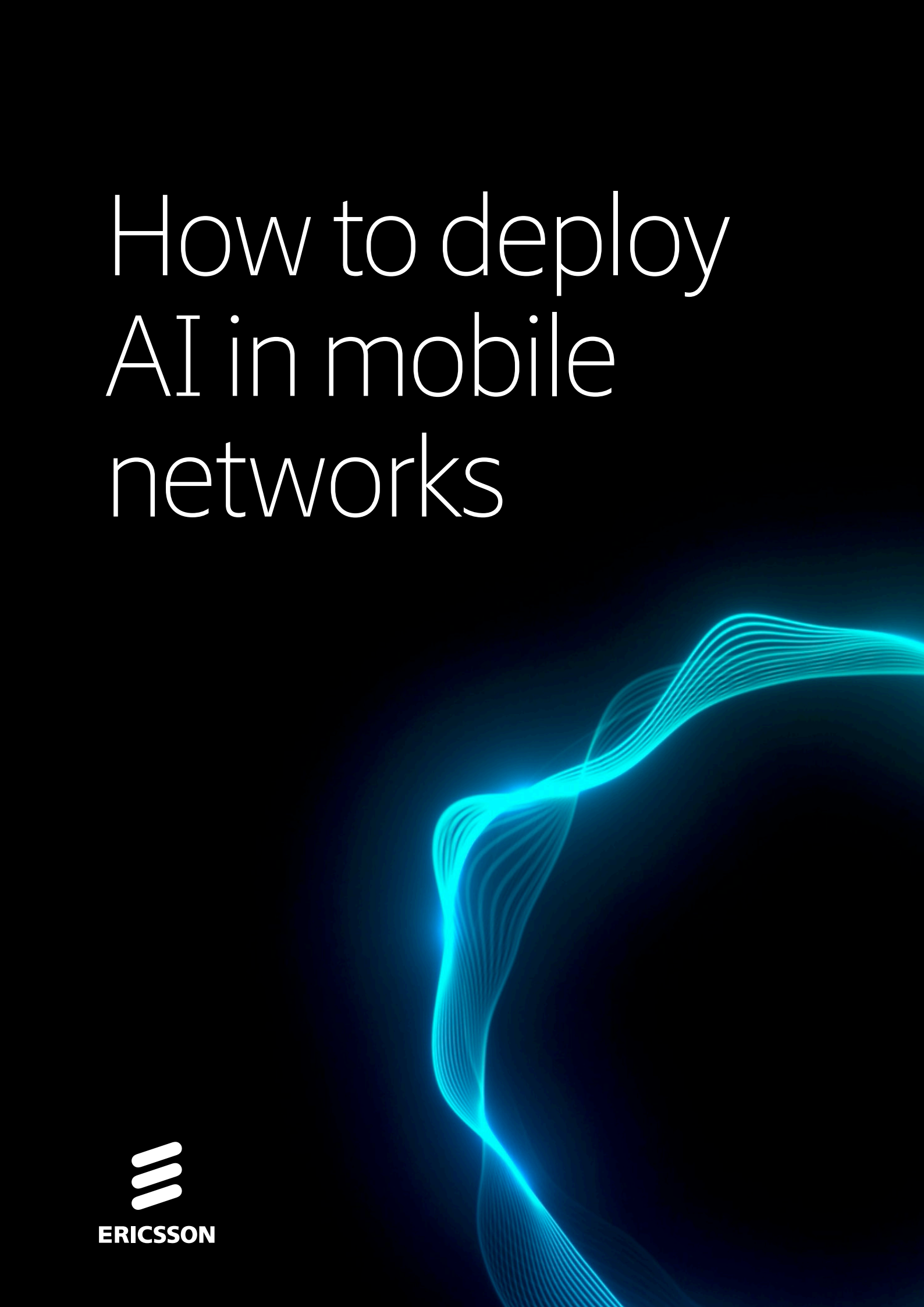
How to deploy AI in mobile networks



ERICSSON

Content

1 Overview

Artificial Intelligence (AI) is transforming how networks are built, operated, and monetized. For Communications Service Providers (CSPs), AI is no longer optional; it is a strategic imperative to manage growing complexity, unlock operational efficiency and boost performance to scale differentiated services.

This report outlines Ericsson's approach to deploying AI in mobile networks, grounded in real-world experience and technical leadership. It distinguishes between "AI for networks"—enhancing network performance and automation using AI —and "networks for AI"—delivering programmable high-performing infrastructure needed to support AI-driven applications.

Ericsson's AI RAN strategy spans both centralized (rApps) and distributed (radio site) deployments, enabling non-real-time and real-time automation and new AI use cases. With a deep integration of AI into the RAN stack and a common software strategy for two hardware architecture tracks (purpose-built and Cloud RAN), Ericsson ensures that AI is executed where it delivers the most value—whether at the edge for ultra-low latency or centrally for network-wide optimization. These AI-powered solutions improve user experience, coverage, mobility, spectrum efficiency, and reduce energy consumption.

For CSPs shaping the future of telecom, this report provides a strategic lens on how to scale AI efficiently, maximize return on investment, and build future-proof networks for 6G and beyond.

2 AI in the telecom landscape: AI for networks and networks for AI

AI transforms networks in two fundamental ways: first, it enhances network efficiency and second, it enables entirely new AI-powered applications that demand consistent, high-performance connectivity, such as AI-powered video feeds, to be uploaded from smart glasses.

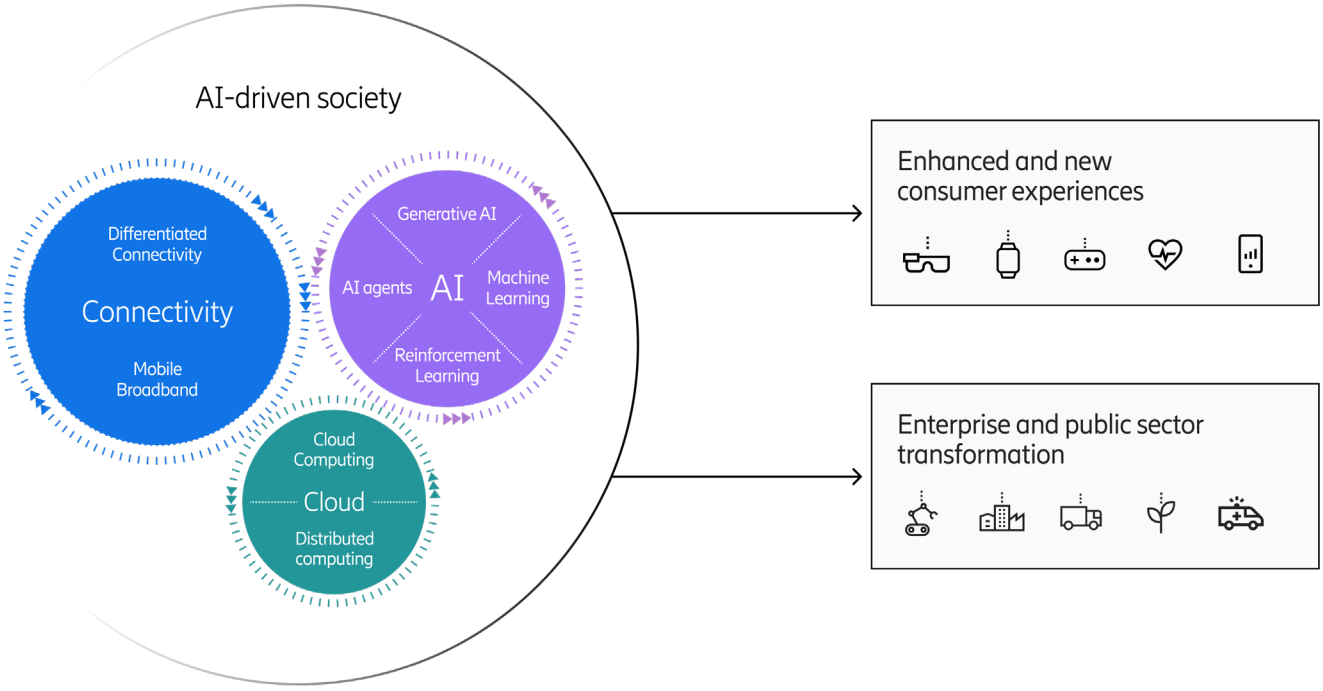


Figure 1: AI connectivity shaping society

Ericsson distinguishes between these two transformative roles through clear terminology: "AI for networks" refers to leveraging AI technology to enhance network operations and performance, while "networks for AI" describes how high-performing programmable networks enable new AI-based applications through differentiated connectivity. The "AI for networks" approach represents the most promising path forward for managing today's exponentially growing complexity of mobile networks—from surging traffic volumes and device proliferation to diverse use cases requiring specialized connectivity to drive business transformation.

3 Strategy for business growth with AI

AI in networks improves performance and unlocks new growth opportunities. Ericsson envisions AI as a key technology enabler for building high-performing programmable networks that are service-aware, AI-powered and intent-driven.

Intent-driven means that CSPs can specify their desired network outcomes without detailing how to achieve them or mentioning the specific configurations required for implementation. This could include optimizing or prioritizing certain traffic types, users, or balancing performance with energy consumption, while a network's self-adaptive approach reduces operations complexity. AI technologies allow networks to understand these intents or CSP's business objectives, process large data sets, make real-time decisions, handle conflicts resolution, and optimize the network accordingly. By building high-performing programmable networks, Ericsson is transforming mobile networks with new capabilities for business growth through differentiated connectivity and more autonomous operations.

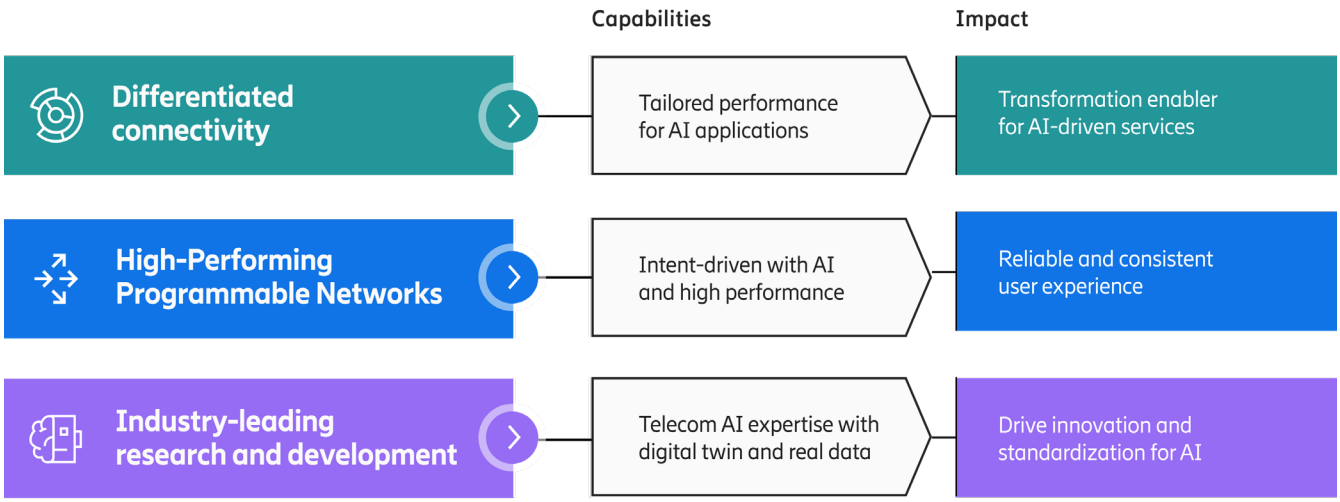


Figure 2: Ericsson's strategy for business growth with AI

3.1 Wide range of AI-powered solutions

Ericsson is a pioneer of AI in telecom. In RAN, the AI journey began with 4G and Ericsson's AI-native approach, laying the foundation well before the advent of 6G.

AI technologies used in Ericsson solutions for automation[1] include generative AI, digital twin, neural networks, reinforcement learning and AI agents. The goal is to deploy AI in networks efficiently and scalable with an architecture that maximizes the return of investment for each use case where AI is applied. Some examples are:

Digital twin is used for simulation and training of the AI models and to deploy new functionality preventing any KPI degradation[2].

Generative AI is used to support networks operations, software development and developers' enablement[3].

Reinforcement learning optimizes radio resource allocation to improve spectrum efficiency and user throughput with real-time network data[4].

Agentic AI will enable networks to make autonomous decisions based on the intents[5].

1. Intelligent RAN Automation managing 5G complexity - Ericsson
2. Network Support Services powered by AI and ML - Ericsson
3. EIAP Ecosystem for automation applications - join now - Ericsson
4. AI Native Link Adaptation
5. Unleash the power of AI intent-based operations - Ericsson

4 The path to AI-native RAN

4.1 Evolution journey ahead of 6G

The AI RAN journey that began in 4G is now rapidly evolving toward AI-native architectures in advance of 6G deployment. This transformation represents a fundamental shift from AI as an add-on enhancement to AI as an integral design foundation.

Ericsson pioneered this evolution in 4G with breakthrough AI features. Notable examples include mobility optimization to improve handover speed and reduced dropped calls, AI MIMO sleep mode to maximize energy efficiency, and sleeping cell detection for improved reliability[6].

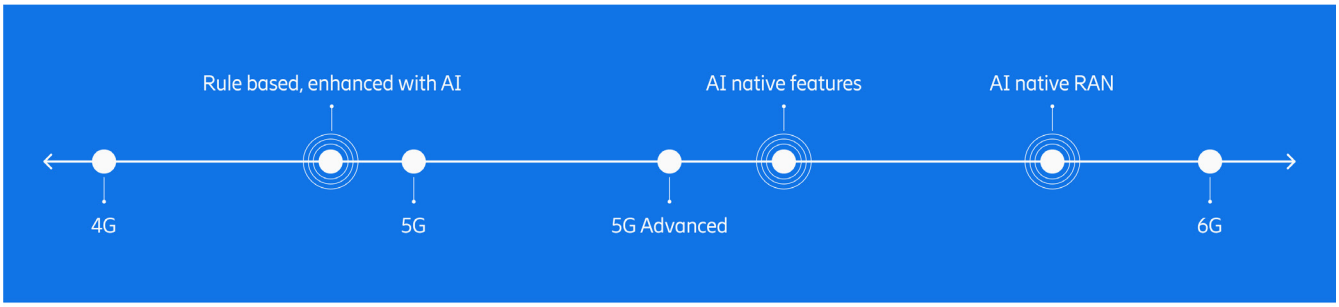


Figure 3: Ericsson's AI RAN native journey

4.2 The three stages of AI integration with RAN

AI started augmenting rule-based features with data and ML algorithms, leading to high-performing AI-powered features. The transformative AI journey continues with AI-native features where AI is an inherent part of design and development[7].

Adopting AI in RAN is now paving the way for high-performing programmable networks and intent-driven RAN as a new operations paradigm[8]. At every step of this journey, the added value of AI increases as it acquires new capabilities and addresses specific challenges, including requirements on high-reliability, low latency, and distributed compute resources.

All these requirements drive the need for enhanced storage and processing capabilities in the hardware and the need for deep telco expertise to handle and minimize complexity, which becomes critical to reduce cost implications. Enabled by Ericsson RAN software evolution, AI-native features can replace traditional algorithms with AI models. Network software is developed with AI built-in from the start.

An example is AI-native link adaptation, a feature of Ericsson 5G Advanced portfolio, where the real-time process of selecting modulation and coding schemes for every transmission is handled by an AI-native model embedded in the RAN software.

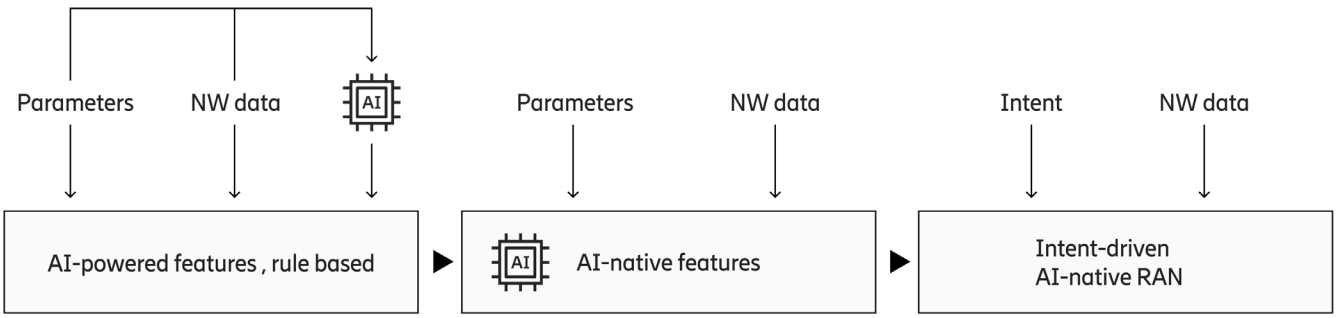


Figure 4: the three stages of AI integration in RAN

6. AI MIMO Sleep
7. AI journey in RAN report
8. Intent driven networks whitepaper

4.3 Data strategy: High-quality data drives performance

Data fuels AI models. More high-quality data means better trained algorithms. A good data strategy is fundamental for selecting the best AI technology for a specific purpose. Data availability and quality can limit the AI technologies that can be used.

With 5G, in a typical network deployment, there will be tens of terabytes of data to process from more than 1,000 distributed sources, opening new opportunities in AI-native networks. Serving RAN data to ML algorithms in an efficient, secure and sustainable way requires a profound understanding of the data operations and the RAN use cases.

Ericsson's AI models offer a unique advantage: they are globally trained on real network data, so when deployed, they already deliver strong performance—further optimized within the CSP's own network environment.

4.4 Selecting the best AI technology depending on the use case

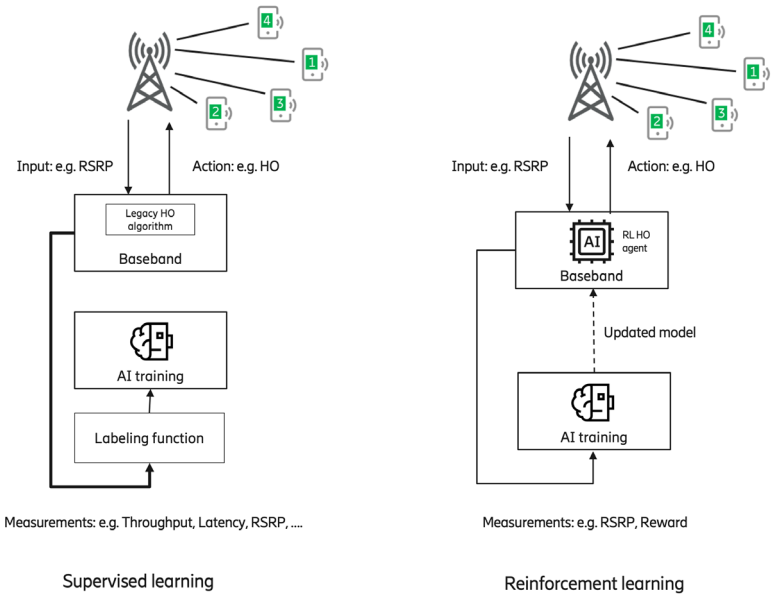


Figure 5: Examples of AI technologies used in RAN

Ericsson's approach emphasizes that no single technology defines the future of AI for networks. Decisions are guided by an end-to-end perspective, balancing innovation with the realities of deployment and operation. This ensures AI does not become just an enhancement but is a foundational element of RAN evolution, empowering CSPs to meet the demands of today while preparing for the opportunities of tomorrow. Ericsson evaluates, for each functionality, the best AI technology to reduce complexity and maximize the benefits.

Different learning paradigms can be applicable depending on the use case. Ericsson implements multiple AI techniques that include supervised learning, reinforcement learning, and neural networks with AI agents. Handling that complexity in the RAN control loops requires deep expertise.

For instance, reinforcement learning-based models require integrating exploration into the control loop. The model must actively try new actions to discover better strategies and optimal solutions in dynamic environments. The results achieved are up to 20% of improved downlink throughput and 10% of improved spectrum efficiency in the first deployments.[9]

9. AI-native link adaptation results in Bell Canada

5 AI-driven hardware evolution

5.1 AI RAN hardware processing approaches in the industry

Different infrastructure vendors have different approaches to execute AI models in RAN. These are the main approaches:

01

Additional hardware boards only for AI workloads. This approach increases the number of processing boards in the radio site rack, increasing the radio footprint.

02

Evolving existing hardware to support AI-native functionality without adding extra boards or units in the radio site rack. This is the Ericsson approach.

5.2 Ericsson's hardware architecture approach

Ericsson solutions support two hardware variants or architectures: purpose-built RAN Compute with Ericsson Silicon and Cloud RAN compute currently based on x86 from Ericsson ecosystem partners.[10]

Hardware-software co-design for AI models provides optimal performance and cost-efficiency. This is why Ericsson invests both in evolving its own hardware platforms as well as working with hardware ecosystem partners on what are the critical requirements for AI in a cloud-based solution. Ericsson Cloud RAN solutions can use hardware acceleration technologies to optimize the efficiency of AI algorithms.

These accelerators can offload the CPUs and improve performance and can be implemented either in the hardware or on the System on Chip (SoC).

10. Cloud RAN compute and acceleration technology

5.3 Ericsson RAN Compute evolution

Ericsson's hardware platform has continuously evolved to support a growing number of AI-native software features and connected devices, delivering enhanced performance with each new generation.

The latest Ericsson RAN Compute features an industry-standard X86 CPU with advanced vector extensions that enhance AI capabilities. This CPU efficiently handles L3 functionalities such as traffic steering, mobility, and energy efficiency. RAN Compute storage has increased 20 times to accommodate extensive AI model management compared to the previous generation.

Ericsson Silicon uses state-of-the-art technology from the silicon industry. For time-critical L1 and L2 processing, Ericsson Silicon utilizes a pool of hundreds of digital signal processors (DSPs) in the innovative Ericsson Many-Core Architecture (EMCA). These DSP cores are optimized with instructions specifically designed for AI inference.

Examples of AI use cases on Ericsson Silicon include AI-native link adaptation and AI-native scheduling, both requiring extremely fast execution (microseconds and milliseconds). To achieve the best processing performance for AI in the industry, Ericsson RAN Compute combines both third-party CPUs and in-house Application Specific Integrated Circuit (ASIC) and a massively parallel DSP pool for AI workloads.

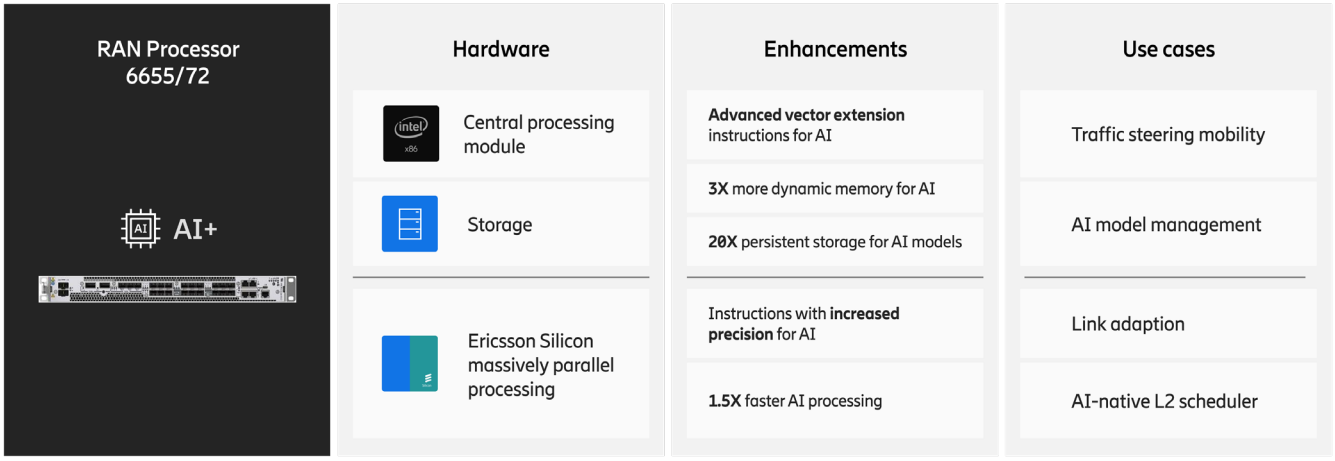


Figure 6: Latest RAN Compute enhancements for AI acceleration

Flexibility and AI acceleration

The same board can execute both AI-native and RAN workloads concurrently thanks to the massively parallel processing capability of Ericsson RAN Compute powered by Ericsson Silicon with EMCA architecture, that allows fast AI inference and execution without requiring additional hardware.

Cost efficiency and sustainability

By integrating all processing capabilities onto the same board, without requiring extra hardware, no additional rack space is needed. This reduces both maintenance and energy consumption costs.

Highest ROI and Time to Market (TTM)

AI capabilities are already available in Ericsson's RAN Compute. This provides a competitive advantage to execute new AI-native functionality that will improve network performance and spectrum efficiency with limited investment.

6 Deployment architecture: Maximizing value

AI is transforming the telecom landscape, offering Communications Service Providers (CSPs) powerful new ways to boost network performance, streamline operations, and reduce complexity. But beyond these internal gains, AI is also opening the door to entirely new revenue opportunities. For instance, personalized service offerings with exclusive event experiences and AI-powered network slicing for enterprise customers.

To unlock this potential, CSPs must ensure their infrastructure is ready. AI models require significant storage and processing power, making the hardware at radio sites a key part of the deployment strategy. However, with thousands of sites in play, this hardware must also be cost effective and energy efficient.

This chapter explores how to design deployment architectures that maximize the return on investment (ROI) from AI. It focuses on how to integrate AI-ready infrastructure, ensuring CSPs can scale AI capabilities efficiently and profitably.

The fundamentals of AI RAN deployment evaluation involve various timescales, processing complexities, and implementation costs.

6.1 Balancing performance and efficiency

Deploying AI in the Radio Access Network (RAN) involves navigating different timescales, processing demands, and cost implications across the RAN stack. Ericsson evaluates the return on investment (ROI) and deployment challenges for each layer—both in purpose-built or Cloud RAN environments.

Lower RAN layers require ultra-low latency and high data throughput, demanding more powerful and localized compute resources. In contrast, higher layers allow for more relaxed latency and can support more complex AI models. This variation drives distinct cost and infrastructure requirements.

Benefits/ requirements for interface	L1	L2 inner loop	L2 outer loop	L3 ue procedures
Benefit	Channel performance	Throughput Cell edge performance	RRM optimization Service awareness Differentiated connectivity	Coverage Prediction Energy-awareness Traffic load balancing
Latency	microseconds	10 x microseconds	miliseconds	10 x miliseconds
Data bandwidth	10 Tbps	250 Gbps	50 Gbps	1 Gbps

Figure 7: RAN layers requirements and benefits

A key factor is the latency of the AI control loop: from microseconds to milliseconds. To meet these demands, compute resources must be placed close to the antenna, making over-dimensioning costly.

This is where hardware-software co-design becomes essential—delivering high performance while optimizing cost and energy use.

To create highly efficient and compressed models, deep expertise in RAN models and training data is fundamental for selecting the appropriate model and to optimize it for the underlying hardware.

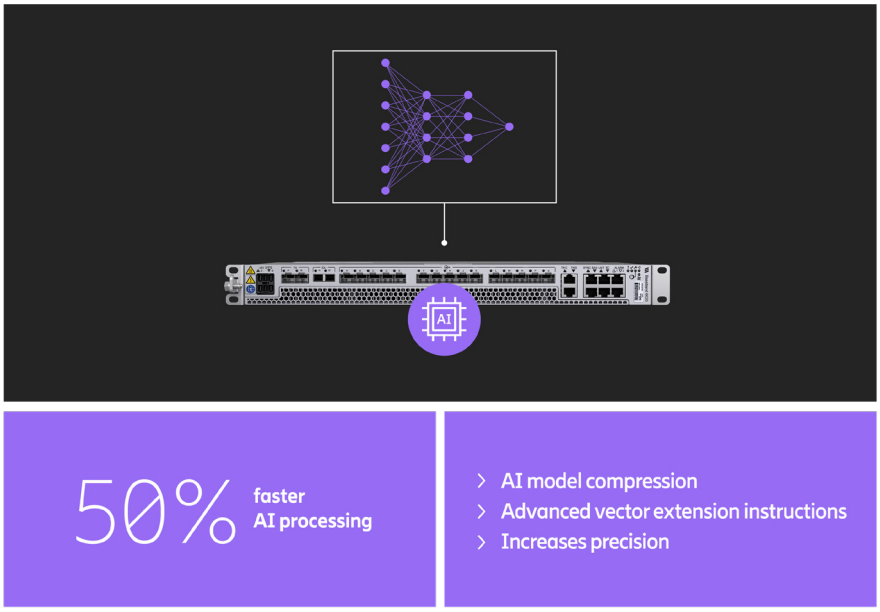


Figure 8: Accelerating AI in with high-efficient and compressed models

Ericsson continuously assesses the cost-benefit of each AI deployment option. Investments span both purpose-built and Cloud RAN solutions, evolving Ericsson's own hardware platforms and collaborating with ecosystem partners to define AI-ready infrastructure.

For example, Ericsson's evaluation of AI-native Link Adaptation showed that even with powerful GPUs, uncompressed models failed to meet latency targets. Compressed models running on DSPs proved more efficient due to lower data transfer overhead.

Currently, L1, L2 and L3 RAN functionality is implemented in the Ericsson RAN Compute, where highly efficient and compressed AI models are executed. It can also be deployed on third-party partners hardware as part of Ericsson's Cloud RAN solutions.

Ericsson is evaluating using Massive MIMO radios to execute AI models. This is possible because Ericsson Massive MIMO radios and Ericsson RAN Compute are built with Ericsson Many-Core Architecture (EMCA), which is AI-ready hardware with parallel processing capabilities. By executing AI in the RAN stack, Ericsson estimates very significant improvements in the areas of spectrum efficiency, uplink and downlink performance, and traffic control.

6.2 Combining the advantages of both centralized and distributed architectures

AI provides high value integrated in all the layers of the RAN stack (distributed) for radio resources management optimization in the radios and in the rApps (centralized) with a key role in optimizing radio network configuration, monitoring, and fault detection.

rApps, hosted in the Service Management and Orchestration (SMO) platform in a data center location, automate operations over seconds to weeks. Ericsson's implementation, the Ericsson Intelligent Automation Platform (EIAP)[11], is an open, multi-vendor SMO that supports rApps from Ericsson, CSPs, and other vendors[12].

AI and automation operate on different timescales in the RAN stack (real-time) and SMO (non-real-time), addressing varying complexity levels. In rApps, AI is used for prediction and analytics to configure, optimize and assure overall RAN performance and reliability.

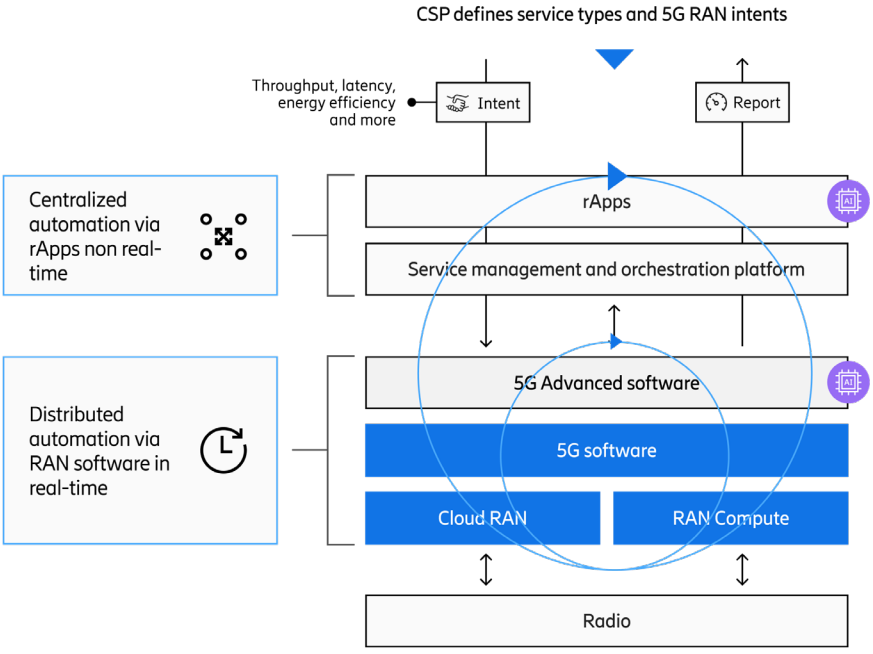


Figure 9: Centralized and distributed architecture for AI and intent-based automation

11. Intelligent Automation Platform (EIAP) - Ericsson
12. rApp Directory - Ericsson

Field results from selected implementations:

Ericsson NSA Traffic Optimizer rApp improves 5G NR frequencies utilization by NR traffic prediction based proactive NSA users distribution, achieving 18.6% downlink (DL) and 9.1% uplink (UL) average NR cell throughput gains (North America).

AI-native link adaptation (Ericsson 5G Advanced solution): Up to 20% downlink (DL) throughput and 10% spectral efficiency gains (Canada).

Ericsson's Service Continuity AI App suite: AI features work in tandem to measure, predict and optimize energy consumption across the network. Reduced energy consumption by 33% (Europe).[13]

Machine Intelligence Enabled Mobility radio feature: 60% faster handover decisions, 11% fewer inter-frequency handover failures (North America), 1.2% reduction in the overall drop rate.

6.3 The potential of edge data centers

Edge computing focuses on bringing computing resources closer to where data is generated. It is best for situations where low latency or real-time processing are required, or where large volumes of data are being unnecessarily transmitted to a central location. Regional and enterprise data centers enable advanced AI applications offering lower latency than centralized locations and with data soverignty.

In a network edge, computing and storage resources are distributed across communication service provider (CSP) premises, between national, regional and local access sites. These can be standalone or integrated with the mobile cloud (running both telecom and third-party workloads). Edge compute can be seen as an extension of the CSPs existing network capabilities.

Manufacturing, healthcare and gaming and entertainment are three of the top vertical industries with enormous potential when it comes to edge computing with a latency required between 50 miliseconds and one second[14]. We have classified into three categories the type of new use cases that would benefit for deployment in edge data centers:

01

Low-latency AI services: Edge proximity allows real-time processing for applications like video analytics, AR/VR, immersive gaming, and autonomous vehicles.

02

Context-aware services: Context awareness enables AI-based security, healthcare, and enterprise use cases including industrial automation.

03

IoT analytics: Local breakout (LBO) mechanisms allow IoT and sensor data to be processed near the source, enabling immediate insights and reducing cloud dependency.

6.4 AI executed where it makes sense

AI model training and inference require advanced storage and processing. To meet these needs, AI should be deployed where it's most effective—centrally as rApps or distributed at radio sites. For network-level tasks like optimization or healing that operate over seconds or longer, centralized deployment in data centers via rApps is ideal. For real-time radio resource optimization tasks such as channel estimation, scheduling and link adaptation, AI should be implemented in the radio sites. Ericsson advocates a balanced approach, with RAN software features complemented with non-RT RIC (rApps). This aims at enabling AI-driven innovation while managing complexity with cost efficiency, ensuring a sustainable networks evolution.

13. Ericsson's Service Continuity AI App suite
14. Edge computing use cases

7 Future strategy and emerging opportunities

Ericsson's AI strategy is designed to unlock value across both dimensions of Telecom transformation: AI for networks and networks for AI.

By embedding AI natively into the Radio Access Network (RAN), Ericsson enhances performance, automation, and energy efficiency. Simultaneously, by evolving the RAN into a high-performing, programmable platform, Ericsson enables new AI-driven applications and services.

This dual focus ensures that AI is not only a tool for operational excellence but also a catalyst for innovation and growth.

New traffic growth in mobile networks is set to be driven by high-performing 5G networks serving new devices, such as AR glasses, together with scalable, multimodal generative AI (GenAI) applications.

Significant network impact will stem from applications that are both data-intensive and widely adopted, including video-based AI assistants that use real-time video feeds for interaction, requiring constant uplink/downlink flow and semantic understanding which can unlikely be provided by a GenAI model on the device.

Data growth prediction will depend on the adoption rate of the new devices for AR/VR (Source: Ericsson Mobility Report June 2025).[15]

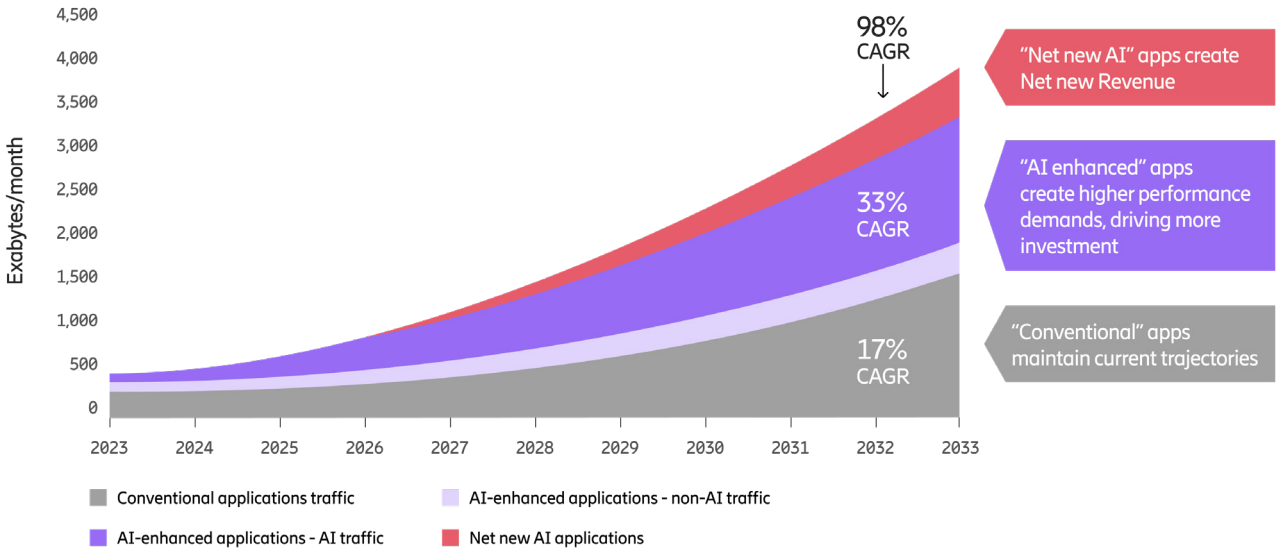


Figure 10: AI impact on traffic forecast (source: Omdia, May 2025)

In the figure we show AI traffic forecast. This prediction anticipates a significant growth on data volumes in the next decade, this growth would be driven by AI-enhanced applications. These applications that existed prior to AI, which upgrade to include AI features and enhancements. Over time, the projection shows that AI elements of these application will eclipse conventional traffic. There are two subcategories: AI-enhanced applications with AI traffic and AI-enhanced applications with non-AI traffic.

Examples of AI enhanced applications include: existing development tools that enable low-code environments; smart image recognition and editing added to existing photo editing software; applications on an e-commerce or media site that make recommendations based on users' past preferences; management systems that incorporate machine learning into their analytics toolkits. Net new AI applications will exist because of new capabilities unlocked by AI. A few applications are revolutionary; most replace basic manual tasks with automation.

Net new AI applications include: cameras installed in a warehouse to monitor and track inventory levels on shelves; virtual assistants that interact with people through natural language; video surveillance that can identify a range of dangers and hazards to trigger alerts and rapid responses.

Net new AI traffic is significant because it adds to network operators' total global traffic loads and may have special performance requirements that will require differentiated connectivity.

15. Ericsson mobility report

Ericsson's strategic areas of development for AI in radio networks

Future proofing and advanced AI models for 6G. While the immediate focus is on integrating AI into 5G networks, the strategy also prepares for future 6G standards. Ongoing research aims to develop sophisticated AI models tailored for 6G, enhancing network functionality such as channel estimation, modulation, and traffic management enabling new use cases[16]. This standardization work builds on the learnings from 5G, ensuring that 6G networks achieve unprecedented levels of performance and reliability.

AI with energy efficiency in mind. Energy efficiency will continue to be a key criterion for hardware and software development. Additional energy consumption from AI models execution should be minimal and be compensated with savings gained through optimized channel and interference management—ensuring the benefits outweigh the costs.

Exploring executing AI functionality in radio units. Executing AI in Massive MIMO radios enhances uplink performance, reduces latency, and supports new use cases. This approach improves cost-efficiency by offloading processing from other network elements, it requires robust infrastructure for data collection, processing, and assurance. This is a line of future development for Ericsson.

Industry collaboration and an open, diverse, strong ecosystem.
The successful implementation and global adoption of AI RAN architectures hinge on fostering industry-wide consensus. This involves agreeing on the most appropriate frameworks, standards, and collaborative efforts to build a healthy and diverse ecosystem.

Such alignment ensures economies of scale and reduces fragmentation, enabling faster innovation and deployment. Ericsson will continue to play a driving role in standardization (3GPP, ETSI, O-RAN, TMF) to define the technical frameworks that ensure interoperability across vendors and operators. Ericsson will also continue to be a key player in other industry collaborations such as AI-RAN Alliance, contributing with its vast experience from the field and its vision to make future networks energy efficient.

Differentiated connectivity.
It is enabled by programmable networks that are service aware, intent-driven and AI-powered. New AI-based use cases and applications that require guaranteed performance levels (throughput, latency) will be unlocked by differentiated connectivity.

For example: new AR/VR immersive experiences, industrial IoT and automation, robotics, real-time video analytics, autonomous vehicles, fraud detection in real-time, and more. In the coming years, AI traffic generated by these new applications is expected to grow significantly.

16. 6G Use cases: Beyond communication by 2030 - Ericsson

8 Recommendations and key takeaways

A forward-looking AI strategy must capture today's deployment opportunities while remaining flexible for future innovations. This chapter outlines key recommendations to help Communication Service Providers (CSPs) deploy AI effectively, maximizing performance, return on investment, and long-term scalability across both centralized and distributed network environments.

There are three primary locations in mobile networks where AI workloads can be deployed, each differing significantly in scale, as illustrated in the figure.

In a typical CSP network, there are between 5 and 10 central sites. These are suitable for network-level automation, such as deploying rApps within the Service Management and Orchestration (SMO), Ericsson Intelligent Automation Platform being Ericsson's implementation. At the other end, close to the end user, radio sites number in thousands. AI models deployed here aim to optimize radio resource performance, where the benefits must outweigh the added complexity and cost.

Edge data centers are located between centralized and distributed deployments. These offer a potential deployment option for future use cases requiring lower latency—though not as low as the microsecond level latency needed at the distributed RAN.

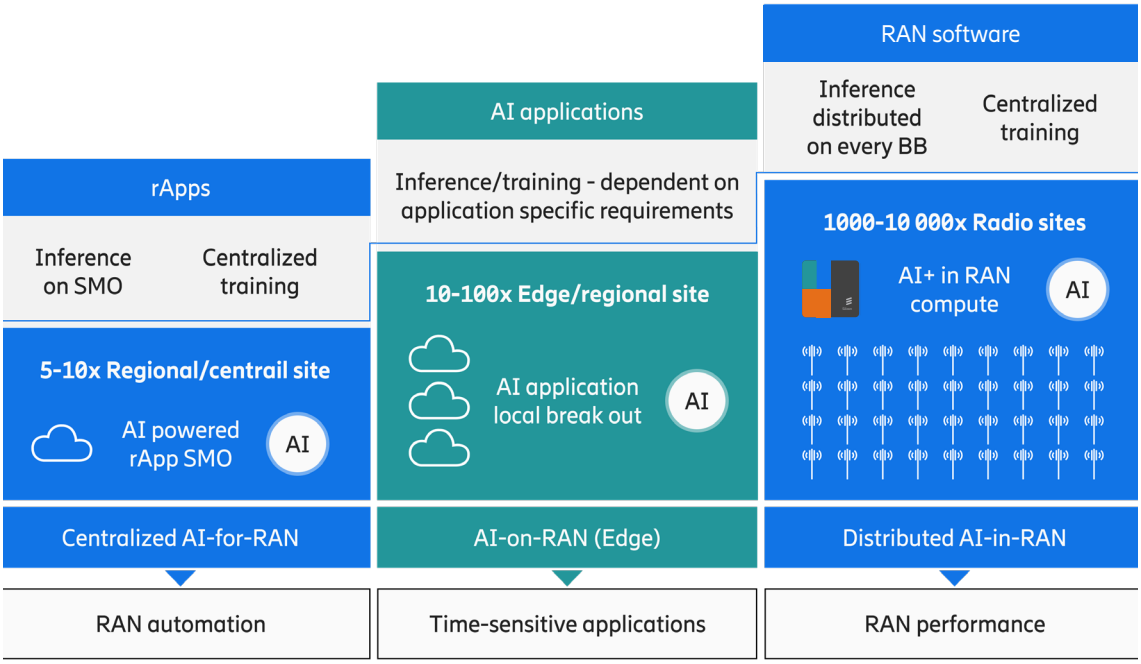


Figure 11: Flexible options for AI deployment adjusted to the use case in radio networks

Four guidelines to define your AI deployment strategy

01

Consider for your vendor evaluation criteria both hardware and software holistically. Compressed models trained with high-quality data will improve performance and accelerate the execution regardless of the selected hardware. Also, hardware and software co-design provides optimal performance, cost and energy efficiency benefits by optimizing AI models for the underlying HW. Ericsson invests both in evolving purpose-built hardware platforms as well as working with Ericsson hardware ecosystem partners as part of Ericsson's software portability strategy[10].

02

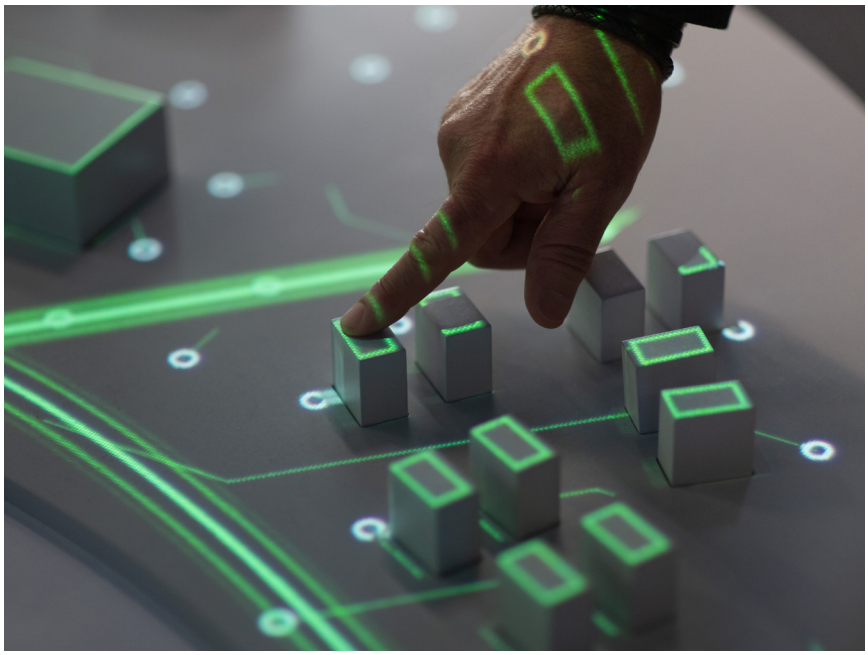
Look for hardware solutions that can give more with less. Minimizing extra hardware and energy consumption can be achieved by leveraging signaling processing capabilities in the Systems on Chip (SoC) as an attractive option for AI inference models. The massively parallel processing capability of the Signaling Processing modules in the Ericsson Silicon with Ericsson Many-Core Architecture (EMCA) allows fast AI inference at microsecond level in Ericsson's RAN Compute. This combined with Ericsson's high-compressed AI models, ensures fast AI execution without requiring additional hardware.

03

Future proof your network investments with strong ecosystem partners. The success of AI in telecom depend on industry-wide alignment around frameworks, standards, and collaboration. This coordination reduces fragmentation, enables faster innovation and scalable deployment, and improves security. Ericsson plays a leading role in standardization (3GPP, O-RAN, TMF) and alliances (AI-RAN Alliance), driving energy-efficient, interoperable AI-RAN networks. Ericsson also collaborates with partners to deploy AI-native in Cloud RAN with multiple options for hardware platforms.

04

Execute AI where it makes more sense depending on the use case to maximize ROI. AI models necessitate enhanced storage and processing capabilities, so to maximize ROI, a flexible, use case-based deployment strategy is recommended: centralized, distributed, and regional edge. Ericsson combines the advantages of both centralized and distributed architectures for AI. When the AI model is intended to act at network level for optimization, deployment or healing purpose in a time range above seconds, a centralized location is the optimal solution (SMO and rApps). When the functionality needs execution at microsecond or milliseconds level, the radio sites are the only placement (Example: Ericsson 5G Advanced solutions).



10. Cloud RAN compute and acceleration technology

Key takeaways

Ericsson's vision is to build high-performing, programmable networks that are energy efficient, service aware and intent driven where AI is a key enabler.

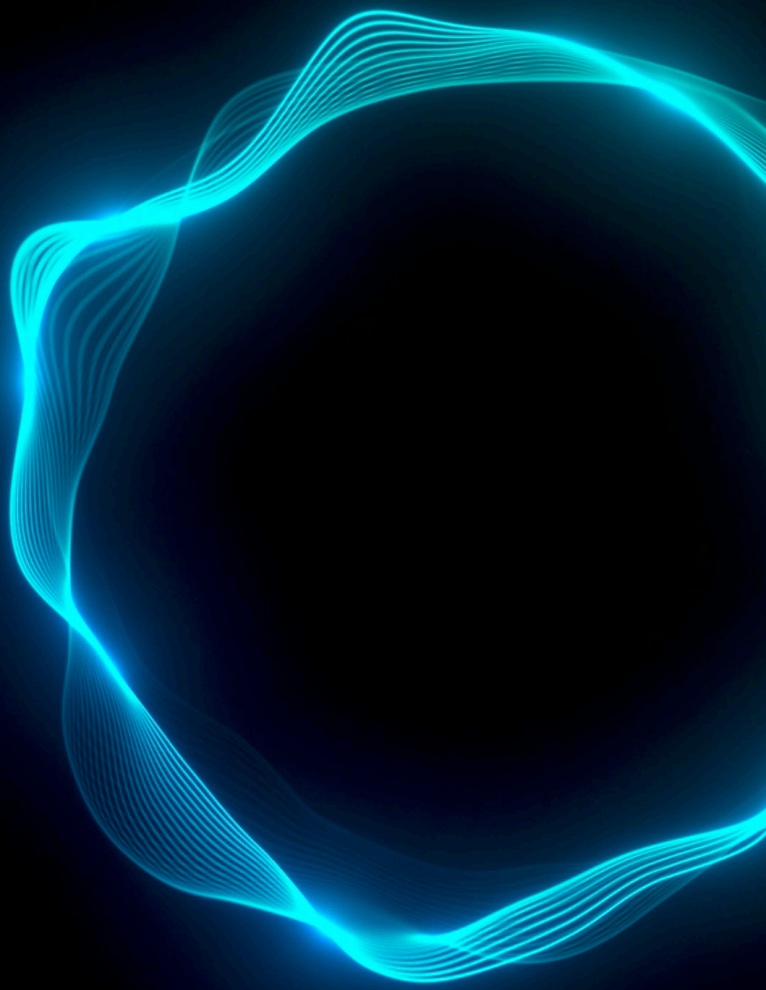
AI, combined with intent-driven networking, is a powerful tool for CSPs to achieve their business objectives and provide new services at scale.

Ericsson's AI-ready platforms powered by Ericsson Silicon with Ericsson Many-Core Architecture (EMCA) are enabling real-time, localized AI execution with massively parallel processing capabilities and fast execution for AI workloads.

Deploying AI in radio access networks (RAN) to maximize ROI requires deep telco expertise balancing latency, compute power, and cost efficiency across different layers with hardware and software co-design.

Ericsson's AI flexible architecture strategy combines both AI in the radio sites (radio software) and in the centralized data centers (rApps) using purpose-built and Cloud RAN architectures and supporting Open RAN interfaces for CSP's freedom of choice.

Accelerating AI in mobile networks driving standardization and industry collaboration towards 6G is an evolutionary step for more advanced, autonomous and energy-efficient networks.



Authors

Melike Erol-Kantarci
Strategic Product Manager for AI in RAN

Anders Söderlund
Strategic Product Manager of RAN Compute

Christoffer Stuart
Strategic Product Manager of RAN deployment and performance

Noelia Lopez
Solution Marketing Manager for Intelligent Networks

Klas Johansson
Head of Networks Automation and Operations

Oscar Toorell
Head of Cloud RAN Technology

Petter Sundberg
Strategic Technology Manager in Networks

Jonas Rosenberg
Product development leader in Networks Technology and Strategy

Joakim Bergström KO
Senior Expert in RAN Standardization

Mathias Sintorn
Expert in RAN traffic handling and service performance

About Ericsson

Ericsson's high-performing networks provide connectivity for billions of people every day. For nearly 150 years, we've been pioneers in creating technology for communication. We offer mobile communication and connectivity solutions for service providers and enterprises. Together with our customers and partners, we make the digital world of tomorrow a reality. www.ericsson.com