AI-led data center networking

White paper

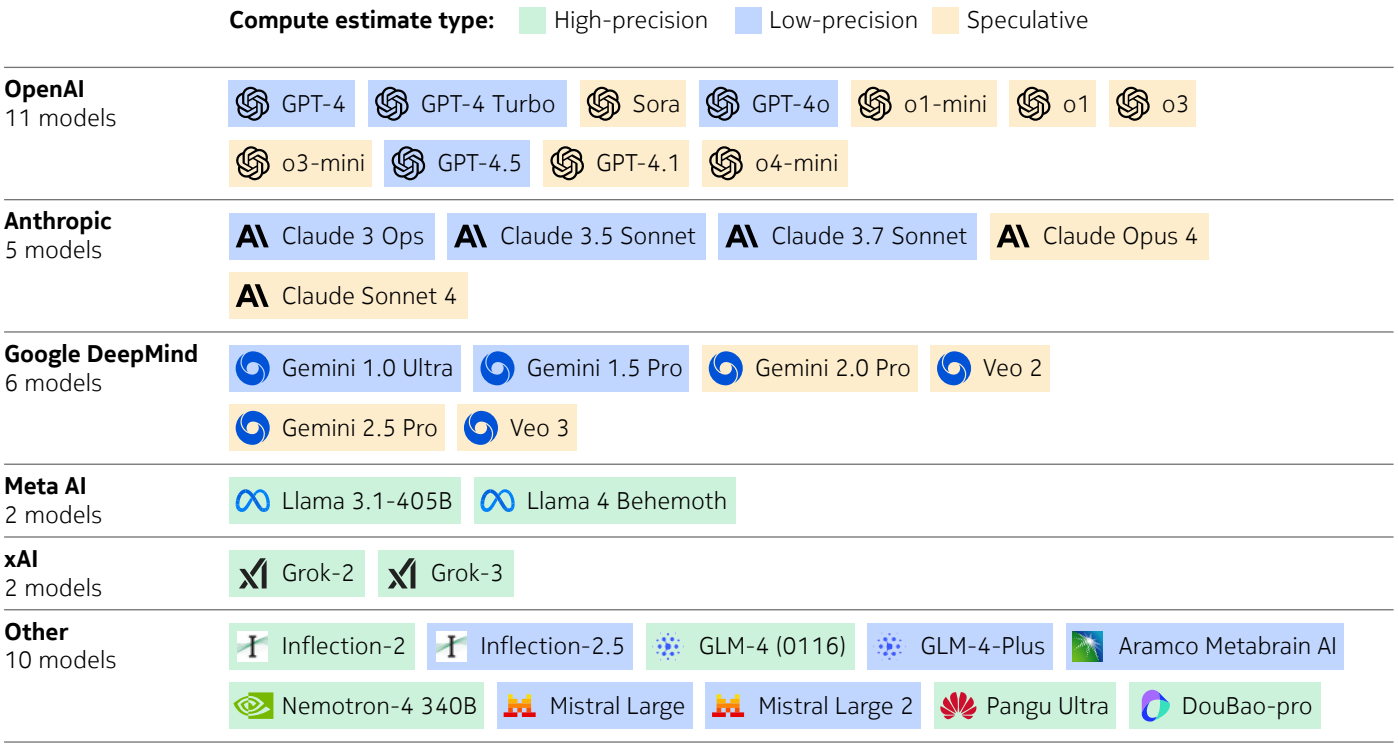
The substantial growth in large language models (LLMs) and parameters that contribute to LLMs as well as AI-driven workloads is reshaping data center (DC) networking demanding unprecedented scalability, low latency and energy efficiency. This paper looks at some of the trends, requirements and challenges that LLM training brings to DC networking.

Contents

Evolution of large language models (LLMs) and GPU requirements

Large language models have grown exponentially, driven by the need for enhanced reasoning and generalization. Similarly, the number of large language models trained each year since 2023 has grown substantially [1]. The following figure presents some of the recent models trained in the recent past and the trends show that almost two large models are being trained each month.

Figure 1. Major LLMs [1]



Outside of the LLMs being trained by the big tech AI giants (e.g., OpenAI, xAI, Google), there is growth happening in sovereign AI and enterprise AIs.

Sovereign AI is a concept that nations and jurisdictions need to build, control and govern their own AI infrastructure, data and models in alignment with local securities, regulations, language and values. Rather than relying exclusively on global hyperscalers, sovereign AI initiatives emphasize domestic cloud capacity, trusted data pipelines and open or locally developed foundation models. For example, the European Union (EU) is pushing sovereign AI through projects like Gaia-X and the European Language Grid, while countries such as France (with Mistral AI) and Germany (with Aleph Alpha) are developing homegrown LLMs. Major countries in Asia-Pacific, like China (with Deepseek and Alibaba) and India (with the Yotta/Savam/Bhashini initiative), are focused on building local AI language platforms across their diverse linguistic ecosystems.

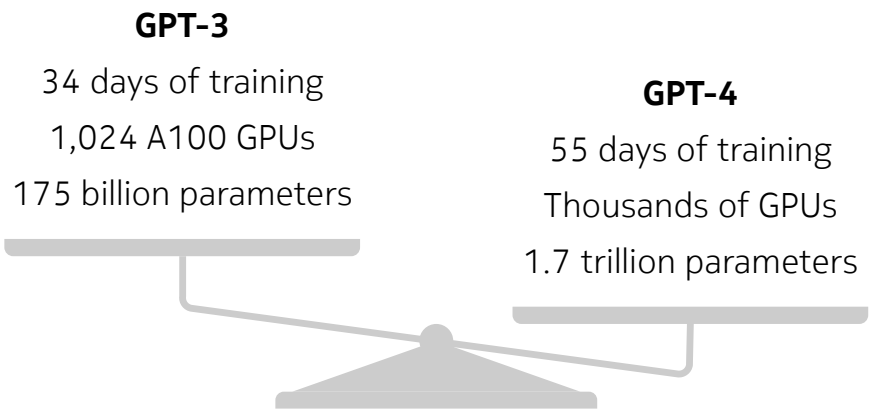
To fast track AI training and inference, a new concept of clouds, called neocloud, is emerging, especially in the EU. Neoclouds are a new generation of cloud providers purpose-built for AI and high-performance computing (HPC) workloads, distinct from traditional hyperscalers. Unlike general-purpose cloud platforms that focus on a broad mix of enterprise IT needs, neoclouds are focusing on offering dense graphics processing unit (GPU) clusters, high-bandwidth/low-latency interconnects, and optimized software stacks designed to accelerate AI training and inference at scale.

LLM size evolution and power needs

Training LLMs presents several challenges, including the immense size of datasets, which can take months to process even on thousands of GPUs. For instance:

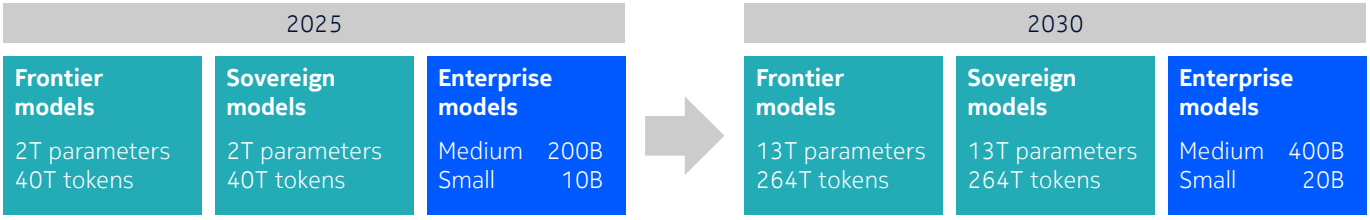
- GPT-3 was trained on 175 billion parameters using thousands of GPUs.
- GPT-4 scaled up to 1.7 trillion parameters, requiring tens of thousands of GPUs.

Figure 2. Comparison of model training requirements



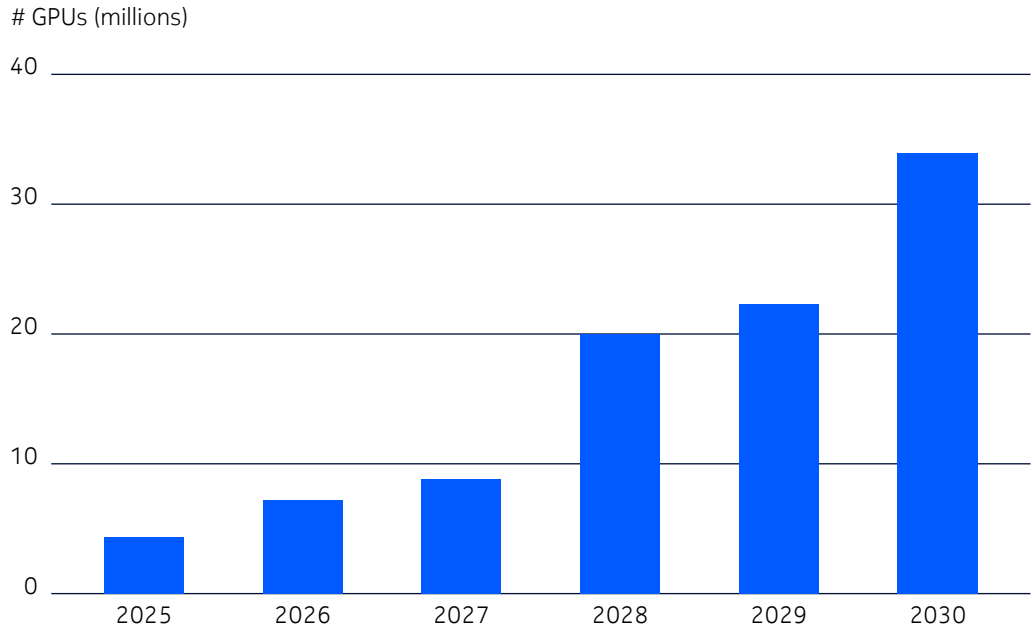
Looking ahead, frontier and sovereign models such as Llama4, which currently uses two trillion parameters and ~40 trillion tokens per LLM, are projected to grow to 13 trillion parameters and ~260 trillion tokens per model by 2030.

Figure 3. Projection of LLM growth



When aggregated across various LLMs, this growth could result in a combined total of 23 quadrillion parameters, necessitating millions of GPUs for training and inference.

Figure 4. Estimated GPU needed for training



Furthermore, AI inference workloads are rapidly increasing. For example, ChatGPT has recorded over 5.2 billion visits per month, leading to significant growth in input and output tokens.

GPU training and its energy demands

Training LLMs and generative AI systems requires immense computational power, which translates into substantial energy consumption. For example:

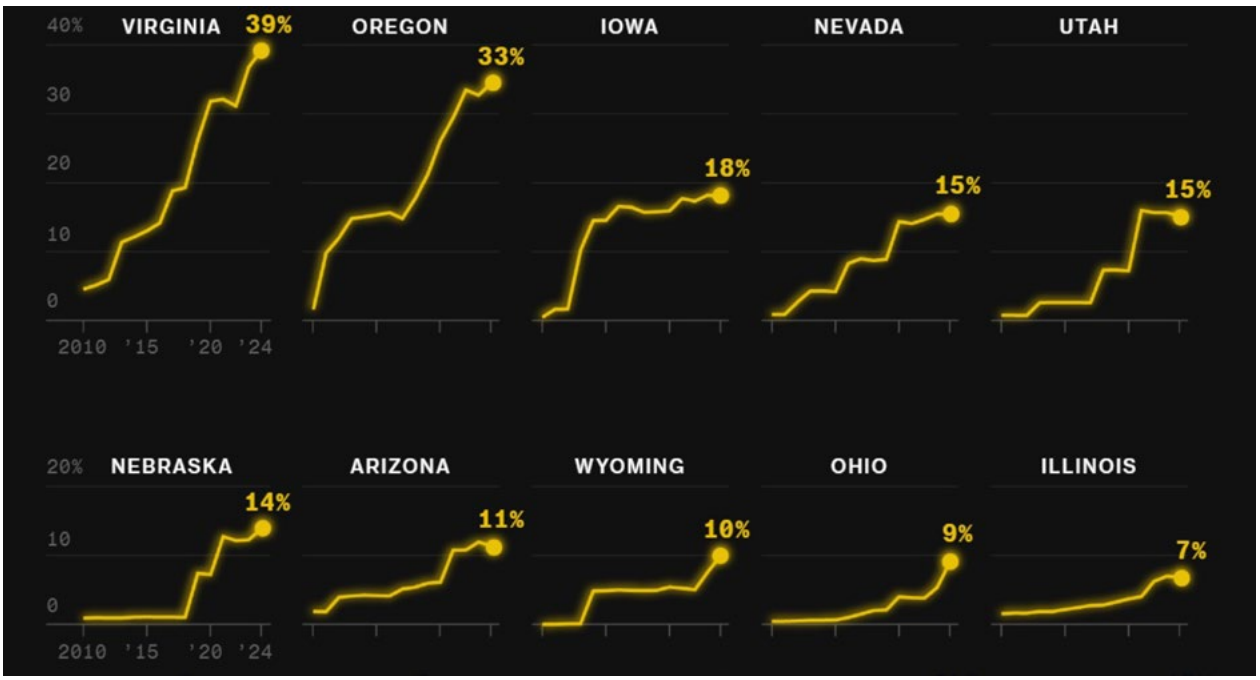
- High-performance GPUs like NVIDIA's H100 can consume up to 700 watts (W) at peak operation, while Blackwell GPUs require over 1400 W—equivalent to the energy use of several households
- At the system level, an NVIDIA HGX H100 system with eight GPUs can draw over 5.6 kW, excluding additional components such as CPUs (central processing units), memory and cooling systems.

Figure 5. Power demands by GPU

Even large telco data centers are not designed to meet such power demands. While typical data centers are built to support ~10 kW per rack, the latest generation of GPUs requires over 10 times this power per rack.

When scaled to the level of AI data centers, the energy demands become even more pronounced. Modern AI facilities, housing thousands of GPUs, require power in the megawatt (MW) range per site. For instance, training a cutting-edge AI model may demand over 4 gigawatts (GW) of power by the late 2020s. Globally, AI data centers are expected to consume 68 GW by 2027—comparable to California's total electricity usage in 2022—and this demand could increase by 165% by 2030. In the US, AI data centers are accounting for 7–39% of total electricity consumption in some states, as reported by Bloomberg [2].

Figure 6. Growth in power demands by US states [2]



One report from Epoch.ai forecasts that the largest individual frontier training runs in 2030 will likely draw 4–16 GW of power, or enough to power millions of US homes [1].

This growing energy demand is driven by several factors, including the exponential increase in model sizes (from billions to trillions of parameters), the need for continuous 24/7 operations, and networking, optics and cooling. Additionally, traditional copper-based interconnects between GPUs and servers become bottlenecks at high speeds, further contributing to energy inefficiency. Addressing these challenges requires innovation in some of the possible strategies:

- Geographically distributed training to overcome local power delivery limits. In future, training workloads may be spread across data centers separated by miles. This will result in a new market for fiber and coherent optics to enable interconnecting AI data centers (AI connect)
- While compute energy efficiency is happening (more FLOPs delivered per W), there is also scope for networking hardware and architectural evolution. For instance, in large AI data centers, the non-IT loads contribute another 10–30% of power demand [3]. Some of the possible options are co-packaged optics (CPO) and linear pluggable optics (LPO).

GPU interconnection

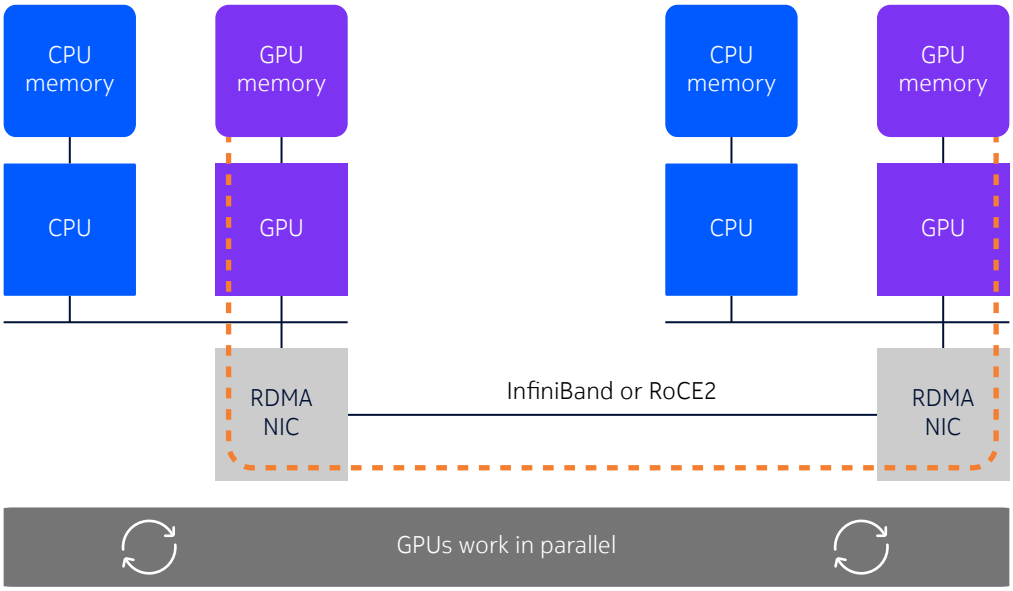
To understand the challenges of the timelines required to train large models, take for example a Llama2 70B model: if trained on a single GPU, it would have taken 1.7 million GPU hours to train.

Table 1. Training time for different LLM models

Model name	Training time (GPU hours)
Llama 2 7B	184,320
Llama 2 13B	368,640
Llama 2 70B	1,720,320

These obstacles can be addressed by leveraging parallelization strategies to distribute the workload across multiple GPUs. For instance, data parallelism involves splitting a large dataset into smaller subsets, allowing each GPU to process a portion independently. Similarly, model parallelism breaks the model into smaller components, enabling different GPUs to handle distinct parts of the model.

Figure 7. Inter GPU communications

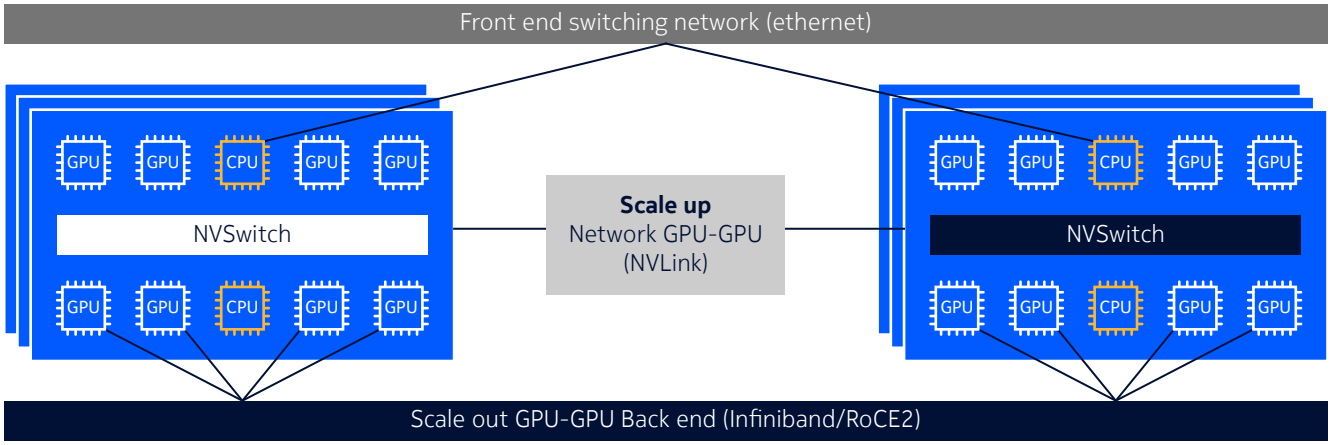


This requires enabling multiple GPUs to communicate and work together. Multi-GPU setups allow for scaling up the computational resources, making it possible to train larger models that wouldn't fit into the memory of a single GPU. This requires that data is split across GPUs. Because CPUs can become the bottleneck, GPUs are connected directly using remote direct memory access (RDMA) to access each other's memory (see "backend networking" below) and typically use protocols such as InfiniBand or RDMA over converged Ethernet (RoCE2) .

GPU parallel processing and networking

The need to connect the GPUs together for parallel processing has given rise to a terminology called “backend networking.” Backend networking connects the GPUs together for parallel processing. Within backend networking there are two ways of interconnecting GPUs: scale-up (vertical scaling within a rack) and scale-out (horizontal scaling across racks).

Figure 8. Scale out and scale up network for GPU connectivity



Scale-up network

Especially promoted by vendors like Nvidia, scale-up networks use NVLink [4], Nvidia’s peripheral component interconnect express (PCIe) protocol alternative. The NVLink protocol addresses the communication limitations between GPUs by interconnecting GPUs within a server. But now the same NVLink protocol has been extended for interconnecting GPUs across servers within racks. This ensures ultra-low latency and provides high throughput speeds for AI workloads.

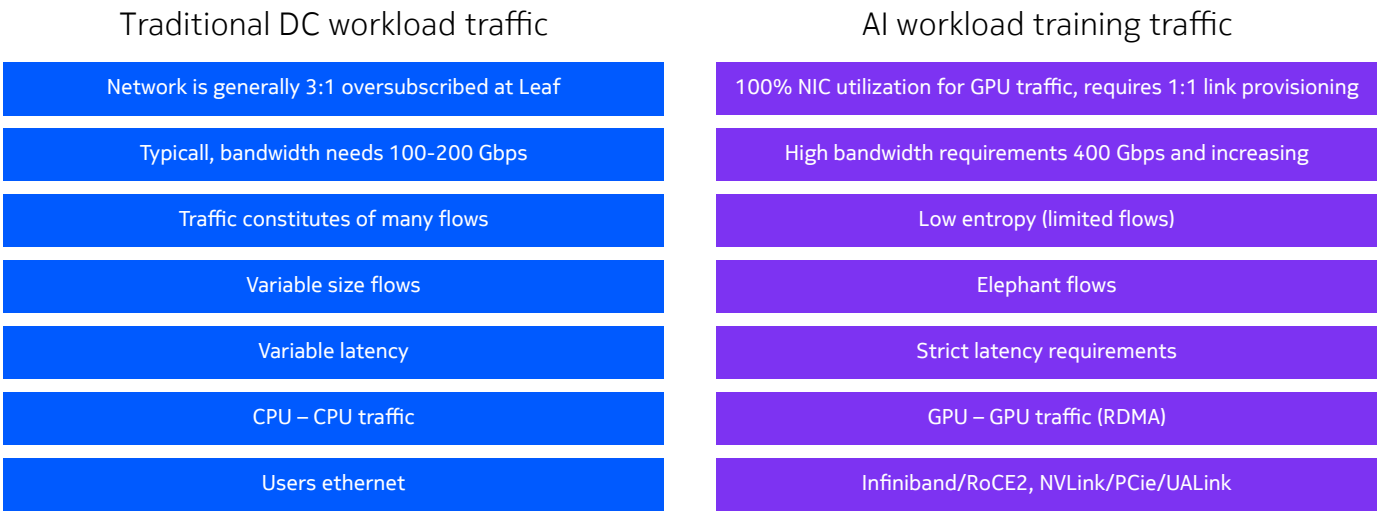
Additionally, a new consortium has been formed by many industry members known as UA Link with an objective to define and form an open standard, for AI scale-up networking.

Scale-out network

By contrast, scale out connects GPUs across different racks spread across in a POD (point of delivery) or data center to create a horizontally expandable infrastructure. It transports GPU payloads utilizing either InfiniBand or RoCEv2 (Remote Direct Memory Access over Converged Ethernet, ds version 2).

Characteristics of AI workload traffic

Figure 9. Difference between traditional and AI workload requirements



A key distinction between AI workload traffic and traditional DC network traffic lies in the characteristics of data flows and bandwidth requirements. AI workloads, especially during the training of large language models, are dominated by “elephant flows”—large, persistent streams of data that transfer vast amounts between GPUs or servers, often saturating link capacity for extended durations (hence requiring high-capacity network bandwidth and 1:1 link provisioning as compared to traditional DCs that are oversubscribed). These flows stem from synchronized operations such as parameter updates and collective communications in distributed training, resulting in predominantly east-west traffic patterns.

In contrast, traditional DC traffic typically comprises a mix of IMIX/variable-sized flows (short, sporadic queries like web requests or database interactions) and occasional elephant flows, with a more balanced distribution of north-south and east-west traffic. This shift in AI workloads demands exceptionally high bandwidth—scaling to 800GE or higher per port—to accommodate terabytes of data without bottlenecks, whereas traditional DCs function effectively at lower speeds, such as 100–200GbE, since their workloads do not require sustained, high-throughput transfers.

Another major difference lies in the protocols used and the focus on lossless transmission. AI traffic frequently utilizes RDMA protocols, such as RoCE (RDMA over Converged Ethernet) or InfiniBand, which facilitate direct GPU-to-GPU data transfers with minimal CPU involvement. This optimizes efficiency for large-scale parallel computations.

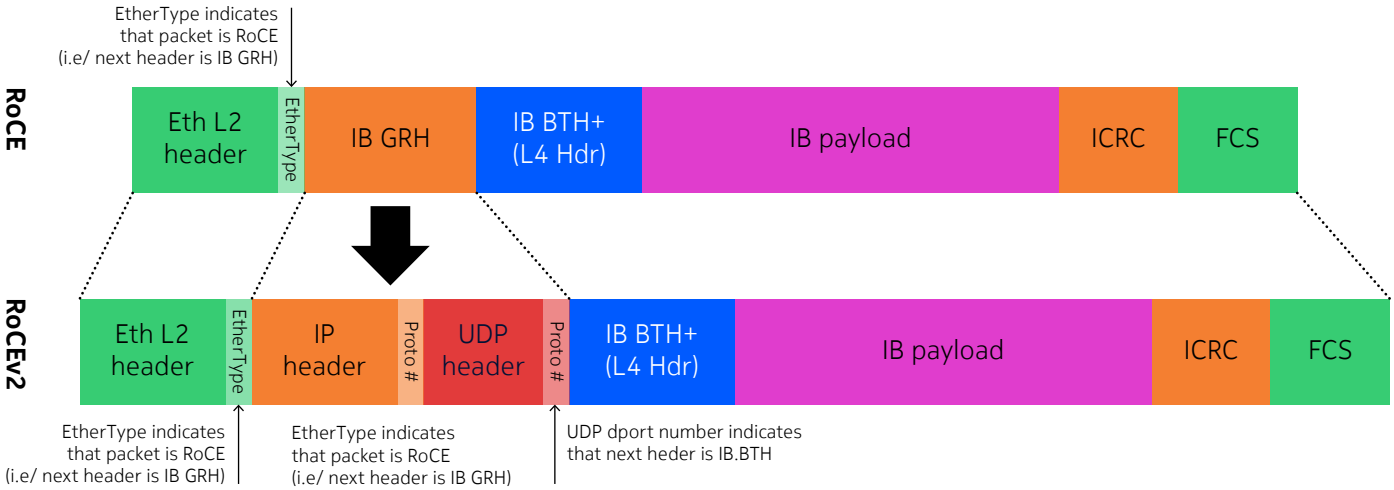
In contrast, traditional DC traffic primarily relies on TCP/IP (transmission control protocol/internet protocol) or UDP (user datagram protocol) over standard Ethernet, where reliability is achieved through retransmission rather than built-in guarantees. For AI workloads, lossless networking is critical; even minor packet losses can cause cascading delays in synchronized operations, necessitating mechanisms like priority flow control (PFC) to pause traffic and prevent data loss. Traditional networks, on the other hand, can tolerate some degree of packet loss, as TCP’s error correction mechanisms mitigate occasional drops without significantly affecting performance, making them more forgiving but less tailored to AI’s precision requirements.

Additionally, AI traffic is characterized by low entropy. This means it consists of fewer flows and hence requires additional capabilities by the network to load balance traffic. In contrast, traditional DC traffic exhibits high entropy, with diverse flows from a wide range of applications, simplifying the load balancing and greater flexibility in managing variability.

RoCEv2 protocol details

The RoCE protocol has evolved to support traffic in layer 3 (L3) environments through a modification of its packet format. In RoCEv2, instead of the global routing header (GRH) used in traditional RoCE, it includes an IP header, which enables it to traverse IP L3 routers. Additionally, a UDP header is incorporated to provide a stateless encapsulation layer for RDMA transport protocol packets over IP.

Figure 10. RoCEv2 header format [5]



RoCEv2 meets the stringent demands of AI workload traffic by utilizing Ethernet infrastructure combined with RDMA capabilities, enabling direct memory access between GPUs. For the large, continuous data streams common in AI training, RoCEv2 employs UDP/IP headers to ensure routability across IP networks, facilitating scalable, all-to-all communication in extensive clusters without fragmentation challenges. Its lossless operation is maintained through PFC, which halts traffic on congested queues to prevent packet drops—an essential feature for synchronized AI processes where even minimal data loss could disrupt progress.

RoCEv2 also requires some development to meet the need for load balancing large datasets that need to be transported in data centers and require the use of all possible data center links to accommodate any link failures. This requires support for per packet load balancing (required reordering at destination) and queue pair ID (QPID) hashing.

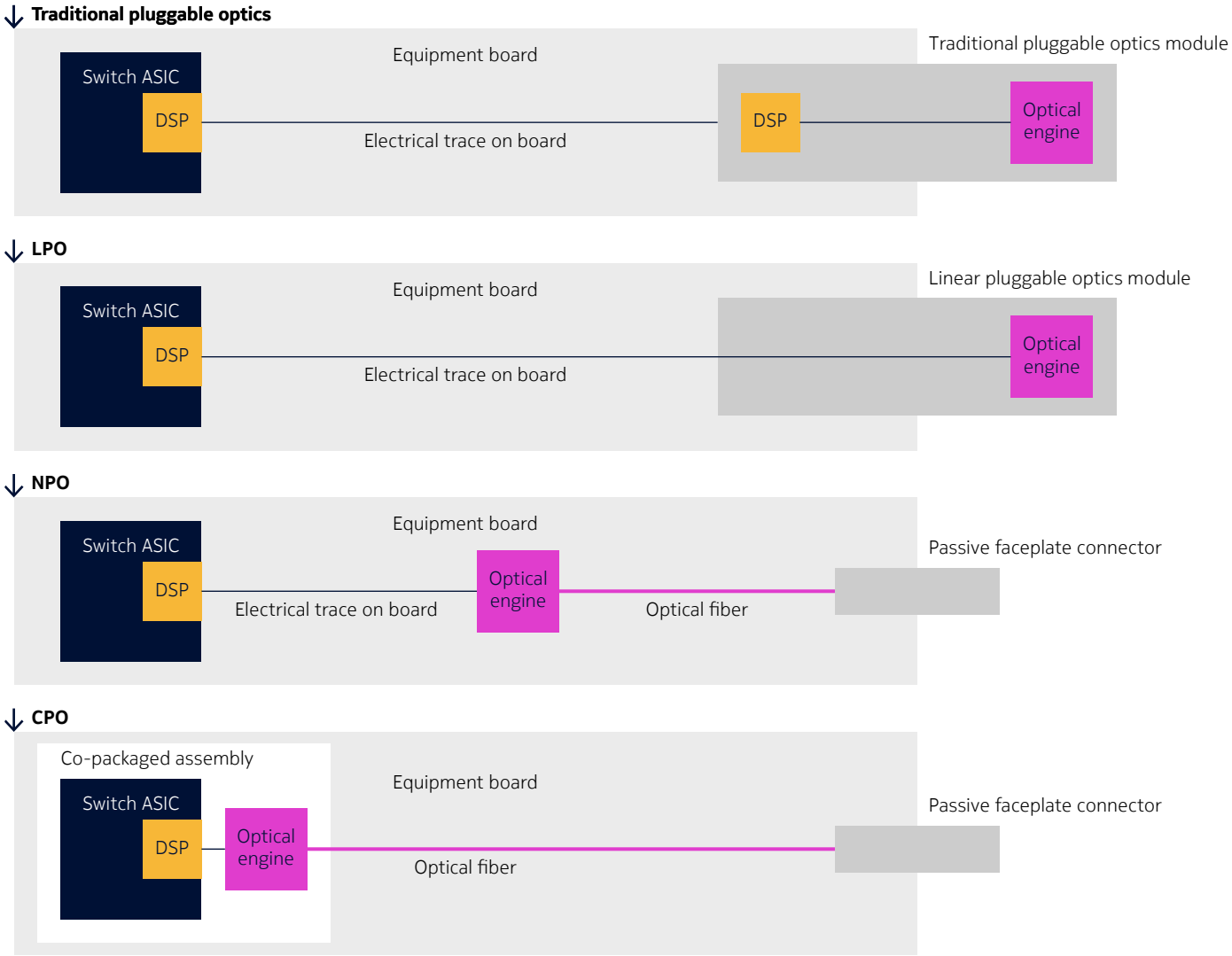
Additionally, in order to make an alternate standard solution, Ultra Ethernet Transport standards are being defined to provide a better alternate to InfiniBand [6].

Optical evolution

As AI data centers continue to grow, so does network connectivity inside the data center. This is resulting in the evolution of networking and optical solutions to help address the need for higher speeds and lower energy requirements.

To alleviate power inefficiencies in AI data centers, particularly in high-bandwidth interconnects emerging with the new GPUs (>800 Gbps), optical technologies like CPO, near-packaged optics (NPO), and linear drive pluggable optics (LPO) are emerging as transformative solutions [7, 8]. Many chip vendors like Broadcom and NVIDIA are investing especially in the CPO solution.

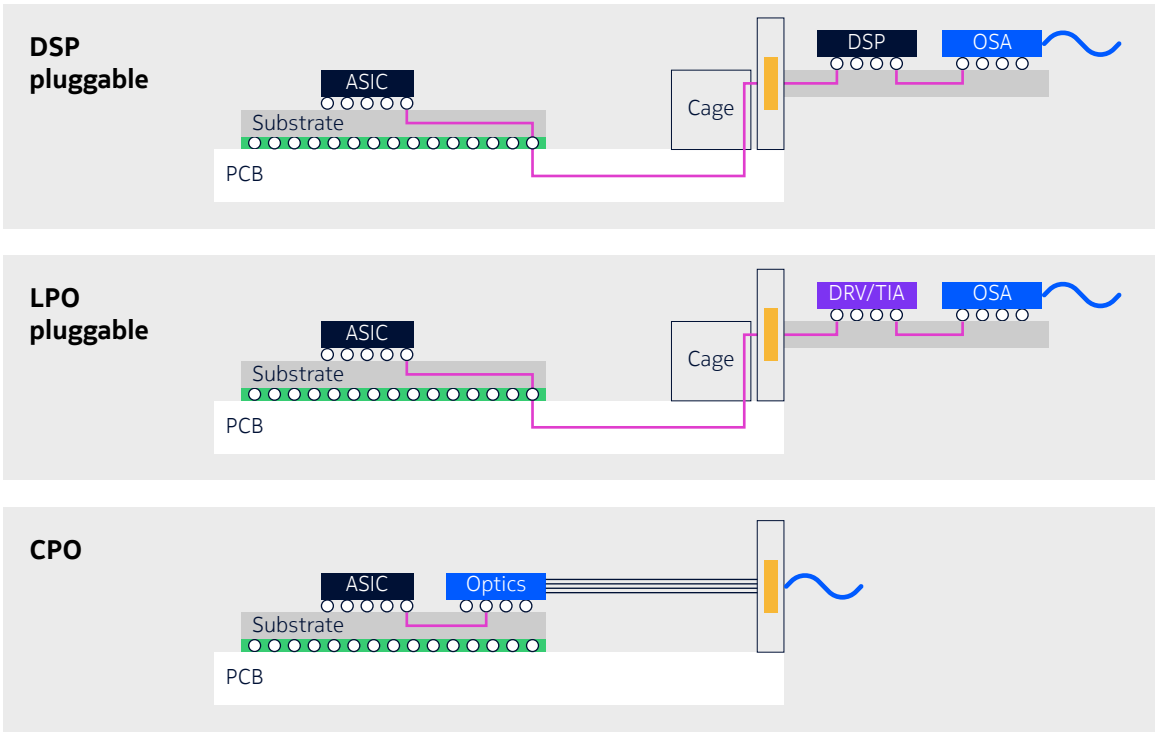
Figure 11. Comparison of optics and evolution (Source: CignalAI)



In a typical data center switch fully loaded with pluggable optics, half of the power is consumed by the pluggable optics. These traditional pluggable optics rely on digital signal processors (DSPs) and electrical interfaces, which consume significant power— up to ~30 W per module at high speeds. CPO and LPO

offload the need for the DSP in the optical transceiver. In the case of CPO, the optical engine is directly built on the switch application-specific integrated circuit (ASIC). This ensures a reduction in latency, enhances bandwidth density, and lowers energy consumption, making them efficient for high-speed, lower latency AI environments.

Figure 12. CPO and LPO removes the DSP in pluggable driving



Source: Broadcom

Summary

The rapid growth of large language models (LLMs) and AI-driven workloads is revolutionizing data center (DC) networking, demanding unparalleled scalability, low latency and energy efficiency. Training LLMs involves massive datasets and computational challenges, requiring parallelization across thousands of GPUs. This necessitates advanced backend networking, leveraging protocols like RoCEv2 and InfiniBand for direct GPU-to-GPU communication, minimizing CPU bottlenecks. Additionally, new scale-up communications are going to be a large part of the AI DC network, and new protocols like Ultra Ethernet might become the key standard alternatives to InfiniBand.

Abbreviations

800GbE	800 Gigabit Ethernet
AI	Artificial intelligence
ASIC	Application-specific integrated circuit
CPO	Co-packaged optics
CPU	Central processing unit
DC	Data center
DSPs	Digital signal processors
EU	European Union
GRH	Global routing header
GPU	Graphics processing unit
GW	Gigawatt
HGX	(NVIDIA high-performance computing system)
HPC	High-performance computing
IP	Internet protocol
L3	Layer three
LLMs	Large language models
LPO	Linear pluggable optics
MW	Megawatt
NPO	Near-packaged optics
NVLink	(NVIDIA's inter-GPU communication protocol)
PCIe	Peripheral component interconnect express
PFC	Priority flow control
POD	Point of delivery
QPID	Queue pair ID
RDMA	Remote direct memory access
RoCE	RDMA over converged Ethernet
RoCE2/RoCEv2	RDMA over converged Ethernet version 2
TCP/IP	Transmission control protocol/Internet protocol
UDP	User datagram protocol

References

[1] Epoch.ai, “Investigating the trajectory of AI for the benefit of society.” [Online]. Available: https://epoch.ai/

[2] J. Saul, L. Nicoletti, D. Pogkas, D. Bass, and N. Malik, “AI data centers are sending power bills soaring,” Bloomberg Technology: The Big Take, Sep. 29, 2025. Available: https://www.bloomberg.com/graphics/2025-ai-data-centers-electricity-prices/

[3] Epri, “Epri home page.” [Online]. Available: https://www.epri.com/

[4] Howard, “An overview of NVIDIA NVLink,” FS web site, Feb. 19, 2024. Available: https://www.fs.com/blog/fs-an-overview-of-nvidia-nvlink-2899.html

[5] InfiniBand Trade Association, “InfiniBand trade association releases updated specification for remote direct memory access over converged ethernet (RoCE),” InfiniBand Press Release, Sep. 16, 2014. Available: https://www.infinibandta.org/infiniband-trade-association-releases-updated-specification-for-remote-direct-memory-access-over-converged-ethernet-roce/

[6] Ultra Ethernet Consortium, “Ultra Ethernet Specification v1.0,” Jun. 11, 2025. [Online]. Available: https://ultraethernet.org/wp-content/uploads/sites/20/2025/06/UE-Specification-6.11.25.pdf

[7] Light Counting, “July 2025 highlights from the 1st conference on co-package optics (CPO),” Lightcounting Research Note, Jul. 2025. Available: https://www.lightcounting.com/research-note/july-2025-highlights-from-the-1st-virtual-conference-on-co-packaged-optics-cpo-406

[8] M. Tan, J. Xu, S. Liu, et al., “Co-packaged optics (CPO): Status, challenges, and solutions,” Front. Optoelectron., vol. 16, no. 1, p. 1, 2023. Available: https://doi.org/10.1007/s12200-022-00055-y

About Nokia

Nokia is a global leader in connectivity for the AI era. With expertise across fixed, mobile, and transport networks, we’re advancing connectivity to secure a brighter world.

Nokia OYJ
Karakaari 7
02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000

Document code: SR1828800EN (December) CID215206