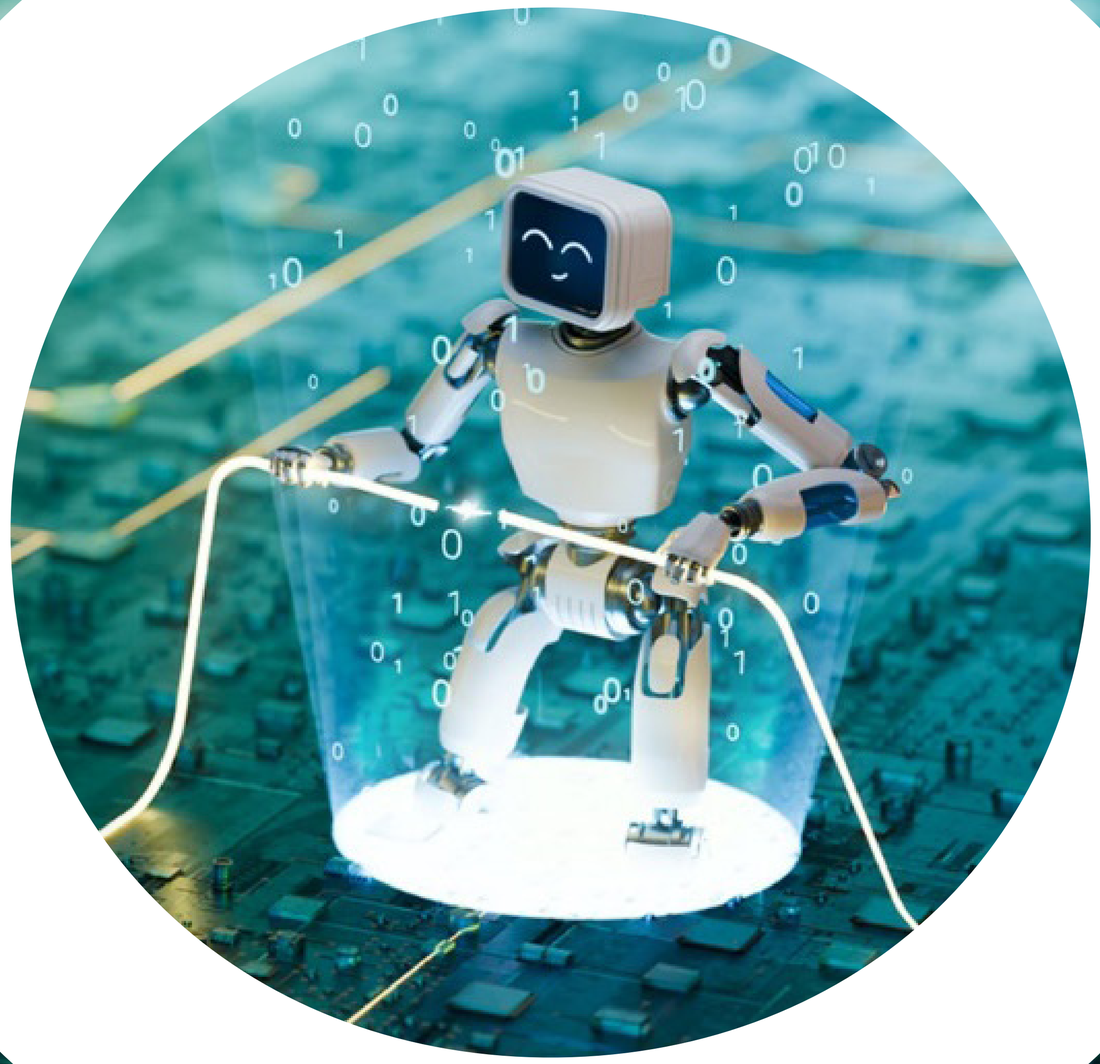
NOKIA

Data center networks for the AI era

Reliable automated fabrics and high-performance network interconnectivity



Contents

The rise of AI and the role of data center networks

How we got here

Artificial intelligence (AI) is dominating the technology landscape and 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 businesses are already exploring a wide range of use cases that leverage AI workloads for applications such as natural language processing (NLP), outcome prediction, visual analysis and personalization. Generative AI (GenAI) in particular is gaining popularity because of its ability to create content such as text, images, code, audio and video.

What data center teams expect from their networks

Modern data center networks need to support traditional workloads and new AI workloads. Data center teams expect their networks to be simple, reliable and easy to scale. They also expect that everything will keep working even while they're upgrading hardware or software. And they crave products that are resilient, secure and perform well. In short, data center networks must just work and be easy to operate.

Current solutions have lost the plot on reliability and simplicity

Data center networking has a well-established market, but existing solutions have not been able to provide the reliability that data center teams need.

For example, an ACM SIGCOMM Computer Communication Review study of more 180,000 switches in data centers across 130 geographical locations revealed that approximately 32 percent of switch failures are caused by hardware issues and 17 percent of switch failures are caused by software bugs in vendor switch operating systems.[2]

A study by Uptime Intelligence found that human error directly or indirectly accounts for between two-thirds and four-fifths of all downtime incidents.[3] The cause of a failure can lie in how well a process was taught, how tired, well trained or resourced the staff are, or whether the equipment itself was unnecessarily difficult to operate.

These problems need an immediate remedy.

It's time to transform data center network infrastructures

Data center networks must evolve so that they can deliver reliable and seamless connectivity within the data center, which includes servers equipped with central processing units (CPUs), graphics processing units (GPUs), network interface cards (NICs) and storage solutions, as well as between data centers, clouds and other networking domains.

1 "Worldwide Spending on Artificial Intelligence Forecast to Reach $632 Billion in 2028, According to a New IDC Spending Guide". IDC press release. 19 August 2024. Accessed 4 September 2024.

2 R Singh, M Mukhtar et al. "Surviving switch failures in cloud datacenters". ACM SIGCOMM Computer Communication Review, Volume 51 Issue 2, April 2021. Accessed 6 September 2024.

3 A Lawrence and L Simon. "Annual outages analysis 2023: The causes and impacts of IT and data center outages". UII keynote report 92M. March 2023. Accessed 6 September 2024.

Networking for AI workloads

AI workloads are classified into two broad categories—AI training and AI inference—based on the tasks they perform. The AI training process creates the initial AI model. It includes data collection, model selection, model training, model evaluation, model deployment and model monitoring and involves intensive use of GPUs. After the model is developed, the process of inference can be initiated. Inference is about enabling end users or things to interact with the model.

The network plays a critical role in ensuring the best possible performance for training and inference tasks. It is essential to implement well designed back-end and front-end network architectures that can meet the stringent requirements of AI workloads, which include high reliability, high speed, high capacity, low latency and lossless networking. These networks must be able to maximize the utilization of compute resources to achieve the shortest possible job completion times (JCTs) for these workloads.

The **back-end network** is used for interconnecting high-value GPUs required for compute-intensive AI training, AI inference and other high-performance computing (HPC) workloads, often at very high scale. The **front-end network** supports connectivity for AI workloads, general-purpose workloads (non-AI compute) and the management of AI workloads.

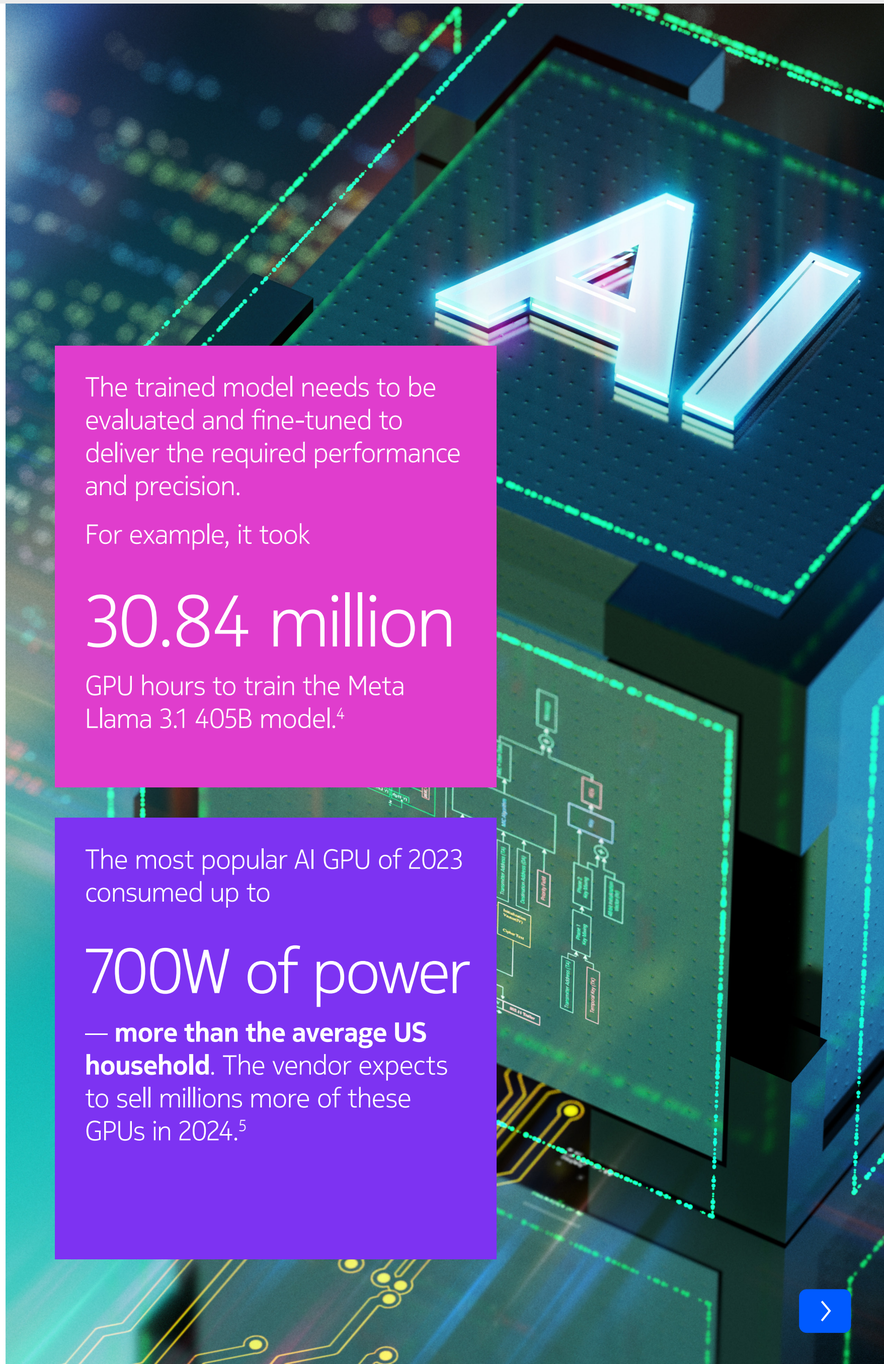
Businesses may choose to implement their own private training infrastructures or use AI platforms and associated services from large cloud providers, an approach known as GPU as a service (GPUaaS) or public AI. Inferencing models that need low-latency end-user access are typically run in edge locations in private AI infrastructure on the business's premises or in colocation provider facilities.

Some massive AI infrastructures may be too large to meet power requirements or legislative constraints within the planned area of deployment. These infrastructures may have to be segmented and moved to locations where power is readily available.

In addition to reliable connectivity within the data center, it is essential to support reliable, high-performance network interconnectivity between data centers that implement AI and high-performance computing (HPC) workloads across multiple locations.

4 Meta Llama 3.1 405B model overview. Hugging Face website. Accessed 4 September 2024.

5 Shilov, A. "Nvidia's H100 GPUs will consume more power than some countries — each GPU consumes 700W of power, 3.5 million are expected to be sold in the coming year". Tom's Hardware. 26 December 2023. Accessed 4 September 2024.



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.[4]

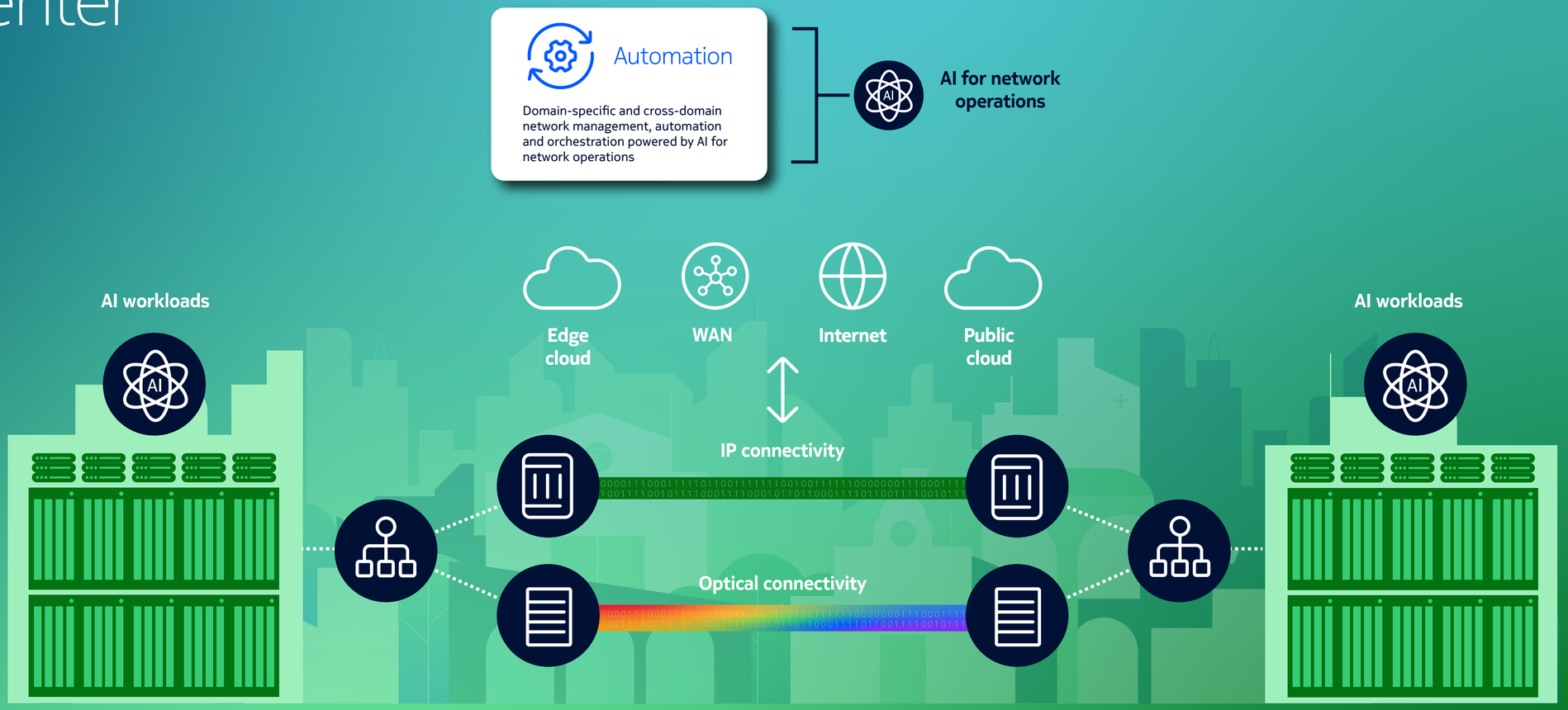
The most popular AI GPU of 2023 consumed up to

700W of power

— **more than the average US household**. The vendor expects to sell millions more of these GPUs in 2024.[5]

Nokia Data Center Networks for the AI era

Nokia addresses the demands of the AI era with a comprehensive data center networking solution. The solution includes data center fabrics that deliver reliable connectivity within the data center and IP/optical data center interconnectivity solutions that connect data centers, clouds, the WAN and the internet.



Data Center Fabric

Reliable data center switching with predictable and simplified operations for your data center environments

Data Center Gateway

High-performance data center to data center, internet, WAN and cloud interconnect and DDoS mitigation

Optical DCI

Dedicated optical networking enables high bandwidth, very low latency and highly secure data transmission for business-critical applications

Key solution components

Component	How it supports networking for the AI era
Data Center Fabric	• Provides reliable data center switching with predictable and simplified operations for data center environments • Supporting products: Data center hardware platforms (7220 IXR, 7250 IXR), network operating system (SR Linux) and Event-Driven Automation (EDA) platform • Supporting solutions: Networking for AI workloads, AIOps for data center networks
Data Center Gateway	• Supports high-performance data center to data center, internet, WAN and cloud interconnect, as well as security solutions for DDoS mitigation • Supporting products: 7750 Service Router portfolio
Optical DCI	• Provides dedicated optical networking for data center interconnect (DCI) to enable high bandwidth, very low latency and highly secure data transmission for business-critical applications • Supporting products: 1830 Photonic Service Interconnect – Modular (PSI-M), 1830 Photonic Service Switch
Automation	• Provides domain-specific and cross-domain network management, automation and orchestration powered by AI for network operations • Supporting products: EDA, Network Services Platform (NSP), WaveSuite

Reliable data center switching fabrics

Data center switching infrastructures are under massive strain because of pervasive demand for all forms of content, evolution to cloud-native applications, growing dominance of AI/ML-based workloads, and increasing adoption of hybrid cloud, edge cloud and multicloud models.

Networking teams want their data centers to work reliably and consistently and be simple to operate. But human errors continue to cause major problems, regardless of whether they are caused by vendor product design/quality issues or mistakes by network operations staff. Current solutions can't deliver the reliability, ease of use and flexibility required for modern data center switching infrastructures.

Our modern Data Center Fabric solution adopts a quality-first approach and brings new levels of reliability, simplicity and adaptability to help data center teams manage accelerating demand with all the freedom and control they need.

We offer a comprehensive portfolio of data center hardware platforms for implementing high-performance back- and front-end networks for AI workloads and traditional workloads. For more information on these network architectures, see our Networking for AI workloads application note.

Our uniquely open, extensible and resilient SR Linux NOS is built on the latest technological innovations, including containers, microservices, open-source projects, model-driven architectures, YANG data models, and streaming and "on change" telemetry. It also uses modern management protocols such as gNMI and REST APIs and provides capabilities for delivering lossless Ethernet networking for AI infrastructures.

The EDA platform provides automation capabilities that simplify network design, deployment and operations and ensure that they work with the expected reliability and predictability.

Why choose Nokia for data center switching?

- Boost reliability and quality with data center switching solutions that aim to eliminate human errors.
- Benefit from the highest levels of reliability, flexibility and openness with a NOS built from the ground up around model-driven management.
- Deploy a next-generation network automation platform designed to make network operations reliable and simple.
- Gain new levels of flexibility with a data center fabric solution that easily adapts to existing staffing, processes, ecosystems and environments.

GigaOm has identified Nokia as a data center switching leader and outperformer for a third straight year.

Read the report

Next-generation data center automation

Our goal with network operations is to reduce human error to zero. Nokia EDA is a next-generation data center network automation platform that combines speed with reliability and simplicity. It makes network automation more trustable and easier to use, from small edge clouds to the largest data centers.

Existing solutions have not realized the promise of automation because they focus on the wrong things and miss out on adding more reliability and predictability to data center network operations. Automation is an amplifier that can make good things better and turn bad things into breakage at scale. Turning on automation at scale can uncover previously unknown weaknesses that expose the fragility of the network.

EDA makes network operations predictable. For example, operations teams can use tools such as pre- and post-automation checks to help identify conflicts before pushing a new configuration to a device. This makes it easy to verify whether the intended configuration has been applied on the device or rejected by it.

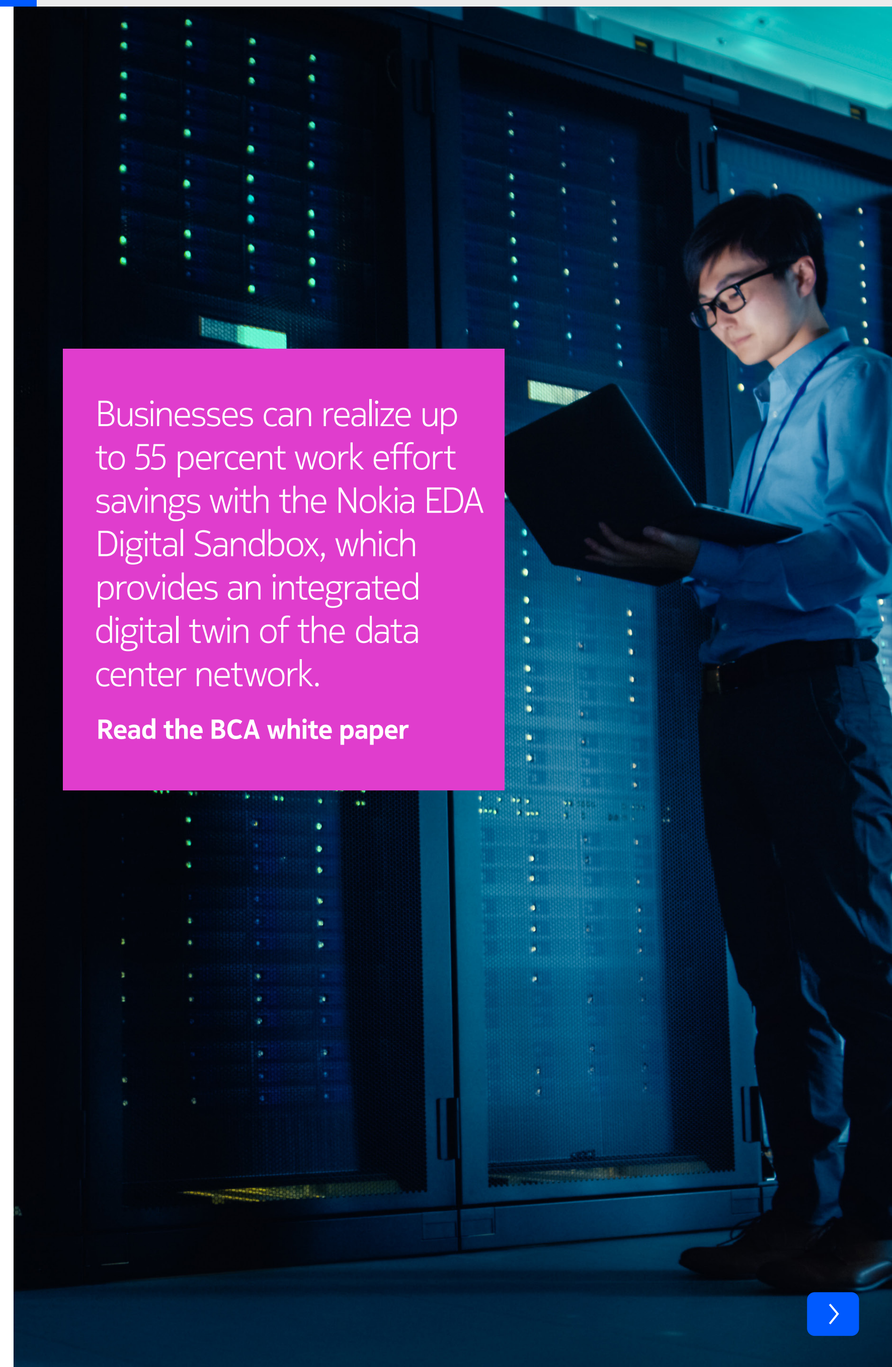
With EDA, operations teams can automate the entire data center network lifecycle from initial design to deployment and daily operations. They can specify what they want to achieve through high-level declarative intents and let EDA determine how best to implement the low-level detailed configurations.

EDA builds on the proven Kubernetes platform and leverages a vast open-source ecosystem. This reduces risk and lowers barriers to entry for adopting automation. EDA provides a built-in Digital Sandbox that creates a virtual digital twin of the live production network. It provides a snapshot of the network and maintains its state to provide a true network emulation at any point in time.

Building on our SR Linux GenAI Assistant, the EDA platform supports natural language query capabilities that help to simplify data center fabric operations. EDA integrates easily with cloud, workload management, event notification and collaboration systems without manual intervention.

Why choose Nokia for data center automation?

- Work with the most automation-forward vendor in the data center space.
- Benefit from an automation platform that ensures that changes do what they are supposed to do, and that the network is operating as intended.
- De-risk operations with a Digital Sandbox that provides a digital twin of the live data center network.
- Automate with confidence using a platform that delivers multivendor, intent-based management with a broad set of reliability-enhancing capabilities environments.



Businesses can realize up to 55 percent work effort savings with the Nokia EDA Digital Sandbox, which provides an integrated digital twin of the data center network.

Read the BCA white paper

Interconnectivity with the Data Center Gateway

Businesses now implement multiple data centers or clouds as part of their digital transformation journey. For example, they may use a mix of their own privately run central, regional, metro or edge data centers. They may also use resources from public cloud or colocation facilities to form their own hybrid clouds. The use of distributed edge clouds is gaining momentum as a means to offer better performance by locating data centers and clouds closer to the end user.

In addition to managing these complex and varied networks, businesses need to connect data centers to each other, to the internet, to the WAN and to edge clouds. The Nokia Data Center Gateway (DCGW) addresses this challenge by enabling high-performance interconnectivity between these entities.

As AI-related workloads become a more dominant part of business transformation initiatives, the role of data center interconnectivity will become even more important, especially in scenarios where it needs to deliver:

- Connectivity to public AI-based cloud frameworks (also known as GPUaaS) to provide the data needed to build the AI models
- 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
- Connectivity across AI infrastructures where workloads are distributed and offered across multiple GPUaaS provider locations to ensure adherence to keep power consumption within specific limits.

The Nokia DCGW combines an industry-leading Border Gateway Protocol (BGP) and Ethernet VPN (EVPN) protocol stack in a wide range of fixed and modular routing platforms based on Nokia or merchant silicon. Our flagship 7750 SR portfolio supports all gateway applications without compromising on performance. Our 7250 IXR for data center fabrics platforms support interconnectivity between data centers and across the WAN.

Why choose the Nokia Data Center Gateway?

- Supports Data Center Interconnect (DCI) over IP-only, MPLS (LDP, RSVP, SR-MPLS) or SRv6 tunnels in the WAN.
- Optimize traffic performance with internet connectivity that uses multiple BGP peering agreements.
- Connect public cloud providers to on-prem data centers to support a hybrid approach that combines the strengths of both.
- Combine DDoS mitigation with traffic encryption to secure the data center fabric and ensure maximum availability.

Interconnectivity with optical Data Center Interconnect

The biggest trends in today's digital market all point to the need for rapid bandwidth growth. Whether it is AI processing and analytics, the rollout of autonomous use cases in transportation and manufacturing, or the deployment of 5G standalone services or residential PON access networks, the focus will be on boosting capacity and reducing latency in metro and interconnect networks over the next decade.

In an increasingly AI-automated world, the need to reduce latency is paramount, especially for autonomous and remote-control applications. Edge computing brings processing power closer to the source, enabling lightning-fast response times and robust local analytics.

Nokia business-critical optical DCI solutions provide two simple ways to connect the optical transport layer to the routing layer.

The first option is to use short-reach or gray optics to enable the router port to connect with a modular optical transponder or muxponder in a separate chassis or line system. For this option, we offer a range of Photonic Service Engines (PSEs) that are performance-optimized for capacity, reach and fiber efficiency. This method is best for applications that require high capacity, flexibility and long reach, up to and including intercontinental submarine links.

The second option is to use pluggable digital coherent optics (DCOs) directly in the router itself. This approach enables routers to connect efficiently and directly over coherent wavelengths and does not require external transponder modules. Pluggable DCOs are profile-optimized for high density and low energy consumption and are suitable for metro–regional transport applications.

Our optical DCI solutions also help you build Quantum-Safe Networks. They combine AES 256-bit encryption with symmetric key distribution to provide the highest levels of security at any data rate. Optical intrusion detection tools, including wavelength tracking and Optical Time-Domain Reflectometry (OTDR), localize problems and detect fiber cuts and network intrusions to complete a highly secure DCI solution.

Why choose Nokia for Optical DCI?

- Work with a leader in optical DCI, coherent optical components and optical line systems.
- Benefit from the innovative power and cooling designs of our QSFP56-DD and CFP2 line cards, which make it easy to equip 400GE coherent pluggable transceivers in existing Nokia routers and line cards.
- Take advantage of our PSE-6s, which ushers in a new frontier in optical transport network evolution, offering 2.4 Tb/s scale, three times the reach of other 800G solutions and 60 percent lower power consumption per bit.

Nokia OYJ
Karakaari 7
02610 Espoo
Finland

Tel. +358 (0) 10 44 88 000

CID:214215

nokia.com

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.

With truly open architectures that seamlessly integrate into any ecosystem, our high-performance networks create new opportunities for monetization and scale. 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.