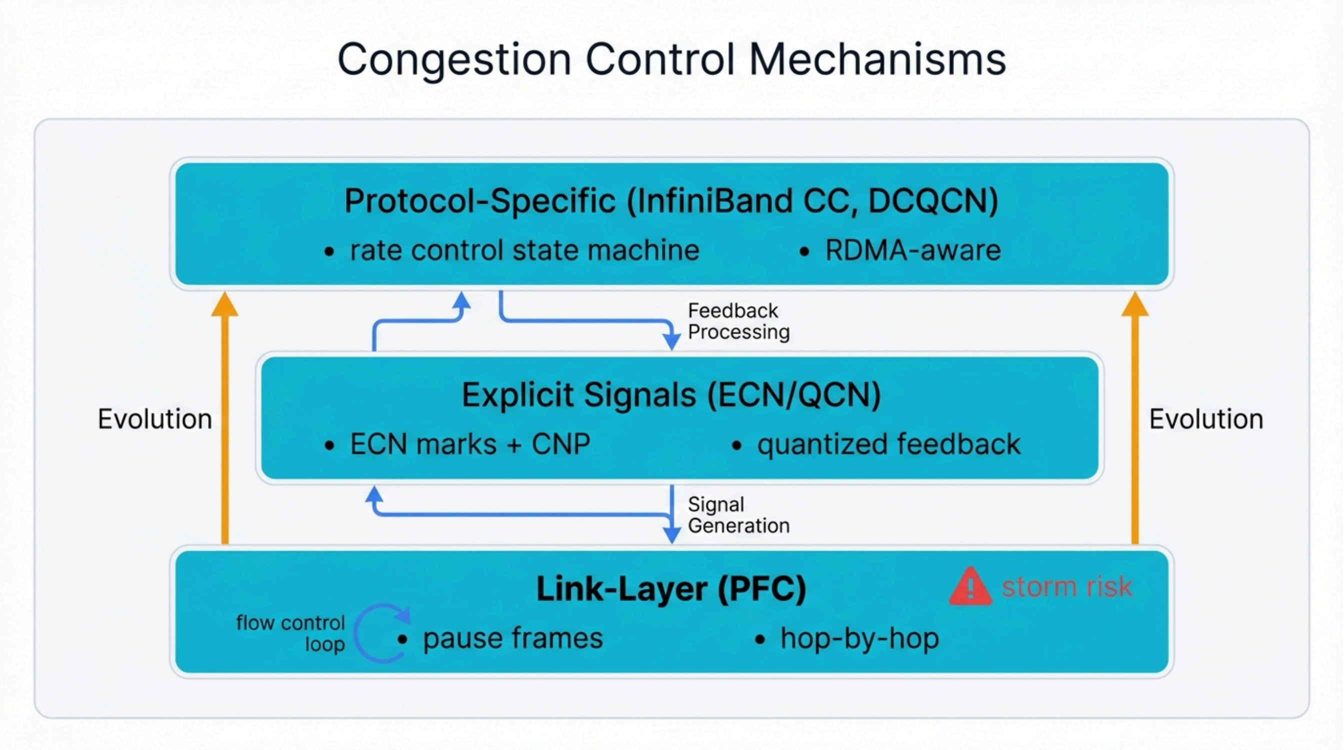
RDMA over Converged Ethernet: RoCEv2 takes a different approach by running InfiniBand transport protocols over standard Layer 3 Ethernet and IP networks, promising massive Ethernet scale with ecosystem leverage and operational familiarity that helps adoption.

Standard Ethernet creates fundamental challenges. It's inherently lossy and best-effort by design, making RoCEv2 non-natively lossless while RDMA's requirements demand auxiliary mechanisms where RoCEv2 relies on Priority-Based Flow Control to prevent congestion-related packet drops.

Architectural Mismatch: Fundamental architectural mismatch emerges here when bolting lossless protocols like RDMA onto lossy fabrics like Ethernet creates problems, and this mismatch is the root cause of many significant challenges in modern AI networking.

Part II: A Deep Dive into Congestion Control Mechanisms

One challenge dominates RDMA-based AI fabrics: managing congestion without dropping packets. This seemingly simple requirement created complex realities where multiple layers of evolving control mechanisms emerged, ranging from brute-force link-layer protocols to highly sophisticated end-to-end algorithms.



Congestion Control Stack: PFC → ECN/QCN → DCQCN/IBCC

Foundational Link-Layer and End-to-End Controls

First-generation solutions took direct approaches. Lossless Ethernet relied heavily on link-layer mechanisms with simple end-to-end signaling providing basic control, and these approaches became foundational to everything that followed, but significant limitations emerged that expose the difficulties of retrofitting losslessness onto Ethernet.

Priority-Based Flow Control (PFC / IEEE 802.1Qbb)

PFC creates "no-drop" Ethernet services. It operates hop-by-hop for RoCEv2 traffic where switch egress buffers monitor specific traffic classes, and when they exceed pre-configured thresholds, switches send IEEE 802.1Qbb PAUSE frames to upstream neighbors that instruct switches or NICs to stop transmitting specific priority traffic for short durations, preventing buffer overflow and avoiding packet drops.

PFC successfully prevents packet loss. But it's a blunt instrument with severe side effects, and the most critical problem is congestion spreading.

Single link congestion points trigger PAUSE frames that cause upstream switch buffers to fill, which triggers further upstream PAUSE frames, creating "PFC storms" from this cascading effect where waves of PAUSE frames propagate backward through the network, freezing traffic on many links and affecting even flows not destined to original congestion points.

2024 Security Research: New vulnerabilities emerged in 2024 research when the ReDMArk study revealed critical flaws in PFC implementations where attackers can exploit RDMA packet injection vulnerabilities to trigger PFC storms deliberately, creating distributed denial-of-service attacks that bring down entire AI training clusters with organizations reporting up to 73% performance degradation during synthetic congestion attacks.

Explicit Congestion Notification (ECN / IETF RFC 3168)

ECN provides more graceful alternatives. Instead of PFC's brute-force pausing approaches, ECN-enabled switches signal incipient congestion early without waiting for near-full buffers by monitoring switch queue depths for particular flows, and when they exceed lower "ECN marking" thresholds, switches set Congestion Experienced bits in traversing packet IP headers without dropping packets.

Marked packets continue to destinations where receiving NICs see CE bits and generate and send special Congestion Notification Packets back to original senders, and senders receiving CNPs understand path congestion and reduce transmission rates accordingly.

Closed-loop feedback systems emerge from this approach where endpoints react to congestion before severity causes drops or triggers PFC, representing significant sophistication over link-layer pause mechanisms.

Quantized Congestion Notification (QCN / IEEE 802.1Qau)

QCN, standardized in IEEE 802.1Qau, builds on the explicit notification concept but operates in Layer 2 bridged domains. Instead of ECN's simple binary congestion signals, QCN provides nuanced feedback.

When bridges detect congestion, they send Congestion Notification Messages back to source hosts that crucially contain quantized values—typically 6-bit fields—indicating congestion severity, enabling senders to make proportional rate adjustments with small reductions for light congestion and larger reductions for severe congestion.

Protocol-Specific Congestion Management

Building on explicit notification's foundational concepts, the InfiniBand and RoCEv2 ecosystems developed protocol-specific standardized congestion control algorithms tailored to their unique characteristics and requirements.

InfiniBand Congestion Control

The InfiniBand architecture specifications include detailed Congestion Control Algorithms that define complete end-to-end management frameworks, conceptually similar to ECN but native to InfiniBand protocols.

Mechanism: When InfiniBand switches detect port congestion through buffer utilization crossing configurable thresholds, they begin marking Forward Explicit Congestion Notification bits in the headers of packets contributing to congestion, and these packets travel to their final destinations.

Destination Host Channel Adapters receiving FECN-set packets notify sources by sending packets back with Backward Explicit Congestion Notification bits set, and source HCAs receiving BECN notifications consult their Congestion Control Tables and increase packet injection delays, thereby throttling transmission rates.

Evolution: July 2023's IBTA Volume 1 Release 1.7 introduced in-band Round-Trip Time measurement primitives that represent significant developments providing standardized hooks for implementing sophisticated algorithms like Timely, HPCC, and Google's Swift directly within InfiniBand ecosystems.

RoCEv2 Congestion Control - DCQCN

DCQCN dominates RoCEv2 networks. Data Center Quantized Congestion Notification represents the de facto standard algorithm—a highly sophisticated system running in NICs that works in conjunction with switch ECN marking.

DCQCN has explicit goals: keep switch queues short and stable while preventing PFC activation. It operates through sender NIC multi-stage rate control state machines where NICs start by rapidly increasing transmission rates, then enter rate reduction phases with multiplicative decreases when CNPs arrive from congested switches' ECN marks, followed by recovery phases with slow additive rate increases.

Additive Increase/Multiplicative Decrease behavior enables convergence where careful timing and parameterization allow large numbers of flows to converge on fair bandwidth shares while maintaining low queue depths simultaneously.

Security Concerns: Security research revealed additional concerns in 2024 when attackers exploit DCQCN's parameter sensitivity through telemetry poisoning attacks that manipulate CNPs to force unnecessary rate reductions, achieving up to 67% performance degradation in controlled environments while remaining virtually undetectable to standard monitoring systems.

Innovations from Academia and Industry Research

Standard protocol limitations when facing AI workloads' punishing demands spurred waves of innovation from academic researchers and hyperscale companies' advanced engineering teams, with efforts focused on three key areas: moving intelligence to hosts, developing more sophisticated control signals, and leveraging programmable hardware capabilities.

Host-Based Load Balancing

Recognizing that static hash-based ECMP is a primary cause of congestion in AI workloads, several key research efforts developed dynamic load balancing implemented at endpoints rather than in the network core.

Hopper: Nosrati and Ghaderi proposed a host-based load balancing system specifically for RDMA in AI clusters that operates entirely in end-host software and requires no special switch hardware, continuously monitoring current network paths for signs of congestion, particularly increasing latency.

Upon detecting congestion, Hopper dynamically and carefully switches traffic to alternative, less-congested paths, with evaluations showing this dynamic path selection approach reducing average flow completion times by 20% versus state-of-the-art host-based methods.

Protective Load Balancing: Google developed another elegant host-based design deployed in production data centers where hosts detecting connection congestion randomly reroute flows to different network paths.

Advanced Control Algorithms

HierCC: Some proposals argue that traffic uncertainty from many short-lived flows creates fundamental challenges for RDMA networks, and to mitigate this, HierCC introduces hierarchical control.

The system aggregates individual flows destined for the same rack into single, stable macro-flows where inter-rack aggregate flow rates are managed using proactive credit-based mechanisms, and the aggregate flow's allocated bandwidth is then distributed among individual flows within the source rack.

AI-Driven Customization: Cutting-edge work explores using machine learning to create adaptive congestion control systems where PCC Vivace employs online-learning protocols that automatically customize congestion control logic based on high-level requirements—such as optimizing for latency versus throughput—and observed network conditions.

Programmable NIC-based Approaches

Perhaps the most significant recent trend involves shifting congestion control execution onto powerful programmable NICs—Data Processing Units or SuperNICs—that contain multi-core processors and hardware accelerators, providing ideal platforms for executing complex line-rate algorithms without burdening host systems.

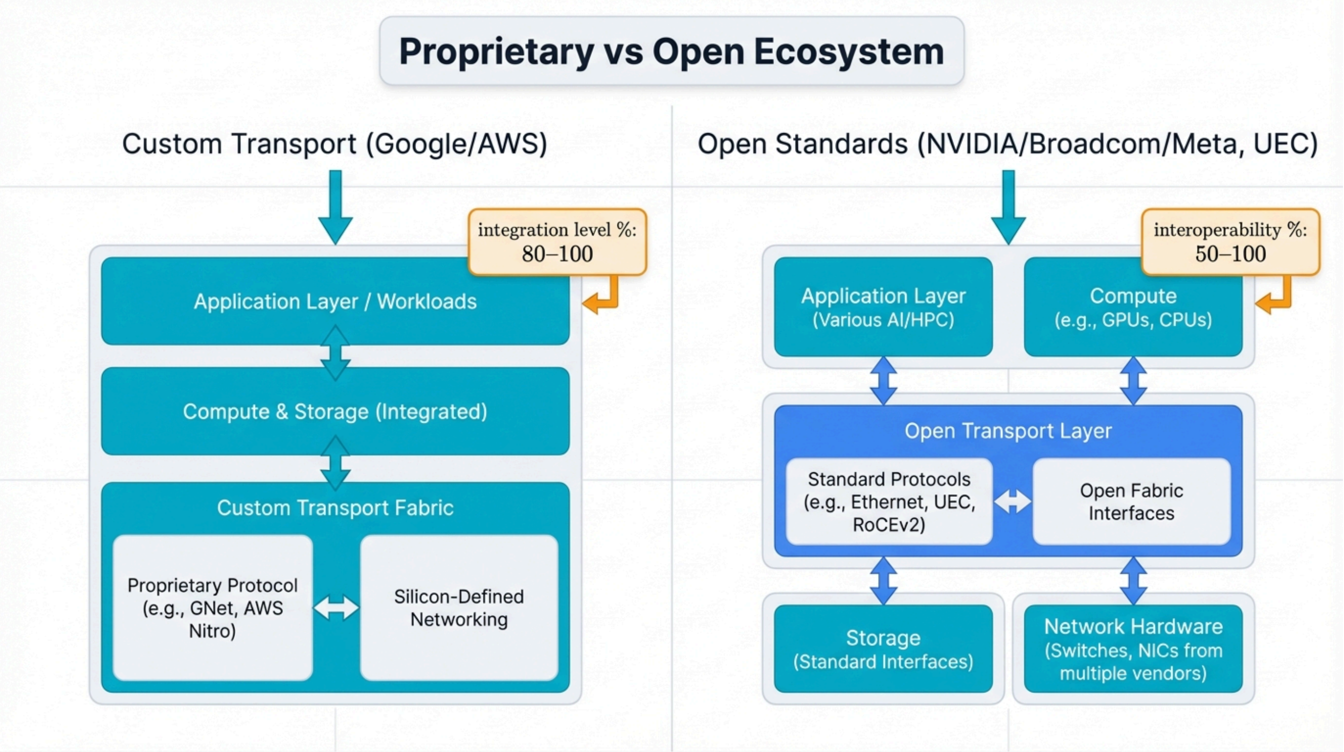
Barre: Alibaba developed and deployed a practical programmable NIC implementation for their 400 Gbps clusters by leveraging NVIDIA BlueField-3 SuperNIC capabilities to implement simple yet highly effective event-based congestion control.

Production Success: The production deployment supporting over 10,000 GPUs improved AI training throughput by an average of 9.6% versus previous solutions and more importantly demonstrated the viability of programmable approaches at hyperscale.

The rise of NVIDIA's Morpheus AI framework for real-time threat detection in 2024 further accelerates this trend by integrating security monitoring directly into programmable NICs, enabling organizations to simultaneously optimize performance and detect the telemetry manipulation attacks that increasingly target AI infrastructure.

Part III: Architectures of Major Industry Stakeholders

Theoretical concepts and individual congestion control technologies ultimately realize themselves within cohesive hardware and software architectures that industry leaders develop and deploy, and the strategic choices made by silicon vendors, systems companies, and hyperscale providers reveal distinct philosophies for solving AI networking challenges.


Proprietary vs Open Ecosystem
Custom Transport (Google/AWS)
integration level %: 80–100
Application Layer / Workloads
Compute & Storage (Integrated)
Custom Transport Fabric
Proprietary Protocol (e.g., GNet, AWS Nitro)
Silicon-Defined Networking
Open Standards (NVIDIA/Broadcom/Meta, UEC)
interoperability %: 50–100
Application Layer (Various AI/HPC)
Compute (e.g., GPUs, CPUs)
Open Transport Layer
Standard Protocols (e.g., Ethernet, UEC, RoCEv2)
Open Fabric Interfaces
Storage (Standard Interfaces)
Network Hardware (Switches, NICs from multiple vendors)

Industry Architecture Divergence

Silicon and Systems Vendor Strategies

NVIDIA (Vertical Integration)

NVIDIA pursues deep vertical integration, offering end-to-end, highly optimized AI infrastructure that spans from GPUs to networking hardware and comprehensive software stacks.

InfiniBand: As the historical leader in HPC interconnects, NVIDIA's Quantum InfiniBand platform remains dominant in the highest echelons of AI training, providing complete, purpose-built solutions that co-design switches, Host Channel Adapters, and software for maximum performance.

Advanced technologies like Scalable Hierarchical Aggregation and Reduction Protocol leverage the network for in-network computing, offloading collective communication operations from GPUs directly into switches. This further reduces latency and frees GPU cycles for computation rather than communication overhead.

Ethernet (Spectrum-X): Recognizing the vast enterprise Ethernet markets, NVIDIA developed Spectrum-X end-to-end platforms that deliver InfiniBand-like performance over Ethernet infrastructure—not simply collections of individual components but holistic architectures combining Spectrum-4 switches with BlueField-3 SuperNICs.

Zero Touch RoCE and RTTCC: A major barrier to RoCE enterprise adoption has been the complexity of configuring entire fabrics—both switches and NICs—for ECN and PFC lossless operation, and addressing this challenge, NVIDIA developed Zero Touch RoCE.

Broadcom (Merchant Silicon Dominance)

Broadcom's strategy centers on providing leading merchant networking silicon where their ASICs form the foundation for numerous vendor switch platforms, including those from Arista and Juniper, which are extensively used in Meta and other hyperscaler data centers.

Jericho4: The latest Jericho generation is purpose-built for distributed, large-scale AI infrastructure, manufactured on 3nm process technology with deep buffering capabilities and intelligent congestion control to ensure lossless RoCE transport over distances exceeding 100 kilometers—crucial for building multi-building or geographically distributed AI clusters.

A key innovation is the 3.2 Tbps "HyperPort" that logically consolidates four 800GE links, simplifying management, improving utilization, and eliminating traditional ECMP inefficiencies that plague AI workloads.

Open Ecosystem Strategy: Contrasting with NVIDIA's vertical integration approach, Broadcom champions open ecosystems by providing powerful, programmable, standards-compliant silicon to a wide range of customers, fostering systems-level competition and innovation.

System Vendors (Building on Silicon)

Arista: As a leader in high-performance data center networking, Arista leverages Broadcom Jericho and Tomahawk silicon in their 7000-series platforms combined with their robust and extensible EOS, providing high-performance lossless Ethernet that fully supports the standard RoCEv2 toolkit, including PFC and ECN with Weighted Random Early Detection.

Juniper: Juniper pursues an "ASIC diversity" strategy, offering architectural choice to customers where their portfolio includes Broadcom Tomahawk-based QFX leaf switches and Juniper's own Express silicon in PTX high-radix spine and super-spine routers.

Cisco: Cisco addresses AI markets with Nexus HyperFabric—new solutions developed in partnership with NVIDIA that represent strategic shifts toward integrated, full-stack approaches combining Cisco Ethernet switching based on Silicon One with NVIDIA accelerated computing including GPUs and BlueField DPUs plus VAST storage systems.

Hyperscaler AI Fabric Architectures

Hyperscale providers operate at scales that often precede off-the-shelf solution capabilities, forcing them to pioneer network architectures where their custom AI fabrics provide glimpses into high-performance networking's future, revealing approaches ranging from clean-slate custom designs to pushing open standards to their absolute limits.

Google (Clean-Slate, Custom Design)

Google's data center networking has long been characterized by willingness to build ground-up custom solutions rather than adopt off-the-shelf technologies, and their AI infrastructure continues this tradition.

Jupiter Fabric: Google's massive-scale, software-defined Jupiter fabric powers all their services, including their largest ML training workloads, where a key innovation involves moving beyond rigid, multi-layered Clos topologies by incorporating data center interconnection using Optical Circuit Switches.

Swift Congestion Control: Complementing their custom hardware, Google developed Swift congestion control, deployed since 2017 as a delay-based algorithm that doesn't rely on ECN marking or switch buffer states.

Instead, Swift uses high-precision hardware timestamps to measure end-to-end latency where control loops employ AIMD to maintain small, constant queuing delays called "delay targets," and this approach provides both extremely low latency for short RPC requests and high line-rate throughput for large AI training data transfers.

AWS (Vertically Integrated Custom Transport)

Similar to Google, AWS leverages control over their entire cloud stack to build completely custom, vertically integrated networking solutions optimized for their specific use cases and scale requirements.

Scalable Reliable Datagram: The custom transport protocol powering EFA represents a fundamental rethinking of transport for multipath-rich data center environments where core logic implements in AWS's custom Nitro hardware.

Key features include:

- **Intelligent Multipath:** "Sprays" packets from single logical flows across many physical paths while continuously monitoring RTT
- **Reliable but Out-of-Order:** Decouples reliability from ordering, guaranteeing packet delivery but not arrival order
- **Proactive Control:** Continuously estimates available bandwidth and RTT, proactively adjusting transmission rates

Meta (Scaling Open Standards to the Limit)

Meta's strategy involves aggressively adopting and scaling open standards, becoming leaders in operating RoCEv2 for massive AI cluster deployments.

RoCE at Scale: Meta has successfully deployed multiple 24,000-GPU Llama training clusters using RoCEv2 over Ethernet, employing traditional Clos topologies built with Arista merchant silicon where their experience demonstrates that careful co-design of network, software, and model architectures can make standard Ethernet and RoCE perform without bottlenecks even for the most demanding generative AI workloads.

Disaggregated Scheduled Fabric: Next-generation Meta clusters are evolving toward DSF—a significant philosophical shift from reactive congestion control to proactive scheduling.

Paradigm Shift: This moves from reactive paradigms where you detect congestion then respond to deterministic paradigms where you schedule traffic to prevent congestion, and the approach promises greater scale, performance predictability, and deterministic behavior critical for AI workloads.

Alibaba (Custom Architecture for LLM Training)

Alibaba Cloud developed custom architectures specifically tailored for Large Language Model traffic patterns and training requirements.

High-Performance Networking: Alibaba's data center network specifically addresses the low-entropy, bursty traffic characteristics of LLM training where traditional 3-tier Clos architectures, when faced with this traffic pattern, become prone to ECMP polarization issues.

HPN mitigates this using a rail-optimized 2-tier dual-plane architecture that interconnects 15,000 GPUs—scales that typically require 3-tier architectures—while significantly reducing reliance on ECMP for load balancing.

Part IV: The Critical Role of Telemetry and Its Security

If congestion control algorithms are the muscles that keep AI fabrics performing under pressure, telemetry systems form the nervous system, and the shift from slow reactive management to fast proactive control depends entirely on gathering accurate, granular, real-time network state data that introduces critical new security challenges around protecting telemetry from exposure and manipulation by increasingly sophisticated threats.

Gaining Fabric Visibility

Your AI fabric's microsecond-scale dynamics render traditional monitoring approaches completely obsolete. Congestion events arise and dissipate faster than SNMP polling cycles can complete, meaning managing these networks effectively requires per-packet, per-hop visibility delivered in real-time to control systems.

Shift to Proactive Monitoring: Modern telemetry's goal is providing data for proactive control—seeing congestion building and reacting before performance degradation occurs—which requires moving far beyond simple link utilization metrics to richer data streams that capture queue depths, packet residence times, and detailed path information.

In-band Network Telemetry (INT/IOAM)

A fundamental shift occurred in how telemetry data is collected. Instead of having central systems poll network devices through out-of-band monitoring, In-band Network Telemetry embeds telemetry directly into live packets traversing the network.

As packets pass through INT-enabled switches, devices append metadata to packet headers—switch IDs, ingress and egress timestamps, current queue occupancy levels—so by the time packets reach their destinations, they carry complete hop-by-hop records of the path traversed and conditions encountered.

This provides unprecedented visibility into the real-time experience of actual workload traffic, rather than synthetic probes or statistical samples that might miss critical congestion events.

Standardization Efforts

The IETF actively standardizes in-band telemetry to ensure interoperability across vendor equipment where the In-situ Operations, Administration, and Maintenance working groups have produced an extensive RFC series including 9197, 9322, and 9326 that define frameworks and data formats.

IETF "Fast CNP" draft proposals suggest using IOAM within congestion notification mechanisms, giving senders detailed information about congestion location and nature that enables much more intelligent responses than simple rate reduction.

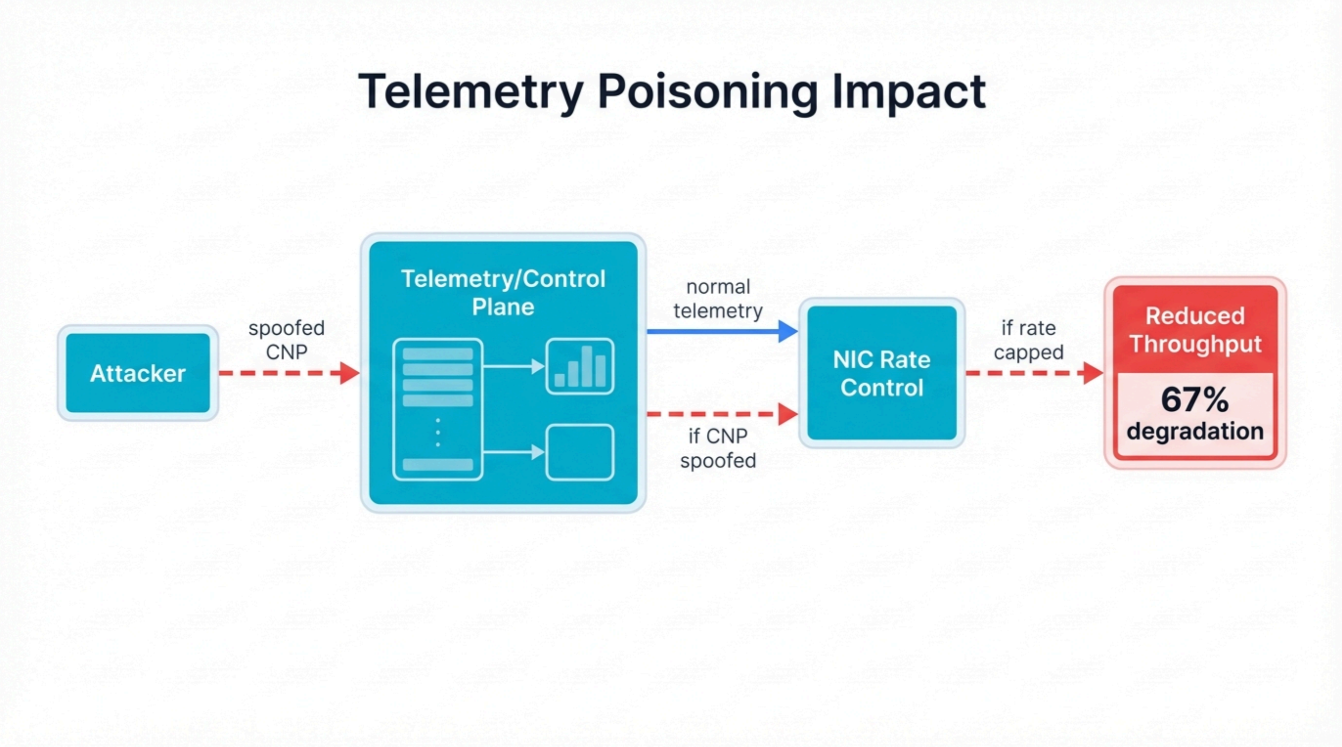
Programmable Data Planes (P4)

The Programming Protocol-independent Packet Processors language gives network operators direct control over switch data plane packet processing, creating powerful opportunities for custom telemetry collection.

Advanced Capabilities: Research demonstrates P4-enabled switches performing novel functions—collecting optical transport telemetry from underlying physical layers, or even deploying simple Deep Neural Networks directly in forwarding planes to perform wire-speed, real-time security analysis of traffic patterns.

Security Posture of AI Fabrics

Telemetry creates double-edged consequences. Modern telemetry's rich data streams offer essential optimization capabilities, yet they create significant new attack surfaces that, if left unsecured, enable disruption or surveillance of AI workloads worth billions of dollars.



Telemetry Security Attack Path

Telemetry as Attack Surface

Granular telemetry data carries high sensitivity from multiple perspectives where attackers gaining access to INT or IOAM streams can achieve several objectives:

- **Map fabric topology** through packet traversal observation where switch IDs reveal detailed physical data center network structures

- **Identify critical paths** via latency and queue data analysis where heavily utilized connections between GPU clusters and storage systems get exposed
- **Infer workload behavior** by correlating traffic patterns with source and destination information where model architectures and training strategies potentially leak

QoS Manipulation and Denial of Service

Tight coupling creates dangerous vulnerabilities. Telemetry and congestion control connect directly, creating direct vectors for performance degradation because modern algorithms explicitly react to telemetry signals like ECN markings, CNPs, and RTT measurements that drive behavior.

Signal manipulation can trick systems into self-destruction where systems cripple their own performance as attackers inject forged CNPs, intercept and modify RTT probe responses, and create artificially inflated latency measurements that result in sender NICs dramatically and unnecessarily throttling transmission rates.

Devastating Effectiveness: Highly effective Denial of Service attacks emerge when recent 2024 studies demonstrated devastating effectiveness by reducing AI training throughput by up to 89% while remaining virtually undetectable to traditional monitoring systems.

Advanced Persistent Threats

Security researchers identified new attack patterns targeting AI infrastructure in 2024-2025 where sophisticated actors use machine learning to analyze telemetry patterns and identify optimal times to launch attacks when they'll cause maximum disruption to training runs.

These "AI-enhanced" attacks can adapt their behavior based on observed network patterns, making them particularly dangerous for long-running AI training jobs that represent massive investments in compute time and resources.

Securing Telemetry Data

Telemetry security remains a nascent field. While existing research has focused primarily on resource-constrained IoT environments, core principles apply to high-performance fabric deployments:

- **Authentication:** Ensuring telemetry and control packets originate from legitimate sources
- **Integrity and Confidentiality:** Protecting telemetry data from unauthorized modification and eavesdropping
- **Anonymization:** Privacy-preserving techniques when sharing or storing telemetry data for analysis

Business Critical Concern: With 93% of security leaders anticipating daily AI-enhanced attacks by 2027, according to recent threat intelligence, the security of telemetry systems becomes a business-critical concern that can no longer be treated as an afterthought.

Part V: Synthesis, Incident Analysis, and Future Outlook

Impact of Network Mismanagement on AI Workloads

Comprehensive understanding of real-world network failure impacts on AI workloads faces significant challenges due to the highly proprietary and competitive nature of large-scale AI development where organizations remain extremely reluctant to publish detailed incident post-mortems that might reveal weaknesses in multi-million or billion-dollar infrastructure investments.

The Scarcity of Public Incident Reports

Unlike mature SRE practices in web services where public post-mortems regularly share lessons learned, AI infrastructure remains largely opaque, and we found no publicly available reports that explicitly detail incidents where misconfigured DCQCN parameters caused PFC storms that failed specific LLM training runs.

Inferring Impact from System Characteristics

The tight coupling and synchronous nature of distributed AI training means that even subtle network issues can have outsized impacts on entire system performance.

Congestion Mismanagement and Wasted Resources: Network congestion has direct, unavoidable consequences that immediately translate to increased tail latency, and for synchronous AI workloads, this immediately translates to longer Job Completion Times.

When you consider that a single H100 GPU hour costs approximately $3-4 in cloud environments, and training runs involve thousands of GPUs for weeks or months, even small network inefficiencies translate to enormous costs.

Hardware Failures as Network Impact Proxy: Meta publicly stated that 66% of large-scale AI training interruptions can be attributed to hardware failures, explicitly including network switches in this category, and this data powerfully underscores networks as critical components in overall cluster reliability.

Grey Failures: Networks are particularly susceptible to "grey failures"—intermittent issues like misconfigured Quality of Service, faulty cables, or flapping links that manifest as subtle performance degradation that's extremely difficult to diagnose and often gets misattributed to other system components.

Case Studies and Frameworks

Telecom Case Study: A major telecommunications operator case study demonstrates the power of AI-driven proactive congestion management where machine learning algorithms analyze real-time network data to predict traffic hotspots and dynamically reallocate resources to prevent congestion before it impacts service quality.

Failure Diagnosis Frameworks: Academic systems like L4—which performs log-based diagnosis of LLM training failures—highlight the significant challenges in root cause analysis for complex distributed systems where the complexity of modern AI systems makes human-driven troubleshooting increasingly untenable.

Future Trajectories and Strategic Recommendations

AI networking continues evolving at breakneck pace, driven by exponential growth in model sizes and complexity, and several key trends will shape the future of fabric architectures and operations, creating both opportunities and challenges for organizations building AI infrastructure.

Ultra Ethernet Consortium

The most significant trend in open ecosystem development is the Ultra Ethernet Consortium's formation, backed by Broadcom, Meta, Microsoft, AMD, and Arista, with a mission involving developing open, interoperable Ethernet standards specifically optimized for AI and HPC workloads.

UEC specifications go beyond current standards by incorporating packet-spraying multipath techniques that avoid ECMP polarization issues, enhanced congestion control algorithms, and improved telemetry capabilities.

Proactive vs. Reactive Control

The future of congestion management will likely synthesize the leading philosophical approaches we've examined where Meta's DSF represents proactive, centrally scheduled approaches that offer deterministic, congestion-free operation by design while NVIDIA's Spectrum-X represents reactive, telemetry-driven approaches that offer extreme agility and workload isolation capabilities.

Future architectures will likely combine both approaches—high-level schedulers orchestrating large traffic patterns and long-term resource allocation, while fast local reactive mechanisms handle microbursts and transient conditions that can't be perfectly predicted.

Rise of AIOps

The growing scale and complexity of AI fabrics makes manual configuration, monitoring, and troubleshooting increasingly untenable, and the future of network operations is AIOps—machine learning platforms that ingest massive telemetry streams and become essential for predicting congestion, detecting anomalies, performing automated root cause analysis, and triggering automated remediation.

Operational Evolution: This represents a shift from reactive human-driven operations to predictive AI-driven operations where the NVIDIA Morpheus AI framework, launched in 2024, exemplifies this trend by providing real-time threat detection and automated response capabilities directly integrated into network infrastructure.

Scaling to 800G, 1.6T, and Beyond

Bandwidth demands show no signs of slowing. As AI models continue growing in size and complexity, the industry rapidly transitions from 400G to 800G interfaces, with 1.6T already on the horizon.

These higher speeds compress latency budgets even further, straining existing congestion control approaches and demanding faster reaction times and more intelligent control loops where the physics of high-speed networking—particularly signal propagation delays—become increasingly constraining factors.

Security Integration

The security landscape for AI infrastructure continues evolving rapidly. By 2025, we expect to see integrated security become a standard feature rather than an afterthought in AI fabric design, including built-in telemetry authentication, encrypted control plane communications, and real-time threat detection capabilities integrated directly into network hardware.

Business Imperative: The projected $25-40 billion in annual losses from AI-enhanced attacks by 2027 makes security integration a business imperative rather than just a technical nicety.

Strategic Recommendations

Two competing visions dominate your AI networking landscape that will define infrastructure architecture, performance, and economics for the next decade.

The first vision gets epitomized by NVIDIA's approach where InfiniBand and NVLink ecosystems represent fully vertically integrated, single-vendor solutions with every component co-designed for maximum performance, and these approaches offer proven, best-in-class performance but create vendor lock-in risks and potentially higher total cost of ownership.

The competing vision involves open, multi-vendor ecosystems where next-generation Ethernet gets built as promised by the Ultra Ethernet Consortium, and this approach offers choice, fosters innovation, and leverages massive ecosystem scale, but fully integrated, proven solutions currently lack availability.

Competition outcomes will fundamentally define AI infrastructure choices where organizations must weigh complex tradeoffs carefully: performance optimization versus vendor flexibility, operational complexity versus feature integration, and current capabilities versus future innovation potential.

AI workloads continue growing in scale and sophistication. Security threats become more advanced and persistent, and the networking infrastructure you choose today determines your organization's competitive ability in the AI-driven future where stakes have never been higher.

Final Insight: Convergence creates both risks and opportunities where high-performance networking, advanced security requirements, and unprecedented scale challenges intersect, and organizations successfully navigating these complexities will gain sustainable competitive advantages in the AI economy while those that don't get constrained by infrastructure decisions made without understanding long-term implications.

Choose wisely. Your AI future depends on it.

Example Implementation

```
# Example: Model training with security considerations
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

def train_secure_model(X, y, validate_inputs=True):
    """Train model with input validation"""

    if validate_inputs:
        # Validate input data
        assert X.shape[0] == y.shape[0], "Shape mismatch"
        assert not np.isnan(X).any(), "NaN values detected"

    # Split data securely
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.2, random_state=42, stratify=y
    )

    # Train with secure parameters
    model = RandomForestClassifier(
        n_estimators=100,
        max_depth=10,  # Limit to prevent overfitting
        random_state=42
    )

    model.fit(X_train, y_train)
    score = model.score(X_test, y_test)

    return model, score
```

Thank You for Reading

Explore more AI security research at **perfecxion.ai**

This document was generated from perfecXion.ai
For the latest updates, visit the online version