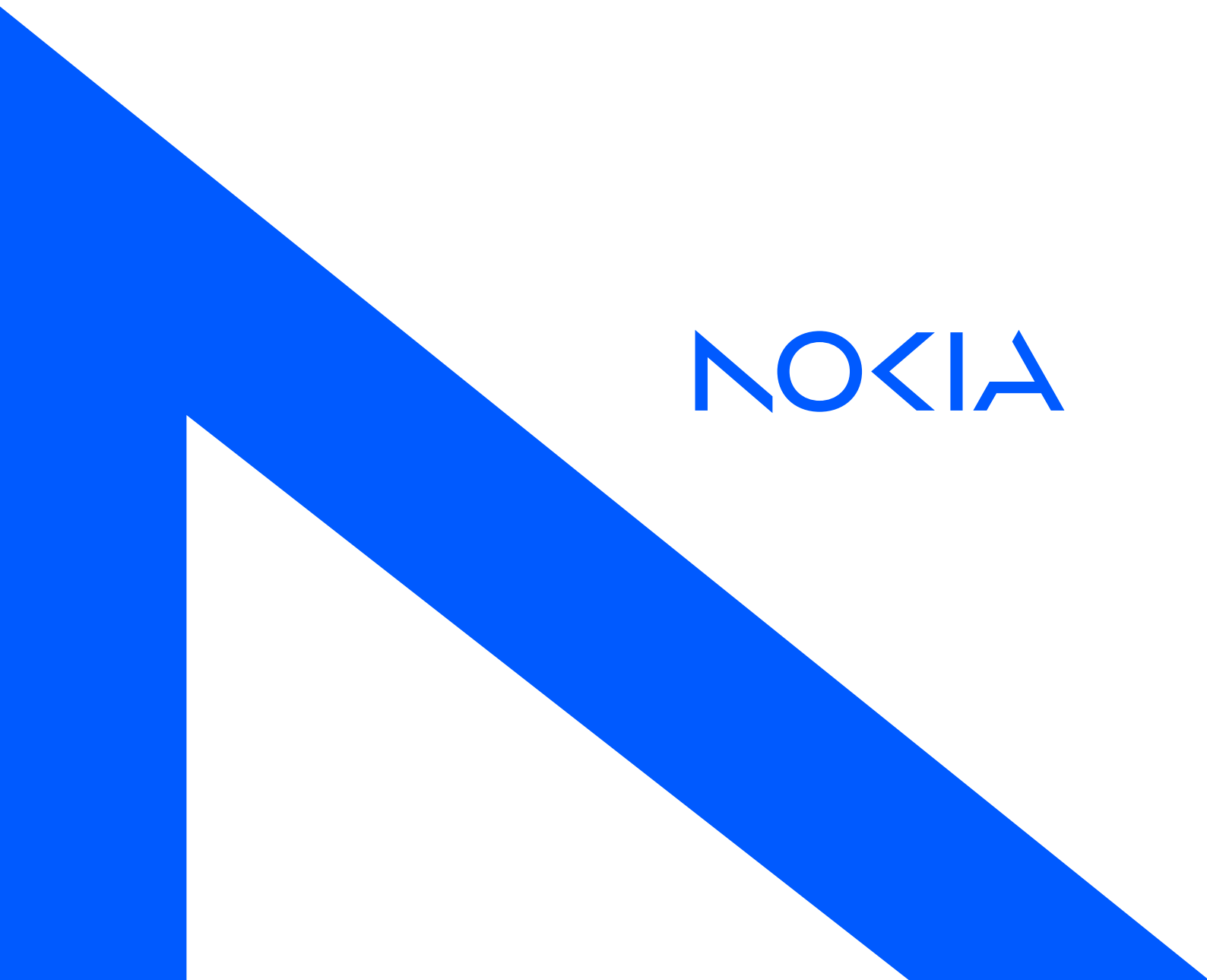


Inference connectivity

A market opportunity created by distributed AI inference



NOKIA

Executive overview

The next phase of AI infrastructure will be shaped not only by the growth of compute, but by where that compute must be deployed and how it must be connected. Today's AI applications generate relatively modest workloads, largely driven by human-to-AI interactions. The transition to agentic AI, followed by physical AI involving devices such as robots, drones, cameras and sensors, is expected to generate a much larger volume of machine-created workloads and tokens requiring processing.

Supporting that growth requires a massive expansion of compute and power. At the same time, the mix of infrastructure is expected to shift from training-dominated deployment toward inference. Roughly 70% of resources deployed over the next five years could be inference-oriented, with the overall training-to-inference mix moving from approximately 80/20 toward 40/60.

Inference has a different geographic and operational model from training. Training resources can be centralized in large, remote sites because location is less important and tightly synchronized processors benefit from proximity. Inference supports end-user applications and therefore benefits from being located closer to users and data sources. Proximity reduces the cost of transporting high-bandwidth data and helps address latency, sovereignty, privacy and regulatory constraints.

Yet building large new metropolitan data centers is difficult. Cost, power availability, permitting, environmental requirements and community opposition can constrain development. The practical alternative is to distribute inference capacity in smaller clusters across existing sites such as warehouses, central offices, office buildings or malls within a metropolitan area.

That distributed model creates the central market opportunity: AI providers need connectivity that allows many physically separate inference clusters to be managed and used as though they were a single logical environment. Enter a new suite of specialized services that enable AI-native connectivity by combining IP and optical networking with deterministic performance, assurance, security, reliability and customer visibility.

Opportunity at a glance

Dimension	Description
Demand catalyst	Rapid growth in AI applications as the market moves from human-to-AI interactions into agentic and physical AI.
Infrastructure shift	A growing share of deployment moves from centralized training toward distributed inference.
Deployment constraint	Large metropolitan facilities face power, cost, construction, permitting and community barriers.
Network requirement	Specialized IP-optical connectivity that unifies smaller, distributed inference clusters.
Proposed market	An incremental market estimated in the meeting at \$10-\$15 billion, subject to further validation.
Commercial outcome	A repeatable service model for telecommunication providers, supported by Nokia technology and potentially adopted by multiple providers.

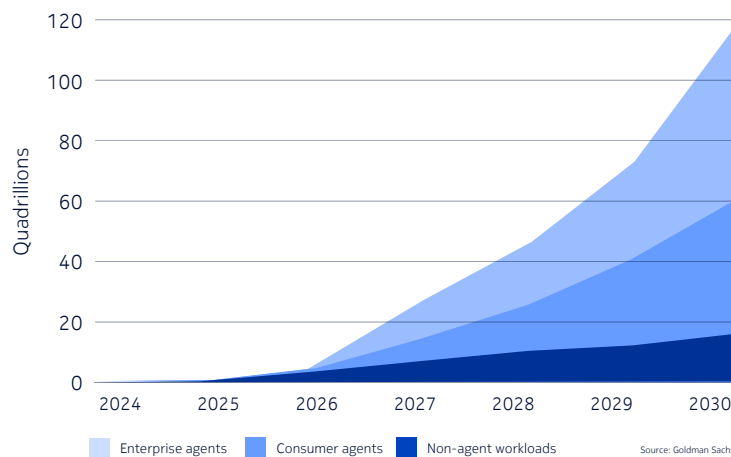
Contents

Executive overview	2
Market trends reshaping AI infrastructure	4
The deployment problem: Metropolitan inference at scale	6
New services for a new era of AI	8
How Nokia can help	10
Conclusion	11

Market trends reshaping AI infrastructure

From human-to-AI to agentic and physical AI

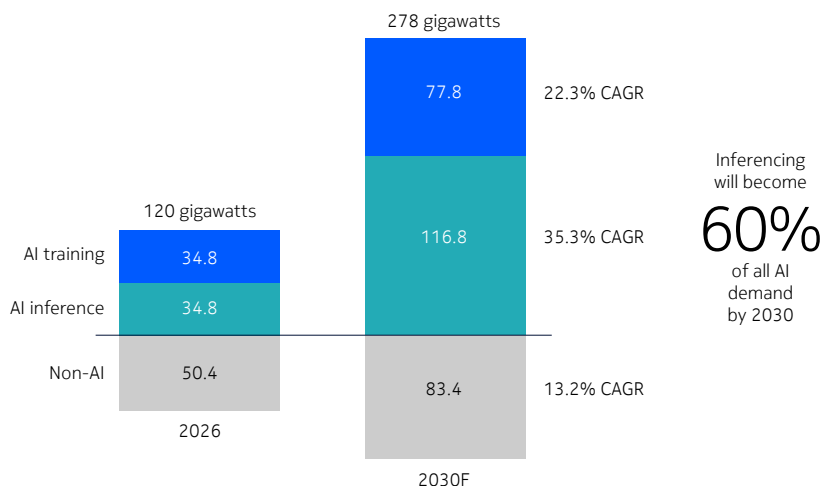
The AI supercycle has been running strong for nearly 5 years. During that time, most of the focus has been on building large training centers and creating large language models. Applications to date have been focused primarily on generative activities such as chat bots. These applications have resulted in a relatively small demand for token processing with the total number of tokens being processed per month at around 5 quadrillion. However, as we enter the agentic era, with many specialized agents communicating with other agents to complete tasks, we are seeing a sharp rise in the amount of tokens being processed. Physical AI is right around the corner and will add connected machines and devices that create their own workloads, pushing that number even higher. Current projections anticipate token demand rising to more than 120 quadrillion tokens per month by 2030 – that’s more than 80% CAGR over the next 5 years.



The next wave of AI compute resource deployments

Processing those tokens requires a major increase in compute resources. While incremental training resources are projected to grow at 22%, deployment of inferencing resources to process those 120Q of tokens per month is projected to grow at 35% CAGR, growing from 34GW in 2025 to more than 115GW in 2030.

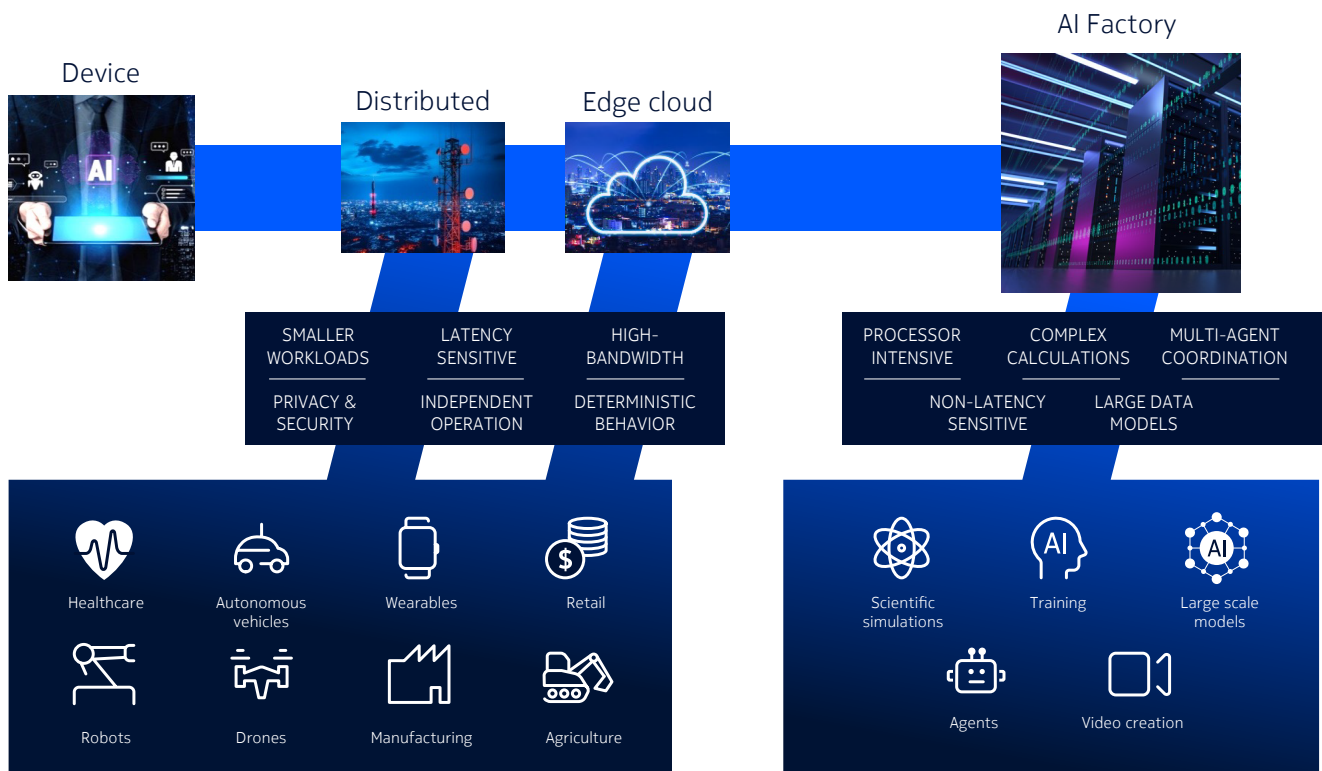
Demand by workload type (continued momentum scenario) %



Source: <https://www.mckinsey.com/industries/semiconductors/our-insights/scaling-ai-inference#/>

Compute moves closer to demand

As inferencing resource deployments begin to take center stage, deployment strategies will also change. Unlike training resources, which are typically deployed in large facilities in remote locations that can support massive numbers of accelerators, inferencing resources often benefit from and are sometimes required to be deployed closer to the end user. For high-bandwidth uses, such as processing thousands of high-definition camera streams or analysis of numerous large MRI files, placing compute nearer to the data source will significantly reduce the cost of transporting data. Additionally, local placement also helps accommodate data sovereignty, privacy and sector-specific handling requirements. And finally, having the processing closer to the end-user will provide an overall better user experience for latency-sensitive applications.



The deployment problem: Metropolitan inference at scale

There are several challenges with bringing large amounts of compute resources closer to end-users.

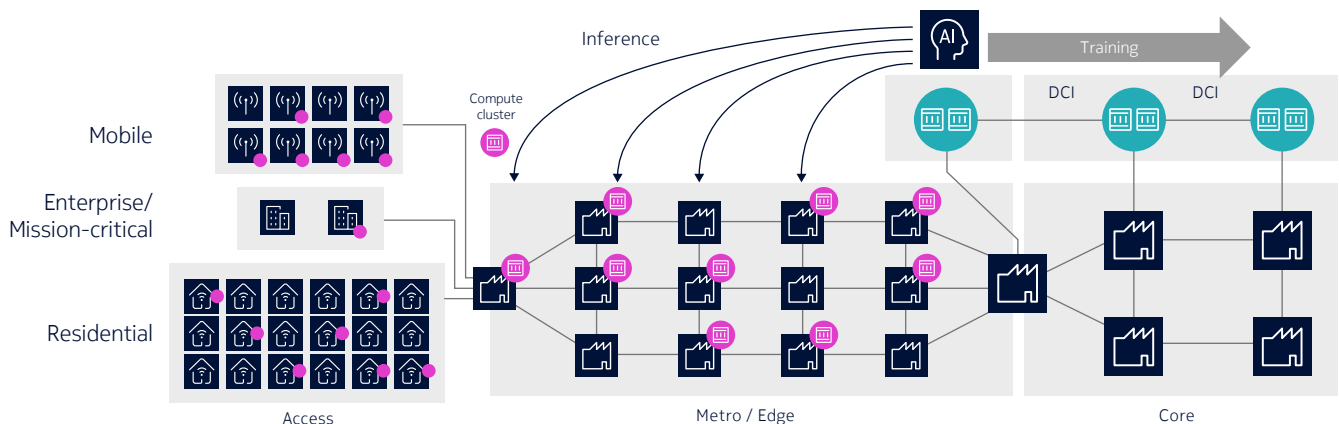
Large facilities are costly and difficult to build

A new large data center can carry substantial construction costs and must also clear permitting, environmental and community requirements. Construction of a new data center can range from \$30M to \$60M per MW, putting the cost of a single 100MW facility in the range of \$600M. Additionally, construction can take 3-5 years before the deployed resources can serve end-customers and begin generating revenue. Beyond the cost and time, opposition to large new data centers can make projects difficult to approve, particularly near major population centers.

The path to deploy the necessary inference resources

The alternative to large new constructions is to place smaller clusters wherever there is usable space and power, including warehouses, central offices, or commercial buildings. The result is that same 100MW of inference resources could be broken into 1-5MW clusters and deployed to numerous locations around a single geographical region.

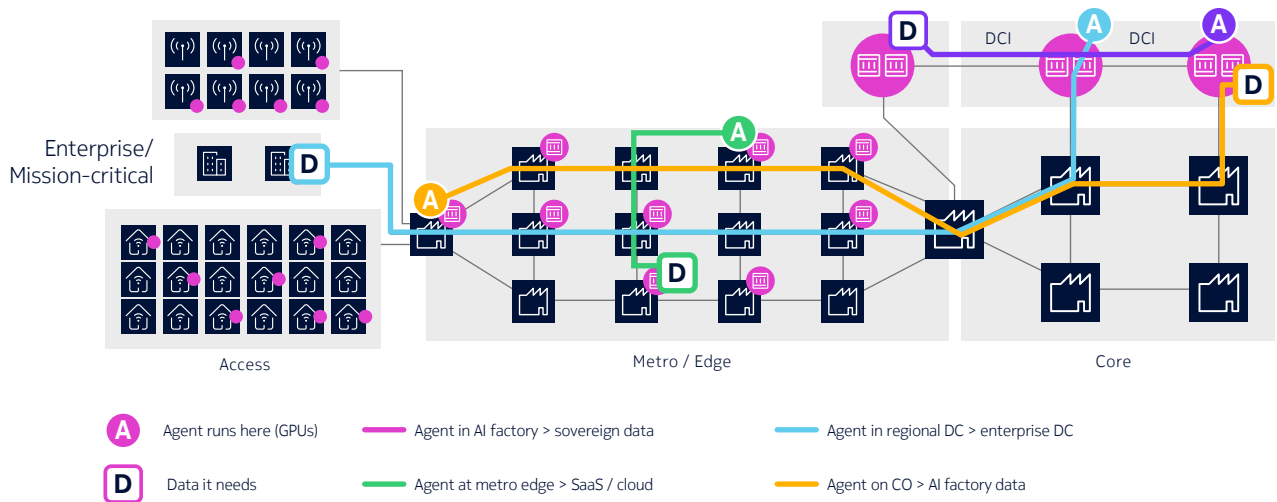
Inferencing clusters will become increasingly distributed



Distribution creates a new operational gap

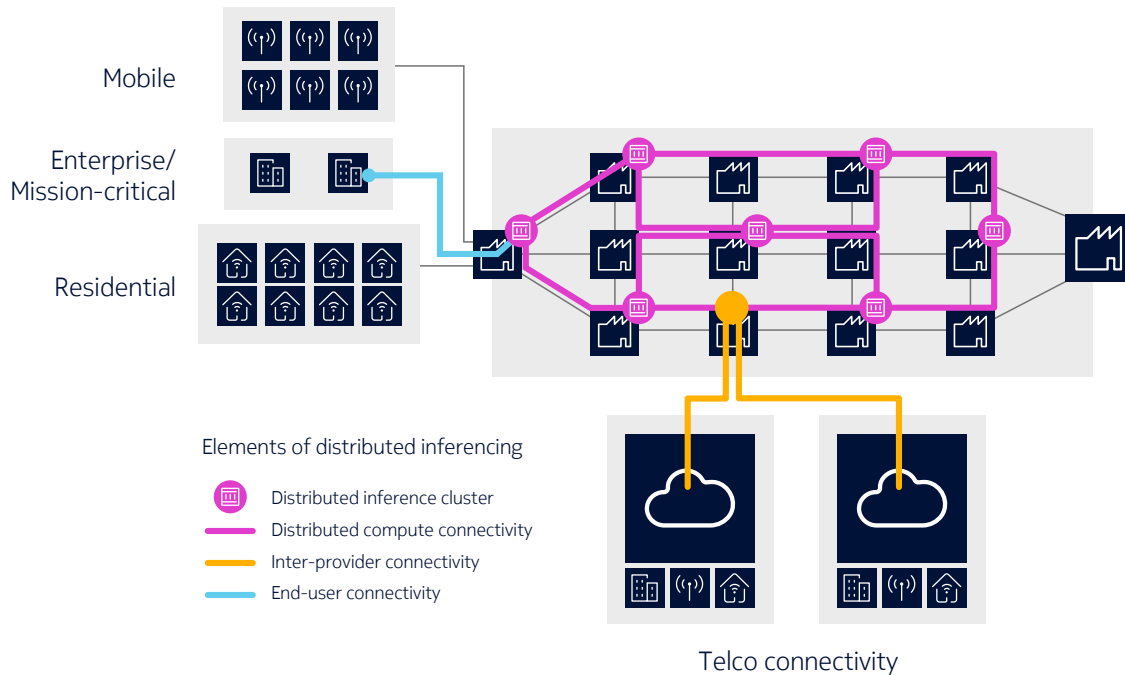
While smaller, distributed clusters can overcome deployment constraints, they create a new challenge: coordinating workloads across multiple independent compute clusters. Even a single workload may require multiple agents operating in different clusters, each accessing data from additional distributed locations.

Multi-agent workload processing over a distributed infrastructure



New services for a new era of AI

Telecommunication providers can play a key role in solving the challenges associated with distributed inferencing. They provide physical locations to host distributed inference resources and network infrastructure. They also have expertise to connect those locations and trusted relationships with end users.



However, traditional connectivity services will not be sufficient to provide the performance and behavior necessary for AI and cloud providers to host new and complex agentic-driven applications. To meet this need, a new suite of AI-ready connectivity will be required to connect these inferencing locations together. This requires a new approach for connectivity called “Inference connectivity.” Inference connectivity delivers new levels of scalability, security, availability, latency and agility which enable the distributed clusters to operate as a single federated system.

To effectively enable distributed inference, several new services will be required to provide the following capabilities:

- Connectivity between distributed inference clusters
- Connectivity to share distributed inference cluster with other service providers
- Connectivity to distributed inference clusters to end users (fixed and mobile)

The following table describes the network challenges and the key attributes of a networking solution.

Challenges	Description	Infrastructure requirement(s)
Geographically distributed compute resources	AI services must connect AI factories, regional sites, metro facilities, and edge clouds into a unified service platform.	Programmable, open, multi-domain, scalable and efficient
Very large inter-site traffic demands	Continuous exchange of prompts, data, model context, KV cache information, telemetry and workload state drives significant bandwidth demands between sites.	Scalable and efficient
Latency-sensitive AI applications	Agentic AI, physical AI and real-time applications require predictable response times regardless of where compute resources are located.	Deterministic and SLA-driven
Service continuity across failures and traffic shifts	AI services cannot tolerate outages, reconvergence delays or disruptions that interrupt active inference sessions and workflows.	Resilient
Data sovereignty, security and multi-tenancy	AI workloads frequently cross organizational, cloud and geographic boundaries while requiring strict isolation, privacy and regulatory compliance.	Secure, isolated and sovereign
Dynamic workload placement across clouds and domains	AI resources may move between sites, enterprises, neoclouds, hyperscalers and service providers, requiring on-demand connectivity and automation.	Programmable, open and multi-domain

An ecosystem model, not a proprietary island

Bringing these new services to the market will require strong partnerships between telcos, AI and cloud providers and network infrastructure suppliers. Services must be designed to meet the specific operational demands of each AI and cloud provider and network solutions must support the necessary features and capabilities to enable those services.



How Nokia can help

For generations, Nokia has designed, deployed, operated, and secured some of the world's largest and most demanding networks. From telecommunications infrastructures and internet backbones to hyperscale cloud connectivity and mission-critical enterprise networks, Nokia has consistently delivered the scale, reliability, and performance required during major industry transitions.

As networking requirements have evolved, Nokia has repeatedly helped customers navigate technology inflection points, including broadband transformation, IPTV, cloud adoption, over-the-top (OTT) services, mobile broadband, 5G, and cloud-native networking. Distributed AI represents the next major infrastructure transition, and Nokia is uniquely positioned to lead it.

A complete networking portfolio for the AI era

Distributed inference requires infrastructure that extends well beyond the boundaries of a single data center. AI services depend on seamless connectivity between compute, storage, models, users and data sources across multiple locations and domains.

Nokia offers one of the industry's most comprehensive networking portfolios, including:

- High-performance IP routing platforms for AI-ready connectivity.
- Optical transport systems for high-capacity inter-data-center connectivity.
- Coherent pluggable optics that simplify deployment while maximizing capacity and reach.
- Complete security solutions, including encryption and DDoS protection, to safeguard AI infrastructure, data and services.
- End-to-end network automation, assurance, and operational platforms.

This allows customers to build a consistent AI network across regional, metro, and edge data centers and across national and global infrastructures.

Leadership in network segmentation and multi-tenancy

Distributed inference introduces new requirements for workload isolation, tenant separation, security, and data sovereignty. These capabilities must extend across enterprises, clouds, operators, and AI service providers.

Nokia has been a pioneer in networking technologies that create secure, scalable, multi-tenant infrastructures, including MPLS, VPLS, Segment Routing, SRv6, EVPN, Layer-2 and Layer-3 VPN services, and network slicing and service-based architectures.

These technologies form the foundation for secure AI service delivery, allowing solutions providers to isolate workloads, enforce policy boundaries, satisfy regulatory requirements, and securely connect distributed AI resources across multiple domains.

Built for the AI inferencing infrastructure

The future of AI will be defined by the ability to federate distributed resources into a unified intelligent system. Success will require networking that is scalable, deterministic, resilient, secure, and programmable.

Nokia's combination of decades of networking expertise, industry-leading routing and optical innovation, advanced security technologies, automation platforms, and open networking architectures makes it uniquely positioned to enable the next generation of AI infrastructure.

As distributed inferencing becomes the dominant model for AI service delivery, Nokia provides the networking foundation that transforms geographically distributed compute resources into a seamless federated system, allowing intelligence to be delivered wherever it is needed.



Conclusion

AI's greatest promise is not simply greater autonomy and intelligence, but its potential to create a more responsive, accessible, and personalized world. AI can help clinicians detect disease earlier and tailor treatments to individual patients, give students adaptive learning support, improve mobility through safer transportation and intelligent traffic management, make workplaces more productive by automating routine tasks, and help communities use energy, water, and public services more efficiently.

Realizing these benefits at scale will require the network to evolve from a best-effort transport layer into an intelligent, distributed service platform that connects users, devices, data and compute wherever they are located. It must provide high-capacity and energy-efficient connectivity, predictable low latency, continuous resilience, strong security and tenant isolation, data-sovereignty controls, and open programmability across cloud, metro, edge and access domains. By enabling AI resources to operate as a coordinated system, the network will become the foundation that turns rapidly advancing AI capabilities into trusted services that improve daily life.

About Nokia

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

Nokia is a registered trademark of Nokia Corporation. Other product and company names mentioned herein may be trademarks or trade names of their respective owners.

© 2026 Nokia

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

Document code: CID215597 (October)