Network the cloud

The critical role of the network in cloud evolution

White paper

The cloud is undergoing a significant transformation, driven by AI and data sovereignty. At the heart of this evolution are data centers, which are proliferating, growing in size and complexity, and becoming increasingly distributed. As data centers expand, the network emerges as a vital component, rivaling the importance of compute, energy and cooling systems.

This white paper delves into the impact of new business- and mission-critical demands on the network, both within and between data centers. It examines the critical role of the network in cloud evolution and the need for a seamless network-cloud continuum that delivers unparalleled performance, scalability and security.

Contents

Introduction

The cloud has become an indispensable part of our daily lives, and it's growing. By the end of this decade the cloud is expected to be a $2 trillion business with a growth rate of over 20% every year[1].

In concept, the cloud enables users to access applications, store data or execute tasks from their devices, but the actual processing, storage and management of those resources occur remotely on servers, often managed by a third-party provider. Behind this massive industry is a vast network of data centers spanning the globe, which can be thought of as the organs of the cloud, performing core functions like compute, storage and processing.

For scaling and economic reasons, the initial success of the cloud was in centralizing servers in large data centers, a model still reflected in the market dominance of the hyperscalers. This is evolving as applications, use cases and regulation demand higher performance, security and privacy. Rather than a limited number of hyperscale-sized data centers, the cloud is becoming a constellation of global, national, regional, metro core, edge and on-premises data centers. Cloud-based applications are disaggregated and distributed, with different functions being performed where necessary to meet the service requirements of the use case or application.

We call this the cloud continuum, and it is both being made possible by advances in network architectures and technologies and is also driving new advances. Cloud interconnection is now not simply a question of connecting large data centers but a matter of connecting a continuum of computing resources from the macro to the local seamlessly with end-to-end continuity, flexibility, scalability, reliability and security.

In what follows, we will explore what is driving the evolution of the cloud and the network capabilities that will be necessary to further this evolution.

1 Source: Goldman Sachs Research

The key drivers of cloud evolution

The cloud evolution is marked by a proliferation of data centers, with smaller, more distributed facilities emerging closer to the sources of data generation or consumption.

Figure 1. Data center locations worldwide

According to Data Center Map, there are more than 10,000 data centers worldwide, as illustrated by Figure 1. This number is growing every day, driven lately by two primary factors: artificial intelligence (AI) and data sovereignty.

AI

According to research from consultancy firm McKinsey, by 2030, 70% of the world's data center capacity will be used to manage AI workloads. Hyperscalers like Amazon, Google, Microsoft and Meta are building AI factories, specialized data centers designed to process and manage the vast amounts of data required to train and deploy AI models. These companies are poised to invest over $315 billion in 2025 to expand their global data center footprints.

Many other major players are also investing heavily in building out AI infrastructure, highlighting the growing importance of cloud-based AI solutions in the industry.

Communications service providers (CSPs) are evaluating potential business opportunities in AI infrastructure, exploring ways to monetize their existing assets and deliver new AI-powered services to customers.

Enterprises seek to build AI workloads on-premises to improve security, reduce latency, and increase control. However, the AI trend and the extreme demand for AI workloads are driving the cost of accelerated compute through the roof as they become increasingly scarce. This is giving rise to an entirely new category of cloud service provider called neocloud, such as CoreWeave and Nscale, who are opening up compute capacity on demand for running enterprise models at scale. For many enterprises this will be the only way to get their hands on the compute they need for training models and running inference.

Data sovereignty

The growing importance of data sovereignty is driving the repatriation of data and the buildout of new data centers, as governments increasingly scrutinize data security risks from non-national providers and update regulations to mandate local data processing and storage and ensure data is confined within their own borders.

There is also a regional concern that without government investment in the development of AI that the economic benefits of this technology will not be realized. An example of this is the European Commission investment in the European High Performance Computing Joint Undertaking (EuroHPC JU) where the EC will invest billions of euros in building 13 AI factories within the region (Figure 2).

Figure 2. AI factories built as part of the EuroHPC JU

This trend is further fueled by the escalating threat of cyber-attacks, which is prompting enterprises to keep their data closer to their core operations, reducing the risk of data breaches and unauthorized access.

Beyond security and privacy, cost can also be a driver for repatriation. Enterprises have over the last decade moved significant portions of their data center activities to the hyperscalers, but some are expected to repatriate and re-balance the mix of private and public clouds. Many have failed to realize the anticipated cost savings, or they have a clearer picture of what data center capacity they need and no longer require the flexibility of the pay-as-you-grow public cloud, which is best suited to the introduction of new services rather than established ongoing operations.

Cloud mix and match

As organizations grapple with the competitive imperatives to digitalize their operations, repatriate their data or unlock the power of AI, they confront a host of different choices.

The first consideration is how to balance the costs of building their own compute infrastructure versus using public clouds. Most organizations will pursue a 'cloud-also' approach with a mix of private and public clouds, known as hybrid cloud.

Where the organization has definitively established what they need in terms of cloud resources, so that they don't anticipate having to scale up or down in a hurry, then it makes economic sense to do it in-house.

The main caveat is that there are other constraints. For instance, compute or real estate availability might make it more advantageous to go with a public cloud service. Depending on what those constraints are, the organization then must choose from a wide range of providers including hyperscalers, hosting providers, neocloud providers, colocation providers, internet exchange providers (IXPs) and CSPs. This will come down to the requirements of the applications and use cases that the organization needs to pursue and how best to reach and connect the various places needing service to central and edge compute and storage assets.

Another consideration that affects the private versus public cloud choice relates to the organization's workforce. Running private clouds might not be in a business's wheelhouse. As well, workforces are going through dramatic demographic shifts in many developed countries, which may be affecting an organization's ability to operate as they have in the past. They must consider not only how they can replace the expertise of those retiring, but the operational skills of newer generations. Even something as seemingly trivial as an aversion to using command line interfaces (CLIs) or an expectation that mobile devices be part of operational processes can limit a company's ability to hire younger talent.

The AI data center hot topics

When it comes to building or planning new data centers, conversations tend to revolve around a few key areas: determining the required compute capacity, ensuring a reliable and sufficient energy supply, implementing effective cooling technologies, and selecting the optimal location. These topics are currently dominating the discussion among data center architects, consuming a significant amount of their time and effort, particularly amongst those who are building the AI factories.

Compute

At the core of the cutting-edge data centers are graphics processing units (GPUs), the tiny silicon chips that are driving the AI revolution. Originally designed for video games by companies like NVIDIA, GPUs have proven to be the perfect solution for running the complex calculations that power AI. Their unique architecture allows them to process vast amounts of data in parallel, making them an essential component of modern AI systems.

To maximize their processing power, GPUs are densely packed into specialized computers, creating a new generation of supercomputers. These behemoths can contain up to 100,000 chips, all working together in harmony to deliver unprecedented levels of computational power. This innovative design enables data centers to tackle even the most demanding AI workloads, from training complex models to running sophisticated simulations.

Energy

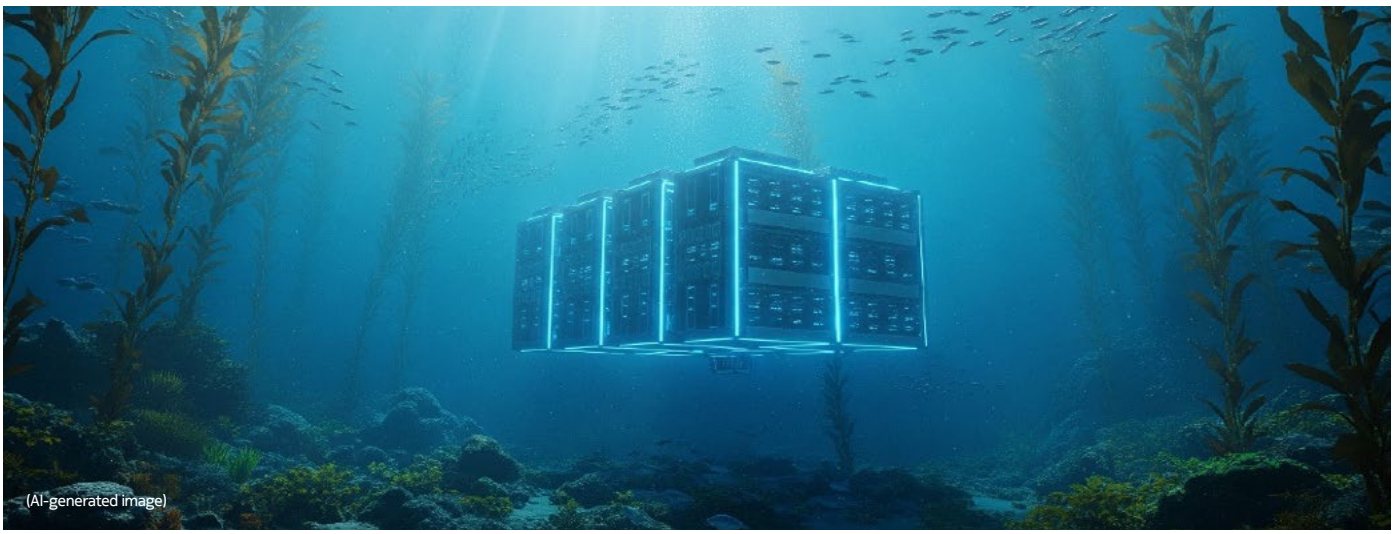
As data centers continue to grow in size and complexity, power availability becomes a critical concern. In the AI factory, a single GPU consumes approximately one kWh of power. To put this into perspective, this is roughly the per capita hourly energy consumption in developing countries like Egypt, Vietnam and Peru.

Multiply that by hundreds or thousands, and the result is an unprecedented power density in a remarkably small space. NVIDIA's upcoming Rubin Ultra NVL platform, for example, integrates 572 GPUs in a single rack that requires a staggering ~600 kW of power. Some in the industry are even anticipating the imminent reality of one MW racks.

These developments signal a dramatic rise in the energy required to fuel modern data centers. In the US, data center power usage is projected to triple by 2028, reaching up to 12% of national electricity consumption, up from 5% today[2].

Cooling

Most of the energy injected into supercomputers is converted into heat. Heat dissipation has become a significant challenge in data centers. To address this, air-based cooling systems are no longer sufficient,



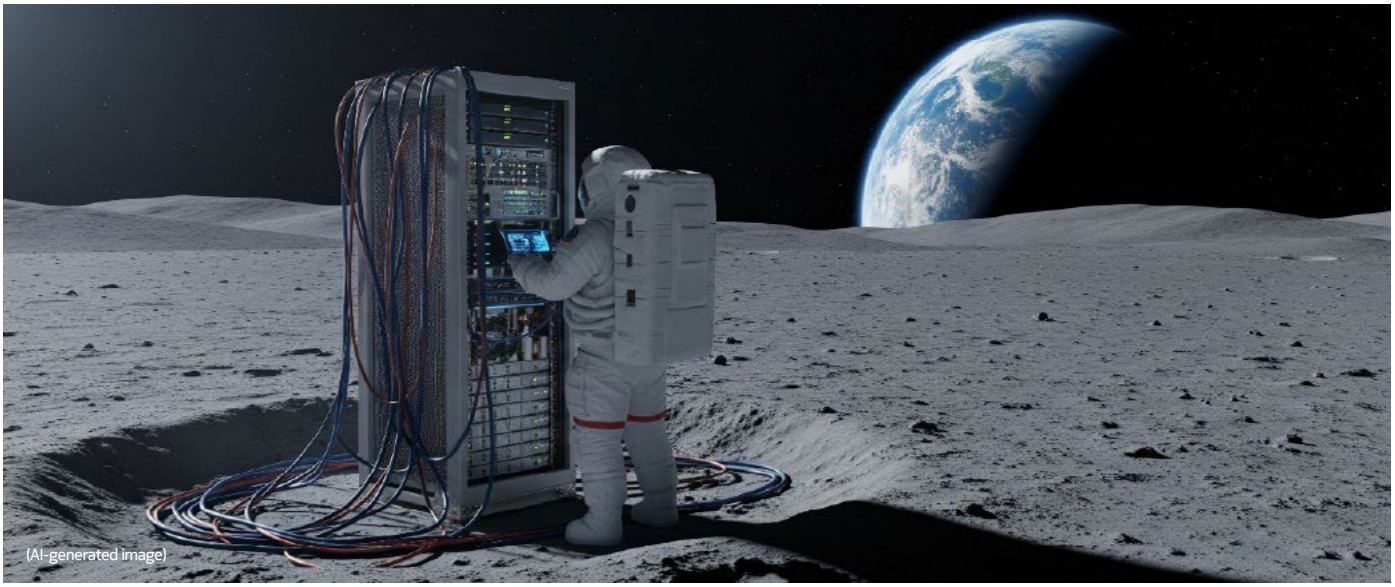
2 Source: U.S. Department of Energy, Dec 2024

and liquid cooling has become a necessity. However, this approach also presents a new set of challenges, including water scarcity and increased water consumption.

In fact, most large-scale data centers now rely on water-based cooling systems, which can consume millions of liters of water per day. To mitigate this issue, some companies are exploring innovative solutions, such as immersing supercomputers in water. For example, the company Subsea Cloud provides containers designed for 25+ years of service to build subsea data centers, while Chinese companies are building data centers under the ocean to utilize seawater as a natural coolant.

Location

Even though the quest for efficient cooling has led some companies to install data centers in the ocean, location is not just about finding a cool spot. Beyond availability of energy and cooling, geopolitical concerns, wars, sabotage, natural disasters, and other risks must be considered when selecting the right location for a data center.



In fact, some companies are taking location to new heights—literally. A company named Starcloud proposes deploying data centers in space, where the vacuum allows for efficient cooling, and abundant solar energy is available. Former Google CEO Eric Schmidt explained that putting data centers in space is the only way to answer the ballooning energy needs of AI data centers by harvesting solar energy directly in space. And if that's not enough, a company named Lonestar Data Holdings is planning to put data centers on the moon.

The network: The cloud's best-kept secret

Regardless of their location—whether under the ocean, on the earth's surface, or even on the moon—data centers require a vital connection to function. Yet, the network that enables this connection often flies under the radar, working behind the scenes to deliver seamless performance.

We tend to take it for granted, but the truth is that the network is just as crucial as compute, energy, cooling, and location in the grand scheme of cloud infrastructures. As organizations invest heavily in building data centers, it's essential that they don't overlook the network's role in optimizing their

performance. If they do, it could become a major bottleneck, causing expensive compute resources to sit idle while they wait for data to be transmitted between them.

Just as the network is necessary for the existence of the cloud in the first place, the evolution of the cloud completely depends on the evolution of the network. Early IPTV services were first made possible by the ubiquity of broadband access (DSL and cable) as well as advances in optical transport networks such as wave division multiplexing (WDM) in the late 1990s. Notably, with the addition of content delivery networks (CDNs), it became possible to deliver over-the-top (OTT) video streaming content to the home, which disrupted the entertainment industry.

The next advance in networks from best-effort IP to MPLS and sophisticated edge routing, made it possible to offer quality of service (QoS), which enabled enterprise applications and as-a-service models from the cloud. Private line T1/E1 services were replaced by virtual private networks (VPNs) and enterprise IT data centers began moving processes into the hyperscale cloud for greater scale and cost efficiencies. One key enabler was the development of specialized network processors (NPUs) that could process traffic at the edge at line speed, allowing for real-time bandwidth management and deterministic performance.

As the cloud evolves, key innovations in IP and optical technologies are driving data center interconnect technologies. For instance, super coherent optical plug-in technologies and digital signal processing (DSP) advances now enable terabit-level speeds for highly scalable network architectures that are more economical and sustainable to operate.

In access networks, the move to 4G and, more recently, to 5G has introduced cloud-native, software-based network architectures that enable broadband connectivity to mobile devices, energy-saving techniques for connecting IoT devices, and highly reliable, low-latency capabilities that can support many new use cases, especially those associated with automation, augmented and virtual reality, and video analytics.

As the cloud continues to evolve, it is imposing new constraints on the network which will necessitate innovative solutions to ensure seamless and efficient cloud connectivity.

Impact of the cloud evolution on the network

Billions of eyeballs, trillions of sensors

While there are still human beings not connected to the internet, that is quickly becoming a thing of the past. It is no exaggeration that content is being consumed in ever-greater quantities by most people on the planet. However, inclusivity remains a challenge we need to overcome. Eyeballs are driving bandwidth (more than ears) as video and games become more accessible, devices more affordable, and broadband more ubiquitous. From cellular to Wi-Fi and satellite networks, we now have the means to satisfy this need in the few places on the earth today where connectivity hasn't reached.

While today the majority of traffic is being driven by billions of eyeballs, potentially, machines talking to machines may be an even bigger driver of network traffic. Cloudflare's data shows that bot traffic accounted for 30% of all HTTP requests[3] in early 2025. From sensors on and in our bodies to deep space probes and everything in between, we are being connected to anything and everything that we see as providing usable information.

3 HTTP requests refers to the number of times a browser, app, or bot asks a server for data (such as loading a web page, image, or API call)

AI agents will become increasingly integral to this vast network of connected devices. As they continuously process contextual information, reason, exchange data with other systems, make decisions, and act upon them in a fully autonomous way, they will generate a massive influx of data that will further fuel the growth of machine-to-machine communication.

These entities have an unlimited capacity to consume and use data, while humans are of course limited in how much they can consume.

The big drivers of this hyper-connectivity are use cases in remote healthcare, finance, security and surveillance, industry-specific applications (Industry 4.0), smart services, and environmental monitoring and sustainability. Sensors gather data that is analyzed by AI and machine learning programs, and decisions and actions are either presented in summarized form to human actors or simply performed autonomously, from regulating pacemakers to moving giant ore haulers.

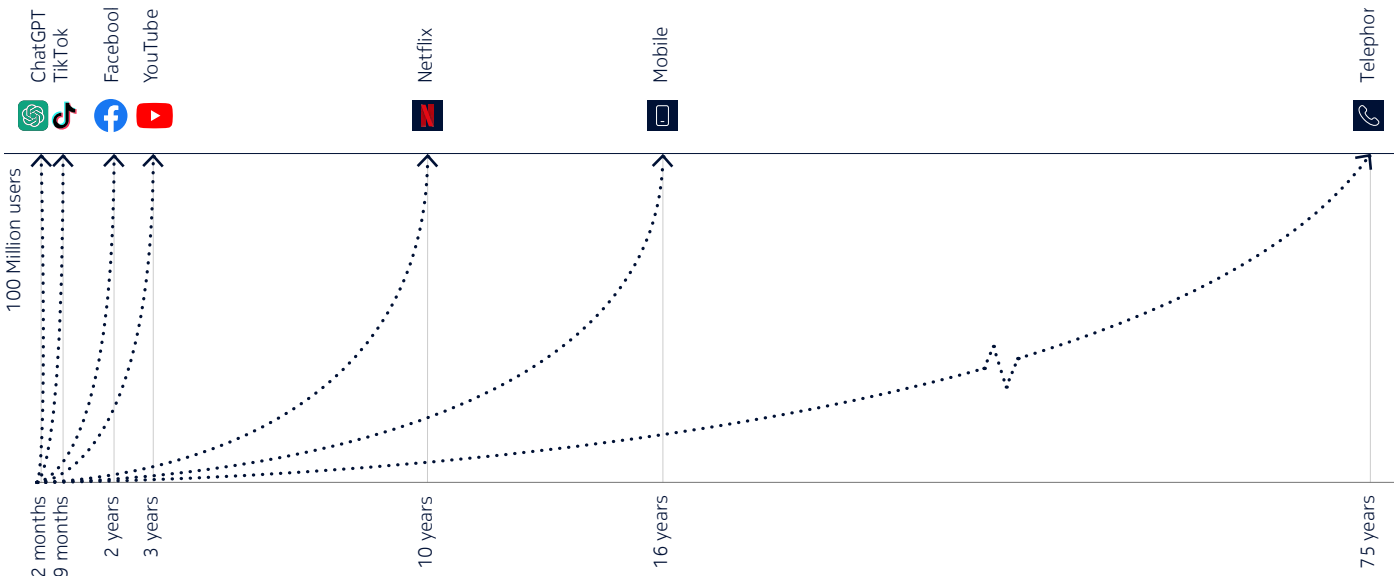
AI driving growth

The most important new factor driving network traffic growth in the next six years will be the enormous demand for AI that is a necessary part of realizing many of these use cases.

We should be in no doubt that the impact on networks is going to pose a challenge for network and data center operators. The first YouTube video, 'Me at the zoo', posted in April 2005, which marked the beginning of the OTT video era, revolutionized the way we consume video content and put unprecedented demands on network infrastructure. The release of ChatGPT in November 2022 may come to be seen as a similar watershed moment.

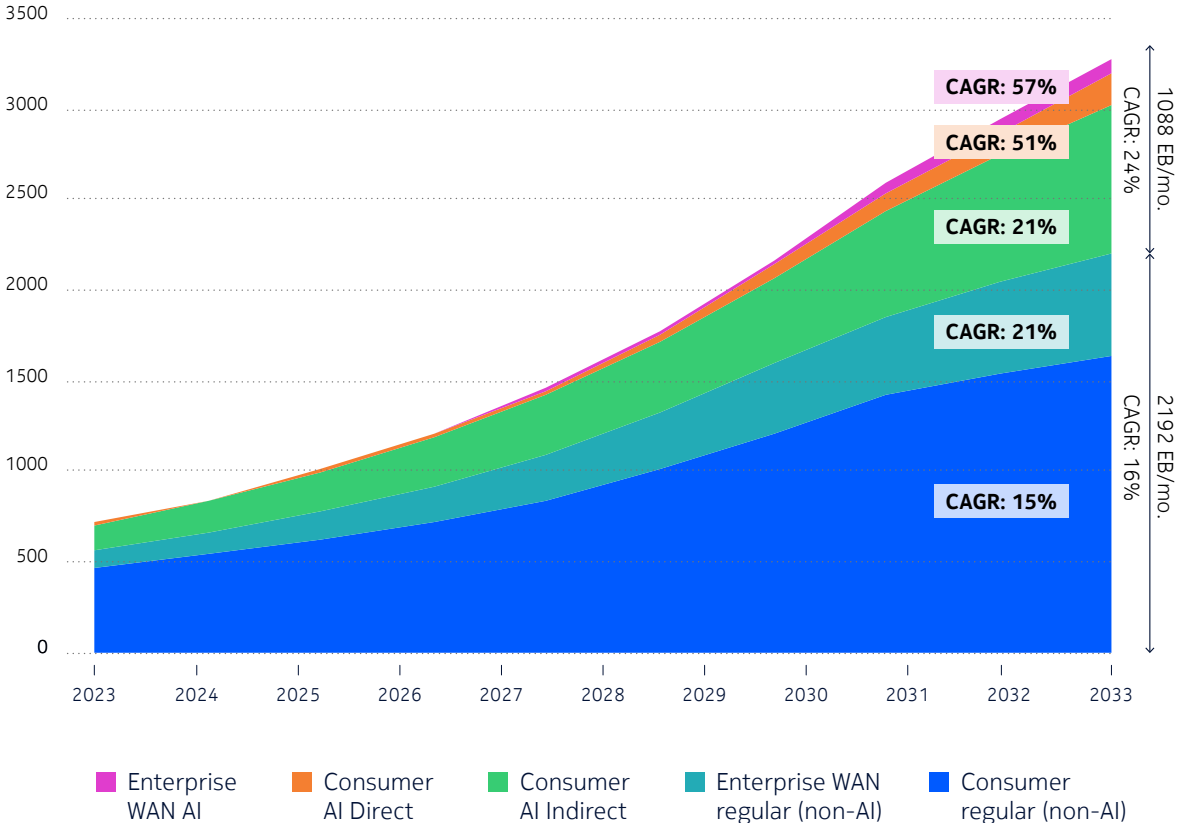
There's a notable difference, however, between the two: while it took YouTube three years to reach 100 million users, ChatGPT achieved this milestone in only two months. This is not an isolated phenomenon; looking back at the history of communication technology, we've consistently seen that each new generation of applications and services is adopted at a faster pace than the last, from the telephone to the internet to social media, as shown in Figure 3. This accelerated adoption rate suggests that the impact of AI on networks will be felt much sooner than it was with OTT video. As a result, the need for upgrades to network infrastructure to support the low-latency, high-bandwidth requirements of AI-driven applications may become pressing much more quickly than anticipated.

Figure 3. Time to reach 100 million users for various applications



According to the latest Bell Labs Global Network Traffic report, global wide area network (WAN) traffic is projected to reach 3,386 EB/mo by 2033 with 1,088 EB for AI traffic alone. This AI traffic is projected to grow at a compound annual growth rate (CAGR) of 24%, as shown in Figure 4.

Figure 4. Global WAN AI traffic projections, EB/month



Consumer AI, including both direct and indirect traffic, will dominate this surge, comprising a substantial portion of overall global WAN traffic:

- Consumer direct AI traffic consists of traffic from user interactions with AI-driven applications, including generative AI, AI-assisted tasks, AI-powered gaming, and extended reality (XR) experiences
- Indirect AI traffic is generated as a result of AI algorithms influencing and increasing user engagement, which reflects traffic growth resulting from personalized AI-driven recommendations across platforms such as video streaming, social media, audio streaming, and online marketplaces.

On the enterprise front, AI traffic will mostly be within the enterprise with lower impact on the WAN. Enterprise direct AI traffic is generated by use cases improving operational efficiency, such as predictive maintenance, autonomous operations, video and image analytics, immersive media applications, AI-enhanced customer interactions, and other enterprise-focused AI solutions.

The network-cloud continuum

Beyond the additional network capacity that will be required to enable new use cases and AI-based applications, the network architecture will evolve. Centralized cloud-based gaming, for instance, is suitable for some types of games, but anything requiring fast reaction times has issues, thus the continued popularity of consoles. Latency simply degrades performance enough that the game winner is often the player with the best connectivity. Consoles will likely be replaced, but edge clouds—not central clouds—will make it possible.

Many AI use cases are also expected to require fast reaction times. AI applications today are fairly limited and mostly text based. Most of the AI traffic is generated to train large language models (LLM) happening in big, centralized AI factories owned by hyperscalers, a few governments, and research institutes and very large enterprises.

While the majority of the AI workloads today are used for training of these LLMs, 60% to 70% of the AI workloads will be used for AI inferencing by 2030[4]. In other words, as AI application adoption grows, the use of the AI workloads will shift from training to making predictions and answering user requests (humans, machines or agents).

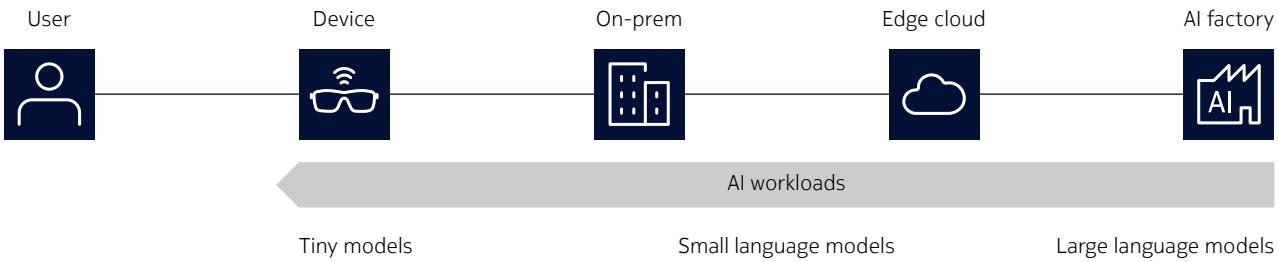
While training involves batch processing of data, inferencing for mission-critical applications involves rapid, real-time processing and analysis of data. These tasks make network speed and low latency vitally important for inferencing.

Inferencing will drive the distribution of the AI workloads closer to the consumers of AI applications. Moving the workloads closer to consumers will reduce the round-trip delay and improve the overall response time of the AI models, hence improving the user experience. It will also reduce bandwidth consumption as data does not have to go all the way up to the central data centers anymore.

This not only reduces the bandwidth consumption by minimizing data transmission over networks, but it also improves privacy and security. This is crucial for applications that require rapid response times or that operate in environments with limited or high-cost connectivity. This is particularly true when large amounts of data are collected, or when there's a need for privacy.

Energy and water supply can also be major drivers for distributing AI workloads closer to sources of power and cooling.

Figure 5. The network-cloud continuum



As these AI workloads get more distributed, they will serve a smaller number of users and will be physically installed in smaller data centers compared to the big AI factories. These low-density AI workloads can run on CPUs instead of costly GPUs, and they can be used for small language models (SLMs) or even tiny models, both of which are less compute intensive. Smaller does not necessarily mean less powerful.

4 Source: McKinsey & Company

While LLMs act as Swiss Army knives with a wide range of knowledge and functionality, their capabilities are relatively shallow compared to smaller, specialized models. Tiny AI models running on devices, as shown in Figure 5, can do specific repetitive tasks faster, cheaper and more reliably than their larger counterparts.

These dynamics operate for robots as well, where on-board processing, although it makes the robot more expensive, also makes it more responsive. Autonomous cars such as those currently being beta-tested by Tesla, for instance, rely on onboard AI inference computers to operate the car.

When more processing power or memory than what the device can provide is needed, more complex processing and reasoning can still happen in the cloud. This is split inferencing where some happens on the device, but more complex things still need to happen in the cloud.

Besides service quality and performance capability, economics will also drive the movement of processing to different levels in the continuum. As data centers expand and grow, cloud arbitrage— dynamically running workloads on whatever cloud offers the best price-to-performance—means that the workload can move anywhere, even mid-process. Using request-response fanouts, the application can determine where the lowest cost compute can be found while still meeting the quality of experience (QoE) required by the end user. This can be a complex calculation and means that some processes are carried out locally, others regionally, and some globally.

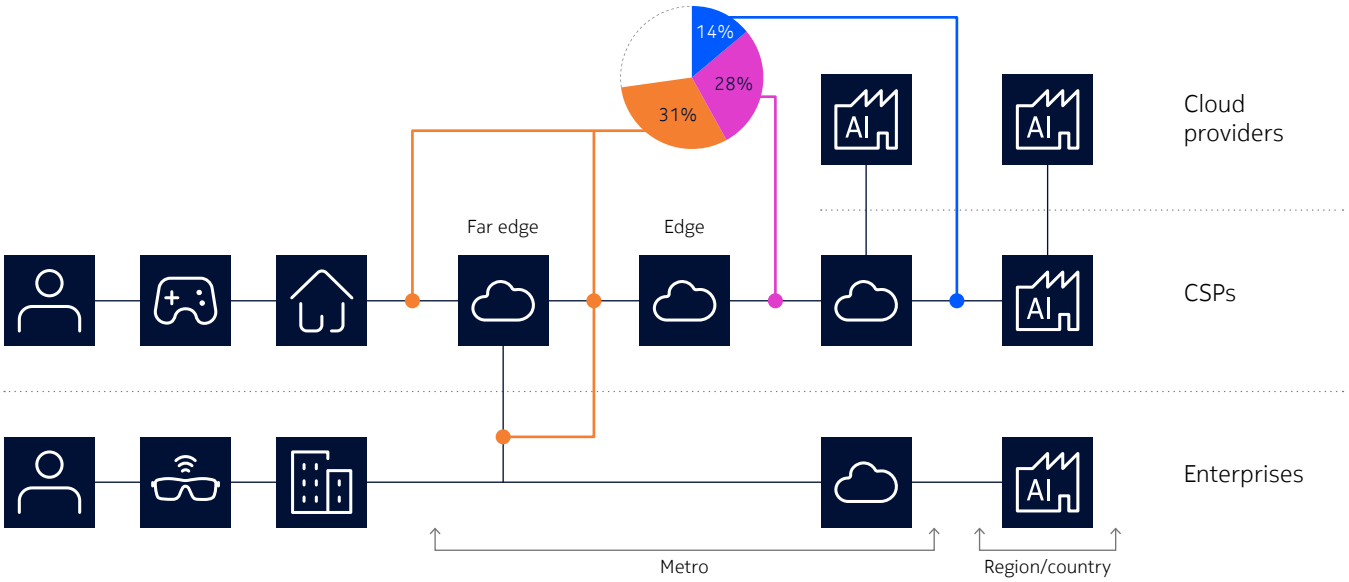
Traffic over data center links in CSP networks

The shift in cloud architectures towards the edge is, in fact, not new. If we look back a few decades, broadcast streaming of content across the best-effort internet was hit or miss. For those that wanted to ensure the quality of the end service, CDNs met the need by caching content locally, reducing the streaming distance. This was an early edge cloud application that was necessary to improve the end user experience without exhausting the capacity of core networks. The streaming cloud solution was good enough to completely disrupt the entertainment console market of the time—VHS and DVD—and businesses like Blockbuster disappeared overnight.

Similarly, to support AI applications, workloads will be placed at various locations—regional core, metro core, metro edge, and even on-premises. The deployment options span hyperscalers, second-tier cloud providers, CSPs, and enterprises, influenced by various considerations, including performance, cost, privacy, security, and resource availability, as discussed in the previous section.

As traffic moves between end users and data centers within the WAN and between data centers, it generates diverse traffic flows across multiple inter-data center links. A single AI traffic flow can potentially traverse several inter-data center links, amplifying the overall traffic volume by a multiplying factor. This dynamic creates an increased demand for efficient routing and inter-data center link capacity to handle the growing complexity of AI traffic patterns.

Figure 6. AI traffic over inter-data center links



As illustrated in Figure 6, from the Bell Labs Global Network Traffic report, the access and aggregation network (orange) of CSPs will bear the heaviest load, roughly 31% of the total AI traffic. The CSP's metro network (pink) is next with 28%, followed by the regional network (blue), with about 14%.

To support this growing demand, significant additional transport network capacity will be needed in the future across CSPs, hyperscalers and enterprise network backbones. This expansion in network capacity will be pivotal in scaling and ensuring the reliable, high-quality delivery of AI traffic across the network.

Building the network-cloud continuum

Building the network-cloud continuum requires a comprehensive approach that delivers connectivity within the data center and interconnectivity solutions that connect data centers, clouds, the WAN and the internet.

Connectivity inside the data center

Within data centers, the network allows the compute resources to work together. This network is under massive strain because of pervasive demand for all forms of content, evolution to cloud-native applications, and the growing dominance of AI/ML-based workloads.

Connecting traditional workloads

First-generation data centers were primarily designed to provide access to content and storage. Early first-generation data center switches evolved from LAN switches, so were fundamentally bridging and virtual LAN (VLAN) systems — a very limited foundation.

In the era of cloud and as-a-service, there has been a shift in data center applications from classic client-server delivery models to microservices-based models with increased communication between servers and processing units. Instead of the emphasis being on north-south traffic patterns, east-west flows are becoming critical, which favors leaf-spine network architectures and Layer 3 (L3) IP fabric designs.

Virtual Extensible LAN (VXLAN) has long been used as an overlay in the data center to provide L2 and L3 connectivity for application workload connectivity within leaf-spine architectures. It is the most widely used tunnelling protocol to overlay L2 connectivity on top of L3 networks as data centers have become more complex.

However, VXLAN alone lacks built-in control plane capabilities, which can lead to scalability and operational challenges in such architectures. Ethernet VPN (EVPN) addresses these issues by using a robust control plane for VXLAN based on Border Gateway Protocol (BGP), enabling efficient address advertisement and minimal resource-consuming network traffic.

In addition, EVPN was designed to support automation and is compatible with model-based programmability—such as YANG models used with BGP and network controllers—which is key to enabling faster, more consistent service deployment, minimizing manual configuration errors, and making the network easier to scale and operate.

For these reasons, EVPN has become the preferred approach to supporting traditional applications on modern data center architectures.

Connecting AI workloads

In the AI factories, the network allows the GPUs to connect and perform efficient training and inferencing. High-speed, lossless networking is crucial for connecting GPUs and servers within a rack or between racks. In this environment, evaluating the performance of AI workloads requires a combination of metrics that captures both efficiency and throughput.

Two key performance indicators (KPIs) for AI workloads are Job Completion Time (JCT) and tokens per second:

- JCT measures the time it takes to complete a job or task, providing insights into the system's ability to process AI workloads efficiently
- Tokens per second measures the system's throughput, indicating how many units of text (such as words or characters) can be processed simultaneously.

By combining these metrics, developers and operators can gain a comprehensive understanding of their AI system's performance, identifying areas for optimization and ensuring that the system can handle the expected volume of user input while meeting latency and throughput requirements.

AI models work with enormous datasets and often involve high-dimensional data (e.g., images, video, or text) that the memory and processing power of a single GPU cannot handle in a reasonable time. The practical way the industry has solved the problem is by using parallel processing. This leads to increased east-west traffic in data centers, rather than the traditional north-south traffic.

Consequently, AI workloads are often deployed on distributed infrastructures like clusters of GPUs. As opposed to traditional workloads, which scale more linearly and with predictable load balancing, AI workloads require dynamic scalability to meet fluctuating demands, especially during training.

To minimize GPU to GPU communication time and achieve high KPIs, optimized network architectures with low-latency, high-bandwidth interconnects will be necessary to efficiently facilitate inter-node communication. By ensuring that the network is fast, reliable and lossless, data centers can optimize the performance of their GPUs, which are the most expensive and critical assets in the AI infrastructure. This, in turn, helps to prevent costly rework, reduce downtime, and maximize the utilization of these valuable resources.

Here are four strategies to ensure high performance of AI workloads:

1. **Use the fastest interconnect**

   Servers for AI workloads are typically designed to support 8 GPUs. These GPUs within a single server can communicate with each other at a far higher speed than they can communicate across different servers through a peripheral component interconnect (PCI) bus. As an example, NVIDIA NVLink5.0 technology used to communicate inside a server is 14 times faster than PCIe5.0, the latest generation of the PCI express interface. Thus, prioritizing intra-server communication maximizes the speed of data flows.

2. **Use the shortest path**

   When GPUs need to be connected across different servers, it's essential to design the network path with the shortest possible distance. This approach gives rise to the concept of "rail-only" or "rail-optimized" designs, where the architecture is configured to ensure that the minimum distance between two GPUs is a single hop through a leaf switch, thereby reducing latency and improving overall system performance.

3. **Use all the available network capacity**

   When GPU-to-GPU communication requires multiple hops, it necessitates traversing a network path. To maximize efficiency and minimize costs, it's crucial to optimize the utilization of the available fabric capacity by distributing traffic as evenly as possible across all the available leaf-spine-(spine)-leaf paths, ensuring that the network is fully leveraged to handle the increased traffic. Various load-balancing techniques can achieve this, including flow-level load-balancing, dynamic load balancing, and per-packet load balancing.

4. **Avoid congestion**

   Congestion avoidance is paramount, as data loss can lead to significant setbacks, including reverting to the last checkpoint, re-syncing GPUs, and incurring hours-long increases in JCT. Common congestion culprits include network oversubscription, uneven load balancing, and incast. While a non-blocking Clos design can mitigate oversubscription issues (albeit at a cost), other challenges persist, underscoring the need for a comprehensive approach to congestion management.

   All these aspects must be considered when building the AI fabric and choosing the networking technology, including ROCE[5] using standard Ethernet network, Ultra Ethernet Consortium (UEC), Data Direct Connect, or Infiniband.

5 RDMA (Remote Direct Memory Access) over Converged Ethernet

Data center interconnect

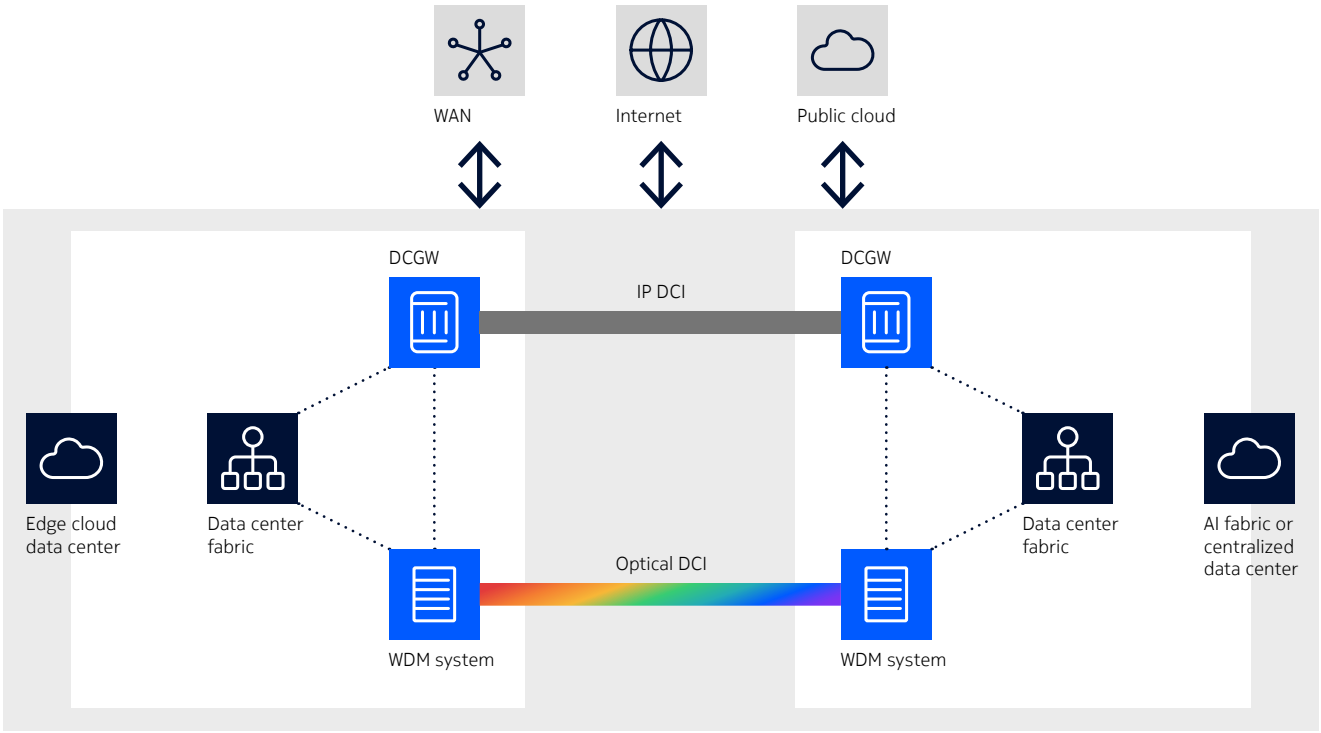
As discussed, the cloud continuum is a mix of central, regional, metro and edge data centers that require interconnect. Data centers also need to connect to the internet, the WAN and public clouds to allow end users to access the data.

As AI-based applications become more popular with consumers and business, the role of data center interconnect (DCI) will become even more important, especially in scenarios where it needs to deliver:

- Connectivity to AI factories to provide the data needed to build the AI models
- High-speed connectivity across AI infrastructures where workloads are strategically distributed across locations, creating a single, cohesive AI compute fabric from geographically dispersed clusters
- Low-latency connectivity for AI inferencing interactions with end users or things, which often means running use cases in private AI infrastructures at enterprise edge locations where data is analyzed in real time.

There are several high-capacity connectivity options available at the IP and optical layers to achieve these scenarios, as illustrated in Figure 7. The choice of technology depends on factors like distance, bandwidth requirements, latency, security and cost.

Figure 7. Data center interconnect with IP and optical options



IP data center interconnect

IP DCI allows the data center fabric to connect to the outside world across the WAN. By leveraging the WAN, organizations can achieve greater network agility, improved resource utilization and enhanced operational efficiency while maintaining the high levels of security and performance required for today's distributed data center deployments.

Diverse WAN technologies can be used, including IP-based routing networks, Multiprotocol Label Switching (MPLS)-based architectures, and tunnels based on the emerging Segment Routing over IPv6 (SRv6). Supporting a diverse range of technologies enables organizations to build more flexible and agile connections between their geographically distributed data centers.

The transport and service technologies used in the data center and the WAN are typically different, so some level of interworking between services such as EVPN and IP-VPN or Virtual Private LAN Service (VPLS) is often required. This interworking ensures consistent service delivery and simplifies network management across the entire infrastructure. EVPN can act as the glue between the data center and WAN environments by providing a unified control plane that supports both L2 and L3 services, enabling data center fabrics and WANs to speak the same "language".

IP DCI is typically done using a data center gateway (DCGW) that allows seamless interworking of services between the data center and WAN environments. The gateway acts as a comprehensive solution for traffic flowing between data centers or between data centers and external networks. The gateway participates in the data center fabric's EVPN/VXLAN infrastructure by connecting to the spine layer in a manner similar to a leaf switch. The gateway also participates in the MPLS and SRv6 infrastructure. This approach is ideal in scenarios where the data center and WAN are operated by the same administrative entity.

Because of the nature of Ethernet or IP and the various interface speeds of these interconnections, buffering and QoS mechanisms are essential for maintaining optimal application performance across interconnected data centers.

Optical data center interconnect

Fiber optic networks offer superior scalability and bandwidth capacity for interconnecting data centers. This transport can be thought of as the freeway that higher level routing networks use to deliver information. Optical DCI networks can be designed as point-to-point, mesh, and ring topologies depending upon the number of data centers and resiliency requirements. Here again, depending on the distance, speed and capacity requirements, several options are possible.

The goal of this layer is, at the lowest cost, to drive the most fiber speed and capacity while extending fiber reach as far as possible.

For maximum capacity over the fiber optic cable and to get even higher reach, the most common method is to connect data centers through an optical physical layer with a traditional WDM line system and embedded optical transponders. Short-reach, gray pluggable optics are used to connect the data center gateway to the optical transponder in a separate chassis.

In applications where the distance between data centers is in the metro-regional range, pluggable coherent optics can be used directly in the data center gateway itself. These modules integrate high-speed optics and coherent detection within a compact, plug-and-play form factor. This enables routers to connect directly at high speeds over moderate link lengths without external optical transponders.

New requirements for data center networking

As data centers evolve to meet new demands, traditional networking approaches are being re-evaluated to ensure they can efficiently support more stringent requirements:

- More business- and mission-critical applications require ultra-low latencies coupled with high reliability, requirements that are more demanding than legacy networks were designed to address
- Increased scalability, as AI applications often use algorithms that can trigger a cascade of data requests and responses, leading to rapid traffic bursts that can overwhelm existing networks
- Improved network responsiveness to adapt dynamically to evolving demands using robust and reliable automation—in contrast to traditional networks, which often rely on manual configurations and find it difficult to prioritize and allocate resources quickly and effectively enough.

Lurking behind the cloud evolution and data center expansion trend is the growing prevalence, volume and sophistication of global cybersecurity threats that make the networks potentially vulnerable. It is critical for organizations to protect data integrity against these threats.

In fact, building the network-cloud continuum to deliver on existing and evolving demands requires a visionary approach to networking both within the data center itself and between distributed data centers. By meeting these requirements, networks can provide the foundation for a seamless and efficient network-cloud continuum that enables organizations to unlock the full potential of their data and applications.

Conclusion

The network is the unsung hero of the cloud, enabling the efficient exchange of data between data centers, end-users and applications. If the cloud is the human body and data centers represent the organs, then the network infrastructure is the nervous system. Just as organs need constant, fast communication via nerves and blood vessels to function as a body, data centers rely on networks to act as a single, coordinated cloud.

The cloud and the network have always evolved in tandem. As we enter the age of AI, it remains the case that despite the extreme discrepancy between the investment in compute, power and cooling as compared to the network, the network will still play a crucial gating role in how far AI and the cloud can evolve. Enterprises evolving their digital operations and embracing AI cannot ignore their data center and interconnect networks.

What we are seeing evolving now is a continuum of cloud capabilities from the center to the far edge, which will be a hybrid of private and public. The network that will make this highly distributed, massively interconnected infrastructure possible will need to be extremely fast, reliable and secure because many of the applications running on it will be business- and mission-critical. It will need to be operationally agile, employing automation wherever possible to reduce costs and respond faster to evolving demands. Finally, it must be architected for extreme scale since we are today only at the beginning of AI, and the cloud will evolve in ways that we can barely anticipate.

What's next?

If you're planning to build an infrastructure for mission-critical applications that demand the highest levels of performance, reliability and security, you need a network that can keep pace. Nokia helps the world's most innovative companies build and operate networks that power general-purpose and AI workloads. These companies demand the highest levels of business continuity and need data center switching fabrics that just work.

Nokia is addressing this need with a new mission for data center networks: Human Error Zero. Our commitment to quality and innovation drives everything we do. We aim to eliminate human error associated with software bugs and hardware issues in networking products, along with errors people make in performing network operations tasks.

To succeed in this mission, our comprehensive portfolio provides high-performance seamless network connectivity within and between data centers and ensures unmatched reliability, scalability, efficiency and security for the most demanding environments.

A wide spectrum of network operators who rely on cutting-edge solutions to build data centers that are more resilient and ready for the future trust Nokia—from neocloud providers like CoreWeave and Nscale, to hyperscalers such as Microsoft (and 9 out of the 10 largest hyperscalers), cloud infrastructure specialists like CoreSite and ResetData, leading enterprises including Hetzner and Kyndryl, research networks like Renater and Surf, and telecom service providers such as Elisa and Maxis.

To find out more, please visit nokia.com/data-center-networks/

Abbreviations

AI	Artificial intelligence
CAGR	Compound annual growth rate
CDN	Content delivery network
CLI	Command line interface
CSP	Communication service provider
DCGW	Data center gateway
DSL	Digital subscriber line
DSP	Digital signal processing
DVD	Digital Versatile Disc
EuroHPC JU	European High Performance Computing Joint Undertaking
DCI	Data center interconnect
EVPN	Ethernet VPN
GPU	Graphic processing unit
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
IPv6	IP version 6
IT	Information technology
IXP	Internet exchange provider
JCT	Job completion time
KPI	Key performance indicator
L2	Layer two
L3	Layer three
LAN	Local area network
LLM	Large language model
ML	Machine learning
MPLS	Multiprotocol Label Switching
NPU	Network processing unit
OTT	Over the top
PCI	Peripheral component interconnect
QoE	Quality of experience
QoS	Quality of service
RDMA	Remote direct memory access
ROCE	RDMA over Converged Ethernet
SLM	Small language model
SRv6	Segmented Routing over IPv6
UEC	Ultra Ethernet Consortium
VHS	Video Home System
VLAN	Virtual LAN
VPLS	Virtual Private LAN Service
VXLAN	Virtual Extensible LAN
VPN	Virtual Private Network
WAN	Wide area network
WDM	Wave Division Multiplexing
YANG	Yet-another next generation

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

Document code: 1507750 (July) CID214961