FUTURIOM
FUTURE OF CLOUD TECH

Cloud Market Trend Report

May 2025

What's Next for Networking Infrastructure for AI

Sponsored by:

hedgehog

Highlights of What's Next for Networking Infra for AI

- **AI infrastructure needs continue to expand.** Enterprises are starting to adapt large language models (LLMs) to fit their specific business requirements. A mixture of infrastructure will be needed to deliver LLMs for training, as well as specialized models including small language models (SLMs).
- **Inferencing needs will expand infrastructure in a variety of ways.** AI inferencing, which enables applications to take input and process output through AI models, will be distributed across cloud and enterprise infrastructure. As models evolve, more inferencing infrastructure will be needed to interpret and serve up results on a variety of devices and infrastructure from the data center to the edge.
- **Ethernet is gaining ground against InfiniBand.** Efforts to shift AI networks from reliance on NVIDIA's proprietary InfiniBand networking technology is showing dramatic results, with adoption of Ethernet solutions being heralded even by NVIDIA as key to inferencing.
- **The Ultra Ethernet Consortium remains relevant.** Efforts by vendors, including NVIDIA, are coalescing around a standard that improves on Ethernet's drawbacks and RDMA's limitations.
- **Speeds are increasing.** While most AI datacenter switches support speeds of 400-Gb/s, 800-Gb/s rates are increasingly on the horizon, with even higher speeds in the works.
- **Optical networking is part of AI's future.** As AI networks grow in scale, speed, and power requirements, optical components will furnish solutions that save power, space, and operational costs.
- **AI specialized processors such as SmartNICs, IPUs, and DPUs are growing in importance for AI infrastructure.** Network interface cards powered by specialized chips are key to enabling better performance of networking, security, and storage functions of AI networks. But switch vendors are finding ways to be "NIC agnostic" to avoid vendor lock-in.
- **SASE, SD-WAN, and network-as-a-service will evolve to support AI networking with more pervasive security.** AI will increase data traffic by orders of magnitude, but a distributed networking infrastructure is needed to connect these resources and provide enterprise controls and compliance.
- **Observability and AIOps are central to AI networking ROI.** The ability to track, analyze, and automate networking efficiency is becoming vital to enterprise adoption of AI networking.
- **Companies included in this report:** Akamai, AMD, Arista, Arrcus, Aryaka, Astera Labs, Aviz Networks AWS, Broadcom, Ciena, Cisco, Cloudflare, CoreWeave, DriveNets, Enfabrica, Equinix, Google Cloud, Hedgehog, Infinera, Juniper Networks, Lambda Labs, Marvell, Meta, Microsoft, Napatech, Netris, NVIDIA, Oracle, Vapor IO, Versa Networks, xAI, ZEDEDA

hedgehog

AI needs a new network

Hedgehog is the AI Network

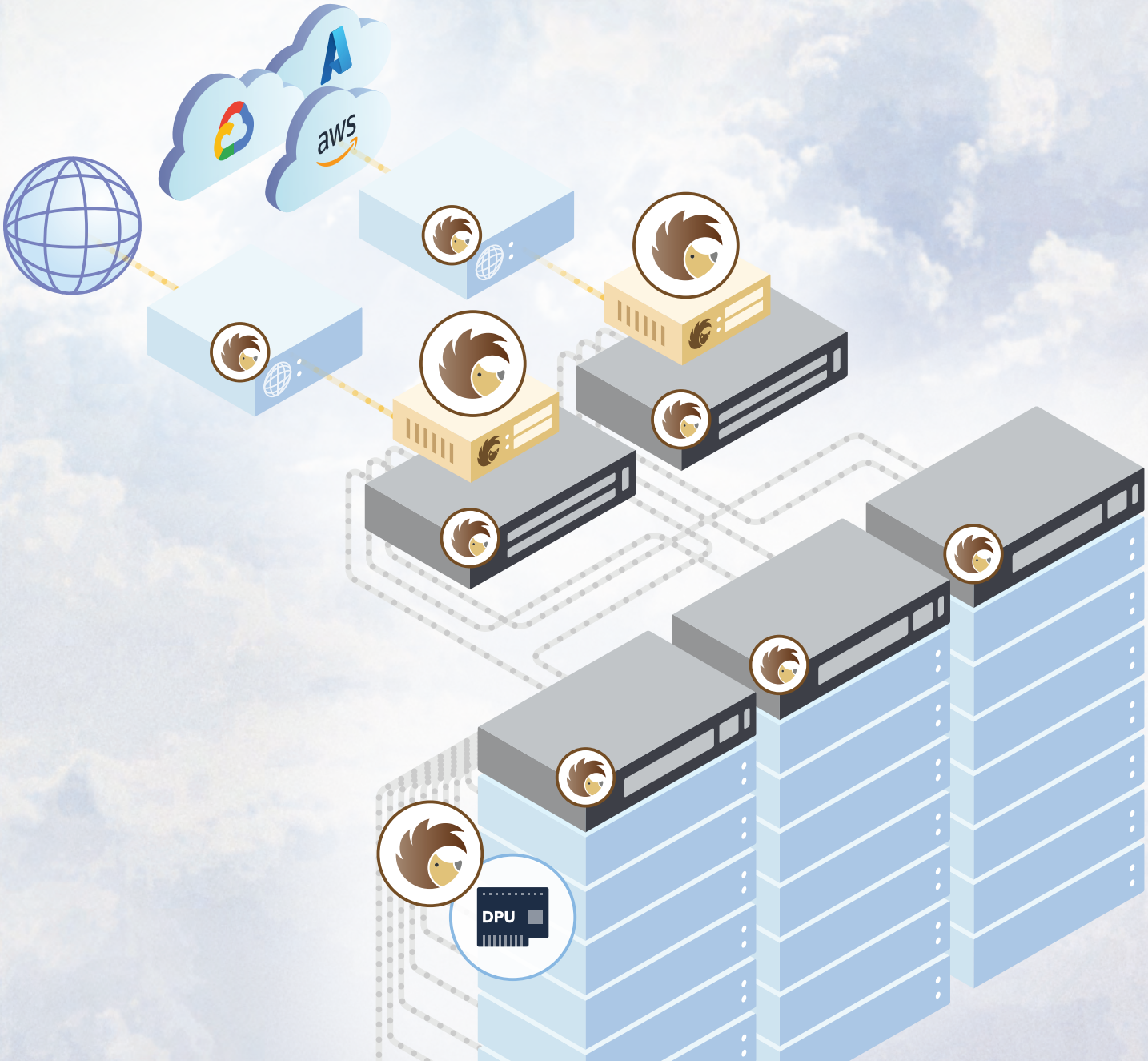


Table of Contents

1. Introduction: AI Boosts the Networking Market

Artificial intelligence (AI) heads the list of corporate goals. Firms everywhere are kicking the tires of AI, engaging in proof-of-concepts, and strategizing to engage the necessary talent to bring generative AI to bear on a range of critical applications. At the same time, it's clear that to reap AI's rewards requires a full-on reworking of the IT infrastructure, with particular emphasis on networking.

In many ways, networking determines AI success. AI requires new levels of performance in bandwidth, latency, and security. Only the most efficient and top-performing networks can deliver the data needs for inferencing, or the adaptation of trained large language models (LLMs) such as ChatGPT, Llama, Claude, and Grok to fit enterprise applications.

More importantly, AI is changing entire network architectures. Enterprises now need to think about what networks are needed to support AI whether that's in the core or at the edge. They also need to think about what impact AI applications will have on corporate networks, datacenters, and governance strategies. If you are putting your most valuable and proprietary data into AI models, you certainly need to think about how they are connected by networks—to partners, your internal resources, and the Internet.

So far, the shifts in networking for AI have fallen into two general categories: 1) Networking infrastructure for AI, or the connectivity infrastructure needed for AI datacenters and inference; and 2) AI for networking—the AI automation to drive operations, often also referred to as AIOps. In this report, we are focused on number 1, the infrastructure needed to support AI datacenter networks.

AI Democratization: From Foundational LLMs to the Edge

This report covers how AI is impacting networking infrastructure from the core the edge, whether it's a datacenter in Seattle or an inference network in Dubai. So far, much of the attention has been placed on the hyperscalers as they build massive datacenters to support clusters of hundreds of thousands of GPUs for training LLMs. But that's just part of the picture. These foundational models, as they are called, are being open-sourced and propagated into many different forms, including SLMs and customized models adapted with Retrieval Augmented Generation (known as RAG, which is the topic of another report we'll publish later this year). AI applications can range from a massive 100,000 GPU cluster for training LLMs to 4-GPU clusters that run a computer-vision application at the edge.

We recently met with a technology executive who asked to remain anonymous who told us that an AI-driven inspection application on the manufacturing floor was delivered with a grand total of four GPUs. The application was custom built. If the arrival of DeepSeek taught us anything, it's that the long trajectory of AI will follow the lead of previous technology markets (PCs, client/server, Internet) toward commoditization and democratization. As prices drop and technology matures, it will be more widely adopted.

This theme will drive the needs for inferencing networks for AI from the core LLMs to the edge. Building models is not enough. You also need infrastructure both inside and outside the data center to interpret and deliver results. AI inferencing calls for revamped, accelerated networks supporting speed, accuracy, and scale throughout the corporate networking infrastructure.

It's inferencing that will drive these networking changes required by so many enterprises. "2025 is the year of inferencing," said Kevin Deierling, NVIDIA's SVP Networking, in an analyst presentation earlier this year. And his prediction is holding true. Some of the largest datacenters in the world are being built not only for AI training of LLMs but to support inferencing workloads.

CoreWeave, for instance, a leader in the GPU-as-a-service market (which also recently went public), offers inference and optimization services as an essential part of the software offerings based on its 250,000-GPU network. And CoreWeave rival **Lambda,** which recently scored $480 million to grow its AI services, is also intent on offering inferencing services to developers. "We're investing in both the underlying infrastructure… but we're also making major investments in terms of the software layer that runs on top of it," said Stephen Balaban, CEO and cofounder of Lambda. "Expect to see more things like the Lambda Inference API that's recently been launched."

As these large datacenters expand to support AI inferencing, they are creating blueprints for other enterprises to follow. From rack to cluster to overall datacenter, innovations are emerging to spur AI adoption. Industry partnerships and acquisitions are creating a range of options for connectivity among clouds and distributed data platforms in AI environments.

Security and Architecture Implications: From Core to the Edge

The expansion of AI needs will have dramatic effects on networking infrastructure across nearly any installation, ranging from datacenters to the enterprise edge as well as IoT. AI is a horizontal application. That means it can be adopted by almost any type of business in just about any industry, from graphic artists to biomedical engineers. As you'll see in this report, that means that it changes the needs of networking in just about any deployment.

This dramatic shift means that network architects and engineers will need to adopt higher-level thinking about how they support networking for AI, from the core layer to the applications layer and beyond. One of the key considerations will be security.

The scale and pace of change brings many challenges and opportunities to the networking market. New capital and changing architectures are already re-sorting decades-old market positions and generating enormous value for leaders such as **Arista Networks** and **NVIDIA**. But as the AI revolution broadens and spreads into enterprise and edge, these opportunities for market-share gains—or risks of market-share losses—will come down to how well networks are adapted for AI.

Another consideration will be implementing the networking architecture that can distribute compute tasks to support AI applications in the the right place. As networks expand to support AI, architects will

have to think holistically about the needs of the distributed infrastructure. Data and applications supported by AI will cross boundaries and mean that compute infrastructure will be tapped for inferencing AI apps anywhere from a security camera to the largest datacenters in the world

The needs of supporting distributed AI are likely to expand for several key reasons. Here are the key benefits for building a distributed networking infrastructure, whether it's a datacenter or edge compute node:

Lower Latency: Efficient, AI-optimized networking will be used to lower the AI inferencing time to process data and deliver decisions, enabling real-time responses for critical applications like threat detection, autonomous systems, or industrial control.

Reduced Bandwidth Consumption: By analyzing and filtering data in the right place, only relevant insights or anomalies need to be sent to the cloud or data center, dramatically lowering the volume of network traffic.

Improved Privacy and Compliance: Organizations will need to keep data secure across the entire infrastructure to comply with data sovereignty, privacy regulations, and internal governance policies.

Operational Resilience: Edge inferencing and optimized networks will ensure that AI-powered functions continue even if connectivity to the cloud is disrupted, providing fault tolerance and local autonomy.

Scalability and Cost Efficiency: Distributing compute tasks across, coordinated by networking, will reduce dependency on centralized infrastructure and avoids the cost and latency of round trips to the cloud for every inference.

So let's dig into more details.

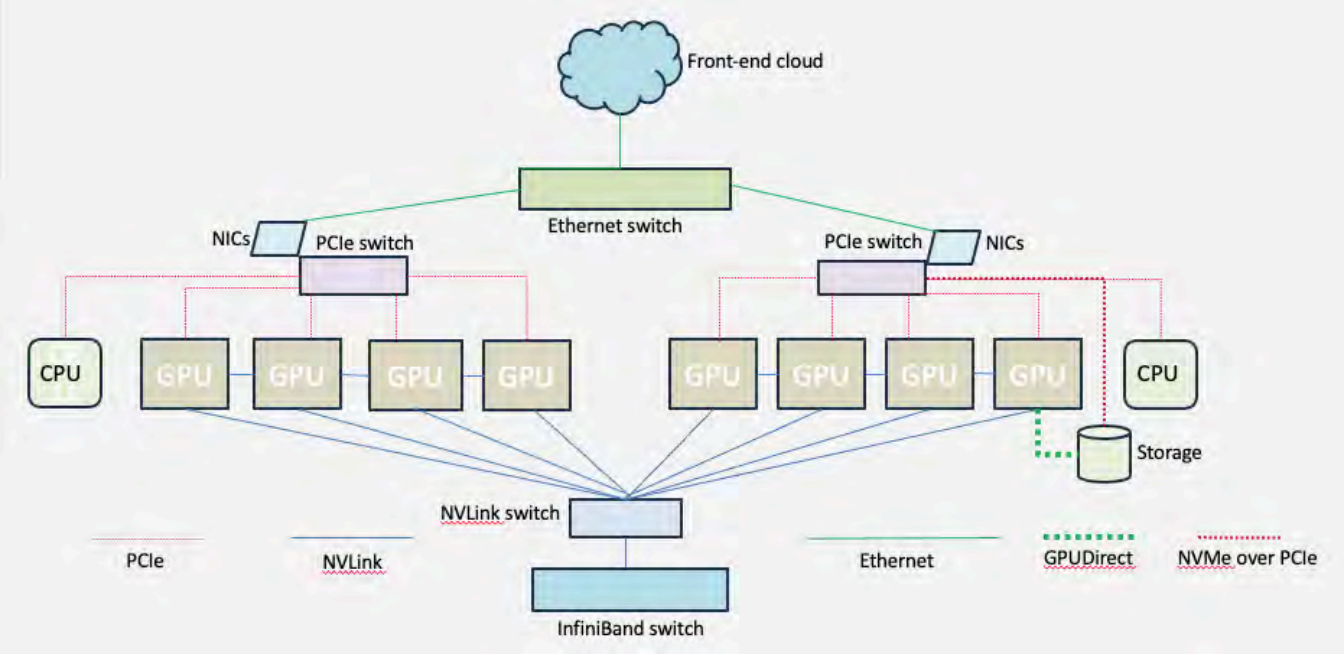
2. AI Networking in the Core: The Arrival of Ethernet

AI starts in the core, with the hyperscalers training foundational LLMs. This process has been enormous and unprecedented, as they have built the largest datacenters in history and packed them with literally as much data as has existed on earth.

AI clusters are costly and complicated to implement, requiring a full-stack reimagining of the datacenter from the ground up. This means revising the networking approach from deep within the rack to the scale-out capabilities between clusters and ultimately between AI datacenters.

Scaling AI Clusters in the Core

Today's networks are largely based on **NVIDIA** GPUs and that vendor's related proprietary connections for "back-end" AI training, including NVLink and InfiniBand. Still, Ethernet is increasingly favored as the fabric connecting front-end clients and clouds. Below is a rudimentary illustration of a typical AI networking environment:



NVIDIA, originally a niche graphical chip provider, was visionary in developing GPU applications other than AI. This is how it discovered networking, as it sought higher performance ways to connect GPUs, resulting in its acquisition of Mellanox for $7 billion in 2019—yet another visionary move. Mellanox was the leader in InfiniBand networking, a high-performance technology targeted at supercomputing. While NVIDIA's InfiniBand and its proprietary NVLink are still solidly entrenched in AI networks, there is a movement away from proprietary technologies, and this is reflected in a range of innovations underway at all levels. Following is a list of ones we've identified:

Support for Ethernet. Ethernet is increasingly popular for inferencing workloads—by NVIDIA's own admission. Further, leading industry vendors such as **Arista, Cisco, Juniper, Nokia,** and **NVIDIA,** have backed the Ultra Ethernet Consortium, leading the march toward improved and standardized Ethernet for networking for AI.

The move to 800-Gb/s. While 400-Gb/s data rates for Ethernet are common in networks for AI, 800-Gb/s is on the horizon. All major switch providers, including **Arista, Cisco, Juniper, Nokia,** and **NVIDIA,** offer an 800-Gb/s option. A further option to 1.6 Tb/s is being discussed by various standards bodies, including the IEEE and the Ethernet Alliance.

Optical innovation. As AI networks grow in size and capacity, they grow in power consumption. To reduce power requirements as well as to sustain faster line rates, AI networking equipment is gradually shifting toward future integration of optical components, including co-packaged optics (COP) and chips that streamline connectivity as networks scale.

Rack-scale density and integration. There's a movement in AI networking toward providing denser numbers of GPUs, compute, and storage elements within a rack. A prime example of this trend is **NVIDIA's** GB200 NVL72, which connects 36 Grace CPUs and 72 Blackwell GPUs in a single rack. (More on that in a bit.)

Movement toward faster, nonproprietary connections in the rack. While NVIDIA's proprietary connectivity solutions predominate at the rack level, where NVLink has been improved to accommodate exponentially more GPU connections (from 8 to 72), momentum toward open standards is increasing. Example: The UALink (Ultra Accelerator Link), recently standardized with a 1.0 specification by a consortium including **AMD, Apple, Broadcom, Cisco, Enfabrica, Google, HPE, Intel, Marvell, Meta, Microsoft,** and **Qualcomm,** among many others, has directly challenged NVLink.

Focus on NICs. The use of network interface cards (NICs) has grown in AI networks because they can offer additional processing power to offload networking functions, including encryption—although they also add cost to the network. **NVIDIA's** BlueField-3 SuperNIC continues to be an essential element of networking for AI, particularly given its capability to work with a range of third-party switches. **Arista** and **DriveNets,** however, offers load balancing for AI networks without NICs.

Observability as key to performance. As networks get larger and more complex, observability tools are on the rise, and their availability for Ethernet networks as opposed to InfiniBand can be a deciding factor for Ethernet. In recent announcements, vendors such as **Arista, Aviz Networks,** and **Juniper Networks** have added AI to improve the performance of network fabrics in AI environments.

The relevance of the network edge. As inferencing spreads, the performance advantages of working at the edge of the network have boosted solutions from content delivery network (CDN) providers and related suppliers. These efforts pair edge networking nodes with GPUs and AI software development to fuel inference and agentic AI applications.

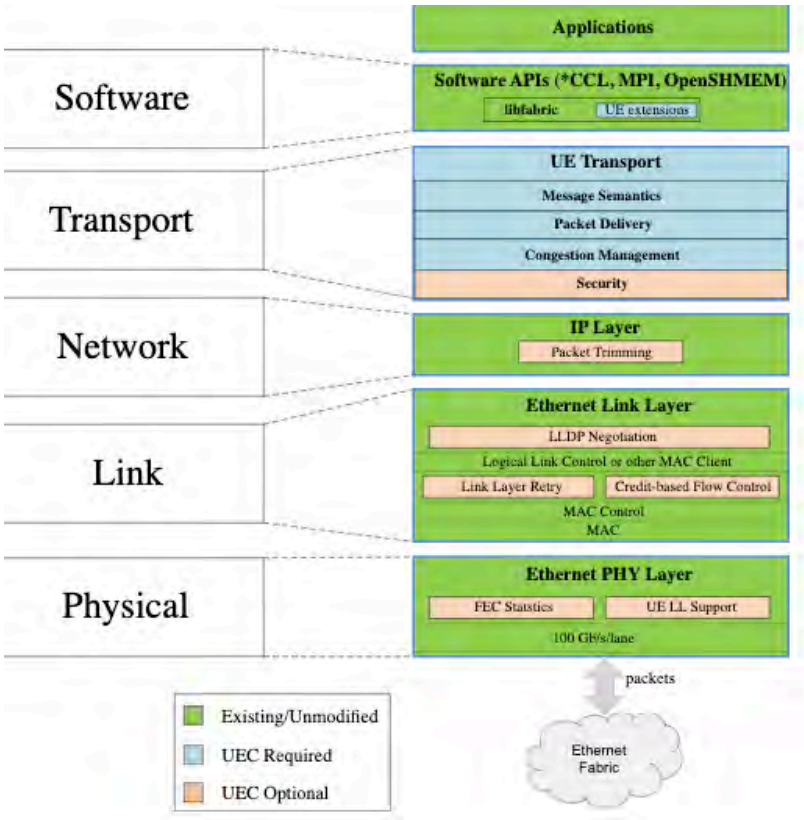
The Arrival of Ethernet and the UEC

One of the major developments in the past 12 months has been the advance of Ethernet to support AI networking from the core to the edge.

Only a year or so ago, there seemed to be some debate about whether standards-based Ethernet could be adopted for the special demands of AI. Those reservations appear to have been dispelled. A growing number of members in the Ultra Ethernet Consortium (UEC) are working toward standardizing improvements to Ethernet for use in AI and HPC (high performance computing) networking.

The UEC includes a series of working groups addressing Ethernet's physical layer, link layer, transport layer, and software layer, along with storage compliance, management, and performance/debug. The group said in January 2025 that it plans a 1.0 specification for imminent release, reflecting work in each of these areas, as illustrated in the diagram from the UEC below:

The UEC Stack *(Source: The Ultra Ethernet Consortium)*



The focus of the UEC's work is to improve Ethernet's performance by addressing the drawbacks that in the past kept it out of high-performance networking. These include packet loss, lack of adequate load balancing, congestion management, and other performance drawbacks.

The goal is to match the job completion time (JCT) of NVIDIA's InfiniBand in HPC and AI environments. Ethernet has already evolved to adapt to these needs, with the leading vendors and the UEC driving

adoption of technologies such as packet spraying and congestion control, including Priority-based Flow Control (PFC) and Explicit Congestion Notification (ECN). All of this has served to upgrade Ethernet for the AI era.

The UEC's work on Ultra Ethernet Transport revises Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE). RDMA was originally developed as part of InfiniBand to allow data transfer from the memory of one GPU to another without having to involve the CPU. Ethernet supporters adapted it to run over IP and Ethernet via RoCE. Today, most networking vendors support RoCEv2, which adds a variety of enhancements. Still, the UEC maintains that RoCEv2 requires even better congestion control. Specifically, the UEC aims to solve incast, or the tendency in TCP networks for packet congestion to happen when many data sources simultaneously transmit to a single source. Since multiple GPUs communicate to servers in parallel, it is vital to solve this potential snarl. All of this work has driven Ethernet to become a foundation for AI networking.

"If the network is the computer and you want to build the best network that is out there, there's nothing like Ethernet," said Ram Velaga, SVP and GM of the Core Switching Group at **Broadcom,** in a panel talk on SiliconANGLE's theCUBE in September 2024. **Broadcom** makes a series of Ethernet NICs and switch chips for leading AI networking used by **Arista, Juniper, Cisco** (which also touts its own Silicon One chips as a Broadcom alternative), and other Ethernet fabric suppliers. It also enables a vibrant market for whitebox AI switches produced by **Accton**, **Dell**, **EdgeCore,** and **Quanta**, among others. This in turn enables innovation in the startup community, whether it's network operating systems and software from **Arrcus**, **DriveNets,** and **Hedgehog,** was well as orchestration solutions from Netris.

Other industry sources agree that Ethernet can compete directly with **NVIDIA's** proprietary InfiniBand fabric, which remains a strong force in AI networking. Even NVIDIA has acknowledged the strength of Ethernet by improving its Spectrum-X series of Ethernet-compatible AI switches. While NVIDIA doesn't publicly release numbers related to Spectrum-X versus InfiniBand sales, it's clear that the vendor is watching and conforming to demand for Ethernet in AI. The vendor acknowledges that Ethernet is becoming the preferred networking fabric for inferencing, and that inferencing is the market to watch this year.

"Most were quite surprised that we came into the Ethernet world," said **NVIDIA** CEO Jensen Huang during his keynote speech at the company's annual GTC conference in March 2025. "And the reason why we decided to do Ethernet is if we could help Ethernet become like InfiniBand, have the qualities of InfiniBand, then the network itself would be a lot easier for everybody to use and manage." He cited ongoing improvements to NVIDIA's Spectrum-X NICs and switches for Ethernet and he noted that clusters comprising **xAI's** Colossus supercomputer are interconnected with Spectrum-X switches. "Spectrum-X is unquestionably a huge home run for us," he said.

Other industry sources agree on the importance of Ethernet in AI networking. "Ethernet—we're just seeing tremendous traction and it's kind of a foregone conclusion I think that Ethernet, from my

standpoint, will be dominant in AI networking going forward," said Hugh Holbrook, VP of Software Engineering at **Arista,** in a presentation late in 2024.

Note: While most Ethernet switches for AI networks today support data rates to 400 Gb/s, with 800 Gb/s increasingly common, a range of component suppliers, switch vendors, and standards bodies are prepping for the next anticipated data rate of 1.6 Tb/s. But the step-up to terabit Ethernet will require innovations in power usage, digital transmission, and other aspects of networking.

To shift to ever-higher link speeds while reducing power draw and environmental impact, it will be necessary to enlist optical technologies. But it's also worth mentioning that great speed requires more power, which in turn generates more heat. It's therefore become important for vendors to support liquid cooling, which uses water or alternative liquids directed to the rack, chip, or server, to cool down machinery and reduce the heat they produce. While specifics are beyond the scope of this report, liquid cooling is gaining ground and should be considered a key element of AI networking.

Incumbent Networking Vendors Supporting the Ethernet Trend

Leading Ethernet switch vendors have revamped their products to accommodate AI workloads, with most AI switches supporting 400-Gbit/s, though 800-Gbit/s is emerging as a back-end cluster option in 2025. Following is a short summary of trends and announcements:

Arista. The vendor's Etherlink AI platforms (including the 7060X6 AI leaf switch, the 7800R4 AI Spine switch, and the 7700R4 AI Distributed Etherlink Switch [DES]), are specifically designed for AI networking. All support 800-Gb/s. Notably, Arista's 7700R4 Distributed Etherlink Switches have been adopted as part of **Meta's** disaggregated scheduled fabric (DSF) for its AI datacenters (see chart following this section).

In March 2025, Arista introduced Cluster Load Balancing (CLB) as "a new Ethernet-based, RDMA-Aware, AI load balancing solution that enables high bandwidth utilization between spines and leaves." Designed to work with Arista switches based on the vendor's Extensible Operating System (EOS) software, which delivers cloud-enabled management across Arista's product portfolio, CLB is aimed at maximizing the efficiency of AI workloads.

Cisco. The vendor's existing Nexus 9000 switch series supports 1 Gb/s to 800 Gb/s in leaf/spine and modular configurations and features a range of functions to improve the reliability of Ethernet packet flows, including PFC (Priority-based Flow Control) and ECN (Explicit Congestion Notification) as well as intelligent buffering. The switch uses RoCEv2 for AI networking. It also supports Fibre Channel and IP storage.

In February 2025, Cisco announced an alliance with NVIDIA whereby Cisco's Silicon One ASICs will be placed in NVIDIA's Spectrum-X Ethernet networking platform. Specifically, Silicon One chips compatible with several Cisco switches and network fabrics will be added to NVIDIA's BlueField-3 SuperNIC, which combined with NVIDIA's Spectrum-4 Ethernet switch comprise the NVIDIA Spectrum-X platform.

Cisco also plans a separate series of switches combining Silicon One with its operating system software in a new line of Spectrum-compatible switches. Those are slated for delivery in the future.

Also in February 2025, Cisco announced a similar deal with **AMD,** in which Cisco Silicon One E100 ASICs will be combined with AMD's Pensando data processing units (DPUs) in a new line of Ethernet-based datacenter switches called the N9300 Series.

Juniper Networks. The vendor offers its QFX switches and PTX routers and line cards, all of which now support lossless 400-Gb/s and 800-Gb/s connectivity and work with Apstra AIOps intent-based software, which also is linked to Juniper's AI-Native Networking Platform and Marvis Virtual Network Assistant. Juniper wants its switches and routers to be "agnostic" and work with a variety of third-party SmartNICs.

For high-end AI networking, Juniper's QFX5240 Ethernet switch (based on **Broadcom's** Tomahawk 5 chip) is designed for leaf/spine configurations in AI/ML clusters. It supports up to 64 800-Gb/s Ethernet connections, or with breakout cabling, 128 ports at 400-Gb/s, 250 ports at 200 G/bs, or 256 port at 1 Gb/s. The switch supports RoCEv2 and features DCQCN (Data Center Quantized Congestion Notification) with ECN and PFC for congestion management. It offers dynamic load balancing and works with Apstra AIOps intent-based software.

Nokia. Nokia provides an Ethernet-based Data Center Fabric approach for AI networking, comprised of a 400-Gb/s Nokia 7220 Interconnect Router (IXR); an 800-Gb/s and 7250 IXR equipped with coherent optics support; and a Nokia 7215 Interconnect System, which provides top-of-rack, out-of-band management, including IP routing, quality-of-service control, telemetry, and model-driven management. All platforms run under the vendor's SR Linux network operating system.

Startup Networking Solutions

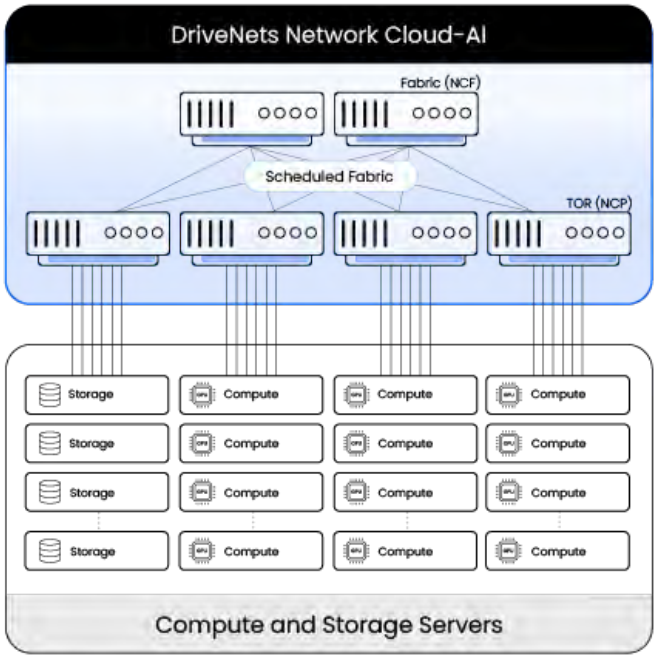
A range of vendors provide networking for AI based on open architectures. These include the following:

Arrcus offers the Connected Edge for AI (ACE-AI), a flexible networking solution that delivers a unified fabric across the network for Distributed AI, from Datacenter to Edge to Multi-cloud. For GPU connectivity in DC, ACE-AI provides lossless, congestion-free architectures: high-speed IP CLOS leaf/spine fabric supporting RDMA over Converged Ethernet (RoCE) v2, as well as Virtualized Distributed Routing (VDR) architecture with scheduled fabric. ACE-AI is hardware-agnostic and supports multiple form-factors, including white-box switches and routers with 400-Gb/s and 800-Gb/s data rates, and SmartNICs/DPUs, based on multiple silicon vendors, including Broadcom and NVIDIA. In March 2025, the vendor also unveiled TGAX, a telco-grade network switch based on NVIDIA's Spectrum Ethernet platform. The new switch is geared to helping telcos add AI capabilities, including AI to the radio access network (RAN). The platform's distributed, cloud-native network operating system, ArcOS, also works with NVIDIA BlueField DPUs for host-based networking at the Edge for inferencing use cases. ACE-AI also connects distributed AI/ML workloads across hybrid and public cloud infrastructures.

Aviz Networks has built the Open Networking Enterprise Suite (ONES), a multivendor networking stack for the open-source network operating system, SONiC, enabling datacenters and edge networks to deploy and operate SONiC regardless of the underlying ASIC, switching, or the type of SONiC. Aviz ONES provides anomaly detection, configuration automation, a RoCE fabric, and observability tools and works with third-party switches at data rates to 800-Gb/s. The vendor also features AIOps, applying AI LLMs to managing the network via an LLM-based network copilot.

DriveNets delivers a scale-out Ethernet-based architecture on industry-standard white boxes, with the potential to lower capital spending costs and provide an alternative to InfiniBand without compromise on performance. On the contrary, the vendor has multiple customers that achieved higher performance and lower Job Completion Time (JCT) using DriveNets solution than using InfiniBand, the vendor claims. DriveNet's cloud-native software serves up telco-grade networking for a series of Tier 1 operators worldwide. Two years ago, the vendor unveiled Network Cloud-AI, a solution based on a Distributed Disaggregated Chassis (DDC) approach to interconnecting any brand of GPUs in AI clusters via a cell-based fabric. Implemented via white boxes based on Broadcom components, including the Jericho-3 AI chip, the solution can link up to 32,000 GPUs at up to 800 Gb/s and deliver better performance than InfiniBand. DriveNets says that recent customer testing demonstrated 30% better performance than standard Ethernet and 5% faster performance than InfiniBand.

Through an approach that DriveNets terms Fabric-Scheduled Ethernet for AI, the vendor offers performance independent of NIC or GPU type; load balancing; support for multi-tenancy; and the elimination of incast switch congestion. The fabric supports backend compute and storage networking together, as shown in the diagram below:



A depiction of an enterprise deployment of DriveNets white boxes in an AI network. Source: DriveNets

According to Dudy Cohen, DriveNets VP, Product Marketing, the DriveNets solution is well positioned for AI due to its design: "Multiple white boxes become a single network entity," he told us recently. "This virtual chassis can support up to 32,000 GPUs in a nonblocking and lossless connectivity with a single Ethernet hop from any GPU to any other GPU in the cluster."

According to DriveNets, the fabric-based scheduling and packet spraying technology enables extremely fast deployment and network fine-tuning. The solution is currently deployed by enterprises in the pharma research and financial industries and by so-called neoclouds.

Hedgehog offers a simple, turnkey AI network solution enabling enterprise customers to network like hyper-scalers. Hedgehog includes a mesh fabric for enterprise inference at the edge, a data center fabric for RAG, and a back-end GPU fabric for training or fine tuning. Delivered as an open-source software appliance, Hedgehog offers Zero Touch Lifecycle Management (ZTLM) from day 0 provisioning to day N configuration, observability, and automated software updates. The product is built for DevOps users, not network engineers. Built on Kubernetes, Hedgehog is cloud native to the core with support for the entire cloud native toolchain. Cloud operations teams use the same people, processes, tools and skills to run private and hybrid clouds as they use for their public cloud. They scale out with Infrastructure as Code and GitOps using the Hedgehog API to provision Hedgehog VPCs for multiple AI workloads, with secure packet routing for multiple edge and cloud locations through the Hedgehog Transit Gateway. The gateway includes a data plane built on DPDK that uses NVIDIA ConnectX NICs to scale up for high performance data I/O in front end AI networks."

Building Out Hyperscaler Stacks

In general, you can describe many networking for AI solutions targeted at two areas: 1) Large core LLMs and training, which require massive scale; and 2) Inferencing at the edge, which might require support of a few dozen GPUs and CPUs to combinations of thousands.

As mentioned earlier, many UEC members, including **Arista, Broadcom, Cisco, Juniper,** and **NVIDIA,** already claim to solve Ethernet's issues in the large LLM cores by adopting technologies such as packet spraying and congestion control based on Priority-based Flow Control (PFC) and Explicit Congestion Notification (ECN). But the goal of the UEC is to unify these many proprietary innovations in a standardized way that is anchored in the economics of Ethernet and commercial off-the-shelf (COTS) hardware.

All this attention to Ethernet doesn't mean that **NVIDIA's** proprietary InfiniBand is going away anytime soon. "InfiniBand for the compute+storage network, then a 1-10-Gbs Ethernet for management and/or internet access. This is the tried-and-true setup for most HPC clusters," wrote one engineer on a recent Reddit thread. "This is almost always going to be cheaper than an Ethernet solution with comparable speeds, latency, and blocking ratios."

One thing is sure, the landscape is broadening and the market is expanding. With the growth of GPU clouds such as **CoreWeave, Lambda, Crusoe** and others as well as the emergence of hyperscaler alternatives such as **Vultr,** there is a growing landscape of cloud providers looking to host AI services.

Below is a sampling of the key elements of the leading hyperscaler AI networks. It shows that InfiniBand maintains a firm position in the largest AI datacenters, though Ethernet is gaining ground.

Leading AI Networks: A Sampling*								
	AWS (UltraClusters)	**CoreWeave**	**Google (AI Hypercomputer)**	**Lambda**	**Meta**	**Microsoft (Azure Eagle)**	**Oracle (OCI Supercluster)**	**xAI (Colossus)**
GPUs	NVIDIA H200, H100, A100; AWS Trainium1 and Trainium2	NVIDIA GB200 NVL72/HGX B200, HGX H100/H200, HGX A100, PCIe A100, L40, L40S, A40, RTX	NVIDIA H200	NVIDIA GB200 NVL72, HGX B200, H200 NVL, H200 SXM, H100 SXM, H100 PCIe, RTX Pro 6000, GH200 Superchip	Proprietary GPU platforms Catalina (with NVIDIA GB200 NVL72) and Grand Teton (with AMD Instinct MI300X)	NVIDIA H100; AMD MI300X and EPYC Genoa	NVIDIA GB200 NVL72, B200, H200, H100, A100; AMD MI300X	NVIDIA H100, H200
Accelerators	Proprietary Trainium and Inferentia chips	Proprietary CoreWeave Tensorizer	Proprietary Google Trillium TPU	N/A	Meta Training and Inference Accelerator (MTIA) ASICs	Proprietary Maia	N/A	N/A
NICs/DPUs	AWS Elastic Fabric Adapter (EFA)	NVIDIA ConnectX, BlueField-3 DPU	Google Titanium ML Network Adapter; NVIDIA ConnectX	NVIDIA ConnectX	NVIDIA ConnectX for InfiniBand; Ethernet based on Arista switches is "NIC agnostic"	NVIDIA ConnectX for InfiniBand	NVIDIA ConnectX for InfiniBand and Ethernet	NVIDIA BlueField-3 SuperNIC
Networking	AWS EC2 NeuronLink (for Trainium); NVIDIA Quantum InfiniBand, GPUDirect RDMA; Ethernet (via AWS EFA)	NVIDIA Quantum InfiniBand; NVLink	Proprietary Jupiter networking featuring optical circuit switches (OCS), wave division multiplexing (WDM), software-defined networking (SDN); NVIDIA Quantum InfiniBand	NVIDIA Quantum InfiniBand, NVLink, Ethernet	Arista 7700R4 Distributed Etherlink Switch; NVIDIA Quantum InfiniBand	NVIDIA Quantum InfiniBand	NVIDIA Quantum InfiniBand; Arista Ethernet; NVLink	NVIDIA Spectrum-X Ethernet
Software	AWS Direct Connect	Open-source development tools; NVIDIA	JAX, OpenXLA TensorFlow, PyTorch; other	Managed Kubernetes; managed	PyTorch	Open-source Ubuntu;	OCI Generative AI Agents; OCI	TensorFlow, PyTorch;

		software, including Run:ai; Weights & Bases (acquired by CoreWeave)	open-source development tools	Slurm (H100, B200); Lambda Stack (includes TensorFlow, PyTorch, Keras, Ubuntu, CUDA); observability plug-in (opt-in)		NVIDIA software	Kubernetes Engine; NVDIA software	NVIDIA software
Storage	NVMe SSD instances; FSx for Lustre; Amazon Elastic Block Store; Amazon S3	Proprietary local, file, object options; third-party options; GPUDirect functionality	Hyperdisk ML block storage, Cloud Storage FUSE, Filestore, Parallestore	VAST Data Platform; persistent storage using solid state disk drives; direct-attach storage; S3 Adapter for filesystems	Linux FUSE API; proprietary Tectonic platform; Hammerspace network file system	Azure Blob object storage; open-source Lustre parallel file system; NetApp files; SSD	NVMe SSD; fiile, block and objecct options	DDN; VAST Data Platform; Supermicro hardware
Other features	Graviton CPUs	Kubernetes services; virtual private cloud; AMD EPYC Genoa and Intel Emerald Rapids CPUs	Integration with Google Kubernetes Engine (GKE)	Inference API; Private Clouds deployments for Enterprise and hyperscalers; GPUDirect RDMA; on-premise GPU desktop & server systems	N/A	Intel Xeon CPUs	Azure Kubernetes Service	Kubernetes support for GPU resource allocation; Intel Xeon CPUs

**This chart has been updated from a version in the last Futuriom report on Networking for AI in June 2024.*

Getting Smart About NICs

Network interface cards (NICs), which connect servers and storage to the AI network, are another key element of networking for AI. These components are classified as smartNICs if they contain processors or processes that offload a range of tasks from the AI server CPUs, including storage, security, and network management functions.

NICs don't necessarily have to have special chips or data processing units (DPUs) to offload tasks from the CPU in high-throughput AI networks. **Broadcom's** 400-Gb/s Ethernet adapters, for instance, aren't classified by the vendor as smartNICs, though they do feature a proprietary TruFlow functionality that classifies packets to obviate the need for extra processing in the CPU.

In contrast, **NVIDIA's** BlueField-3 SuperNICs DPUs offload not just networking but also storage and security tasks such as encryption from the CPUs in an AI cluster.

Most NICs and smartNICs are designed to work with a range of third-party switches—though some products, such as **NVIDIA's** ConnectX and BlueField-3 SuperNICs, work optimally with the vendor's Quantum InfiniBand and Spectrum-X Ethernet switching platforms. SmartNICs typically support up to 400-Gb/s total bandwidth today, though some, such as NVIDIA's ConnectX-8, offer 800-Gb/s bandwidth. And **Enfabrica's** Accelerated Compute Fabric SuperNIC Chip offers 3.2 Tb/s through its innovative design, which supports multiple 800-Gb/s ports.

Napatech, which specializes in NICs and smartNICs based on Intel Xeon x86 processors, notes that programmability is a key feature of these components. But the vendor specifies that it's not programming the NIC itself that defines programmability necessarily. Instead, Napatech notes that the ability to upgrade a NIC without having to swap it out is vital to ensuring cost efficiency.

Interestingly, **Arista Networks** has adopted a "NIC-agnostic" approach via its recent announcement of Cluster Load Balancing (CLB) for its switches. CLB can work with or without third-party NICs, which is an important differentiator from NVIDIA, whose networking offerings are heavily geared to the use of its BlueField NICs. The use of NICs has grown in AI networks, because they can offer additional processing power to offload networking functions including encryption. But they also add cost to the network, and Arista is saying—we can work with NICs or not, but we think we can also load balance without them, if you want to.

NICs and SmartNICs for AI Networking; A Sampling

Vendor	Product	No. of Network Ports	Supported Networks	Total Bandwidth	Host Interface	RoCE Support	General-purpose Programmability	Other Features
AMD	AMD Pensando Pollara 400 AI NIC	4	Ethernet, UEC Ethernet	400 Gb/s	PCIe 5, 16 lanes	Yes	Yes	Supports UEC Ethernet; GPUDirect
	AMD Pensando Salina 400 DPU	2	Ethernet, UEC Ethernet	400 Gb/s	PCIe 5, 16 lanes	Yes	Yes	Load balancing, routing, storage acceleration, advanced security; supports UEC Ethernet
Broadcom	N1400GD	1	Ethernet	400-Gb/s	PCIe 5, 16 lanes	Yes	N/A	Proprietary TruFlow flow processing and TruManage for management
Enfabrica	Accelerated Compute Fabric SuperNIC Silicon	32	Ethernet	8 Tb/s	PCIe, 160 lanes	Yes	Yes	Proprietary Resilient Message Multipathing for RoCE and TCP; RDMA over TCP; network-attached scalable memory
Napatech	N3070X AI SmartNIC	1, 2, 4	Ethernet	400 Gb/s	PCIe 5, 16 lanes, +CXL 2.0	Yes	Yes	2 additional PCIe 5 x 16 ports

NVIDIA	BlueField-3 SuperNIC	1, 2	Ethernet, InfiniBand	400 Gb/s	PCIe 5, 32 lanes	Yes	Yes	GPUDirect/GPUDirect Storage connectivity; SNAP storage virtualization
	ConnectX-8 SuperNIC	1, 2 (can be split up to 8 ports)	Ethernet, InfiniBand	800 Gb/s	PCIe 6, 48 lanes	Yes	No	GPUDirect/GPUDirect Storage; NVIDIA SHARP protocol support; port splitting into multiple virtual ports; telemetry-based congestion control
	ConnectX-7	1, 2, or 4	Ethernet, InfiniBand	400 Gb/s	PCIe 5, 32 lanes	Yes	No	ASAP2 packet processing

Source: Company reports and information.

What About Optics?

AI networks have increased demands on power, speed, and space. While copper connectivity remains the standard way to link rack-level elements, vendors are looking toward faster fiber interconnections to connect AI clusters together, both within datacenters and between them. This calls for technology that eliminates the need for optical-to-electronic pluggable transceivers, which can quickly multiply cost and complexity as AI networks grow.

At its GTC 2025 conference, **NVIDIA** announced new versions of its Quantum InfiniBand and Spectrum-X switches based on co-packaged optics (CPO). This technique replaces optical-to-electronic transceivers with silicon photonics placed directly on the same substrate as the ASIC associated with the CPU or GPU.

In a press release, NVIDIA claims its new CPO switches will deliver "3.5x more power efficiency, 63x greater signal integrity, 10x better network resiliency at scale and 1.3x faster deployment compared with traditional methods." The new Quantum switch, due later in 2025, will feature 144 ports of 800-Gb/s InfiniBand. The new Spectrum-X Photonics Ethernet switch, due in 2026, will sport 512 ports of 800-Gb/s Ethernet or 128 ports of 800 Gb/s Ethernet. Both products will be liquid-cooled.

NVIDIA isn't alone in developing optical switches for AI networking. In March 2024, **Broadcom** announced the Bailly 51.2T Ethernet CPO switch system, which packages eight silicon photonics-based, 6.4-Tb/s optical engines with Broadcom's Tomahawk 5 switch chip. The "Bailly" has been in development with a number of suppliers, including **Micas Networks,** whose CPO switch was released in March 2025. The Micas product features 128 ports of 400-Gb/s fiber connectivity in a 4U, air-cooled system.

Google has created its own optical networking technology, Jupiter, for its AI Hypercomputer service offering. This features optical circuit switches (OCS), wave division multiplexing (WDM), and software-defined networking (SDN). It forms the network for Google's AI Hypercomputer services.

Datacenter Interconnect Trends

When it comes to datacenter interconnection (DCI), fiber is the rule. **Ciena** has innovations in this area. The vendor's WaveLogic 6 Nano 1.6T Coherent-Lite pluggable uses coherent optics, a technology that manipulates light to achieve greater data capacities over fiber. The transceiver is compatible with third-party routers and switches, including ones that support 200G/lane electrical interfaces on the host side. The lead application for Coherent-Lite is for campus applications, interconnecting datacenters across distances from 2 to 20 kilometers, though Ciena also has its eye on intra-datacenter solutions.

Nokia supports 400- and 800-Gb/s coherent optics in its 7250 Interconnect Router for DCI. Nokia also recently announced a series of low-power 1.6-Tb/s pluggable optical Intra-Data Center Connectivity Solutions that it says will drive down power by as much as 70% when deployed in transceivers and other networking options for AI datacenters. The chips, acquired with Nokia's purchase of Infinera for $2.3 billion in February 2025, can be used in digital signal processors (DSPs) and CPO products, Nokia says.

Many enterprises look to datacenter service providers for AI DCI. **Equinix** is a key player in this space and features worldwide fiber interconnections. It offers Platform Equinix, a complex ecosystem of partners in a variety of areas, including cloud services, SaaS providers, GPU vendors, and networking infrastructure providers. Equinix provides Platform Equinix in 73 metros across 34 countries. Equinix also provides Equinix Fabric, a private AI networking solution that bypasses the global Internet with thousands of worldwide physical and virtual endpoints, including those of digital partners and cloud providers.

Overall, it's clear that optical technologies will continue to integrate into AI networks at all levels, creating innovations in power savings, speed, and transmission reliability.

Rack-level Interconnection Trends

When it comes to the most basic I/O interconnections linking CPUs, GPUs, storage devices, and NICs within datacenter racks, a few trends are apparent. First is the effort to improve the Peripheral Component Interconnect Express (PCIe), which remains essential to connecting rack components. PCIe, currently in its 5.0 release with 6.0 emerging, is also set for a new release 7.0 this year from the PCI-SIG group responsible for the spec's upgrades. Release 7.0 will increase PCIe's bidirectional speed to 512 Gbytes/second compared to the present 6.0 capability of up to 256 Gbytes/second and 5.0's 128 Gbytes/second.

The problem with PCIe is that as speeds increase, signal reliability and strength decrease across distances or in complex networking environments. This requires the use of retimers, which act to retransmit I/O signals to relieve jitter and distortion. **Broadcom** offers PCIe switches and retimers and recently released a series of products compatible with PCIe 6.0. **Marvell** recently struck a five-year deal to supply AWS with PCIe retimers along with a range of other AI datacenter interconnect products. Marvell also is demonstrating a Marvell Alaska P PCIe Gen 6 retimer and a PCIe Gen 7 SerDes technology for running PCIe transmission over optical fiber. According to Marvell's press release: "With PCIe over

optics, system designers will be able to take advantage of longer links between devices that feature the low latency of PCIe technology."

Another spec in play is Compute Express Link (CXL), implemented and governed by the CXL Consortium, whose members include **Broadcom** and **Marvell** as well as **NVIDIA, Astera Labs,** and **Enfabrica,** among many others. Based on PCIe's physical layer, this connectivity option unifies the memory space between GPUs, CPUs, NICs, and other components, making it possible to allocate resources dynamically, reducing latency and reducing cost and complexity in AI networks. The CXL Consortium in December 2024 released its latest specification 3.2, adding management and monitoring functionality along with a Trusted Security Protocol (TSP).

There are also rack-level interconnection innovations coming from non-standard sources. As part of the movement to improve on traditional PCIe interconnection scenarios, startup **Enfabrica** replaces the smart NICs and PCIe switches in an AI networking rack, offering faster connections from the network to the AI system; reducing latency associated with traffic flows between NICs and the GPUs; and

supporting a lower total cost of ownership (TCO) for AI systems. Thanks to its design, Enfabrica's Accelerated Compute Fabric (ACF) SuperNIC Chip can support up to 32 Ethernet ports on the networking side and 160 PCIe lanes, substantially outstripping capacity for other NICs.

Astera Labs, which went public in 2024, provides a series of products designed to streamline AI processing in both scale-up and scale-out functions. Its Scorpio Smart Fabric Switches are software-definable, meaning the firmware can be tweaked according to customer requirements. At NVIDIA's GTC conference in March, Astera previewed its Scorpio Smart Fabric Switches for PCIe 6-ready NVIDIA Blackwell-based MGX platforms. Astera also makes a series of PCIe/CXL retimers, which shorten the distance between CPUs, DPUs, storage, and other elements in AI servers by retransmitting physical-layer signals, improving overall performance.

NVLink and Its Challengers

Linking GPUs and CPUs in AI networks requires high-speed connectivity capable of supporting multiple parallel ports. Leading the pack is **NVIDIA's** proprietary NVLink, which has been the chief connectivity choice in HPC racks since its introduction in 2014. NVIDIA has increased NVLink's profile from an 8-channel supporting 900-Gbytes/second to a 72-channel format for use in NVIDIA's GB200 NVL72 Blackwell configuration. The expanded channel capacity is designed to support the powerful GB200 platform combination of 36 Grace Hopper CPUs with 72 Blackwell GPUs with a total data throughput rate of 1.8 Tbytes/second.

Other vendors offer similar solutions, and work is underway to provide an open-source alternative to NVLink. **AMD,** for instance, offers an Infinity Fabric architecture to connect its Instinct accelerators and EPYC processors in AI clusters. AMD is also a leader in the UALink Consortium, developers of an Ultra Accelerator Link (UALink) standard aimed at eliminating the need for proprietary interconnects such as NVLink.

Founded in October 2024, the UALink Consortium has an impressive list of members, including, besides **AMD, Apple, Astera Labs, AWS, Broadcom, Cisco, Enfabrica, Google, HPE, Intel, Juniper Networks, Meta, Microsoft,** and **Synopsys,** to name just a few.

On April 8, 2025, the UALink Consortium released its first 200G 1.0 Specification. The spec, aimed at cloud hyperscalers, OEMs, and chip providers, creates what the group calls a "switch ecosystem" that is designed for high performance, low latency, low power, and a cost-efficient form factor. The spec defines an interconnect for GPUs in back-end networks that supports 200-Gb/s bidirectional data rates for 1, 2, or 4 lanes connected to up to 1,024 accelerators in a pod. Hence, maximum bidirectional bandwidth is 800 Gb/s.

3. Enterprise AI at the Edge: What Does AI Mean for Inferencing and Security?

The current boom in AI has been focused on infrastructure to build LLMs, where the bulk of money is being spent now. But this is changing, as AI impacts the inference market, which in turn affects network connectivity at the edge.

At the same time, we are in the very early days of Enterprise AI adoption. This runs the gamut from how the large software as-a-service companies such as Adobe, Salesforce, and ServiceNow adopt AI and AI agents for their customers, as well has how vertical enterprises in industries ranging from energy to agriculture develop targeted AI applications for productivity.

How the infrastructure and networking will be built to support these wide-ranging applications will vary widely, but one thing is sure: They'll need new infrastructure, with new features, and that will have to be supported with networking and new forms of security.

RAG and SLMs at the Edge

The need for edge processing for inferencing is clear: Edge processing can reduce latency and improve throughput. For inferencing, which will entail RAG and agentic AI to optimize processing, frontier LLMs won't be required as much as emerging small language models (SLMs) that will enable enterprises to quickly adapt their data to AI applications.

Implementing RAG and SLMs at the edge requires development tools optimized for the purpose. This approach will also have a huge impact on how core executive teams think about infrastructure and security. How do you think about data and data sovereignty? Do you build your AI infrastructure in a partner cloud, or do you build private infrastructure? Who do you partner with on network connectivity?

SD-WAN, SASE, and NaaS Adapt for AI

Developing trends in the architecture of distributed applications such as the use of hybrid cloud, cloud networking, Kubernetes, and AI are accelerating the need for distributed network security.

How do you build and secure this enterprise AI infrastructure? As we've covered on Futuriom.com, we think the recent series of "typhoon" attack by nation states on critical infrastructure in the United States and elsewhere have serious implications for security architectures and how enterprises think about hosting networking infrastructure. Can your networking partners be trusted? What are the security measures?

Traditional networking vendors such as **Cisco,** as well as dominant network security vendors such as **Palo Alto Networks,** will have challenges as they try to integrate their portfolios to meet the challenges of distributed AI network security. Integration is a heavy lift with fragmented product portfolios that were accumulated by M&A. Emerging SASE and network security vendors such as **Versa Networks** and **Aryaka,** as well as NaaS providers such **Alkira,** are providing AI-targeted networking infrastructure and services.

Futuriom sees a sea-change ahead in the SASE and SD-WAN market, as network architects think about providing pervasive, cloud-native networking security to support any network endpoint in any environment, whether that's a cloud network or an enterprise data center.

When using network-as-a-service or private datacenters, enterprises will look to service providers and infrastructure companies that can support distributed, secure network fabrics or NaaS. Emerging technologies such as eBPF, Cilium, Calico, and Istio, as well as multicloud networking (MCN), can come in to help connect diverse network assets. Enterprises will be looking for services aimed at supporting AI workloads at the edge using a software-defined, secure network for cloud applications across multiple private clouds, datacenters, and public clouds.

These distributed networking solutions should combine application layer and network layer functions to eliminate the complexity of traditional routed networks while ensuring lower latency, higher throughput, and better security and observability. These features allow edge applications and devices to securely and efficiently process and deliver information, including AI inferencing. We also think enterprises will take different approaches, whether it's using a do-it-yourself approach by building networks with vendor tools or adopting a NaaS approach.

For example, **Aryaka** is a provider of converged network and network security delivered as-a-service in over 100 countries. Lately, it has been further evolving the unified power of both a private network as well as a SASE platform to help large organizations lock down their security in the age of AI, wrapping this strategy around unified SASE-as-a-service. By owning its own network assets, Aryaka has the additional flexibility of offering layer 2 or layer 3 network services integrated with security functions. The recent February 2025 release of the platform includes the addition of powerful AI-powered data observability and security services as well as the capability to add worldwide dynamic points-of-presence (PoPs).

Versa Networks offers a variety of SASE, SD-WAN, and security analytics features that run on a single platform and operating system, which can be used to build a secure distributed networking fabric that can be hosted either on premises or in the cloud. This secures application and user connectivity across the entire enterprise.

Versa supports AI-ready networking and infrastructure by enabling distributed processing of AI workloads across the network edge. Through intelligent traffic steering, application-aware routing, and integrated security, Versa ensures that AI data flows are prioritized, optimized, and protected as they

move between cloud, core, and edge locations. This distributed approach reduces latency, improves inference performance, and enables real-time responsiveness—key requirements for AI-powered applications. By embedding AI-aware capabilities directly into the network fabric, Versa helps organizations push compute closer to where data is generated and decisions are made, laying the foundation for scalable, edge-driven AI architectures.

Greater Focus on MCN and Hybrid Networking

With AI arriving, distributed fabrics and MCN functionality becomes more important. Partnerships among MCN vendors, cloud providers, and colocation providers will become more important. For example, Futuriom recently looked at partnerships among **Aviatrix, Equinix,** and **Megaport** to provide more flexible multicloud to deliver a secure global network fabric. Developing trends in the architecture of distributed applications such as the use of hybrid cloud, cloud networking, Kubernetes, and AI are about to accelerate these trends.

Arrcus, for instance, offers Arrcus's ACE-AI, which deploys the vendor's distributed, cloud-native network operating system—ArcOS—to help build networks on the fly that are optimized for AI workloads at the edge. ACE-AI deploys Ethernet along with VDR, PRC, intelligent buffering and congestion control, and visibility functions to deliver edge services that are hardware-agnostic and will work with white-box switches and routers. In March 2025, Arrcus announced a partnership with OEM Lanner Electronics that pairs the ACE-AI network with Lanner's implementation of NVIDIA's MGX Server to power telco AI-RAN and enterprise inference applications.

Orchestration at the Edge

Netris is focused on the infrastructure orchestration layer, building a virtual private cloud that furnishes routing, load balancing, security, and other essential network services along with monitoring and analysis and automated remediation of network anomalies. It enables NCPs (NVIDIA Cloud Partners) and other tier 2/3 cloud providers to build and operate multi-tenant and highly automated networks for launching GPU and CPU cloud services.

ZEDEDA specializes in orchestrating devices and applications associated with edge nodes by providing only the operating system needed to run the application at the edge; there are no runtimes or libraries to slow things down. ZEDEDA also works via its own API with Kubernetes, Docker, and virtual machines to speed up edge applications and reduce operational costs – all of which play into more efficient networking of AI inferencing. In 2025, the vendor added integration with NVIDIA's Jetson systems, NGC catalog, and TAO toolkit.

CDNs Will Need to Adapt

At the edge, content delivery network (CDN) providers are also looking to get in the game with AI infrastructure and services. So far, they're rolling out AI-based services and infrastructure as well as

talking a good game about inferencing, although investors don't appear to have bought into the story that CDNs are a major component of the AI boom.

Akamai, for example, in March 2025 unveiled Akamai Cloud Inference, a series of tools that give developers the means to run inference tasks over Akamai's network of edge nodes—comprising more than 4,200 points of presence across over 1,200 networks in over 130 countries worldwide. Akamai has equipped its Akamai Cloud with NVIDIA GPUs and NVIDIA Enterprise tools, as well as data management capabilities from VAST Data. A Linode Kubernetes Engine delivers containerization, while edge compute is powered by WebAssembly (Wasm) technology powered by **Fermyon.**

"While the heavy lifting of training LLMs will continue to happen in big hyperscale data centers, the actionable work of inferencing will take place at the edge," said Adam Karon, Chief Operating Officer and General Manager, Cloud Technology Group at Akamai, in the press release.

Akamai investors, however, remain unimpressed, just by judging this stock chart in comparison to, say, NVIDIA's stock. Of course, much of this is down to the revenue growth reported by both companies. Akamai has yet to see a substantial financial benefit.

Cloudflare, likewise, has equipped its worldwide datacenters with NVIDIA GPUs so that the vendor's network can extend inference services to enterprise customers via its CDN. Cloudflare offers a range of LLMs for use at the edge, an object storage service, and a Vectorize vector database that facilitates the creation and storage of embeddings, which are crucial to inference processing.

4. Conclusion

When the so-called AI revolution started two years ago, AI training of LLMs took center stage. Only the cloud hyperscalers and the world's wealthiest enterprises could afford the components and IT infrastructure that training required. But 2025 has shifted the focus to inferencing, or the adaptation of enterprise data to AI models to create specific AI applications.

However, AI infrastructure needs are expanding as its clear that inferencing for AI will drive many use cases. Many enterprises we have spoken with will be adopting AI for very specific use cases, some of which will require the large firepower of hyperscalers and others will only require a few GPUs at the edge. In addition, RAG and agentic AI will change how architects think about the fundamental design of their IT infrastructure. Requirements for improved latency and parallel processing by GPUs calls for a shift from traditional centralized computing to distributed, accelerated computing. This architecture is characterized by racks of concentrated networking, compute, and storage resources linked by high-speed interconnections to form so-called AI factories.

At the rack level, networking for AI requires links that not only feature low latency but support the high-speed parallel processing of data that GPUs provide for training and inferencing. Between rack-level devices such as switches and storage, PCIe continues to be a suitable connection, while **NVIDIA's** proprietary NVLink serves as the preferred method of linking GPUs together. For linking racks or clusters in formations that allow multiple units to act as one, HPC networking has long favored **NVIDIA's** InfiniBand and its associated adapters.

As AI networks expand and inferencing grows in popularity and demand, a trend has emerged toward standardizing each of these networking elements with alternatives to **NVIDIA's** dominance. Groups have coalesced to address links within racks, between racks, between GPUs and CPUs, and across multiple racks or pods. These include the following:

- **PCI-SIG:** This group has standardized the PCIe 5.0 release with 6.0 emerging and is preparing release 7.0, which will increase PCIe's bidirectional speed to 512 Gbytes/second compared to the present 6.0 capability of up to 256 Gbytes/second and 5.0's 128 Gbytes/second.
- **CXL Consortium:** This group is developing the Compute Express Link (CXL) connectivity option, which is based on PCIe's physical layer and unifies the memory space between GPUs, CPUs, NICs, and other components, ensuring dynamic allocation of resources, reduced latency, and lower cost and complexity in AI networks.
- **UALink Consortium**: This group has developed an Ultra Accelerator Link (UALink) standard aimed at eliminating the need for proprietary interconnects such as NVLink. The spec's first release is aimed at cloud hyperscalers, OEMs, and chip providers and creates what the group calls a "switch ecosystem" that is designed for high performance, low latency, low power, and a cost-efficient form factor.

- **Ultra Ethernet Consortium (UEC).** This group aims to make Ethernet the standard for scaling out racks in AI datacenters by improving the networking protocol's performance and speed through ameliorating packet loss, lack of adequate load balancing, congestion management, and other drawbacks. The goal is to match the job completion time (JCT) of NVIDIA's InfiniBand in HPC and AI environments.

All of these efforts point the way toward networking for AI that reduces reliance on costly proprietary solutions, supports open-source standards, and encourages competitive product offerings. The ultimate goal is to democratize networking for AI, making it increasingly accessible to enterprises of all sizes.

Along with the standardization trend in AI networking is a focus on improving security and automation. SASE, SD-WAN, and network-as-a-service (NaaS) are evolving to eliminate the complexity of traditional routed networks while ensuring lower latency, higher throughput, and better security and observability.

AI networking also heralds an increase in deployment of MCN to link distributed and hybrid fabrics supporting Kubernetes and AI. Partnerships among MCN vendors, cloud providers, and colocation providers will become more important, as will involvement of specific AI providers.

Increased networking for AI will also spur the growth of services. So-called neocloud providers such as **CoreWeave** and **Lambda** are expanding their roster of services beyond GPU provisioning to meet developer needs for agentic AI, RAG, and other inferencing tasks. **Lumen** and **IBM,** for instance, recently unveiled a partnership that pairs IBM's watsonx AI solutions with Lumen's global network (which features GPUs for networking for AI) to bring inferencing closer to customers, improving performance and JCT.

All of these trends point to the increase of networking for AI across enterprises worldwide. As inferencing increases, drawing with it a focus on agentic AI and RAG, the trends will grow. Futuriom anticipates that the inferencing market will advance by double digits to hundreds of billions of dollars by the end of the decade, powered by innovations that continue to emerge.