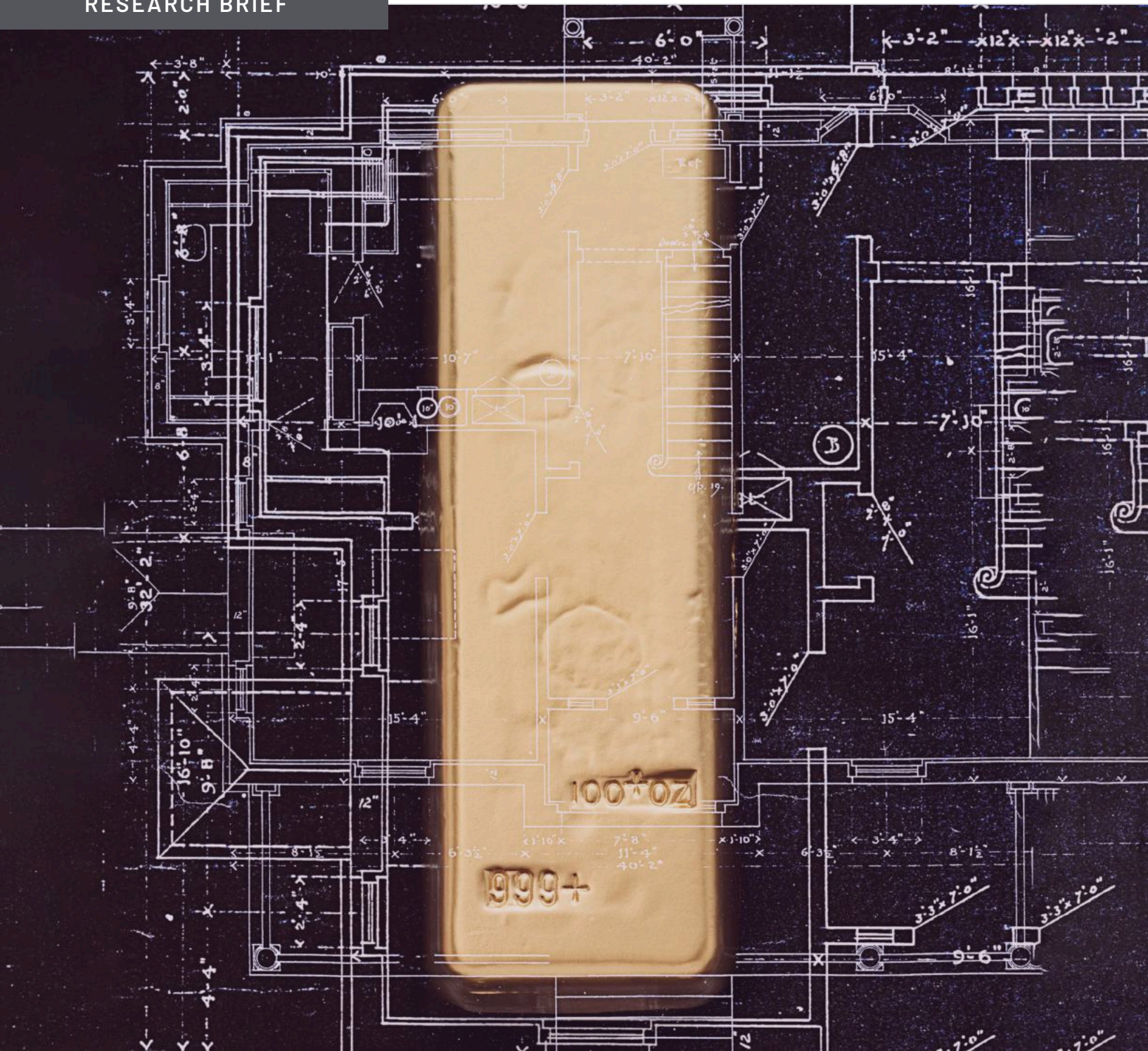


# Prospecting for Performance: Data Center Networking in 2025

Networking Blueprints for the AI and Cloud Era

RESEARCH BRIEF



# Table of Contents

<b>Report Background .....</b>	<b>2</b>
<b>Data Center Workload Evolution.....</b>	<b>2</b>
AI Drivers.....	2
Meanwhile, on the Cloud Front, Growth Continues .....	4
<b>Data Center Architecture Evolution .....</b>	<b>5</b>
The Shift from Training to Inference: Infrastructure Implications .....	5
Distributing Workloads – Distributed, Federated Training and Inferencing.....	6
What is the New Unit of (AI) Computing? .....	6
<b>Data Center Networks – Scale Up to Scale Out to Scale Outside .....</b>	<b>6</b>
Die-to-die: UCIe and Chiplets .....	6
System Level - PCIe.....	7
System Level – CXL.....	8
System Level – NVLink.....	8
System Level – UALink.....	8
Scale-Out Networking – Preamble .....	9
Scale-Out – InfiniBand .....	11
Scale-Out – Ethernet and RoCE .....	12
Scale Out – UEC and UET .....	12
Scale-Out/Scale Outside – Frontend Ethernet Networks.....	14
Scale Outside – Data Center Interconnects (DCIs) .....	14
<b>Other Data Center Considerations.....</b>	<b>15</b>
Security and the DPU.....	15
Open Source and SONiC .....	15
CPO, LRO, and LPO: Transforming Data Center Interconnects .....	16
<b>Vendor Landscape.....</b>	<b>17</b>
Select Networking Vendors .....	17
Hyperscalers .....	19
Other Ecosystem Vendors .....	20
<b>Observations and Recommendations .....</b>	<b>22</b>
Key Observations on Data Center Networking.....	22
AvidThink Recommendations .....	23
<b>Wrap-Up .....</b>	<b>24</b>

# Prospecting for Performance: Data Center Networking in 2025

## Networking Blueprints for the AI and Cloud Era

### Executive Summary

The data center networking landscape is undergoing rapid transformation driven by generative AI workloads. Major industry players are directing unprecedented investments into AI infrastructure – Microsoft alone plans USD 80B in data center buildouts for 2025. This massive investment fundamentally reshapes traditional architectures as AI training and inference demands push networks to new performance extremes: 400/800Gbps backend speeds, sophisticated congestion control, and ultra-low latency. However, this expansion faces significant energy generation constraints, with projections indicating a need for 47 GW of incremental power generation capacity in the US through 2030.

The industry is responding with innovations across both scale-up and scale-out networking solutions. At the silicon level, companies like Broadcom and Marvell provide the foundational technology powering these connections. Within racks, proprietary interconnects like NVIDIA's NVLink compete with emerging open standards such as UALink, while at the scale-out level, both InfiniBand and Ethernet solutions are evolving to meet AI workload demands. The Ultra Ethernet Consortium's development of the UET protocol signals strong industry momentum toward open standards, though proprietary solutions maintain significant performance advantages.

A key architectural shift is emerging: the fundamental unit of AI computing is moving from individual servers to integrated rack-scale systems, exemplified by NVIDIA's GB200 NVL72 platform and AWS's Trainium2 UltraServer. The networking vendor landscape is adapting rapidly, with traditional players like Arista, Cisco, Juniper, and Nokia being joined by innovative startups like Arrcus and DriveNets, while NVIDIA maintains a unique position offering both networking solutions and AI accelerators. These companies are developing new architectures optimized for AI workloads, incorporating features like cell-based switching, advanced congestion control, and sophisticated telemetry. The industry is also seeing significant advancement in data center interconnect (DCI) technologies, with rapid adoption of 400ZR/ZR+ modules and development of 800ZR/ZR+ solutions, critical for enabling distributed AI training across geographically dispersed facilities.

Successful data center networking strategies will need to balance competing priorities of performance, openness, scalability, power efficiency, and security while maintaining flexibility for an increasingly distributed AI computing future.

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](mailto:research@avidthink.com).

### About AvidThink

AvidThink is a research and analysis firm focused on providing cutting-edge insights into the latest in infrastructure technologies. Formerly SDxCentral's research group, AvidThink launched as an independent company in October 2018. AvidThink's coverage includes 5G infrastructure, enterprise networks, private wireless, edge computing, SD-WAN, SASE, SSE, ZTNA, AI/cloud infrastructure, and infrastructure security. Our clients range from Fortune 500 enterprises and hyperscalers to tier-1 communications service providers, fast-growing unicorns, and 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 [avidthink.com](http://avidthink.com).

# Prospecting for Performance: Data Center Networking in 2025

## Networking Blueprints for the AI and Cloud Era

### Report Background

This report examines how generative AI (GenAI) and continued cloud adoption are transforming data center networking in 2025. Our analysis focuses on how AI is reshaping network architectures and hardware requirements in both new and existing data centers, while considering the impact of cloud-centric workloads served by hyperscalers and web giants. The report investigates the demands of GenAI workloads, including the surge in AI training, the rise of AI inference, and their implications for data center networks. While not every enterprise will run a 100K+ GPU or xPU cluster<sup>1</sup>, the learnings from leading foundation model builders and hyperscalers will be valuable. Our target audience is CxOs, network architects, and network engineers at enterprises and service providers seeking actionable insights for their data center strategies. We look forward to hearing your feedback on this report at [research@avidthink.com](mailto:research@avidthink.com).

### Data Center Workload Evolution

2024 continued to see a rapid shift in data center workloads toward GenAI. While cloud workloads, including web applications, collaboration, and streaming media, maintained modest growth, AI training workloads have fundamentally disrupted the data center market.

### AI Drivers

The AI landscape has expanded beyond OpenAI's ChatGPT (GPT4 family) to include Google's Gemini, Anthropic's Claude, xAI's Grok, and numerous members of Meta's Llama family. International players with new AI models and services like Mistral's Le Chat, Alibaba's Qwen, and DeepSeek-R1 from the Chinese hedge fund High-Flyer-backed company have emerged as strong competitors to US West Coast-based AI leaders.

In parallel, GenAI user interfaces and co-pilot tools saw rapid adoption throughout 2024, with major enterprise software/SaaS vendors aggressively investing in foundation model (FM)/large language model (LLM) integrations. The second half of 2024 saw growing enthusiasm for autonomous agentic AI, enabled by new generations of foundation models with improved reasoning capabilities. OpenAI and Anthropic led advances in chain-of-thought reasoning and test-time scaling, allowing models to break tasks into smaller steps with traceable reasoning. The drive for GenAI monetization has accelerated the agentic wave, with 2025 expected to be the breakthrough year for commercial agentic AI deployments<sup>2</sup>.

### Hyperscalers and GPU-as-a-Service

This proliferation and intensifying arms race in model building is driving unprecedented demand for AI-dedicated data center capacity, outstripping the availability of xPUs, power, space, and talent. This has fueled record growth in NVIDIA's enterprise value and steep rises in valuations of dedicated xPU data center operators like CoreWeave, Lambda, and Crusoe.

**Goldman Sachs estimates about 47 GW of incremental power generation capacity will be required to support US data center power demand growth cumulatively through 2030.**

<sup>1</sup> We will use GPU and xPU interchangeably in this report where xPU refers to GPUs and other AI accelerators (NPU, LPU, etc).

<sup>2</sup> "Jensen Huang Declares the Age of "Agentic AI" at CES 2025 – A Multi-Trillion-Dollar Shift in Work and Industry", Yahoo Finance, January 13, 2025

Major investments showcase the scale of commitment:

- **Microsoft:** USD 80B planned for data center buildouts in 2025, primarily for AI<sup>3</sup>
- **Meta:** USD 60-65B investment "primarily on data centers and servers,"<sup>4</sup> representing a 60-70% increase from 2024
- **AWS:** USD 11B committed in Georgia for "infrastructure to support AI and cloud technologies" plus Project Rainier, which is an initiative to build an ultra cluster with hundreds of thousands of its Trainium chips to serve customers like Anthropic

### The AI Data Center Gold Rush

The infrastructure expansion continues at an unprecedented scale. For example, co-location provider Equinix is involved in a USD 15B joint venture with GIC and CPP Investments to build over 1.5 gigawatts of new AI data center capacity. Meanwhile, OpenAI, SoftBank, Oracle, and GPU-as-a-service player Crusoe announced project Stargate, funded by a planned investment of USD 500B over five years with a USD 100B immediate commitment<sup>5</sup>.

This expansion requires massive investment in data center networking to connect servers within racks, between racks in rows, and across data centers. For example, xAI's Colossus at 100K GPUs represents an estimated USD 3B to 4B systems CAPEX, with an estimated USD 300M (using a 10% networking to system cost ratio<sup>6</sup>) dedicated to networking infrastructure.

### Balancing Drivers for AI Workloads: Known Unknowns

Despite the exuberance, several factors could moderate AI workload growth:

- **Power Availability:** Current commitments suggest approximately 20 GW of data center capacity in 2025 alone. Goldman Sachs estimates about 47 GW of incremental power generation capacity will be required to support US data center power demand growth cumulatively through 2030.<sup>7</sup> Some operators are pursuing innovative solutions – Crusoe Energy combines GPUaaS with technology to capture oil and gas flares for power generation, while European providers like **NexGencloud** and **Taiga Cloud** differentiate through green power strategies. Microsoft's commitment to use 100% of the revived Three Mile Island nuclear power plant's output for AI data centers is another example of the search for sustainable power sources<sup>8</sup>.
- **DeepSeek Impact:** The release of the DeepSeek-V3 base model and DeepSeek-R1 reasoning model with open weights has prompted a reevaluation of AI infrastructure investment economics. While it's too early to determine long-term implications, this development has sparked industry-wide discussions about infrastructure strategy.
- **Monetization Challenges:** Despite OpenAI's growth to USD 3.6B in revenue by September 2024 (3x year-over-year), returns remain disproportionate to infrastructure investment. Analyst firm Forrester reports that only 20% of businesses reported earnings benefits from AI in 2024, despite billions invested in infrastructure<sup>9</sup>.
- **Inference Focus:** The shift toward inference workloads and a potential slowdown in LLM training infrastructure spending may affect investment patterns. The trend toward test-time scaling, highlighted by NVIDIA CEO Jensen Huang at CES 2025, and demonstrated by DeepSeek and OpenAI's o1/o3 reasoning models, could impact data center architecture and networking requirements.
- **Efficiency Innovations:** The rise of open-source models, sparsification techniques, and improvements in GPU efficiency could temper the rate of infrastructure growth. While the impact that wide availability of open-source/open-weight models may have on in-house training investments remains unclear, greater efficiency and test-time scaling may influence AI cluster architecture and networking requirements.

---

<sup>3</sup> "The Golden Opportunity for American AI," Microsoft on the Issues, January 3, 2025

<sup>4</sup> M. Bobrowsky, "Meta Spending to Soar on AI, Massive Data Center," WSJ, Jan. 24, 2025.

<sup>5</sup> S. Nellis and A. Tong, "Behind \$500 billion AI data center plan, US startups jockey with tech giants," Reuters, Jan. 23, 2025.

<sup>6</sup> T. P. Morgan, "Arista Networks Conservatively Awaits Its AI Boom," The Next Platform, Feb. 13, 2024.

<sup>7</sup> "AI, data centers and the coming US power demand surge" Goldman Sachs, April 28, 2024

<sup>8</sup> "Three Mile Island nuclear power plant to return as Microsoft signs 20-year, 835MW AI data center PPA", DataCenter Dynamics, September 20, 2024

<sup>9</sup> "Predictions 2025: Accelerated Demand For AI-Powered Infrastructure And Operations", Forrester, October 22, 2024

## Meanwhile, on the Cloud Front, Growth Continues

Traditional cloud workloads continue to evolve predictably from previous years. Cloud workloads served by regional and edge data centers include SaaS web applications, which are increasingly adopting micro-services architectures. This shift drives greater east-west traffic and demands improved agility and programmability. Security remains a key focus, with emphasis on zero-trust frameworks and micro-segmentation, alongside enhanced observability for both security and troubleshooting.

Latency-sensitive applications continue to shape network requirements in two key areas:

- **Collaboration platforms** drive demand for reduced latency across network fabrics with enhanced end-to-end QoS.
- **Gaming applications** require consistent, low-latency performance throughout the data center segment.

Video content remains the dominant driver of internet and data center traffic, particularly at edge locations hosting CDNs. Two distinct categories show strong growth:

- **Short-form video** traffic continues to surge, boosted by AI-driven personalization that amplifies user engagement through curated content matching. Look no further than TikTok to demonstrate the power of this format, and we expect continued growth throughout 2025.
- **Long-form streaming** shows sustained strength, evolving toward hybrid business models that combine subscription, ad-supported, and pay-per-view formats. These HD and 4K streams consume significant egress bandwidth from edge data centers to viewing devices.

**Social media** platforms are expected to see increased traffic as generative AI tools enhance video clip creation, incentivizing content creators to produce and share more video content.

Cloud workloads will continue expanding as enterprises progress in their **digitization** journeys, including migrating legacy IT workloads to cloud environments. The drivers we've discussed will push global internet traffic growth in 2025 beyond the 17.2% growth seen in 2024<sup>10</sup>. Volume-wise, video remains the dominant component, accounting for 70% of all fixed and mobile data consumption in 2023<sup>11</sup>.

The key takeaway is that while cloud workloads continue their steady evolution, the most significant changes in data center networking are being driven by AI rather than traditional cloud applications.

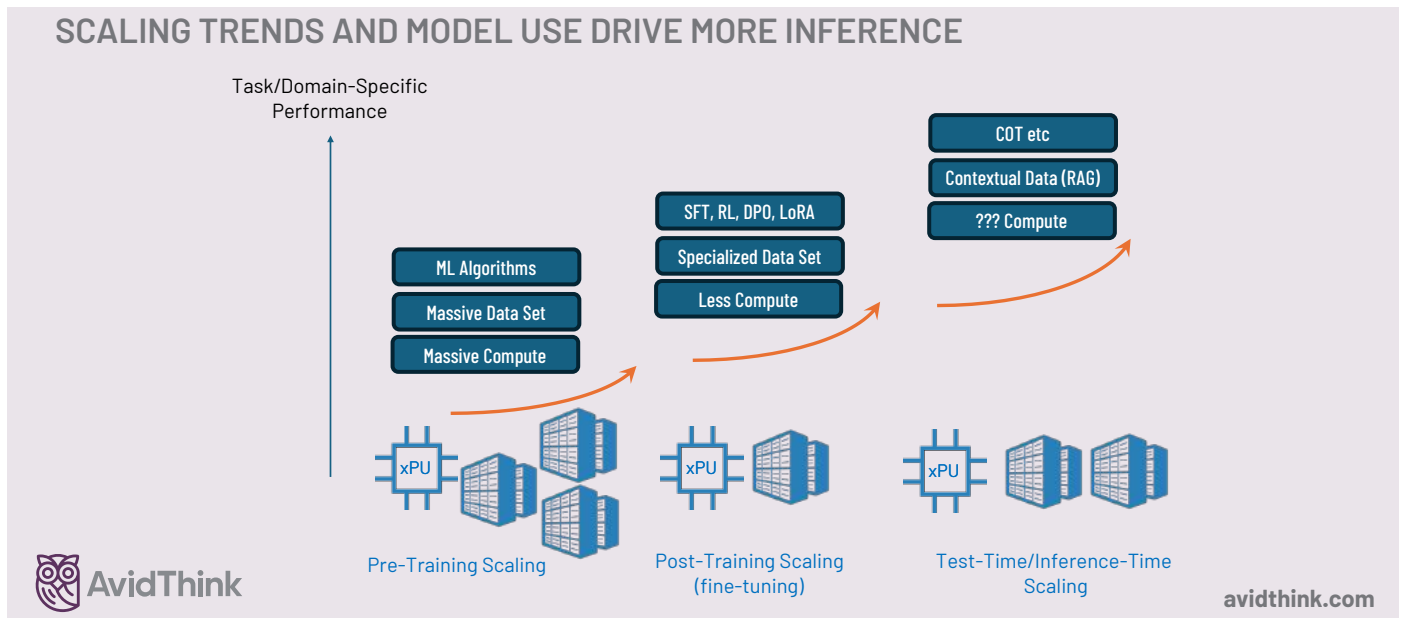
### CLOUD WORKLOADS DRIVING DATA CENTER NETWORKING

- Online Collaboration
- Gaming
- Social Media
- Short-form Video
- Long-form Streaming
- Digital Transformation

**AvidThink** avidthink.com

<sup>10</sup> "Cloudflare 2024 Year in Review," The Cloudflare Blog, Dec. 09, 2024.

<sup>11</sup> Sandvine, "Global Internet Phenomena," Sandvine.com, 2024



## Data Center Architecture Evolution

There are two types of AI workloads that impose extreme requirements on data center networks: training and inferencing. During training, large neural networks continuously adjust the weights of their hundreds of billions (or trillions) of parameters in response to massive training datasets.

As xPUs recalculate billions to trillions of neural network weights, they exchange large quantities of data at high frequencies. These data flows must occur rapidly and reliably. Backend AI networks must support high data speeds (400Gbps to 800Gbps) at nanosecond-level latencies without packet loss. Delays can impact training time, as model training is sensitive to packet loss and tail latencies.

### The Shift from Training to Inference: Infrastructure Implications

While FM training drives AI infrastructure investments, we anticipate a gradual shift toward inference-focused architectures:

- Computing investment is moving from pre-training workloads to post-training fine-tuning, inference, and test-time scaling
- The next wave of agentic AI relies more on chain-of-thought reasoning models leveraging test-time scaling
- Mixed models incorporating reinforcement learning and other methodologies like state space approaches could improve efficiency over existing transformer architecture models
- Sparsification techniques, exemplified by Neural Magic's (acquired by IBM/RedHat in January 2025) open-source vLLM project, enable inference workloads to run on less powerful hardware, including CPUs

Enterprise adoption patterns support this transition:

- Organizations increasingly prefer customizing smaller models with proprietary data over using larger FMs.
- ServiceNow Research demonstrates that RAG-enhanced smaller models can outperform larger alternatives<sup>12</sup>.
- Gartner projects enterprise use of industry-specific GenAI models to grow from 1% in 2023 to 50% by 2027<sup>13</sup>.
- Open-source model adoption is accelerating, with Llama seeing over 400 million downloads in the first three quarters of 2024 – ten times the previous year<sup>14</sup>, meaning less organizations have an incentive to train their own models.

<sup>12</sup> O. Marquez, "Reducing hallucination in structured outputs via Retrieval-Augmented Generation," ServiceNow Research, June 2024.

<sup>13</sup> "3 Bold and Actionable Predictions for the Future of GenAI", Gartner, April 2024

<sup>14</sup> "The enterprise verdict on AI models: Why open source will win" Venture Beat, October 24, 2024

Other recent developments are also reshaping the landscape:

- OpenAI's o1 and o3 models leverage inference-time scaling rather than expanded pre-training.
- NVIDIA's Blackwell GPU announcement in January 2025 promises greater efficiencies with 3-4x performance improvements per watt and dollar.
- The transition opens opportunities for alternative silicon solutions, including Google's TPUs, AWS Trainium, Azure Maia 100, and offerings from Groq and Cerebras.
- While NVIDIA maintains market leadership with their GPUs, they are apparently exploring other AI acceleration approaches<sup>15</sup>

## Distributing Workloads – Distributed, Federated Training and Inferencing

AI model builders and hyperscalers are preparing for scenarios where the largest new FMs will exceed single campus capacity due to power limitations. They are developing plans for clusters of geographically dispersed data center campuses, each near low-cost power sources, interconnected with ultra-high-bandwidth, low-latency fiber links to function as a single massive data center<sup>16</sup>. While most models are currently trained and served from single data centers, AI technology vendors, particularly those in optics like Marvell, report strong customer interest in high-speed optical links for short to mid-distance inter-data center connections.

## What is the New Unit of (AI) Computing?

An architectural evolution is occurring in AI computing, shifting from individual processors and servers to integrated rack-scale systems. Examples include:

- NVIDIA's GB200 NVL72 platform which combines 72 Blackwell GPUs and 36 Grace CPUs in a single rack, connected via an NVLink switch, delivering 1.44 exaFLOPS (at FP4 precision) of AI compute power with 13.5TB of high-bandwidth GPU memory
- Amazon's Trainium2 UltraServer spanning two racks that features 64 Trainium2 chips in a 4x4x4 3D-torus configuration using AWS's proprietary NeuronLink<sup>17</sup>

This move toward vertical integration and rack-level computing units is driven by the need to maximize performance through tightly coupled systems. Traditional component boundaries are dissolving as vendors optimize every aspect of the computing stack. NVIDIA's achieving 30x faster real-time inference for trillion-parameter models demonstrates both their GPU improvements and the benefits of this integrated approach. AMD's acquisition of ZT Systems reinforces this trend<sup>18</sup>, bringing in expertise to help them optimize the entire stack from chips to rack-level infrastructure.

## Data Center Networks – Scale Up to Scale Out to Scale Outside

While previous editions of this report focused on Ethernet networks and SmartNIC/DPU evolution, AI workloads require a broader examination of interconnect technologies. This edition will cover the spectrum from die-to-die communication to scale-out interconnects and data center interconnects (DCI).

## Die-to-die: UCle and Chiplets

The Open Compute Project (OCP) continues to foster an open chiplet ecosystem through its Open Domain-Specific Architecture (ODSA). With intensified competition in GPUs and AI accelerators, and silicon release cycles accelerating to 12 months from 18/24 months, chiplets' time-to-market advantages are becoming more attractive for AI SoCs. The January 2025 Chiplet Summit in Santa Clara showed an uptick in activity around chiplets and UCle interconnect.

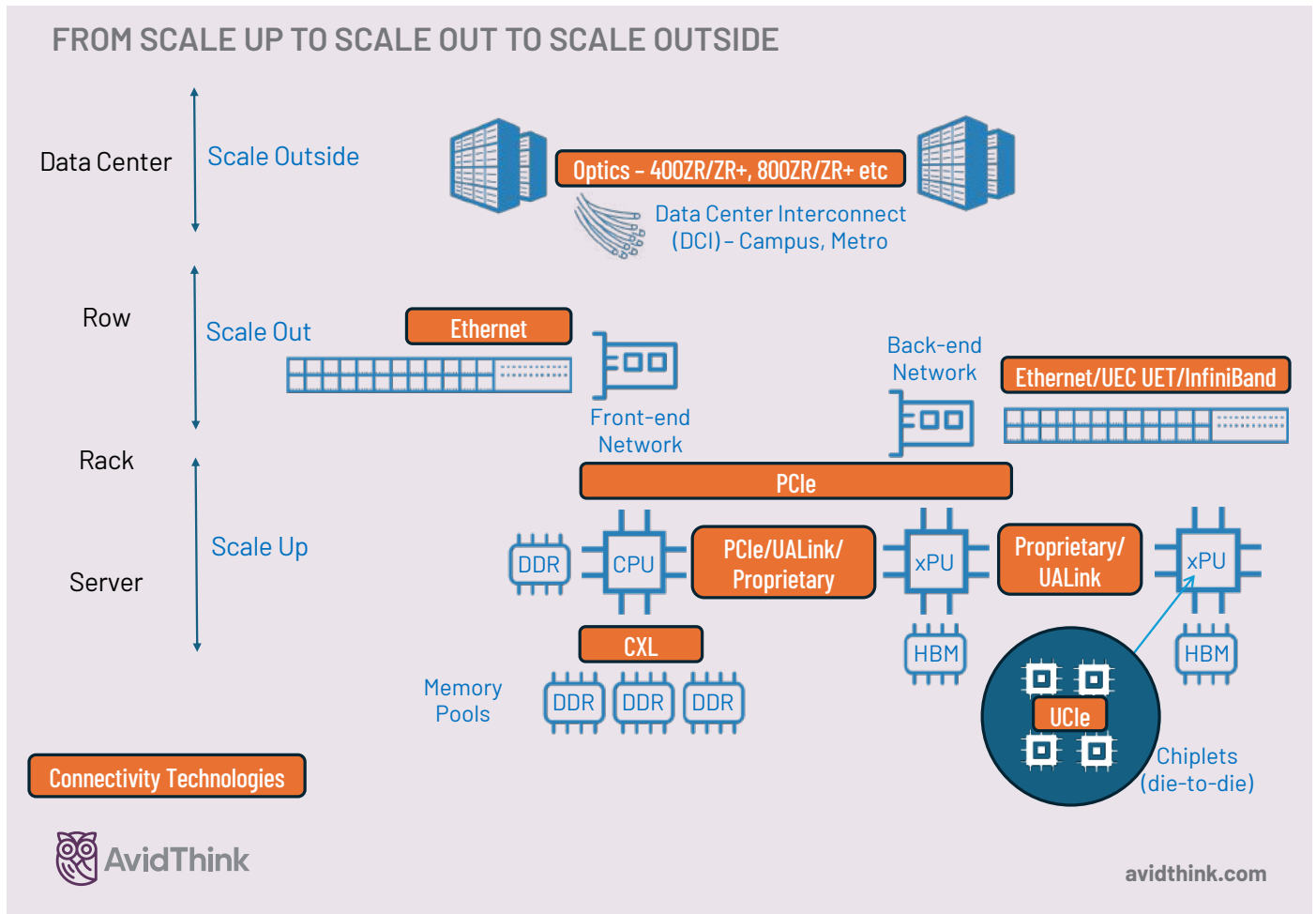
---

<sup>15</sup> "NVIDIA makes a major change as it prepares to enter a hot new market," The Street Jan 15, 2025

<sup>16</sup> "Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure," SemiAnalysis, September 4, 2024

<sup>17</sup> D. Patel et al., "Amazon's AI Self Sufficiency | Trainium2 Architecture & Networking," SemiAnalysis, Dec. 03, 2024.

<sup>18</sup> D. O'Shea, "AMD already leveraging soon-to-be acquired ZT Systems," Fierce Electronics, Dec. 05, 2024.



Universal Chiplot Interconnect Express (UCle) is an open industry standard that defines the interconnect between chiplotlets within a package, enabling a modular approach to system-on-chip design. Unlike CXL, which focuses on processor-to-device connections across a motherboard, UCle specializes in die-to-die communications within a single package. The standard is backed by a consortium of major technology companies, including AMD, Arm, Intel, Qualcomm, Samsung, and TSMC, indicating broad industry support for creating an open chiplot ecosystem.

Key characteristics of UCle that allow mixing and matching different vendor chipsets while maintaining high performance include:

- Support for data rates up to 32 Gbps per pin
- Compatibility with both standard and advanced packaging technologies
- Leveraging of existing PCIe and CXL protocols
- Specialized layers for die-to-die adaptation and physical connectivity

### System Level - PCIe

Standards-based PCIe has made significant progress in addressing the demands of AI systems. PCIe 6.0 now supports 256 GBps for 16 lanes, while PCIe 7.0, which released version 0.7 of its specification in January 2025 for member review, targets an impressive 512 GBps. The primary advantage of PCIe remains its standards-based approach, enabling interoperability across diverse silicon ecosystems.

However, this evolution comes with challenges. As we progress through PCIe generations, the maximum reach of PCIe shrinks. To address this limitation, PCIe retimers from vendors like Marvell have become essential for extending reachability and enabling connectivity between CPUs, GPUs, and I/O devices on servers.

### System Level – CXL

Compute Express Link (CXL) has emerged as a forward-looking open standard for granular connectivity within xPU server clusters and nodes. Built on the PCIe 5.0/6.0 standard, it adds coherence capabilities that allow xPUs to share common memory pools with state synchronization. This technology has become an enabler for efficient AI compute architectures and composable infrastructure.

The CXL architecture is built on three fundamental protocols:

- **CXL.io** serves as the foundation, handling device initialization, discovery, and basic I/O operations.
- **CXL.cache** enables cache-coherent communication between host and device memory through an ultra-low latency request-response mechanism.
- **CXL.mem** allows host processors to directly access device memory through load/store commands, supporting both volatile and persistent memory.

The CXL Consortium has seen strong growth since its inception. Founding members Intel, Microsoft, Google, Dell, HPE, Facebook (Meta), and Huawei are joined by key contributors AMD, NVIDIA, Arm, and Micron. The group numbers 150 members across the industry spectrum. It is finding its way into enterprise environments where workloads require heterogeneous computing hardware and memory disaggregation. Recent milestones include the December 2024 release of the CXL 3.2 specification, which enhanced security and compliance features<sup>19</sup>, and Marvell's introduction of their CXL-based Structera near-memory accelerator in July 2024<sup>20</sup>.

#### Scaling Up: NVLink and UALink

**NVIDIA's CEO Jensen Huang has noted that by the time UALink achieves commercial adoption, NVLink may have progressed to even higher performance levels.**

Previously, PCIe switches and CXL looked promising in facilitating composable computer architectures – shared memory pools, shared storage, reconfigurable architectures. However, the increased focus on large scale AI training clusters has relegated CXL to shared memory access, with vendors betting more on alternatives like UALink for xPU-to-xPU links<sup>21</sup>.

### System Level – NVIDIA NVLink

We briefly mentioned NVLink earlier when introducing NVIDIA's flagship GB200 NVL72 system. The GPUs within each node are interconnected using NVIDIA's proprietary ultra-high-speed, low-latency NVLink technology, which provides exceptional bandwidth and latency characteristics.

2024 marked significant advancements in NVLink capabilities. The 5th-generation NVLink doubled throughput from its predecessor's 900GBps to 1.8 TBps (per GPU). NVIDIA also introduced NVLink Switch, a breakthrough architecture that enables

bidirectional full-speed connectivity between every pair of GPU across 576 GPUs.

### System Level – UALink

UALink (Ultra Accelerator Link) represents a significant industry push toward open standards in high-speed, low-latency die-to-die interconnects for AI and HPC accelerators. Formed in May 2024, the consortium has attracted major industry players

---

<sup>19</sup> CXL 3.2 Specifications Announcement

<sup>20</sup> "Marvell Introduces Breakthrough Structera CXL Product Line to Address Server Memory Bandwidth and Capacity Challenges in Cloud Data Centers," Marvell News, July 30, 2024

<sup>21</sup> K. Heyman, "CXL Thriving As Memory Link," Semiconductor Engineering, Sep. 16, 2024.

	UALink	NVLink
<b>Developer</b>	UALink Consortium	NVIDIA
<b>Specification</b>	Open standard	Proprietary
<b>Scalability</b>	Up to 1,024 xPUs	Up to 576 GPUs (NVLink 5.0)
<b>Memory Semantics</b>	Supported	Supported
<b>Maturity</b>	Specification release in Q1 2025	Mature technology with multiple generations
<b>Adoption</b>	Early stages, expected to grow	Widely adopted in NVIDIA's ecosystem

**Table 1: Comparing UALink and NVLink**

including AMD, Intel, Google, Microsoft, Meta, HPE, Cisco, and Broadcom<sup>22</sup>. The initiative gained further momentum with Marvell and Qualcomm joining as contributor members, and recent board expansion to include Alibaba Cloud, Apple, and Synopsys<sup>23</sup>. Apple's involvement may signal its development of in-house data center processors for AI.

To accelerate development, AMD has contributed their Infinity Fabric shared memory protocol and GPU-to-GPU interface xGMI to the UALink effort, with consortium members agreeing to use Infinity Fabric as the standard protocol for accelerator interconnects<sup>24</sup>.

UALink offers several key advantages architecturally, including:

- Scalability to support up to 1,024 accelerators in a single AI pod
- Competitive high bandwidth and low latency performance
- 40% improvement in energy efficiency
- Support for both AI training and inference solutions

UALink 1.0 specification is expected to be publicly available in Q1 2025, and multiple semiconductor companies are rumored to be developing Ultra Accelerator Link switches. Meanwhile, Synopsys announced the first UALink IP solution, offering 200 Gbps throughput per lane, and linking up to 1,024 accelerators. The solution is scheduled for 2H 2025 availability.<sup>25</sup>

While UALink presents a promising open approach to xPU-to-xPU connectivity with compelling bandwidth per lane performance and scale, NVIDIA maintains several key advantages through NVLink's maturity and deployment experience (see table 1). NVIDIA's CEO Jensen Huang has noted that by the time UALink achieves commercial adoption, NVLink may have progressed to even higher performance levels<sup>26</sup>.

### Scale-Out Networking Preamble

Many AI networks today implement a split architecture with distinct frontend and backend networks. The frontend network connects the xPU cluster to external systems like applications and storage using straightforward 100/200Gbps Ethernet in 2/3-tier standard Clos topologies. Backend networks, operating at higher speeds (400/800Gbps), support the intensive data transfer required during AI training or computation tasks.

Backend networks, whether InfiniBand or Ethernet, rely on RDMA (Remote Direct Memory Access) protocol for performance

<sup>22</sup> UALink Consortium poised to compete with Nvidia's NVLink – AMD and Intel-led group opens doors to contributor members | Tom's Hardware

<sup>23</sup> Alibaba, Apple, and Synopsys become latest members of UALink Consortium - DCD

<sup>24</sup> T. P. Morgan, "Key Hyperscalers And Chip Makers Gang Up On Nvidia's NVSwitch Interconnect," The Next Platform, May 30, 2024.

<sup>25</sup> UALink IP Solution | Synopsys

<sup>26</sup> S. Sharwood, "Nvidia CEO brushes off Big Tech's attacks on NVLink network tech," Theregister.com, Jun. 04, 2024

optimization. RDMA enables GPU nodes to read from and write to each other's memory without CPU involvement<sup>27</sup>. This direct memory access is crucial for AI workloads as it reduces latency and CPU overhead. While originally developed for InfiniBand networks, RoCE (RDMA over Converged Ethernet) has gained widespread support across major vendor offerings.

Unsurprisingly, the industry anticipates significant growth in backend networks with analyst firm 650 Group projecting RDMA-related revenue to grow from USD 6.9B in 2023 to USD 22.5B in 2028<sup>28</sup> and another analyst firm, Dell'Oro, predicting that data center switches for AI backend networks will drive nearly USD 80B in spending over the next five years<sup>29</sup>

While InfiniBand has traditionally dominated high-performance computing (HPC) workloads requiring demanding bandwidth and latency specifications, Ethernet with RoCEv2 is gaining momentum due to:

- Lower cost (estimated 40-50% cheaper than InfiniBand)
- Greater familiarity among network engineers
- Opportunity for converged and streamlined operations
- Extensive ecosystem of tools and expertise

The increasing viability of Ethernet for AI workloads is demonstrated by major deployments:

- Meta, which has InfiniBand AI clusters, used an Ethernet-switched cluster for training their open-weights Llama model<sup>30</sup>
- xAI's Colossus supercomputer employed NVIDIA's Spectrum-X Ethernet-based fabric across 100K GPUs<sup>31</sup>

These examples, along with the Ultra Ethernet Consortium's efforts, could rapidly close the gap between Ethernet and InfiniBand.

### Scale-Out Considerations for Backend Networks

Data center operators have identified several critical lessons for scaling AI clusters. The following lessons were gathered from Meta's networking@scale conference, IEEE Hot Interconnects 2024, OCP Global Summit 2024, and our discussions with industry leaders:

#### Smart Parallelism Implementation

- Training large models like Llama 3 requires sharding across many GPUs due to memory constraints. This introduces significant communication overhead between GPUs, making efficient parallelism crucial.
- Key techniques include:
  - **Data Parallelism:** Duplicates the entire model on each GPU and splits mini batches. While this requires expensive all-reduce operations for gradients, it's the most straightforward to implement.
  - **Pipeline Parallelism:** Distributes model layers across GPUs, requiring communication between layers. This reduces memory requirements but needs careful balancing of pipeline stages.
  - **Tensor Parallelism:** Splits matrix operations across GPUs using all-reduces. This approach minimizes memory usage but requires high-bandwidth, low-latency connections.

While InfiniBand has traditionally dominated high-performance computing (HPC) workloads requiring demanding bandwidth and latency specifications, Ethernet with RoCEv2 is gaining momentum for AI workloads. UEC efforts can help further close the gap.

<sup>27</sup> The Evolution of Data Center Networking for AI Workloads | Kentik Blog

<sup>28</sup> 650 Group RDMA Networking and AI Research Report, June 2024

<sup>29</sup> Dell'Oro Press Release "AI Backend Data Center Switch Spending over the Next Five Years", July 16, 2024

<sup>30</sup> Adithya Gangidi et al., "RDMA over Ethernet for Distributed Training at Meta Scale," pp. 57-70, Jul. 2024

<sup>31</sup> "NVIDIA Ethernet Networking Accelerates World's Largest AI Supercomputer, Built by xAI," NVIDIA Newsroom, 2024.

- **Multi-Dimensional Parallelism:** Combining these techniques becomes necessary when scaling to 10s of thousands of GPUs. For example, Meta's Llama 3 training used all three approaches, creating complex communication patterns with many smaller concurrent communications.

### Network Optimization Strategies

- **Network-aware parallelism:** Places latency-sensitive techniques like Tensor Parallelism within the same rack where high bandwidth is available, while communication-tolerant techniques like Fully Shared Data Parallelism (FSDP) can span different zones.
- **Topology-aware scheduling:** Physical GPU location can impact communication speed. Scheduling systems must understand network topology (rack, row, AI zone) to minimize communication overhead by placing frequently communicating ranks closer together.
- **Advanced scheduling considerations:** Must balance multiple objectives including data locality, GPU quotas, fault tolerance, and scheduling overhead. This requires sophisticated algorithms that handle both hard and soft constraints.
- **Control message prioritization:** Critical control messages like CTS and ACK receive higher priority in network queues to prevent delays in completion signals and data transfer operations.
- **Spine switch optimization:** Tuning Virtual Output Queuing (VOQ) of spine switches reduces their latency impact on data forwarding.
- **Switch architecture considerations:** Higher radix switches with shared buffer outputs can address potential bottlenecks in high-density deployments.
- **Enhanced telemetry:** Communication libraries must handle multiple concurrent collectives that may be unaware of each other, requiring sophisticated monitoring and routing optimization.

### Physical Infrastructure Considerations

- **Power density implications:** The increase from traditional 12-15KW per rack to 120KW+ for AI workloads creates new thermal and spacing challenges.
- **Facility adaptation:** Data centers without liquid cooling capabilities need increased rack spacing, which directly impacts cable type selection and length requirements.
- **Cable selection strategy:** Use passive copper when possible for cost and simplicity, use active cables as a fallback option, and use optical solutions when distance and bandwidth requirements demand it. Topics like the selection of optical modules, Active Optical Cables (AOC), or Direct Attach Cables (DAC), including ACC, AEC, or passive DAC, are important but beyond this report.

### Scale-Out – InfiniBand

NVIDIA, through its Mellanox acquisition, remains essentially the main provider of datacenter scale InfiniBand solutions. Having proven itself in HPC environments, InfiniBand has established itself as a high-performance fabric for AI training.

The InfiniBand standard made significant advances in 2024. The InfiniBand Trade Association's (IBTA) September release of Volume 1, Specification 1.8 substantially enhanced RDMA capabilities. The specification introduced XDR (Extended Data Rates), pushing data speeds to ~200Gbps per lane while enhancing transmission reliability through XDR FEC (Forward Error Correction) support. It also expanded support for next-generation interfaces, including 4 Lane QSFP 800 Gbps and 8 Lane QSFP-DD and OSFP 1,600 Gbps. Additionally, the release strengthened security features for RDMA fabrics in data-intensive environments, improved congestion management, and enabled switches with up to 256 ports<sup>32</sup>, facilitating higher radix switches.

---

<sup>32</sup> "InfiniBand 1.8 Spec Brings 200Gbps Per Lane Speed, Scalability for RDMA Fabrics", Converge Network Digest, September 10, 2024

## Scale-Out – Ethernet and RoCE

The evolution of Ethernet for AI workloads has seen significant progress. IBTA improved RoCEv2 interoperability with InfiniBand, while the OpenFabrics Alliance enhanced its RoCE support to further reduce latency and boost data speeds. Combined with Linux kernel enhancements, these improvements have substantially improved RoCEv2 performance on Linux systems.

Most Ethernet fabrics used for AI training support RoCEv2 and implement additional scheduling and load-balancing capabilities. A key differentiator among vendors is their approach to intelligent congestion control for packet loss prevention and latency (including tail latency) reduction. Different vendors have developed strategies to improve scaling-out architectures<sup>33</sup>:

- **Endpoint and notification-based methods** focus on mitigating congestion impact after occurrence. These systems use Priority Flow Control (PFC) where receiving nodes message originating nodes to slow incoming data flows when certain queue-depth thresholds are reached.
- **Multipath-based approaches** take a preventive stance through ECMP (Equal Cost Multi-Path), identifying available paths to destinations with identical routing metrics and using hashing mechanisms for data flow load-balancing. While a DPU or SmartNIC can enforce these, they still face challenges with in-cast patterns common in AI training workloads.
- **Scheduling-based solutions** represent a more comprehensive approach, preventing congestion through end-to-end flow scheduling from input to output ports. These systems deliver deterministic latency and throughput while eliminating packet loss and reducing jitter. Some vendors have enhanced this further with packet spraying, which distributes traffic at the packet level rather than flow level, providing more granular load balancing than traditional flow-based approaches.
- **Virtual chassis approaches** that treat multiple switches as parts of a single logical chassis can help with coordination across the fabric. Networking vendors, including Arista, Arccus, and DriveNets, have implemented architectures where switches are interconnected to function as a single logical switch or router. These solutions employ centralized control planes to ensure consistent routing, scheduling, and management across all network nodes. They're designed to scale elastically and use advanced scheduling and load-balancing techniques to prevent congestion and optimize utilization.

There are other techniques and topologies for backend xPU networks. For example, researchers at MIT and Meta have proposed a Rail-Only Network that eliminates spine switches, leveraging high-bandwidth interconnects within nodes. Their approach is based on the observation that FM training traffic is sparse and remains within "rails" (GPUs with the same rank across nodes). When non-sparse traffic occurs, such as in Mixture of Experts models where expert parallelism requires each expert to communicate with the rest of the model, any cross-rail traffic is forwarded within nodes using high bandwidth intra-node interconnects like NVLink. This innovative architecture demonstrates impressive efficiency gains, with 38% to 77% cost savings and 37% to 75% power savings compared to existing state-of-the-art solutions<sup>34</sup>.

## Scale Out – UEC and UET

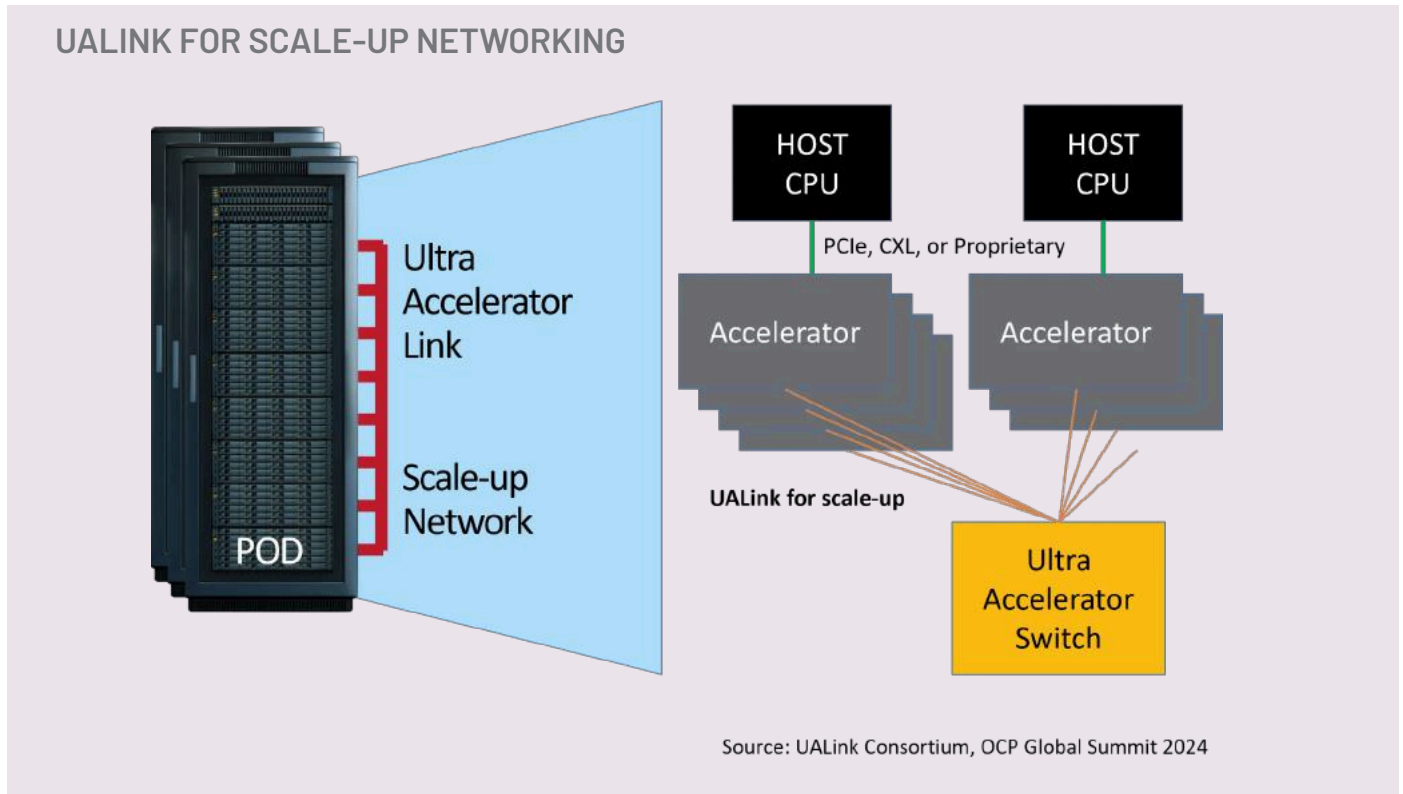
The Ultra Ethernet Consortium (UEC) represents a significant industry initiative to create an open Ethernet-based transport protocol optimized for AI and HPC workloads. The consortium has grown rapidly since its formation, now encompassing several hundred members, including major industry players across multiple sectors:

- Semiconductor companies: NVIDIA, AMD, Intel, Broadcom
- Network equipment providers: Cisco, Arista, Juniper, Nokia
- System manufacturers: Dell, Huawei, Lenovo, HPE
- Hyperscalers: Meta, Microsoft, Alibaba, Baidu, Tencent

The UEC announced the development of the Ultra Ethernet Transport (UET) protocol, designed to eventually supersede RoCE as the open Ethernet transport protocol for AI and HPC workloads.

<sup>33</sup> Optimize data center networking for AI workloads, Nokia, September 2024

<sup>34</sup> 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/HOT163208.2024.00013.



In its 2023 white paper, the UEC announced the development of the Ultra Ethernet Transport (UET) protocol, designed to eventually supersede RoCE as the open Ethernet transport protocol for AI and HPC workloads. UET's design objectives are comprehensive, aiming to solve current limitations in existing protocols while preparing for future scaling challenges.

Key technical objectives of UET include:

- Scaling capability to 1 million connected endpoints
- Achievement of up to 1.6 Tbps data speeds
- Embedded security at the transport layer
- Minimized connection establishment time
- Reduced connection state overhead

The consortium made substantial progress throughout 2024, and as of January 2025, is finalizing the 1.0 specification. Leading member vendors are preparing to release UET-ready network cards and switches concurrent with the specification release. The protocol introduces several innovative features that set it apart from existing solutions<sup>35</sup>:

- Multi-pathing capabilities for improved reliability and performance
- Advanced packet spraying techniques for optimal resource utilization
- Flexible delivery ordering that eliminates the need for packet reordering before delivery
- Real-time automated congestion controls driven by telemetry data
- Built-in security features that specify authentication, authorization, and confidentiality capabilities without compromising performance

<sup>35</sup> "Ultra-Ethernet Specification Update," Ultra-Ethernet Consortium, August 29, 2024

## Scale-Out/Scale Outside – Frontend Ethernet Networks

Frontend networks employ traditional 2/3-tier Clos Ethernet architectures, but the rise of inference workloads is driving new requirements. The increased north-south bound traffic to AI clusters demands higher data speeds of 100–400Gbps with more stringent end-to-end QoS requirements as multi-modal inferencing catches on (voice response, visual models supporting augmented reality).

Security has become a critical concern for these networks. Modern frontend architectures incorporate multiple layers of protection:

- Encryption of data-in-motion
- Least privilege/zero-trust frameworks
- Role-based network access control
- Intelligent firewalling capabilities

IPv6 Segment Routing (SRv6) has emerged as a key technology for enhancing frontend network performance. SRv6 embeds segment information supporting QoS prioritization and granular traffic steering directly into IPv6 packet headers, offering greater flexibility compared to MPLS. This approach enables rich orchestration capabilities while reducing control plane overhead. Major network vendors, including Cisco, Juniper, Arista, Nokia, and Arrcus, have embraced SRv6, with vendors like Arrcus leveraging it for end-to-end QoS across multiple network domains.

**IPv6 Segment Routing (SRv6) has emerged as a key technology for enhancing frontend network performance. SRv6 embeds segment information supporting QoS prioritization and granular traffic steering directly into IPv6 packet headers, offering greater flexibility compared to MPLS.**

## Scale Outside – Data Center Interconnects (DCIs)

Meta revealed at their Networking@Scale 2024 conference that the impact of AI model training on backbone networks has exceeded initial projections. Major carriers like Lumen and Zayo corroborate this trend in their investor presentations, highlighting the growing importance of robust inter-data center connectivity.

The optical interconnect technology landscape is evolving rapidly to meet these demands. The industry has seen widespread adoption of 400ZR/ZR+ modules, with distinct characteristics for different use cases:

- Standard ZR modules optimize for 400Gbps over distances up to 120km
- ZR+ variants support flexible modulation and extended reach up to 400km through OpenZR+ and OpenROADM standards

The next frontier in DCI technology is 800ZR/ZR+, marked by the OIF's release of the 800ZR Implementation Agreement in October 2024. This new standard brings significant advancements:

- Enables 800Gbps transmission over 520km using 16QAM modulation
- ZR+ variants extend reach beyond 1,000km
- Achieves 30% power per bit reduction compared to 400ZR
- Recent multi-vendor interoperability demonstrations have validated these capabilities with solutions from providers like Marvell, Lumentum, and Coherent

These developments are crucial for AI workloads, with 800ZR+ shipments expected to see rapid growth. The technology's improved energy efficiency is valuable for power-constrained data centers supporting AI infrastructure.

## Other Data Center Considerations

Beyond networking and interconnect technologies, several other factors shape modern data center architecture. Security and isolation through SmartNICs/DPUs play an increasingly crucial role, both in fabric management and security enforcement.

### Security and the DPU

AI frontend and cloud datacenter networks share common security requirements, particularly in network segmentation. This includes:

- Robust tenant isolation
- Prevention of unauthorized lateral data movement
- Enforcement of application-based network policies

DPUs and SmartNICs serve dual purposes in modern architecture. Beyond offloading RDMA functions (segmentation, reassembly, and advanced congestion control), they provide sophisticated security capabilities:

- Hardware-accelerated encryption/decryption
- Secure key management
- Advanced firewalling
- Granular micro-segmentation

AWS's Nitro architecture has demonstrated the value of this approach, serving as a foundational secure platform for cloud services through its bare metal hypervisor implementation. Following this model, vendors like NVIDIA (with BlueField DPUs) and AMD (with Pensando DPUs) are enhancing their offerings for enterprise data centers. Google has taken a similar approach with their Titanium custom silicon security microcontrollers, which perform comparable security and offload functions.

NVIDIA has been aggressive in building partnerships to showcase BlueField DPU capabilities. Notable collaborations include:

- Arrcus: Running ArcOS on BlueField-3 DPUs for network function offload and SRv6 support across 5G networks<sup>36</sup>
- VMware/Broadcom: Supporting DPUs within their hybrid cloud architecture for AI and data-intensive applications
- Red Hat: Integrating BlueField DPUs and the NVIDIA Morpheus AI framework with Red Hat Enterprise Linux and OpenShift
- Security vendors: Fortinet and Guardicore leveraging BlueField-3 for enterprise network security
- Edge computing: Cloudflare, F5, and Juniper Networks adopting BlueField for edge platforms

### Open Source and SONiC

SONiC (Software for Open Networking in the Cloud) continues to gain traction in data center networking, expanding beyond its hyperscale origins to become viable for enterprise deployments. Major vendors are now positioning SONiC for AI workloads, with several notable implementations:

- Arista, an early SONiC contributor, enables the platform on their 7050X/7060X series switches, providing customers with flexibility in network operating system choice.
- Cisco has partnered with startup Aviz Networks to provide enterprise-grade SONiC support on its 8000 Series routers. The routers leverage Silicon One ASICs for high-performance 400G networks.
- Juniper offers SONiC as an option on their QFX and PTX platforms, integrating it with their Apstra automation platform for enhanced management capabilities.

---

<sup>36</sup> "Arrcus Delivers Next Generation of High-Performance, Zero-Trust Networking for Datacenters Enabled by NVIDIA BlueField," Arrcus, September 2023

- Nokia has adopted a dual strategy, offering both community SONiC and its proprietary SR Linux. It notably supports Microsoft Azure's transition from 100G to 400G infrastructure.
- NVIDIA has emerged as a key SONiC supporter, enabling it on their Spectrum open Ethernet switches and Spectrum-X networking platform for AI. As a SONiC governing board member and major contributor, NVIDIA is helping shape the platform's future.

Silicon vendors Broadcom and Marvell both support and contribute to SONiC, with Broadcom shipping and supporting an Enterprise SONiC release.

This growing adoption reflects a broader industry shift toward open networking solutions that reduce vendor lock-in while maintaining enterprise-grade reliability. This trend is relevant for AI-focused data centers, where flexibility, scalability, and cost efficiency are paramount.

**The growing adoption of SONiC reflects a broader industry shift toward open networking solutions that reduce vendor lock-in while maintaining enterprise-grade reliability.**

### **CPO, LRO, and LPO: Transforming Data Center Interconnects**

The growing demands of AI workloads for bandwidth, coupled with concerns about power consumption and cost, have driven innovation in optical interconnect technologies. Three approaches have emerged, each offering unique solutions to these challenges.

- Co-packaged optics (CPO) integrates the optical engine directly with the switching silicon on the same substrate. This tight integration reduces signal losses through shorter electrical traces while enabling lower power consumption through lower-power SerDes. The approach shows promise for AI data centers, where enhanced system efficiency and improved scalability are crucial. However, CPO faces several adoption challenges, including questions about technical maturity and complex business model considerations.
- Linear Receive Optics (LRO) takes a more evolutionary approach, maintaining the DSP on the transmit path while eliminating it from the receive path. This hybrid strategy combines analog and digital processing to reduce power consumption while maintaining interoperability with industry standards. LRO could serve as an important transitional technology between traditional optical modules and more advanced solutions, offering a practical balance between performance and power efficiency.
- Linear Drive Pluggable Optics (LPO) pursues simplification by removing traditional DSP and CDR (clock and data recovery) chips. This direct drive approach reduces both power consumption and latency while maintaining hot-swap capability – a crucial feature for data center operations. The technology has proven effective for short-range applications within data centers, offering an attractive combination of performance and cost benefits.

The industry is making significant advances in complementary technologies. Silicon Photonics (SiPh) could be an enabler for both LPO and CPO implementations (in addition to other interconnects), offering a path to compact, low-power modules that can scale to 1.6Tbps. Meanwhile, thin-film lithium niobate technology shows promise for enhancing modulation efficiency in hybrid LPO solutions. While multiple vendors have demonstrated these technologies in early designs, commercial adoption patterns are still evolving.

## Vendor Landscape

The data center networking ecosystem spans merchant silicon providers like Broadcom and Marvell to switching vendors and xPU providers. In this section, we'll start by covering scale-out networking vendors and recent hyperscaler AI datacenter networking developments. We've selected both established incumbents and innovative challengers in the space, highlighting key developments rather than providing exhaustive coverage.

### Select Networking Vendors

#### Arista



Arista's **AI networking portfolio** centers on their 7700R4 Distributed Etherlink Switch (DES) architecture, which extends beyond traditional chassis limitations while maintaining deterministic performance. The system supports over 27,000 800GbE ports or 31,000 400GbE ports through a distributed scheduler design that creates a single logical switching domain. This architecture combines four key technologies to achieve its performance:

- Cell-based traffic spraying for uniform load distribution
- Virtual output queues (VOQ) to prevent head-of-line blocking
- Distributed credit scheduling to eliminate noisy neighbor effects
- Deep buffering to handle micro-bursts

Arista's **CloudVision** platform and AI Analyzer provide microsecond-level visibility and management capabilities. Their EOS (Extensible Operating System) enables consistent operation, with features like Smart System Upgrade for seamless software updates during long-running AI training jobs. The platform supports LPOs to reduce power consumption compared to traditional DSP-based solutions.

#### Arrcus



As an upstart in the world of networking giants, Arrcus looks to differentiate itself by providing an end-to-end network fabric (**ACE-AI**) that can meet the multiple needs of AI workloads, from data management to training to inferencing. Arrcus believes that AI will require distributed and federated learning to continue to scale, and that inference will be inherently a distributed and more edge-centric operation. As a software-based offering, Arrcus brings its ArcOS NOS, ACE (Arrcus Connected Edge) Platform, and FlexMCN (multi-cloud networking) to bear in their multi-domain offering.

At the data center level, Arrcus offers a choice of a Virtualized Distributed Fabric or a standard IP Clos leaf/spine offering with high radix switches, 800Gbps support, and ROCEv2. In Arrcus's solution for IP Clos networks, the ArcOS network operating system leverages PFC, ECMP and adaptive load balancing to distribute traffic and minimize congestion. In the Distributed Fabric for massive GPU clusters, ArcOS controls distributed white-box switches in a virtual chassis architecture. This also provides switch-based scheduling for load balancing and minimizing congestion.

Arrcus also leverages SRv6 extensively to optimize network end to end, from data center network through the WAN to edge/access networks including 5G RAN. Supporting a range of hardware platforms, from multiple Broadcom silicon families (Jericho, Tomahawk, Trident) to NVIDIA switches (NVIDIA Spectrum) and DPUs (NVIDIA BlueField 3), Arrcus intends to provide customers with a flexible unified approach for distributed AI needs.

#### Cisco



Cisco addresses **AI networking demands** through a three-pronged strategy targeting different deployment scales. Their Nexus HyperFabric AI, built around their 6000 Series switches, provides a cloud-managed approach with automated design and deployment capabilities. The system supports speeds from 10Gbps to 400Gbps, built on EVPN-VXLAN underlays, with Nexus Dashboard providing comprehensive telemetry and automation.

The Nexus 9000 series, particularly the 9364E-SG2 switch, forms the foundation for mainstream AI deployments, supporting speeds from 100Gbps to 800Gbps. The platform implements essential AI/ML features including dynamic load balancing, priority flow control (PFC), and Data Center Quantized Congestion Notification (DCQCN).

For hyperscale deployments, Cisco's 8122-64EH router leverages their Silicon One G200 processor, a 5nm 51.2T chip enabling 64 ports of 800G connectivity in a 2RU form factor. Advanced features include improved flow control, congestion awareness, hardware-based link-failure recovery, and sophisticated packet-spraying functionality.

### DriveNets



**DriveNets** is another upstart in the networking world that leverages disaggregation and white-box networking hardware. DriveNets **Network Cloud-AI solution** utilizes a wide set of capabilities including leveraging cell-switching abilities in Broadcom's merchant silicon. The DriveNets Network Operating System (DNOS) orchestrates traffic scheduling, optimizes link usage through packet spraying, and manages disaggregated white box hardware as a virtual chassis. Their solution uses an approach where large packets are broken down into smaller fixed-sized "cells," which are then load-balanced across switching elements. The packet order is restored at the destination. This pre-scheduling of flows combines with their end-to-end virtual output queue (VOQ) traffic management to enable predictable and low latencies. In summary, the DriveNets solution "sprays" small, uniform cells across multiple paths within a virtual chassis. This allows for improved utilization and faster job completion time for AI training.

### Juniper



**Juniper addresses AI networking** through a standards-based Ethernet approach, leveraging their PTX and QFX switches for leaf-spine connectivity. The architecture delivers up to 460.8Tbps of throughput with 576 ports of 800GbE in their largest configuration, enabling over 18,000 GPUs to connect in a two-tier Clos fabric. Central to their strategy is the use of both custom silicon (Juniper Express 5) and merchant silicon (Broadcom Tomahawk), providing customers with architectural flexibility.

The platform implements congestion management through a combination of Explicit Congestion Notification (ECN), Priority-Based Flow Control (PFC), and Data Center Quantized Congestion Notification (DCQCN) to ensure lossless transmission.

For operational simplicity, Juniper touts its **Apstra** intent-based networking software. Apstra automates the fabric lifecycle from Day 0 through Day 2+, with particular optimization for GPU-intensive workloads through its rail-optimized design capabilities. Apstra provides the ability to manage backend, frontend, and storage fabrics from a single pane of glass while allowing customers a choice of network hardware vendors (including Juniper). Telemetry from routers and switches is used to automatically calculate and configure optimal parameter settings for congestion control in the fabric using closed-loop automation capability in Juniper Apstra to deliver peak AI workload performance.

Juniper's Ops4AI Lab provides advanced GPU computing, storage, and networking infrastructure from vendors, including NVIDIA, AMD, Weka, and Vast Data, for qualified customers and partners to test AI workloads with expert support.

### Nokia



Nokia is leveraging its IP routing and optical transport equipment for data center backbone buildouts (bolstered by **Nokia's announcement to acquire Infinera**), **data center switching solutions** (with recently added support for open-source SONiC), and Event-Driven Automation (EDA) within the data center. Their portfolio of Nokia 7215 IXS, 7220, and 7250 IXR series data center platforms powered by their SR Linux NOS and complemented by their 7750 service router (data center gateway) provide a scalable solution that offers fixed and modular configurations for deployment flexibility.

Nokia recently secured several significant data center wins, positioning itself as a key player in the AI-driven data center market. In September 2024, Nokia **announced a commercial agreement with CoreWeave**, a leading AI infrastructure provider. The company followed that in November with a **five-year expansion of its agreement** to supply Microsoft Azure with data center routers and switches. The Azure deal will increase Nokia's global footprint to over 30 countries and support Microsoft's migration from 100Gbps to 400Gbps. In December 2024, it secured an agreement to provide IP networking equipment for Nscale's new data center in Stavanger, Norway, focused on GPU-as-a-service for AI workloads. The momentum for Nokia puts it in a good starting position in 2025 to grow in AI-related data center buildouts<sup>37</sup>.

---

<sup>37</sup> "Nokia suddenly has a growth story again – and it's all about AI," Lightreading.com, 2025.

## NVIDIA



NVIDIA's **InfiniBand platform** centers on the Quantum-X800 line, purpose-built for HPC and "trillion parameter scale" AI workloads requiring 800Gbps throughput. The flagship Q3400-RA switch provides 144 ports of 800Gbps across 72 OSFP cages, enabling two-tier fat-tree topologies supporting over 10,000 endpoints. The platform's fourth-generation SHARP (Scalable Hierarchical Aggregation and Reduction Protocol) technology accelerates workloads by offloading compute operations to the network. This latest version adds FP8 precision support and new collective operations like ReduceScatter and ScatterGather. For smaller deployments, the Q3200 offers two independent 36-port switches in 2RU. Both switches integrate with NVIDIA's Unified Fabric Manager software and feature dedicated InfiniBand management ports. NVIDIA combines the Quantum switches with their ConnectX-8 SuperNICs and LinkX interconnects for a complete AI networking solution.

In Ethernet environments, **NVIDIA's Spectrum-X platform** combines Spectrum-4 switches with BlueField-3 SuperNICs to optimize AI workloads. The SN5600 switch delivers 64 ports of 800G with 51.2Tb/s throughput in 2RU, implementing RoCE extensions for adaptive routing and telemetry-based congestion control using deep learning models. When paired with BlueField-3 SuperNICs, the system provides microsecond-level visibility from GPU to network while managing out-of-order packets. NVLink's broader SN5000 series leverages a shared 160MB packet buffer architecture for predictable latency across ports. Their Spectrum-4 switches provide multiple NOS options, including Cumulus Linux, SONiC, and standard Linux distributions.

NVIDIA demonstrated their technology's scalability with xAI's Colossus, deploying 100,000 H100 GPUs interconnected via Spectrum-X Ethernet networking platform. The system achieved 95% data throughput compared to typical 60% throughput with standard Ethernet<sup>38</sup>.

## Hyperscalers

### Amazon Web Services



Amazon Web Services (AWS) showcased their **10p10u network fabric** at re:Invent 2024<sup>39</sup>, designed specifically for AI workloads. The architecture delivers ten petabytes of network capacity with sub-ten-microsecond latency across its data centers. The system incorporates several novel hardware solutions:

- A proprietary trunk connector that bundles 16 fiber optic cables
- Firefly Optical Plug system for pre-deployment testing, resulting in a 54% reduction in installation time
- Scalable Intent-Driven Routing (SIDR) protocol, which combines centralized planning with decentralized execution to achieve sub-second failure response times
- NeuronLink, a proprietary scale-up interconnect technology for Trainium2 servers provides low latency chip-to-chip communication at 1 Tbps in a 2D torus

The 10p10u infrastructure is built on 800 Gbps Ethernet technology and tightly integrated with AWS's UltraServer compute technology and Trainium2 AI chips. This integration is enhanced by their custom Elastic Fabric Adapter (EFA), which employs the Scalable Reliable Datagram protocol for efficient data transfer, enabling the network to scale from individual racks to multi-campus clusters while maintaining its performance characteristics.

<sup>38</sup> "NVIDIA Ethernet Networking Accelerates World's Largest AI Supercomputer, Built by xAI," NVIDIA Newsroom, 2024

<sup>39</sup> S. M. Kerner, "AWS upgrades its 10p10u network to handle massive AI clusters," Network World, Dec. 04, 2024.

## Google Cloud



Google Cloud has enhanced its networking infrastructure through several recent developments<sup>40</sup>:

- Introduction of the **Titanium ML network adapter**, which leverages NVIDIA ConnectX-7 hardware and a 4-way rail-aligned network architecture to power their A3 Ultra VMs and deliver non-blocking 3.2 Tbps of GPU-to-GPU traffic with RoCE
- Preview of their sixth-generation TPU, Trillium, which delivers 4x training and 3x inference throughput over current generation TPU v5e, while providing a 67% increase in energy efficiency. The system scales up to 256 chips in a single high-bandwidth, low-latency pod using their high-speed inter-chip interconnects. For further scaling, pods can be connected via their 13 Petabits per second Jupiter data center network
- Launch of Hypercompute Cluster, a solution for large-scale AI workload management that enables dense resource colocation, provides targeted workload placement across thousands of accelerators and incorporates advanced maintenance features to minimize workload disruptions. It allows customers to deploy and manage large accelerator pools as a single unit while maintaining ultra-low-latency networking performance.
- Enhancement of its Cloud Interconnect service with new application awareness capabilities, optimizing traffic prioritization during network congestion and improving bandwidth utilization.

These developments from AWS and Google Cloud demonstrate the increasingly sophisticated and specialized nature of hyperscale networking solutions for AI workloads, with particular emphasis on low latency, high bandwidth, and efficient resource utilization at a massive scale.

## Other Ecosystem Vendors

### Marvell



**Marvell** has established itself as a key player in data center networking through its comprehensive semiconductor solutions, particularly for AI workloads. The company's flagship Teralynx 10 Ethernet switch, which entered volume production in July 2024, delivers 51.2 Tbps of switching capacity with latency as low as 500 nanoseconds. Its high radix design enables significant infrastructure efficiencies, reducing network layers in large AI clusters and requiring up to 40% fewer switches in 64K xPU deployments. The solution achieves laudable power efficiency at 1W per 100Gbps, representing a 50% reduction from previous generations.

Beyond switching, Marvell's portfolio addresses critical data center connectivity needs through several innovative product lines. Their Structera CXL products tackle memory bandwidth challenges in AI-driven environments, while their Alaska P PCIe retimer products enable high-speed connectivity between AI accelerators, GPUs, and CPUs. In the optical domain, Marvell's new Aquila coherent-lite DSP supports 1.6 Tbps optical transceiver modules for data center interconnects, optimized for O-band wavelengths to reduce cost per link compared to traditional C-band solutions. These innovations are validated by strategic partnerships, including a five-year agreement with AWS covering custom AI products, optical DSPs, and Ethernet switching solutions.

### Broadcom



**Broadcom** maintains its dominant position in data center networking through its comprehensive switching portfolio, particularly with its Tomahawk series. The Tomahawk 5 represents a big leap forward with:

- 51.2 Tbps of throughput supporting configurations up to 64x800G ports
- Power efficiency under 1W per 100Gbps through 5nm process technology
- Cognitive Routing feature optimizing latency for AI/ML workloads
- Validation for hyperscale AI clusters supporting over 1 million xPU infrastructures

---

<sup>40</sup> "Google Cloud turbocharges its AI Hypercomputer stack with next-gen TPUs and bigger GPU clusters" SiliconANGLE, Oct. 30, 2024.

According to JP Morgan, Broadcom's market leadership is estimated at 80% share of the \$5-7B data center/AI Ethernet switching market. Beyond the Tomahawk line, the company continues to innovate across multiple product families. The Trident 5-X12, targeting enterprise and cloud environments, offers 16 Tbps bandwidth with 800G uplink ports and features an on-chip neural network for traffic pattern analysis. The upcoming Tomahawk 6, expected to be announced at OFC 2025, aims to achieve 102.4 Tbps throughput using 3nm process technology. Strategic partnerships with major hyperscalers like AWS, Google, and Meta, combined with their comprehensive portfolio including Jericho, Ramon, and Qumran product lines, position Broadcom as a crucial enabler of next-generation AI and cloud infrastructure.

### AlphaWave Semi



**AlphaWave Semi**, a public company listed on the London Stock Exchange, focuses on high-speed connectivity IP, offering comprehensive solutions for data center environments. Their PCIe and CXL subsystems combine controller IPs with multi-standard SerDes PHY solutions to create flexible, high-performance architectures. The KappaCORE CXL controller stands out for its highly configurable design, supporting parallel TLP/DLLP processing and seamless integration with Alphawave's PipeCORE PCIe PHY IP. For next-generation interconnects, their AresCORE UClie Die-to-Die PHY IP enables high-bandwidth density and low-latency connections between chiplets within a package.

### Astera Labs



**Astera Labs**, which completed its IPO in March 2024, has established itself through purpose-built connectivity solutions for data-centric systems in data centers. Their Intelligent Connectivity Platform integrates high-speed connectivity ICs, modules, and boards with their COSMOS connectivity system management software.

The company's product portfolio include:

- Aries PCIe/CXL Smart DSP Retimers: Address signal integrity challenges in AI and general-purpose servers while maintaining low power consumption
- Aries PCIe/CXL Smart Cable Modules: Provide long-reach copper cabling connectivity for disaggregated cloud and AI infrastructure
- Leo CXL Smart Memory Controllers: Support memory expansion and pooling, enabling optimized memory utilization in cloud servers

### Credo



**Credo**, a public company listed on the NASDAQ, has developed a suite of solutions for data centers, particularly focusing on AI computing and CXL applications. At the OCP Global Summit 2024, they showcased their PCIe 6 retimers, including the Toucan x16, which allows designers to extend PCIe trace lengths while maintaining optimal system performance. The retimers include sophisticated debug tools providing visibility into PCIe port/lane status, Link Training and Status State Machine monitoring, waveform analysis capabilities, and real-time eye diagram visualization.

Their HiWire Active Electrical Cables (AECs) represent a significant advancement in connectivity solutions, providing extended range with lossless end-to-end PCIe connectivity. These cables have proven particularly valuable in AI and data center applications where reliability and performance are critical.

The success of these ecosystem vendors demonstrates the increasing specialization within the data center networking supply chain. Their continued innovation in areas like signal integrity, power efficiency, and system monitoring will be crucial as data center networks evolve to meet the growing demands of AI workloads.

## Observations and Recommendations

We recognize that this report runs longer than our typical research brief due to the expansiveness of the topic. There are more details that we haven't delved into, but we hope the above content provides our readers with broad context around the space. This section will provide our concise suggestions on data center networking for CxOs at enterprises and service providers.

### Key Observations on Data Center Networking

The transformation of data center networking by AI workloads has revealed several critical trends:

#### AI is Fundamentally Reshaping Data Center Architecture

The impact extends across all scales in networking – from training to inference workloads and from scale-up to scale-out to scale-outside networking. Critical developments include:

- The shift from individual servers to integrated rack-scale systems, exemplified by NVIDIA's GB200 NVL72 and AWS's Trainium2 UltraServer.
- Power is becoming a critical constraint, with projections indicating a need for 47 GW of incremental power generation through 2030 in the US alone.
- The evolution of both training and inference architectures driving new connectivity requirements.

#### Network Architecture Complexity Demands New Approaches

Today's AI workloads require sophisticated solutions for:

- Multi-dimensional parallelism (data, pipeline, tensor) requiring network-aware scheduling.
- Emerging distributed and federated training across geographically dispersed data centers requiring high bandwidth low-latency interconnects.
- Real-time tracing and monitoring for debugging performance issues, including subtle problems like bit flips.
- Complex topology management and workload placement.

#### Standards and Open Solutions Are Gaining Momentum

We're seeing strong industry commitment to open standards through:

- The Ultra Ethernet Consortium's UET protocol development.
- UALink initiatives for high-speed interconnects.
- Growing adoption of SONiC across major vendors.
- Balance between open solutions and proprietary performance advantages.

#### Network Performance Requirements Continue Intensifying

Key trends we expect in network performance include:

- Backend networks moving to 400/800Gbps speeds with RDMA/RoCE support.
- Critical importance of lossless networking with sophisticated congestion control.
- Emergence of novel topologies like rail-only networks optimized for AI training.
- Growing focus on end-to-end latency optimization.

## AvidThink Recommendations

### Embrace Architectural Flexibility

Organizations should develop network architectures that can adapt to evolving AI workload requirements. This means:

- Implementing separate frontend and backend networks optimized for different workload types. Frontend networks can focus on both traditional enterprise traffic and interactions with AI clusters, while backend networks support intensive AI communication patterns for training or inferencing.
- Considering data center deployments with Ethernet and InfiniBand as appropriate (e.g., Meta, Microsoft Azure). While Ethernet with RoCE is gaining ground, InfiniBand may be optimal for specific high-performance requirements.
- Planning for distributed AI training and inference capabilities across multiple data centers. This requires thinking beyond single-site architectures to include inter-data center connectivity and orchestration.
- Maintaining flexibility to support training and inference workloads as their balance evolves. Infrastructure should be adaptable enough to accommodate shifts in workload patterns.

### Prioritize Operational Excellence

Success in AI-driven data centers requires robust operational practices:

- Implement comprehensive monitoring and telemetry systems that provide end-to-end visibility across the entire network stack. This visibility becomes crucial for troubleshooting complex AI training issues.
- Develop expertise in network-aware workload scheduling and placement. Teams need to understand both traditional networking concepts and AI workload characteristics.
- Invest in automation tools and platforms to manage increasing complexity. Manual operations are impractical at AI scale.
- Build teams with cross-functional expertise spanning both traditional networking and AI infrastructure. This may require investment in training and recruitment.

### Focus on Future-Readiness

Long-term success requires careful consideration of future requirements:

- Plan for power constraints in the design process. Power availability may be the limiting factor in data center expansion.
- Understand how direct liquid cooling will impact the data center architecture.
- Adopt open standards where possible while maintaining performance requirements. This provides flexibility for future technology adoption.
- Consider the impact of emerging technologies like optical switching and co-packaged optics in future architecture plans.
- Build infrastructure that can scale vertically (within racks) and horizontally (across data centers) as demands grow.

### Enhance Security and Reliability

Modern AI-driven data centers require robust security and reliability measures:

- Implement zero-trust security frameworks and micro-segmentation from the outset. The high value of AI training data and models makes security paramount.
- Deploy DPU/SmartNIC solutions for enhanced security and network function offload. These provide hardware-accelerated security features while improving network performance.
- Build redundancy and fault tolerance into initial network designs. Extended AI training jobs require exceptional reliability.
- Consider the impact of soft errors and implement appropriate detection and mitigation strategies. High-speed networks supporting AI workloads are susceptible to these issues.

## Optimize for Cost and Efficiency

Financial optimization requires a holistic approach:

- Evaluate the total cost of ownership, including power, cabling, cooling, and operational expenses. Power costs, in particular, may become a dominant factor in TCO calculations.
- Consider location-based strategies to access low-cost, sustainable power sources. This may involve distributed architectures across multiple regions.
- Implement efficient workload placement and scheduling to maximize resource utilization. AI infrastructure is too expensive to run at low utilization.
- When choosing between proprietary and open solutions, balance capital expenses against operational flexibility. Sometimes, paying more upfront for flexible solutions reduces long-term costs.

## Wrap-Up

AI workloads, particularly training runs for LLMs and other FMs, are reshaping the landscape of data center networking in 2025. This transformation occurs alongside continued cloud evolution, creating a complex environment in which traditional networking approaches are being upended.

The industry is seeing unprecedented investment in AI infrastructure, with major players committing hundreds of billions of dollars to new data center buildouts. However, this expansion faces significant constraints, particularly in power availability and the need to demonstrate commercial returns on these massive investments. The shift from training to inference workloads, coupled with innovations in model efficiency and deployment strategies, suggests future architectures may need to be more flexible and distributed than current designs.

The networking industry is innovating across multiple fronts, from new standards and open initiatives like UALink and UEC/UEP to continued development of proprietary solutions like NVIDIA NVLink. The traditional boundaries between scale-up and scale-out networking are blurring, with rack-scale computing emerging as a new architectural paradigm. Meanwhile, significant strides are being made in optical networking, congestion control, and network automation.

Looking ahead, successful data center networking strategies will need to balance multiple competing priorities: performance versus openness, scalability versus power efficiency, and security versus operational simplicity. Organizations must prepare for a future where AI workloads may be increasingly distributed across multiple facilities and regions, requiring new network architecture and management approaches.

As we move through 2025 and beyond, the pace of innovation shows no signs of slowing. The networking industry's ability to meet the demands of AI workloads while addressing practical constraints will be crucial in enabling the next generation of AI applications and services. Those who can successfully navigate these challenges while maintaining flexibility for future developments will be best positioned for success in this rapidly evolving landscape.

Reach out to us at [research@avidthink.com](mailto:research@avidthink.com) with your feedback on this research brief. We look forward to hearing from you!

## INTERVIEW SECTION FOLLOWS

Please support our sponsors. Check out in-depth interviews with data center networking thought leaders and experts. We encourage you to visit their websites to learn more about their solutions.

# Interview with Dudy Cohen VP of Product Marketing, DriveNets

**AI workloads, like high-performance computing (HPC) before, demand extremely high-performance networking. Why consider anything else but InfiniBand for AI infrastructure in data centers?**

InfiniBand is a well-known option for AI clusters, particularly for its low latency and high throughput. However, it's not well suited for enterprise and cloud data centers because it is complex to configure, requires a specialized skill set, and lacks native support for multi-tenancy of different types of workloads. Plus it's typically more costly than alternatives like Ethernet. This makes adopting InfiniBand challenging for large-scale, multi-tenant environments such as public cloud providers or large enterprises looking to build scalable AI infrastructure.

**If InfiniBand isn't the ideal solution for these environments, why hasn't standard Ethernet been able to fully meet AI networking requirements?**

Traditional Ethernet, or "vanilla" Ethernet, wasn't designed for the needs of AI workloads. AI training clusters require networks with zero packet loss, ultra-low jitter, and fast convergence times in case of faults. Standard Ethernet lacks the performance needed to ensure seamless communication between thousands of GPUs. However, advancements in Ethernet architectures are addressing these gaps.

**What then are the key characteristics of an Ethernet-based solution that can effectively support AI workloads?**

The right Ethernet solution for AI needs to incorporate scheduling mechanisms to handle congestion and optimize traffic flows. There are two main approaches to scheduled Ethernet:

- 1. Endpoint-Scheduled Ethernet**, where the network interface cards (NICs) or endpoints manage traffic distribution based on telemetry data. This approach is still evolving, with solutions like Ultra Ethernet and NVIDIA's Spectrum-X pushing the boundaries.
- 2. Fabric-Scheduled Ethernet**, which takes a more holistic approach by handling load balancing at the network level. We believe this is the most effective method for AI workloads because it ensures perfect load balancing, eliminates head-of-line blocking, and provides lossless packet delivery.

**Why is a scheduled Ethernet fabric much better for AI workloads?**

A **scheduled Ethernet fabric** optimizes the entire data path from the top-of-rack switches to the spine, ensuring efficient traffic distribution. It does this by:

- Using **cell-based switching**, where packets are broken into fixed-size cells and evenly distributed across the network to eliminate congestion.
- Implementing **Virtual Output Queues (VOQs)** and **grant-based mechanisms** to prevent packet drops and ensure lossless transmission.
- Creating a **single Ethernet hop** architecture, meaning that every GPU is essentially connected to the same switch, even when scaling up to massive clusters like 32,000 GPUs.

This ensures that AI workloads operate with maximum efficiency, delivering the low-latency, high-performance fabric needed for training large models.

**How does DriveNets' approach to AI networking differentiate itself?**

DriveNets' **Network Cloud AI** solution takes the principles of a scheduled Ethernet fabric and implements them using a disaggregated, cloud-like architecture. It runs on commercial off-the-shelf (COTS) white-box hardware, allowing enterprises and cloud providers to select the best ODM and optical vendors for their needs. This provides:

- A highly **scalable** and **open** architecture that avoids vendor lock-in.
- **Cost efficiency** by leveraging white-box economics.
- **High performance** without compromising Ethernet's flexibility and scalability.

With this approach, data centers can move beyond the limitations of InfiniBand and standard Ethernet, achieving a truly optimized AI fabric that scales efficiently.



[Learn more about DriveNets AI Networking Solution](#)

**DRIVE**  **NETS**

# NOKIA

## Interview with Mike Bushong, Vice President, Data Center



**AI is often treated as a singular market, but you've pointed out that it's actually multiple markets with distinct challenges. Can you elaborate on that?**

Absolutely. AI isn't a monolithic market—it's a series of interconnected but distinct markets. At one end, you have the hyperscalers — Amazon, Google, Microsoft, Meta — who operate under very different conditions than everyone else. Even within that segment, their strategies differ based on geography. What works in North America may not apply in Brazil, India, or the Middle East, where real estate and power availability dictate deployment models. For example, a Google data center might support 100-kilowatt racks, while a Mumbai facility may be limited to much smaller, distributed deployments. These variations drive different architectural considerations, from power distribution to networking and workload placement.

**How do these constraints impact data center architecture and networking?**

The biggest factors are power and real estate. Hyperscalers build to their own specifications, but others must work within existing infrastructure — maybe a 20-megawatt site instead of a gigawatt-scale campus. That affects everything: rack density, cabling choices, cooling strategies, and even the way compute clusters are distributed. Do workloads stay local, or extend across multiple locations? If multiple sites are involved, WAN connectivity becomes critical. These factors drive architectural decisions that directly shape networking strategies.

**You mentioned a shift toward multi-tenant AI infrastructure. What are the networking implications?**

The rise of GPU-as-a-Service and colocation-based AI models introduces complexity. If an anchor tenant only consumes 60% of capacity initially but scales over time, how do you design for flexible workload migration? Do you create separate clusters, or build a unified infrastructure that allows seamless transitions? Networking plays a major role here — latency, congestion management, and interconnect topology all matter when workloads are shifting dynamically.

**Ethernet vs. InfiniBand has been a longstanding debate in AI networking. Where do you see the market going?**

There's no absolute winner-takes-all scenario, but we believe Ethernet will dominate over time. The economic advantages are clear — merchant silicon from vendors like Broadcom delivers superior cost efficiencies at scale. Silicon from Broadcom like Tomahawk 4, 5, and 6, combined with next-gen NICs like NVIDIA ConnectX-7 and ConnectX-8, push Ethernet performance forward. Congestion management and lossless networking will evolve, but in a world that values interoperability and multi-vendor flexibility, Ethernet has the edge. The UEC (Ultra Ethernet Consortium) is playing a key role in driving standardization, which reinforces this trend.

**What's your perspective on interconnecting AI data centers beyond just switching inside racks?**

AI data centers won't exist in isolation — they'll need high-bandwidth, low-latency interconnects. That means DCI (Data Center Interconnect) and optical networking are just as important as the internal switching fabric. Our recent acquisition of Infinera strengthens our position in this space, allowing us to integrate optics directly with our networking solutions. The future of AI networking isn't just about what happens inside the rack or even across the row; it's about how those units connect to a wider AI computing infrastructure with ultra-reliable, high-speed transport.

**Finally, let's wrap up with the management of potentially massive networking infrastructure. What role can AI-specific infrastructure management play?**

Hyperscalers handle their own operations, but for everyone else, operational complexity becomes a critical challenge. That's why we built Nokia Event-Driven Automation (EDA), a multi-vendor, intent-based management platform that abstracts complexity. Instead of focusing on low-level configurations, enterprises can define high-level intent, which the system translates across platforms like SR Linux or open-source SONiC.

**[Discover how Nokia can help you implement networking for AI workloads.](#)**

# An Interview with Shekar Ayyar, Chairman & CEO, Arccus



## **AI is transforming many industries, but what does that mean for networking infrastructure? How does networking play a role in AI's evolution?**

AI's impact on industries has been immense, but a critical piece often overlooked is the underlying infrastructure needed to support AI workloads. Initially, AI was all about training — bringing together massive datasets, processing them in high-performance computing clusters, and refining large-scale models. The next challenge that we'll face is inferencing, where AI models are deployed in real-world applications across industries like finance, healthcare, and autonomous systems.

Training can occur in a few large-scale data centers, but inferencing requires AI models to be deployed and executed across distributed locations — at the edge, across multiple data centers, and even in specialized processing environments. This means that networks need to efficiently handle data movement, ensure low-latency connectivity, and optimize power consumption. AI's success now depends on how well the network can support these distributed inferencing workloads.

## **You've previously mentioned the growing role of Ethernet in AI networking. Can you expand on that?**

Historically, proprietary networking technologies dominated within tightly-coupled GPU clusters used for high-performance computing and training large AI foundation models. However, Ethernet is rapidly evolving, offering comparable performance and low latency. Advancements in networking silicon from Broadcom, NVIDIA, and others are pushing Ethernet into new frontiers, making it capable of supporting AI workloads across the entire infrastructure — from GPU-to-GPU communications to inter-data-center traffic and beyond. We envision a future where Ethernet spans the entire network, from within GPU stacks to inter-data center connectivity and edge inferencing, offering a flexible, open alternative that supports AI at scale.

## **What specific innovations is Arccus bringing to AI networking?**

At Arccus, we are focused on building a networking fabric that is highly flexible, distributable, and optimized for AI workloads. One example is our work with NVIDIA's Bluefield DPUs, where we enable

IPsec offloading directly at the network layer. This allows AI workloads to operate with higher efficiency, freeing up GPU resources that would otherwise be used for encryption tasks.

Another key innovation is our Egress Cost Control solution, which can optimize the cost of moving data between hyperscaler data centers. AI applications involve enormous amounts of data transfer, and managing egress costs is critical for ensuring economic efficiency. By embedding cost-aware routing intelligence directly into our software, we help organizations minimize unnecessary expenses while maintaining performance.

## **With AI expanding across industries, how should enterprises and hyperscalers think about security in AI-driven networks?**

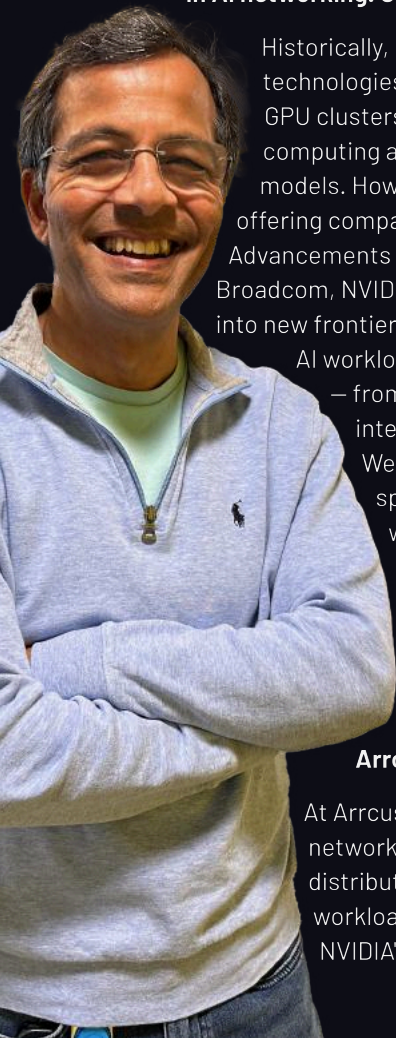
Security is an integral part of AI infrastructure, and it needs to be built into the network itself rather than treated as an afterthought. As AI workloads become pervasive, security vulnerabilities also expand, making it essential for networking platforms to provide native security capabilities.

At Arccus, we incorporate security at the routing level, such as implementing route origin validation to verify the authenticity of data flows. This ensures that traffic entering the network is authenticated and policy-controlled at the routing layer. Additionally, by exposing security functions as APIs, we enable organizations to integrate security policies directly into their AI operations, reducing risk and ensuring compliance.

## **Looking ahead, how do you see networking evolving to support AI's continued growth?**

AI is pushing infrastructure to its limits, and networking must evolve to keep up. We're moving toward a world where AI workloads are no longer confined to centralized clusters but are distributed across a broad spectrum of environments — from cloud data centers to on-premises deployments and edge locations.

As this shift occurs, networking will need to be intelligent, cost-efficient, and secure. Disaggregated network architectures, software-driven optimizations, and the rise of Ethernet-based AI fabrics will all play a crucial role in shaping the next generation of AI infrastructure. Arccus is excited to be at the center of this transformation, providing the networking foundation that enables AI to scale seamlessly and efficiently.





## Interview with Kevin Deierling Senior Vice President of Networking

**NVIDIA has been pushing the boundaries of scale-up architecture. What's the strategy behind GB200 NVL72, and how does it fit into the broader AI infrastructure picture?**

Scale-up is fundamental to how we're approaching AI infrastructure. With GB200 NVL72, we're enabling our Blackwell processors to work together as if they were a single GPU, which is crucial for modern AI workloads. This unified approach simplifies programming and supports critical capabilities like tensor parallelism and pipeline parallelism – features that are essential for today's large language models. With NVLink interconnects in NVL72, we follow a "copper where you can" philosophy, leveraging its advantages of lower power consumption, cost efficiency, and reliability for scale-up, while using optics for scale-out beyond that.

**We're seeing a shift in how inference workloads are being deployed. How is this changing the infrastructure requirements compared to traditional expectations?**

The industry initially thought inference would be simple, something you could run on a single CPU. That's proven dramatically wrong. Inference is becoming the peer of training, especially with what we call test-time scaling or test-time inference. We're seeing the emergence of Agentic AIs and complex reasoning that requires significant computational power. This shift means inference platforms need the same scale-up capabilities we originally designed for training. It's driving a fundamental change in how we think about infrastructure design.

**Could you explain the differences between your approach to InfiniBand and Ethernet for AI workloads?**

InfiniBand remains the gold standard for AI scale-out networking, offering about 30% better performance than Ethernet. However, we recognize Ethernet's importance in the ecosystem. The key is understanding that traditional Ethernet wasn't designed for AI workloads. AI training involves highly correlated operations across thousands of GPUs, with all nodes needing to communicate simultaneously through collective operations.

That's why we developed our Spectrum-X networking platform specifically for AI, incorporating features like adaptive routing and congestion control. You can't just use traditional Ethernet infrastructure and expect optimal AI performance – you need purpose-built solutions that understand these unique workload characteristics.

**What's NVIDIA's position on industry consortiums like UALink and the Ultra Ethernet Consortium?**

We are a member of the Ultra Ethernet Consortium but not yet a member of the UALink Consortium. We've supported open standards when they meet customer needs and can move at the pace the industry requires. The challenge with standards bodies is often the speed of development and achieving consensus among partners with different priorities. That's why we continue to innovate independently with technologies like NVLink and our Spectrum-X platform. However, as standards mature, and customers and partners express interest, we're always ready to participate. We've demonstrated this approach with technologies like OpenGL, while also maintaining our own frameworks like CUDA to drive innovation forward.

**As we look toward 2025, what major shifts do you see in the AI infrastructure landscape?**

While training has dominated the conversation, 2025 will be the year of inference. We're seeing interesting parallels to the Jevons paradox – as we make inference more efficient, instead of reducing demand, we're enabling broader adoption across new use cases. This expansion will drive unprecedented infrastructure requirements. The scale we're talking about now is staggering – customers are discussing deployments of not just hundreds of thousands, but millions of GPUs. It's the most exciting phase I've seen in technology, beyond emergence of the web or cloud computing, and it's driving innovation across the entire data center stack.



## Q&A with Nick Kucharewski

SVP and GM, Network Switching Business Unit

**The AI industry has seen rapid growth in computing infrastructure. Can you define what "scale-up" means in this context and how it differs from "scale-out"?**

Scale-up refers to adding resources to a system while ensuring software perceives it as a single computer – expanding compute capacity or memory within a unified architecture. Scale-out, in contrast, involves distributed computing, running multiple software instances across many machines.

**How is Marvell positioning itself for the AI data center?**

Marvell is investing in a comprehensive scale-up and scale-out interconnect portfolio, spanning electrical and optical solutions. Innovations include PCIe retimers, high-density SerDes, memory accelerators, optical modules, high-bandwidth Ethernet switches, and custom silicon for AI fabrics. Our co-design and custom capabilities are critical for next-generation AI infrastructure.

**What are the key challenges in designing scale-up compute architectures?**

Unlike networking, which handles diverse interconnects and congestion, compute chipsets prioritize highly reliable, low-latency channels. Key challenges include power, cooling, distance, and time. AI accelerators (xPUs) require vast power, making delivery and cooling difficult. Spacing them apart introduces another issue – high-speed signals degrade over distance. Signal regeneration can help, but each added device increases latency, affecting performance. Ensuring deterministic, predictable response times at scale is exponentially complex.

**What are the dominant interconnect architectures being considered for scale-up?**

We're seeing multiple fundamental approaches emerging. Circuit switching can support arbitrary domain sizes with deterministic, high-bandwidth connections, but it's semi-static. Flit (flow control unit) switching, used in protocols like PCIe and CXL, provides any-to-any connectivity with bounded latency but has a limited domain size. Then there's packet switching with Ethernet, which offers large domains but needs additional constraints and features to deliver the determinism required for scale-up applications. The choice depends on whether you're performing training or inference, running single or multi-tenant applications, and your specific data set characteristics.

**What innovations are shaping the future of scale-up interconnects?**

The industry is shifting toward custom interconnects for AI workloads. **Co-packaged copper (CPC)** extends its reach while maintaining low power. **Linear pluggable optics (LPO)** reduce latency and power versus traditional DSP-based optics. **Co-packaged optics (CPO)** integrate optical links directly into ASICs, improving efficiency. As AI clusters scale beyond racks into row-based architectures, transitioning from copper to optical is inevitable.

**What's the ultimate vision for scale-up computing?**

AI clusters will scale to 1000+ xPUs, driven by custom silicon integrating compute, fabric, and optical connectivity. Industry collaboration is key to optimizing manufacturing and ensuring ecosystem adoption. As we advance fully customized, lossless scale-up fabrics, we're pushing the boundaries of AI infrastructure.

**Let's step back from scale-up and talk about scale-out; what are the architectural considerations?**

There's no single perfect architecture. At the rack level, customers can choose between top-of-rack switches, middle-of-row architectures, or end-of-row configurations. Choices are driven by cluster size, interconnect speeds, and physical distance constraints. At the data center level, we see many approaches: single buildings maximizing available power, campus distributions spanning 10-20 Kms, and regional data centers separated by thousands of kilometers. Each requires different solutions, from modulation schemes to wavelength division multiplexing options.

**Any recent Marvell scale-out innovation you'd like to touch on?**

My colleagues developed **'Coherent-lite'** silicon photonics technology, bridging traditional PAM4 and long-haul coherent solutions. It's optimized for 2-20 kilometer campus-style links at 400G per wavelength, providing lower cost and power consumption than traditional coherent solutions. For longer distances, we've pioneered pluggable coherent modules capable of 800G per wavelength, with a path to 1.6T.





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

©2025 AvidThink LLC. All Rights Reserved.

This material may not be copied, reproduced, or modified in whole or in part for any purpose except with express written permission from an authorized representative of AvidThink LLC. No part of this work may be used or reproduced in any manner for the purpose of training artificial intelligence technologies or systems. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgment of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.