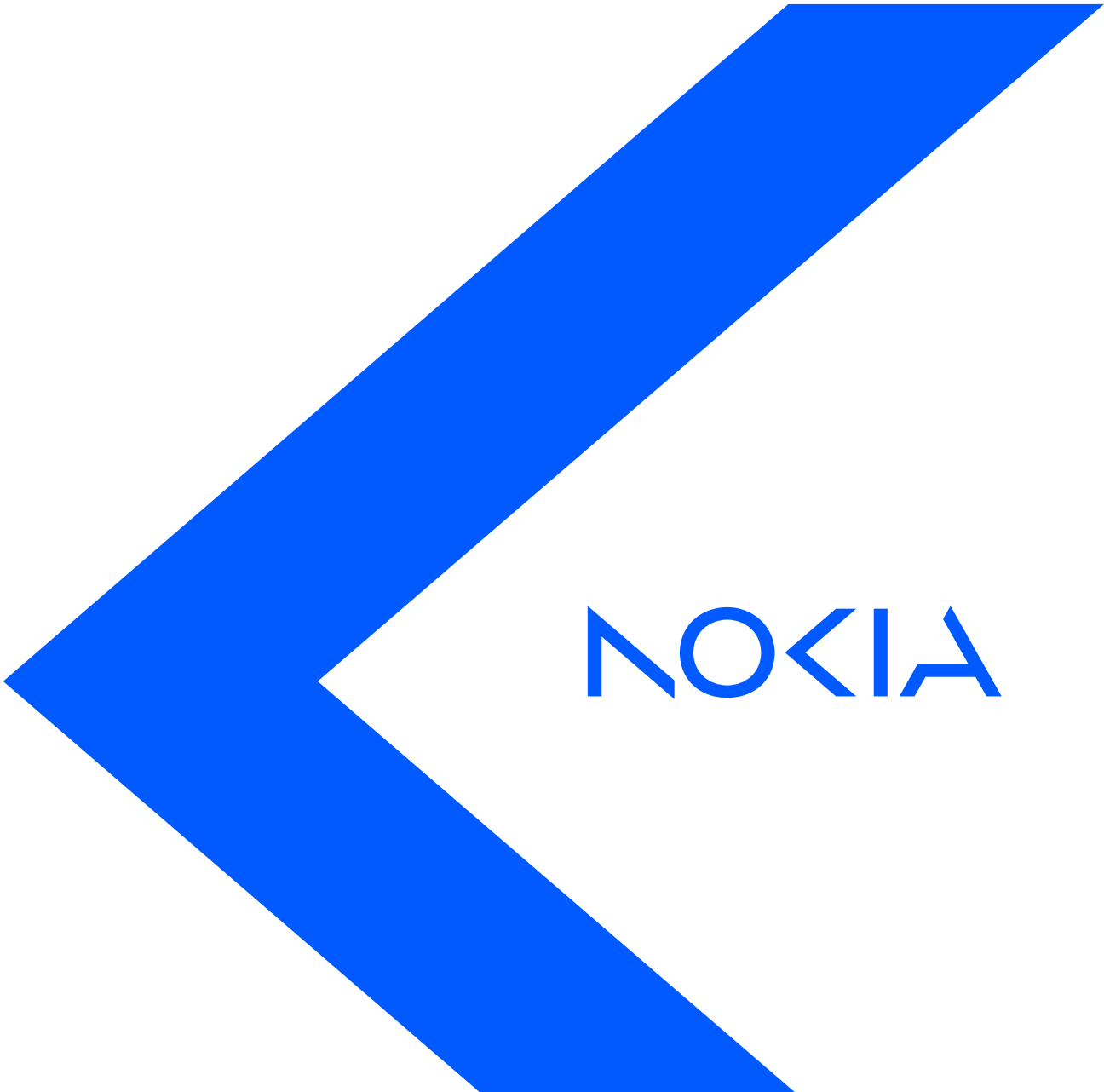
Networking for AI workloads

Reliable fabrics and simplified network operations for all data center environments

Application note



Abstract

Artificial intelligence (AI) technologies are changing the way the world works. Businesses in every sector want to use AI to boost operational efficiency, generate more revenue and revolutionize the user experience. Many have already taken the leap and are exploring AI applications such as natural language processing (NLP), outcome prediction, visual analysis and personalization.

To get the most from these applications, businesses need networks that can efficiently handle the compute- and data-intensive nature of AI workloads and complete jobs in the shortest amount of time. This white paper explores the unique characteristics of AI workloads, the key components of AI infrastructure and the factors that organizations need to consider as they evolve their data center networking for AI workloads. It describes how the comprehensive Nokia Data Center Fabric solution can help them implement high-capacity, lossless infrastructures that are ready to meet the demands of any current or future AI applications and use cases.

Contents

Introduction

Artificial intelligence (AI) has become a mainstream topic in the technology landscape and will continue to play a dominant role in our daily lives going forward. AI and machine learning (ML) workloads harness the power of modern accelerated computing, storage and networking to learn and interpret data, make decisions and enhance problem solving.

While these technologies are still in their early days, AI and ML will continue to transform the way industries and businesses work. They promise to help improve operational efficiency and foster innovation while providing new business opportunities, generating new revenue streams and revolutionizing the user experience across a range of sectors.

The applications and use cases for AI continue to increase at a rapid pace. The following list, while not comprehensive, helps provide a view of the possibilities that may be explored with AI/ML technologies.

- Natural language processing (NLP): The arrival of ChatGPT was a turning point for the AI/ML space and a key validation of the NLP use case. NLP improves the end user experience and helps enhance communication for applications such as chatbots or voice-enhanced applications.
- Outcome prediction: The ability to enhance the prediction of outcomes based on analysis of historical data can be beneficial to enterprises across all segments.
- Personalization: The ability to provide customized recommendations based on analysis of user behaviors and preferences can be beneficial to e-commerce based companies as well as social media platforms
- Visual analysis: The ability to analyze and interpret human, machine or process that is visual in nature can enhance applications such as facial recognition, medical imaging and manufacturing quality control.
- High-performance computing (HPC): HPC can provide the scalability and computational power that organizations need to leverage and adopt AI technologies for scientific research, intelligent simulation and modeling applications.

AI workloads differ from traditional workloads because they are more computationally intensive and typically require the exchange of large blocks of data between iterations. AI workloads require HPC and often need specialized computing and processing hardware as well as large-scale storage to manage the stringent needs of workload processing. The networking infrastructure that supports AI workloads plays a critical role in maximizing the utilization of compute resources to achieve the shortest possible job completion times (JCTs) for AI workloads. This paper focuses primarily on networking aspects related to supporting AI workloads.

Characteristics of AI workloads

In the AI world, everything revolves around the concept of a model. Today's well-known models include Open AI's Generative Pre-trained Transformer (GPT), Meta's Llama, Mistral's Mixtral, Anthropic's Claude and Google AI's Gemini. Models can be trained to do any number of things. For example, a large language model (LLM) is designed to process natural language requests from users and provide humanlike responses to text queries.

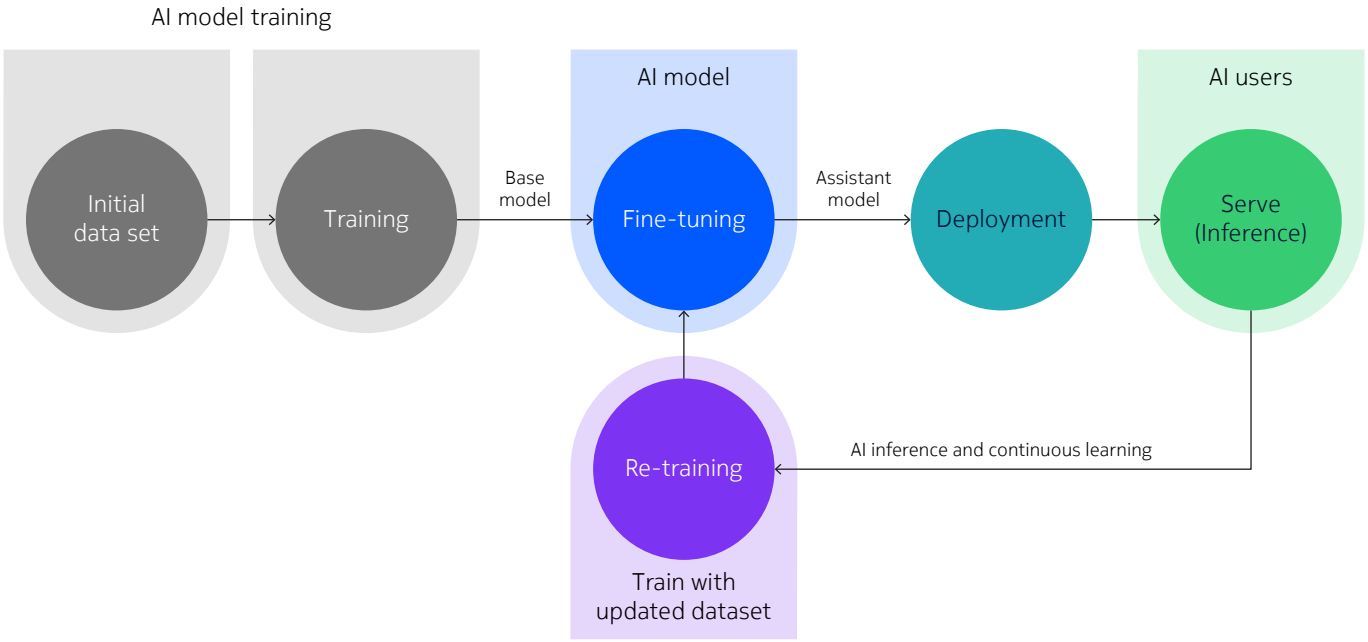
A model can be a general-purpose model or be highly trained for a business outcome. For example, the finance and healthcare segments could have their own training models built around sets of data relating to their particular business cases.

AI workloads are classified into two broad categories based on the tasks they perform: AI training and AI inference.

AI training

Figure 1 depicts the key stages of AI training and AI inference. The key stages of AI training include data collection, model selection, model training and model deployment.

Figure 1: Stages of AI training and AI inference



A training set is all the data that goes into a model. For the biggest models, such as GPT, the training set involves a tremendous amount of data scraped from sources such as the World Wide Web, Wikipedia articles, novels and news sites. This data is used to teach the model how to perform the expected tasks.

LLMs represent a subset of generative AI, with a focus on text generation. Generative AI is not limited to text, and can include outputs such as images, audio, video and code. Additional use cases may dictate the use of specific data. For example, the training set may include historical log files for various Internet of Things (IoT), operational technology (OT) or information technology (IT) devices, application data captured by operations and business support system (OSS/BSS) tools, traffic patterns, security threats, application usage patterns and more.

Model selection involves selecting the type of optimizing algorithms or architecture for the model. During model training, the model learns patterns and the relations between data sets. The training process can take a period of hours to weeks with smaller models and data sets. It can take months or up to a year to train models that are larger or that use data sets that may require greater precision.

The trained model needs to be evaluated and fine-tuned to deliver the required performance and precision. For example, it took 30.84 million GPU hours to train the Meta Llama 3.1 405B model. (Source: https://huggingface.co/meta-llama/Meta-Llama-3.1-405B).

AI inference

After a model is developed, it can be deployed to serve end users at scale, in a process called inference. This process applies the trained model to respond to input data and provide outputs based on the requested queries.

The first step is to deploy the trained model for inference tasks. This involves packaging the model as part of an app or web page or within a software library or as an executable file. Once the trained model is deployed, inference parses the input, adds some preprocessing and feeds the data into the model to produce the desired output.

Based on the application and desired outcomes, the output may need to be generated in real time (e.g., for voice assistants or autonomous vehicle responses). This imposes a requirement for quick response times.

Key components of AI infrastructure

AI infrastructure includes the necessary resources to support AI workloads. AI training workloads are data and compute intensive. They involve the processing of exceptionally large data volumes that may include text, audio, databases ,tables and other types of data. These data sources provide the basis for processing and interpreting the data. AI workloads involve complex mathematical models and operations that require extensive computational power.

AI applications often require parallel processing that spans multiple compute nodes or processing units. This provides better scale and faster processing. The processing of AI workloads is not a one-and-done event. The processed data may need to be modified or improved, and the entire process of learning and performance evaluation is repetitive and iterative.

Table 1 describes the key components of an AI infrastructure.

Table 1: Key AI infrastructure components

Compute nodes	• These are individual servers or nodes that include the compute and memory resources required to support the necessary computational processing tasks and frequently accessed storage. They include central processing units (CPUs), auxiliary processing units (XPUs) such as graphics processing units (GPUs), tensor processing units (TPUs), accelerated processing units (APUs) and language processing units (LPUs), and network interface cards (NICs). The servers also support the large amounts of high-speed memory that AI workloads need to rapidly access and process data. • GPUs are currently the market's most commonly deployed XPUs. Unlike CPUs, which support general-purpose computing and control operations, GPUs are specialized hardware processing units that carry thousands of cores and are better suited for the advanced requirements of AI training and inference workloads. They are deployed for their ability to perform complex mathematical computations at scale and speed. • The number of GPUs in GPU clusters can range from hundreds of GPUs on the low end to tens of thousands of GPUs on the high end. **"xAI's Memphis Supercluster, which recently went live in Tennessee, is equipped with 100,000 GPUs, making it the most powerful AI training cluster in the world." (Source: Data Center Dynamics)**
Storage systems	• Storage systems store data within the AI infrastructure. They are used to store the datasets, parameters and intermediate results related to the model being trained. • Storage technologies may include network-attached storage (NAS), storage area networks (SANs), NVM Express (NVM-e) or distributed file systems such as Hadoop or Google File System.
Networking	• To meet existing and evolving AI needs, compute nodes must be interconnected by high-speed, lossless and low-latency networking. This is essential to reduce JCT, a metric used to measure the time it takes to complete a task, such as training a model or performing an inference operation. • Networking technologies may include InfiniBand and Ethernet for providing reliable, high-capacity interconnects within the AI infrastructure. • Networking provides the access to servers hosting the GPUs to orchestrate learning and inference jobs.
Software	• Software plays a critical role within the AI infrastructure. It typically includes software tools for deploying and managing the compute, storage and networking resources within the infrastructure. Containerized AI workloads may use platforms such as Kubernetes and Docker to manage and orchestrate workloads. • In addition to management and orchestration software, AI workloads require software to support model development as well as distinct phases related to AI training and AI inference tasks.
Power and cooling	• GPUs consume a lot of power and drive the need for a much greater level of cooling for required non-AI workloads. Power consumption and cooling are important deciding factors in the AI cluster design. • The average power consumption of a GPU is five times that of a CPU. GPUs consumed up to 700 W of power at the end of 2023, and that number is inching up even higher in 2024. • French firm Schneider Electric estimates that power consumption of AI workloads totals around 4.3 GW in 2023, which is slightly lower power consumed by the nation of Cypress in 2021 (4.7 GW). **"Power consumption of AI workloads will grow at a CAGR of 26 to 36 percent. By 2028, AI workloads will consume from 13.5 GW to 20 GW, which is more than what Iceland consumed in 2021." (Source: tomsHardware)**
Price	• In current market conditions, specialized servers with GPUs, accelerated storage and high-speed interconnects pose a big challenge to optimizing the budgets, especially for small and medium enterprises (SMEs) and publicly funded universities. • A carefully designed data center infrastructure goes a long way in reducing CAPEX and OPEX and driving down the overall total cost of ownership (TCO) without compromising on the performance required for the workloads.

Networking considerations for AI workloads

Networking plays a critical role in supporting AI workloads. For example, the data center network fabric needs to deliver reliable and seamless connectivity within the AI infrastructure. The network must also consistently deliver the best possible performance for training and inference tasks and operations.

AI training workloads require lossless networks that can provide high capacity, high speed and very low latency. InfiniBand is widely used today, but Ethernet is beginning to gain strong traction because it provides several advantages when it comes to AI workloads.

The networking capacity, speed and latency requirements are less stringent for AI inference than for AI training. Since inference is about serving the AI model to the end users, response times and proximity to the end user become key considerations for network design. “Ethernet is well suited to meeting these requirements, as illustrated by Meta’s use of Ethernet to train Llama 3-450B, which delivered equivalent performance to InfiniBand.” (Source: https://arxiv.org/abs/2407.21783)

It is essential for organizations to implement well-designed network architectures that can meet the challenging reliability, speed, capacity and latency requirements of AI workloads within their price and power budget constraints. This paper will first explore some key technologies and the roles they play in supporting networking for high-performance AI workloads, and then examine some of the architectural models required to achieve the right performance for high-performance workloads and applications.

InfiniBand and RDMA

The InfiniBand architecture[1] emerged in 1999 as an interconnect technology designed for HPC and data-intensive applications. InfiniBand is an industry-standard specification that defines an input/output architecture used to interconnect servers, communications infrastructure equipment, storage and embedded systems.

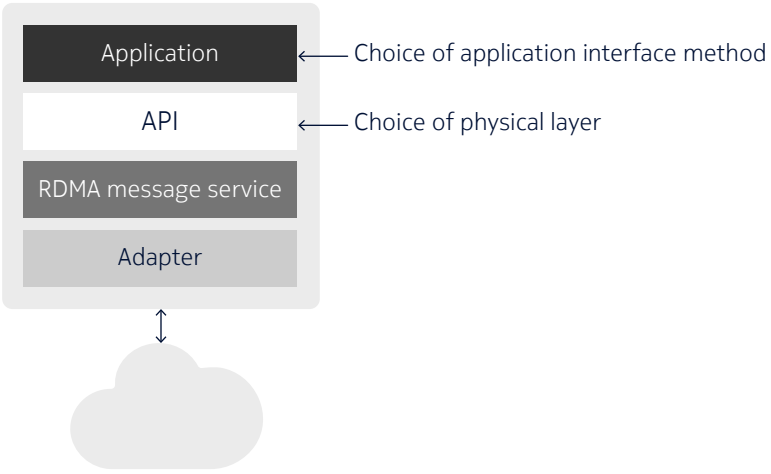
InfiniBand supports Remote Direct Memory Access (RDMA).[2] RDMA technology enables direct memory access from the memory of one server to another without involving either server’s operating system or processor. Instead of using valuable CPU processing time to manage communications between the application and the network, RDMA directly passes data (files, messages, blocks, etc.) between different application memory spaces, eliminating CPU involvement. The InfiniBand Architecture (control stack) is made up of the following parts, as illustrated in Figure 2 below:

- An application programming interface (API) that enables applications to take advantage of RDMA through the RDMA message service
- The RDMA message service, which is enveloped in the RDMA software and provides access to the RDMA hardware
- A Host Channel Adapter (HCA) that provides InfiniBand network connectivity
- Interconnect, which is a network of cabling, switches and routers (InfiniBand).

1 The InfiniBand® Trade Association.

2 “Enabling the Modern Data Center – RDMA for the Enterprise”. InfiniBand Trade Association white paper. 20 May 2019. Retrieved 9 August 2024.

Figure 2: The ‘parts’ of RDMA (control stack)



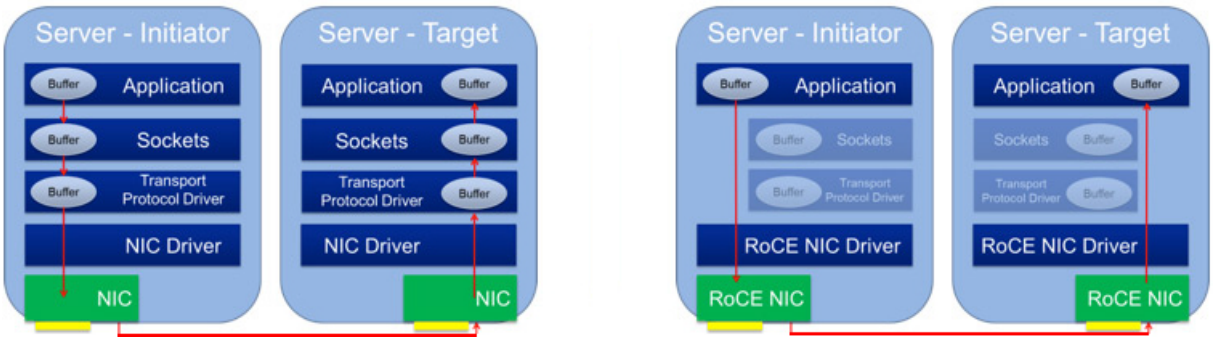
Source: IBTA white paper, 20 May 2019

RDMA over Converged Ethernet

The InfiniBand specification describes two network stack implementations for RDMA technology: RDMA over InfiniBand (or simply InfiniBand) and RDMA over Converged Ethernet (RoCE).

RoCE benefits from the ubiquity of, and advancements made in, IP/Ethernet over the past decades. It places the InfiniBand transport layer inside IP/Ethernet frames, providing the RDMA capability, kernel bypass and other benefits that are not part of traditional TCP/IP. As shown in Figure 3, RoCE offers the ability to directly “read from” or “write to” an application’s memory, in contrast to traditional client-server interactions that use TCP/IP and involve many copies and significant CPU overheads.

Figure 3: Traditional and RoCE data movement

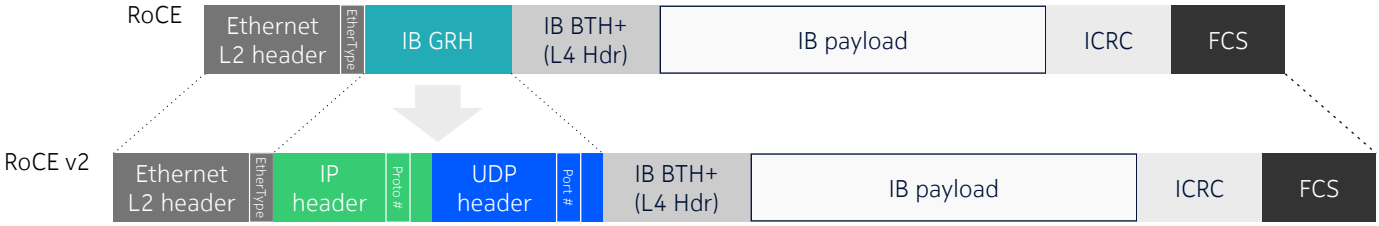


Source: RoCE Introduction

The RoCE specifications have two variants: RoCEv1 and RoCEv2. RoCEv1 was introduced in 2010. As shown in Figure 4, it uses regular Ethernet frames with an Ethertype value that indicates these are RoCE-related headers (global routing and base transport headers). The RoCE frame does not include an IP header, so it can only provide connectivity within a layer 2 Ethernet domain.

RoCEv2 was introduced in 2016 and is an extension to the RoCE specification. It replaces the InfiniBand global routing header (GRH) with IP and User Datagram Protocol (UDP) headers. This allows communication across IP subnets.

Figure 4: RoCE and RoCEv2 frame formats



RoCE and lossless networks

InfiniBand supports link layer flow control, and the hardware dynamically tracks buffer usage.[3] This allows InfiniBand to be lossless because flow control can pause data transmission to preempt buffer overflows.

RoCE has a few limitations compared to pure InfiniBand and needs to address congestion and flow control. Data Center Quantized Congestion Notification (DCQCN) is an end-to-end congestion control scheme for RoCEv2.[4] It is implemented on the server NICs (endpoints) and works in conjunction with Explicit Congestion Notification (ECN) and Priority Flow Control (PFC) features implemented in the data center switches.

- The ECN mechanism monitors buffer usage on the network elements and allows endpoints to be notified of imminent congestion without dropping packets when the buffer usage exceeds a configured threshold. This is achieved with RoCE Congestion Notification Packets (CNPs) sent to the sender endpoint.
- PFC enables drop-free Ethernet fabrics by sending per-priority PAUSE frames to upstream devices. This priority-specific back-pressure mechanism ensures that the network provides lossless transmission.

The correct operation of DCQCN requires balancing PFC and ECN requirements to ensure that PFC is not triggered too early, before ECN can send congestion feedback. While PFC delivers a lossless network, ECN helps achieve maximal network resource utilization. For RoCE-based deployments, it is essential to support the congestion and flow control feature enhancements discussed above. This will help data center teams deliver lossless networks when they use Ethernet instead of InfiniBand for supporting AI workloads.

Architecture considerations for deploying AI infrastructures

Hyperscalers are leading the charge in AI workload deployments. Their main drivers are to gain a competitive advantage and meet demands for AI services and tools that will enable their customers to deploy AI-based models and use cases. Their incumbent roles and ability to deliver cloud-like scale and service options (e.g., pay as you grow and need) will progress naturally as they evolve their AI offers.

Enterprises, communications service providers (CSPs) and cloud providers of all types understand the importance of AI and are looking for ways to apply it to their business transformation initiatives.

3 Grun, P. "Introduction to InfiniBand for End Users". InfiniBand Trade Association white paper. Retrieved 9 August 2024.

4 Zhu, Y et al. "Congestion control for large-scale RDMA networks". SIGCOMM15 paper. Retrieved 9 August 2024.

Implementing AI infrastructures for training is a highly challenging process because it requires special technical expertise and can be cost prohibitive. Some companies may implement their own training infrastructures, while others may prefer to use AI platforms and associated services from large cloud providers. IDC refers to this model as "public AI."[5]

AI inferencing requirements are typically less stringent than AI training requirements. It is likely that more companies will look at a do-it-yourself (DIY) approach to implementing AI inference infrastructures that support their specific business needs. IDC refers to this model as "private AI," which is the use of enterprise data center and AI platforms for actioning enterprise-specific generative AI workflows.

According to IDC, large enterprise IT departments prefer utilizing public AI frameworks for AI use cases that are cost, time, scale and performance efficient to implement on cloud foundation models. IDC also says that inferencing models that need low-latency end-user access are typically run in enterprise edge locations in private AI infrastructures, while batch-mode inferencing and global-scale inferencing models are better suited for implementation in public AI frameworks.

Power consumption is another key consideration for organizations deploying AI workloads. Some massive AI infrastructures may be too large to meet power requirements or legislative constraints within the planned area of deployment. For such scenarios, a model where AI workloads are distributed across multiple locations can help ensure compliance with the required power budgets and constraints.

Back-end and front-end networks

As shown in Figure 5, the back-end network is used for interconnecting high-value GPU resources required for high-computation tasks such as AI training, AI inference or other HPC workloads. The back-end network delivers lossless, low-latency and high-performance connectivity for the AI training compute and dedicated storage resources.

Figure 5: Back-end network

"AI training/inference" infrastructure

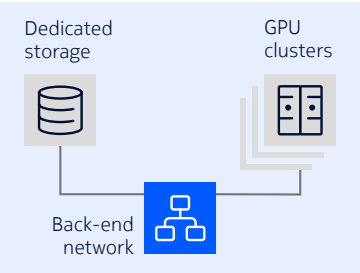
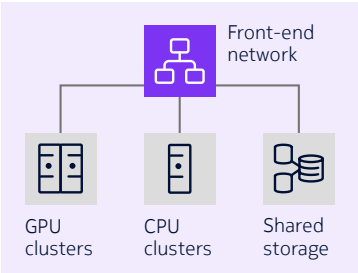


Figure 6: Front-end network

"AI inference" and "non-AI" infrastructure



The front-end network (Figure 6) supports connectivity for AI workloads, general-purpose workloads (non-AI compute) and management of AI workloads. In the context of AI inference, the front-end network supports connectivity for compute and shared storage resources to enable communication with end users and devices.

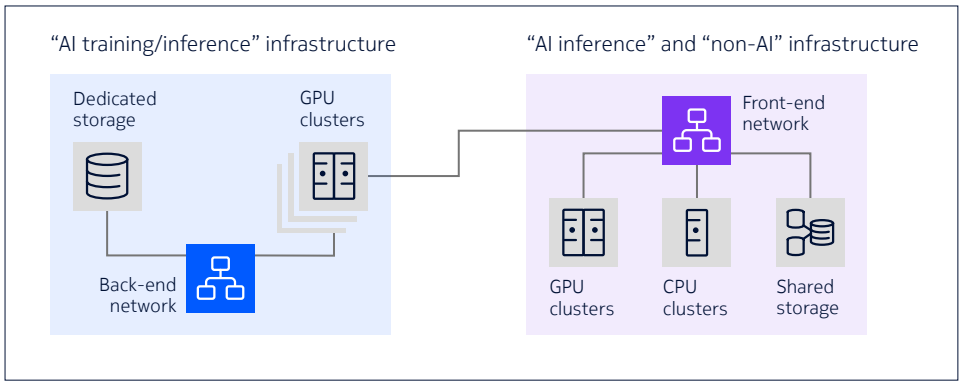
5 Bhagavath, V and Mehra, R. "Multicloud networking will inflect in 2024". IDC Market Perspective. Retrieved 9 August 2024.

Separate vs. converged back-end and front-end networks

Based on its cost and power budgets, an organization can choose to deploy separate front-end and back-end networks as shown in Figure 7, or have a converged design, as shown in Figure 8. In both scenarios, the front-end and back-end networks are co-located within the same data center location.

Figure 7: Separate back-end and front-end networks

Co-located back-end and front-end networks

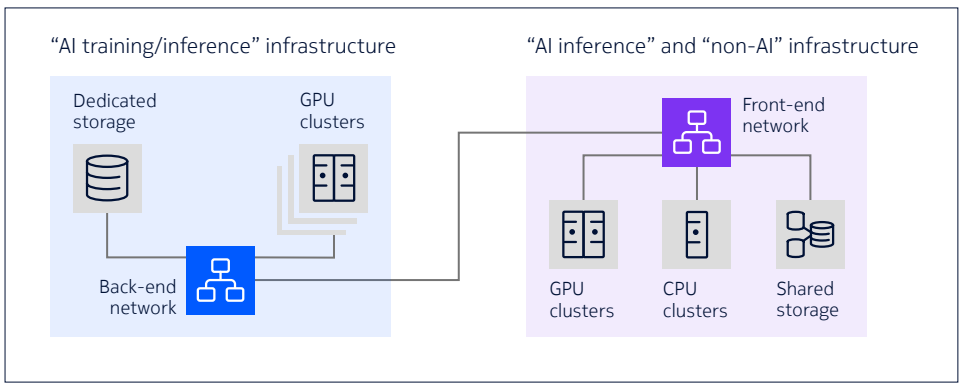


In Figure 7, the back-end and front-end networks implement dedicated and separate leaf-spine architectures. The GPU cluster and dedicated storage are interconnected by a high-speed, non-blocking, lossless, low-latency back-end network. This network enables GPUs to interconnect with each other and read disks with high performance.

The CPU cluster, shared storage and GPU cluster—some AI inference use cases may dictate the need for GPUs, albeit with lower performance criteria—are interconnected through a lower-speed, front-end network that serves that serves AI inference, general-purpose data center, and AI training and inference management workloads. To enable connectivity for AI training and inference management, the GPU cluster (in the left-side AI training/inference infrastructure block) connects to the back-end and front-end networks.

Figure 8: Converged back-end and front-end networks

Co-located back-end and front-end networks



In the converged design (Figure 8), the back-end network connects to the front-end network. The back-end leaf nodes and front-end leaf nodes are interconnected through spine nodes. This lowers CAPEX by reducing the number of NICs required on GPU nodes and reducing the number of leaf and spine nodes and interconnections required.

In the back-end, GPU interconnect is supported by high-speed, typically non-blocking, lossless and low-latency networking. But there can be oversubscription in the front-end network on the links connecting leaf nodes to the rest of the front-end data center network.

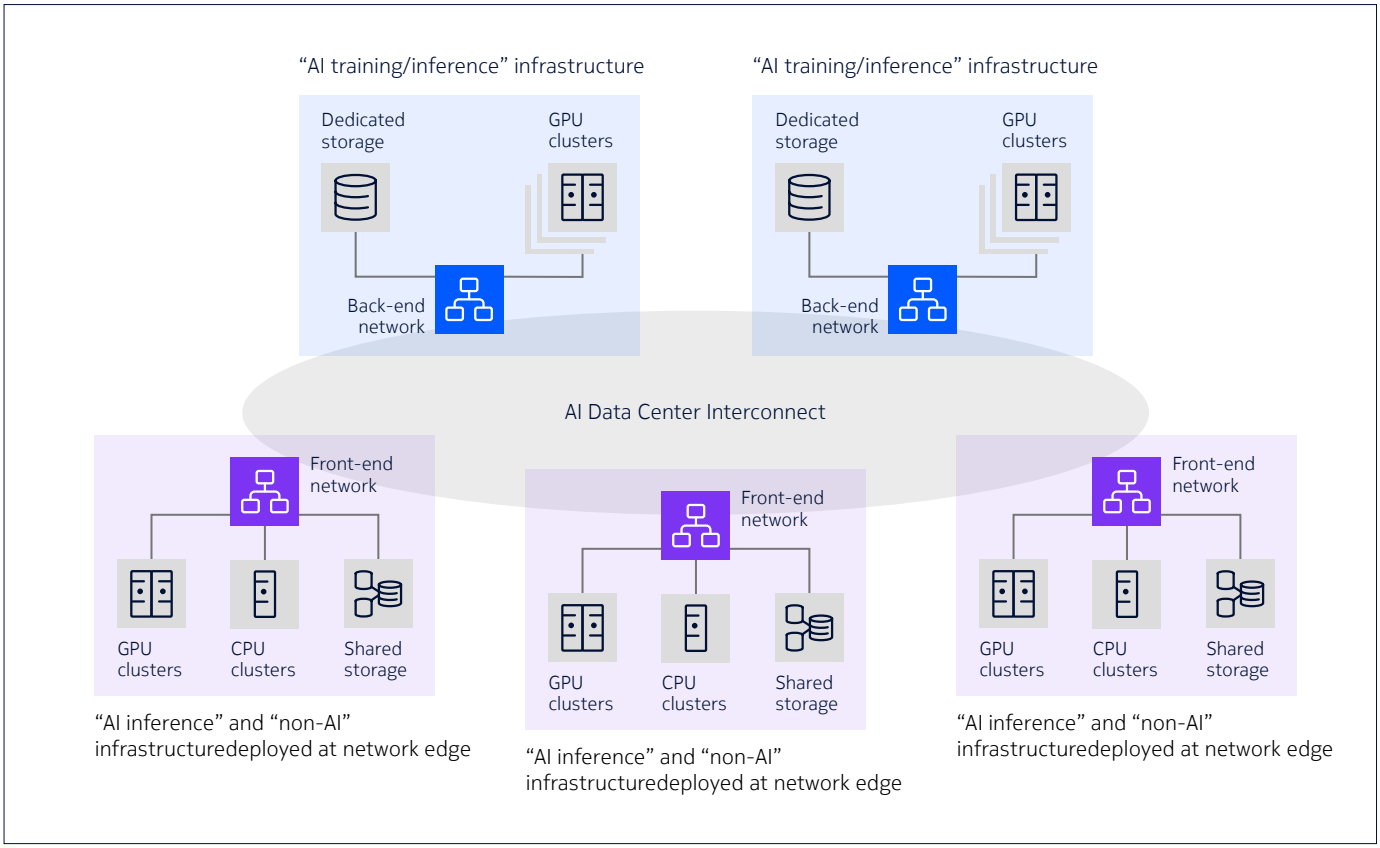
These design choices are based on customer workload requirements, and their cost and power budgets.

Distributed AI infrastructures

AI infrastructures may be co-located, as shown in Figure 7 and Figure 8, or distributed across different locations, as shown in Figure 9.

Figure 9: Distributed AI infrastructures

Co-located back-end and front-end networks



The distributed deployment model may unlock some potentially interesting use cases. For example, several large cloud providers are beginning to offer GPU as a service (GPUaaS). GPUaaS offerings deliver scalable, on-demand access to GPUs to support AI training and inference. Training infrastructure imposes large and highly stringent power consumption requirements. To meet these requirements, organizations must carefully consider the AI infrastructure designs and the possible need to implement training infrastructure at multiple locations to comply with power consumption limits and constraints.

Distributed inference locations can enable cloud providers to extend their service footprints to meet the needs of AI inference use cases that are deployed at the network edge, closer to the end user, to meet real-time response requirements.

Another example involves an enterprise utilizing public AI solutions such as GPUaaS for AI training use cases. In this case, inferencing models run in enterprise edge locations in private AI infrastructures (physically at the enterprise's premises or deployed in a colocation provider's facility) to meet the real-time response and low-latency requirements for AI inference.

The distributed deployment model requires exceptionally reliable and high-performance AI data center interconnect solutions.

Ultra Ethernet Consortium: Ethernet for AI and HPC workloads

More and more organizations are considering Ethernet as an alternative to InfiniBand for the networking portion of AI infrastructures. While Ethernet offers several advantages, there are areas that will need improvements to minimize "tail latency" within AI infrastructures.

The Ultra Ethernet Consortium (UEC) is working to deliver an open, interoperable, high-performance, full-communications-stack architecture based on Ethernet to meet the growing network demands of AI and HPC at scale.

The UEC aims to define a modern transport protocol for AI and HPC applications. While InfiniBand and RoCE are deployed today, they require careful tuning, operation and monitoring. For example, RoCE relies on DCQCN for end-to-end congestion control. DCQCN is sensitive to latency, buffering capabilities and types of workloads and often needs manual tuning so that it can meet performance expectations. This requires a level of expertise and investment, which leads to a high TCO.

UEC members aim to leverage the ubiquity, performance curve and cost benefits of Ethernet to evolve the legacy RoCE protocol with Ultra Ethernet Transport (UET). This modern transport protocol is designed to enhance network performance to meet the requirements of AI and HPC applications while preserving the advantages of the Ethernet/IP ecosystem.

Nokia is a member of the UEC and will continue to actively participate in the following areas of work identified by current UEC specifications and their future evolutions:

- Multi-pathing and packet spraying
- Flexible delivery order
- Modern congestion control mechanisms
- End-to-end telemetry
- Increased scale, stability and reliability.

Nokia data center fabrics for AI workloads

The Nokia Data Center Fabric solution is designed to deliver the reliability and simplicity required to implement high-performance, lossless AI infrastructures, while providing the flexibility to allow the network designs to adapt to evolving business needs.

A complete portfolio of data center hardware platforms

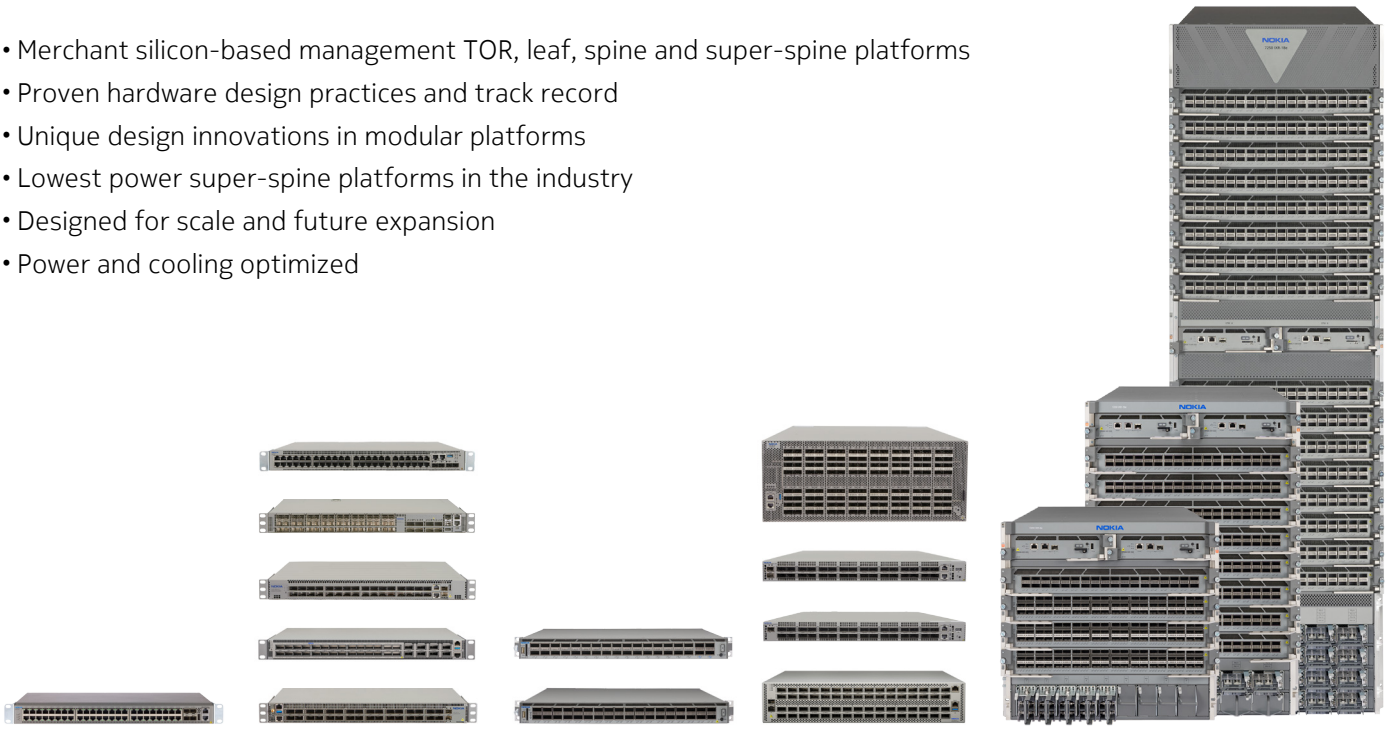
Nokia is the leading builder of cutting-edge IP networking platforms. More than 1.8 million Nokia IP routers have been deployed in mission- and business-critical network environments worldwide, and Nokia brings this expertise to the design of platforms for data center switching. The field-hardened protocol applications that support the world's largest and most demanding IP networks also run on Nokia data center switches.

Nokia offers a comprehensive portfolio (Figure 10) of data center hardware platforms for implementing high-performance leaf–spine designs for back-end and front-end networks.

This portfolio is designed and optimized to support high-capacity, low-latency and lossless back-end networks for the most stringent AI training requirements. It offers an extensive choice of hardware platforms in varying form factors to support front-end network designs that interconnect AI inference compute, non-AI compute and shared storage resources based on an organization's deployment needs.

Figure 10: Nokia hardware portfolio for data center switching

- Merchant silicon-based management TOR, leaf, spine and super-spine platforms
- Proven hardware design practices and track record
- Unique design innovations in modular platforms
- Lowest power super-spine platforms in the industry
- Designed for scale and future expansion
- Power and cooling optimized



7215 IXS-A1 | 7220 IXR-D series | 7250 IXR-X1b/X3b | 7220 IXR-H series | 7250 IXR-6e/10e/18e

Leading-edge modular platform design

Nokia 7250 IXR-6e/IXR-10e/IXR-18e Interconnect Routers are differentiated, modular platforms designed for data center spine, super-spine, aggregation and wide area network (WAN) deployments. These platforms are based on the latest versions of Broadcom Jericho merchant silicon and deliver massive scalability, flexibility and operational simplicity. This makes them an optimal choice for designing very-high-capacity networks for AI training and inference and HPC workloads.

The 7250 IXR-6e/10e/18e platforms provide native hardware support for 800GE, 400GE, 100GE, 50GE, 40GE, 25GE and 10GE interfaces, including breakout support for intra-fabric, WAN and server connectivity. The four-slot 7250 IXR-6e platform supports a system capacity up to 115.2 Tb/s full duplex (FD). The eight-slot 7250 IXR-10e platform supports a system capacity up to 230.4 Tb/s FD. The 16-slot 7250 IXR-18e platform supports a system capacity up to 460.8 Tb/s FD.

In addition to supporting high-availability control, fabric, fan and power configurations, these platforms support industry-leading and unique hardware design innovations and capabilities, including:

- High-quality, midplane-less, orthogonal direct cross-connect—a critical design element for successfully moving to future faster Serializer/Deserializer (SerDes) speeds and beyond
- An architecture without retimers across multiple generations of ASICs, driving low power and ultra-high reliability with a component-minimizing design
- A power- and cooling-optimized design
- Support for 800GE and 400GE coherent optics with support for 400GE ZR+ optics in all pluggable optics positions
- High-capacity 800GE density and efficiency in a 16-slot configuration
- A generational chassis design that can start with Broadcom J2C+ Integrated Media Modules (IMMs) and upgrade to Broadcom J3 while preserving Control Processor Modules (CPMs), power supply units (PSUs) and fans with full backward compatibility for J2C+ IMMs.

These leading hardware design attributes, combined with a full suite of Nokia SR Linux network operating system (NOS) features and the Nokia Event-Driven Automation (EDA) platform, help data center and cloud teams achieve their high-availability design and operations efficiency goals.

A comprehensive portfolio of fixed configuration platforms

As part of the Nokia Data Center Fabric solution, the Nokia 7250 IXR-X1b/X3b, 7220 IXR H series and 7220 IXR D series platforms provide multiple fixed configuration chassis variants with support for 800GE,[6] 400GE, 100GE, 50GE, 40GE, 25GE, 10GE or 1GE port speeds.

The Nokia 7250 IXR-X1b and IXR-X3b are based on Broadcom merchant silicon and provide high speed and density in a 1RU form factor. These platforms support low-latency applications while providing a large buffer memory for delay-tolerant applications.

Nokia 7220 IXR-H series routers are based on Broadcom merchant silicon and designed for the leaf and spine layers of data center fabrics. They deliver very-high-scale interconnectivity for enterprise, service provider and webscale data center and cloud environments. The 7220 IXR-H series consists of the 7220 IXR-H2, 7220 IXR-H3 and 7220 IXR-H4.

Nokia 7220 IXR-D series platforms are based on Broadcom merchant silicon and designed for the leaf and spine layers of data center fabrics. They deliver high-scale interconnectivity for enterprise, service provider and webscale data center and cloud environments. The 7220 IXR-D series consists of the 7220 IXR-D1, 7220 IXR-D2L, 7220 IXR-D3L, 7220 IXR-D4 and the 7220 IXR-D5 platforms.

The 7250 IXR-X1b/X3b, 7220 IXR-H and IXR-D series platforms support redundant DC or AC power options and support either front-to-back or back-to-front airflow configuration options with redundant hot-swappable fans.

The Nokia 7215 IXS-A1 Interconnect System is designed for leaf and spine data center fabric management connectivity in enterprise, service provider and webscale data center and cloud environments.

6 800GE support will be offered in an upcoming release.

SR Linux: The industry's most advanced NOS

Unlike closed operating systems, which are monolithic by design and lack modularity and flexibility, an open NOS is consciously designed to implement an architecture that is modular, extensible, extremely customizable and ready for network automation. The Nokia SR Linux NOS delivers an open, extensible framework that enables automation with advanced software features that provide proven quality and resiliency.

SR Linux opens the NOS infrastructure with a unique architecture built from the ground up around model-driven management and modern interfaces. With SR Linux, organizations benefit from:

- A cloud-native design approach that offers superior programmability, unrivaled flexibility and resilient IP routing
- A Linux-based NOS and kernel that enable network teams to build applications that are modular and isolated into their own failure domains
- A fully modular design where each network application (e.g., BGP, EVPN, LLDP) has its own YANG data structure, which ensures complete openness and consistent operation across all system applications
- A microservices-based, state-efficient design that makes it easy to enable hitless per-application upgrades and resilient networking
- An open, scalable telemetry framework that uses gRPC, gNMI and protobufs and does not require any translation layers
- A NetOps Development Kit (NDK) that allows third-party network applications to be fully integrated into the system with their own YANG models (similar to Nokia network applications).

Lossless Ethernet networks powered by SR Linux features

Ethernet is gaining momentum as a suitable choice for interconnecting high-performance AI workloads. Several factors make Ethernet appealing when it comes to AI. For example, Ethernet is ubiquitous and widely implemented, and has a vast ecosystem of vendors and proven interoperability. It is also cost-effective compared to alternative technology choices.

RoCE enables RDMA technology (which originally was introduced for InfiniBand-based networks) to run over existing Ethernet infrastructures. AI training infrastructure requires lossless network connectivity, and Ethernet networks must support features that help ensure bandwidth is prioritized for AI workload traffic. As part of its comprehensive QoS feature set, SR Linux provides ECN and PFC capabilities for delivering lossless Ethernet networks in RoCE-based deployment models.

ECN is a congestion management mechanism. It reduces packet loss during network congestion scenarios by marking packets to flag congestion within the network. This, is turn, helps notify the endpoints and trigger subsequent rate adjustment actions as required. It ensures lossless behavior and helps maintain throughput.

PFC helps enable lossless Ethernet networks by ensuring that AI networking traffic is given the highest priority based on queue markings. PFC provides flow control on a per-priority basis by allowing pause frames (between endpoints) to be prioritized only on queues experiencing congestion. There is no impact on other priority traffic. It can work with ECN to provide comprehensive congestion control for lossless AI fabrics and support end-to-end congestion control mechanisms such as DCQCN. Most DCQCN capability is implemented in the endpoints (NICs) but needs to work in conjunction with ECN and PFC.

In addition to congestion management features, SR Linux offers superior telemetry, manageability, ease of automation and resiliency features that are relevant and necessary for supporting high-performance AI infrastructures.

The UEC is working on several initiatives to enhance lossless Ethernet technologies for AI and HPC workloads. One of its key initiatives is to define a new transport layer, called Ultra Ethernet Transport (UET). Nokia is an active member and participant of the UEC and will continue to enhance and introduce features within SR Linux to align with the UEC initiatives and market directions related to lossless Ethernet connectivity for AI workloads.

Fabric management and automation platform

Managing data center fabrics that interconnect AI infrastructure is no different from managing data center fabrics that interconnect non-AI or general-purpose workloads. While the hardware platform and software feature requirements may be more stringent to meet the needs of lossless fabric design, the fabric for AI workloads needs to be designed, deployed and operated. The key is to make operations reliable and simple.

SR Linux is designed to enable scalable, easy integration and efficient automation for data center networks. The EDA fabric management and automation platform complements SR Linux. It delivers a modern, innovative solution that increases network agility by using declarative, intent-based approaches to automate all phases of data center fabric operations—from Day 0 design through Day 1 deployment and Day 2+ operations.

EDA builds on the proven Kubernetes platform and leverages a vast open-source ecosystem. This reduces risks and lowers barriers to entry for adopting automation .

EDA includes a cloud-native Digital Sandbox (an integrated network digital twin) that provides a true emulation of the production network. It creates container instances of a subset of the actual live network elements, maintaining both configuration and state.

The Digital Sandbox allows data center teams to represent the design and configuration of the data center fabric in an intent-based, declarative way. Design, fabric and workload intent can be validated on the Digital Sandbox, enabling operations teams to quickly and confidently manage the risk associated with a change. The Digital Sandbox allows teams to try out the changes, perform detailed validations and then apply the changes to the production network.

An abstract, intent-based approach simplifies Day 0 design. The data center operator can focus on high-level aspects of the design, identifying the basic information needed to build a data center fabric. For example, the operator can build the fabric simply by specifying a few parameters, such as the number of racks and the number of servers per rack.

Day 1 deployment uses workload intents and abstracts the complexity of the EVPN configuration by enabling the data center operator to focus on specifying high-level parameters. This can be as simple as identifying the set of downlinks an application workload uses to connect to the fabric. Complexities such as switch-to-switch EVPN and allocation of VXLAN network identifiers, route distinguishers, route targets, Ethernet segment IDs and Ethernet virtual interfaces are all abstracted. Workload intent can be validated using the Digital Sandbox before being deployed into the production network.

The Nokia automation platform adopts DevOps approaches to deliver enhanced NetOps. It uses multi-dimensional telemetry to monitor and gain deep insights into all network traffic. It also supports easy integration with third-party systems and multiple cloud environments with flexible next-generation interfaces such as REST APIs and the Nokia Connect microservice.

AI data center interconnect solution

Nokia offers a complete and comprehensive solution for interconnecting AI infrastructures between data centers and across the WAN. The cloud era demands a new network architecture that interconnects edge, regional and core cloud data centers. Data center interconnect ecosystems are being built based on agile, flexible and automated IP/optical infrastructures that can support current and future cloud service requirements.

The two main types of interconnect ecosystems are optical data center interconnect (DCI) and IP data center interconnect. Optical DCI involves connecting AI infrastructures using optical networking or optical transport. Optical DCI can be designed as a point-to-point, mesh or ring topology depending on the number of data centers and the resiliency requirements. The Nokia sixth-generation super-coherent Photonic Service Engine (PSE-6s) enables massive network scale with the industry's first 2.4 Tb/s coherent transport solution. It enables network operators to scale transport capacity to unprecedented levels across metro, long-haul and subsea networks.

IP DCI connects distributed AI infrastructure domains through IP technology. The Nokia family of FP5-based routers offers industry leading capacity and speed. These routers are the first to introduce 800GE routing. Deploying these routers can enable an operator to triple IP network capacity in the same space and energy footprint.

Reference designs with Nokia data center fabric solution

Optimizing AI infrastructure designs for scale, performance and cost

Networking for AI workloads is highly dependent on the processing requirements for AI applications. The network design considerations differ significantly between AI training and AI inference infrastructures. Emerging AI use cases and applications typically need to handle a massive number of parameters and data sets as part of the process for training the AI model. This can drive the need for tens to hundreds to thousands of specialized GPUs and similar accelerated processors, which are installed in AI servers (compute). The back-end network that connects these servers is dependent on the number of GPUs installed per server as well as the number of servers implemented for a particular AI training deployment.

With AI training, the workflows associated with the GPUs need to communicate with each other without any network delay and loss. The GPUs are high-cost, processing-intensive devices. Any idle time caused by to network delay has a negative impact on GPU performance and utilization as well as the overall JCT.

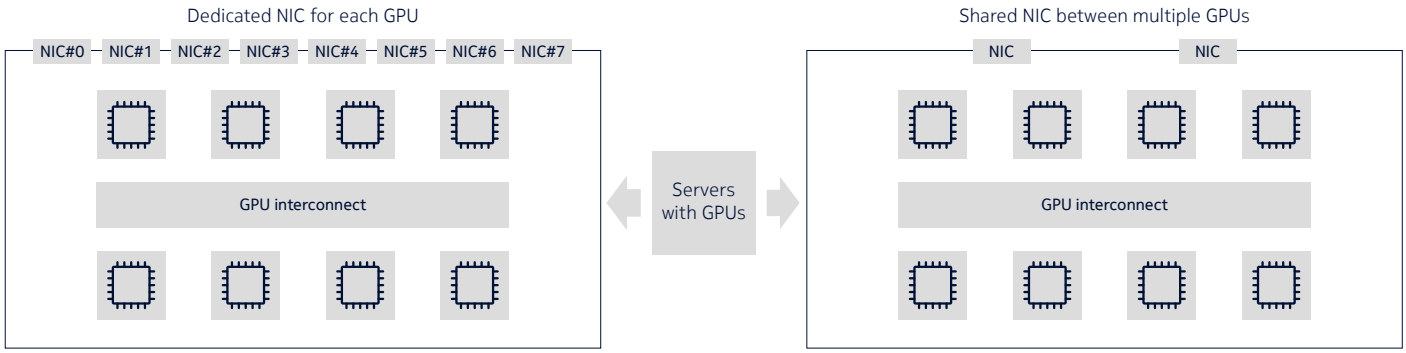
As shown in Figure 11, the network for AI workloads will need to support varying AI server and GPU requirements. A performance-optimized design will typically have a building block of up to eight GPUs per server, with each GPU associated with a NIC, requiring a total of eight NICs for this peformance-optimized design option. A design that has less stringent performance and processing requirements may adopt building blocks with fewer GPUs and NICs.

High-performance GPUs typically support internal GPU-to-GPU communication through GPU interconnects.[7] Traditionally, inter-GPU communication shares the same bus interconnect as CPU-to-GPU communication such as Peripheral Component Interconnect Express (PCIe). PCIe is a high-speed

7 Li, A et al. "Evaluating Modern GPU Interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect". White paper. Retrieved 9 August 2024.

serial communication computer expansion bus standard that may become a bottleneck for GPU-to-GPU communication. For example, NVIDIA NVLink is a high-speed connection for GPUs and CPUs formed by a robust software protocol that typically rides on multiple pairs of wires printed on a computer board.[8] It lets processors send and receive data from shared pools of memory at lightning speed. Now in its fourth generation, NVLink connects hosts and accelerated processors at rates up to 900 Gb/s.

Figure 11: Performance- and cost-optimized design options



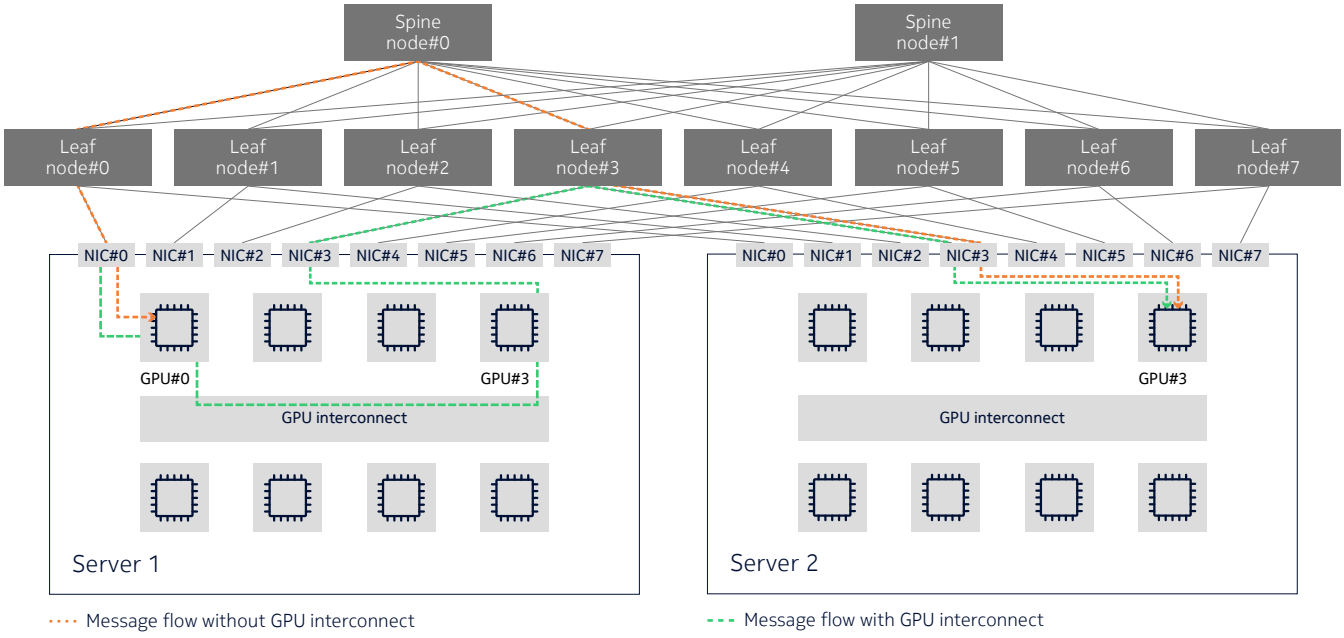
Performance-optimized design

Cost-optimized design

Rail-optimized design

A rail-optimized design (Figure 12) uses NVLink technology to enable GPUs within a server to communicate directly with each other, as well as enable optimized inter-server GPU-to-GPU communications.[9]

Figure 12: Rail-optimized design for maximum performance



8 Merrit, R. "What is NVLink?". NVDIA blog post. Retrieved 9 August 2024.

9 Mandakolathur, K and Jeaugey, S. "Doubling all2all Performance with NVIDIA Collective Communication Library 2.12". NVIDA blog post. Retrieved 9 August 2024.

In Figure 12, NIC#0 in Server 1 and NIC#0 in Server 2 are connected to the same switch: Leaf node#0. Similarly, NIC#1 in Server 1 and NIC#1 in Server 2 are connected to Leaf node#1, and so on for each NIC within each of the servers to create dedicated rails. This design approach is referred to as a rail-optimized design.

The dotted orange arrow shows message flow between GPU#0 in Server 1 to GPU#3 in Server 2 where no NVIDIA Connection Communications Library (NCCL) are used. The message flow would originate from GPU#0 in Server 1 and traverse through Leaf node#0, Spine node#0 and Leaf node#3 on the way to GPU#3 in Server 2, through three hops of network nodes, resulting in traffic slowdown.

The dashed green arrow shows message flow between GPU#0 in Server 1 to GPU#3 in Server 2 where NCCL features are used. The NCCL features leverage connectivity between GPUs within Server 1 to first move data from GPU#0 to GPU#3 in Server 1 on the same rail as the destination, and then send it to the destination (GPU#3 in Server 2) without crossing rails. This enables optimized traffic flows and reduces traffic flow through the spine layer, helping to reduce overall network design costs.

This rail-optimized topology, also called a rail-strip, is a fundamental building block of the back-end network cluster for supporting AI applications. An operator can easily scale out this topology by adding multiple rail strips interconnected over a layer of spine switches.

Nokia reference design for a single-stage back-end network

Figure 13 depicts the available Nokia data center platform choices for a single-stage back-end network design implemented using fixed-configuration platforms. These platforms support redundant power and fan subsystems. This back-end network design may implement a pair of fixed-configuration platforms to address additional redundancy considerations for the control plane.

Figure 13: Single-stage back-end network with fixed-configuration platforms

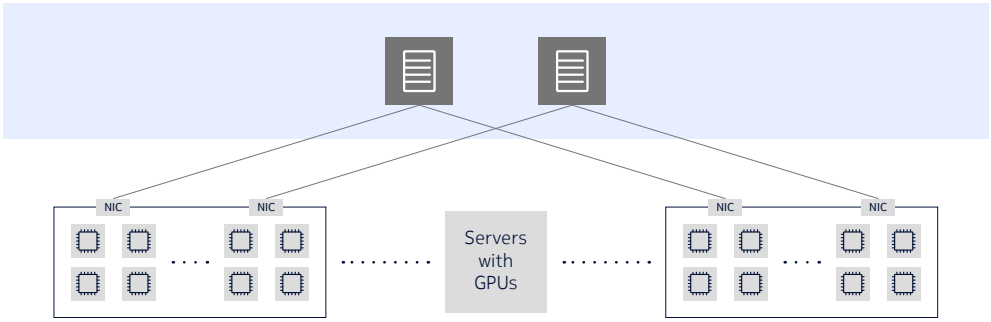
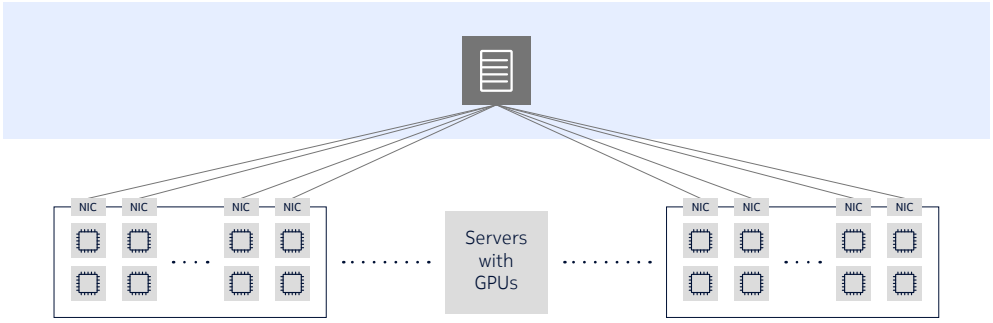


Figure 14 depicts a variant of the single-stage back-end network design implemented using a modular configuration chassis. These platforms support redundant control, fabric, power and fan subsystems, offering a compact and redundant platform footprint.

This design approach can support network connectivity for a small to medium configurations with tens to hundreds of GPUs, along with server connections from 100GE up to 800GE speeds.

Figure 14: Single-stage back-end network with modular platform

Back-end network
Platform choices: Nokia 7250 IXR-6e/10e/18e



Nokia reference design for rail optimized back-end network

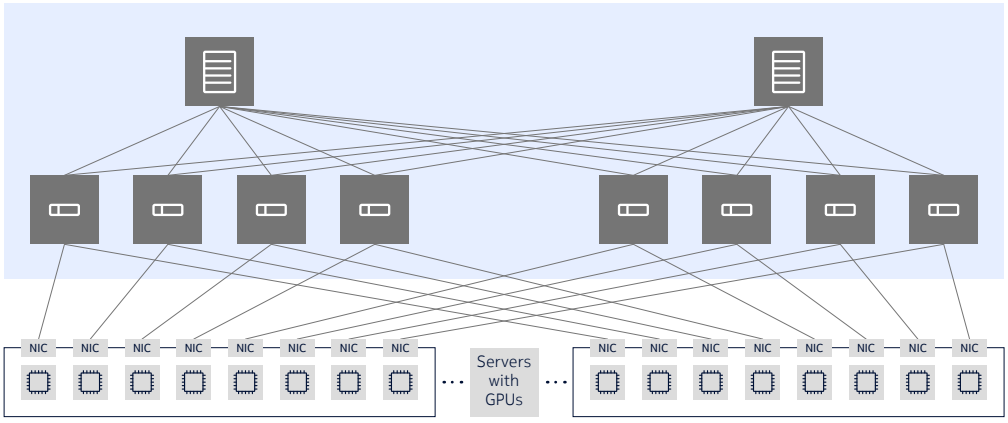
Figure 15 depicts the available Nokia data center platform choices for a rail-optimized back-end network design implemented using a choice of fixed-configuration platforms or modular configuration chassis. Fixed-configuration platforms support redundant power and fan subsystems. Modular configuration chassis support redundant control, fabric, power and fan subsystems.

Rail-optimized designs help deliver lower-latency, high-performance and efficient GPU communication. They use GPU interconnect technologies to optimize the way traffic flows through the networking interconnects.

These designs enable the implementation of very-large-scale GPU designs and high-capacity network configurations that can support network connectivity for medium to very large configurations that include thousands to tens of thousands of GPUs, along with support for server connections at 100GE up to 800GE speeds. The basic building block depicted in Figure 15 can be extended to scale in a modular manner to support GPU designs that include tens of thousands of GPUs.

Figure 15: Rail-optimized back-end network design

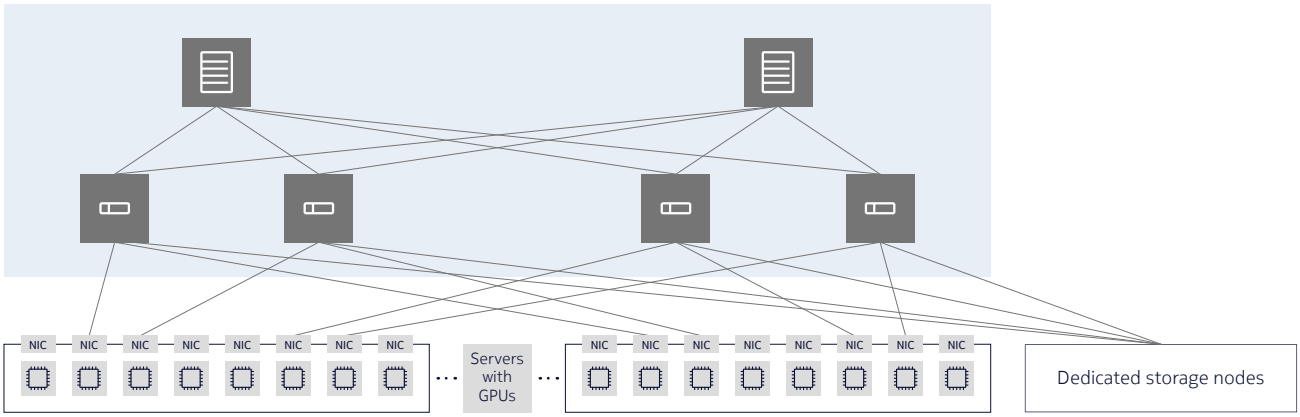
Back-end network
Platform choices: Nokia 7250 IXR-6e/10e/18e,
Nokia 7250 IXR-X1b/X3b, Nokia 7220 IXR-H4, Nokia 7220 IXR-D4/D5



Nokia reference design for dedicated storage back-end network

Figure 16 shows the available Nokia data center platform choices for a dedicated storage back-end network design implemented using a choice of fixed-configuration platforms or modular configuration chassis. Fixed-configuration platforms support redundant power and fan subsystems. Modular configuration chassis support redundant control, fabric, power and fan subsystems. This design supports server connections at 100GE up to 800GE speeds.

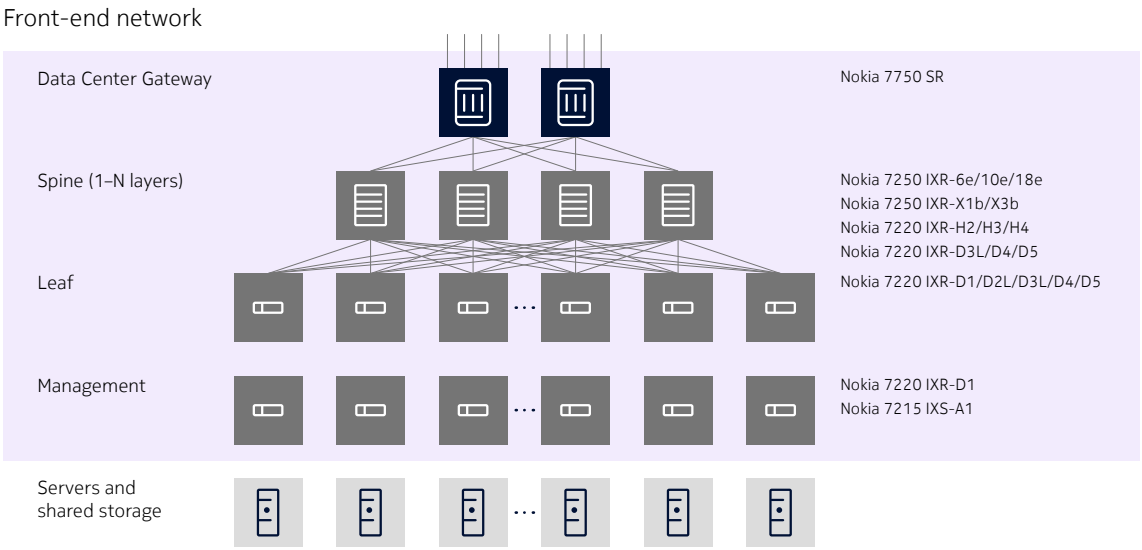
Figure 16: Dedicated storage back-end design



Nokia reference design for a front-end network

Figure 17 shows the available Nokia data center platform choices for implementing a front-end network. The front-end network supports connectivity for general-purpose and AI inference workloads. Nokia provides a comprehensive family of next-generation data center platforms that support 800GE, 400GE, 100GE, 50GE, 40GE, 25GE, 10GE and 1GE port speeds for data center spine, leaf and management roles. The Data Center Fabric portfolio is complemented by the flagship Nokia 7750 Service Router (SR) platforms, which can be deployed at the data center gateway to support IP backbone connectivity, IP DCI and IP peering roles.

Figure 17: Front-end network leaf–spine design



Conclusion

AI and ML will continue to transform the way industries and businesses are run. They promise to help improve operational efficiency and foster innovation while providing new business opportunities, unlocking new revenue streams and revolutionizing the user experience across a range of sectors.

Networking plays a critical role in the implementation of AI training and inference infrastructure. Ethernet is rapidly becoming a preferred choice for back-end networks, which complements its ubiquity and current dominance in front-end network designs. The evolving work of the Ultra Ethernet Consortium (UEC) will continue to drive enhancements that will make Ethernet the best option for implementing AI infrastructures.

The Nokia Data Center Fabric solution includes a comprehensive portfolio of data center hardware platforms for implementing high-performance leaf–spine designs for back-end and front-end networks. This portfolio is designed and optimized to support high-capacity, low-latency and lossless back-end networks that meet the most stringent AI training requirements. It includes an extensive choice of hardware platforms in varying form factors to support front-end network designs that interconnect AI inference compute, non-AI compute and shared storage resources based on deployment needs.

Additionally, the requirement for AI models to be served closer to end users (AI inference at the edge of the network) is driving the need for reliable, high-performance interconnect across the AI infrastructure domains. The Nokia Optical DCI and Nokia Data Center Gateway solutions are ready to meet all current and evolving distributed AI connectivity requirements.

Abbreviations

AI	artificial intelligence
API	application programming interface
APU	accelerated processing unit
BGP	Border Gateway Protocol
BSS	business support system
CNP	Congestion Notification Packet
CPM	Control Processor Module
CPU	central processing unit
CSP	communications service provider
DCI	data center interconnect
DCQCN	Data Center Quantized Congestion Notification
DIY	do it yourself
ECN	Explicit Congestion Notification
EVPN	Ethernet Virtual Private Network
gNMI	gRPC Network Management Interface
GPT	Generative Pre-trained Transformer
GPU	graphics processing unit
GPUaaS	GPU as a service
GRH	global routing header
gRPC	gRPC Remote Procedure Calls
HCA	Host Channel Adapter
HPC	high-performance computing
IDC	International Data Corporation
IMM	Integrated Media Module
IoT	Internet of Things
IP	Internet Protocol
IT	information technology
JCT	job completion time
LLDP	Link Layer Discovery Protocol
LLM	large language model
LPU	language processing unit

ML	machine learning
NAS	network-attached storage
NCCL	NVIDIA Connection Communications Library
NDK	NetOps Development Kit
NIC	network interface card
NLP	natural language processing
NOS	network operating system
NVM-e	NVM Express
OSS	operations support system
OT	operational technology
PCIe	Peripheral Component Interconnect Express
PFC	Priority Flow Control
PSE	Photonic Service Engine
RDMA	Remote Direct Memory Access
RoCE	RDMA over Converged Ethernet
SAN	storage area network
SerDes	Serializer/Deserializer
SME	small and medium enterprise
TCO	total cost of ownership
TCP	Transmission Control Protocol
TPU	tensor processing unit
UDP	User Datagram Protocol
UEC	Ultra Ethernet Consortium
UET	Ultra Ethernet Transport
VXLAN	Virtual Extensible Local Area Network
XPU	auxiliary processing unit
YANG	Yet Another Next Generation

About Nokia

At Nokia, we create technology that helps the world act together.

As a B2B technology innovation leader, we are pioneering networks that sense, think and act by leveraging our work across mobile, fixed and cloud networks. In addition, we create value with intellectual property and long-term research, led by the award-winning Nokia Bell Labs.

Service providers, enterprises and partners worldwide trust Nokia to deliver secure, reliable and sustainable networks today – and work with us to create the digital services and applications of the future.

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

Document code: 1039350 (September) CID214186