InfiniBand vs Spectrum-X vs RoCE vs Ethernet for AI Clusters

Published July 21, 2026 39 min read

Executive Summary

Choosing a network fabric for an artificial intelligence (AI) cluster in 2026 means choosing among four overlapping options: **InfiniBand**, **NVIDIA Spectrum-X** (an Ethernet platform), **RDMA over Converged Ethernet** (**RoCE**, specifically RoCEv2), and generic or **Ultra Ethernet Consortium** (**UEC**) standards-based Ethernet. These are not four independent technologies so much as two link-layer choices (InfiniBand versus Ethernet) crossed with an implementation question (whose RDMA stack and whose silicon). InfiniBand remains the lowest-latency, most deterministic fabric, with switch hop latency in roughly the 200 to 300 nanosecond range against a typical Ethernet hop of 300 to 3,000 nanoseconds (Source: datacenterdynamics.com), and it still connects the majority of the highest-ranked systems on the semiannual TOP500 supercomputer list (Source: nvidianews.nvidia.com). But Ethernet, running RoCEv2 with priority flow control (PFC) and explicit congestion notification (ECN), has become the dominant choice for the largest hyperscale AI training deployments by volume: Dell'Oro Group reported that Ethernet switch sales in AI back-end networks "more than doubled and accounted for about two-thirds of data center switch sales in AI clusters during the first quarter 2026" (Source: delloro.com), even as InfiniBand sales "more than tripled during the quarter" on the ramp of NVIDIA's Blackwell Ultra platform (Source: delloro.com).

Meta operates two 24,576-graphics processing unit (GPU) clusters used to train Llama 3 entirely on RoCEv2 over Ethernet, describing the approach in a peer-reviewed paper at ACM SIGCOMM 2024 (Source: engineering.fb.com). xAI's Colossus supercomputer, which trains the Grok model family, scaled to 100,000 NVIDIA Hopper GPUs on NVIDIA's Spectrum-X Ethernet platform rather than InfiniBand (Source: nvidianews.nvidia.com), and has since doubled to 200,000 GPUs (Source: x.ai). Microsoft Azure's ND H100 v5 virtual machines and Oracle Cloud Infrastructure's (OCI) newest Zettascale10 cluster illustrate the split further: Azure still wires each GPU to a dedicated 400 Gigabit-per-second (Gb/s) NVIDIA Quantum-2 InfiniBand connection (Source: learn.microsoft.com), while Oracle's OpenAI-partnered Stargate site in Abilene, Texas, targeting up to 800,000 GPUs, runs on Oracle's own "Acceleron RoCE networking architecture" (Source: oracle.com).

The market context explains why both camps are growing simultaneously rather than one replacing the other. Dell'Oro Group forecasts that spending on data center switches in AI back-end networks will "surpass $100 billion by 2030" (Source: delloro.com), while 650 Group projects that the total AI-era data-center-networking market could "approach and potentially exceed $200B" annually by decade's end (Source: 650group.com), with AI scale-out Ethernet revenue alone "surging to over $100B by 2030" (Source: 650group.com). The Ultra Ethernet Consortium, founded in 2023 by AMD,

Arista, Broadcom, Cisco, Eviden, Hewlett Packard Enterprise (HPE), Intel, Meta, and Microsoft (Source: linuxfoundation.org), published its 1.0 specification in June 2025 (Source: ultraethernet.org) and now counts AMD's Pensando Pollara 400 as the first UEC 1.0-compliant network interface card (NIC) to ship, with Oracle as the first cloud provider to deploy it (Source: networkworld.com). For most buyers building multi-thousand-GPU clusters in 2026, the practical decision is not "InfiniBand or Ethernet" in the abstract but a function of tenancy model, in-house networking skills, vendor lock-in tolerance, and how many nanoseconds of tail latency actually change job completion time; this report works through each fabric's architecture, adoption, and quantified trade-offs to make that decision legible, including how the underlying topology, congestion control, and switch-silicon choices interact with each option.

Introduction and Background

Training and serving large-scale AI models depends on moving enormous volumes of data between accelerators, not just on the accelerators themselves. A distributed training job synchronizes gradients across tens of thousands of GPUs after every batch, and Arista's AI networking whitepaper notes that in this "compute-exchange-reduce cycle, approximately 20-50% of the job time is spent communicating across the network," so bottlenecks have "a substantial impact on job completion time" (Source: arista.com). This is the reason a specialized "back-end" or "east-west" fabric, separate from the front-end network that serves ordinary enterprise traffic, has become a standard feature of large AI clusters. Meta, for instance, explicitly separated GPU-based training into its own backend network when it built its 24,576-GPU clusters (Source: cs.stanford.edu).

At the center of this back-end fabric is remote direct memory access (RDMA), a mechanism that lets one server's network interface card write directly into another server's memory, bypassing the operating system kernel and the central processing unit (CPU) on both ends. **InfiniBand**, introduced in the late 1990s, was purpose-built around RDMA and credit-based, lossless flow control from its inception (Source: networkworld.com). **RoCE**, and specifically **RoCEv2**, retrofits the same RDMA semantics onto standard Ethernet by encapsulating RDMA traffic inside User Datagram Protocol (UDP) and Internet Protocol (IP) packets, which is what allows it to traverse ordinary IP routers and coexist with the rest of a data center's Ethernet infrastructure. **NVIDIA Spectrum-X** is not a new wire protocol; it is a commercial platform, combining NVIDIA's Spectrum Ethernet switches with its BlueField and ConnectX/SuperNIC network adapters and proprietary RoCE extensions, engineered as a tuned, matched pair to make RoCEv2 Ethernet perform closer to InfiniBand at very large scale (Source: docs.nvidia.com).
The fourth term in the query, generic "Ethernet," typically refers to the emerging **Ultra Ethernet** standard: an effort by the UEC to standardize the RoCE-like transport innovations that individual vendors (NVIDIA, Broadcom, AMD, Oracle) have each been building proprietarily, so that AI fabrics can be built from multiple, interoperable vendors rather than one.
The physical topology underneath all four fabrics is usually some variant of a non-blocking Clos or "fat tree" network, in which every leaf (top-of-rack) switch connects to every spine switch so that any GPU can reach any other GPU with the same number of hops and the same available bandwidth, avoiding hot spots. NVIDIA documents that its Quantum-X800 InfiniBand switches "support a two-level, fat-tree topology capable of connecting up to 10,368 network interface cards (NICs) with minimal latency and optimal job locality" (Source: networking-docs.nvidia.com), while a vendor engineering paper on AI cluster back-end fabrics frames the basic building block of these designs as a "scale unit," a repeatable rack-level module that is replicated to grow the cluster to its final size (Source: datacenterdynamics.com). Regardless of link-layer choice, most production AI fabrics also implement "rail-optimized" wiring, in which each GPU's dedicated NIC connects to a separate, parallel plane of switches rather than sharing a single uplink, a pattern visible in both NVIDIA's Spectrum-X Multiplane architecture and Oracle's Acceleron design discussed later in this report.

The stakes of this choice are large. Dell'Oro Group's Sameh Boujelbene has framed the shift bluntly: "just as Ethernet prevailed over InfiniBand in large-scale scale-out environments, we are now seeing alternative technologies such as UALink and Ethernet gain momentum in scale-up compute fabrics" (Source: delloro.com). Yet InfiniBand has not disappeared, and in the most recent quarter for which Dell'Oro has published data, InfiniBand sales "more than tripled" even as Ethernet held roughly two-thirds of unit volume (Source: delloro.com). The remainder of this report examines each fabric individually, compares them directly on a feature matrix and in published benchmarks, walks through named production deployments, and closes with implications for buyers planning clusters through the rest of the decade.

InfiniBand

Capabilities

InfiniBand is a purpose-built interconnect architecture, standardized by the InfiniBand Trade Association (IBTA), that was designed from the ground up around RDMA and credit-based flow control rather than best-effort packet delivery. Its defining architectural trait is that InfiniBand "prevents packet loss by ensuring receiving buffers have space before transmission begins," using link-level credits rather than the drop-and-retransmit model of conventional Ethernet (Source: networkworld.com). NVIDIA's current flagship platform, **Quantum-X800**, delivers "the world's first end-to-end 800 Gb/s networking" with 144 ports of 800 Gb/s connectivity per switch (Source: nvidia.com), while the prior-generation **Quantum-2 (NDR)** platform ships 64

ports of 400 Gb/s in a single 1U chassis with 51.2 terabits per second (Tb/s) of aggregated bidirectional throughput (Source: networking-docs.nvidia.com). The IBTA's Volume 1 Release 2.0 specification, published July 31, 2025, supports "XDR speeds up to 200Gb/s per lane, enabling total link bandwidths of 800Gb/s (QSFP) and 1.6Tb/s (QSFP-DD and OSFP)" (Source: infinibandta.org).

Beyond raw bandwidth, InfiniBand includes several features not natively present in Ethernet. Adaptive routing lets switches reroute packets around congested links using real-time queue-depth telemetry rather than a static hash (Source: datacenterdynamics.com). **SHARP** (Scalable Hierarchical Aggregation and Reduction Protocol), now in its fourth generation on Quantum-X800, offloads collective operations such as all-reduce directly into switch silicon, and NVIDIA reports service providers seeing "10-20% performance improvements for in-house AI workloads" from SHARP integration with the NVIDIA Collective Communication Library (NCCL) (Source: developer.nvidia.com). Quantum-X800 also markets up to 9x higher in-network computing throughput over prior generations, alongside the doubled 800 Gb/s per-port bandwidth described above.

Adoption

InfiniBand's institutional stronghold remains high-performance computing (HPC) and the frontier of AI training. Multiple named hyperscale deployments run on it: Microsoft Azure's **ND H100 v5** instances give "each GPU within the VM ... its own dedicated, topology-agnostic 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand connection" (Source: learn.microsoft.com), and Azure documented that "the inclusion of NVIDIA Quantum-2 CX7 InfiniBand with 3,200 Gbps cross-node bandwidth ensures seamless performance across the GPUs at massive scale" (Source: blogs.nvidia.com). GPU cloud provider CoreWeave offers InfiniBand as one of two RDMA-enabled fabrics, describing it as "the fastest, lowest-latency, and most reliable way for GPUs to exchange information in traditional HPC environments" (Source: docs.coreweave.com). On the TOP500 list of the world's fastest publicly benchmarked supercomputers, NVIDIA (then Mellanox) reported that its "InfiniBand solutions accelerate six of the top ten" systems, including "the top three, and four of the top five" (Source: nvidianews.nvidia.com), and while that data point is from June 2019, InfiniBand's tendency to dominate the raw-performance tier (rather than the systems-count tier) of TOP500 has persisted for years, since its typical customers are single-tenant national laboratories and research clusters optimizing for pure floating-point throughput rather than multi-tenant cloud economics.

Strengths and Limitations

InfiniBand's principal strength is a deterministic, low, and well-understood latency floor: switch hop latency in the "~200ns to 300ns" range according to one vendor engineering analysis, versus a considerably wider and more variable range for Ethernet. Its native congestion management and adaptive routing are mature and purpose-built rather than adapted. Its principal limitation is vendor concentration: NVIDIA (through its 2019 acquisition of Mellanox) is effectively the sole supplier of InfiniBand switch silicon and host channel adapters at scale, a dynamic that industry commentary characterizes as "single-vendor pricing" carrying a cost premium; one vendor analysis estimates "the InfiniBand premium typically lands in the +30% to +60% range over open-hardware Ethernet for equivalent capacity" (a claim that should be read as an interested Ethernet-hardware vendor's estimate rather than an audited figure) (Source: ipinfusion.com). InfiniBand also typically requires a dedicated subnet manager and specialized administration skills distinct from a conventional Ethernet network operations team.

NVIDIA Spectrum-X

Capabilities

Spectrum-X is described by NVIDIA as "the world's first Ethernet networking platform for AI," combining the Spectrum-X Ethernet switch with the Spectrum-X SuperNIC to run RoCE at higher effective bandwidth than commodity Ethernet (Source: nvidia.com). NVIDIA claims the platform accelerates "AI network performance by 1.6x over off-the-shelf (OTS) Ethernet" (Source: nvidia.com), a vendor benchmark that should be read alongside independent testing discussed later in this report. Under the hood, Spectrum-X combines the NVIDIA Spectrum-4 (and now Spectrum-6) switch application-specific integrated circuit (ASIC) with the BlueField-3 SuperNIC, ConnectX-7, and newer ConnectX-8 SuperNIC families, and NVIDIA documents that its "advanced RoCE extensions for scalable AI communications" include RoCE adaptive routing, congestion control, and performance isolation designed to hold "up to 95% effective bandwidth across hyperscale systems at load and at scale" (Source: developer.nvidia.com). NVIDIA's "Spectrum-X Multiplane" architecture splits each GPU SuperNIC across two or more independent network planes and can scale "Spectrum-X Ethernet networks up to 128K GPUs in two tiers" (Source: nvidia.com). In August 2025 NVIDIA added a third tier, **Spectrum-XGS Ethernet**, a "scale-across" layer for linking multiple physically separate data centers into a single "giga-scale" AI training domain; NVIDIA reports it "delivers up to 1.9x higher NCCL all-reduce bandwidth compared to off-the-shelf Ethernet" for that use case (Source: developer.nvidia.com).

Adoption

The signature Spectrum-X deployment is xAI's **Colossus** supercomputer in Memphis, Tennessee. NVIDIA announced that "xAI's Colossus supercomputer cluster comprising 100,000 NVIDIA Hopper Tensor Core GPUs ... achieved this massive scale by using the NVIDIA Spectrum-X Ethernet networking platform" (Source: nvidianews.nvidia.com), and that the entire facility "achieves unprecedented network performance" with "zero application latency degradation or packet loss due to flow collisions" across all three network tiers (Source: nvidianews.nvidia.com). The Register independently confirmed that "unlike most AI training clusters, xAI's Colossus with its 100,000 Nvidia Hopper GPUs doesn't use InfiniBand," describing the choice as notable specifically because it broke with convention (Source: theregister.com). xAI's own site now lists **200,000** H100 GPUs "in a single interconnected cluster," reached in 276 days from groundbreaking (Source: x.ai). CoreWeave also offers Spectrum-X (RoCE) as one of its two primary HPC interconnect options, alongside InfiniBand, on newer GPU generations such as GB300 (Source: docs.coreweave.com) and is the launch partner for Spectrum-XGS scale-across networking (Source: nvidianews.nvidia.com).

Strengths and Limitations

Spectrum-X's core value proposition is capturing much of InfiniBand's tuned, matched-hardware performance while running standard, routable Ethernet frames that integrate with the rest of a hyperscaler's data-center Ethernet fabric, support multi-tenancy, and use SONiC or NVIDIA's Cumulus Linux as an open network operating system (Source: nvidia.com). The limitation is that, despite running on Ethernet, Spectrum-X is still a largely NVIDIA-defined, matched-silicon platform (Spectrum switch plus BlueField/ConnectX SuperNIC); it narrows but does not eliminate the vendor-concentration dynamic associated with InfiniBand, since achieving the advertised 1.6x uplift depends on NVIDIA's specific switch-and-NIC pairing rather than being guaranteed on arbitrary third-party Ethernet gear.

RoCE (RDMA over Converged Ethernet)

Capabilities

RoCEv2 encapsulates the InfiniBand transport's RDMA verbs inside UDP/IP packets, allowing GPUs to perform direct memory-to-memory transfers over conventional, routable Ethernet. As one technical comparison frames it, "both protocols expose RDMA verbs at the top of the stack... The verb layer is identical. The difference is what happens between the NIC and the wire" (Source: factryze.ai). Losslessness on RoCEv2, unlike InfiniBand's native link-level credit system, is achieved through a combination of **priority flow control (PFC)**, a mechanism that pauses specific traffic classes on a link rather than the whole link when a downstream buffer nears capacity, and **explicit congestion notification (ECN)**, which marks packets so endpoints can throttle transmission before buffers overflow, historically managed end to end by **Data Center Quantized Congestion Notification (DCQCN)**. Meta's peer-reviewed SIGCOMM 2024 paper is the most detailed public account of running RoCEv2 at scale: it documents that Meta built its backend fabric "using RDMA Over Converged Ethernet version 2 (RoCEv2) as the inter-node communication transport for the majority of our AI capacity" (Source: engineering.fb.com), that this backend fabric "utilizes the RoCEv2 protocol, which encapsulates the RDMA service in UDP packets for transport over the network" (Source: engineering.fb.com), and, notably, that Meta "initially attempted to use DCQCN for congestion management but then pivoted away from DCQCN to instead leverage the collective library itself to manage congestion" (Source: cs.stanford.edu). Meta reported that after moving to 400 Gigabit deployments, "default DCQCN settings and doubled ECN thresholds compared to 200G networks" degraded performance, and that the firm has since run "over a year of experience with just PFC for flow control, without any other transport-level congestion control," observing "stable performance and lack of persistent congestion for training collectives" (Source: engineering.fb.com).

Adoption

RoCEv2 is the transport underneath both NVIDIA's Spectrum-X and most non-NVIDIA hyperscaler Ethernet AI fabrics, making it arguably the single most widely deployed RDMA transport for AI clusters by GPU count. In addition to Meta's 24,576-GPU Llama 3 clusters (Source: engineering.fb.com), Oracle's OCI Zettascale10 platform, underpinning its Stargate partnership with OpenAI, is "built on next-generation Oracle Acceleron RoCE networking architecture" and provides "ultra-low-latency RoCEv2 networking" targeting up to 800,000 GPUs per deployment (Source: oracle.com). Oracle's earlier OCI cluster network used "RDMA over Converged Ethernet (RoCE) on top of NVIDIA ConnectX RDMA NICs" to connect A100 GPU nodes, each with "1.6Tbps (1600Gbps) of full-duplex connectivity" per node (Source: blogs.oracle.com). CoreWeave, AWS (via its Elastic Fabric Adapter, discussed below), and every Spectrum-X deployment (including xAI's Colossus) also ultimately run RoCEv2 as the wire-level RDMA transport. Amazon Web Services (AWS) takes a related but distinct approach: its **Elastic Fabric Adapter (EFA)** is "a network device that you can attach to your Amazon EC2 instance to accelerate Artificial Intelligence (AI), Machine Learning (ML), and High Performance Computing (HPC)

applications," and AWS states EFA "provides lower and more consistent latency and higher throughput than the TCP transport traditionally used in cloud-based HPC systems," integrating with NCCL and the Message Passing Interface (MPI) rather than exposing raw InfiniBand or RoCEv2 verbs directly to the customer (Source: docs.aws.amazon.com).

Strengths and Limitations

RoCEv2's core strength is that it runs on the same Ethernet fabric, tooling, and (largely) the same operator skill set as the rest of a hyperscale data center, avoiding a separate subnet manager and specialized InfiniBand administration. Its core limitation, as Meta's own engineering account illustrates, is that lossless Ethernet requires active, hands-on tuning: PFC alone risks head-of-line blocking, and off-the-shelf congestion control algorithms designed for storage workloads (DCQCN) were found by Meta to behave unpredictably under AI collective-communication traffic patterns at 400G speeds, forcing a bespoke, in-house solution built on "careful coordination between the collective communication library and the network" (Source: engineering.fb.com). A 2022 academic study by researchers from the University of Texas at Austin, Georgia Institute of Technology, and Meta similarly found that "the native RoCE congestion control scheme, based on Priority Flow Control (PFC), suffers from many drawbacks such as unfairness, head-of-line-blocking, and deadlock," which is precisely why the paper set out to evaluate whether general-purpose schemes such as DCQCN, DCTCP, TIMELY, and HPCC actually suit the narrower, more predictable traffic patterns of distributed training collectives (Source: arxiv.org). This means RoCEv2's operational simplicity relative to InfiniBand is real but conditional on the scale and sophistication of the operator; Meta's paper is explicit that this tuning took years of iteration.

Standards-Based and Ultra Ethernet

Capabilities

"Ethernet" as a fourth, distinct option in this comparison generally means multi-vendor, standards-based Ethernet running RoCEv2 (or, increasingly, the emerging Ultra Ethernet transport) on switch silicon that is not tied to a single networking vendor's proprietary platform. The switch-silicon layer is dominated by a small number of merchant-silicon vendors. Broadcom's **Tomahawk 5**, shipping in production volume since March 2023, was "the industry's highest bandwidth switch chip" at launch, delivering "51.2 Terabits/sec of Ethernet switching capacity in a single, monolithic device" and enabling "single-hop connectivity between 256 high-performance AI/ML accelerators, each having 200Gbps of network bandwidth" (Source: investors.broadcom.com). Cisco's competing **Silicon One G200**, also a 51.2 Tb/s device, is "purpose-built" for "high-bandwidth, AI networking, and web-scale switching," offering "deterministic, low-latency, and power-efficient" performance and now underpins Meta's own OCP-inspired Cisco 8501 platform (Source: cisco.com) (Source: blogs.cisco.com). Arista's **Etherlink** portfolio delivers "an 800G end-to-end technology platform across front-end, training, inference, and storage networks," supporting cluster sizes "ranging from thousands to 100,000s of XPUs" with one- and two-tier topologies (Source: investors.arista.com).

Layered on top of this silicon, the **Ultra Ethernet Consortium**, operating under the Linux Foundation, released **UEC Specification 1.0** on June 11, 2025, "a comprehensive, Ethernet-based communication stack engineered to meet the demanding needs of modern Artificial Intelligence (AI) and High-Performance Computing (HPC) workloads," covering "all layers of the networking stack, including NICs, switches, optics, and cables, enabling seamless multi-vendor integration" (Source: ultraethernet.org). The specification, updated to version 1.0.2 in January 2026 (Source: ultraethernet.org), brings InfiniBand-like transport techniques ("packet spray across all paths, multi-path RDMA, out-of-order delivery, and selective retransmission") to standard Ethernet without requiring single-vendor lock-in (Source: ipinfusion.com).

Adoption

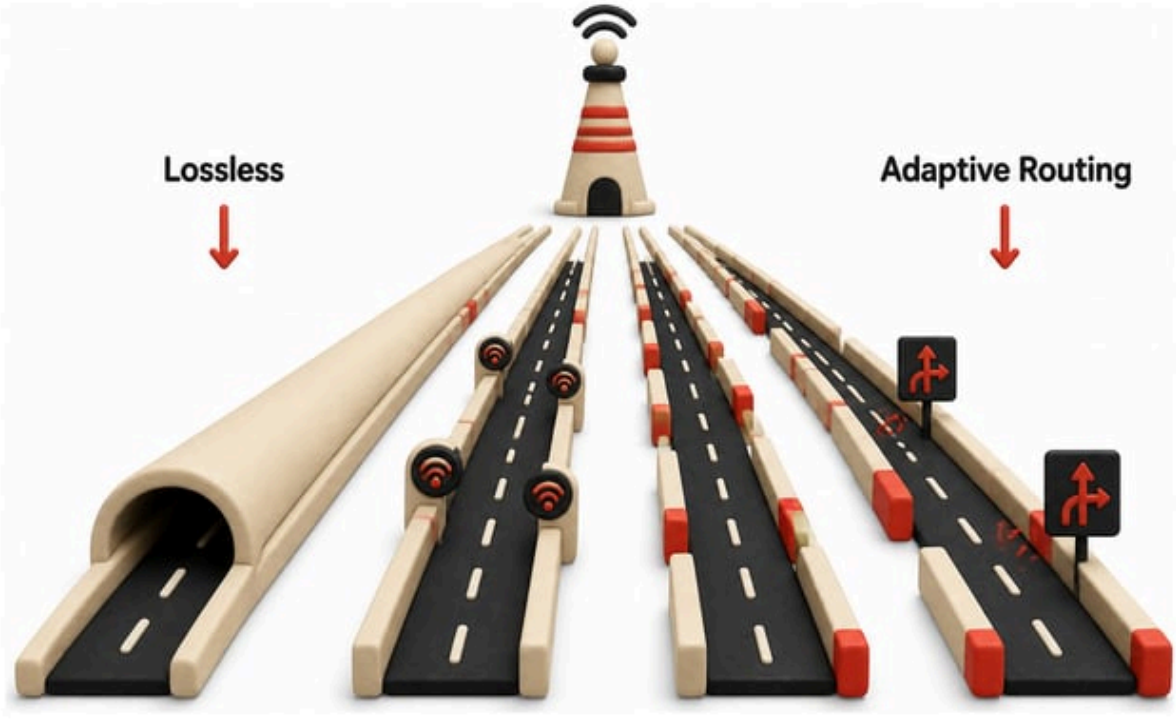
AMD's **Pensando Pollara 400** shipped as "the industry's first NIC that's compliant with the Ultra Ethernet Consortium's (UEC) 1.0 specification," with Oracle announced as its first cloud-provider customer (Source: networkworld.com). AMD followed with **Vulcano**, an 800 Gb/s, fully UEC 1.0-compliant NIC (Source: networkworld.com). Broadcom separately shipped **Thor Ultra**, which it calls "the industry's first 800G Ethernet NIC" and describes as "fully feature compliant with UEC specification" (Source: broadcom.com). The UEC's founding-member roster, AMD, Arista, Broadcom, Cisco, Eviden, HPE, Intel, Meta, and Microsoft (Source: ultraethernet.org), grew to include NVIDIA, Oracle, Alibaba Cloud, Baidu, Bytedance, Dell, Huawei, Juniper, Nokia, and Tencent among 97 total member organizations as of August 2024 (Source: ultraethernet.org).

Strengths and Limitations

Standards-based Ethernet's core value is supply-chain diversity: buyers can source switch ASICs from Broadcom, Cisco, or Marvell; NICs from AMD, Broadcom, Intel, or NVIDIA; and network operating systems from SONiC, Cumulus, or a vendor's proprietary stack, avoiding dependency on a single company's product roadmap and pricing. Its principal limitation, at least as of mid-2026, is immaturity of interoperability: UEC 1.0 was only ratified in mid-2025, compliant silicon only began shipping in volume through 2026, and (per IBTA's own November 2025 announcement) the InfiniBand and Ethernet RDMA ecosystems still run separate, if increasingly overlapping, interoperability testing (Plugfest) programs (Source: infinibandta.org). A separate technology worth flagging for completeness, though outside the scope of this report's four named contenders, is Cornelis Networks' revived **Omni-Path**, originally an Intel design later spun out into an independent company; Network World reports Cornelis "targets price-sensitive deployments where cost-performance optimization matters more than absolute performance" with its CN5000 series aimed at 400 Gb/s AI deployments, illustrating that InfiniBand and Ethernet are not the only two link layers under active development for AI clusters, even if they remain the two that matter for the overwhelming majority of production deployments (Source: networkworld.com).

Feature Comparison

Table 1 below summarizes how InfiniBand, NVIDIA Spectrum-X, generic RoCEv2 Ethernet, and standards-based/Ultra Ethernet compare across the axes that matter most for AI cluster design: latency, loss handling, multi-tenancy, vendor ecosystem, and typical cost posture.



ATTRIBUTE	INFINIBAND	NVIDIA SPECTRUM-X	ROCEV2 (GENERIC ETHERNET)	ULTRA ETHERNET / STANDARDS-BASED
Underlying link layer	Native InfiniBand	Ethernet	Ethernet	Ethernet
Loss handling	Native credit-based flow control, lossless by architecture (Source: networkworld.com)	PFC + NVIDIA RoCE congestion/adaptive-routing extensions tuned to Spectrum switches (Source: developer.nvidia.com)	PFC, with ECN/DCQCN optional; Meta ultimately runs PFC-only with collective-library-driven congestion management	UEC 1.0 defines packet spray, out-of-order delivery, and selective retransmission to reduce PFC dependence (Source: ipinfusion.com)
Typical switch hop latency	Approximately 200 to 300 nanoseconds (Source: datacenterdynamics.com)	Vendor-optimized, tighter than generic Ethernet but not independently disclosed in nanoseconds	Approximately 300 to 3,000 nanoseconds depending on hardware and mode (Source: datacenterdynamics.com)	Same silicon base as generic Ethernet; UEC aims to close remaining gap
Max single-fabric scale (public claims)	Up to 10,368 NICs in a two-level fat tree per switch generation (Source: networking-docs.nvidia.com)	Up to 128,000 GPUs in a flat, two-tier topology via Multiplane (discussed above)	Deployed at 24,576 GPUs (Meta) (Source: engineering.fb.com); Oracle targets 800,000 GPUs	Not yet demonstrated at comparable single-fabric scale as of mid-2026
Vendor ecosystem	Effectively single-vendor (NVIDIA/Mellanox silicon)	NVIDIA-defined switch and SuperNIC pairing	Broad: any RoCEv2-capable NIC and switch	Open, multi-vendor by design (Broadcom, Cisco, AMD, Arista, Marvell, and others) (Source: ultraethernet.org)
Illustrative large deployment	Azure ND H100 v5 (Source: learn.microsoft.com)	xAI Colossus, 100,000 to 200,000 GPUs (Source: x.ai)	Meta 24,576-GPU Llama 3 clusters (Source: engineering.fb.com)	AMD Pollara at Oracle, first UEC-compliant deployment (Source: networkworld.com)

The comparison illustrates that the sharpest technical differentiator is still latency and loss-handling determinism, where InfiniBand retains an edge measured in the hundreds of nanoseconds per hop, but that this edge is increasingly outweighed, for hyperscale buyers, by scale, multi-tenancy, and ecosystem considerations that favor Ethernet-family fabrics. Notably, no single fabric wins on every axis: InfiniBand still leads on raw per-hop latency and single-vendor performance tuning, Spectrum-X leads on demonstrated single-fabric GPU count under one commercial platform, and Ultra Ethernet leads on long-term vendor optionality, though it remains the least battle-tested of the four as of July 2026. Buyers evaluating this table should also weigh operational factors that resist tabular summary, such as the depth of in-house Ethernet versus InfiniBand administration expertise, the number of existing vendor relationships, and whether the cluster will ever need to share physical switches with non-AI, multi-tenant workloads, since InfiniBand's subnet-manager model is designed around single-tenant HPC assumptions that do not map cleanly onto multi-tenant public cloud economics.

Performance and Benchmarks

Independent, apples-to-apples benchmarking of InfiniBand against Ethernet-based fabrics is scarce, because most large training runs are proprietary and run on a single fabric rather than being repeated across two for comparison. The most concrete public independent test comes from World Wide Technology (WWT), which ran matched generative-AI and inference workloads across identical original equipment manufacturer (OEM) hardware, varying only the fabric. WWT found that "across generative tests and OEMs, the performance delta between InfiniBand and Ethernet was statistically insignificant (less than 0.03 percent)," that "Ethernet was faster than InfiniBand's best time in three out of nine generative tests," and that "in inference

tests, Ethernet averaged 1.0166 percent slower" than InfiniBand (Source: wwt.com). WWT frames the broader competitive picture as one in which "the ultra-high-performance domain (perhaps top 3-5 percent of the total market?) still belongs to InfiniBand," while "the vast majority of current InfiniBand deployments can actually be handled by Ethernet" (Source: wwt.com).

On the vendor side, NVIDIA's own Spectrum-X benchmarks claim a "1.6x" network performance improvement over off-the-shelf Ethernet and a target of sustaining "breakthrough 95% efficiency across deployments exceeding 100,000 GPUs" (Source: nvidia.com); these figures compare Spectrum-X specifically to generic, unoptimized Ethernet, not to InfiniBand, and should be read as a vendor claim rather than an independent measurement. Similarly, Oracle's Acceleron RoCE architecture is described in Oracle's own materials as providing "5 times higher network bandwidth performance" and "up to five times lower network latency compared to competitors" for its original Zettascale cluster (Source: blogs.oracle.com), reporting "as low as 2 µs (microseconds) latency" for a 131,072-GPU supercluster (Source: blogs.oracle.com), and Oracle's newer Zettascale10 architecture claims "sub-10 microsecond latency" across the fuller RDMA fabric (Source: blogs.oracle.com). These are competitor-facing claims from Oracle rather than third-party audits, and the "competitors" being compared against are not specified in a peer-reviewed setting.

On academic ground, a 2022 study by researchers from the University of Texas at Austin, Georgia Tech, and Meta documented that "the native RoCE congestion control scheme, based on Priority Flow Control (PFC), suffers from many drawbacks such as unfairness, head-of-line-blocking, and deadlock" in general datacenter environments, motivating the design of training-specific congestion control (Source: arxiv.org). A November 2025 benchmarking study by Qualcomm researchers concluded that "RoCE can achieve near-linear scaling performance comparable to InfiniBand when properly configured," while cautioning that InfiniBand "imposes significant infrastructure and operational costs" relative to "commodity Ethernet and RDMA" (Source: doi.org). A June 2026 AMD engineering analysis comparing the network traffic patterns of four production LLM training runs found that "while GPT-4 and DeepSeek-V2 predominantly leveraged InfiniBand for its low-latency guarantees, Llama 3 adopted a hybrid RoCE and IB approach, and Grok 4.0 fully embraced Ethernet-based RoCE through NVIDIA's Spectrum-X platform," concluding that despite the differing fabric choices, "the core traffic patterns ... remain consistent across models, rooted in shared distributed training frameworks such as PyTorch, JAX, and NCCL" (Source: rocm.blogs.amd.com), a useful reminder that fabric choice is a design decision layered on top of, not a substitute for, the collective-communication software stack. Read together, the independent and peer-reviewed evidence base points toward convergence: both fabrics can deliver comparable end-to-end training throughput at scale, provided the RoCEv2 deployment is competently engineered, while InfiniBand retains a smaller, largely latency-bound edge that matters most for the narrowest, most latency-sensitive HPC and frontier-training workloads.

Data Analysis and Evidence

The market data confirms the trend implied by the individual deployment case studies: InfiniBand still leads in dollar terms in some periods and in raw switch-hop latency, but Ethernet-family fabrics (RoCEv2, Spectrum-X, and emerging Ultra Ethernet) now carry the majority of unit volume and are forecast to carry the majority of long-run dollar growth in AI back-end networking.

Table 2 below places five of the named deployments discussed throughout this report side by side, making the relationship between fabric choice and achieved scale concrete before turning to the aggregate market figures that follow. Each figure in this table is drawn from, and separately cited at, the primary source discussed in the corresponding section above.

ORGANIZATION / DEPLOYMENT	NETWORK FABRIC	SCALE (GPUS OR CHIPS)	STATUS AS OF JULY 2026
Meta (Llama 3 training clusters)	RoCEv2 over Ethernet	24,576 GPUs per cluster	In production since 2024
xAI Colossus	NVIDIA Spectrum-X Ethernet	100,000 to 200,000 GPUs	In production, still expanding
Oracle OCI Zettascale10 / Stargate (with OpenAI)	Oracle Acceleron RoCE (RoCEv2)	Up to 800,000 GPUs targeted per deployment	Announced October 2025, deploying
Microsoft Azure ND H100 v5	NVIDIA Quantum-2 InfiniBand	Thousands of GPUs per deployment, 3.2 Tbps interconnect bandwidth per virtual machine	Generally available since August 2023
Google TPU v4 supercomputer	Custom optical circuit switching (neither InfiniBand nor Ethernet RoCE)	4,096 chips	In production since 2020

This side-by-side view underscores that scale alone does not predict fabric choice: the two largest deployments in the table, xAI's Colossus and Oracle's Zettascale10, both run Ethernet-based RDMA rather than InfiniBand, while the smallest and most tenant-constrained deployment, Azure's ND H100 v5, keeps InfiniBand as its default. Fabric choice appears to track tenancy model, in-house engineering capacity, and vendor relationship more closely than it tracks raw GPU count.

Dell'Oro Group, a market-research firm specializing in telecommunications and data-center networking, published its most recent quarterly snapshot on June 2, 2026, reporting that "Ethernet switch sales in AI back-end networks more than doubled and accounted for about two-thirds of data center switch sales in AI clusters during the first quarter 2026," even as "InfiniBand sales more than tripled during the quarter, supported by the ramp of 800 Gbps switches shipping with NVIDIA's Blackwell Ultra platform" (Source: delloro.com). Dell'Oro's analysts caution that this InfiniBand rebound may partly reflect "upgrades of the installed base in brownfield deployments rather than greenfield expansion" (Source: delloro.com), a distinction that matters for buyers trying to infer future architecture preference from quarterly sales alone. On the vendor-share side of the same quarter, Dell'Oro reported that "Celestica regained the leading position" in Ethernet AI back-end networks, "followed closely by NVIDIA," with "Arista ranked third" and "Cisco recorded the largest share gain during the quarter and ranked fourth" (Source: delloro.com). Also noteworthy: Dell'Oro reported that "800 Gbps switches accounted for the vast majority of the Ethernet switch shipments and revenues in AI back-end networks during the quarter," with "1600 Gbps switches" only beginning to sample and expected to "ramp in the second half of 2026" (Source: delloro.com), which places the current generation of both InfiniBand (Quantum-X800) and Ethernet (Spectrum-X, Tomahawk 5/6, Silicon One G200) switch silicon at roughly the same 800 Gb/s per-port speed tier for direct comparison.

Looking further out, Dell'Oro's February 4, 2026 report forecasts that "spending on data center switches deployed in AI back-end networks is forecast to surpass $100 billion by 2030," with "Ethernet ... projected to dominate both the scale-up and scale-out segments of the market" (Source: delloro.com). This is a marked shift from Dell'Oro's July 2024 report, which found that "InfiniBand is currently dominating the AI back-end network market" even while forecasting that "Ethernet is poised to take over soon," in a market then projected to "approach $80 B over the next five years" (Source: delloro.com).

Independent researcher 650 Group corroborates the same trajectory. Its January 2026 forecast states that "AI scale-out Ethernet revenue" will surge "to over $100B by 2030," and that, factoring optics, the total data-center-networking market is "poised to approach and potentially exceed $200B" a year by the end of the decade (Source: 650group.com). Separately, 650 Group's scale-up ("inside-the-rack") networking forecast projects PCIe/UALink switching revenue reaching "$3B+," Ethernet-based scale-up at "$8B+," and NVLink at "$25B+" by 2030, indicating that even within GPU-to-GPU scale-up fabrics, where NVIDIA's proprietary NVLink still dominates, Ethernet-based alternatives (branded Scale-Up Ethernet, or SUE) are gaining a material and growing share (Source: 650group.com). An earlier 650 Group forecast from June 2024 had projected the total data-center AI networking market to "surge to nearly $20B in 2025" (Source: wwt.com), illustrating how rapidly the addressable market has been revised upward within roughly 18 months.

On the silicon side, Broadcom's Tomahawk 5, the first 51.2 Tb/s Ethernet switch chip, was shown to replace "forty-eight Tomahawk 1 switches in the network, resulting in over 95 percent reduction in power requirements" for equivalent aggregate bandwidth (Source: investors.broadcom.com), a data point that underscores how much of the AI-networking cost and power story is now driven by switch-silicon generational improvement rather than by the InfiniBand-versus-Ethernet choice per se. Google's decade-long production experience with its Jupiter datacenter network fabric, presented at ACM SIGCOMM 2022 and covering both AI and general cloud workloads, similarly shows how much headroom exists in topology and switching improvements independent of link-layer choice: Google reports that "in this period Jupiter has delivered 5x higher speed and capacity, 30% reduction in capex, 41% reduction in power" through a shift from a traditional Clos fabric to a "direct-connect topology" using optical circuit switches (Source: gribble.org). On adoption breadth, the IBTA reported a "record-breaking Plugfest event that expanded interoperability testing" spanning "both InfiniBand and Ethernet RDMA-based products" in November 2025, a sign that the two ecosystems, while commercially distinct, are increasingly tested and validated together rather than as hermetically separate worlds (Source: infinibandta.org).

Case Studies and Real-World Examples

Meta: RoCEv2 Ethernet at 24,576-GPU Scale for Llama 3

Meta's GenAI infrastructure announcement of March 2024 disclosed "two 24k GPU clusters," used to train Llama 3, built "on top of Grand Teton, OpenRack, and PyTorch" and part of a broader 2024 roadmap that aimed to reach "350,000 NVIDIA H100 GPUs" by year end (Source: engineering.fb.com). Meta's follow-up SIGCOMM 2024 paper, presented at the conference in Sydney, Australia in August 2024, is one of the most detailed public engineering accounts of any hyperscale AI network, describing routing, transport, and operational lessons from years of running RoCEv2 at scale (Source: engineering.fb.com). It stands as the primary evidence that RoCEv2 Ethernet, properly engineered, can support frontier-scale training of a widely used open model family.

xAI Colossus: Spectrum-X Ethernet at 100,000-to-200,000-GPU Scale

xAI's Colossus supercomputer in Memphis, Tennessee, was built to train the Grok family of models and, according to NVIDIA, was constructed by xAI and NVIDIA "in just 122 days, instead of the typical timeframe for systems of this size that can take many months to years," with training beginning just "19 days from the time the first rack rolled onto the floor" (Source: nvidianews.nvidia.com). The Register reported the system's peak performance at "98.9 exaFLOPS of dense FP/BF16" before accounting for sparsity (Source: theregister.com). xAI's own site now cites the cluster at 200,000 H100 GPUs, "170PB/s" of aggregate memory bandwidth, and "2.8Tb/s" of per-server network bandwidth for distributed training (Source: x.ai). Colossus is the clearest public counterexample to the assumption that only InfiniBand can scale to six-figure GPU counts.

Oracle Cloud Infrastructure Zettascale10 / Stargate: RoCE at Multi-Gigawatt Scale

Oracle's OCI Zettascale10, announced October 14, 2025, is described by Oracle as "the largest AI supercomputer in the cloud," connecting "hundreds of thousands of NVIDIA GPUs across multiple data centers" for "up to an unprecedented 16 zettaFLOPS of peak performance," and it is "the fabric underpinning the flagship supercluster built in collaboration with OpenAI in Abilene, Texas, as part of Stargate" (Source: oracle.com). OpenAI's own vice president of Infrastructure and Industrial Compute, Peter Hoeschele, is quoted describing the "highly scalable custom RoCE design" as one that "maximizes fabric-wide performance at gigawatt scale while keeping most of the power focused on compute" (Source: oracle.com). This is a direct, named validation from a leading AI lab that RoCE-based Ethernet, not InfiniBand, can meet its requirements for one of the largest publicly disclosed AI training campuses in the world. Oracle also describes the underlying "wide, shallow, resilient fabric" as using "the GPU NIC as a mini-switch and connecting to multiple physically and logically isolated planes," a design intended to boost "scale while reducing network tiers, cost, and power" (Source: oracle.com).

Microsoft Azure ND H100 v5: InfiniBand as the Default for Cloud GPU Training

In contrast, Microsoft's ND H100 v5 series, generally available since August 2023, standardizes on InfiniBand rather than Ethernet for GPU-to-GPU communication. Microsoft documents that "ND H100 v5-based deployments can scale up to thousands of GPUs with 3.2 Tbps of interconnect bandwidth per VM," with "each GPU within the VM ... provided with its own dedicated, topology-agnostic 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand connection" (Source: learn.microsoft.com). OpenAI's then-president Greg Brockman is quoted describing the co-design relationship with Azure as "crucial for scaling our demanding AI training needs, making our research and alignment work on systems like ChatGPT possible" (Source:

azure.microsoft.com). This case illustrates that even as Ethernet-based fabrics gain unit-volume share elsewhere, InfiniBand remains the default architecture for a major hyperscaler's flagship GPU-cloud product line as of the most recent Microsoft Learn documentation update in April 2026 (Source: learn.microsoft.com).

Google Jupiter and TPU v4: A Non-InfiniBand, Non-Ethernet-RoCE Alternative Worth Noting

Not every large-scale AI cluster chooses between InfiniBand and Ethernet RoCE at all. Google's TPU v4 supercomputer, described in a peer-reviewed ISCA 2023 paper, uses optical circuit switches (OCSes) to dynamically reconfigure its interconnect topology and is explicitly benchmarked by its own authors as much cheaper, lower power, and faster than InfiniBand, with OCSes and underlying optical components representing less than 5 percent of system cost and less than 3 percent of system power, and the TPU v4 supercomputer, deployed since 2020 and outperforming its TPU v3 predecessor by 2.1x, scaled to 4,096 chips sharing 256 tebibytes (TiB) of high-bandwidth memory (Source: arxiv.org). The Jupiter network fabric that underlies Google's broader data center estate, including its AI and TPU infrastructure, achieved comparable generational gains through topology engineering rather than a link-layer switch, delivering, as noted above, "5x higher speed and capacity, 30% reduction in capex, 41% reduction in power" over a decade of production operation (Source: gribble.org). This case is included as a reminder that the InfiniBand/Ethernet binary, while dominant for GPU-based clusters, is not the only architecture in production at hyperscale; Google's custom silicon and optical interconnect approach is a genuine third path, albeit one unavailable to buyers who are not Google itself.

Implications and Future Directions

Several structural trends will shape which fabric buyers choose over the next several years. First, the network is proportionally a small but non-trivial share of total cluster cost: one vendor estimate places it at "roughly 5 to 8 percent of AI cluster total cost of ownership over five years," which means fabric choice, while consequential, is rarely the single largest line item in a build (Source: ipinfusion.com). Second, the boundary between "scale-up" (inside a rack, historically NVIDIA NVLink territory) and "scale-out" (between racks, the traditional InfiniBand/Ethernet battleground) is itself becoming contested, as Broadcom's Scale-Up Ethernet (SUE) initiative, now standardized as part of Ethernet for Scale Up Networking (ESUN) with backing from a "who's who group of hyperscalers and vendors" announced at the OCP Summit in October 2025, brings Ethernet-based transport into the scale-up domain that NVLink and AMD's Ultra Accelerator Link (UALink) have historically owned (Source: 650group.com).

Third, co-packaged optics (CPO), which integrate optical interfaces directly with switch silicon rather than relying on pluggable transceivers, is expected to reshape both InfiniBand's and Ethernet's power and density profiles; 650 Group projects that "pluggable optics at 800G and 1.6T, plus emerging CPO, could exceed 50% of the annual switch revenue by 2030" (Source: 650group.com), and NVIDIA has already begun shipping "co-packaged silicon photonic networking switches" designed "to scale to millions of GPUs" and "multi-site AI factories" on both its Quantum-X800 InfiniBand and Spectrum-X Ethernet lines (Source: nvidia.com). Fourth, the frontier of scale is shifting from single-building clusters to multi-site "AI super-factories": NVIDIA's Spectrum-XGS Ethernet, explicitly positioned as a "third pillar of AI computing beyond scale-up and scale-out," anticipates that individual data centers are "reaching the limits of power and capacity within a single facility" and that future frontier training runs will span multiple physically separate sites (Source: nvidianews.nvidia.com). Oracle's Zettascale10, targeting multi-gigawatt data center campuses and up to 800,000 GPUs across the Stargate program, is the clearest existing evidence of this shift already occurring in production.

Fifth, buyers should expect networking teams to need broader, not narrower, skill sets over the next several years, since the largest operators increasingly run mixed fleets rather than standardizing on one fabric. CoreWeave, for example, documents that it maintains "InfiniBand and RoCE labels" across its fleet with a shared, "fabric-agnostic schema" so that workload placement and monitoring tooling work identically regardless of which fabric a given node pool uses (Source: docs.coreweave.com), an operating model that treats fabric choice as a per-cluster procurement decision rather than a company-wide architectural commitment. Finally, the maturation of the Ultra Ethernet Consortium's specification and first compliant silicon (AMD's Pollara and Vulcano NICs, Broadcom's Thor Ultra NIC) over 2025 and 2026 suggests that within the next product cycle or two, buyers will be able to build genuinely multi-vendor, InfiniBand-competitive Ethernet fabrics without depending on any single company's matched-hardware platform, a development that, if it plays out as the UEC's founding members intend, would narrow the remaining rationale for choosing single-vendor InfiniBand to the smallest, most latency-critical segment of the market that WWT and other observers already describe as "top 3-5 percent" of deployments (Source: wwt.com).

Frequently Asked Questions (FAQs)

Is InfiniBand better than Ethernet for AI training? InfiniBand offers a lower and more deterministic latency floor, roughly 200 to 300 nanoseconds per switch hop against a wider and more variable range for Ethernet, but independent testing by WWT found the end-to-end performance delta on real generative-AI and inference workloads to be under 1 to 1.02 percent in most cases (Source: wwt.com). "Better" depends on whether the extra nanoseconds and mature tooling justify InfiniBand's single-vendor cost premium for a given cluster's scale and tenancy model.

What is NVIDIA Spectrum-X networking? Spectrum-X is NVIDIA's commercial Ethernet platform for AI, pairing NVIDIA Spectrum switches with NVIDIA SuperNICs and proprietary RoCE extensions to claim "1.6x" performance over generic Ethernet (Source: nvidia.com); it powers xAI's 200,000-GPU Colossus cluster (Source: x.ai).

Is RoCE the same as InfiniBand? No. RoCEv2 runs InfiniBand's RDMA transport semantics over standard Ethernet using UDP/IP encapsulation, whereas InfiniBand uses its own dedicated link layer with native credit-based flow control (Source: factryze.ai).

What is the best network fabric for GPU clusters? There is no single answer independent of scale and organizational context. Deployments in the narrowest, most latency-sensitive tier of scale still favor InfiniBand, while Meta, xAI, and Oracle's OpenAI-partnered Stargate site demonstrate that RoCEv2-based Ethernet fabrics now support some of the largest training clusters in the world.

Is the Ultra Ethernet Consortium meant to replace InfiniBand? UEC's own founding framing is to standardize an "Ethernet-based, open and interoperable, high-performance, full-communications stack architecture" for AI and HPC, explicitly to avoid the vendor lock-in associated with single-vendor fabrics (Source: linuxfoundation.org), which functionally positions it as a long-term, multi-vendor alternative to InfiniBand rather than a modification of InfiniBand itself.

How much does an AI network fabric cost relative to the rest of the cluster? Estimates place networking at roughly 5 to 8 percent of five-year total cost of ownership for an AI cluster, with InfiniBand carrying, per one Ethernet-hardware vendor's estimate, a 30 to 60 percent premium over open-hardware Ethernet for equivalent capacity (Source: ipinfusion.com); this figure should be treated as indicative rather than audited.

Does AWS use InfiniBand or Ethernet for AI training? AWS's Elastic Fabric Adapter runs over the company's own Ethernet-based Scalable Reliable Datagram transport rather than native InfiniBand, and AWS documents that it "supports RDMA (Remote Direct Memory Access) write on most supported instance types that have Nitro version 4 and later," integrating with NCCL and MPI so that customers do not need to manage InfiniBand-specific tooling directly (Source: docs.aws.amazon.com).

Can InfiniBand and Ethernet coexist in the same organization? Yes, and this is increasingly the norm rather than the exception among GPU cloud providers; CoreWeave, for instance, operates both fabrics side by side, applying a shared labeling schema so that InfiniBand and RoCE node pools are managed with the same tooling (Source: docs.coreweave.com).

Conclusion

The choice among InfiniBand, NVIDIA Spectrum-X, RoCEv2, and standards-based or Ultra Ethernet for an AI cluster is not a single binary decision but a layered one: link layer (InfiniBand versus Ethernet), then, if Ethernet, whose matched-hardware platform (NVIDIA's Spectrum-X) versus whose open, multi-vendor stack (generic RoCEv2 today, Ultra Ethernet increasingly going forward). InfiniBand remains the technically purest, lowest-latency option, and continues to anchor major cloud GPU products such as Microsoft Azure's ND H100 v5 series and a large share of the world's highest-performing public supercomputers. But the market has clearly moved. Ethernet-based fabrics, whether NVIDIA's proprietary Spectrum-X or the increasingly standardized RoCEv2 and Ultra Ethernet stacks, now carry roughly two-thirds of AI back-end switch unit sales as of the first quarter of 2026, underpin frontier training clusters at Meta, xAI, and Oracle's Stargate site with OpenAI, and are forecast by independent analysts to dominate long-run dollar growth in AI networking spending through the end of the decade.

For organizations building or buying AI cluster infrastructure, the practical framework that emerges from this evidence is straightforward. Choose InfiniBand when the workload sits in the narrow, latency-critical tier that WWT estimates at roughly the top 3 to 5 percent of deployments, when the cluster is single-tenant, and when a single-vendor operational model is acceptable in exchange for a mature, deterministic fabric. Choose NVIDIA Spectrum-X when the buyer wants Ethernet's routability and multi-tenancy with NVIDIA's tuned, matched-hardware performance uplift, and is comfortable staying within NVIDIA's switch-and-SuperNIC ecosystem. Choose generic RoCEv2 Ethernet, following the model Meta has published in detail, when in-house Ethernet operations expertise exists and the organization is prepared to invest in the congestion-control and routing tuning that hyperscale RoCEv2 deployments require. And watch Ultra Ethernet closely: with its 1.0 specification ratified in mid-2025, updated through version 1.0.2 in January 2026, and its first compliant NICs from AMD and Broadcom already shipping to a launch customer in Oracle, it is the technology most likely to determine whether the AI networking market of the early 2030s looks like today's InfiniBand-dominated HPC world, today's NVIDIA-dominated Spectrum-X world, or a genuinely open, multi-vendor Ethernet market that the UEC's founding members explicitly set out to build.

Tags: infiniband, spectrum-x, roce, rocev2, ethernet, ai cluster networking, ultra ethernet consortium, gpu networking, rdma, nvidia networking

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will GPUSmith or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.