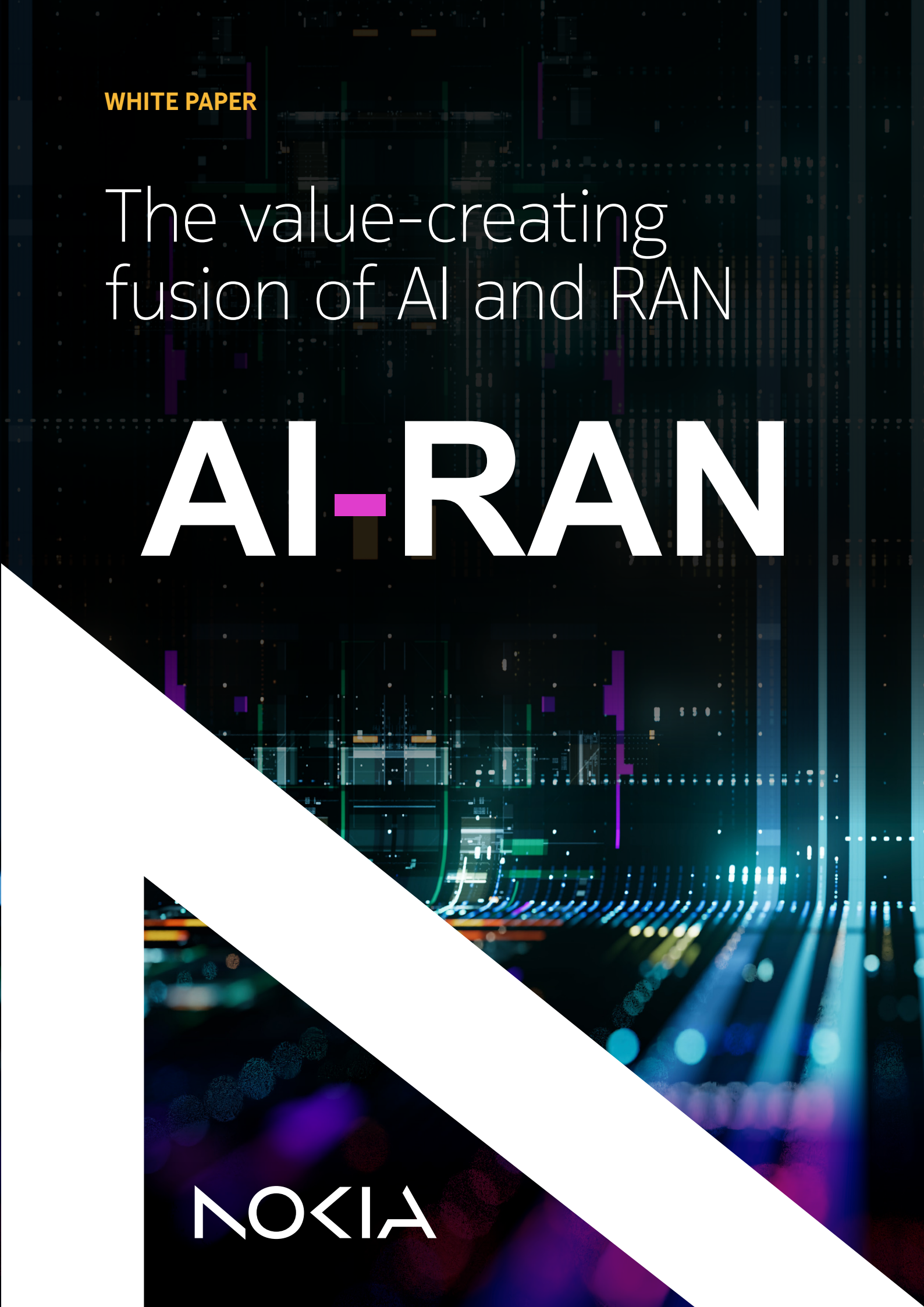
WHITE PAPER

The value-creating fusion of AI and RAN

AI-RAN



NOKIA

Executive summary

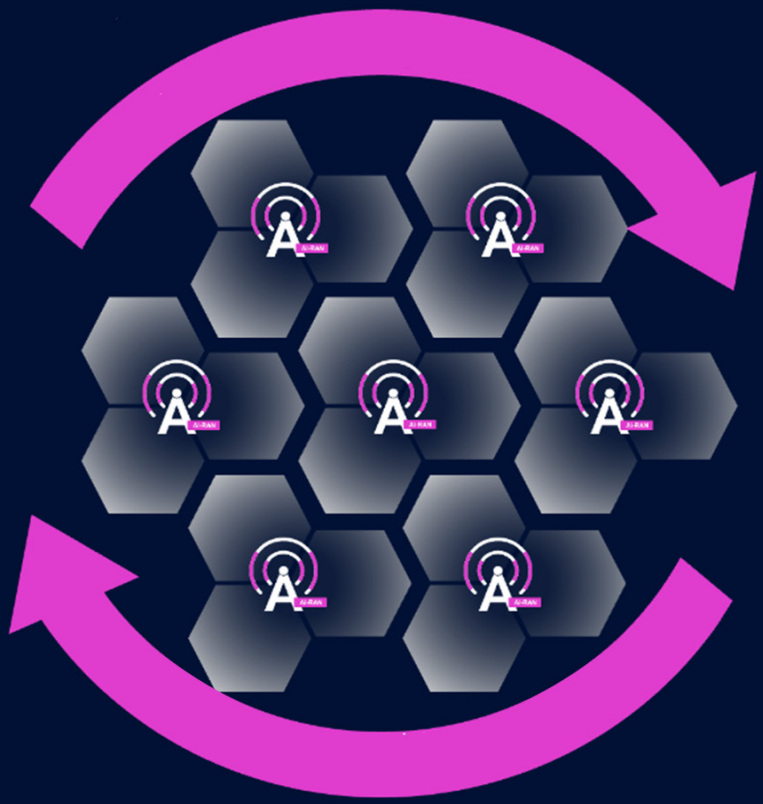
Artificial intelligence is becoming increasingly integrated into the services that we consume and that enterprises rely on. This will not only change how we interact with our devices and the digital world but also impact how we build radio access networks (RAN), the critical infrastructure that powers wireless services. On the one hand, AI applications will benefit from enhanced RAN capabilities and on the other hand, RAN itself will benefit from AI. A value-creating fusion of AI and RAN is emerging: AI-RAN.

With AI-RAN, the deeper and wider use of AI is stepping alongside openness and cloud technologies to enhance RAN. Nokia is driving this evolution. We are a founding member with leading roles in the AI-RAN Alliance and work closely with the wide industry ecosystem to research and develop the technologies that enable leveraging AI-RAN's potential.

As generative AI, agentic AI and robotic AI become more widespread, token throughput will become an increasingly important differentiating factor of mobile broadband networks. Token throughput measures how efficiently AI in user devices, robotic systems and inference factories can process, exchange and generate tokens as well as input and output data. The efficient, fast and reliable exchange of AI inputs, AI tokens and AI outputs will become a critical part of societies and economies and depend on mobile broadband networks.

In this whitepaper, we first explore the emerging landscape of the connected AI world and highlight the pivotal role of wireless connectivity in enabling agentic and robotic AI. In the latter part, we focus on AI-RAN itself, beginning with introductory chapters followed by more in-depth discussion of the three foundational, well-established pillars of AI-RAN: AI-on-RAN, AI-for-RAN and AI-and-RAN.

Innovations in these three areas will unlock new business opportunities, further efficiencies, and synergetic effects for the investments needed to transform RAN into the most widely distributed AI computing grid.



Introduction to the connected token world

Spurred by generative AI, tremendous progress has been made in AI models and related computing since around 2017. Investments and adoption have followed in many areas. Generative AI for human users is evolving to agentic AI, where domain-specific, small language models (SLM) and larger, foundational language models (LLM) communicate with each other. This multi-LLM approach can help ensure data privacy and data ownership while tapping the full potential of the latest and greatest LLMs.

Additionally, small language models can run on devices such as smartphones and emerging XR glasses for human users, as well as robotic systems in manufacturing, autonomous vehicles and humanoid robots, which assist human workers and people in need of help with daily tasks. Immersive user experience and many robotic applications require very short control cycles. Device affordability and battery life both limit the computing power on the device, which in turn limits the size of the AI models on the device. The immersive XR and robotic devices will feature smaller domain-specific AI models that communicate with other models through fast and reliable wireless broadband connections.

The communication between LLMs is based on tokens, abstractions of meaningful pieces of information. Tokens can be simply parts of words, as in text prompting, but can also be optimized for images, voice and sounds, video or control systems. As human communication did not culminate with the use of the telephone, we will see multi-modal LLMs, which communicate with each other using different sorts of data representations and tokens. While an individual token is small, we cannot yet predict the number of tokens that will be exchanged. Traditional traffic volume estimates that are based on the number of human users and their activity will not work anymore when an unknown number of autonomous machines and agentic systems exchange tokens.

The exchange of tokens between human users' devices and robotic systems through mobile broadband networks will make mobile networks an ever more valuable and critical part of societies and economies. Just like mobile broadband added new use cases and business opportunities to cellular networks about 20 years ago, we will now see token throughput enabling another, fundamentally new wave of use cases and business opportunities.

The introduction and adoption of mobile broadband took approximately two decades, from early visions in the mid-1990s to the widespread adoption of 4G by the mid-2010s. At the beginning, the value of mobile networks was in mobile voice. With 4G, the value shifted to mobile broadband data packages, which allowed people to watch videos, engage on social media, get relevant information and stay in touch with each other and the world.

We are now on the verge of what could become the second big value transformation of mobile networks. Initially, the volume of token exchange will be small compared to mobile broadband. Over time, fast, secure and reliable token exchange will grow to huge economic value. Token throughput will become the "new currency" of mobile networks, as Alex Choi, head of the AI-RAN Alliance, has phrased it [1].

We use the term "token" here as a placeholder for all sorts of data that is typically exchanged with and between AI systems. Those include tokens in text prompts, conditioning inputs such as images and patches, or abstractions such as embeddings and features. Their reliable and fast exchange is foundational for value-creating, interworking AI agents and robotic systems.

Figure 1 illustrates the drivers of value transformation in mobile networks in the 21st century.

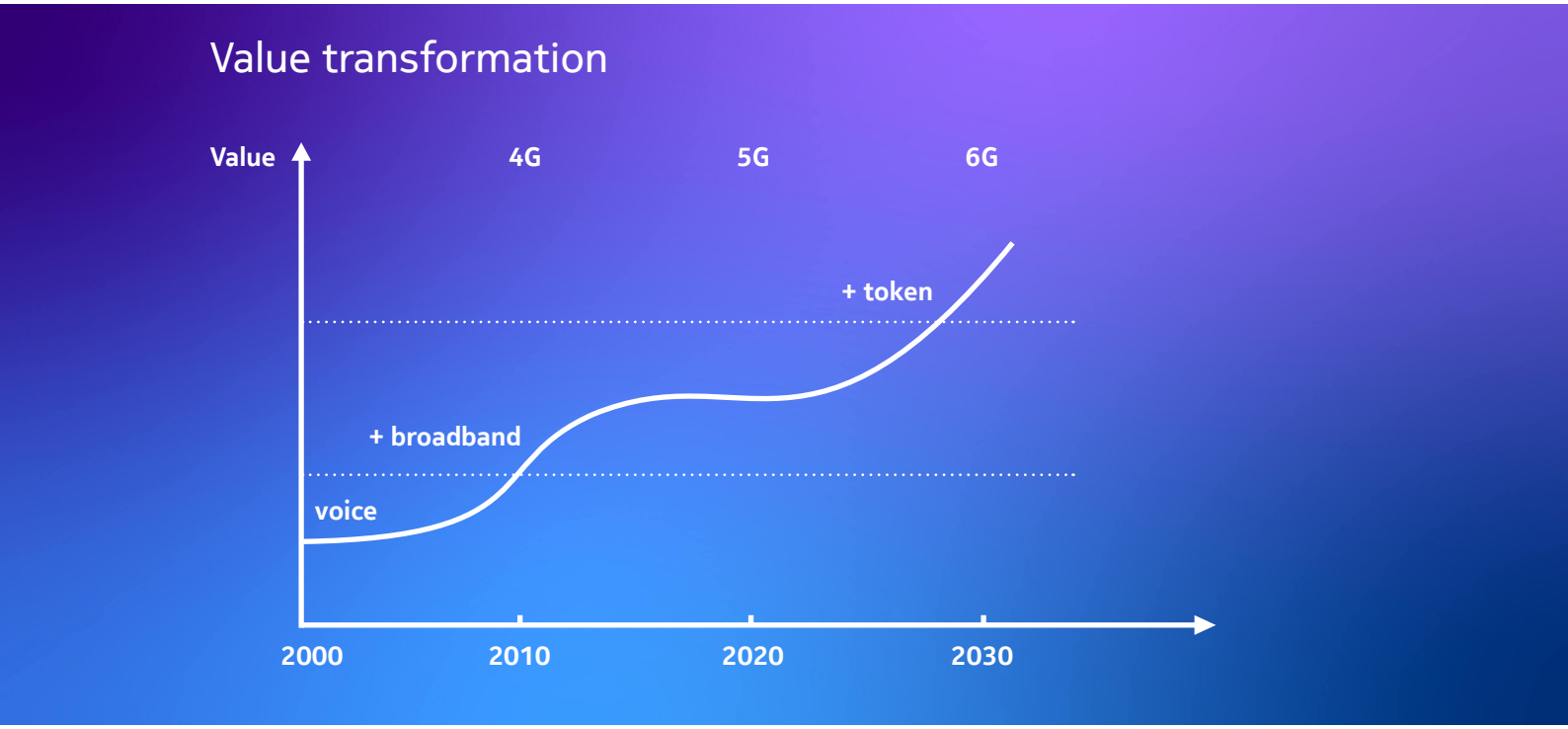


Figure 1. Value transformation drivers in mobile networks

Toward multi-agent, multi-device AI

Most of the AI applications we engage with today have two things in common: a human triggers the inference and the response time is multiple seconds or longer. Consequently, the large language models used by generative AI reside in large AI factories, where processing power, memory and power supply can scale up efficiently. The AI factories are operated by established hyperscale companies or companies specializing in AI-specific cloud computing, referred to as neocloud providers.

Over time, this centralized approach to AI will change; central AI factories will be complemented by more AI processing on devices and at the network edge. Among the drivers for this development are data privacy concerns, latency requirements, data transmission optimization and device capabilities.

Smartphones and wearables set some limits on the size of the AI models running locally, both in terms of processing speed and in terms of memory size. These devices must be energy-efficient enough to run an entire day with a single battery charge and must remain affordable for the consumer mass market. While over time the devices become more powerful, they will never be able to compete with the computing power of the latest server farms. Instead, consumer devices will run some smaller AI models locally and communicate with the largest, most advanced and most comprehensive AI models on AI factories that hyperscalers or sovereign neocloud companies provide.

The same is true for robotic AI, which will drive the number of AI devices far beyond the number of devices used by humans. The pressure for cost efficiency, the persistent target of long device lifecycles and the necessity to leverage the latest AI innovations will drive the need for unified connectivity for robotic AI. Cellular networks are the optimal connectivity solution.

Figure 2 illustrates how the centralized AI factory approach will be complemented by local AI processing on devices or the edge cloud.

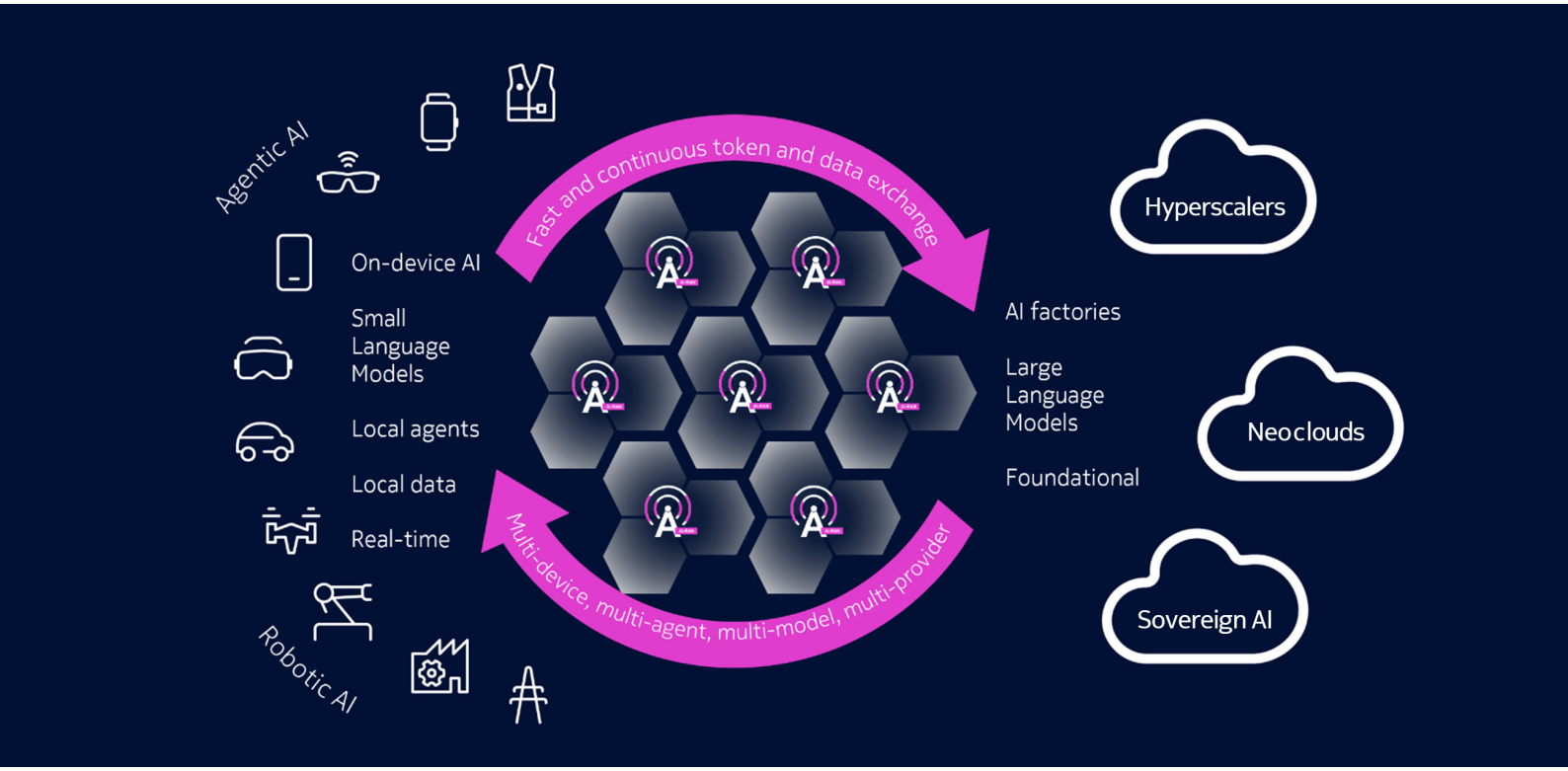


Figure 2. Local and centralized AI processing

Access to the latest and most trustworthy AI also means that an AI agent on a device will communicate with not only one large language model but with multiple large language models, as these provide potentially different answers to the same prompt or display expertise in different areas. Some models are better at reasoning, while others might have larger context windows or a more recent training cut-off date. Reasoning-based approaches are more like a dialogue between AI agents and models than just a simple one-shot prompt-response sequence. In combination with a geographic distribution of involved devices and servers, this leads to further token exchange.

Multi-agent and robotic AI workflows are decoupled from human busy hours. Robots are designed to work 24/7, so a possible implementation for complex AI processing is that on-device agents can communicate overnight with large language models to have results for the device owner the next morning.

All these trends drive a massive token and data exchange between consumer and industrial devices and AI factories, which generates significant downlink and uplink traffic in radio access networks. Already today, we can observe an acceleration of the uplink traffic growth in mobile networks driven by AI apps on smartphones.

Token exchange impact on mobile networks and AI-RAN

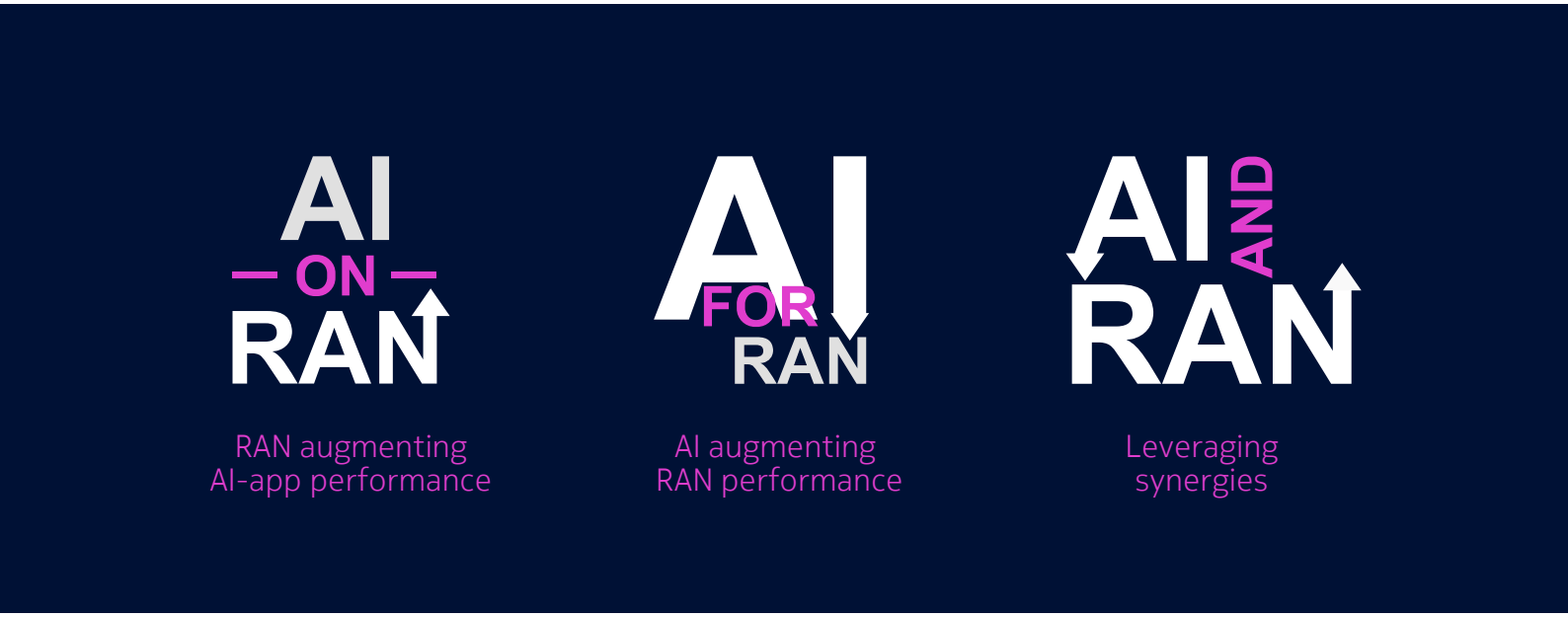
While mobile broadband is generally a best-effort business, token throughput comes with various requirements for delay budget for the first and subsequent tokens and network availability. Consequently, there will be a need for more and more specific 5G slices and supporting network capacity.

Immersive experiences and robot control cycles will define the allowed delay budget for end-to-end latency, consisting of the network latency and AI computing latency. We can reduce the RAN latency with 5G Ultra-Reliable Low Latency Communications (URLLC) technology, and end-to-end network latency by bringing AI workloads to the network edge. This will allow running larger and therefore more precise and comprehensive LLM models within the delay budget.

The return on investment (ROI) of edge AI computing can be accelerated by sharing the AI computing resources with Cloud RAN. Today, Cloud RAN itself is among the network edge systems with the shortest delay budget. Cloud RAN will develop or merge into a multi-purpose computing platform, perfectly located for latency-limited AI workloads.

Context is key for many LLM applications. With 6G, networks will integrate communication and sensing capabilities. While today's networks know about device location, geographic device density and usage intensity, 6G sensing will add environmental context information. This context could further enhance the decision-making of self-driving vehicles and robotic systems. The exposure of network and sensing context data is an area of ongoing exploration, both from a 6G perspective as well as from an AI-on-RAN perspective.

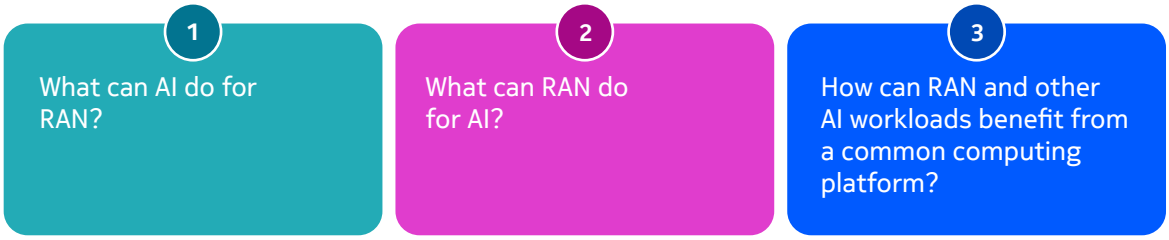
AI-for-RAN, AI-and-RAN and AI-on-RAN each drive efficiency, performance and capabilities that benefit both mobile broadband use cases and token throughput. Together, the three areas will unlock further synergies and business potential for mobile network operators as they embrace the next value transformation.



The fundamentals of AI-RAN

What is AI-RAN?

AI is a very wide field and so is RAN. To approach the mutual benefits, we ask three basic questions:



These questions lead directly to the well-established AI-RAN framework, which consists of 1) AI-for-RAN, 2) AI-on-RAN, and 3) AI-and-RAN.

First, let's consider what AI can do for RAN. When AI augments RAN performance, efficiency or capabilities, we call it AI-for-RAN. AI-for-RAN is not limited to specific AI models, computing hardware or computing location. Initial AI-for-RAN functionalities have already delivered double-digit gains. We explored this area in more detail in our previous whitepaper, AI for Radio Access Networks [2].

In terms of latency, low-layer RAN functionalities typically have delay budgets in the range of microseconds that mandate execution at or near the cell site, while high-layer functionalities often feature somewhat more relaxed delay budgets that allow for execution at more centralized locations. 6G will bring AI-for-RAN to the next level with the introduction of an AI-native design to RAN.

Second, let's explore what RAN can do for AI. Most of today's consumer and enterprise applications already feature some AI capabilities. We already see an impact of AI apps on RAN traffic; most notably, the relatively faster growth of uplink traffic. Over time, the width and depth of AI in applications will increase with the progress of AI models and AI computing. When RAN augments the AI application performance or capabilities, we call it AI-on-RAN. AI-on-RAN has the potential to enable new or enhanced consumer and enterprise applications, with RAN providing enhanced capabilities and exposing data to these AI applications.

Third, let's discuss the benefits of a common computing platform for RAN and other AI workloads. The key to lowering the cost of computing is to remove unproductive idle times and to increase the average utilization rate of computing hardware. From a RAN perspective, a potential use case is utilizing RAN computing power for other AI workloads when mobile traffic is low. From an Edge AI perspective, RAN can be seen as just another workload. Regardless of the perspective, we call this area of seeking mutual synergies AI-and-RAN. In the long term, the decoupling of AI inference from direct human prompts and the partial transition of centralized AI inference factories to the far edge can propel the synergetic effects between AI and RAN.

Sharing AI computing resources between RAN and other AI workloads will ensure high resource utilization and help reduce the total cost of ownership (TCO) per use case. Offering spare capacity under a Platform as a Service (PaaS) model can further accelerate the break-even point for mobile network operators, which is important considering the pace with which AI computing currently progresses.

There is a symbiotic relation between these three areas: AI-and-RAN facilitates the build-up of powerful AI computing capabilities, which enables enhanced AI-for-RAN model inference. This, in turn, unlocks throughput, latency and capabilities for advanced mobile broadband use cases and token exchange. The token exchange, together with data enablers from AI-on-RAN, will unlock new AI application capabilities for consumers and businesses, of which some will run on the AI edge built with AI-and-RAN.

Why is AI-RAN needed?

AI-RAN has huge potential to optimize RAN TCO:

- AI-for-RAN can reduce TCO by enabling higher levels of automation and increased resource efficiency, including more efficient spectrum utilization and energy savings.
- AI-on-RAN can enable AI applications to adapt to the network conditions. An example is adjusting their throughput demands during peak traffic.
- AI-and-RAN has the potential to reduce the cost per computing task and the network upgrade costs.

AI-RAN can help unlock new monetization opportunities:

- AI-for-RAN can enhance the capacity and capabilities of the network to meet the requirements of increasing mobile broadband and token traffic.
- AI-on-RAN can inject data and insights that enable new use cases and business opportunities.
- AI-and-RAN can reduce end-to-end latency of AI model inference and the network, enabling new agentic and robotic AI applications.

Together, the three areas of AI-RAN have high potential to increase mobile operators' profitability and to augment the value of private wireless networks. Figure 3 illustrates the benefits.

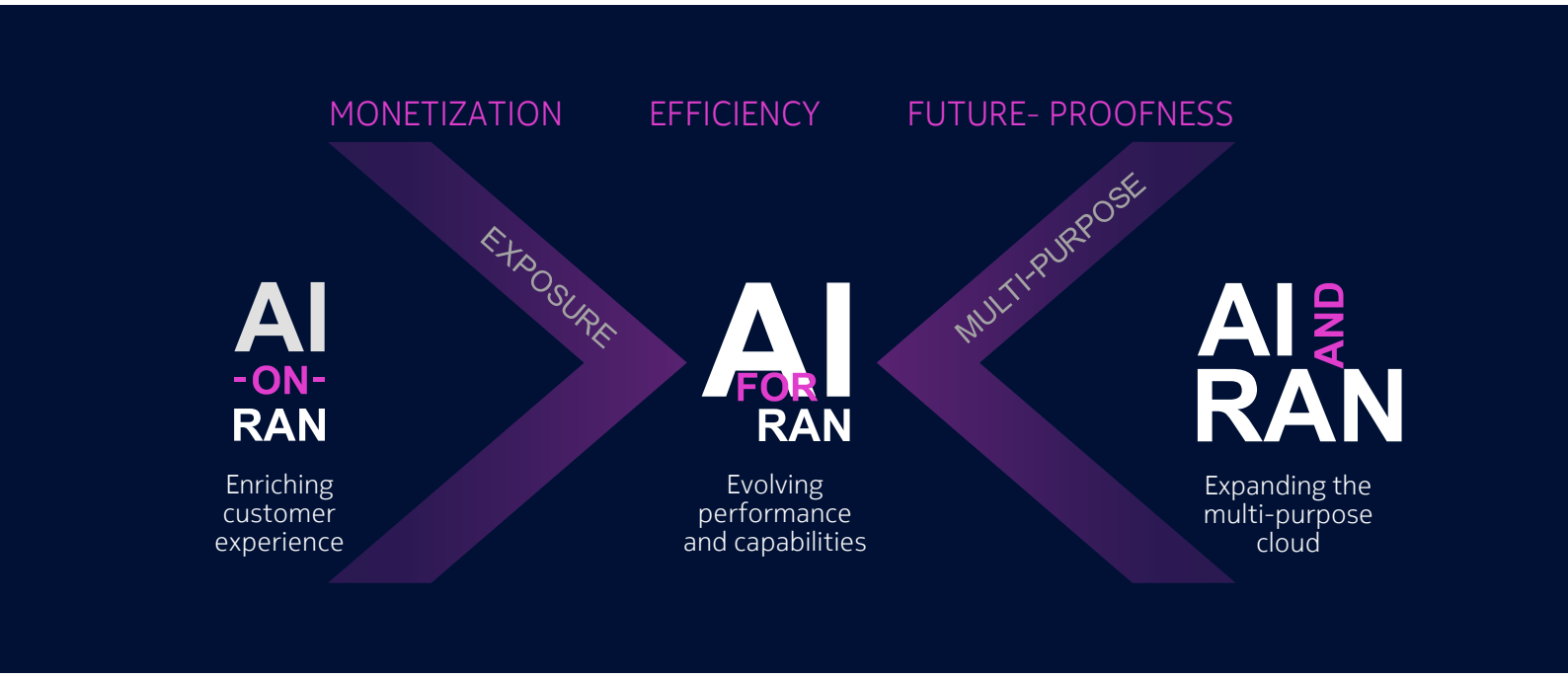


Figure 3. The benefits that AI-RAN enables

How to make AI-RAN a success?

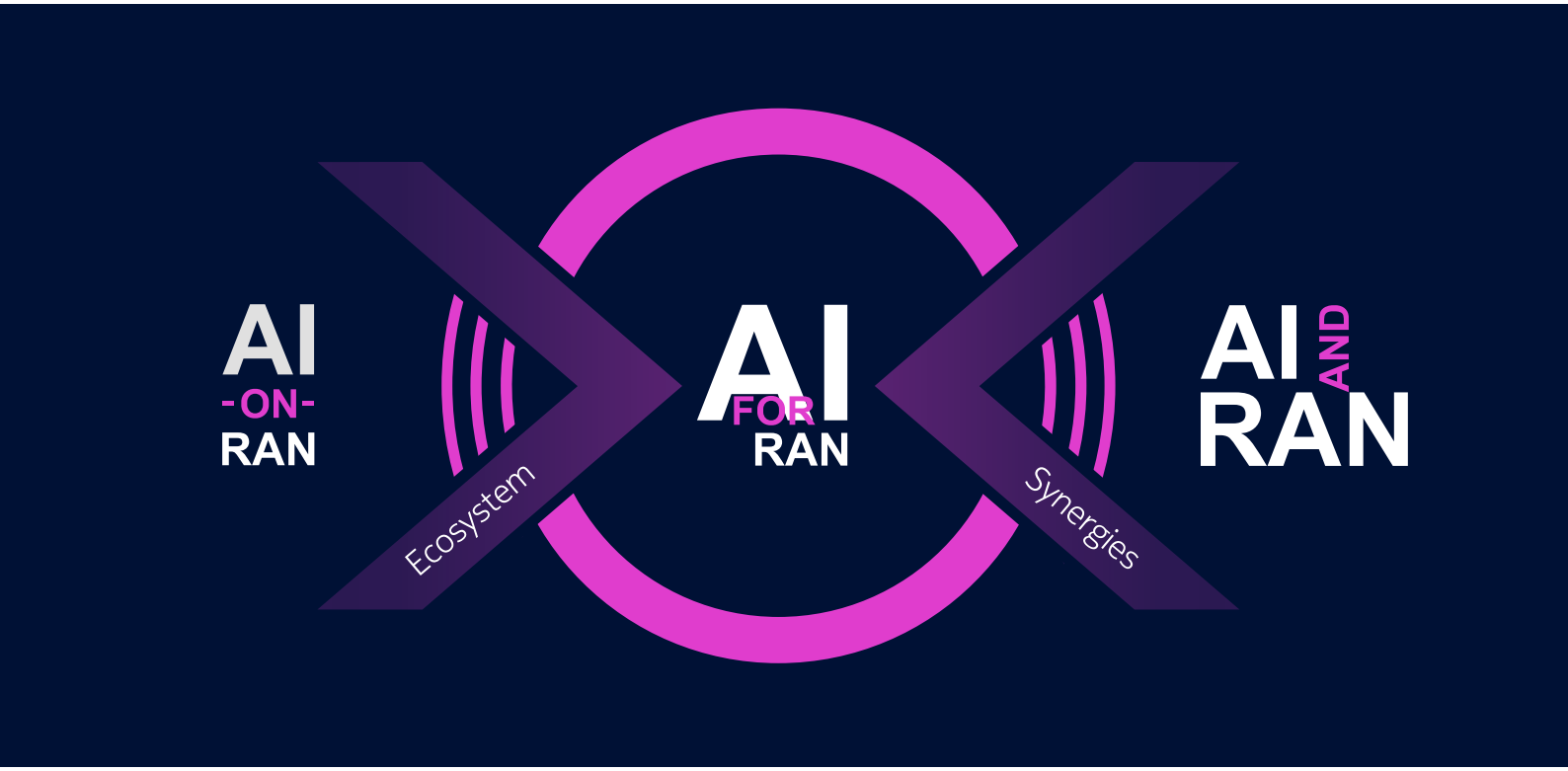
AI-RAN can only be successful if data, computing power, AI models and RAN expertise come together. This requires close collaboration of leading technology companies with their respective domain knowledge, similar to Nokia's anyRAN approach to Cloud RAN.

Collaboration is the foundation of AI-and-RAN, which will lead to the integration of Edge AI and Cloud RAN. To identify smart investment strategies, it takes a thorough understanding of the predicted AI computing power needs and PaaS market dynamics. As RAN sites cover nearly the entire population and most geographical areas in many countries, RAN provides an optimal foundation for expediting time-critical or data-sensitive AI inference from central AI factories to the locations where end-users reside and enterprises operate.

The AI-on-RAN collaboration will eventually bring together multiple partners. The goal is to enable a global ecosystem of AI applications that operate across networks built on technologies from multiple RAN suppliers. To achieve this, it's essential to establish shared foundations and position the RAN as a common platform toward the AI application ecosystem.

Nokia has an active role in driving industry collaboration:

- We are a founding member of the AI-RAN Alliance, which brings together mobile network operators, AI industry leaders and academia [3].
- In 2024, we announced AI-RAN collaboration with T-Mobile US [4], SoftBank in Japan [5] and NVIDIA [6].
- At Mobile World Congress 2025, we showcased a joint AI-RAN proof-of-concept with SoftBank and announced further collaboration with several operators [7].
- At our Tech Winter Horizon 2024 event, we brought together leading industry experts from mobile network operators and innovative technology companies to explore different technological aspects of AI-RAN and its potential to provide a future monetization platform for operators [8].



AI-for-RAN enables enhanced efficiencies and autonomous operations

Already today, AI-for-RAN addresses a multitude of mobile network operator challenges and opportunities along the entire RAN life cycle, which include:

- Discovery of previously hidden quality of experience (QoE) issues, inefficiencies and risks
- Prediction of demand growth or imminent failures for timely action
- Selection of the best options from site design to beam set level for optimization
- Enhancing spectral efficiency and resource utilization for increased capacity

Figure 4 illustrates concrete examples of the optimizations and enhancements that AI-for-RAN enables throughout the network lifecycle from planning and design to deployment, operations and maintenance.

Figure 4. AI-for-RAN enables enhanced efficiencies throughout the network lifecycle

Some AI functionalities have a real-time (microsecond) delay budget and, consequently, the inference needs to run within the base station. Other functionalities can be run more efficiently and effectively with a network-wide view from a centralized data center, either within the operator network or provided as a service.

Over time, the number of AI-powered functionalities and their related AI models is growing. This growth in breadth and depth of AI workloads requires base station designs that can scale up AI-computing power with the right granularity.

As the self-optimization of the RAN becomes more dynamic, we need to ensure an optimal balance between potentially conflicting optimization intents, such as user experience versus energy efficiency. In a commercial network, the AI-powered Nokia MantaRay SON coordinated the energy efficiency features in base stations, achieving a 6% average reduction in energy consumption, and up to 30% reduction at certain sites [9]. This is a great example of how agentic interworking of intelligent systems can shift the whole system closer to the optimum, which only advanced AI-powered solutions can accomplish without compromising user experience.

There is a clear trend in RAN operations towards intent-based autonomous operations. TM Forum's autonomous network framework defines six levels of autonomous operations from 0 to 5. While most operators today are at level 3, which is characterized by closed-loop automation of selected tasks, Nokia MantaRay AutoPilot reaches TM Forum's level 4 today, which means highly autonomous operations involving many individual tasks. Levels 4 and 5 of autonomous network operations are only possible with AI.

In 2025, MantaRay Autopilot performed 15,000 autonomous operations per hour with massive ongoing network traffic in a commercial RAN network. This degree of dynamic network optimization is simply not possible without AI. The resulting efficiency gains in the network were significant: a 60% reduction in cell outages and an increase of 10% in the peak throughput [10]. Already in 2024, MantaRay Autopilot delivered substantial operational enhancements in a customer's network during a mass event while overall traffic grew by 40% [11].

At our Midsummer Launch 2025 [12], we introduced MantaRay SMO, our AI-powered RAN automation solution. It extends the potential of MantaRay SON and MantaRay AutoPilot by enabling the integration, management and orchestration of O-RAN compliant rApps, which facilitates the application development for AI-for-RAN. Compliance with O-RAN standards future-proofs investments in platform and application development, both in terms of innovation speed and supply chain security.

6G will mark the transition from AI-powered features to an AI-native end-to-end network. Deep neural network-based transmitters and receivers are a key area of research for spectral efficiency in 6G. In 2024, Nokia demonstrated the world's first AI-native air interface prototype [13], which is a significant steppingstone toward 6G.

Figure 5 illustrates the architecture of the AI-native air interface prototype.

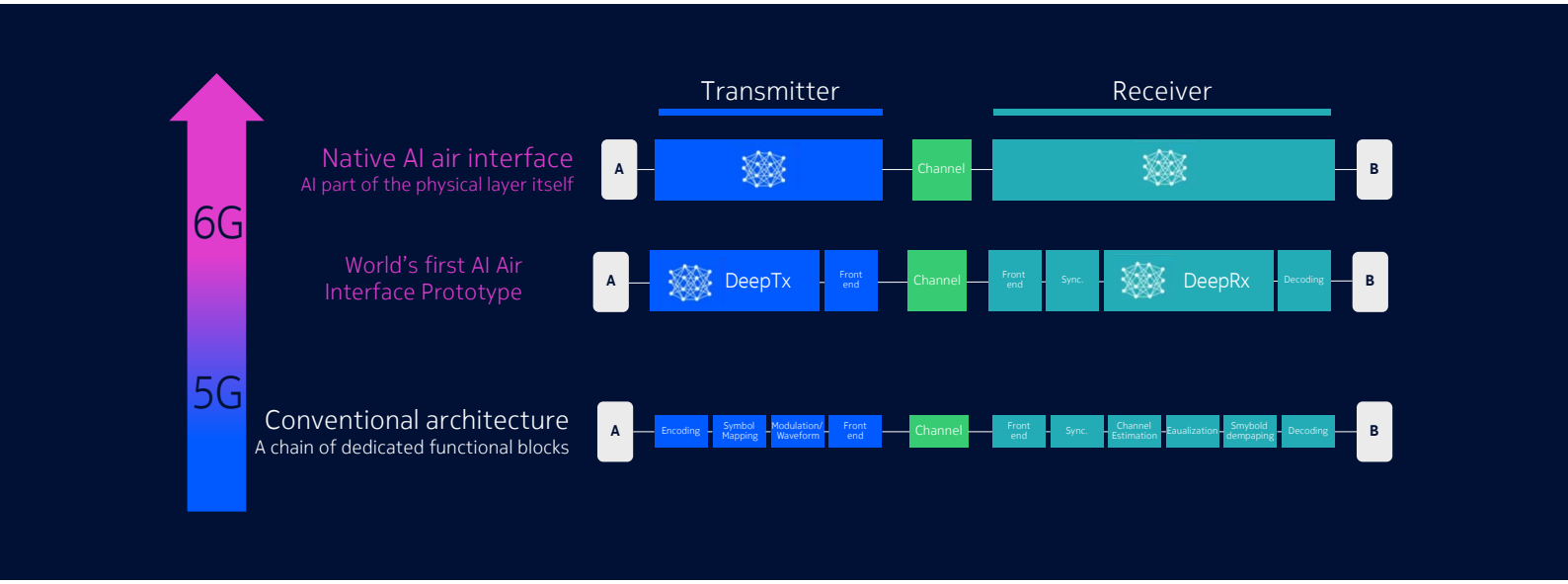


Figure 5. The AI-native Air Interface prototype introduced by Nokia Bell Labs

AI-on-RAN enables new and enhanced AI applications

Already today, the most frequently used AI apps on smartphones have a measurable impact on traffic volumes in mobile networks. Created with or without AI, video continues to dominate as a traffic driver. With the help of AI features, more people can express themselves with self-made videos. AI-generated and AI-enhanced viral videos increase downlink network traffic. However, the relative traffic increase in mobile networks driven by AI-enhanced applications is more marked in uplink than in downlink.

Based on Nokia's lab testing and analysis of commercial network clusters, wider and deeper use of the most popular AI apps of 2025 can lead to an uplink traffic growth of 10 to 70%. At an annual growth rate of 70%, total uplink data volumes could reach downlink levels in about five years. That would be a dramatic change from today's network reality, where capacity limits are first approached in the downlink, while coverage limits appear first in the uplink.

Radio access networks with good uplink coverage and sufficient flexibility to enhance uplink capacity are ideal enablers of enhanced AI application experience. But there is more to AI-on-RAN.

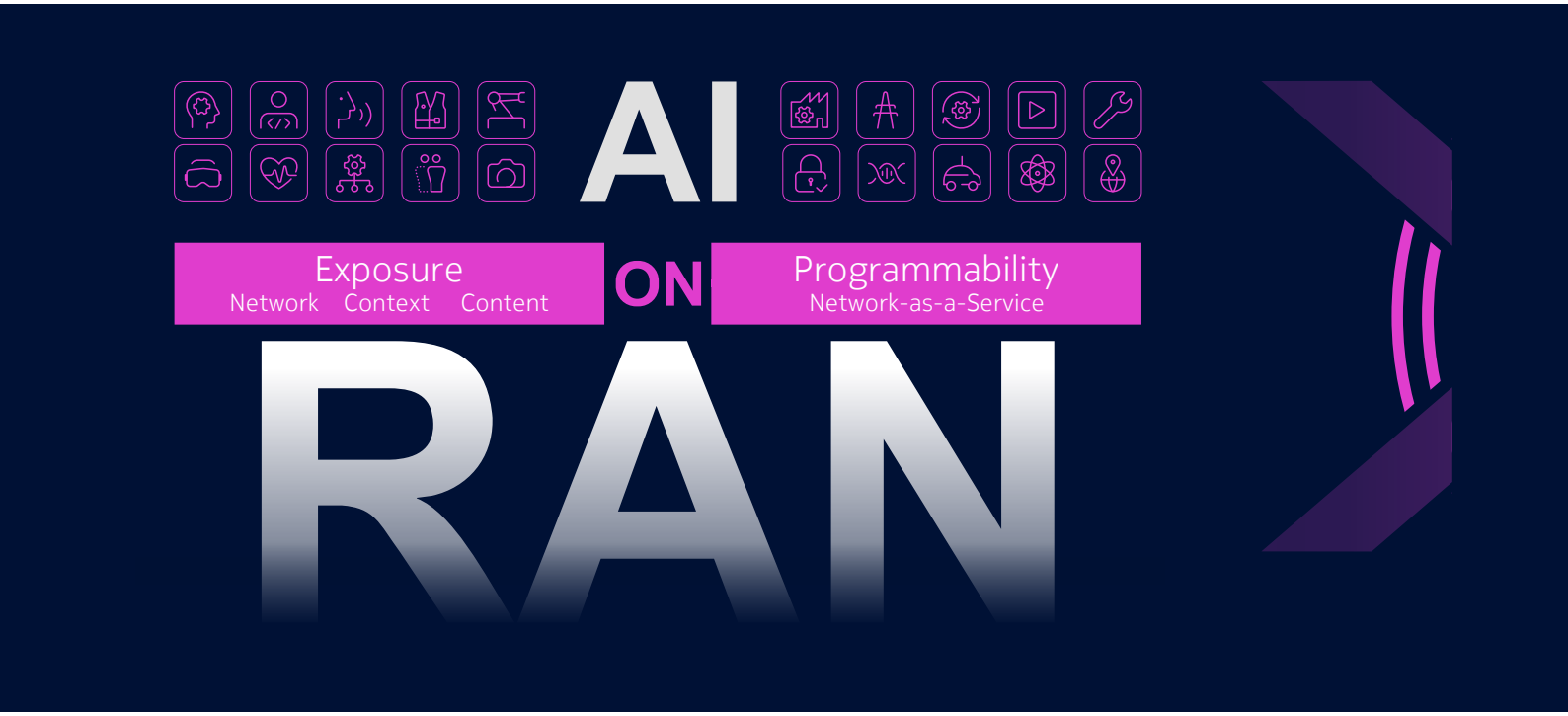
With AI-on-RAN, the radio access networks transition one step further from a connectivity solution to a platform that serves consumer and enterprise applications and devices. These applications and devices are increasingly powered by AI, which increases their capability to make use of data. RAN can provide unique data from the network, like network load levels, user location, user density and additional context information. This data will become more relevant, as AI enables applications to turn it into action to perform better or provide new value-added services to the users. Likewise, network programmability and Network-as-a-Service (NaaS) models will become more relevant once AI-powered applications can define their demand more dynamically.

From a functional standpoint, AI-on-RAN means that RAN functionalities can enhance AI applications, irrespective of where the data center that is providing the computational resources is located. From a performance perspective, co-locating RAN functions, relevant core network functions and application servers will accelerate the performance of AI applications that rely on user plane data. The core network functions include User Plane Function (UPF) in 5G and Evolved Packet Core (EPC) in 4G. This approach improves application responsiveness by reducing end-to-end latency, and it can also help limit data processing and exposure to where the data is needed and owned.

6G will introduce integrated sensing capabilities, which can provide valuable contextual data for autonomous vehicles and robotic systems. Sensing is a key research topic within the AI-on-RAN area.

AI-on-RAN could enable advanced use cases in the area of semantic communications, another important 6G research topic. Semantic compression could become an enabler for device and application use cases that involve large data volumes, which would otherwise limit their large-scale adoption.

To scale up the ecosystem of AI applications and robotic systems that can make use of AI-on-RAN, harmonizing RAN capabilities across different networks and suppliers is essential. This alignment would ensure broader compatibility and accelerate AI-on-RAN capabilities.



AI-and-RAN unlocks multi-purpose computing synergies

In the telecom industry, Cloud RAN has often been discussed in the context of multi-purpose edge computing. The challenge has been to identify other computing workloads that need CPU cores at the far-edge RAN sites.

The GPU-as-a-Service (GPUaaS), or more generally, Platform-as-a-Service (PaaS) business model has developed with the rise of generative AI. This evolution has led to a new breed of Cloud: the neocloud. The neocloud and the AI inference factory business models illustrate how fast the computing demand and related solutions can scale up and progress.

As generative AI evolves beyond delay-tolerant tasks such as text prompts or image and video generation, it will develop toward immersive, conversational and multi-modal generative AI. These advanced capabilities will require much lower latency, shrinking the overall delay budget of network transmission and AI inference to just milliseconds. This is also the case for many robotic AI use cases. To meet these demands, accelerated AI computing will need to happen at the network edge, a concept referred to as Edge AI.

At some point, the providers of AI factories will shift some of their inference capacity to the network edge. This development is linked to the consumer and enterprise demand for real-time inference, immersive experiences and data-sensitive, data-rich use cases, which will grow in volume and addressable business value.

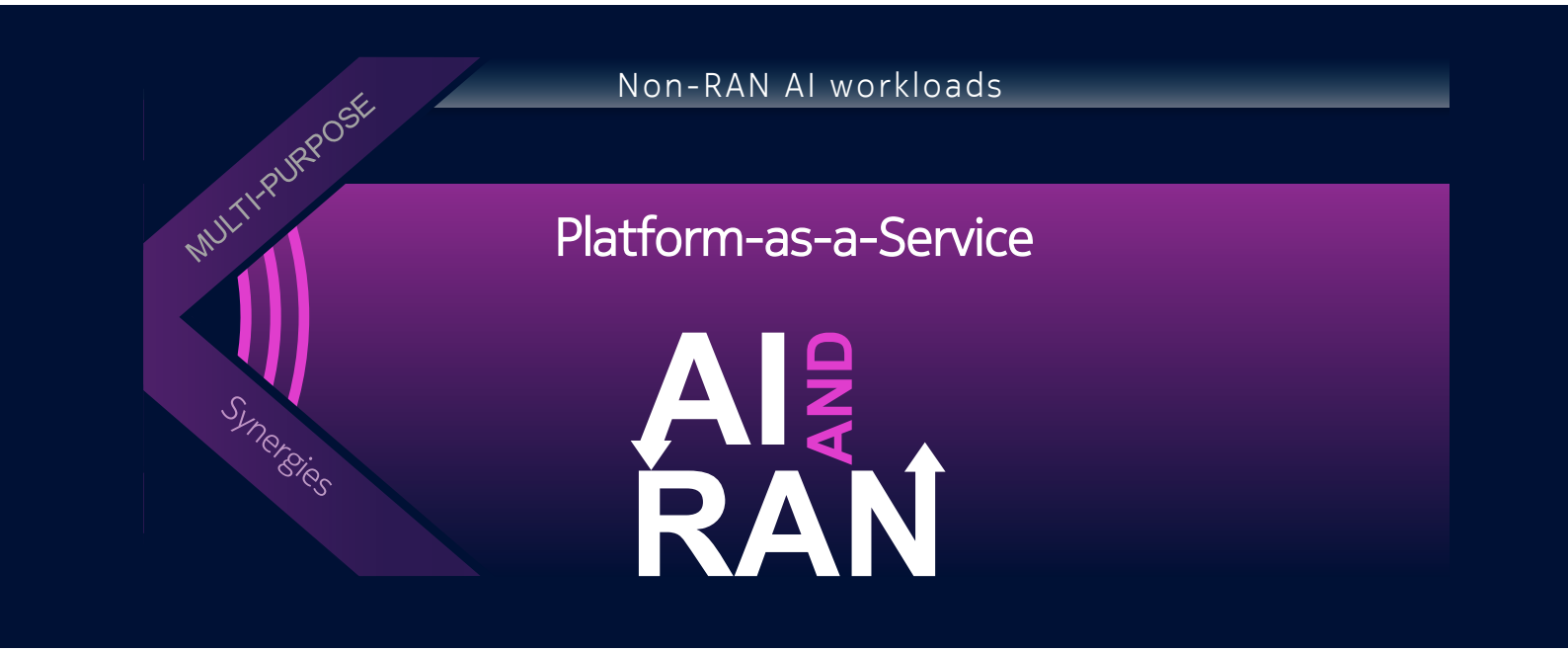
Similarly, the evolution of radio access networks across cellular generations reflects the growing demand for RAN computing power. This includes support for larger carrier bandwidths, more cells and transceivers per cell, increasingly sophisticated algorithms and models to drive efficiency and performance. With AI-native 6G and macro cells using 200 MHz carrier bandwidth, RAN computing requirements will reach the next level.

Sharing computing resources between RAN and other AI workloads can significantly improve the utilization rates of CPU/GPU systems. High utilization rates mean lower computing costs per task and help avoid the inefficiencies of idle computing capacity.

While today's human-triggered generative AI workloads often coincide with the network peak hours, the future can look quite different. Thinking forward to massive multi-agent AI collaboration and robotic AI, the computing-resource sharing model of AI-and-RAN will benefit from evolving into a continuous 24/7 workload environment. This will allow less real-time critical, maybe agentic AI workloads, to be scheduled into low-traffic hours of the RAN.

At that point in the future, computing synergies between RAN workloads and other drivers of edge neocloud environments could lead to a new techno-economic situation. From the RAN perspective, we could call it the neoRAN. This refers to a RAN that provides geographically extensive connectivity and a distributed, AI-accelerated computing grid.

In a pioneering initiative, Nokia and SoftBank focused on enabling AI workloads to efficiently share computing resources with RAN workloads [14]. We also showcased a joint AI-and-RAN proof-of-concept at Mobile World Congress 2025.



Conclusions

The impact of the fusion of artificial intelligence and radio access networks and the subsequent business potential for mobile operators will be massive:

- The exchange of tokens through mobile broadband networks will become the driver for the **second big value transformation** in mobile networks.
- AI-for-RAN will drive further efficiencies and performance enhancements and enable **intent-based autonomous capabilities**.
- With AI-on-RAN, the radio access networks transition to a platform that serves consumer and enterprise applications and devices by **providing valuable data from the network**.
- The rise of generative **AI has had an impact on business models** such as GPUaaS, PaaS, AI inference factory, and the latest new model: the neocloud. Further developments in generative AI, agentic AI and robotic AI, and the increasing infrastructure synergies of AI-and-RAN could lead to a new paradigm: the neoRAN.

Nokia is committed to the AI-RAN vision and has taken a lead in driving industry collaboration:

- We play a **leading role in shaping the development of AI-RAN** through active participation in the AI-RAN Alliance, of which we are a founding member.
- With its extensive AI capabilities, our **MantaRay AutoPilot delivers TM Forum's level 4 autonomous operations** in commercial networks today.
- We are leveraging our **extensive RAN domain knowledge and research capabilities** to make AI an integral part of RAN and build a concrete technology evolution path to AI-native 6G.
- Our **collaborative approach and successful proof-of-concepts** with leading operators and technology partners such as T-Mobile US, SoftBank in Japan and NVIDIA demonstrate our leadership in building the AI-RAN ecosystem.

We conclude by emphasizing our **commitment to responsible AI** [15], which ensures that AI-RAN technologies are developed and deployed with trust, transparency and societal benefit in mind.

References

1. Whitepaper: AI for Radio Access Networks, December 2024.
   https://www.nokia.com/asset/214372
2. Alex Choi, Head of AI-RAN Alliance, at Mobile World Congress, March 2025, Nokia video interview "Measuring AI-RAN performance with token throughput".
   https://youtu.be/9xRDmXCZYi8
3. AI-RAN Alliance website, Founding Members.
   https://ai-ran.org/#members
4. T-Mobile US Press Release: "T-Mobile Announces Technology Partnership with NVIDIA, Ericsson and Nokia to Advance the Future of Mobile Networking with AI at the Center", Sept 2024.
   https://www.t-mobile.com/news/business/t-mobile-launches-ai-ran-innovation-center-with-nvidia
5. SoftBank Press Release: "SoftBank Corp. and Nokia partner to research AI-RAN and 6G Network Technologies". Sept 2024.
   https://www.softbank.jp/en/corp/news/press/sbkk/2024/20240911_01/
6. Nokia Press Release: "Nokia to Revolutionize Mobile Networks with Cloud RAN and AI Powered by NVIDIA", Feb 2024.
   https://www.nokia.com/newsroom/nokia-to-revolutionize-mobile-networks-with-cloud-ran-and-ai-powered-by-nvidia/
7. Nokia Press Release: "Nokia and industry partners accelerate AI-RAN development #MWC25", March 2025.
   https://www.nokia.com/newsroom/nokia-and-industry-partners-accelerate-ai-ran-development-mwc25_20250422153504787/
8. Event: Tech Winter Horizon 2024.
   https://www.nokia.com/mobile-networks/tech-winter-horizon/
9. Case Study: "stc improves RAN energy efficiency with Artificial Intelligence powered energy savings management", November 2023.
   https://www.nokia.com/asset/213678
10. Comms MEA article: "AI-Powered Network Automation Slashes Outages in Makkah by 60% During Hajj Season", July 2025.
    https://www.itp.net/commsmea/ai-powered-network-automation-slashes-outages-in-makkah-by-60-during-hajj-season
11. Case Study: "stc starts autonomous RAN operations journey with AI-powered MantaRay AutoPilot", November 2024.
    https://www.nokia.com/asset/214374
12. Event: Nokia Midsummer Launch 2025.
    https://www.nokia.com/mobile-networks/midsummer-launch/
13. Nokia Bell Labs video: "6G AI native air interface proof of concept".
    https://youtu.be/VA95nQOVAtY
14. SoftBank Press Release: "SoftBank Corp. and Nokia Achieve AI and vRAN Coexistence with Automated Optimal Resource Allocation on a Single Server", March 2025.
    https://www.softbank.jp/en/corp/news/press/sbkk/2025/20250303_02/
15. Webpage: Responsible AI.
    https://www.nokia.com/bell-labs/research/air-lab/responsible-ai/

Glossary

AI	Artificial Intelligence
CPU	Central Processing Unit
EPC	Evolved Packet Core
GPU	Graphics Processing Unit
GPUaaS	GPU-as-a-Service
LLM	Large Language Model
NaaS	Network-as-a-Service
PaaS	Platform-as-a-Service
QoE	Quality of Experience
RAN	Radio Access Network
ROI	Return on Investment
SLM	Small Language Model
SON	Self-Organizing Networks
TCO	Total Cost of Ownership
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
XR	Extended Reality

Nokia OYJ
Karakaari 7
02610 Espoo
Finland

Tel. +358 (0) 10 44 88 000

CID: 215037

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, which is celebrating 100 years of innovation.

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.