AvidThink

Data Center Networking in 2026

Scaling to New Heights: Up, Out, and Across

RESEARCH BRIEF

Table of Contents

About AvidThink

AvidThink is a research and analysis firm focused on providing cutting-edge insights into the latest in infrastructure technologies. AvidThink's coverage includes cloud and AI infrastructure, 5G and mobile network infrastructure, enterprise and data center networks, private wireless, edge computing, SD-WAN, SASE, ZTNA, and infrastructure security. Our clients include Fortune 500 enterprises, hyperscalers, tier-1 communications service providers, as well as innovative startups. AvidThink's research has been quoted by Forbes, the Wall Street Journal, Light Reading, Fierce Networks, Mobile World Live, and other major publications. Visit AvidThink at www.avidthink.com.

Research Briefs are independent content created by analysts working for AvidThink LLC. These reports are made possible through the sponsorship of our commercial supporters. Sponsors do not have any editorial control over the report content, and the views represented herein are solely those of AvidThink LLC. For more information about report sponsorship, please reach out to us at research@avidthink.com.

RESEARCH BRIEF

Data Center Networking in 2026

Scaling to New Heights: Up, Out, and Across

Executive Summary

AI workloads are fueling a massive expansion of data center infrastructure. Leading technology companies are projected to spend over $630 billion on AI infrastructure in 2026. A key component of this growth is data center networking, which enables AI cluster scaling from thousands to hundreds of thousands of GPUs.

Five transitions define the evolution of data center networking for AI:

1. Scale-up interconnects are growing quickly in size and importance. NVIDIA's NVLink is advancing within high-bandwidth domains, expanding toward rack-level and multi-rack setups with over 1,000 GPUs on its 2028 roadmap. Open alternatives such as UALink and Ethernet-based options are moving from specification to announced hardware, with production in late 2026 and into 2027.
2. Ethernet is the dominant choice for new scale-out AI fabric builds. By mid-2025, it exceeded InfiniBand in new deployments. Cost, broad ecosystem, and multi-vendor support drove this shift. Loss-tolerant, multipath Ethernet fabrics now approach InfiniBand-like efficiency at lower cost, though InfiniBand retains relevance in tightly coupled environments.
3. The optical interconnect landscape is undergoing a structural shift. Traditional pluggable optics are being supplemented by new architectures, including linear pluggable optics (LPO), near-packaged optics (NPO), and co-packaged optics (CPO). The XPO Multi-Source Agreement, announced at OFC 2026, defines a liquid-cooled pluggable that could bridge the gap between conventional pluggables and CPO.
4. The software stack continues to drive system performance. Advances in collective communication libraries (Meta's NCCLX), distributed training frameworks (PyTorch FSDP2, Megatron-LM), fault-tolerant training (TorchFT, TorchPass), and better inference runtimes all play a significant role.
5. Disaggregated inference — separating compute-bound prefill from memory-bandwidth-bound decode — is critical to scaling. NVIDIA's integration of Groq LPUs into Vera Rubin (GTC 2026) pairs purpose-built silicon for each phase, introducing new latency-sensitive traffic patterns that are shifting fabric design assumptions.

One through line across these changes is extreme co-design. Compute, networking, memory, and software are now architected as integrated platforms, bringing measurable gains but increasing ecosystem lock-in. The main bottleneck has also shifted from bandwidth to power. AI accelerators may soon exceed 4 kW per device, and US data centers could consume 12% of national electricity within the decade, driving investment in power-efficient optics, liquid cooling, and multi-campus (scale-across) architectures.

Building on these shifts, principal findings include: multi-vendor scale-out competition at 102.4 Tbps (Broadcom, Cisco, NVIDIA, and Marvell), creating increased buyer leverage; SONiC adoption on the rise; fault-tolerant training, saving hundreds of millions annually at scale; and a multi-front optics form factor competition at varying maturity levels. Some of these elements are production-proven in 2026; others remain announced, sampling, or roadmap items.

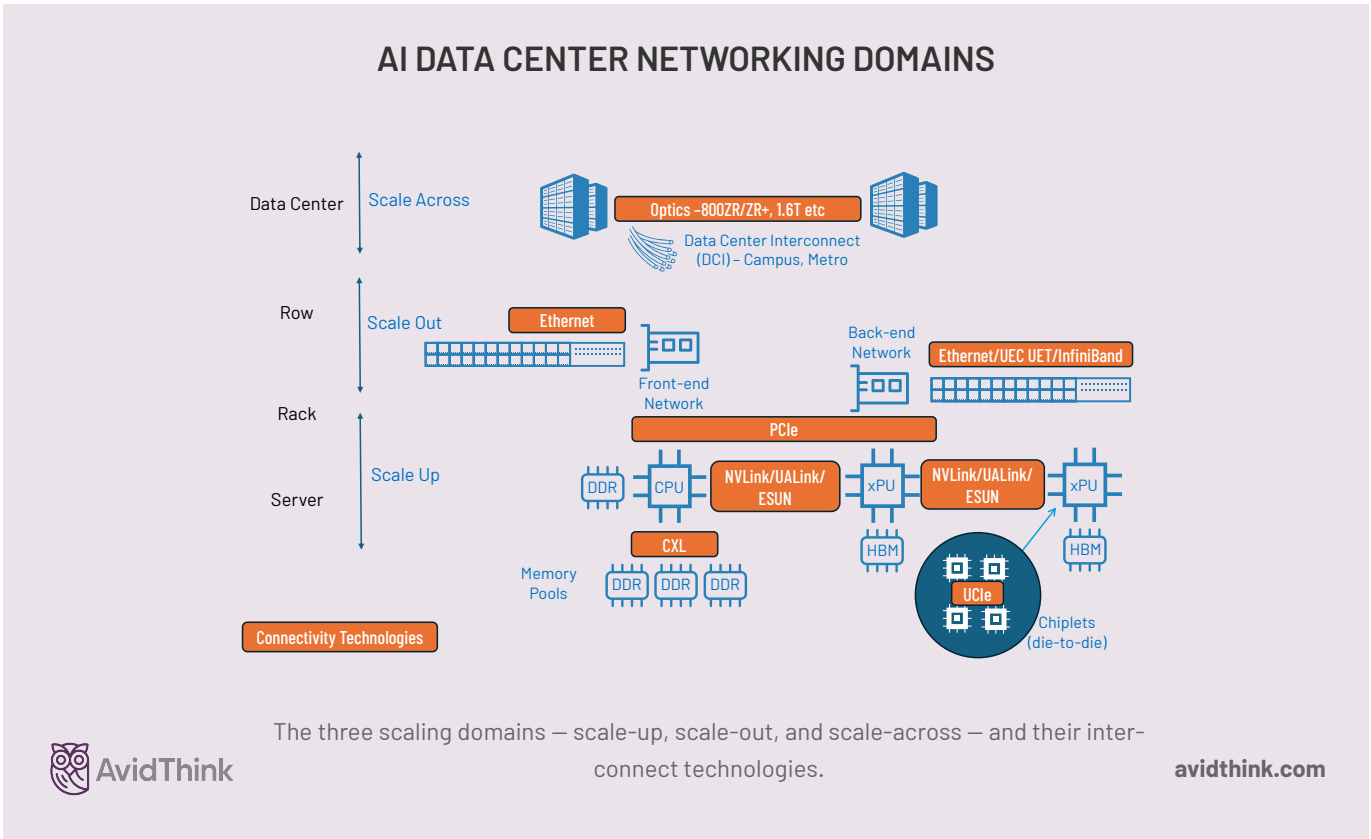
Introduction — AI Workloads and their Networking Demands

Capital spending on AI infrastructure has hit new highs, with leading technology companies projected to invest over $630 billion in 2026.[1] This boom is driven by large language models, which need far more computing and networking than older deep learning systems. One of the current leading beneficiaries of this boom, NVIDIA, anticipates $1 trillion in GPU orders through 2027.[2]

Connectivity remains a key component in the AI data center. Moving from scale-up within the rack to scale-out across racks, and ultimately to scale-across clusters of adjacent data centers, the underlying network shapes the performance of these "AI factories" that consume and produce tokens. The network is no longer one design problem; it is three coupled design problems, each with different physics, economics, and operational constraints.

While this report draws extensively on hyperscaler deployments — Meta's 129,000-GPU clusters, xAI's Colossus, Google's TPU fabric — the analysis is directed at data center operators, cloud service providers, and enterprises building clusters at the scale of 1,000 to 10,000+ GPUs for training, fine-tuning, or inference. We've seen that technologies that are cutting-edge at a frontier lab can trickle down to the enterprise 12–18 months later. And where hyperscaler innovations are relevant to smaller-scale deployments, we try to highlight a bridging path.

In this 2026 edition of our data center networking report, we examine the advances and challenges across these three network domains, along with the latest in communication libraries, robustness, and fault tolerance. Before delving into each networking domain, we first outline the AI training and inference workloads that fundamentally impact network traffic patterns.



The three scaling domains — scale-up, scale-out, and scale-across — and their inter-connect technologies.

[1] K. Kwok, "How Big Tech's $630 billion AI splurge will fall short," *Reuters*, 26 Mar. 2026. [Online]. Available: https://www.reuters.com/commentary/breakingviews/how-big-techs-630-bln-ai-splurge-will-fall-short-2026-03-26/

[2] NVIDIA, "NVIDIA GTC 2026 Keynote," J. Huang, 16 Mar. 2026. [Online]. Available: https://blogs.nvidia.com/blog/gtc-2026-news/

Training: Communication Patterns at Scale

A large transformer-based model with a trillion parameters, trained at scale, creates sustained communications. These push traditional networks to their limits. For example, for a workload spread across 100,000 GPUs using **tensor parallelism**, each forward pass produces AllReduce operations on hidden states. One transformer layer may need collective operations involving tens of thousands of devices, and the cross-device traffic can total petabytes per training day.

Training traffic depends on methodology:

- **Data-parallel** training creates coarse, synchronized AllReduce collectives.
- **Tensor-parallel** training requires fine-grained, latency-sensitive AllReduce operations that finish in microseconds to remain efficient.
- **Pipeline parallelism** adds point-to-point bandwidth between ranks.
- **Expert parallelism** in mixture-of-experts (MoE) architectures forms sparse AllToAll communication, with tokens sent to experts based on learned routing.
- PyTorch's **fully-sharded data parallelism** (FSDP), used for large models, manages model states using AllGather and ReduceScatter, reducing memory use but increasing network traffic.

Whatever the method, networks must deliver low latency, no packet loss, and handle irregular communication patterns efficiently.

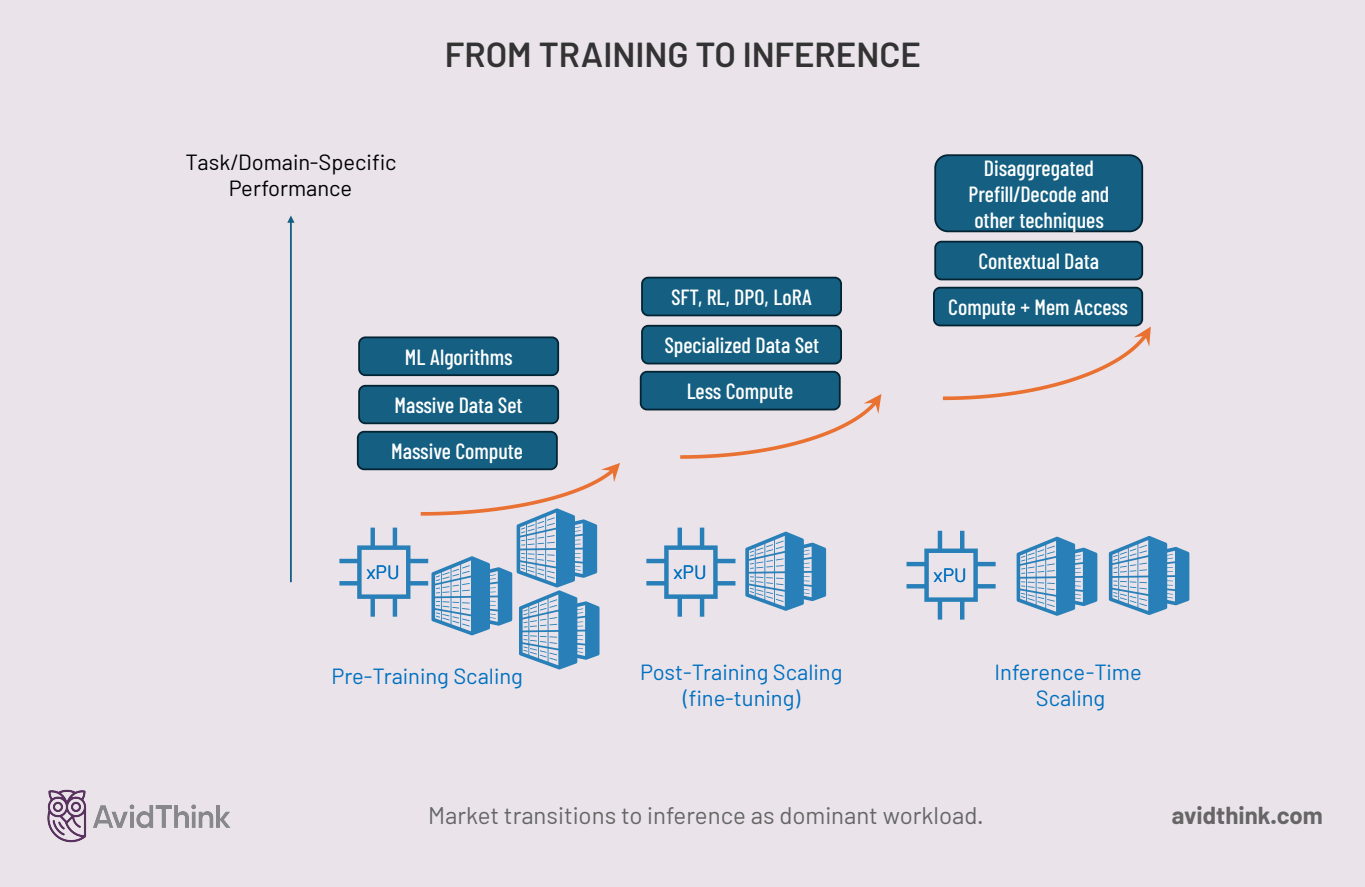
Model FLOPs Utilization remains stalled at 35–40% in many deployments, with networking overhead among the primary constraints alongside software overheads, kernel efficiency, and stragglers.

Bandwidth and latency budgets for training are tighter now. In large tensor-parallel training, the compute-to-communication ratio for AllReduce operations is low, so networking greatly affects throughput. A 10% reduction in collective latency can significantly shorten job completion times. Training runs can consume months of compute and hundreds of millions in capital, so even small gains matter. Meta's Llama 3 training had over 400 interruptions in just 54 days.[3] In Alibaba's testing, they found that communication failures (mainly congestion-related) were a frequent and persistent problem (in a sample of 107 jobs, 43 experienced slow communication).[4] These highlight network reliability challenges in scaled AI training. Unfortunately, model FLOPs Utilization remains stuck at 35–40% in many deployments due to networking (and other) overhead.

Inference: A Distinct Networking Challenge

Inference workloads add a different dimension of complexity. Unlike training, where batching and synchronization are the norm, inference requires serving individual or small-batch user requests with strict latency targets. An LLM inference engine must generate tokens sequentially, meaning the latency of a single AllGather operation translates directly to milliseconds of user-facing delay.

[3] A. Grattafiori et al., "The Llama 3 Herd of Models," *arXiv*, 31 Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
[4] T. Wu et al., "FALCON: Pinpointing and Mitigating Stragglers for Large-Scale Hybrid-Parallel Training," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2410.12588



Two structural trends are altering the requirements for inference networking:

- The first is the continuing rise of test-time compute and reasoning models. Chain-of-thought models, starting with OpenAI's o1, o3, DeepSeek-R1, and many of today's frontier models, shift substantial compute from training to inference, generating far longer output sequences (10–100x the input length) with irregular, data-dependent parallelism that differs from training's predictable traffic profiles. These models generate bursty, asymmetric traffic patterns in which the ratio of compute to communication varies dynamically with the complexity of the reasoning task. Network architects must accommodate workloads where a single inference request may trigger thousands of tokens of internal reasoning before producing a response.
- The second is disaggregated inference, where the prefill phase (processing the input prompt) and the decode phase (generating output tokens) execute on separate XPU pools optimized for their particular computational profiles. The prefill phase is compute-bound and can efficiently utilize high-throughput XPUs; the decode phase is memory-bandwidth-bound and benefits from different hardware characteristics. This disaggregation creates a new networking bottleneck: the transfer of KV caches between the prefill and decode pools.

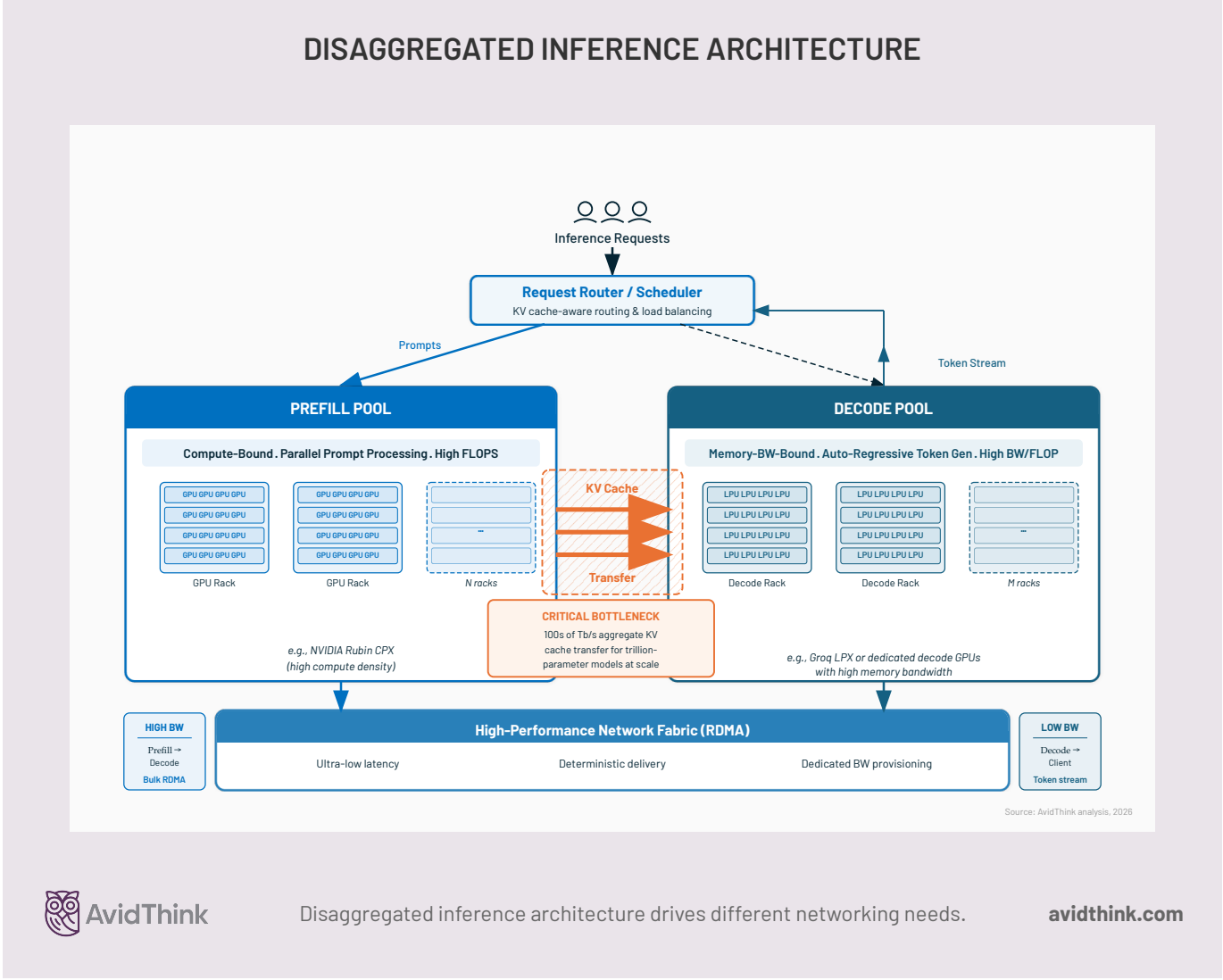
As a reminder, many inferencing use cases will not require multi-GPU clusters. Often, limited task- and domain-specific use cases can use models that fit within a single or few GPUs. However, the current wave of sophisticated coding agents, co-working agents, autonomous agentic workflows tend to benefit from the increased capabilities that large frontier models provide.

Dedicated Inference Hardware and Disaggregated Architectures

The disaggregated inference architecture — separating the compute-bound prefill phase from the memory-bandwidth-bound decode phase — has moved from academic concept to production deployment, with multiple operators running prefill/decode

architectures and software like the open-source llm-d project supporting the same. At GTC 2026, NVIDIA announced the integration of Groq 3 LPUs (Language Processing Units) into the Vera Rubin platform[5], targeting a purpose-built disaggregated inference system with dedicated silicon for each phase (shipping H2 2026). The Groq 3 LPX rack houses 256 LPUs delivering 128 GB of SRAM with 40 PB/s of memory bandwidth across the full rack. In this architecture, Rubin GPU racks handle prefill (compute-intensive prompt processing), while LPX racks handle decode (token generation). NVIDIA claims the LPX platform provides 35x higher throughput (per unit power) than GB 200 NVL72 for large-model inference.[6]

The architectural split has ramifications for the underlying network. The KV cache transfer between prefill and decode pools becomes the critical interconnect bottleneck: for a trillion-parameter model serving thousands of concurrent requests, the aggregate KV cache transfer bandwidth requirement can reach hundreds of terabits per second. The prefill-to-decode handoff requires ultra-low-latency RDMA with deterministic delivery guarantees, as any variation in transfer time directly affects time-to-first-token latency. Network architects must provision dedicated bandwidth between prefill and decode pools.



AvidThink Disaggregated inference architecture drives different networking needs. avidthink.com

[5]NVIDIA, "Vera Rubin Opens Agentic AI Frontier," *NVIDIA Newsroom*, 16 Mar. 2026. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-vera-rubin-platform

[6]"GTC 2026: With Groq 3 LPX, Nvidia adds dedicated inference hardware," *The Decoder*, 17 Mar. 2026. [Online]. Available: https://the-decoder.com/gtc-2026-with-groq-3-lpx-nvidia-adds-dedicated-inference-hardware-to-its-platform-for-the-first-time/

The existence of purpose-built decode hardware changes the economics of fabric design. When training clusters demand symmetric bisection bandwidth (every GPU communicates with every other GPU), disaggregated inference clusters exhibit asymmetric traffic patterns: high-bandwidth, low-latency transfers from prefill to decode pools, followed by lightweight token-streaming traffic from decode pools to clients. This asymmetry may favor topology designs that over-provision specific paths rather than providing uniform bisection bandwidth — a departure from the fat-tree and rail-optimized architectures optimized for training.

The coexistence of training and inference — now with dedicated decode hardware, adding a third traffic profile — creates a multidimensional fabric-optimization problem. The emerging consensus favors unified fabrics with workload-aware scheduling and QoS differentiation over separate networks, though the engineering complexity should not be underestimated.

Software Stack: Training Frameworks, Collectives, and Inference Runtimes

Before examining each networking domain, we outline the software stack that generates and shapes network traffic. The collective operations, parallelism strategies, and runtime decisions described here directly determine the traffic patterns that the fabric must support.

Networking Libraries and Communication Collectives

Collective communication libraries define how distributed workloads utilize the network, translating operations such as AllReduce, AllGather, ReduceScatter, and AllToAll into concrete traffic patterns throughout the fabric.

- **NCCL** (NVIDIA Collective Communications Library) has evolved into the de facto standard for GPU collective operations. Recent versions (v2.27–v2.29) have added Symmetric Memory support (7.6x latency decrease for small-message AllReduce)[7], SHARP in-network reduction (10–20% gains per NVIDIA), GPU-Initiated Networking that eliminates CPU-GPU synchronization latency, and a Put API providing zero-SM-overhead one-sided communication optimized for Blackwell.[8] The networking implication: each NCCL generation shifts more traffic initiation to the GPU, reducing host-side bottlenecks but increasing demands on NIC and fabric responsiveness.
- Meta's **NCCLX** extends NCCL with production-hardened optimizations at 100,000+ GPU scale. Evaluated on Llama 4, NCCLX reduced per-step training latency by up to 12%, accelerated startup time by up to 11x at 96,000 GPUs, and reduced decode latency by 15–80%.[9] Meta has open-sourced NCCLX alongside TorchComms, a lightweight PyTorch Distributed communication API with NCCLX as its production backend.
- At a lower level, **NVSHMEM** provides a Partitioned Global Address Space model where GPU kernels perform put/get operations on remote GPU memory without collective synchronization — well-suited to the sparse AllToAll patterns in MoE models.[10]
- Meanwhile, **AMD's RCCL** seeks to maintain feature parity with NCCL; RCCL v2.28.3 integrates MSCCL for hardware-aware MoE AllToAll optimization, delivering 15–25% throughput improvement on MI300X.[11] And AMD's Pensando Pollara NIC and Salina DPU provide a vertically integrated alternative to NVIDIA's ConnectX/BlueField ecosystem.

The mapping of collective algorithms to physical topology has become a critical design consideration. AllReduce — the workhorse of data-parallel training — generates symmetric, many-to-many traffic patterns that benefit from fat-tree topologies with full bisection bandwidth. AllToAll — dominant in MoE architectures — creates sparse, point-to-point traffic that can cause

[7] NVIDIA, "NCCL v2.27 Release Notes," *NVIDIA Developer*, 2025. [Online]. Available: https://docs.nvidia.com/deeplearning/nccl/release-notes/index.html
[8] NVIDIA, "NCCL v2.28/v2.29 Release Notes," *NVIDIA Developer*, 2025. [Online]. Available: https://docs.nvidia.com/deeplearning/nccl/release-notes/index.html
[9] M. Si et al., "Collective Communication for 100k+ GPUs," arXiv preprint, arXiv:2510.20171, Oct. 2025. [Online]. Available: https://arxiv.org/abs/2510.20171.
[10] NVIDIA, "NVSHMEM 2.0 Documentation," *NVIDIA Developer*, 2025. [Online]. Available: https://developer.nvidia.com/nvshmem
[11] AMD, "RCCL v2.28.3 Release Notes," *AMD ROCm*, 2025. [Online]. Available: https://rocm.docs.amd.com

congestion on rail-optimized topologies. ReduceScatter and AllGather, used in FSDP training, generate asymmetric traffic patterns, with different phases of training creating distinct bandwidth demands. Network architects must account for those mappings to avoid overprovisioning some paths while starving others.

Pre-Training Frameworks and Their Impact on Networking

Popular training framework **PyTorch's** FSDP2 introduces Hybrid Sharded Data Parallel, combining data parallelism across nodes with FSDP sharding within nodes — reducing inter-node bandwidth demand while improving memory efficiency.[12] Other pipeline parallelism innovations (PipeOffload[13], ZeroBubble, DualPipe, TD-Pipe[14]) overlap forward passes, backward passes, and communication to deliver 5–10% end-to-end training throughput gains at 96,000 GPUs in frameworks such as Megatron-LM.[15]

At the networking level, **Megatron-LM** tensor parallelism requires high-bandwidth, low-latency intra-node connections; **DeepSpeed's** ZeRO-3 distributes communication more uniformly across the scale-out fabric. Research from PyTorchCon 2025 showed that combining FSDP communication overlap with SHARP-based in-network reduction and multicast-accelerated AllGather reduced backward execution time by 10.7%.[16]

Inference Runtimes and Their Network Demands

vLLM and **SGLang** have become the dominant open-source inference engines, each generating network traffic patterns distinct from training. vLLM's PagedAttention eliminates 60–80% of KV cache memory waste, while its continuous batching creates bursty traffic requiring coordination across inference servers.[17] SGLang's RadixAttention achieves 50–99% cache hit rates in production, with deployments spanning 400,000+ GPUs worldwide.[18]

Both runtimes generate token-by-token, sequential communication with millisecond-level latency budgets. As inference scales to exceed training by orders of magnitude and the prefill/decode split introduces new intra-fabric traffic patterns, fabrics must adapt from training-optimized symmetric designs to accommodate these asymmetric, latency-sensitive workloads.

In-Network Computing — Offloading XPUs

In-network computing — using programmable switching fabric to perform collective operations — was historically an InfiniBand exclusive. The trajectory in 2025 moved toward Ethernet-native support, though broad production ecosystem deployment is still emerging. NVIDIA's SHARP v3 introduces multi-tenant collectives with 10–20% aggregate throughput improvement and Broadcom's Tomahawk Ultra performs operations in switch buffer memory without XPU overhead.[19]

UEC 1.0 defines transport and congestion-management capabilities such as packet trimming and atomic-operation support, while broader in-network collectives remain future or optional extensions. The convergence of in-network computing with FSDP optimization points to a future where framework-level collective libraries and switch-level acceleration are co-designed rather than independently optimized.

[12] Y. Zhao et al., "PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel," *Proc. VLDB Endowment*, vol. 16, no. 12, pp. 3848–3860, 2023. [Online]. Available: https://dl.acm.org/doi/10.14778/3611540.3611569

[13] "PipeOffload: Improving Scalability of Pipeline Parallelism with Memory Optimization," *arXiv*, 2025. [Online]. Available: https://arxiv.org/html/2503.01328v1

[14] DeepSeek-AI, "DeepSeek-V3 Technical Report," *arXiv*, Dec. 2024. [Online]. Available: https://arxiv.org/pdf/2412.19437

[15] NVIDIA, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," *GitHub*, 2025. [Online]. Available: https://github.com/NVIDIA/Megatron-LM

[16] N. Tateiwa, "FSDP Communication Optimization with In-Network Computing," presented at *PyTorchCon 2025*, 2025.

[17] vLLM Project, "vLLM: Easy, Fast, and Cheap LLM Serving," *vLLM Documentation*, 2025. [Online]. Available: https://docs.vllm.ai/

[18] SGLang Project, "SGLang: Efficient Execution of Structured Language Model Programs," *SGLang Documentation*, 2025. [Online]. Available: https://sgl-project.github.io/

[19] "Broadcom Ships Tomahawk Ultra: Reimagining the Ethernet Switch for HPC and AI Scale-up | Broadcom Inc." Broadcom Inc., 2025, https://investors.broadcom.com/news-releases/news-release-details/broadcom-ships-tomahawk-ultra-reimagining-ethernet-switch-hpc.

Scale-Up Networking: Raising the Ceiling

Scale-up interconnects provide ultra-high-bandwidth, low-latency connectivity between accelerators within a node, rack, or tightly coupled cluster — typically spanning 8 to 1,000+ devices. This domain is transitioning from a single-vendor model to a competitive, multi-standard landscape. Three distinct approaches — NVIDIA's NVLink, UALink, and Ethernet-based scale-up (ESUN/SUE-T), complemented by CXL for memory operations — are competing to define how the next generation of AI clusters connect their accelerators.

NVLink: The Incumbent

NVIDIA's NVLink architecture dominates scale-up with a multi-generational lead. NVLink 5, shipping in volume with the Blackwell architecture, provides 1.8 TB/s of aggregate bandwidth per GPU. The GB200 NVL72 configuration aggregates 72 Blackwell GPUs with 130 TB/s of total NVLink bandwidth through a specialized switch fabric, creating a coherent compute unit capable of executing substantial training workloads.[20]

NVLink 6, now in production with the Vera Rubin platform (GTC 2026), doubles per-GPU bandwidth to 3.6 TB/s, with the Vera Rubin NVL72 delivering 3.6 exaflops of NVFP4 inferencing and 260 TB/s aggregate NVLink 6 bandwidth across 72 GPUs — representing a 2x bandwidth increase over the Blackwell NVL72. Last year, NVIDIA introduced NVLink Fusion as a licensable chiplet-interface program that allows qualified partners to incorporate NVLink connectivity, rather than as an open industry standard. Partners committed to NVLink Fusion include SiFive (RISC-V processors), AWS (Trainium4), Fujitsu, Qualcomm, Marvell, MediaTek, and Astera Labs.[21] NVLink Fusion represents a strategic expansion of NVIDIA's ecosystem, though questions remain about whether third-party implementations will achieve performance parity with NVIDIA's integrated designs. Nevertheless, Marvell received an investment of $2B from NVIDIA in March 2026, with NVLink Fusion as a pillar in that relationship.[22]

The most significant scale-up announcement at GTC 2026 was the Kyber rack architecture. Kyber uses a vertical-insertion, cable-free design with an NVLink backplane at the rear. It will provide customers with three options for Vera Rubin Ultra NVLink scale-up domains: NVL72, NVL144, and the flagship NVL576.

NVIDIA's Feynman architecture (2028 roadmap, not yet in production) extends the scale-up roadmap further with NVLink 8, both copper and co-packaged optical scale-up (NVLink 8 CPO brings silicon photonics to the scale-up interconnect for the first time), and a projected 204.8T Spectrum-7 switch for scale-out. For Feynman, eight Kyber racks combine to form the NVL1152, connecting 1,152 GPUs in a single NVLink domain via direct optical interconnects for rack-to-rack scale-up. This will represent a 16x increase in NVLink domain size compared to the original NVL72, blurring the boundary between scale-up and scale-out: workloads that previously required scale-out fabric traversal can execute within the scale-up domain. The Feynman generation targets the "angstrom era" (1.6 nm) and signals NVIDIA's intent to maintain its scale-up leadership through at least 2029.[23]

UALink: The Open Alternative

The UALink Consortium released UALink 1.0 in April 2025, with promoter members including AMD, Astera Labs, AWS, Cisco, Google, HPE, Intel, Meta, and Microsoft. The specification supports custom memory-semantic protocols optimized for accelerator-to-accelerator communication with ultra-low latency. Notably UALink leverages the Ethernet PHY at the physical layer.

[20] NVIDIA, "NVIDIA GB200 NVL72," *NVIDIA Data Center*, 2025. [Online]. Available: https://nvidia.com/en-us/data-center/gb200-nvl72/
[21] NVIDIA, "NVIDIA Unveils NVLink Fusion for Industry to Build Semi-Custom AI Infrastructure with NVIDIA Partner Ecosystem," *NVIDIA Investor Relations*, 19 May 2025. [Online]. Available: https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Unveils-NVLink-Fusion-for-Industry-to-Build-Semi-Custom-AI-Infrastructure-With-NVIDIA-Partner-Ecosystem/default.aspx
[22] NVIDIA, "NVIDIA AI Ecosystem Expands as Marvell Joins Forces through NVLink Fusion," *NVIDIA Newsroom*, 2026. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-ai-ecosystem-expands-as-marvell-joins-forces-through-nvlink-fusion/
[23] A. Shilov, "Nvidia updates data center roadmap with Rosa CPU and stacked Feynman GPUs — optical NVLink, Groq LPUs with NVFP4, and NVLink also on deck," Tom's Hardware, Mar. 17, 2026. [Online]. Available: https://www.tomshardware.com/pc-components/gpus/nvidia-updates-data-center-roadmap-with-rosa-cpu-and-stacked-feynman-gpus-optical-nvlink-groq-lpus-with-nvfp4-and-nvlink-also-on-deck

The UALink consortium has enrolled over 115 member companies. Critically, UALink has moved from specification to announced hardware: AMD's Helios rack architecture, unveiled at CES 2026, integrates up to 72 MI400-series GPUs with 260 TB/s of scale-up bandwidth via UALink — matching NVIDIA's Vera Rubin NVL72 aggregate bandwidth. Helios pairs AMD EPYC Venice CPUs with Pensando Vulcano AI NICs and supports both native UALink and "UALink over Ethernet" for customers preferring Ethernet compatibility. Oracle has committed to being the first hyperscaler to offer publicly available MI450 GPU clusters (50,000 GPUs, Q3 2026), and Meta has signed a 6GW deal with AMD.[24] The first UALink-capable switches are expected from Marvell (leveraging its XConn acquisition for PCIe/CXL switching silicon) and Upscale AI (supporting UALink and ESUN in scale-up) in late 2026 or early 2027.

In April 2026, before any UALink 1.0 hardware has shipped, the consortium published four additional specifications that expand UALink's scope.[25] UALink Common 2.0 introduces in-network compute, defining collective operations (Broadcast, Reduce) executed on UALink switches — closing a key feature gap with NVLink's in-switch reduction capability. UALink Manageability 1.0 adds centralized control and management planes using standard protocols (gNMI, Yang, SAI, Redfish), and UALink Chiplet 1.0 (UCIe 3.0 compliant) enables integration into multi-die SoCs. A separate 200G DL/PL 2.0 specification decouples the physical layer from the common protocol, allowing independent speed bumps — UALink 3.0 with increased bandwidth is roadmapped for 2027.

ESUN/SUE-T: Never Bet Against Ethernet

The combination of ESUN and SUE-T represents the Ethernet community's answer to proprietary scale-up interconnects, offering the potential for multi-vendor interoperability and commodity economics. ESUN provides the framework and broader architectural specification, while SUE-T contributes specific link-level mechanisms.

ESUN (Ethernet for Scale-Up Networking) was announced in October 2025 at the OCP Global Summit. It defines Ethernet L2/L3 frame extensions enabling high-performance Ethernet switches to provide scale-up fabric capabilities. ESUN operates as an open technical forum within OCP, with members spanning the breadth of the AI ecosystem, including AMD, Arista, ARM, Broadcom, Cisco, HPE Networking, Marvell, Meta, Microsoft, NVIDIA, and OpenAI.[26] NVIDIA's participation in ESUN (even as it continues to invest in NVLink) is similar to their joining of UEC (Ultra Ethernet Consortium) while promoting InfiniBand. We expect they want to provide customers with alternatives but also as a technology hedge. ESUN's technical focus areas include Ethernet framing and switching optimization, XPU network interface interoperability, Ethernet switch ASIC specifications for scale-up, and robust, lossless, error-resilient topologies supporting both single-hop and multi-hop configurations.

Complementing ESUN, **SUE-T** (Scale-Up Ethernet Transport) 1.0 defines three key technical features: an AI Fabric Header (AFH) that minimizes per-packet overhead, Link Level Retry (LLR) for hop-by-hop packet reliability, and Credit-Based Flow Control (CBFC) for congestion prevention. SUE targets up to 1,024 ports per switch and is designed to enable conventional Ethernet silicon to support scale-up workloads through firmware and driver enhancements rather than hardware redesign.[27]

CXL: Memory Pooling and Coherency

Compute Express Link (CXL) continues to complement the scale-up interconnects discussed above, enabling memory pooling and coherency across accelerators rather than direct accelerator-to-accelerator compute traffic. CXL 3.1, released in November 2023, has entered volume production. CXL 4.0, released in November 2025, increased bandwidth to 128 GT/s (data transfer

[24]"Advanced Micro Devices, Inc.," Advanced Micro Devices, Inc., Feb. 24, 2026. [Online]. Available: https://ir.amd.com/news-events/press-releases/detail/1279/amd-and-meta-announce-expanded-strategic-partnership-to-deploy-6-gigawatts-of-amd-gpus
[25]L. Clark, "UALink 2.0 arrives before version 1 chips have even shipped," *The Register*, 7 Apr. 2026. [Online]. Available: https://www.theregister.com/2026/04/07/ualink_2_specs/
[26]Cisco, "Cisco Joins Forces with OCP in the Ethernet for Scale-Up Networking (ESUN) Collaboration," *Cisco Blogs*, 13 Oct. 2025. [Online]. Available: https://blogs.cisco.com/news/cisco-joins-forces-with-ocp-in-the-ethernet-for-scale-up-networking-esun-collaboration
[27]Synopsys, "Ethernet Standards for Scale-Up AI: An Overview of ESUN, SUE, and UALink," *Synopsys*, 15 Jan. 2026. [Online]. Available: https://synopsys.com/articles/ethernet-standards-scale-up-ai.html

rates of up to 512 GB/s) via PCIe 7.0 physical layers. Microsoft launched CXL-enabled instances on Azure in November 2025, and commercial CXL memory pools totaling 100 TB have been deployed. CXL provides latency approximately 3.8x faster than equivalent 200G RDMA-based remote memory access, with typical latencies in the 200–500 nanosecond range — making it attractive for models exceeding per-accelerator memory capacity.[28]

Marvell's acquisition of XConn Technologies for $540 million and launch of the Apollo 2 hybrid CXL 3.1/PCIe 6.2 switch reflects the growing importance of flexible switching across CXL and PCIe protocols.

Scale-Up Competitive Landscape and Convergence Outlook

Attribute	UALink 1.0	NVLink 5	ESUN/SUE
Per-lane data rate	200 Gbps	~900 Gbps (per link)	100–200 Gbps
Max cluster size	1,024 accelerators	72 (NVL72) / 576 (multi-pod)	1,024+ ports
Protocol semantics	Memory (load/store)	Memory (proprietary)	Network (Ethernet L2/L3)
Switch / fabric aggregate	115 Tbps	130 TB/s (NVL72 fabric)	51.2–102.4 Tbps
Consortium members	115+	NVIDIA + Fusion partners	11 founding (ESUN)
Production hardware	Late 2026/early 2027	Shipping (NVLink 5)	Late 2026+

The proliferation of scale-up standards has created technical competition for the first time in this market segment. The existence of multi-protocol support hardware signals market expectation of coexistence rather than winner-take-all. Both Credo's Blue Heron and Marvell's Celestial AI CPO technology support multiple protocols. Blue Heron is a 224G multiprotocol retimer endorsed by AMD and Upscale AI that supports UALink, ESUN, and Ethernet simultaneously. Upscale AI is similarly protocol-agnostic, supporting UALink, ESUN, and custom memory protocols.[29] Nevertheless, NVLink is the de facto standard to beat, and it has the advantage of NVIDIA's focus on extreme co-design. UALink has AMD's backing (along with many other consortium players) and will be adopted at scale with AWS Trainium 4 (which will also use NVLink). ESUN will benefit from Broadcom's backing and its extensive ecosystem of switching partners, plus its alignment with UEC/UET and the simplicity of having Ethernet across all domains make it a formidable competitor to NVLink.

NVIDIA's NVLink and NVSwitch represent the dominant existing scale-up interconnect today, with NVLink scale-up shipments crossing $1 billion in Q1 FY2026 alone and contributing materially to NVIDIA's record $8.2 billion of networking revenue in Q3 FY2026.[30] The merchant-silicon, multi-vendor scale-up switching market (switching ASICs sold separately from a vertically integrated GPU platform) is currently near-zero, but Astera Labs projects it to reach $20B by 2030 as clusters scale beyond the direct-connect limits of current architectures.[31] NVIDIA's NVL72 and AMD's Helios (72 GPUs with UALink at 260 TB/s) represent the first generation of competing rack-scale architectures, with both vendors now demonstrating aggregate scale-up bandwidth in the hundreds of terabytes per second. The pragmatic recommendation for infrastructure planners is to maintain optionality: design clusters with modular architectures that can accommodate different interconnect standards, and

[28] CXL Consortium, *Compute Express Link Specification 4.0*, Nov. 2025. [Online]. Available: https://computeexpresslink.org/cxl-specification/

[29] Credo, "Credo Introduces Industry's First 224G Multiprotocol AI Scale-Up Retimer Supporting UALink, ESUN and Ethernet," *Business Wire*, 29 Jan. 2026. [Online]. Available: https://businesswire.com/news/home/20260129771160/en/

[30] NVIDIA, "NVIDIA Announces Financial Results for Third Quarter Fiscal 2026," *NVIDIA Newsroom*, 19 Nov. 2025. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-third-quarter-fiscal-2026

[31] "Astera Labs Broadens Scorpio X-Series Smart Fabric Switch Roadmap to Address Expanding Scale-Up Market," *Astera Labs*, 22 Jan. 2026. [Online]. Available: https://asteralabs.com/news/astera-labs-broadens-scorpio-x-series-smart-fabric-switch-roadmap-to-address-expanding-scale-up-market-opportunities/

phase deployments to leverage emerging alternatives. The competitive dynamics will intensify through 2027–2028 as Feynman (NVLink 8) and MI500-series (likely UALink 2.0+) reach production.

Scale-Out Networking: The Architecture of Modern AI Clusters

Whereas scale-up focuses on intra-rack connections, **scale-out** networking connects racks and pods within a data center, forming the backbone of large-scale AI clusters. This domain has undergone a structural shift: where InfiniBand once dominated AI training settings, Ethernet has emerged as the preferred architecture for new hyperscale deployments, driven by economics, ecosystem breadth, and rapid innovation. Dell'Oro projects that the AI back-end switch market will exceed $100 billion by 2030.[32]

Scale-Out Architecture Choices: Networking Topologies and Scheduling

Fabric topology choice has major cost and performance impacts. Competing architectures offer different trade-offs between switch cost, traffic flexibility, and scalability.

Properly dimensioned **fat-tree (folded Clos)** topologies that are not oversubscribed provide full bisection bandwidth through multi-tier switching hierarchies. At 10,000 GPUs, a three-tier fat-tree delivers predictable performance for all collective operations but requires substantial switch investment — particularly at the spine tier, where high-radix switches aggregating hundreds of 800G ports become the cost center. Fat-tree remains the default choice for mixed workloads with unpredictable traffic patterns.

Rail-optimized topologies reduce switch cost by exploiting the observation that most training traffic follows GPU-rank ordering within rails — GPU 0 in each server communicates primarily with GPU 0 in other servers, GPU 1 with GPU 1, and so forth. This allows each "rail" to use a separate, smaller switch fabric rather than a single large one. The cost savings are significant — 30–50% fewer switch ports at the spine tier — but rail-optimized topologies require strict cabling discipline and perform poorly when traffic deviates from the assumed pattern, as may occur in MoE AllToAll operations. Rail-optimized designs have been adopted at hyperscale for pure training clusters, with **hybrid-rail** approaches that may leverage scale-up links for cross rank communication. An extreme example of this is the Rail-only approach that MIT/Meta investigated and that we covered in last year's report.[33]

3D torus topologies connect each node directly to its neighbors in three dimensions, eliminating switches from the data path. The approach lowers network diameter, substantially lowering worst-case hop count. Google's TPU architecture is the most prominent production deployment: TPU v4 interconnects 4,096 chips in a 3D torus using direct-attached copper within racks and optical circuit switches (OCS) between cubes, with OCS consuming less than 5% of system cost and 3% of system power.[34] TPU v5p scales to 8,960 chips achieving 4,800 Gbps per-chip interconnect bandwidth. Twisted torus variants provide up to 70% higher bisection bandwidth over non-twisted configurations.[35] The trade-off is rigidity: torus topologies are difficult to partition efficiently, require topology-aware workload placement, and multi-hop communication between distant nodes consumes intermediate-node bandwidth — making them best suited for purpose-built accelerator pods with predictable communication patterns rather than general-purpose GPU clusters.

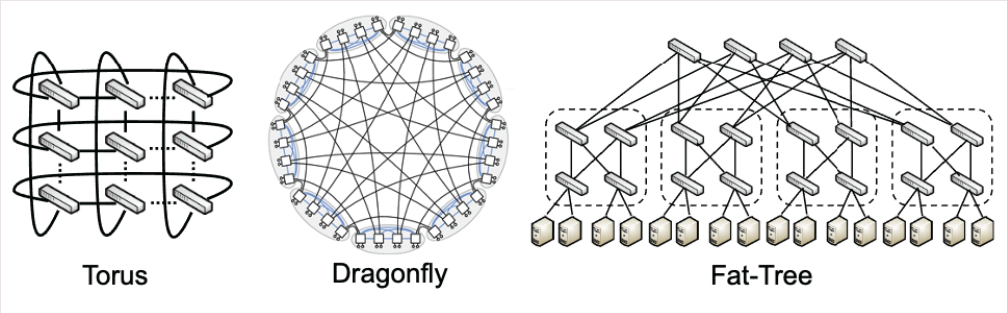
[32] "AI Back-End Switch Market Will Push Past $100 Billion by 2030," Dell'Oro Group, Feb. 04, 2026. https://www.delloro.com/news/ai-back-end-switch-market-will-push-past-100-billion-by-2030/

[33] Wang, Weiyang & Ghobadi, Manya & Shakeri, Kayvon & Zhang, Ying & Hasani, Naader. (2024). Rail-only: A Low-Cost High-Performance Network for Training LLMs with Trillion Parameters. 1-10. 10.1109/HOTI63208.2024.00013. Available: https://people.csail.mit.edu/ghobadi/papers/rail_only_hoti_2024.pdf

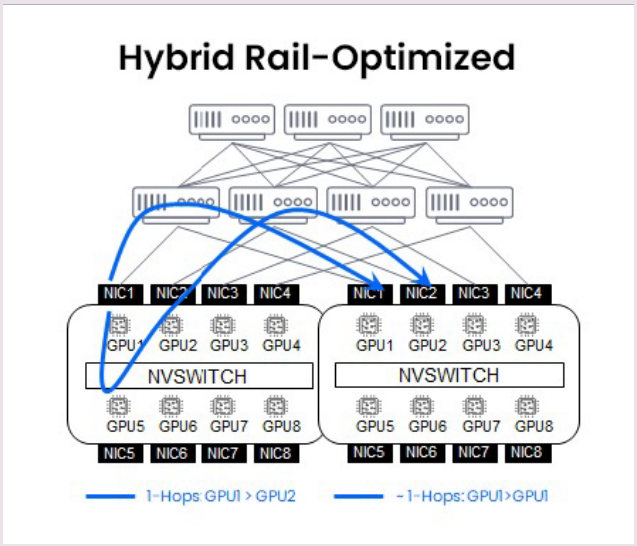
[34] N. P. Jouppi et al., "TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings," *arXiv preprint*, arXiv:2304.01433, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.01433

[35] Google Cloud, "TPU v5p Documentation," *Google Cloud*. [Online]. Available: https://cloud.google.com/tpu/docs/v5p

EXAMPLES OF SCALE-OUT TOPOLOGIES



Source: Z. Chen, Z. Zhao, Z. Li, J. Shao, S. Liu, and Y. Xu, "SDT: A Low-cost and Topology-reconfigurable Testbed for Network Research," IEEE Cluster vol. 23, pp. 343–353, Oct. 2023, doi: https://doi.org/10.1109/cluster52292.2023.00036.



Source: DriveNets

AvidThink

Topologies have different costs, flexibility, and scale.

avidthink.com

Dragonfly topologies organize switches into fully connected local groups, with inter-group optical links providing low-diameter system-wide connectivity. Typically, there are three hops maximum between any two endpoints. HPE's Slingshot interconnect, built on 64-port switches at 200 Gbps per port (12.8 Tbps per switch), implements a dragonfly topology across three exascale-class supercomputers: Frontier (over 36,000 AMD MI250X GPUs), El Capitan (43,808 AMD MI300A GPUs), and Aurora (over 60,000 Intel Data Center Max GPUs with nearly 85,000 Slingshot NICs and 5,600 switches, which is the largest Slingshot deployment to date).[36] Dragonfly's cost advantage over fat-tree grows with scale, as inter-group links scale sub-linearly rather than linearly. However, dragonfly requires adaptive routing to distribute traffic effectively. Misconfigured routing can perform worse than

[36] Argonne National Laboratory, "Scaling MPI Applications on Aurora," *arXiv preprint*, arXiv:2512.04291, Dec. 2025. [Online]. Available: https://arxiv.org/abs/2512.04291

static alternatives, and workloads generating uniform AllToAll traffic (common in MoE training) can stress inter-group link capacity.[37] In practice, dragonfly has proven its value primarily in HPC/supercomputing deployments; adoption in commercial AI GPU clusters remains limited compared to fat-tree and rail-optimized designs.

Scheduled fabrics are a scheduling method applied to a physical topology, typically Clos-based. Their main trade-off is between complexity and performance: centralized scheduling and virtual output queueing achieve lossless, predictable transport and near-circuit-switched behavior, but at the cost of more complex control methods. The approach can, in principle, extend to any topology, but requires sophisticated management.

Meta's Disaggregated Scheduled Fabric (DSF), supporting approximately 18,000 GPUs, employed Broadcom Jericho3-AI and Ramon3 switches, implementing VOQ to prevent head-of-line blocking.[38] Meta simultaneously operates a Non-Scheduled Fabric (NSF) supporting approximately 20,000 GPUs, built on shallow-buffer Ethernet switches with adaptive routing. Importantly, Meta qualified three distinct ASIC platforms for NSF, including Broadcom TH5, Cisco G200, and NVIDIA Spectrum-4, all of which are interoperable through Meta's FBOSS (Facebook Open Switching System) and SAI, thereby providing supply chain resilience.[39]

DriveNets' Network Cloud-AI implements fabric scheduling through a switch-only approach (no NIC involvement), using Broadcom Jericho/Ramon chipsets for the fabric-scheduled variant (similar to Meta DSF) and Broadcom Tomahawk 6 (TH6) with select NICs from Broadcom and AMD for a cost-efficient endpoint-scheduled variant. In NCCL testing, DriveNets claims 5–10% higher bus bandwidth than InfiniBand, with production deployments of up to 8,000 GPUs and theoretical scaling to 100,000+ GPU clusters.[40]

Topology	Switch Cost	Traffic Flexibility	Max Proven Scale	Best Fit
Fat-tree (Clos)	Highest	Any pattern	100K+ GPUs	Mixed workloads
Rail-optimized	30–50% lower	Rail-aligned only	100K+ GPUs	Pure training
3D Torus	Lowest (switchless)	Neighbor-optimized	~9K chips (TPU v5p)	Purpose-built accelerator pods
Dragonfly	Moderate	Adaptive-routing dependent	85K+ NICs (Aurora)	HPC/supercomputing

InfiniBand Update

InfiniBand has not disappeared from the landscape, but its market share trajectory has reversed. In 2023, InfiniBand commanded approximately 80% of AI backend network deployments; by mid-2025, Ethernet captures the majority of new builds. Yet, InfiniBand continues to mature with each generation: NDR at 400 Gbps is production-mature, while XDR at 800 Gbps represents the current frontier. The Quantum-X800 Q3400-RA delivers 144 ports at 115 Tbps aggregate bisection bandwidth.[41]

Unfortunately, the economics remain challenging. InfiniBand carries a meaningful capital-cost premium over Ethernet at the network fabric level (e.g., NICs, switches, transceivers, and cabling). A 2024 SemiAnalysis modeling of a 100K H100 cluster observed that Infiniband was 1.3 to 1.6 times more expensive than other options (including NVIDIA Spectrum-X and Broadcom silicon-based).[42] In multi-100K GPU clusters, that gap represents hundreds of millions of dollars in incremental capital expendi-

[37] K. J. Barker et al., "Evaluation of Topology-Aware All-Reduce Algorithm for Dragonfly Networks," in *Proc. NPC 2021*, Springer, 2021, pp. 231–243.
[38] Meta Engineering, "Disaggregated Scheduled Fabric: Scaling Meta's AI Journey," *Engineering at Meta*, 20 Oct. 2025. [Online]. Available: https://engineering.fb.com/2025/10/20/data-center-engineering/disaggregated-scheduled-fabric-scaling-metas-ai-journey/
[39] Perves and L. Lingjun, "Evolution of Ethernet-Based Switch Platforms and Fabrics to Meet Meta's AI Training Clusters," presented at *OCP Global Summit 2025*, 2025. [Online]. Available: https://youtube.com/watch?v=ELtoxc4g9JI
[40] Haim, Shai. "Ethernet Beats InfiniBand in Production. See the Numbers." DriveNets, 14 Jan. 2026, https://drivenets.com/blog/ethernet-beats-infiniband-in-production-see-the-numbers/.
[41] NVIDIA, "Quantum-X800 InfiniBand Platform," *NVIDIA Networking*, 2025. [Online]. Available: https://nvidia.com/en-us/networking/products/infiniband/quantum-x800/
[42] SemiAnalysis, "100,000 H100 Clusters: Power, Network Topology, Ethernet vs InfiniBand, Reliability, Failures, Checkpointing." [Online]. Available: https://newsletter.semianalysis.com/p/100000-h100-clusters-power-network

ture, which is difficult to justify. Recent pricing quotes from network equipment resellers in late 2025 indicate that the Infiniband premium remains for even small sub-1K enterprise GPU clusters. Nevertheless, InfiniBand retains relevance for tightly coupled clusters where microsecond latency improvements warrant the premium, but the installed base of Ethernet expertise, multi-vendor availability, and ecosystem momentum make broad trend reversal unlikely.

Ethernet Fabrics and RoCEv2 at Scale

The migration to Ethernet accelerated sharply through 2025, and the share of 800GbE ports, which help power AI backend networks, has been growing. Crehan Research, which named Arista as the leader in branded market share for Ethernet data center switching, projected in 2025 that the sector would grow at an average of 90% per year over the next five years.[43] RoCEv2 (RDMA over Converged Ethernet v2) remains the production workhorse for AI Ethernet fabrics, encapsulating InfiniBand semantics atop Ethernet frames, with UEC UET (see next section) primed to take over. In addition, the arrival of high-radix 102.4 Tbps switches with multiple vendors providing merchant silicon will continue to drive Ethernet forward (details later in the report), as will proliferation of Ethernet at frontier labs and hyperscaler deployments. xAI's Colossus deployment — which began with 100,000 H100 GPUs and has since scaled past 200,000 GPUs (with H200s and GB200s added)[44] on NVIDIA Spectrum-X Ethernet fabric — demonstrated lossless Ethernet at unprecedented scale: NVIDIA reports zero packet loss in steady state operation and 95% measured data throughput across the entire cluster. Spectrum-X achieves this through adaptive routing, sophisticated congestion control, and higher RDMA bisection bandwidth compared to off-the-shelf Ethernet with lower latency.[45]

xAI's Colossus: scaled past 200,000 GPUs (a mix of H100, H200, and GB200) on Spectrum-X Ethernet, challenging InfiniBand's traditional advantage at unprecedented scale.

Ultra Ethernet Consortium (UEC/UET) 1.0

The June 2025 release of UEC 1.0 represented a significant specification milestone. The 562-page standard, authored by the Ultra Ethernet Consortium with over 100 members, represents a comprehensive re-engineering of Ethernet transport for accelerated computing.[46] UET (Ultra Ethernet Transport) is designed to supersede RoCEv2 with architectural improvements:

- **Native multipathing** at the transport layer, enabling packet spraying across multiple fabric paths without destination reordering overhead.
- **Out-of-order delivery reception**, eliminating the head-of-line blocking that limits RoCEv2 performance during asymmetric load.
- **Rapid loss recovery** with sub-millisecond retransmission, reducing the impact of transient link errors.
- **Pluggable congestion control**, permitting hyperscalers to select algorithms customized for their specific workload characteristics.

The critical question is production readiness. Broadcom's UEC-compliant Thor Ultra NIC (now sampling) was announced in late 2025. UEC-compliant switch hardware from multiple vendors is expected in late 2026, with volume deployments likely in 2027. In the interim, NVIDIA's Spectrum-X with Spectrum-6 CPO (announced at GTC 2026, delivering 102.4 Tbps with integrated silicon photonics at up to 5x power efficiency)[47] provides a vertically integrated Ethernet alternative that delivers many of UEC's

[43] "Arista Networks Unveils Next Generation Data and AI Centers," Arista.com, 29 October. 2025. https://investors.arista.com/Communications/Press-Releases-and-Events/Press-Release-Detail/2025/Arista-Networks-Unveils-Next-Generation-Data-and-AI-Centers/ (accessed Apr. 03, 2026).
[44] "Colossus AI Hits 200,000 GPUs as Musk Ramps Up AI Ambitions," *HPCwire*, 13 May 2025. [Online]. Available: https://www.hpcwire.com/2025/05/13/colossus-ai-hits-200000-gpus-as-musk-ramps-up-ai-ambitions/
[45] NVIDIA, "Spectrum-X Ethernet Networking Platform Whitepaper," *NVIDIA*, 2025. [Online]. Available: https://nvidia.com/en-us/networking/spectrumx/
[46] Ultra Ethernet Consortium, *UEC 1.0 Specification*, Jun. 2025. [Online]. Available: https://ultraethernet.org/
[47] NVIDIA, "NVIDIA GTC 2026 Keynote," J. Huang, 16 Mar. 2026. [Online]. Available: https://blogs.nvidia.com/blog/gtc-2026-news/

architectural benefits today, including adaptive routing, congestion control, and multipath. While Spectrum-X uses standard Ethernet and RoCEv2 at the wire level, its full performance profile depends on tight NIC-switch coordination across NVIDIA's ConnectX/BlueField + Spectrum portfolio. This complicates decisions for data center operators, who must choose between waiting for multi-vendor open standards and deploying NVIDIA's vertically integrated stack today.

Optical Circuit Switching in Scale-Out Fabrics

Google has deployed tens of thousands of optical circuit switching (OCS) ports in production. And OCS is making the move from Google-internal technology toward commercial viability, though it remains niche. The OCP established a dedicated OCS subproject in August 2025, co-led by iPronics and Lumentum, to develop standardized commercial solutions[48]. And at OFC 2026, Marvell and Lumentum demonstrated an integrated, rack-level OCS system combining Marvell's Ara 1.6T PAM4 DSPs, Aquila 1.6T coherent-lite DSPs, and COLORZ 800 ZR/ZR+ DCI modules with Lumentum's R300 OCS platform. This demonstration is significant because it validates OCS interoperability across scale-up (Ara), scale-out (Aquila), and scale-across (COLORZ) connectivity in a single rack. It moves OCS from a Google-only capability to a multi-vendor, multi-tier solution. Lumentum's R300 OCS platform claims 5–10x lower latency and up to 65% power savings compared to electronic switching in 100K-GPU deployments.[49]

Scale-Out Key Takeaways

Scale-out networking is undergoing a fundamental transition: Ethernet has displaced InfiniBand as the default for hyperscale AI clusters, driven by cost and ecosystem advantages; topology choice remains workload-dependent, with Clos and rail-optimized designs dominating commercial deployments; UEC/UET represents the next major evolution, aligning Ethernet transport with AI-specific requirements; OCS is upcoming and worth watching as an alternative; and multi-vendor silicon competition (Broadcom, Cisco, NVIDIA, Marvell) is changing cost and procurement dynamics. For infrastructure planners, attention shifts away from selecting a single "best" architecture toward weighing performance versus cost, determinism versus flexibility, and proprietary integration versus an open ecosystem.

Scale-Across Networking: Interconnecting Distributed Data Centers

As individual facility constraints (power, cooling, floor space) create size ceilings for AI clusters, distributed architectures become the path to ever-larger GPU clusters. The transition from single-facility scale-out clusters to geographically distributed AI infrastructure introduces different challenges.

Within a data center, engineers control the fiber plant, carefully manage latency, and design fabrics around predictable propagation delays measured in microseconds. Across buildings and campuses, fiber latency becomes significant, and engineers must contend with the fundamental physics of light propagation through fiber. In the domain of scale-across, optics and standards such as 400ZR/ZR+, 800ZR/ZR+, and 1.6T and beyond are key drivers.

Multi-Campus Architectures Driven by Power Constraints

The economics of power are forcing architectural progression. US electricity generation totals 4,400–4,500 TWh annually; Lawrence Berkeley National Laboratory predicts that data center demand will grow from 176 terawatt hours (TWh) in 2023 (or, about 4.4% of total U.S. electricity consumption) to between 325-580 TWh (6.7-12.0%) by 2028. This growth rate exceeds the

[48] B. Okonkwo and R. Liu, "OCP Optical Circuit Switching Subproject Update," presented at *OCP Global Summit 2025*, 2025. [Online]. Available: https://youtube.com/watch?v=_UCnVV6cs9Y

[49] "Lumentum Optical Circuit Switch to Improve Next-Generation AI Data Center Scalability," Nasdaq.com, Mar. 2025. [Online]. Available: https://www.nasdaq.com/press-release/lumentum-optical-circuit-switch-improve-next-generation-ai-data-center-scalability

pace at which new generation capacity can be brought online, creating per-site power ceilings that limit cluster size.[50]

The architectural response is multi-campus design: rather than concentrating all GPUs in a single facility limited by its power allocation, operators distribute GPU pools across multiple sites connected by high-capacity DCI. For example, DriveNets' Network Cloud-AI delivers multi-site capabilities extending to 80 km between fabric nodes, enabling distributed GPU clusters to operate as unified compute entities.[51]

Meanwhile, as inferencing clusters grow in size and complexity beyond a few GPU servers, smarter network fabrics will play a critical role. Networking software vendor Arrcus differentiates with its ArcOS by providing sub-second convergence, CDN-like routing intelligence for GPU resource scheduling, and dynamic bandwidth reconfiguration across data centers. In addition, their recently announced Arrcus Inference Network Fabric extends multi-site inferencing, considering XPU loading, model awareness, KV cache locations, and other policy constraints when routing inferencing requests across sites.[52]

Regardless of the routing and intelligence at higher networking layers, the underlying data center interconnect (whether campus scale, metro scale or larger) is fiber-based. Optical vendors in the space are seeing significant demand. For example, Coherent Corp stated Q2 FY2026 revenue of $1.69 billion (17% increase) while Lumentum reported Q2 FY2026 revenue of $665.5 million (65% YoY growth), noting during their earnings call that demand is outpacing supply, with Lumentum currently 'under shipping the market' by 25% to 30%. These revenue scales — hundreds of millions to nearly $2 billion quarterly — reflect that coherent optics has become mainstream datacenter technology.[53]

400G to 800G and the Path to 1.6T

The 400ZR/ZR+ specification saw broad hyperscaler DCI deployment in 2024–2025, becoming widely used. 400ZR modules (QSFP-DD form factor) provide 80 km reach over single fiber pairs, making them the standard for campus and metro DCI. The 800ZR/ZR+ specifications, following the OIF Implementation Agreement released in November 2024, represent the next progression. Meta is a major 800ZR+ customer, reflecting an ongoing emphasis on deploying advanced optical technologies for distributed cluster architectures.[54]

Other players in the space include Ciena and Marvell. Ciena's WaveLogic 6 (WL6) family includes two major variants. WL6e (Extreme) is the performance-optimized engine, delivering 1.6 Tb/s on a single carrier, and supporting unregenerated 800 Gbps across 4,000–5,000 km of terrestrial fiber. WL6n (Nano) is the footprint-optimized variant powering 800ZR/ZR+ coherent pluggables for 800G metro and regional DCI up to 1,000 km, as well as 400G long-haul pluggables.[55] Marvell's coherent portfolio spans the 400G Canopus and Deneb DSPs, the 800G Orion generation, and — announced in March 2026 — the 2 nm Electra 1.6T ZR/ZR+ coherent DSP paired with the COLORZ 1600 1.6T pluggable (sampling in H2 2026), addressing campus (20 km), metro (120 km), and regional (up to 1,000 km) links.[56] Separately, earlier at OFC 2025 Marvell also demonstrated Ara, a 3 nm 1.6T PAM4 (direct-detection) DSP running at 200 Gbps per electrical lane, which targets short-reach 1.6T-DR8 intra-data-center optics.[57]

[50] A. Shehabi et al., "2024 United States Data Center Energy Usage Report," Lawrence Berkeley National Laboratory, LBNL-2001637, Dec. 2024. [Online]. Available: https://doi.org/10.71468/P1WC7Q

[51] DriveNets, "DriveNets Extends AI Networking Fabric with Multi-Site Capabilities," *Network World*, Feb. 2026. [Online]. Available: https://networkworld.com/article/3992283/drivenets-extends-ai-networking-fabric-with-multi-site-capabilities-for-distributed-gpu-clusters.html

[52] "Arrcus delivers record breaking 3x bookings growth in 2025, and introduces AI-policy aware Arrcus Inference Network Fabric | Arrcus," Arrcus, 18 Feb 2026. https://arrcus.com/news/arrcus-introduces-arrcus-inference-network-fabric

[53] Coherent Corporation, "Second Quarter Fiscal Year 2026 Results," *Coherent Corp*, 4 Feb. 2026. [Online]. Available: https://coherent.com/news/press-releases/second-quarter-fiscal-year-2026-results

[54] A. Schmitt, "800ZR/ZR+ Forecast Update - 2Q25 - Cignal AI," *Cignal AI*, 28 Jul. 2025. [Online]. Available: https://cignal.ai/2025/07/800zr-zr-forecast-update-2q25/

[55] Ciena, "Ciena Brings Data Center Connectivity Innovations to OFC 2025," *Ciena Newsroom*, 25 Mar. 2025. [Online]. Available: https://ciena.com/about/newsroom/press-releases/ciena-brings-data-center-connectivity-innovations-to-ofc-2025

[56] Marvell Technology, "Marvell Extends ZR/ZR+ Leadership with Industry-first 1.6T ZR/ZR+ Pluggable and 2nm Coherent DSPs for Secure AI Scale-across Interconnects," *Marvell Newsroom*, Mar. 2026. [Online]. Available: https://www.marvell.com/company/newsroom/marvell-1-6t-zr-zr-plus-pluggable-2nm-coherent-dsp-ai-interconnects.html

[57] "Marvell Technology, Inc.," Marvell Technology, Inc., Mar. 31, 2025. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/106/marvell-advances-interconnect-portfolio-for-scale-up-and-scale-out-fabrics-at-ofc-2025

At OFC 2026, Nokia announced a suite of application-optimized coherent transport solutions. These reflect the industry's shift toward workload-specific optical designs. The portfolio includes a 1.6T-capable coherent pluggable optimized for IP-over-DWDM DCI and scale-across applications. A 2.4T-capable coherent pluggable serves thin transponder deployments across terrestrial and subsea networks. A 3.2T-capable coherent-lite solution is optimized for energy-efficient, short-reach campus and enterprise applications. There is also a new class of full-band transponders that combine hundreds of coherent components into a single, operationally simplified solution for hyperscale capacity.

Nokia also introduced a multi-rail in-line amplifier supporting up to 160 fiber pairs in a single rack — directly addressing the fiber density demands of multi-campus AI deployments. Nokia claims up to 70% lower total cost of ownership across the suite[58]. Ciena also introduced a hyper-rail photonics architecture at OFC 2026, providing up to 128 fiber pairs per rack (32 times current density), a 75% power reduction, and an 85% space reduction. The presence of multi-fiber amplification platforms from both Nokia and Ciena points to the strong need for multi-campus AI connectivity infrastructure.

Technology	Reach	Data Rate	Status (2026)	Primary Use Case
400ZR pluggables	80 km	400 Gbps	Mature, volume	Campus/metro DCI
400ZR+ pluggables	300+ km	400 Gbps	Mature, volume	Regional DCI
800ZR/ZR+/ZR++ pluggables	80–1500+ km	800 Gbps	Ramping, 200K+ units	Next-gen DCI
1.6T ZRx pluggables	10–2,000+ km	1.6 Tbps	Development/sampling	Future DCI
2.4T ZRx pluggables	Terrestrial/subsea	2.4 Tbps	Sampling mid-2027	Thin transponder DCI

DCI Architecture: Capacity, Oversubscription, and Parallelism Mapping

The relationship between DCI capacity and parallelism mapping is bidirectional: the available cross-DC bandwidth constrains which parallelism dimensions can span site boundaries, and the chosen parallelism strategy determines the DCI bandwidth requirement.

Data parallelism is the most DCI-friendly approach. Each site holds a full model replica and synchronizes gradients, requiring only gradient aggregation traffic. This traffic is proportional to model size, not dataset size. Pipeline parallelism can span sites if inter-stage latency budgets can absorb propagation delay. Tensor parallelism requires microsecond-level latency, so it generally cannot span sites. This makes it the main constraint on per-site cluster size.

For a 129,000-GPU cluster spanning five buildings with a maximum GPU-to-GPU distance of 3 km (as Meta has deployed), the approximately 15 µs one-way propagation delay requires distinct congestion management strategies for cross-building versus intra-building communication. NVIDIA's Spectrum-XGS tackles this with auto-adjusted distance congestion control that directly incorporates measured latency into response curves, nearly doubling NCCL performance for distributed training across geographically separated clusters.[59]

Copper's Relevancy, Optical Technologies, and the Power Imperative

Moving on from the three networking domains, we'll examine a topic that continues to make the rounds: if and when copper will yield to optics. First, we note that power consumption now dominates photonics innovation. In earlier periods, optical innovations focused on reducing costs per terabit or improving latency. In 2025–2026, power is a top selection criterion. XPU

[58] Nokia, "Nokia launches suite of application-optimized optical solutions for AI-era networks," *Nokia Newsroom*, 16 Mar. 2026. [Online]. Available: https://nokia.com/newsroom/nokia-launches-suite-of-applicationoptimized-optical-solutions-for-ai-era-networks/

[59] NVIDIA, "NVIDIA Introduces Spectrum-XGS Ethernet to Connect Distributed Data Centers Into Giga-Scale AI Super-Factories," *NVIDIA Newsroom*, Feb. 2026. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-introduces-spectrum-xgs-ethernet-to-connect-distributed-data-centers-into-giga-scale-ai-super-factories

power consumption is rising fast, from below 1,000W in 2023 toward over 4,000W by 2027. That is roughly a 500W increase per year. This creates tight system-level power budgets, demanding efficiency throughout the interconnect hierarchy.[60]

Copper Reach Limits at 800G and 1.6T

Copper interconnects remain relevant at short distances but face physical limitations as data rates increase.

Speed	Passive DAC Reach	Active Copper Reach	Power (pJ/bit)
400G (112 Gbps/lane)	3–5 m	7–10 m	~10–12
800G (112 Gbps/lane)	2–3 m	3–7 m	~10–13
1.6T (224 Gbps/lane)	~1 m	~2–3 m (early)	~10–15
3.2T (400 Gbps/lane)	<1 m	~1 m (highly constrained)	Not competitive vs optics

Copper interconnects are viable within physics-dictated limits, which are tightening as signaling rates increase. At 224 Gbps per lane (supporting 1.6-terabit ports), passive copper links shrink to about one meter, restricting use to intra-rack or intra-chassis connections. Active copper increases this distance but requires more complex equalization and retiming, driving up power and cost. At 400 Gbps per lane, copper is limited to environments such as trays and backplanes, or to electrical domains with short traces.[61]

Rack adjacency, cable routing density, and thermal envelopes must all be reconsidered as copper links shorten.

This reach limitation shapes data center design. What were once simple top-of-rack connections now require either denser rack placement or a transition to optical interconnects. At the same time, copper's economic advantage is eroding: at 800G, active electrical cables offer savings over optics for very short distances, but at 1.6T, signal-conditioning complexity brings costs closer to those of optical modules. By 3.2T, optics becomes the default for nearly all interconnect distances.

Even so, copper is not disappearing. At OFC 2026, Credo demonstrated 1.6T AECs designed into NVIDIA's Vera Rubin NVL144 and Kyber Ultra NVL576 platforms, confirming that copper AECs will remain the preferred interconnect for the shortest-reach, highest-bandwidth scale-up links even as optics dominate longer distances.[62]

Micro-LED Interconnects

Micro-LED technology constitutes a different approach to short-reach interconnect, using wide, slow parallel lanes rather than narrow, fast serial lanes. Avicena's gallium nitride LED arrays with fiber bundle coupling claim to achieve less than 1 pJ/bit power consumption at 4–16 Gbps per lane with 10–20m reach. A TSMC partnership to optimize photodetector (PD) arrays for Avicena's revolutionary LightBundle microLED-based interconnects increases the credibility of such an offering reaching production. Likewise, Credo's acquisition of micro-LED provider Hyperlume in September 2025 also signals wider industry interest in this alternative approach.[63]

[60] C. Koopmans, "Industry Analyst Day 2025," *Marvell Technology*, 9 Dec. 2025.
[61] R. Nagarajan, "Optical Engineering for High-Speed Interconnects," presented at *Marvell Industry Analyst Day 2025*, 9 Dec. 2025.
[62] Credo, "Credo to Showcase Optical Solutions for AI Scale-Out Fabrics at OFC 2026," *Credo Investor Relations*, Mar. 2026. [Online]. Available: https://investors.credosemi.com/news-events/news/news-details/2026/Credo-to-Showcase-Optical-Solutions-for-AI-Scale-Out-Fabrics-at-OFC-2026/
[63] Chris, "Ultra Low Power MicroLED Based Interconnects," presented at *OCP Global Summit 2025*, 2025. [Online]. Available: https://youtube.com/watch?v=NPgvUMtACHQ

Traditional Pluggable Optics (DSP-Based)

DSP-based pluggable transceivers — the workhorses of data center connectivity for the past decade — continue to evolve. These pluggables offer multi-vendor interoperability, field-replaceability, and a mature supply chain, but their power consumption — typically 15–25 pJ/bit, including the DSP — makes them the least power-efficient optical option.

Regardless, the 800G generation, using PAM4 modulation at 100G per lane, is now mainstream for backend AI fabrics. The 1.6T generation, pairing Marvell's Ara 3 nm PAM4 DSP with Coherent's or Lumentum's optical engines, has reached volume production — Marvell confirmed at OFC 2026 that Ara devices are now shipping in volume to hyperscale and cloud customers deploying 1.6T pluggable optics for AI data centers. Marvell also expanded the Ara family: Ara T (the first 8×200G transmit-retimed optics DSP), Ara X (1.6T with advanced link reliability), and Petra (the first 3 nm 8×100G to 4×200G gearbox), alongside Aquila M, the first O-band coherent-lite DSP with integrated MACsec security.[64]

The path to 3.2T is now visible. At OFC 2026, Broadcom debuted Taurus, the industry's first 400G/lane optical DSP, paired with the first-to-market 400G electro-absorption modulated laser (EML) and photodiodes. Coherent demonstrated 3.2T transceivers using 400G/lane PAM4 with both differential EML and silicon photonics implementations. Lumentum showed a 1.6T DR4 OSFP prototype using four 400G differential EML lasers as a stepping stone to 3.2T. These demonstrations establish 400G/lane as the next signaling transition, with 3.2T pluggable modules expected in the 2027–2028 timeframe.[65]

Linear-Drive Pluggable Optics (LPO)

LPO eliminates the DSP retimer, replacing it with linear analog equalization that preserves signal fidelity while dramatically reducing power. Typically, LPO achieves 30–40% power savings versus retimed pluggables while retaining multi-vendor pluggable flexibility.[66]

LPO 800G modules began shipping in September 2025, with projections that over one-third of intra-data-center 800G deployments will adopt LPO by 2026–2027. Nokia's optical components division reports a massive 80% power savings versus fully-retimed pluggables — with 20% power advantage over competing LPO solutions. Nokia also claims that its monolithic InP chip design provides vertical integration advantages that competing module vendors cannot match.[67]

LPO continues to show adoption momentum in the industry. The LPO MSA (50 networking, semiconductor, and optics member companies) continues to drive forward with new specifications (e.g., 400G-FR4-LPO).

Near-Packaged Optics (NPO)

NPO occupies a middle ground between CPO's deep integration and pluggable optics' modularity. NPO places optical engines adjacent to, but not on, the switch ASIC, preserving field serviceability while capturing much of CPO's power advantage. NewPhotonics announced its NPC50503 1.6T NPO module in January 2026, providing 1.6 Tbps in a compact form factor. Ciena's acquisition of Nubis Communications for $270 million brought CPO/NPO technology supporting up to 6.4 Tbps of full-duplex connectivity, extending Ciena's domain from inter-DC to intra-DC connectivity.[68]

NPO's serviceability advantage — damaged optics can be replaced without discarding the switching silicon — addresses a practical concern that has slowed CPO adoption among operators accustomed to pluggable transceiver replacement workflows.

[64] Marvell, "Marvell Ushers In the 1.6T Era with Expanded Optical DSP Platform Portfolio," *Marvell Investor Relations*, Mar. 2026. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/1013/

[65] Coherent, "Coherent to Unveil Breakthrough AI-Scale Optical Innovations and Industry Leadership at OFC 2026," *GlobeNewsWire*, 17 Mar. 2026. [Online]. Available: https://globenewswire.com/news-release/2026/03/17/3257303/11543/en/

[66] LPO-MSA, *LPO 800G Specifications*, 2025. [Online]. Available: https://lpo-msa.org

[67] Nokia, Technology analyst briefing, Jan. 2026.

[68] Ciena, "Ciena to Acquire Nubis Communications to Expand Its Inside the Data Center Strategy and Further Address Growing AI Workloads," *Ciena Newsroom*, 22 Sept. 2025. [Online]. Available: https://ciena.com/about/newsroom/press-releases/ciena-to-acquire-nubis-communications-to-expand-its-inside-the-data-center-strategy-and-further-address-growing-ai-workloads

Co-Packaged Optics (CPO)

CPO reached a milestone in 2026, transitioning from perpetual promise to production deployment. The power advantage is clear: CPO achieves approximately 2.5 pJ/bit, compared to 10–12 pJ/bit for copper passive DAC and 15–25 pJ/bit for DSP-based pluggables. This is a 4–5x improvement over copper and up to 10x over DSP optics.

Meta's validation (using Broadcom's Tomahawk 5, TH5) is the most significant evidence of CPO maturity: Broadcom reports over 1 million link-hours of production operation with zero link flaps and 65% power reduction compared to pluggable transceivers, based on Meta's deployment (vendor-characterized; Meta has not independently published these figures). For a single 128-port 800G switch, CPO reduces optical transceiver power from approximately 40 kW to 14 kW, yielding annual operating expense savings of $100,000–200,000 per switch at typical hyperscale power costs.

Broadcom's Tomahawk 6 "Davisson" — the CPO variant of TH6, announced in October 2025 — integrates CPO directly, achieving 102.4 Tbps with up to 512 × 200G ports and a 36.4% power reduction per 800G port versus the prior generation.[69]

NVIDIA's CPO switch portfolio spans both InfiniBand and Ethernet domains. The Quantum-X InfiniBand switch delivers 115 Tbps across 144 × 800G ports, while the Spectrum-X Photonics SN6800 Ethernet switch delivers 409.6 Tbps across 512 × 800G ports.[70] At GTC 2026, NVIDIA indicated that Spectrum-X CPO switches would be in production in H2 2026. They shared that these switches will provide a 5x power efficiency and 10x reliability improvement versus pluggable optics. NVIDIA's roadmap also extends CPO to the scale-up domain: the Feynman architecture (2028) will introduce NVLink 8 CPO, bringing silicon photonics directly to the GPU-to-GPU scale-up interconnect for the first time. This is architecturally significant — until now, CPO has been applied to scale-out switch-to-switch connections; applying it to the latency-critical scale-up domain signals that CPO has achieved the maturity and performance required for the most demanding interconnect tier.[71]

AMD's January 2025 acquisition of Enosemi adds 1.6 Tbps of CPO silicon photonics technology, signaling that CPO is now considered strategic by GPU vendors, not just for optical component companies.[72]

In March 2025, startup Lightmatter announced Passage L200, the world's first 3D co-packaged optics product, available in 32 Tbps and 64 Tbps versions, representing a 5–10x improvement over existing CPO solutions.[73] And in March 2026, Lightmatter demonstrated record-breaking 1.6 Tbps per fiber throughput using a 16-wavelength DWDM architecture, providing 8x more bandwidth per fiber than current NPO and CPO offerings.[74]

Marvell's acquisition of Celestial AI (estimated $3.25–5.5 billion) brings photonic fabric technology with per-chiplet capacity of 16 Tbps (8 Tbps bidirectional). The Celestial AI architecture is protocol-agnostic with sub-200 ns end-to-end latency and a cost target of 5–8 cents per Gbps. A committed hyperscaler customer is deploying Celestial AI CPO with next-generation processors, with Marvell projecting a $500 million run rate by the end of CY27 and $1 billion by the end of CY28.[75]

Coherent demonstrated multi-technology CPO at OFC 2026, combining silicon photonics, VCSEL, and InP-on-silicon technologies within a single co-packaged architecture — validating that CPO is not limited to a single photonic technology and can

[69] T. P. Morgan, "The Third Time Will Be The Charm For Broadcom Switch Co-Packaged Optics," nextplatform, Oct. 17, 2025. https://www.nextplatform.com/connect/2025/10/17/the-third-time-will-be-the-charm-for-broadcom-switch-co-packaged-optics/1635588.
[70] "SN6000 Datasheet," NVIDIA, 2026. https://resources.nvidia.com/en-us-accelerated-networking-resource-library/ethernet-datasheet-spectrum-sn6000-switch.
[71] NVIDIA, "NVIDIA GTC 2026 Keynote," J. Huang, 16 Mar. 2026. [Online]. Available: https://blogs.nvidia.com/blog/gtc-2026-news/
[72] AMD, "AMD Acquires Enosemi to Accelerate Co-Packaged Optics Innovation," *AMD Blogs*, May 2025. [Online]. Available: https://amd.com/en/blogs/2025/amd-acquires-enosemi-to-accelerate-co-packaged-optics-innovation.html
[73] Lightmatter, "Lightmatter Announces Passage L200, the Fastest Co-Packaged Optics for AI," *Lightmatter*, Mar. 2025. [Online]. Available: https://lightmatter.co/press-release/lightmatter-announces-passage-l200-the-fastest-co-packaged-optics-for-ai/
[74] Lightmatter, "Lightmatter Achieves Record 1.6 Tbps Per Fiber," 11 Mar. 2026. [Online]. Available: https://lightmatter.co/press-release/lightmatter-achieves-record-1-6-tbps-per-fiber-to-accelerate-ai-optical-interconnect/
[75] Marvell Technology, "Marvell to Acquire Celestial AI, Accelerating Scale-Up Connectivity for Next-Generation Data Centers," *Marvell Investor Relations*, Dec. 2025. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/1000/

leverage the best-suited approach for each application.[76]

Also at OFC 2026, Ayar Labs and Wiwynn announced a joint reference design for optically connected, rack-scale AI systems: a 100% liquid-cooled, HVDC-enabled rack integrating TeraPHY optical engines with SuperNova ELSFP light sources, designed to scale to 1,024+ accelerators operating as a unified system across multiple racks.[77]

Addressing CPO's long-standing field-serviceability concern directly, Ciena, Coherent, Marvell, Molex, Samtec, and TeraHop announced the Open CPX MSA at OFC 2026 to standardize a pluggable socket and electrical connector interface for co-packaged and near-package optical engines. Open CPX will define connector mechanicals, thermals, electrical pinout, mechanical form factors, and electrical/optical/management interfaces — enabling damaged engines to be replaced without discarding the switch ASIC, and decoupling optical-engine supplier choice from the switch ASIC vendor. This is a multi-vendor counterweight to the vertically integrated CPO path taken by Broadcom and NVIDIA, and it positions socketed CPO as a credible near-term deployment model.[78]

XPO: Extra-Dense Pluggable Optics (12.8 Tbps)

A significant new entrant in the optical form factor landscape emerged at OFC 2026: the XPO (eXtra-dense Pluggable Optics) Multi-Source Agreement, organized by Arista Networks with 45 founding members including Ciena, TeraHop and Amphenol (co-chairs with Arista) as well as vendors such as Coherent, Marvell, Lumentum, Lightmatter, and Linktel. XPO defines a 64-channel, 12.8 Tbps liquid-cooled pluggable module — an 8x bandwidth increase over 1.6T OSFP — achieving 204.8 Tbps per OCP rack unit, a 4x density improvement over current 1.6T OSFP implementations.[79]

XPO addresses the thermal and density gap between conventional pluggable optics and CPO. By integrating a liquid-cooled cold plate capable of dissipating up to 400W per module, XPO enables the bandwidth density needed for next-gen AI fabrics while retaining the field-replaceability and multi-vendor sourcing that operators value in pluggable form factors. XPO supports all industry optical standards (DR, FR, LR, SR, ZR/ZR+) as well as next-generation coherent-lite and emerging interconnect technologies. The infrastructure impact is substantial: Arista projects that XPO-based AI data centers would require 75% fewer racks, saving over 1,050 racks — a 44% reduction in floor space — and translating into significant reductions in construction, power distribution, plumbing, and installation costs. Multiple founding members demonstrated working 12.8T XPO modules at OFC 2026, including Linktel and TeraHop. However, XPO remains at the MSA-plus-demos stage; production modules are not yet available, and volume deployment timelines are unclear.

XPO defines a 12.8 Tbps liquid-cooled pluggable module — an 8x bandwidth increase over 1.6T OSFP — achieving 204.8 Tbps per rack unit.

XPO creates a new option in the optical hierarchy: operators who worry about CPO's serviceability limitations but need density beyond current OSFP can adopt XPO as an intermediate step. Whether XPO delays or accelerates CPO isn't clear — it could act as a bridge technology that buys time for CPO, or it could become the preferred long-term solution if its density proves sufficient for next-gen switching silicon.

[76] Coherent, "Coherent to Unveil Breakthrough AI-Scale Optical Innovations and Industry Leadership at OFC 2026," *GlobeNewsWire*, 17 Mar. 2026. [Online]. Available: https://globenewswire.com/news-release/2026/03/17/3257303/11543/en/

[77] "Ayar Labs and Wiwynn Partner to Bring Co-Packaged Optics to Rack-Scale AI Systems," *Ayar Labs*, Mar. 2026. [Online]. Available: https://ayarlabs.com/news/ayar-labs-and-wiwynn-partner-to-bring-co-packaged-optics-to-rack-scale-ai-systems

[78] Open CPX MSA, "Leading Optical Connectivity Solutions Providers form New Organization to Support Optical Interconnects for AI Data Center Applications," *BusinessWire*, 12 Mar. 2026. [Online]. Available: https://www.businesswire.com/news/home/20260312221163/en/Leading-Optical-Connectivity-Solutions-Providers-form-New-Organization-to-Support-Optical-Interconnects-for-AI-Data-Center-Applications; Open CPX MSA consortium website. [Online]. Available: https://www.opencpxmsa.org/

[79] Arista Networks, "Arista Announces XPO High Density Liquid Cooled Pluggable Optics," *Arista*, 12 Mar. 2026. [Online]. Available: https://arista.com/en/company/news/press-release/23697-pr-20260311

Takeaways: Looking Across the Connectivity Hierarchy

The connectivity stack inside the data center will evolve as a portfolio of purpose-built interconnect domains. At the short-reach extreme, passive copper DAC remains unmatched in energy efficiency and serviceability, but its usefulness is eroding for sub–3-meter top-of-rack links as lane speeds approach 200G and beyond. Active electrical cables extend copper's life modestly into the mid-rack domain, but at a steep energy penalty that increasingly erodes their advantage over optics. From there, DSP-based pluggable optics continue to dominate for their operational-level simplicity and reach flexibility, even as they carry a persistent power tax driven by retiming and signal conditioning. Early deployments for LPO suggest meaningful efficiency gains without sacrificing the hot-swappable operational model that operators value.

Technology	Power (pJ/bit, per end/ module)	Practical Reach	BW per Module / Endpoint	Serviceability	Production Status
Copper passive DAC	~0.1–0.3	1–3 m (800G); limited viability at 1.6T	800G volume; 1.6T not mainstream	Excellent	Volume (800G)
Copper AEC / active electrical cable	~10–13	~3–7 m	800G volume; 1.6T emerging	Excellent	Volume (800G); 1.6T emerging
DSP pluggable (OSFP / QSFP-DD / OSFP-XD)	~15–25	500 m–2 km+	800G volume; 1.6T ramping	Excellent (hot-swap)	Volume
LPO pluggable	~10–13	30 m MMF to 500 m / 2 km SMF	800G shipping; 1.6T sampling / early ramp	Excellent (hot-swap)	Early production / ramp
NPO	~4–8 (target)	~500 m–2 km (DC optics class)	3.2T–6.4T-class optical engines emerging	Fair to good (depends on design)	Demo / sampling / early ecosystem
CPO	~5–6 (current gen, per 1.6T port)	~500 m–2 km (DC optics class)	102.4T shipping; up to 409.6T announced (2026)	Limited, improving (socketed optics emerging)	Early production / initial volume

Beyond pluggables, the industry is pivoting toward architectural shifts rather than incremental module optimization. Near-packaged optics (NPO) and co-packaged optics (CPO) move optical engines closer to the switch ASIC to eliminate long electrical traces, which are now the dominant source of energy loss at 1.6T and beyond. Serviceability becomes the central issue: while pluggables can be replaced in seconds, CPO introduces tighter coupling between optics and silicon, raising questions around failure domains, sparing strategies, and field maintenance. Efforts such as socketed optical engines — including Ciena's recently announced Vesta 200 6.4T CPX[80] — and standardized optical interfaces are emerging to mitigate these concerns. In practice, the data center will likely settle into a heterogeneous architecture: copper for the shortest links (for as long as possible), LPO and DSP pluggables for flexible leaf-spine fabrics, and NPO/CPO for the bandwidth-dense scale-up and scale-out.

Switch Platforms and Silicon Roadmap

The transition to 102.4 Tbps switching capacity is the next step up in scale-out. Broadcom, Cisco, NVIDIA, and now Marvell have announced or are shipping 102.4 Tbps switching ASICs, creating the broadest competitive field at any single switching generation.

[80] "Ciena Unveils the Industry's Highest-Density, Lowest-Power Pluggable Optical Engine to Meet Data Center AI Demands," Ciena Corporation, 2026. https://investor.ciena.com/news-releases/news-release-details/ciena-unveils-industrys-highest-density-lowest-power-pluggable

51.2T → 102.4T: The Current Generation

Four-vendor silicon competition at 102.4 Tbps — Broadcom, Cisco, NVIDIA, and Marvell — creates unprecedented buyer leverage.

By 2025, Broadcom's Tomahawk 5, Cisco's Silicon One G200, and NVIDIA's Spectrum-4 were qualified by Meta for use in AI training fabrics through FBOSS and SAI (switch abstraction interface). This established compatibility between multiple architectures and set the stage for evolution to the next generation of multi-vendor 102.4 Tbps products.

Broadcom Tomahawk 6 (TH6, 102.4 Tbps, 3 nm process technology) reached volume production in Q1 2026, becoming the company's 102.4 Tbps entry. Separately, the CPO variant — Tomahawk 6 "Davisson," announced in October 2025 — integrates sixteen 6.4 Tbps optical engines into the switch substrate, providing up to 512 ports of 200G or 128 ports of 800G with integrated CPO and a claimed 70% reduction in optical interconnect power versus traditional pluggable approaches.[81]

Cisco Silicon One G300, unveiled in February 2026,[82] uses a unique architecture for collective communication. The 102.4 Tbps chip features 200 Gbps SerDes, a 252 MB shared buffer (2.5x larger than G200), and "Intelligent Collective Networking" that Cisco claims improves utilization by 33% and reduces job completion time by 28% versus non-optimized baselines (vendor-reported, based on Cisco's internal simulation comparisons). Cisco indicates the G300 can connect 128,000 GPUs with 750 switches (vs. 2,500 previously). The chip is liquid-cooled. As indication of Cisco's traction in the AI infrastructure business, its Q1 FY2026 hyperscaler orders reached $1.3 billion, with projected FY2026 AI networking revenue of $3 billion.[83]

NVIDIA Spectrum-6, in production as of GTC 2026, is the third vendor at 102.4T. Spectrum-6 uses 200G PAM4 SerDes and silicon photonics. NVIDIA cites up to 5x better power efficiency per port versus traditional pluggable transceiver-based networks and a 10x resiliency improvement from a reduced optical-component count. The SN6810 model supports 128 800G ports in 2U. Spectrum-6 is central to the Vera Rubin platform and the Spectrum-X ecosystem. It brings close NIC-switch coordination with ConnectX-9 and sets the path to Spectrum-7 at 204.8 Tbps.[84]

Marvell's next-generation Teralynx at 102.4 Tbps was announced at the company's analyst day in early 2026, with sampling expected in H1 2026. This makes Marvell the fourth vendor to enter the 102.4T switching tier, building on Teralynx 10's position in AI cloud deployments. Production timeline and detailed specifications have not yet been disclosed, but the entry adds competitive pressure and gives customers another option for 2027 fabric builds.[85]

Custom Silicon and Vertical Integration

Custom silicon work is moving much faster now. Broadcom's main partnership is with OpenAI to create AI accelerators.[86] Together they aim to deliver 10 GW of AI accelerators for computing by 2029. Broadcom also works with Meta and Google. Marvell's custom silicon team, strengthened by Celestial AI and XConn deals, now offers enhanced custom connectivity to major customers. As silicon vendors work across the spectrum on XPUs and connectivity with hyperscalers or frontier labs, it increases the likelihood of co-design and vertical integration.

[81] Broadcom, "Broadcom Ships Tomahawk 6," *Broadcom Investor Relations*, Jun. 2025. [Online]. Available: https://investors.broadcom.com/news-releases/
[82] Cisco, "Cisco Announces New Silicon One G300, Advanced Systems and Optics to Power and Scale AI Data Centers for the Agentic Era," *Cisco Newsroom*, 10 Feb. 2026. [Online]. Available: https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m02/cisco-announces-new-silicon-one-g300.html
[83] Cisco Systems Inc. "CISCO REPORTS FIRST QUARTER EARNINGS." Cisco.com, Cisco Systems Inc., 12 Nov. 2025, https://investor.cisco.com/news/news-details/2025/CISCO-REPORTS-FIRST-QUARTER-EARNINGS/default.aspx. Accessed 1 Apr. 2026.
[84] NVIDIA, "NVIDIA Kicks Off the Next Generation of AI with Rubin — Six New Chips, One Incredible AI Supercomputer," *NVIDIA Newsroom*, 5 Jan. 2026. [Online]. Available: https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer
[85] C. Koopmans, "Industry Analyst Day 2025," *Marvell Technology*, 9 Dec. 2025.
[86] Broadcom, "Broadcom and OpenAI Collaboration," *Broadcom Investor Relations*, 13 Oct. 2025. [Online]. Available: https://investors.broadcom.com/news-releases/

NVIDIA's Vera Rubin platform (a poster child for co-design) began production at GTC 2026. It combines seven chips: Rubin GPU, Vera CPU, NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU, Spectrum-6 Ethernet switch, and Groq 3 LPU. This creates a comprehensive, single-vendor AI stack. Also at GTC 2026, NVIDIA introduced BlueField-4 STX: a modular reference design for faster storage.[87] It solves data-access limits for AI inference and claims up to 5x higher token throughput and 4x better energy efficiency over CPUs. This level of integration enables NIC-switch co-design, allowing functions to shift between the NIC, DPU, and switch for optimal performance and efficiency.

The Feynman roadmap for 2028 extends NVIDIA's deep integration approach. Spectrum-7 will reach 204.8 Tbps. NVLink 8 will come in both copper and CPO versions, followed by ConnectX-10, BlueField-5, and the Rosa CPU. Feynman also brings silicon photonics CPO for NVLink scale-up.[88]

In-Network Compute Integration

Related to the in-network computing topic earlier, today's switch ASICs are adding more compute to speed up collective operations. In addition to NVIDIA in-network collective operations support in its networking solutions, Broadcom's Tomahawk Ultra can also perform native AllReduce in switch buffer memory. And Cisco's G300 puts "Intelligent Collective Networking" into its silicon. The UEC 1.0 standard sets interfaces for these features to enable multi-vendor compatibility. With these advancements, in-network computing is moving from single-vendor to ecosystem-wide support.

Network Operating Systems

SONiC has matured into the leading open network operating system for AI data center networks. 650 Group projects SONiC-based data center switching revenue will surpass $5 billion in 2026, with continued strong growth expected through 2027.[89] Azure's early support, Meta's FBOSS framework, Neoclouds' desire for openness, and the Switch Abstraction Interface (SAI) all drive this progress. Ongoing evidence around TCO cost savings helps, with operators reporting up to 50% TCO reduction versus proprietary alternatives.[90]

What AI Fabrics Need from a NOS

AI-scale fabrics impose distinct requirements on network operating systems that differ materially from traditional enterprise or cloud networking. The critical NOS functions for AI clusters include: RDMA/RoCEv2 transport configuration (PFC, ECN thresholds, DCQCN tuning) — misconfiguration here directly degrades collective operation performance; adaptive routing and load balancing across ECMP paths, which is essential for avoiding hot spots in rail-optimized and fat-tree topologies; fast convergence after link or switch failures, where recovery times of seconds rather than minutes determine whether a training job can continue or must restart; telemetry integration with GPU-level observability, correlating fabric events with NCCL timeouts and training stalls; and multi-vendor silicon abstraction, allowing operators to mix Broadcom, Cisco, and NVIDIA ASICs without maintaining separate NOS stacks.

SONiC addresses multi-vendor abstraction well through SAI, and its open-source model enables operators to customize RDMA transport tuning and telemetry pipelines. However, proprietary NOSes will likely retain advantages in specific areas around

[87] NVIDIA, "NVIDIA Launches BlueField-4 STX Storage Architecture With Broad Industry Adoption," *GlobeNewsWire*, 16 Mar. 2026. [Online]. Available: https://globenewswire.com/news-release/2026/03/16/3256640/0/en/

[88] A. Shilov, "Nvidia updates data center roadmap with Rosa CPU and stacked Feynman GPUs — optical NVLink, Groq LPUs with NVFP4, and NVLink also on deck," Tom's Hardware, Mar. 17, 2026. [Online]. Available: https://www.tomshardware.com/pc-components/gpus/nvidia-updates-data-center-roadmap-with-rosa-cpu-and-stacked-feynman-gpus-optical-nvlink-groq-lpus-with-nvfp4-and-nvlink-also-on-deck

[89] N. Weinberg, "8 hot networking trends for 2026," Network World, Feb. 03, 2026. https://www.networkworld.com/article/4126582/8-hot-networking-trends-for-2026.html

[90] SONiC Foundation, "SONiC Foundation Accelerates Ecosystem Growth and Global Adoption as the Leading Open Source NOS Optimized for Enterprise AI Workloads," *Linux Foundation*, Jul. 2025. [Online]. Available: https://www.linuxfoundation.org/press/sonic-foundation-accelerates-ecosystem-growth-and-global-adoption-as-the-leading-open-source-nos-optimized-for-enterprise-ai-workloads

scalable routing, proprietary AI acceleration hooks, and vendor-specific telemetry and troubleshooting. The operational burden of open NOS adoption is real — organizations without hyperscaler-level engineering resources should evaluate commercial SONiC distributions (see below) rather than running community SONiC directly, as production AI fabrics require validated configurations, tested upgrade paths, and vendor-backed SLAs for hardware-software interoperability.

SONiC Adoption Trajectory

SONiC lets organizations run different switching silicon — Broadcom, Cisco, NVIDIA, Marvell — in a single network. This eliminates the need for custom drivers or extensive NOS work. This shift moves networks from closed, vertically integrated stacks to flexible, modular systems. Now, chips, operating systems, and applications can advance independently.

The SONiC ecosystem has expanded considerably beyond its Microsoft origins, with Broadcom providing enterprise-grade SONiC through multiple partners, and community SONiC advancing rapidly. Commercial distributions address market segments that traditionally depended on proprietary solutions:

- **Aviz Networks** offers the Open Networking Enterprise Suite (ONES) with broad chip support: NVIDIA Spectrum-X, Cisco Silicon One, Marvell, and Broadcom. Their AI-powered Network Copilot provides proactive automation. Aviz's Certified Community SONiC gives production-ready software, enterprise SLAs, and 500+ automated feature tests.
- **BE Networks** is a company focused on intent-based networking. It provides network management with SONiC support across Broadcom and Spectrum-X platforms.
- **Dell Enterprise SONiC Distribution** extends SONiC to NVIDIA Spectrum-X platforms with enterprise SLAs.
- **Hedgehog** productizes SONiC into the "Open Network Fabric" for enterprise and cloud-builder deployments, with active deployments across financial services, healthcare, biotech, telecommunications, energy, manufacturing, and logistics sectors.
- **Upscale AI** is likewise extending SONiC to support their scale-out switch offerings based on NVIDIA Spectrum-X silicon.[91]

Other switching startups, such as Aria Networks and Nexthop AI, also use SONiC with their own added features as part of their network operating system strategies.

Vendor NOS Landscape

Beyond SONiC, several vendor-specific NOSes remain strong. Nokia's SR Linux includes an Event-Driven Automation (EDA) framework. This allows agentic AI to automate network responses to congestion, failures, or anomalies.[92] Arista's EOS is dominant in hyperscale and cloud environments, with 150 million ports shipped. Arrcus's ArcOS deeply integrates with NVIDIA BlueField-4 DPUs to enable hardware-accelerated networking operations using the DPU's 64-core Grace ARM CPU.[93] DriveNets DNOS, part of the DriveNets AI Fabric solution, provides a scalable programmable solution spanning edge to core networks, and into the data center. Meta's FBOSS, though not sold commercially, sets an architectural template that many NOS vendors emulate.[94]

[91] Upscale AI, "Upscale AI Supercharges Open, Heterogeneous Scale-Out AI Clusters with NVIDIA Ethernet Switch Silicon," *PR Newswire*, 11 Mar. 2026. [Online]. Available: https://prnewswire.com/news-releases/upscale-ai-supercharges-open-heterogeneous-scale-out-ai-clusters-with-nvidia-ethernet-switch-silicon-302710175.html

[92] Nokia, "Nokia Strengthens AI Data Center Performance," *Nokia Newsroom*, 13 Nov. 2025. [Online]. Available: https://nokia.com/about-us/newsroom/

[93] Arrcus, "Arrcus Harnesses NVIDIA BlueField-4 to Power Gigascale AI Factories," *Arrcus*, 28 Oct. 2025. [Online]. Available: https://arrcus.com/news/arrcus-harnesses-nvidia-bluefield-4

[94] Aviz Networks, "Aviz to Accelerate AI Networking with ONES and NVIDIA Spectrum-X," *Business Wire*, 4 Mar. 2025. [Online]. Available: https://businesswire.com/news/home/20250304144656/en/

Intent-Based Automation and AI-Powered Operations

Intent-based networking has moved from aspiration to production. Juniper's Apstra (now HPE Data Center Director), Nokia's EDA, and Aviz's Network Copilot all provide varying degrees of automated fabric provisioning, anomaly detection, and remediation. The underlying trend is the convergence of AI workload management with AI-powered network operations, using the same machine learning techniques deployed on the cluster to manage its network infrastructure.

Robustness, Timing, Observability, Simulation, and Verification

At scales over 100,000 GPUs, network reliability is now essential. Failures may happen every few minutes, with downtime costing thousands per minute in wasted computing power. So, tools that keep clusters healthy are crucial for keeping large AI operations profitable.

AI Workload Profiling and Network Telemetry

Standard switch monitoring (port counters, SNMP traps, syslog) is not enough for AI tasks. To get full visibility, you need to track GPU-to-GPU path quality, collective operation delays, and flow congestion across the NIC, network, and application layers.

NVIDIA's NCCL profiler plugins break down timing for each collective operation. This helps operators spot if AllReduce or AllGather actions are slowed by network or compute time. Arista's EOS telemetry provides sub-second flow-level stats. Startup Clockwork's FleetIQ offers hardware-independent telemetry that works with NVIDIA and AMD GPUs, InfiniBand, Ethernet, and accelerators such as AWS Trainium.[95]

The emerging architectural pattern is three-tier telemetry: application-level collective profiling (NCCL plugins, RCCL profiler), fabric-level flow telemetry (switch counters, INT/in-band telemetry), and infrastructure-level health monitoring (memory ECC errors; NIC link status; physical connectivity conditions including optics, transceivers, cables; power supply health). Correlating events across these tiers — e.g., failing optics that triggers an NCCL timeout that appears as a training stall — requires unified observability platforms that few organizations have deployed. This is where we believe recent switching startups such as Aria Networks intend to differentiate, using deeper telemetry coupled with specialized AI and agentic capabilities, or what Aria calls "deep networking," to optimize training and inference in data centers.

GPU-to-GPU Observability and Troubleshooting

Meta's NCCLX extends observability into the collective library itself: fault localization identifies the specific GPU, NIC, or switch responsible for a training slowdown, reducing mean time to diagnosis from hours to minutes.[96]

Other alternatives include the previously mentioned startup Clockwork's probe-mesh architecture, which addresses the GPU-to-GPU visibility gap by deploying lightweight monitoring agents on every GPU node, creating a full-mesh measurement fabric that continuously probes latency, jitter, and packet loss across all GPU pairs.

Fault Tolerance and Self-Healing Mechanisms

The economic case for fault tolerance is compelling at a 100,000-GPU scale. The mechanisms for minimizing blast radius and recovery time have matured:

- **Pre-deployment health checks**: Comprehensive GPU, NIC, and link validation before admitting nodes to the training cluster, reducing in-job failures.

[95] Clockwork, "FleetIQ Platform," *Clockwork*, 10 Sept. 2025. [Online]. Available: https://clockwork.io/platform/
[96] M. Si et al., "Collective Communication for 100k+ GPUs," arXiv preprint, arXiv:2510.20171, Oct. 2025. [Online]. Available: https://arxiv.org/abs/2510.20171.

- **NCCL Communicator Shrink**: Removes failed GPUs from the communication group without restarting the entire job, enabling training to continue with a reduced device count.
- **NCCLX (Meta) Fault-Tolerant AllReduce**: Completes collective operations even when individual participants fail mid-operation.
- Meta's **TorchFT** (Fault-Tolerant training framework), released in 2025, provides automatic checkpointing and recovery mechanisms that allow training to continue past GPU failures, NIC failures, and network partitions without human involvement.[97]
- **Clockwork TorchPass**: Combines network path failover (via NCCL/RCCL net plugin) with live GPU migration to sustain training through link flaps, GPU faults, and full node crashes without checkpoint restarts. Independent benchmarking on a GPT-OSS-120B run (TorchTitan, 64 H200 GPUs) showed higher MFU than both checkpoint-restart and TorchFT, with planned migrations completing in under two minutes. Clockwork claims 95% reduction in wasted training progress.[98]

Simulation, Digital Twins, and CPU Emulation

Validating network changes at 100,000-GPU scale is prohibitively expensive when done on production hardware. Several approaches have emerged for pre-validation:

CPU emulation (NCCLX): Validates collective algorithms at 96,000-GPU scale using only 3,000 servers by emulating GPU communication modes on CPUs. This allows engineers to test fault scenarios — network delays, packet loss, stragglers — at a fraction of the cost of GPU testing. The approach was used to validate NCCLX's fault-tolerant algorithms before production deployment.[99]

Network digital twins replicate the fabric topology, traffic patterns, and failure modes in software, enabling pre-validation of routing changes, firmware upgrades, and topology modifications. Several vendors offer digital twin capabilities, though production readiness varies: most deployments remain in the pilot stage as of early 2026.

Hardware-in-the-loop simulation pairs physical switch hardware with emulated endpoints, providing higher fidelity than pure software simulation while remaining more cost-effective than full-scale testing.

CPU emulation for collective algorithm validation is production-proven. Full network digital twins that accurately predict fabric behavior under realistic workloads remain aspirational, with today's implementations capturing topology and routing behavior but struggling to model congestion dynamics plus timing-sensitive interactions. We expect this to improve as vendors such as NVIDIA continue to invest in network and infrastructure digital twins. NVIDIA's portfolio includes the Vera Rubin DSX AI Factory Reference Design, the Omniverse DSX Digital Twin Blueprint[100], and DSX Air — a network-specific simulation platform that lets operators model and validate fabric topology, connectivity, and bandwidth constraints before hardware deployment.[101]

[97] Meta Engineering, "TorchFT: Fault-Tolerant Training for PyTorch," *GitHub*, 2025. [Online]. Available: https://github.com/pytorch/torchft
[98] Clockwork, "Clockwork.io Introduces A New Class of Fault Tolerance to End Failure-Driven GPU Waste in AI Training," ACCESSWIRE Newsroom, Mar. 11, 2026. [Online] Available: https://www.accessnewswire.com/newsroom/en/computers-technology-and-internet/clockwork.io-introduces-a-new-class-of-fault-tolerance-to-end-fai-1145681
[99] M. Si et al., "Collective Communication for 100k+ GPUs," arXiv preprint, arXiv:2510.20171, Oct. 2025. [Online]. Available: https://arxiv.org/abs/2510.20171.
[100] NVIDIA, "NVIDIA Releases Vera Rubin DSX AI Factory Reference Design and Omniverse DSX Digital Twin Blueprint with Broad Industry Support," *NVIDIA Newsroom*, 2026. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-releases-vera-rubin-dsx-ai-factory-reference-design-and-omniverse-dsx-digital-twin-blueprint-with-broad-industry-support
[101] NVIDIA, "NVIDIA DSX Air Boosts Time to Token With Accelerated Simulation for AI Factories," *NVIDIA Blog*, 2026. [Online]. Available: https://blogs.nvidia.com/blog/dsx-air-simulation-ai-factories/; product page: https://www.nvidia.com/en-us/networking/ethernet-switching/air/

Observations and Recommendations

Ethernet's Dominance in New AI Fabric Builds

Ethernet's capture of the majority of new AI fabric builds by mid-2025 represents a structural shift in the market.[102] As detailed in the Scale-Out section, deployments such as xAI's Colossus validate lossless Ethernet at multi-100K GPU scale, and the InfiniBand cost premium has become difficult to justify at hyperscale cluster sizes. InfiniBand retains relevance for tightly coupled clusters where microsecond-level latency improvements justify the premium, and it remains the incumbent in many existing deployments, but the direction of new builds has moved decisively toward Ethernet.

Multi-Vendor Competition at 102.4 Tbps Changes Buyer Dynamics

For the first time at the 102.4T tier, four silicon vendors are competing: Broadcom Tomahawk 6 (shipping), Cisco G300 (shipping H2 2026), NVIDIA Spectrum-6 (in production), and Marvell's next-generation Teralynx (announced, sampling H1 2026 — see Switch Silicon section). This makes design wins contestable, gives pricing negotiations credible alternatives, and forces continuous innovation. System vendors and startups further expand the competitive ecosystem by offering differentiated software stacks and optics integration atop merchant silicon. However, software qualification and supply-chain integration still reduce real-world interchangeability — multi-vendor strategies require ongoing engineering investment.

The Optics Form Factor Deluge

The optical interconnect landscape is a multi-front competition spanning DSP pluggables, LPO, XPO, NPO, and CPO — each at a different maturity level and targeting a different combination of bandwidth density, power efficiency, and serviceability (see Copper and Optical section for details). The practical implication: organizations should evaluate their optics strategy across this spectrum rather than making a binary pluggable-vs.-CPO decision, recognizing that CPO is entering production while XPO and NPO are still emerging.

Scale-Up Competition Is Real

The scale-up market has moved from single-vendor dominance to genuine competition. NVIDIA's NVLink remains the incumbent with shipping hardware across multiple configurations; AMD's Helios has matched NVL72's aggregate bandwidth with open-standard UALink (announced, with production hardware expected late 2026). The April 2026 publication of four UALink 2.0 specifications — including in-network compute and manageability — signals the consortium's ambition to match NVLink's feature set.[103] Multiprotocol silicon from Credo, Marvell, and Upscale AI signals market expectations of coexistence. The pragmatic recommendation: maintain optionality with modular cluster architectures that can accommodate different interconnect standards as they mature.

[102] M. Cooney, "Nvidia networking roadmap: Ethernet, InfiniBand, co-packaged optics will shape data center of the future," *Network World*, 3 Sep. 2025. [Online]. Available: https://networkworld.com/article/4050881/nvidia-networking-roadmap-ethernet-infiniband-co-packaged-optics-will-shape-data-center-of-the-future.html

[103] L. Clark, "UALink 2.0 arrives before version 1 chips have even shipped," *The Register*, 7 Apr. 2026. [Online]. Available: https://www.theregister.com/2026/04/07/ualink_2_specs/

Extreme Co-Design Deepens NVIDIA's Moat

Extreme co-design — architecting GPU, CPU, networking, memory, storage, power delivery, cooling, and software as a single integrated system — has become NVIDIA's defining competitive strategy.[104] Since the $6.9 billion Mellanox acquisition in 2020,[105] NVIDIA's networking revenue has grown from roughly $3 billion in FY2021 to over $31 billion in FY2026.[106] The Vera Rubin platform, with seven co-designed chips, and the Kyber rack architecture, which arrives as a factory-integrated unit, exemplify this approach. The competitive implication: AMD's Helios, UALink, and merchant Ethernet are technically credible alternatives at the component level, but they face an integration gap in 2026–2027. NVIDIA's systems arrive validated; open-ecosystem alternatives require customers to assemble and qualify multi-vendor stacks. ESUN, UALink, and UEC are designed to close this gap, but it will take time.

> **"The reason why extreme co-design is necessary is because the problem no longer fits inside one computer to be accelerated by one GPU." — Jensen Huang**

For enterprises and smaller operators, the trade-off is concrete: NVIDIA's co-design premium buys integration risk reduction and faster time-to-production, while open-ecosystem approaches trade integration effort for vendor optionality and potentially lower TCO.[107] Organizations should make this trade-off explicit rather than defaulting to either approach.

Power Is the Binding Constraint

The shift from bandwidth to power as the limiting factor pervades every technical decision (see Copper and Optical section). XPU power escalation toward 4,000W+ by 2027, copper reach compression at higher data rates, and AI's growing share of national electricity consumption (projected to reach 8–12% by 2030) make efficiency the primary evaluation criterion across the stack. The emergence of XPO, liquid-cooled optics, and CPO at multiple tiers reflects the industry's response to this imperative.

Recommendations for Enterprises

Enterprises building clusters for fine-tuning or inference should consider SONiC-powered switches along with proprietary solutions. In particular, commercial distributions from third parties make this viable without hyperscaler engineering resources.

Deploy now: 800G Ethernet with RoCEv2; SONiC-based NOS; rail-optimized topologies for training-dominated clusters (30–50% switch cost reduction) or fat-tree for mixed workloads; DSP pluggable or LPO optics.

Pilot in 2026–2027: Evaluate Broadcom Thor Ultra (now sampling) as a UEC-compliant NIC alternative to ConnectX-9. Assess where disaggregated inference — separating prefill and decode phases — may improve hardware efficiency for your workload mix; NVIDIA's Dynamo provides orchestration. Evaluate CPO for high-density tiers as Broadcom and NVIDIA ship production CPO switches.

Watch for 2027+: UEC/UET-compliant fabrics as ecosystem matures. Dedicated decode hardware (Groq LPU and potential competitors). 1.6T optics and XPO form factor evolution.

[104] J. Huang, "Jensen Huang on Extreme Co-Design," interview, *Lex Fridman Podcast*, Mar. 2026. [Online]. Available: https://lexfridman.com/jensen-huang-transcript/#chapter1_extreme_co_design_and_rack_scale_engineering

[105] NVIDIA, "NVIDIA Completes Acquisition of Mellanox, Creating Major Force Driving Next-Gen Data Centers," *NVIDIA Newsroom*, 27 Apr. 2020. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-completes-acquisition-of-mellanox-creating-major-force-driving-next-gen-data-centers

[106] D. Martin, "Nvidia Hits Record Quarterly Growth, Says It's The 'World's Largest Networking Business,'" Crn.com, 2026. https://www.crn.com/news/ai/2026/nvidia-hits-record-quarterly-growth-says-it-s-the-world-s-largest-networking-business

[107] NVIDIA, "NVIDIA Enterprise AI Factory Validated Design," *NVIDIA Blogs*, Mar. 2026. [Online]. Available: https://blogs.nvidia.com/blog/bluefield-cybersecurity-acceleration-enterprise-ai-factory-validated-design/

Recommendations for Data Center Operators and CSPs

Deploy now: Qualify 102.4T ASICs — Broadcom TH6 (shipping), Cisco G300 (shipping H2 2026), and NVIDIA Spectrum-6 (in production), with Marvell Teralynx as a fourth option for 2027 builds — for competitive leverage, and evaluate system vendors (Nokia IXR-H6, Juniper/HPE QFX5250, DriveNets 2600SL) on merchant silicon. Plan 800ZR+ DCI capacity for multi-campus architectures. Invest in fault-tolerant training infrastructure — the ROI is compelling at significant annual savings from improved effective training time.

Pilot in 2026–2027: Begin optics form factor evaluation across the full hierarchy (LPO, NPO, CPO); XPO may yield a viable intermediate path for operators requiring field serviceability at high density in 2027. For scale-up, UALink switches arriving starting late 2026 can offer an alternative to NVLink.

Watch for 2027+: ESUN-based scale-up alternatives. OCP OCS standardization for fabric-tier optimization. Full UEC/UET ecosystem deployment.

Decision Framework: Networking Architecture Guidance for 2026

For reader convenience, here's a unified summary of our recommendations:

Domain	Sensible Choices for 2026	Watch in the Near Term
Scale-up	NVLink (NVL72, etc., for NVIDIA GPUs) and AMD's Helios with UALink (for AMD GPUs) are the practical choices today. Both deliver rack-scale bandwidth in the hundreds of TB/s.	UALink ecosystem hardware (switches from Marvell, Upscale AI) arriving late 2026/early 2027. ESUN-based alternatives likely a 2027 story. Evaluate as they mature for future multi-vendor optionality.
Scale-out fabric	400G/800G 51.2T Ethernet with RoCEv2. Rail-optimized topologies reduce switch cost by 30–50% for training-dominated clusters; fat-tree for workloads with unpredictable traffic (MoE, mixed training/inference). SONiC for vendor-independent NOS.	102.4T switching silicon from Broadcom (TH6, shipping), Cisco (G300, shipping H2 2026), NVIDIA (Spectrum-6, in production), and Marvell (Teralynx, sampling H1 2026) creates buyer leverage — adopt TH6 now if needed, or wait for all four to reach volume. UEC/UET for next-generation transport as compliant hardware becomes available.
Optics	DSP pluggables (400G/800G) and LPO where power savings justify early adoption. Both are production-proven with multi-vendor supply chains.	CPO from Broadcom (TH6 "Davisson" in customer qualification, TH5 "Bailly" in volume production since May 2025) and NVIDIA (Spectrum-6 CPO, expected H2 2026) for high-density spine tiers. XPO as an emerging high-density pluggable alternative — promising but early. NPO as a middle ground. The form factor landscape is evolving quickly; avoid long-term lock-in to a single approach.
Scale-across (DCI)	800ZR/ZR+ for campus and metro interconnects between buildings. 400ZR/ZR+ remains a good option too.	1.6T coherent optics in development; multi-rail amplification (Nokia, Ciena) for high fiber density in multi-campus/multi-span deployments. Ciena shipping optimized photonic line system configuration in volume

		in 2026. Plan DCI capacity for data-parallel gradient sync across sites.
Software & Resilience	SONiC (commercial distributions available from multiple vendors). NCCL/RCCL for collectives. vLLM or SGLang for inference serving. Dynamo orchestration for disaggregated inferencing. Invest in deep telemetry and observability stacks.	Fault-tolerant training tools (TorchFT, TorchPass) maturing rapidly — pilot for clusters where job restart cost is significant. NCCLX if running at Meta-like scale. UEC/UET software ecosystem as it develops.

The technology landscape is shifting fast and many announced products have not yet reached production. Where this report describes emerging technologies, inline maturity markers indicate their current status. Organizations should maintain architectural flexibility and avoid overcommitting to any single roadmap.

Wrap-Up

The data center networking industry for AI faces the convergence of multiple trends that collectively represent a structural market shift.

Several questions that were open when our research program began a few years back have now been resolved. Ethernet has displaced InfiniBand at scale — driven by economics, not technical inadequacy. CPO has transitioned from laboratory curiosity to multi-vendor production deployment. Four-vendor competition at 102.4 Tbps creates greater buyer leverage despite NVIDIA's rack- and data-center-level lock-in. Fault-tolerant training works at the scale of 100,000+ GPUs. Disaggregated inference has moved from concept to dedicated silicon, and open-standard scale-up interconnects have progressed from consortium specifications to announced hardware and emerging ecosystem silicon, with production expected in late 2026 and into 2027.

New questions have emerged. Will UALink and ESUN break NVLink's hold over scale-up? Will ESUN and Ethernet win across all domains eventually? Can NVL1152 (1,152 GPUs in one NVLink domain) deliver on its promise, or will the optical and thermal challenges of 8-rack scale-up prove intractable? Will XPO (12.8T liquid-cooled pluggable) succeed, and is it a bridge to or competitive with CPO? Can Broadcom's Thor Ultra 800G UEC-compliant NIC break NVIDIA's ConnectX dominance in AI NICs? How will the prefill/decode split in disaggregated inference reshape fabric topology requirements? Can UEC 1.0 achieve production readiness in time to prevent further Spectrum-X entrenchment? And will NVIDIA's Feynman roadmap set a pace that open-standards alternatives cannot match? These milestones will determine the vendor landscape and competitive dynamics up to 2028–2029.

The binding constraint has shifted from bandwidth to power. When bandwidth was limited, the solution was straightforward: faster links and more connections. When power is the constraint, the optimization space becomes multi-dimensional — from optical form factor to fabric topology to inference architecture to cluster geography. This shift will impact every technical decision and stimulate innovation over the next decade.

With standardization and different flavors of merchant silicon for scale-up and scale-out, several emerging startups are challenging the incumbents with novel architectural approaches. Organizations building major AI infrastructure should track the milestones identified in this report. What we saw at GTC 2026 and OFC 2026 has accelerated the timeline: the decisive period is not the next 18 months but the next 12 months, as NVL1152, Helios, Spectrum-6 CPO, UALink, ESUN, and UEC-compliant hardware all converge in production during 2026–2027.

Interviews

The following pages contain interviews with key vendors and industry participants, providing additional context and perspective on the technologies and trends discussed in this report.

ARRCUS
NETWORK DIFFERENT™

A discussion with **Keyur Patel**, Founder & CTO, and **Sanjay Kumar**, VP Product Management & Marketing

Most AI networking discussions focus on what happens inside the data center. Arrcus has been making the case that AI is fundamentally a distributed problem. Why?

Sanjay: The assumption that AI workloads will consolidate into a few massive centralized clusters is not holding up. Sovereignty requirements, privacy regulations, power grid constraints, and the momentum toward inferencing are driving workloads to be far more distributed than originally envisioned. Most organizations are running fine-tuning and inference across diverse infrastructure: edge locations, regional data centers, private facilities, and public clouds. The network problem is not just connecting GPUs within a rack — it is stitching distributed resources across geographies into a seamless, secure fabric. We outlined this vision about a year and a half ago, and the industry is now moving exactly in this direction.

Arrcus talks about a unified fabric spanning scale-up, scale-out, and scale-across. What does that mean architecturally?

Keyur: Each point in an AI network has different requirements. Scale-up demands ultra-low latency and lossless transport. Scale-out requires congestion management, entropy handling, and multi-tenancy. Scale-across demands high-performance routing with sub-second convergence, traffic engineering, and core link protection. We have built a single fabric addressing all three — fully programmable, orchestrated on demand, with unified visibility, policy, and configuration management. Operators get a single fabric that eliminates silos between backend GPU connectivity, frontend IT networks, and storage.

One of Arrcus's differentiators is hardware abstraction. Why does that matter in AI networking?

Sanjay: The pace of change sets AI networking apart. We are transitioning from 400G to 800G to 1.6T, and GPU architectures evolve every year. Our operating system abstracts from the underlying hardware, whether running on Broadcom silicon, NVIDIA Spectrum, AMD Pensando, NVIDIA Bluefield NICs, Broadcom Thor, or x86 and ARM CPUs. That gives operators the flexibility to upgrade without re-architecting the network. We are a hardware diversity enabler, giving customers genuine choice and the ability to innovate at the pace of AI.

Scale-across networking seems to be where Arrcus claims the strongest differentiation. What sets you apart?

Keyur: Routing becomes a first-class citizen beyond the data center. Our BGP implementation was built from the ground up, not based on FRR or open-source SONiC, giving us significant advantages in convergence time, scale, and reliability. When a NeoCloud has GPU resources across multiple colocation sites, the challenge resembles a CDN problem: intelligent traffic direction, dynamic bandwidth reconfiguration based on workload patterns, and fast reconvergence when links fail. Inter-data-center links are precious and require core link protection, traffic mirroring, and end-to-end visibility. We believe we are the only independent solution vendor capable of routing at this scale on merchant silicon.

The standards landscape, including UEC/UET, ESUN/SUE-T, is evolving rapidly. How does Arrcus navigate that?

Sanjay: We are firmly in the Ethernet camp and active contributors to UEC and emerging scale-up standards. But not all hardware vendors implement identical UEC specifications, creating interoperability challenges. We normalize those differences through intelligent pod design within our unified fabric, so pods with different switch types and NIC vendors communicate tightly at the software layer. Customers can adopt new standards as they mature without waiting for the entire ecosystem to align.

How do you see AI networking architecture evolving over the next several years?

Keyur: Power consumption will make the single mega-data-center approach increasingly unsustainable, and scale-across networking is becoming mandatory. Within data centers, we expect current scale-up technologies to migrate toward scale-out topologies. This is similar to how core routing chips historically evolved toward edge deployment. Each generation pushes for lower latency in scale-out, while new innovations replace current scale-up approaches as GPU capabilities increase. Operators who invest in a unified, hardware-independent fabric today will be best positioned to ride these transitions without re-architecting every cycle.

Interview with Dudy Cohen
VP, Product Marketing

AI infrastructure spending is dominated by GPUs and accelerators. Why should the industry pay more attention to networking?

Networking is roughly 10 percent of the cost of an AI cluster, but it is about 80 percent of the headache. The network determines whether GPUs sit idle waiting on data or run at full utilization. As clusters scale from thousands to hundreds of thousands of GPUs, the networking fabric becomes the binding constraint on job completion time, bandwidth utilization, and ultimately cost per million tokens. Get the network wrong and you are burning GPU cycles, which is like burning money.

DriveNets advocates strongly for Ethernet in AI. What makes Ethernet viable for workloads that were traditionally InfiniBand territory?

Ethernet is an open standard with a broad supplier ecosystem, and it is evolving rapidly. The critical advancement is the addition of scheduling. We offer two approaches: fabric scheduling, where packets are broken into cells, sprayed across the fabric, and reassembled at the destination with zero packet loss; and endpoint scheduling, where NICs manage traffic using network telemetry. Fabric scheduling delivers the highest performance: on a production 512-GPU H200 cluster, our NCCL benchmark results showed up to 18 percent better performance than InfiniBand and up to 15 percent better than Spectrum-X across all-reduce, all-gather, all-to-all, and reduce-scatter operations. This is a significant milestone: performance was the last argument against moving from InfiniBand to Ethernet, and that argument no longer holds.

How do customers choose between fabric scheduling and endpoint scheduling?

It comes down to scale and performance requirements. For clusters at the current sweet spot of around 8,000 GPUs, which is where we see the strongest demand from NeoCloud providers and enterprises, scheduling on Broadcom Jericho3AI and Ramon3 chipsets in open white-box platforms delivers the highest performance.

We have demonstrated this in production and can scale to 32,000 GPUs. For very large clusters in the tens or hundreds of thousands of GPUs, endpoint scheduling on Tomahawk-based platforms provides a more cost-efficient and near-unlimited scaling path. The two approaches coexist, and we work with customers to select the optimal architecture based on cluster size, workload type, multi-tenancy requirements, and day-one versus long-term scaling needs.

DriveNets started in service provider networking. How does that heritage translate to AI data centers?

DriveNets has been building scheduled fabrics for over a decade, starting with customers like AT&T, where we implemented cell-based fabrics to build highly scalable routers from white-box building blocks. When AI infrastructure demand accelerated, we recognized that the same scheduled fabric technology was ideal for GPU backend networks. That allowed us to move quickly and deploy production solutions with hyperscalers, NeoCloud providers, and enterprise customers. Our architecture also extends naturally into scale-across connectivity between data centers, leveraging our deep buffering and L3 routing capabilities — the same strengths that serve the world's largest carrier networks.

As AI clusters scale and diversify, what metrics should infrastructure builders focus on?

Two metrics define the business. One is "time to first production token" — how quickly you go from investment to production. That is driven by supply chain availability, setup simplicity, and the amount of tuning required. Open Ethernet-based networking with proven scheduling shortens that timeline versus proprietary alternatives. Another is "cost per million tokens" — marginal cost at scale. An efficient, high-utilization network directly reduces that cost by ensuring GPUs spend less time idle. We believe the combination of open Ethernet, scheduling intelligence, and a unified software stack across scale-out, scale-up, and scale-across fabrics is the architecture that optimizes both.

Learn more about DriveNets AI Networking Solution

Interview with Rishi Chugh

VP & GM, Data Center Switching

AI is transforming data center networking. How do you frame the architectural shift that is underway?

The entire networking space is being disrupted. Traditional three-tier architectures are giving way to flatter, two-tier designs that prioritize predictability and reliability. AI workloads, especially pre-training runs that can take months, demand a level of network reliability that was never previously required. If the network cannot sustain lossless, deterministic performance over that duration, the entire job has to be restarted. That changes the engineering calculus fundamentally. At the same time, the protocols themselves are evolving. Standard Ethernet was never designed for these demands, which is why we are seeing innovations like the Ultra Ethernet Consortium driving dynamic load balancing, link-level retry, and credit-based flow control into the fabric.

Scale-out networking remains central to AI clusters. What is driving the requirements, and where does Marvell's Teralynx family fit?

Scale-out is predominantly driven by bandwidth. As we transition from 800G to 1.6T and look ahead to 3.2T and 6.4T connectivity, the switching fabric has to keep pace. Today that means 51 Tbps switches — our Teralynx 10, which is in production — and we have previewed our 100 Tbps fabric, which we will be announcing very soon. That will be followed by a 200 Tbps switch. These are all purpose-built for scale-out, delivering the port density and radix that AI clusters require as they grow from thousands to hundreds of thousands of GPUs.

Scale-up is a newer concept that has gained significant attention. How is a scale-up fabric fundamentally different from scale-out?

Over a year ago nobody was talking about scale-up — now it is critical. The distinction is that scale-out is a network fabric, while scale-up is a compute fabric. In scale-up, you are unifying GPUs into a common pod where they share a pool of memory. The semantics are entirely different: instead of quality of service, broadcast, and multicast, you are dealing with load-store operations, atomics, and memory sharing. Latency becomes paramount. Think of it like the multi-core CPU transition — workloads are distributed across GPUs that must operate in perfect synchronicity. These are the most expensive components in the cluster, and you need a fabric that ensures no GPU is starved or oversubscribed so you get maximum return on that investment.

Marvell builds both custom XPUs for hyperscalers and merchant switching silicon. How do those capabilities reinforce each other?

That combination is a significant advantage. As a custom compute enabler, we understand the workload demands from the GPU and accelerator side — the bandwidth, the memory semantics, the latency budgets. As a switching silicon provider, we understand the fabric requirements. Very few companies sit at that intersection. It means we can co-optimize across the compute-network boundary rather than engineering each side in isolation, and it informs our roadmap across both scale-up and scale-out.

Multiple scale-up protocols are competing — NVLink, UALink, ESUN. How does Marvell navigate that landscape?

Marvell is in a unique position: we enable all three. We support NVLink through NVLink Fusion. We are contributing members of the UALink consortium. And we are active contributors to the ESUN committee. The reason is straightforward — there is no one-size-fits-all solution. The market is split based on workload requirements, ecosystem preferences, and customer timelines. Our engineering resources are shared across UALink and ESUN switching platforms at the same capacity levels, which gives customers confidence that they are not betting on a stranded roadmap regardless of which protocol prevails.

What should the industry expect from Marvell's data center switching roadmap over the next 12 to 18 months?

We are committed across both scale-out and scale-up (as well as scale-in within the XPU). On scale-out, we are delivering the 100 Tbps and 200 Tbps fabric switches building on Teralynx. On scale-up, we will have UALink fabric swithces at 115 Tbps and 57 Tbps, alongside a dedicated 115 Tbps ESUN fabric. All of these are targeted for availability in the first half of 2027. Combined with our custom XPU capabilities, Marvell is positioned to be a foundational silicon provider across every major AI networking architecture.

Interview with Neel Patel, GM, Optical Components

Nokia has a broad portfolio of AI data center solutions. How should the industry think about what Nokia brings?

Nokia addresses AI data center infrastructure across multiple dimensions. We offer high-performance IP routing built on our own custom silicon, as well as open data center switching solutions leveraging cutting-edge merchant silicon. But where we are uniquely differentiated, and where this conversation will focus, is in optical. With the acquisition of Infinera, Nokia now has one of the deepest vertically integrated optical capabilities in the industry, spanning indium phosphide photonics, coherent DSP technology, tunable laser sources, and module production. We can apply these assets across scale-out, scale-up, and scale-across architectures within and between AI data centers.

How has the Infinera acquisition changed Nokia's position in optical components specifically?

Infinera brings over 20 years of expertise in high-performance, low-power indium phosphide photonic integrated circuits. Historically, that capability served long-haul optical transport. Now, as a fully integrated company, we are entering the data center as a merchant supplier of optical components, including lasers, modulators, PICs, and high-bandwidth driver ICs, to module partners and OEMs serving hyperscale AI infrastructure. We are focused on solving two critical problems: reducing interconnect power consumption inside the data center, and delivering InP manufacturing scale at a time when industry supply is constrained.

As lane speeds scale to 200G and 400G, why does indium phosphide (InP) matter more?

This is a renaissance for indium phosphide. With each generation of speed, silicon photonics faces higher drive voltages, which directly increase power consumption. Our InP modulators operate at significantly lower voltages (low V_{π}), and that advantage compounds at 200G and 400G per lane. InP is also one of the few materials that can generate light, so even silicon photonics platforms require an InP laser source. We demonstrated 200G-per-lane capability early and are confident our modulator technology scales to 400G. Nokia has a mature, proven platform with a clear and growing power advantage.

The industry is debating LPO, NPO, and CPO. How does Nokia play across these approaches?

Reports of the front-panel pluggable's demise are exaggerated. We will see coexistence. Linear pluggable optics are a near-term target; removing the in-module DSP dramatically cuts power, and our integrated PIC and driver IC combination is a strong LPO solution. Near-package optics offer a compelling middle ground: meaningful power savings while preserving supply chain disaggregation, so customers can source their switch ASIC, NPO module, and external laser independently. CPO delivers the most optimized power profile, but comes with tighter vendor coupling. Nokia can participate across all three, and we believe hyperscalers will deploy a mix based on cluster generation, scale, and application.

As coherent optics pushes into shorter distances, how does that position Nokia beyond traditional long-haul?

Coherent technology was once reserved for distances above 80 kilometers, but as speeds increase, it is encroaching closer to—and inside—the data center. Whether it is scale-across DCI links, inter-building campus connections, or emerging intra-data-center applications, Nokia is one of the few companies that owns the entire coherent stack: tunable lasers, DSP technology, and module production. That end-to-end capability allows us to deliver purpose-built coherent solutions optimized for each distance and application as AI workloads span ever-wider infrastructure.

Nokia has onshore US fabrication supported by CHIPS Act funding. How does that change the conversation with hyperscalers?

Five years ago, nobody cared about an onshore fab. Today, it is a critical strategic asset. With geopolitical uncertainty around semiconductor supply chains, our customers find tremendous comfort in US-based fabrication of photonic components. We are honored to have received CHIPS Act funding and are investing significantly beyond that to expand our indium phosphide capacity. As a merchant supplier with onshore production, we offer hyperscalers both supply diversity and supply chain resilience—two things they are increasingly prioritizing as AI infrastructure scales to hundreds of thousands of GPUs per cluster.

Interview with Aravind Srikumar, SVP Products & Marketing, Upscale AI

Most AI networking today is built on cloud-era infrastructure. Upscale AI argues that it is fundamentally insufficient. Why?

AI data centers are being built on cloud principles, not AI realities. AI workloads move differently. They require a completely lossless, synchronized, and highly fault-tolerant network. Networking is the heart that makes a cluster behave as one unit — a single packet drop can stall thousands of GPUs. Today's scale-up and scale-out networks evolved separately, forcing different operational principles and complexity on customers. The network operating systems managing these devices were never built to handle collective communications. Telemetry was an afterthought. Upscale AI exists to deliver AI-specific networking silicon, systems, and software. We do not build for general-purpose, we build for AI.

What is Upscale AI's scope? Where in the AI networking stack do you focus?

Scale-up ties GPUs within a rack or a tightly coupled domain into a single functional unit with unified memory. Scale-out extends those same principles, uniting multiple scale-up domains over larger distances. Scale-across introduces the WAN element, interconnecting data centers across a campus or a metro network. We focus on scale-up and scale-out because that is where most AI efficiency gains are achieved. Model sizes are exploding, making communication the critical path that defines the boundary of a cluster. Copper reach is shrinking, power constraints are tightening, and optical is moving closer to compute. Open standards for AI fabrics are real, and packaging technologies now enable larger radix switches. Five years ago, these advantages did not align. Now they do, and if we do not act, progress stalls.

What does Upscale AI actually deliver, and what is the co-design advantage?

We deliver a complete AI networking solution stack: purpose-built AI networking silicon, AI-optimized systems, and an AI-specific network operating system — all validated across pod and rack architectures. Production systems are scheduled to ship in the second half of 2026 to 2027. The biggest differentiator is co-design. We optimize ASICs, systems, and software together for AI by removing the general-purpose assumptions and delivering deterministic performance. That is fundamentally different from taking a merchant switch designed for cloud workloads and repurposing it for AI.

Openness is a strong theme in Upscale AI's messaging. Why is that so central?

AI customers want control and long-term flexibility. Heterogeneity in compute is inevitable because no two clusters will be the same, and without open, standards-based networking stacks, it is impossible to build clusters optimized for the diverse AI workloads businesses will run in the future. Openness is not optional; it is mandatory. We partner with all GPU and XPU vendors. None are competitors. NVIDIA, for example, is an ecosystem partner that enables our heterogeneous compute vision. We focus on providing the best networking fabric that unites the compute elements from all our partners into a single functioning unit.

Upscale AI takes an agnostic position on copper versus optical connectivity. Why?

Some clusters need the reach that optical provides; others need the efficiency of copper. Both are valid connectivity options, and we are impartial. That principle is built directly into our silicon; it does not limit connectivity choices. Customers can scale their clusters through electrical or optical connectivity as their requirements dictate. As power constraints tighten and optical moves closer to compute, having silicon that accommodates both paths without compromise becomes a critical enabler.

What should potential customers and partners take away from engaging with Upscale AI?

We treat customers as partners, which means enabling true heterogeneous compute. Open standards are mandatory to get there. Partnering with Upscale AI means the networks you build today will not only scale for the future but work with every other network and compute element without concerns about interoperability or vendor lock-in.

upscale ai

Interview with Mansour Karam CEO of Aria Networks

Aria Networks is launching with a strong premise that the network is the highest-leverage point in an AI factory. Can you explain more?

The network touches every component in an AI cluster. It connects all the accelerators, all the storage arrays, and it is the fabric across which every token is ultimately produced. It represents 10 to 15 percent of total cluster cost but can significantly impact overall performance. That asymmetry is the leverage. Jensen Huang has said that AI factories need to be the lowest-cost producers of intelligence. If that is the goal, then the network, which sits at the intersection of every other system, is where the highest-leverage optimization lives.

You frame your discussions around MFU. Why is that the right metric?

Model FLOP Utilization or MFU is simply how many of the theoretical FLOPs in a cluster are actually being applied to the job at hand, whether training or inference. In state-of-the-art systems today, that number hovers around 38 to 40 percent for a 10,000 XPU cluster; and it drops further as clusters scale. That gap represents paid-for capacity sitting idle. MFU is directly proportional to token efficiency, so improving it means you either reduce cost per token or increase throughput. Both matter enormously to how AI factories compete.

How much can network issues actually degrade MFU?

The range is quite sobering. Based on our analysis and simulation work, adding just four milliseconds of uniform latency can degrade MFU by around 13 percent in a 10K XPU cluster. Packet loss compounds that further. Jitter is the worst offender — it can drive degradation as high as 80 percent, which is effectively catastrophic. And these effects can originate from something as mundane as a single bad NIC in a 10,000-unit fleet dropping MFU by 1.7% in a collective operation, or a transceiver with a higher bit-error rate from a speck of dust. If we can help customers achieve a 3 percent MFU improvement, that translates to roughly 7 to 9 percent incremental revenue or about $50 million annually for a 10,000 XPU cluster. Our network more than pays for itself.

What is "Deep Networking," and why did you have to build it from scratch?

Deep Networking is a multi-layered architecture built around telemetry as the first-class citizen — not one focused on configuration state, which is what every prior approach has optimized for. We collect fine-grained telemetry end-to-end: from ASICs, transceivers, cables, NICs, and hosts, down to the microsecond level. Agents embedded at each layer evaluate that telemetry and close the loop at the appropriate resolution and speed — reflexive and near-deterministic at the ASIC layer, more strategic and reasoning-based at the cluster-management layer, and fully agentic and LLM-powered at the operator layer. That architecture did not exist before and could not be bolted onto a legacy system. We started Aria 15 months ago and built it entirely from scratch.

Why does the software model matter as much as the architecture?

Because in AI, innovation is daily — not biannual. Traditional networking vendors ship two releases a year. That cadence was sufficient in the cloud era. Today, inference architectures are evolving faster than most roadmaps can track: a year and a half ago, people were saying inference was a single-GPU problem with no meaningful networking requirements. Now distributed inference, prefill-decode disaggregation, and KV-cache transfers have made it more network-intensive than pre-training in many respects. We have built a continuous upgrade flywheel — the more telemetry we collect, the faster our models improve, and the more we can deliver seamlessly to customers. That compounding loop is core to the architecture, not an afterthought. That's why Aria can deliver outcomes for our customers that other networking vendors cannot.

Interview with Gilad Shainer
Senior Vice President of Networking

AI data centers are increasingly described as "AI factories." What does that change architecturally?

An AI factory is not a traditional data center with one network serving many applications. AI training and inference are distributed computing problems, where thousands, hundreds of thousands, and eventually millions of GPUs and CPUs must operate as one synchronized compute engine. That requires multiple purpose-built network architectures. Scale-up networks connect GPUs into a rack-scale GPU with shared memory semantics. Scale-out connects those rack-scale systems into a larger AI factory. Scale-across connects multiple AI factories so a single workload can span locations. More recently, storage has become a fourth leg of the infrastructure, especially for inference and context memory.

Where does NVLink fit in that model?

NVLink is the foundation for scale-up. Its job is to make multiple GPUs behave like a single virtual GPU, with extremely high bandwidth, very low latency, and support for load-store operations. Today, we focus on NVLink connectivity over copper because it delivers the lowest power and best economics where distances allow. As scale-up domains grow beyond the rack in future generations, optical connectivity becomes necessary, and co-packaged optics will play a larger role.

NVIDIA offers both InfiniBand and Spectrum-X Ethernet for scale-out. Why continue to support both?

InfiniBand was built for distributed computing from the beginning, and it remains a great technology for AI. But accelerated computing is going everywhere, and many customers are more familiar with Ethernet operations, tooling, and management. Spectrum-X is purpose-built Ethernet for AI. We brought many InfiniBand principles into Ethernet: lossless behavior, RDMA optimization, low jitter, synchronization, and the ability to scale to hundreds of thousands, or even millions, of GPUs. Our goal is not generic Ethernet. It is Ethernet engineered for AI factories.

Demand for bandwidth in AI is rising quickly. How does that affect the network roadmap?

Scale-out bandwidth is doubling with each GPU generation. Hopper required 400 Gb/s per GPU, Blackwell moved to 800 Gb/s, Rubin moves to 1.6 Tb/s, and that trajectory continues. When multiplied across hundreds of thousands of GPUs, the fabric becomes enormous. That makes latency, jitter, resiliency, and power consumption central design constraints. Co-packaged optics is important because it places the optical engine much closer to the switch ASIC, reducing electrical distance, lowering power consumption, and improving reliability.

Why does scale-across matter?

Sometimes the compute capacity required for a workload exceeds what can be built in a single location due to power, space, or deployment constraints. Scale-across connects multiple AI factories over longer distances while minimizing latency and jitter. Spectrum-XGS is designed for that gigascale environment, enabling distributed AI factories to operate as one larger compute resource.

Storage is not always part of the AI networking discussion. Why is NVIDIA adding it to your reference architectures?

Inference changes the storage problem. As context windows grow and agentic workloads reuse more KV cache, context memory becomes a first-class infrastructure tier, not just data sitting inside a GPU server. Traditional networked storage was not optimized for that use case, because KV cache is performance-critical but often derived and recomputable. That is why NVIDIA introduced BlueField-4 STX, a modular reference architecture for accelerated storage, and CMX, the first rack-scale implementation of that architecture. CMX creates a purpose-built context memory tier for inference, using BlueField-4 and Spectrum-X Ethernet to move KV cache efficiently across the pod. Our goal is to improve tokens per second, tokens per watt, and GPU utilization by keeping reusable context close to the compute, rather than forcing the system to recompute or retrieve it from conventional shared storage.

What should customers take away from NVIDIA's networking direction?

AI networking is becoming a full-system architecture. NVLink, Spectrum-X, Spectrum-XGS, ConnectX SuperNICs, BlueField DPUs, storage controllers, and NVLink Fusion are all part of our platform. The objective is to let customers build larger AI factories, use custom CPUs or XPUs where needed, reduce development burden, and optimize performance, power, and scale generation after generation.

Appendix: Vendor Profiles

The data center networking ecosystem for AI workloads continues to expand and consolidate simultaneously. This appendix profiles key vendors across six categories. For each vendor, we provide a brief description and a summary of the most significant developments over the past 12 months. Vendor-specific detail is covered in this section; the main report references vendors where they illustrate architectural points.

APPENDIX | A. Integrated Systems Providers

Vendors with end-to-end silicon-to-software stacks spanning GPUs/accelerators, NICs, switches, and networking software.

AMD

AMD's vertically integrated stack spans Pensando DPU/NIC, RCCL collective library, Enosemi CPO (acquired January 2025), and leadership roles in UALink (co-chair) and ESUN (founding member). The Helios rack architecture (see Scale-Up section) targets rack-scale parity with NVIDIA's NVL72. Oracle is committed as the first hyperscaler to offer publicly available MI450 clusters (50,000 GPUs, Q3 2026), and Meta's 6GW deal further validates AMD's position.[108] Platform integration remains less comprehensive than NVIDIA's full-stack co-design.[109]

Intel

Intel participates through Gaudi 3 (24 × 200 Gbps RoCEv2 natively), IPU E2200 (400 Gbps), and UEC standards leadership. A January 2026 Cisco partnership pairs Gaudi 3 with Nexus 9000 for Ethernet-native AI clusters. Meanwhile, Intel Foundry Services is ramping 18A node with advanced packaging capacity exceeding $1B annually, as they attempt to remain relevant in the custom XPU ecosystem.[110]

[108] "Advanced Micro Devices, Inc.," Advanced Micro Devices, Inc., Feb. 24, 2026. [Online]. Available: https://ir.amd.com/news-events/press-releases/detail/1279/amd-and-meta-announce-expanded-strategic-partnership-to-deploy-6-gigawatts-of-amd-gpus

[109] AMD, "AMD Pensando Pollara 400 AI Network Interface Card (NIC)," *AMD Networking*, 2025. [Online]. Available: https://amd.com/en/products/network-interface-cards/pensando.html

[110] Intel, "Intel IPU E2200 Mount Morgan," *Intel*, Jan. 2025. [Online]. Available: https://servethehome.com/intel-ipu-e2200-400g-dpu-at-hot-chips-2025/

NVIDIA

The Vera Rubin platform[111] — in production with seven co-designed chips (see Custom Silicon section) — is the industry's most vertically integrated AI platform, with Jensen Huang projecting $1 trillion in cumulative Blackwell and Vera Rubin orders through 2027.[112] Key GTC 2026 networking announcements — Kyber rack architecture, Spectrum-6 CPO, Groq 3 LPX disaggregated inference[113], BlueField-4 STX storage[114] — are covered in their respective main body sections. The Spectrum-X ecosystem has attracted production commitments from Meta, Oracle, xAI, and AWS. NVLink Fusion partners include SiFive, AWS Trainium4, Fujitsu, Qualcomm, MediaTek, and Astera Labs. NVIDIA holds a $1 billion equity investment in Nokia (2.9% stake). The Feynman roadmap (2028) is discussed in the Scale-Up and Switch Silicon sections.[115]

APPENDIX | B. Merchant Silicon and Semiconductor Providers

Companies providing switching ASICs, SerDes IP, connectivity silicon, and optical components used by multiple system vendors.

Alphawave Semi (now part of Qualcomm)

Qualcomm completed its $2.4 billion Alphawave acquisition (December 2025), integrating 224G SerDes, UCIe chiplet interconnect, and PCIe 7.0 IP. Combined with Arm-based Oryon CPUs, Qualcomm targets energy-efficient AI inference with 224G deployment across retimers, switches, and optical modules starting in 2026.[116]

Astera Labs

Astera Labs provides connectivity solutions across four product families: Scorpio (fabric switches), Aries (PCIe/CXL retimers), Taurus (Ethernet smart cable modules), and Leo (CXL memory controllers). The company has shifted from server-centric to rack-centric AI compute, with Scorpio fabric switches integrating NVLink Fusion support. Scorpio X began volume shipments in January 2026, positioning it as a key PCIe/CXL fabric switch for NVIDIA's NVLink Fusion ecosystem, with Astera collaborating on custom connectivity solutions for Vera Rubin deployments.[117]

[111] NVIDIA, "Inside the NVIDIA Vera Rubin Platform: Six New Chips, One AI Supercomputer," *NVIDIA Technical Blog*, 16 Mar. 2026. [Online]. Available: https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/
[112] NVIDIA, "NVIDIA GTC 2026 Keynote," J. Huang, 16 Mar. 2026. [Online]. Available: https://blogs.nvidia.com/blog/gtc-2026-news/
[113] "GTC 2026: With Groq 3 LPX, Nvidia adds dedicated inference hardware," *The Decoder*, 17 Mar. 2026. [Online]. Available: https://the-decoder.com/gtc-2026-with-groq-3-lpx-nvidia-adds-dedicated-inference-hardware-to-its-platform-for-the-first-time/
[114] NVIDIA, "NVIDIA Launches BlueField-4 STX Storage Architecture With Broad Industry Adoption," *GlobeNewsWire*, 16 Mar. 2026. [Online]. Available: https://globenewswire.com/news-release/2026/03/16/3256640/0/en/
[115] A. Shilov, "Nvidia updates data center roadmap with Rosa CPU and stacked Feynman GPUs — optical NVLink, Groq LPUs with NVFP4, and NVLink also on deck," Tom's Hardware, Mar. 17, 2026. [Online]. Available: https://www.tomshardware.com/pc-components/gpus/nvidia-updates-data-center-roadmap-with-rosa-cpu-and-stacked-feynman-gpus-optical-nvlink-groq-lpus-with-nvfp4-and-nvlink-also-on-deck
[116] Qualcomm, "Qualcomm Completes Acquisition of Alphawave Semi," *Qualcomm*, 18 Dec. 2025. [Online]. Available: https://qualcomm.com/news/releases/2025/12/qualcomm-completes-acquisition-of-alphawave-semi
[117] Astera Labs, "Astera Labs Reports Fourth Quarter and Full Year 2025 Financial Results," *Astera Labs Investor Relations*, 10 Feb. 2026. [Online]. Available: https://asteralabs.gcs-web.com/news-releases/

Broadcom

Broadcom offers a complete Ethernet portfolio spanning scale-up (Tomahawk Ultra), scale-out (TH6/TH6 Davisson CPO), and scale-across (Jericho 4).[118] Broadcom's Tomahawk 6 (TH6, 102.4 Tbps) is in volume production, with the "Davisson" CPO variant in customer qualification.[119] Broadcom Jericho 4 provides deeper buffers for scale-across deployments and Broadcom Thor Ultra delivers the first 800G UEC-compliant NIC (now sampling).[120] At OFC 2026, Broadcom debuted Taurus, the first 400G/lane optical DSP, establishing the pathway to 3.2T optics. Broadcom's CPO validation and switch silicon details are covered in their respective main body sections. Meanwhile, the OpenAI partnership with Broadcom targets 10 GW of custom compute through 2029, continuing to strengthen's Broadcom position in the adjacent custom XPU space.

Credo

Credo specializes in active electrical cables (AECs), optical DSPs, and SerDes IP. The Bluebird 1.6T optical DSP ramped in September 2025, and the Blue Heron 224G multiprotocol retimer (supporting UALink, ESUN, and Ethernet simultaneously) launched in January 2026, endorsed by AMD and Upscale AI. At OFC 2026, Credo showcased significant NVIDIA platform design wins: 1.6T AECs powering Vera Rubin NVL144 and the Kyber Ultra NVL576 platform, confirming Credo's position as a key interconnect supplier for NVIDIA's scale-up architecture. Credo also demonstrated 400G/800G ZeroFlap optical transceivers with real-time link telemetry.[121]

Marvell

Aggressive M&A — Celestial AI ($3.25–5.5B, photonic fabric) and XConn ($540M, PCIe/CXL switching) — gives Marvell span across PAM4 DSPs, custom ASICs, CXL switches, photonic fabric, and now switching silicon in all three scaling domains. The next-generation Teralynx at 102.4 Tbps (announced at analyst day early 2026, sampling H1 2026) extends Marvell's Ethernet switch portfolio alongside the Teralynx 10 (51.2T, in volume production), which drove over $300M in switch ASIC sales in FY2026.[122] Celestial AI CPO is protocol-agnostic (UALink, ESUN, NVLink Fusion), with committed hyperscaler customer orders.[123] At OFC 2026, Marvell confirmed Ara volume shipments and expanded the family[124], joined both the XPO and Open CPX MSAs as a founding member, and demonstrated rack-level OCS with Lumentum (see Physical Layer and OCS sections).[125]

[118] Broadcom, "Broadcom Showcases Industry-Leading Solutions for Scaling AI Infrastructure at OFC 2026," *Broadcom Investor Relations*, 12 Mar. 2026. [Online]. Available: https://investors.broadcom.com/news-releases/news-release-details/broadcom-showcases-industry-leading-solutions-scaling-ai

[119] Broadcom, "Broadcom Ships Tomahawk 6," *Broadcom Investor Relations*, Jun. 2025. [Online]. Available: https://investors.broadcom.com/news-releases/

[120] Broadcom Inc., "Broadcom Introduces Industry's First 800G AI Ethernet NIC," Broadcom.com, 2026. https://www.broadcom.com/company/news/product-releases/63641

[121] Credo, "Credo to Showcase Optical Solutions for AI Scale-Out Fabrics at OFC 2026," *Credo Investor Relations*, Mar. 2026. [Online]. Available: https://investors.credosemi.com/news-events/news/news-details/2026/Credo-to-Showcase-Optical-Solutions-for-AI-Scale-Out-Fabrics-at-OFC-2026/

[122] C. Koopmans, "Industry Analyst Day 2025," *Marvell Technology*, 9 Dec. 2025.

[123] Marvell Technology, "Marvell to Acquire Celestial AI, Accelerating Scale-Up Connectivity for Next-Generation Data Centers," *Marvell Investor Relations*, Dec. 2025. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/1000/

[124] Marvell, "Marvell Ushers In the 1.6T Era with Expanded Optical DSP Platform Portfolio," *Marvell Investor Relations*, Mar. 2026. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/1013/

[125] Marvell, "Marvell and Lumentum to Demonstrate Optical Circuit Switching for Next-generation AI Scale-up Infrastructure," *Marvell Newsroom*, 16 Mar. 2026. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/1015/marvell-and-lumentum-to-demonstrate-optical-circuit-switching-for-next-generation-ai-scale-up-infrastructure

APPENDIX | C. Hyperscalers and Cloud Providers

Organizations that both consume and shape networking technology at massive scale.

AWS

AWS builds AI cluster infrastructure through custom Trainium accelerators integrated with NVIDIA NVLink Fusion, with Trainium4 delivering 6x FP4 throughput while integrating NVLink 6 and MGX rack architecture for hybrid Trainium/NVIDIA GPU clusters. Its EC2 UltraClusters (accelerated EC2 instances) interconnect thousands of instances via EFA in petabit-scale non-blocking networks, with P6e-GB200 UltraServers delivering up to 28.8 Tbps of EFAv4 networking and its proprietary Nitro System v5 showing 35% latency improvement over P5 instances.

Google / Alphabet

Google operates the largest custom OCS fabric in the industry with its Jupiter data center networking operating at 13 Petabits/sec. It is also a founding UALink Consortium member driving the 1.0 specification for open 200 Gbps per-lane scale-up interconnects, and co-led the OCP OCS subproject. Its seventh-generation Ironwood TPU enters general availability in 2026, with details of its eighth-generation TPU expected this year. Parent company Alphabet directed 60% of its record $91–93 billion 2025 CapEx toward servers and 40% towards data center and networking equipment; its 2026 CapEx is expected to be $175 billion to $185 billion.

Meta

Meta operates one of the largest publicly disclosed AI training clusters (notwithstanding the other frontier labs such as OpenAI, Anthropic, and xAI), with a 129,000-GPU deployment spanning five buildings and a maximum GPU-to-GPU distance of 3 km. Meta qualifies three ASIC vendors (Broadcom TH5, Cisco G200, NVIDIA Spectrum-4) for its non-scheduled fabric, and operates the Disaggregated Scheduled Fabric (DSF) on Broadcom's Jericho3-AI/Ramon3 combination. Importantly, Meta's CPO validation with Broadcom (1M+ link-hours, zero flaps, 65% power reduction) provided the industry's most significant CPO production evidence. Likewise, Meta was the earliest and largest 800ZR+ customer for DCI.[126]

Microsoft Azure

Microsoft created and maintains SONiC, now powering 180,000+ Azure switches with a reported 2% device failure rate (lower than proprietary alternatives). Azure launched CXL-enabled instances (November 2025) and deployed the first production GB300 NVL72 cluster with 4,600+ Blackwell Ultra GPUs on Quantum-X800 InfiniBand at 800 Gbps per GPU. A founding UEC member with a five-year Nokia fabric agreement across 30+ countries, Microsoft's "Fairwater" AI superfactories aim to integrate hundreds of thousands of Vera Rubin Superchips into a single flat network via a dedicated AI WAN backbone.

[126] Meta Engineering, "Disaggregated Scheduled Fabric: Scaling Meta's AI Journey," *Engineering at Meta*, 20 Oct. 2025. [Online]. Available: https://engineering.fb.com/2025/10/20/data-center-engineering/disaggregated-scheduled-fabric-scaling-metas-ai-journey/

Oracle Cloud (OCI)

Oracle announced its OCI Zettascale10 targeting 800,000 GPU platforms with up to 16 zettaFLOPS (H2 2026), with existing Superclusters supporting up to 16,384 AMD MI300X GPUs in ultra-low-latency RDMA networks. It also partnered with AMD on a 50,000-GPU MI300X supercluster for Q3 2026. OCI has also adopted NVIDIA Spectrum-X Ethernet (achieving a reported 95% data throughput) alongside Quantum-2 InfiniBand for a hybrid networking strategy. Meanwhile, its Acceleron networking suite provides dedicated fabrics, converged NICs, and zero-trust routing with line-rate encryption.

APPENDIX | D. Networking System Vendors

Companies building switch platforms, network operating systems, or fabric solutions for AI data centers.

Aria Networks

Founded in 2024 by Mansour Karam (founder of Apstra, ex-Arista) and Subhachandra Chandra (ex-Arista), Aria Networks raised $125 million in its first series funding round from Sutter Hill Ventures, Atreides Management, Valor Equity Partners, and Eclipse Ventures. Aria's Deep Networking solution combines hardened SONiC, end-to-end telemetry, intelligent multi-layer agents, and cloud-delivered updates to maximize token efficiency at neoclouds and hyperscalers building GPU clusters. Initial customer shipments began October 2025 with full GA in April 2026 and current products include Aria Switch 1600G with 64 ports of 1.6 Tbps (102.4 Tbps aggregate) on Broadcom Tomahawk 6.[127]

Arista Networks

Arista is the leading independent data center switch vendor, with AI networking revenue projected to double in 2026. The 7800R4 AI Spine (576 ports of 800 GbE on Jericho3-AI) and HyperPort 3.2 Tbps line card (44% faster JCT, Q1 2026 availability) anchor its switching portfolio. Arista's most significant 2026 move was organizing the XPO MSA at OFC 2026 — detailed in the XPO section — representing a strategic bet that liquid-cooled pluggable density can compete with CPO.[128]

Arista is advancing the OCP Open Rack v3 wide-rack proposal for 120kW liquid-cooled AI clusters. Co-founder Andreas Bechtolsheim continues championing LPO. The Etherlink platform spans leaf-spine and horizontal-scale architectures with CloudVision and EOS.[129]

[127] Aria Networks, "Building for Scale: Expanding the Aria Portfolio with Industry Leading 1.6Tbps Performance," *Aria Networks*, 12 Feb. 2026. [Online]. Available: https://arianetworks.com/blog/th6-announcement

[128] Arista Networks, "Arista Announces XPO High Density Liquid Cooled Pluggable Optics," *Arista*, 12 Mar. 2026. [Online]. Available: https://arista.com/en/company/news/press-release/23697-pr-20260311

[129] Arista Networks, "Arista Networks, Inc. Reports Fourth Quarter and Year End 2025 Financial Results," *Arista Investor Relations*, 12 Feb. 2026. [Online]. Available: https://investors.arista.com/

Arrcus

Arrcus provides the ArcOS network operating system and ACE-AI Ethernet fabric, with particular strength in scale-across where it differentiates through sub-second convergence and CDN-like GPU resource routing. The company supports NVIDIA Spectrum, Broadcom, AMD Pensando NICs, and BlueField accelerators, with comprehensive BlueField-4 support (6x compute, 800 Gbps) announced October 2025. With $150M+ in funding (including $30M from NVIDIA) and customer wins including Actapio, Arrcus reported 3x bookings growth in 2025. Its most significant 2026 development is the Arrcus Inference Network Fabric (AINF), integrated with NVIDIA's Dynamo inference framework and BlueField DPUs — a purpose-built, AI-policy-aware fabric for inference workloads that distinguishes Arrcus from competitors primarily targeting training fabrics.[130]

Cisco

Cisco's AI data center portfolio is built on custom Silicon One ASICs, Nexus switch platforms, and Intelligent Collective Networking capabilities. The G300 (102.4 Tbps, detailed in the Switch Silicon section) is shipping commercially in H2 2026, with Nexus 9000 and Cisco 8000 systems delivering 70% energy efficiency improvement and 100% liquid cooling. AI infrastructure revenue has reached scale: Q1 FY2026 hyperscaler orders hit $1.3 billion, with projected $3 billion in AI networking revenue for FY2026. Cisco has committed to LPO for Silicon One (800G LPO modules with 50% power reduction), and Meta's FBOSS/SAI qualification of G200 validates Cisco as a merchant silicon alternative. Cisco is also an ESUN founding member and partnered with Intel for Gaudi 3 Ethernet-native clusters.[131]

DriveNets

DriveNets crossed the $1 billion booking milestone in 2025 with commitments extending through 2029. Built on Broadcom Jericho/Ramon (fabric scheduling) and TH6 (endpoint scheduling) and with both air-cooled and liquid-cooled variants (e.g., Tomahawk 6-based 2600SL), DriveNets positions itself as the open-standards alternative to InfiniBand-like performance. Its fabric-scheduled Ethernet approach — detailed in the Scale-Out Topologies section — delivers deterministic, lossless transport with 5–10% bus bandwidth improvement over InfiniBand and 10–30% JCT improvement versus conventional Ethernet. WhiteFiber operates the solution in production for backend and storage networking.[132]

DriveNets' roadmap extends across all three scaling domains: multi-tenancy and multi-site capabilities spanning up to 100 km, endpoint-scheduling for cost-efficient deployments beyond 32,000 GPUs, and we are anticipating an ESUN-based scale-up solution in 2026.

[130] Arrcus, "Arrcus Inference Network Fabric (AINF) Announces Integration With NVIDIA Dynamo Framework, NVIDIA BlueField DPUs and NVIDIA Spectrum Networking," *Business Wire*, 16 Mar. 2026. [Online]. Available: https://businesswire.com/news/home/20260316991472/en/

[131] Cisco, "Cisco Announces New Silicon One G300, Advanced Systems and Optics to Power and Scale AI Data Centers for the Agentic Era," *Cisco Newsroom*, 10 Feb. 2026. [Online]. Available: https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m02/cisco-announces-new-silicon-one-g300.html

[132] DriveNets, "Multi-Site AI Fabric," *DriveNets Newsroom*, Feb. 2026. [Online]. Available: https://drivenets.com/news-and-press

Juniper Networks / HPE

HPE completed its $14 billion Juniper acquisition on July 2, 2025, doubling its networking business. The QFX5250 (102.4 Tbps on Broadcom Tomahawk 6) targets GPU-to-GPU connectivity with general availability early 2026 alongside HPE liquid cooling. Apstra is being extended as Data Center Director with OpsRamp integration (Q2 2026) for full-stack observability, and Juniper's Mist AI/Marvis platform automates root-cause analysis across the combined networking portfolio including wired switches.[133]

Nexthop AI

Founded by Anshul Sadana (former COO of Arista Networks), Nexthop AI builds networking infrastructure purpose-built for hyperscalers and neoclouds.[134] The company launched from stealth in March 2025 and has raised $610 million total (Lightspeed Venture Partners, Andreessen Horowitz, Altimeter, and other investors).[135] Nexthop's product portfolio, unveiled at OFC 2026, spans three switches: NH-4010 (Broadcom Tomahawk 5), NH-4220 (102.4 Tbps on Broadcom Tomahawk 6), and NH-5010 (deep-buffer scale-across spine switch). The NH-5010 enables Nexthop's Disaggregated Spine architecture — developed in collaboration with a large hyperscaler — which decomposes traditional monolithic chassis into independent, optimized functional tiers (scale-across leaf and scale-across spine), claiming 30% lower cost and 30% lower power versus legacy chassis systems.

Nokia

Nokia addresses AI data center infrastructure across IP Networks (switching) and Optical Networks. The 7220 IXR-H6 (102.4 Tbps, liquid- and air-cooled variants) entered production Q1 2026 with UEC/UET support. Nokia offers a dual-OS strategy — SR Linux with EDA (agentic AI-powered AIOps) alongside SONiC — and has secured partnerships with Microsoft Azure, Nscale, Supermicro, and Lenovo. NVIDIA's $1 billion equity investment (2.9% stake, October 2025) signals the deepening partnership.

Nokia's Optical Networks unit claims a 2 market share position in data center and AI/cloud providers for DCI and scale-across applications, leveraging two decades of InP photonics leadership. For scale-across and DCI applications, Nokia is leveraging its ICE-X 800G ZR/ZR+ coherent pluggable (3 nm CMOS, >1,700 km reach), Hyperscale OLS and multi-rail ILA. For intra-DC connectivity, the near-term focus is LPO (ICE-D InP PIC targeting 75% power reduction for intra-DC), with parallel NPO and CPO investments. With on-shore InP fabs and CHIPS Act-backed expansion, Nokia targets both scale-out and scale-up interconnects.

Nokia's OFC 2026 optical suite — including the multi-rail amplifier supporting 160 fiber pairs per rack — is detailed in the Scale-Across section. Coherent pluggables begin sampling mid-2027; multi-rail ships H2 2026.[136]

[133] Hewlett Packard Enterprise, "Hewlett Packard Enterprise Closes Acquisition of Juniper Networks," *HPE Newsroom*, 2 Jul. 2025. [Online]. Available: https://hpe.com/us/en/newsroom/press-release/2025/07/hewlett-packard-enterprise-closes-acquisition-of-juniper-networks.html

[134] Nexthop AI, "Nexthop AI Unveils Transformative, Industry-Leading Scale-out and Scale-across Switches Engineered for Hyperscalers and NeoClouds," *Business Wire*, 10 Mar. 2026. [Online]. Available: https://businesswire.com/news/home/20260310990709/en/

[135] "Nexthop AI accelerates into Hypergrowth with Oversubscribed $500M Series B Funding, catapulting the company's valuation to $4.2 Billion – Nexthop.ai," Nexthop.ai, 2024. https://nexthop.ai/news-and-event/nexthop-ai-accelerates-into-hypergrowth-with-oversubscribed-500m-series-b-funding-catapulting-the-companys-valuation-to-4-2-billion/

[136] Nokia, "Nokia launches suite of application-optimized optical solutions for AI-era networks," *Nokia Newsroom*, 16 Mar. 2026. [Online]. Available: https://nokia.com/newsroom/nokia-launches-suite-of-applicationoptimized-optical-solutions-for-ai-era-networks/

NVIDIA

Details for NVIDIA as a networking vendor appear under their entry in the System Vendors category (given its strength as a vertically integrated provider of accelerated computing stacks). Notably, NVIDIA is one of the fastest growing data center networking vendors. Its data center networking business reached $31B in the fiscal year ending Jan 2026, a 142% YoY increase (NVIDIA earnings report for 2026).

Upscale AI

Founded in 2024, Upscale AI raised $300M+ rapidly ($100M seed, $200M Series A) to build open-standard AI networking infrastructure[137]. Co-founders Barun Kar and Rajiv Khemani bring backgrounds from Innovium (acquired by Marvell) and Cavium.

Upscale AI's SkyHammer is a clean-slate ASIC that supports all three emerging scale-up standards: UALink, ESUN, and UEC. SkyHammer features deterministic flow control, real-time telemetry, and collective communication acceleration with SONiC NOS and SAI integration.[138] Expected to ship late 2026, it positions as one of the first dedicated scale-up switches supporting UALink at rack scale, directly enabling AMD's Helios architecture. At GTC 2026, Upscale AI expanded scope by announcing scale-out Ethernet systems on NVIDIA Spectrum-X silicon, making it a two-product company covering both tiers of the AI fabric hierarchy.[139]

APPENDIX | E. Optical and Photonics Vendors

Companies focused on optical interconnect, coherent optics, and photonic fabric technologies.

Ayar Labs

Ayar Labs develops co-packaged silicon photonic interconnects ($870M+ total funding, $3.75B valuation as of March 2026, backed by AMD Ventures, Intel Capital, and NVIDIA).[140] Its TeraPHY optical I/O chiplet delivers up to 2 Tbps at under 5 pJ/bit, co-packaging with AI accelerators via UCIe. At TSMC OIP 2025, Ayar Labs demonstrated the first TSMC COUPE-based CPO solution targeting 100+ Tbps of optical scale-up bandwidth per accelerator.[141] At OFC 2026, Ayar Labs and Wiwynn announced a joint reference design for optically connected, rack-scale AI systems scaling to 1,024+ accelerators.[142]

[137] Upscale AI, "From $100M Seed to Unicorn in Months: Upscale AI Closes Oversubscribed $200M Series A," *Upscale AI*, 21 Jan. 2026. [Online]. Available: https://upscaleai.com/from-100m-seed-to-unicorn-in-months-upscale-ai-closes-oversubscribed-200m-series-a/

[138] "Upscale AI Eyes Late 2026 for Scale-Up UALink Switch," *HPC Wire*, 2 Dec. 2025. [Online]. Available: https://hpcwire.com/2025/12/02/upscale-ai-eyes-late-2026-for-scale-up-ualink-switch/

[139] Upscale AI, "Upscale AI Supercharges Open, Heterogeneous Scale-Out AI Clusters with NVIDIA Ethernet Switch Silicon," *PR Newswire*, 11 Mar. 2026. [Online]. Available: https://prnewswire.com/news-releases/upscale-ai-supercharges-open-heterogeneous-scale-out-ai-clusters-with-nvidia-ethernet-switch-silicon-302710175.html

[140] "Co-packaged optics startup Ayar Labs raises $500M round backed by Nvidia, AMD," *SiliconANGLE*, 3 Mar. 2026. [Online]. Available: https://siliconangle.com/2026/03/03/co-packaged-optics-startup-ayar-labs-raises-500m-round-backed-nvidia-amd/

[141] "Alchip and Ayar Labs Unveil Co-Packaged Optics for AI Datacenter Scale-Up," *Ayar Labs*, Sep. 2025. [Online]. Available: https://ayarlabs.com/news/alchip-and-ayar-labs-unveil-co-packaged-optics-for-ai-datacenter-scale-up/

[142] "Ayar Labs and Wiwynn Partner to Bring Co-Packaged Optics to Rack-Scale AI Systems," *Ayar Labs*, Mar. 2026. [Online]. Available: https://ayarlabs.com/news/ayar-labs-and-wiwynn-partner-to-bring-co-packaged-optics-to-rack-scale-ai-systems

Ciena

The $270M Nubis acquisition (Q4 FY2025) extended Ciena from inter-DC coherent optics into CPO/NPO (6.4 Tbps full-duplex). The WaveLogic 6 family spans long-haul (WL6e, 1.6 Tbps), and DCI (WL6n, 800G) for 800ZR/ZR+ and Coherent-Lite pluggables. At OFC 2026, Ciena announced their hyper-rail photonics solution (128 fiber pairs per rack, 75% power reduction, 85% space reduction) which competes directly with Nokia's multi-rail solutions for multi-campus/multi-span deployments. Ciena also demonstrated WL6e 1.6T with quantum-safe encryption.[143] As part of Ciena's ecosystem participation, it joined both the Open CPX (co-chair) and XPO MSA (one of 4 co-chairs) as a founding member.

Coherent Corp

Vertically integrated photonics company (formerly named II-VI) with strong 800ZR/ZR+ DCI traction (Meta as earliest and largest customer). Ramping 1.6T-DR8 with Marvell Ara DSP; expanding Sherman, Texas InP facility with $33M CHIPS Act funding. At OFC 2026, Coherent demonstrated 400G/lane PAM4 links and multi-technology CPO (silicon photonics, VCSEL, InP-on-silicon in a single package) — as detailed in the CPO section. Joined both the XPO and Open CPX MSAs as a founding member.[144]

Eridu AI

Co-founded by Drew Perkins (serial networking entrepreneur, co-founded Lightera Networks sold to Ciena for $500M+, co-founded Infinera sold to Nokia for $2.3B in 2025, and co-founded Gainspeed also acquired by Nokia) and Omar Hassen (networking chip design veteran with roots at Broadcom and Marvell), Eridu emerged from stealth in March 2026 with $230 million in funding. The company targets what it calls the "network wall" — the growing mismatch between AI compute scaling and data center networking capabilities. By consolidating switching, optics, and network intelligence into fewer, more capable chips, Eridu aims to reduce latency, lower power consumption, and improve network reliability. Product details and partnership announcements are expected later in 2026.[145]

Lightmatter

Well-funded innovator in the optical domain with a $400M Series D (approx $850M raised to date) at a $4.4B valuation. Lightmatter Passage L200[146] (32 Tbps) and L200X (64 Tbps) — described as the first 3D CPO products — target 5–10x improvement over existing CPO solutions.[147] At OFC 2026, Lightmatter demonstrated 1.6 Tbps per fiber using 16-wavelength DWDM.[148] Manufacturing partnerships with GlobalFoundries and Amkor provide the production pathway. XPO MSA founding member.

[143] Ciena, "Ciena Brings AI Networking Expertise to OFC 2026," *Ciena Newsroom*, Mar. 2026. [Online]. Available: https://ciena.com/about/newsroom/press-releases/ciena-brings-ai-networking-expertise-to-ofc-2026
[144] Coherent, "Coherent to Unveil Breakthrough AI-Scale Optical Innovations and Industry Leadership at OFC 2026," *GlobeNewsWire*, 17 Mar. 2026. [Online]. Available: https://globenewswire.com/news-release/2026/03/17/3257303/11543/en/
[145] Eridu, "Eridu Emerges from Stealth with Over $200M in Funding To Break Through the Network Wall and Unlock Faster AI," *Business Wire*, 10 Mar. 2026. [Online]. Available: https://businesswire.com/news/home/20260310994409/en/
[146] Lightmatter, "Lightmatter Announces Passage L200, the Fastest Co-Packaged Optics for AI," *Lightmatter*, Mar. 2025. [Online]. Available: https://lightmatter.co/press-release/lightmatter-announces-passage-l200-the-fastest-co-packaged-optics-for-ai/
[147] Lightmatter, "Lightmatter Raises $400M Series D; Quadruples Valuation to $4.4B as Photonics Leader for Next-Gen AI Data Centers," *Lightmatter*, 16 Oct. 2024. [Online]. Available: https://lightmatter.co/press-release/lightmatter-raises-400m-series-d/

Lumentum

Key supplier of EML and LIPD laser technology, ramping 1.6T DR8 modules and pioneering OCS through the R300 platform (5–10x latency reduction, up to 65% power savings in 100K-GPU deployments). Collaborating with NVIDIA on Spectrum-X photonics and sampling R300 OCS with multiple hyperscalers. At OFC 2026, demonstrated a 1.6T DR4 prototype (stepping stone to 3.2T) and the Marvell/Lumentum OCS integration detailed in the Scale-Out section.[149]

Nokia

Nokia (particularly with the Infinera acquisition) is a leading supplier of optics solutions. For details on all their offerings (including optical components), see their entry under Networking System Vendors above.

APPENDIX | F. Software and Emerging Vendors

Network software, automation, resilience, and emerging technology companies.

Aviz Networks

Aviz provides the ONES Certified Community SONiC distribution with the comprehensive switching ASIC coverage (NVIDIA, Cisco, Marvell, Broadcom) and 500+ automated tests in their Fabric Test Automation Suite. Network Copilot offers LLM-native AI operations for multi-vendor fabrics, and NVIDIA Spectrum-X support (March 2025) extends orchestration and telemetry to Ethernet-based AI cluster networking.[150]

BE Networks

Texas-based intent-based networking company whose Verity platform provides Day 0–Day N network management with SONiC support across Dell, Edgecore, Celestica, and NVIDIA switches. SensAI (December 2024) adds LLM-powered natural language provisioning; Verity 6.3 introduced NVMe/TCP storage network support and cut-through switching for AI fabric latency optimization.

[148] Lightmatter, "Lightmatter Achieves Record 1.6 Tbps Per Fiber," 11 Mar. 2026. [Online]. Available: https://lightmatter.co/press-release/lightmatter-achieves-record-1-6-tbps-per-fiber-to-accelerate-ai-optical-interconnect/

[149] Marvell, "Marvell and Lumentum to Demonstrate Optical Circuit Switching for Next-generation AI Scale-up Infrastructure," *Marvell Newsroom*, 16 Mar. 2026. [Online]. Available: https://investor.marvell.com/news-events/press-releases/detail/1015/marvell-and-lumentum-to-demonstrate-optical-circuit-switching-for-next-generation-ai-scale-up-infrastructure

[150] Aviz Networks, "Aviz to Accelerate AI Networking with ONES and NVIDIA Spectrum-X," *Business Wire*, 4 Mar. 2025. [Online]. Available: https://businesswire.com/news/home/20250304144656/en/

Clockwork

Founded by Stanford faculty and researchers ($40.5M total funding), Clockwork provides the FleetIQ Software-Driven Fabric (SDF) delivering sub-microsecond clock synchronization, cross-stack AI observability, dynamic traffic control, and TorchPass Workload Fault Tolerance across NVIDIA, AMD, and custom accelerators on Ethernet, InfiniBand, and RoCE. Production deployments at Uber, DCAI Denmark's Gefion supercomputer, and European neoclouds (Nebius, NScale, WhiteFiber) validate the platform's ability to reduce network issue detection from hours to minutes.[151]

Hedgehog

Founded by former Cisco executives, Hedgehog's Open Network Fabric is a production-ready SONiC distribution that treats switches and smartNICs as a Kubernetes cluster, providing VPC-style multi-tenancy and declarative intent-based networking through standard Kubernetes APIs. At GTC 2026, Hedgehog announced support for NVIDIA Spectrum-X Ethernet and alignment with the NVIDIA Cloud Partner reference architecture, with availability in Q2 2026.[152] Active deployments span financial services, healthcare, biotech, telecom, energy, manufacturing, and logistics at a fraction of legacy solution costs.[153]

[151] Clockwork, "Clockwork Launches FleetIQ, the Software Layer That Recasts GPU Economics," *SiliconANGLE*, 10 Sept. 2025. [Online]. Available: https://siliconangle.com/2025/09/10/clockwork-raises-20-5m-synchronize-gpu-clusters-accelerate-ai-workloads/

[152] Hedgehog, "Hedgehog Announces Support for NVIDIA Spectrum-X Ethernet and NVIDIA Cloud Partner Reference Architecture at GTC 2026," *PR Newswire*, 9 Apr. 2026. [Online]. Available: https://www.prnewswire.com/news-releases/hedgehog-announces-support-for-nvidia-spectrum-x-ethernet-and-nvidia-cloud-partner-reference-architecture-at-gtc-2026-302737673.html

[153] Hedgehog, "Open Network Fabric," *Hedgehog*, 2025. [Online]. Available: https://hedgehog.cloud/

Glossary

AEC (Active Electrical Cable)	A copper cable with built-in signal conditioning electronics, extending reach beyond passive direct attached copper cables.
ASIC (Application-Specific Integrated Circuit)	A custom chip designed for a particular function such as switching or AI acceleration.
Bisection Bandwidth	The aggregate bandwidth available across the narrowest cut that divides a network into halves; a measure of worst-case throughput.
Clos / Fat-Tree	A multi-stage, non-blocking switch topology used in data centers.
CPO (Co-Packaged Optics)	Optical engines integrated directly onto the switch ASIC package, reducing power and latency.
CXL (Compute Express Link)	An open interconnect standard for cache-coherent memory pooling and sharing across CPUs and accelerators.
DAC (Direct Attach Copper)	A passive copper cable assembly for short-reach, low-power connections between adjacent devices.
DCI (Data Center Interconnect)	High-capacity optical links connecting separate data center buildings or campuses.
DCQCN (Data Center Quantized Congestion Notification)	A congestion-control protocol used with RoCEv2 Ethernet fabrics.
DPU (Data Processing Unit)	A programmable network processor that offloads infrastructure tasks (security, storage, telemetry) from the host CPU.
DSP (Digital Signal Processor)	In optics, the component that performs high-speed analog-digital (and vice-versa) conversion, digitally compensates for fiber impairments, handles advanced modulation and polarization multiplexing, applies forward error correction, and provides real-time link telemetry.
DWDM (Dense Wavelength Division Multiplexing)	Technology that transmits multiple optical signals on different wavelengths over a single fiber.
ECN (Explicit Congestion Notification)	A mechanism where switches mark packets to signal congestion, allowing endpoints to throttle before drops occur.
ECMP (Equal-Cost Multi-Path)	A routing strategy that distributes traffic across multiple paths of equal cost for load balancing.
EML (Electro-Absorption Modulated Laser)	A laser type used in high-speed optical transceivers, usually at 400G/lane and above.
ESUN (Ethernet for Scale-Up Networking)	An OCP forum defining Ethernet extensions for high-performance scale-up fabrics.

FBOSS (Facebook Open Switching System)	Meta's internally developed network operating system for its data center switches.
FSDP (Fully Sharded Data Parallelism)	A PyTorch training strategy that shards model parameters, gradients, and optimizer states across GPUs to reduce memory use.
InfiniBand	A high-bandwidth, low-latency interconnect technology historically dominant in HPC and AI training clusters.
KV Cache (Key-Value Cache)	Cached attention states from prior tokens in LLM inference. This is what is transferred between prefill and decode stages in disaggregated architectures.
LPO (Linear-Drive Pluggable Optics)	Pluggable transceivers that replace the DSP retimer with analog equalization, reducing power consumption.
LPU (Language Processing Unit)	Groq's (core team acquired by NVIDIA, which licensed the IP) custom silicon optimized for the decode (token-generation) phase of LLM inference.
MFU (Model FLOPs Utilization)	The fraction of a system's theoretical peak compute actually used by a training workload; a key efficiency metric.
MoE (Mixture of Experts)	An AI model architecture where only a subset of "expert" sub-networks is activated per input.
NCCL (NVIDIA Collective Communications Library)	NVIDIA's standard library for GPU-to-GPU collective operations in distributed training.
NIC (Network Interface Card)	Hardware that connects a server or accelerator to the data center network.
NOS (Network Operating System)	The software that runs on a network switch, managing routing, telemetry, and configuration.
NPO (Near-Packaged Optics)	Optical engines placed adjacent to (but not on) the switch ASIC, balancing CPO's power advantage with improved serviceability.
Open CPX MSA (Open Co-Packaging Multi-Source Agreement)	A March 2026 MSA led by Ciena, Coherent, Marvell, Molex, Samtec, and TeraHop that defines a pluggable socket and electrical connector interface for co-packaged and near-package optical engines, enabling field replacement of optical engines and decoupling engine supplier from switch ASIC vendor.
NVLink	NVIDIA's proprietary high-bandwidth, low-latency interconnect for GPU-to-GPU communication within a node or rack.
OCS (Optical Circuit Switching)	Reconfigurable optical switches that establish dedicated light paths between endpoints without electronic conversion.
PAM4 (4-level Pulse Amplitude Modulation)	A signaling scheme that encodes two bits per symbol, doubling data rate per lane versus NRZ.
PFC (Priority Flow Control)	An Ethernet mechanism that pauses specific traffic classes to prevent buffer overflow, essential for lossless RDMA.

RCCL (ROCm Collective Communications Library)	AMD's communications library equivalent of NCCL for AMD GPUs.
RDMA (Remote Direct Memory Access)	A technology allowing one machine to read/write another's memory without involving the remote CPU, reducing latency.
RoCEv2 (RDMA over Converged Ethernet v2)	A protocol encapsulating InfiniBand transport over standard Ethernet, enabling RDMA on commodity networks.
SAI (Switch Abstraction Interface)	A standardized API that decouples the NOS from the underlying switch ASIC, enabling multi-vendor interoperability.
SerDes (Serializer/Deserializer)	High-speed I/O circuitry on a chip that converts parallel data to serial for transmission and vice versa.
SHARP (Scalable Hierarchical Aggregation and Reduction Protocol)	NVIDIA's in-network computing technology that offloads collective operations to InfiniBand switches.
SONiC (Software for Open Networking in the Cloud)	An open-source NOS originally developed by Microsoft, now the leading open NOS for AI data centers.
SUE-T (Scale-Up Ethernet Transport)	A specification defining low-overhead framing, link-level retry, and flow control for Ethernet-based scale-up fabrics.
UALink (Ultra Accelerator Link)	An open consortium standard for memory-semantic, accelerator-to-accelerator scale-up interconnects.
UCIe (Universal Chiplet Interconnect Express)	An open standard for die-to-die interconnect within multi-chiplet packages.
UEC/UET (Ultra Ethernet Consortium / Ultra Ethernet Transport)	An industry consortium and its transport protocol, re-engineering Ethernet for AI workloads with native multipathing, out-of-order delivery, and rapid loss recovery.
VOQ (Virtual Output Queueing)	A switch buffering technique that maintains separate queues per output port, eliminating head-of-line blocking.
XPO (eXtra-dense Pluggable Optics)	A new MSA, initiated by Arista Networks, defining 12.8 Tbps liquid-cooled pluggable modules for high-density AI fabrics.
XPU	A generic term for any processing accelerator (GPU, TPU, LPU, etc.) used in AI workloads.
ZR/ZR+	Coherent pluggable optics standards for data center interconnect, supporting 80 km (ZR) to 500 km (ZR+) reach.

AvidThink, LLC
1900 Camden Ave
San Jose, California 95124 USA
avidthink.com

Rev B3