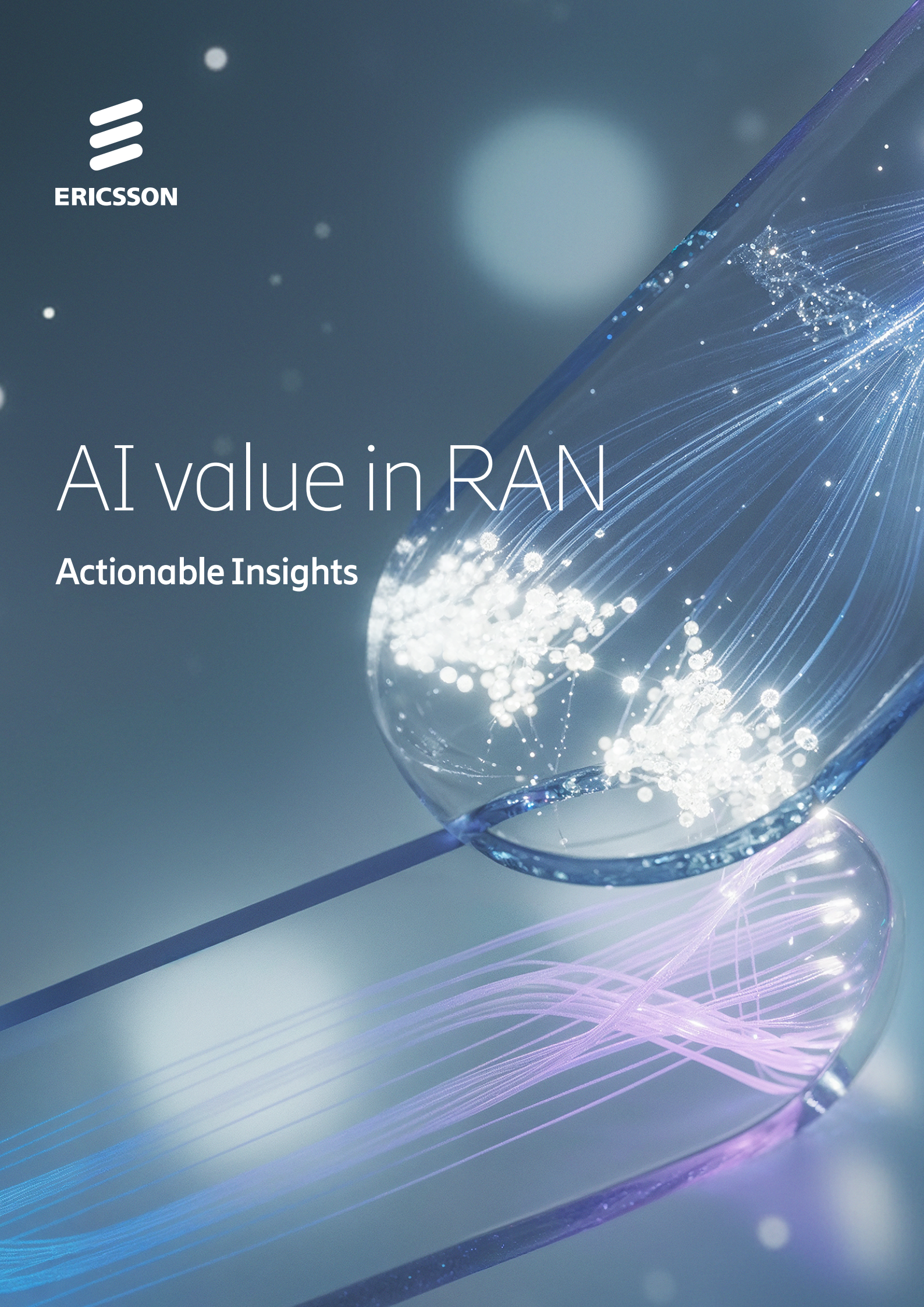
ERICSSON

AI value in RAN

Actionable Insights



Content

Executive summary

Across multiple industries, Artificial Intelligence (AI) has significantly impacted existing solutions by either enhancing their capabilities or disrupting them entirely. This has opened new avenues for innovation and growth.

Telecommunications service providers' spending on AI is expected to grow significantly over the next years, addressing areas such as network performance, operations, and security. AI in the global telecommunication industry is forecasted to grow from a market revenue of USD 2.2 billion in 2023 to USD 19.5 billion in 2030. [1]

AI plays a key role in various aspects of the RAN, including radio functions, predictions, and control loops, which differ from generic plug-and-play AI tools. Specialized knowledge of the radio network data and radio software is crucial for effective AI implementation in this domain. Data-driven AI aims to leverage data from radio access network (RAN) to drive network optimization and management using machine learning (ML) algorithms. ML algorithms, when trained with RAN data, evolve into ML models capable of offering predictions and insights for RAN or taking actions within RAN control loops.

The challenges of implementing AI solutions in the fast control loops of RAN have led to the development of more AI solutions in the high-level, slow-control loops. These challenges include the need for real-time processing and the complexity of verifying solutions quickly. As a result, initial efforts focused on slow-loop solutions where manual verification was more feasible. However, there is an opportunity for AI to deliver greater gains in both the slow and fast control loops, improving overall performance at a reduced network complexity and cost of operations.

Ericsson has successfully developed and deployed solutions in fast control loops for multiple use cases such as the AI MIMO Sleep feature [2]. This feature delivered 14% savings in energy consumption per site, outperforming manual management. Other example is AI powered link adaptation feature which uses information from adjacent cells for medium to high loaded systems with significant improvements in the spectrum efficiency and downlink (DL) throughput (11,6 % improved throughput for heavy users and 50% increased cell edge DL throughput, measured in the field).

We have seen in our collaboration with CSPs that the adoption of AI demands good understanding of data operations and strong development teams. Our extensive experience allows us to share valuable insights for accelerating AI adoption across the entire RAN control loop spectrum and leveraging existing networks architecture to overcome adoption challenges.

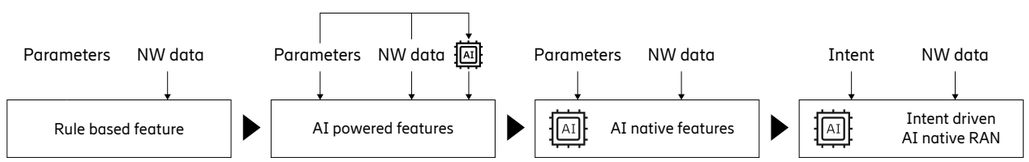
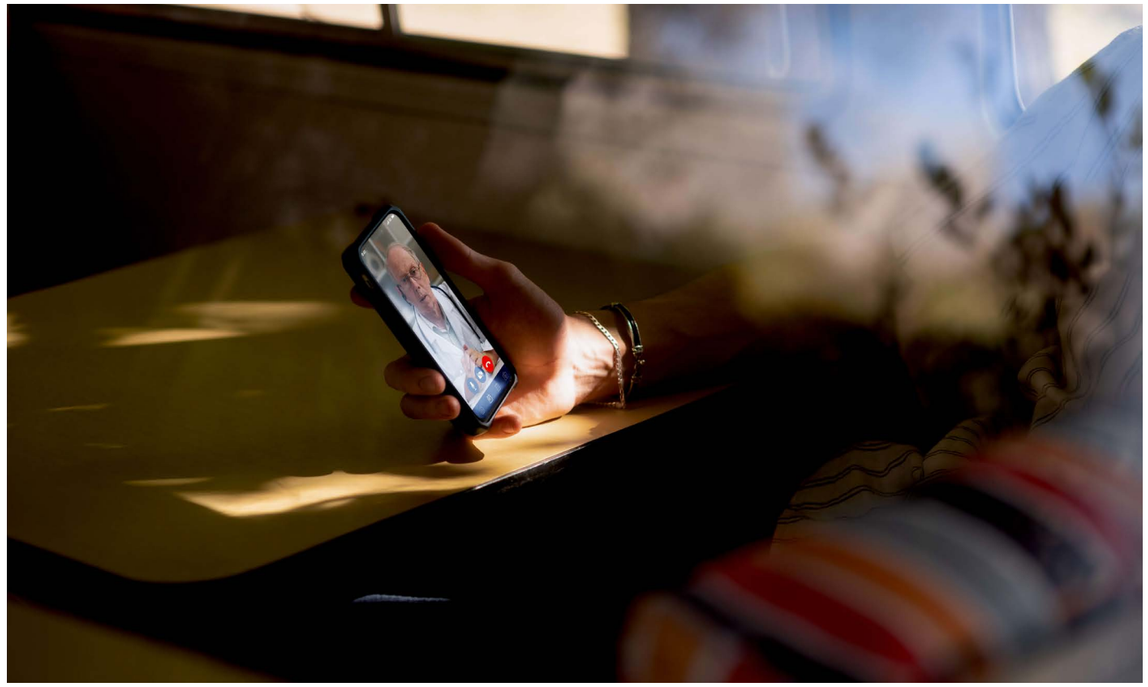


Figure 1: The four phases of AI in RAN journey

1. THE ROLE OF AI IN BUILDING SUSTAINABLE AND ENERGY-EFFICIENT TELCO NETWORKS TECHNOLOGY ANALYSIS REPORT | ABI Research, April 29, 2024
2. Automating MIMO - MIMO Machine Learning and AI - Ericsson

Ericsson | AI value in RAN



At Ericsson, our AI journey started with augmenting our 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. Adopting AI in RAN is now paving the way for intent-driven and AI native RAN. Meanwhile, generative AI has been emerging as a disruptive technology in multiple industries. Generative AI is envisioned as a strong component in supporting our customers' AI in RAN journey starting with intelligent product assistants.

At Ericsson, we envision that mobile networks can significantly benefit from AI technology. In our strategy, AI is an inherent part of the radio networks. This paper aims to serve as a reference point for CSPs who are looking into their AI and automation journey. Networks operations will shift towards intent-driven networks, where AI models continuously adapt to meet specific user experience and energy efficiency goals. This paradigm shift aims to cost effectively deliver superior customer experience and support performance-based business models.

In the CSPs' AI journey, it will be essential to focus on leveraging AI whenever gains are more significant and legacy solutions can be seamlessly replaced to maximize CSP's return of investment (ROI). At every step of this journey, the value-add of AI increases as it acquires new capabilities and addresses specific challenges, including requirements on high-reliability and low latency, as well as the distributed compute.

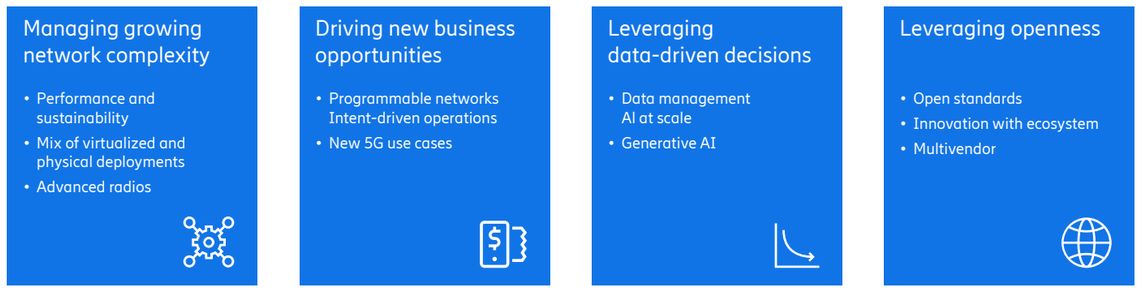


Figure 2: AI in RAN opportunities

AI journey started with 4G and continues with 5G

The telecommunications industry started integrating AI during long-term evolution (LTE) networks for optimization. Ericsson was at the forefront developing AI features such as mobility optimization and link adaptation to improve spectrum efficiency, multiple input multiple output (MIMO) sleep mode to maximize energy efficiency and sleeping cell detection for improved reliability. Leveraging our global market presence to optimize our customers' network to gain insights, we have developed telco-tailored machine learning operations (MLOps) to assure and retrain AI models for best in class AI-native features.

Ericsson leverages AI in the following areas:

- **Enhanced throughput for heavy users (link adaptation):** AI predicts how to best handle the radio channel conditions traffic patterns and user demands, allowing the network to dynamically adjust radio link configurations achieving 11,6 % improved throughput for heavy users measured in field by CSP.
- **Improved handover speed and reduced dropped calls (mobility):** AI anticipates best target for handover due to poor coverage user movement and network congestion. Machine Intelligence Enabled Mobility has achieved 1,2% reduction in overall drop rate and 10,9% reduction in inter frequency handover failure.
- **Increased energy efficiency (MIMO sleep time):** AI predicts traffic patterns and cell usage. Based on these predictions, the network can activate or deactivate MIMO functionalities in cells, optimizing energy consumption without compromising performance. we have estimated up to 14% of reduced energy consumption per radio site.
- **Network anomaly detection:** AI continuously analyzes network behavior to identify unusual patterns that might indicate potential issues like sleeping cells (cells with no traffic). Early detection allows for corrective actions to maintain optimal network performance. We have measured more than 95% of detection rate of sleeping cells with AI powered advanced cell supervision.

Through these AI-powered features, Ericsson has significantly improved network performance, addressing the challenges of user experience, network efficiency, reliability, and resource management. Transitioning from 4G to 5G, we are leveraging the robust RAN expertise and AI learnings from 4G features. We have already started with new features in advanced phases of AI and automation.

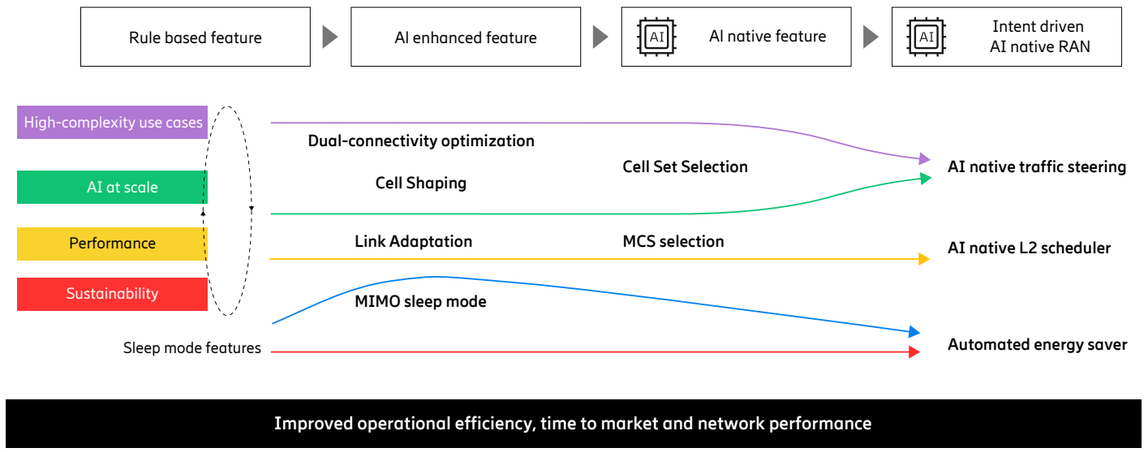


Figure 3: Radio features evolution for advanced AI

A new architecture for AI and Automation

Wherever global applications are downloaded, installed, and launched, they will need to be informed about the available connectivity and current purchased plans. Programmable networks enable not only application service providers to utilize 5G networks for improved applications, but also allow communications service providers (CSPs) to offer customized plans and flexible payment options. It opens doors for CSPs to manage their revenues and investments in a new way and enables mobile networks to evolve into innovative platforms through network exposure and APIs.

As part of the Open RAN architecture, rApps are the software elements controlling the slow loops. They are hosted in Ericsson Intelligent Automation Platform (EIAP), which represents Ericsson's implementation of the Open RAN-specified service management and orchestration platform (SMO) with additional capabilities.

Ericsson´s concept of high-performing programmable networks are mobile networks with enhanced capabilities that will allow CSPs to create new business models for more efficient operations and agility to create new services. For higher levels of automation, networks are moving towards a concept called intent-driven networks. In this approach, CSPs simply specify their business goals and the network itself translates those goals into the necessary actions.

These networks are built with 5G standalone (SA) as the foundation, following the standards architecture (3GPP, TM Forum, and Open RAN) with new software technologies and AI for advanced automation and SLA awareness.

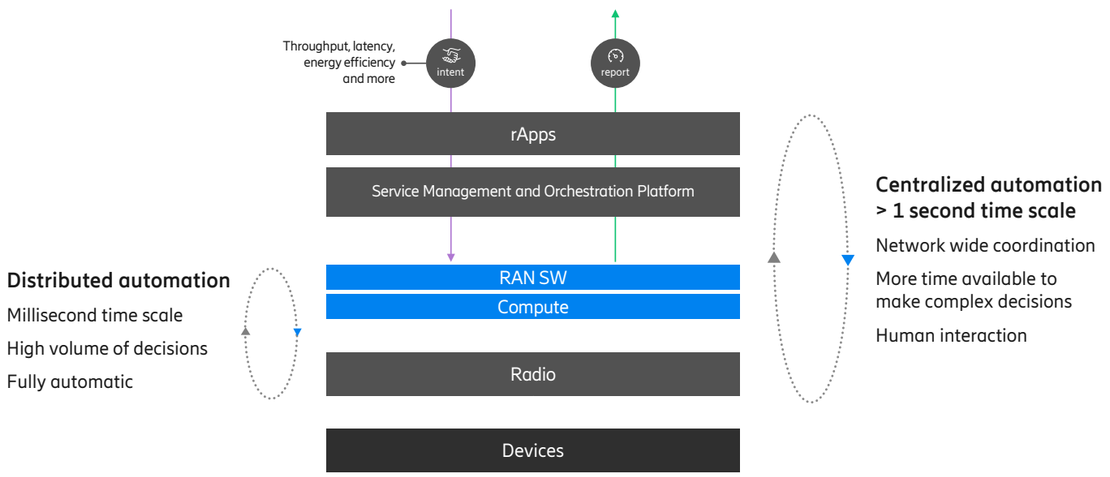


Figure 4: AI-powered architecture of programmable networks (RAN domain)

We have multiple examples of the benefits of implementing AI in this architecture which leverage on the advanced radio capabilities (distributed automation) to further boost improvements at network level via rApps (centralized automation). Cell shaping is one example of this approach. Cell shaping is the technique of letting the system adapt and update the beam shape of the radio signals. We have multiple radio features for optimization such as the Interference Sensing [3] feature which received a GTI award, this feature minimizes inter-cell interference and improves 40% the cell capacity that can work in combination with Ericsson Cell shaping rApp.

Ericsson Cell Shaping rApp uses reinforcement learning to adapt the cell shape based on the characteristics of each cell and its influencing area. Continuous closed-loop optimization automatically maintains the optimum settings as the network evolves and traffic distributions change resulting in 30% and 5% higher throughput in the uplink and downlink respectively. Broad beamforming technology for Synchronization Signal Block (SSB) is a key enabler [4].

Remote Electrical Tilt (RET) is used and controled by the rApp optimization is used to improve coverage, minimize inter-cell interference and distribute traffic for improved network utilization.

Both tilt (vertical plane) and horizontal beamwith of Active Antenna Systems (AAS) are adjusted autonomously. Field trial results executed on a dense urban network of a tier-1 European operator showed Failure Rate reduction around 47% .

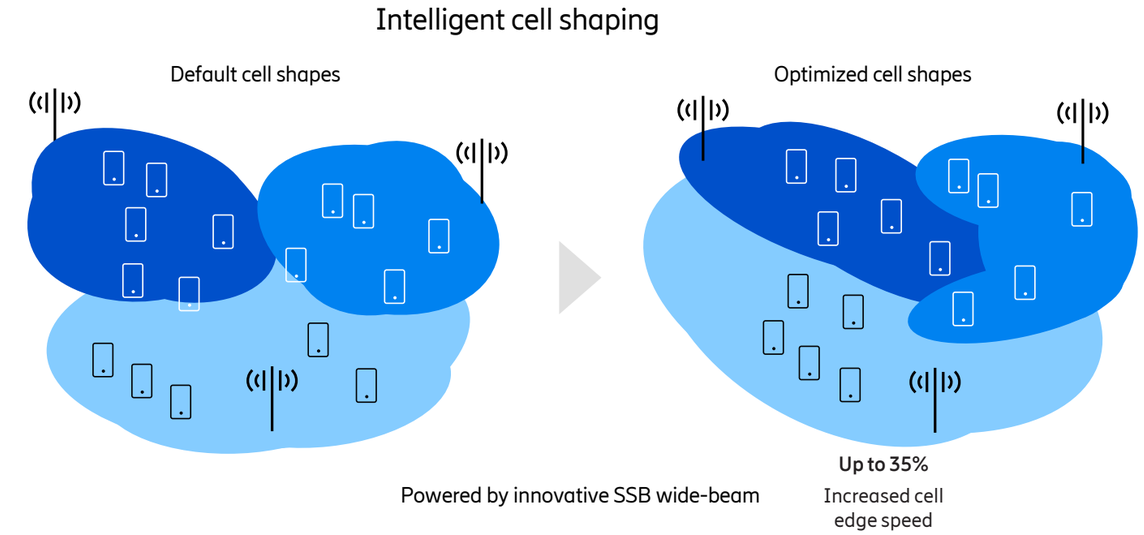


Figure 5: Improvements with Cell Shaping rApp and advanced Massive MIMO radios

AI in RAN aims to use machine learning (ML) algorithms and data from the RAN to reach data-driven decisions, which can be in the form of predictions, insights, or control actions that augment or replace the rule-based approaches used in RAN features.

Effective ML algorithms for RAN require exposure to representative data and training on valid datasets to accurately predict observed and unseen data. In RAN, models can generate predictions on various timescales using fast control loops and capture longer-term predictions using slow control loop models. However, training and inference in the latency and compute stringent RAN environment becomes more challenging with faster control. As a result, the use of AI in RAN needs specialized knowledge of RAN software and hardware, along with customized AI solutions designed specifically for the unique requirements of RAN, rather than generic, plug-and-play AI tools.

3. Ericsson Interference Sensing software recognized by GTI
4. Broad beamforming technology in 5G Massive MIMO - Ericsson

Ericsson´s AI in RAN proposal

Ericsson is building AI in RAN portfolio across L1, L2, L3 (according to OSI model) that leverages key AI capabilities to continue the evolution as described in chapter 3.

Based on Ericsson's long experience on working with AI, there are three main categories of use cases where AI adds unparalleled value and high ROI:

AI unique use cases

AI's uniquely applicable scenarios means that specific AI tools will give added benefits, which include prediction, anomaly detection, analysis, and so on.

High complexity use cases

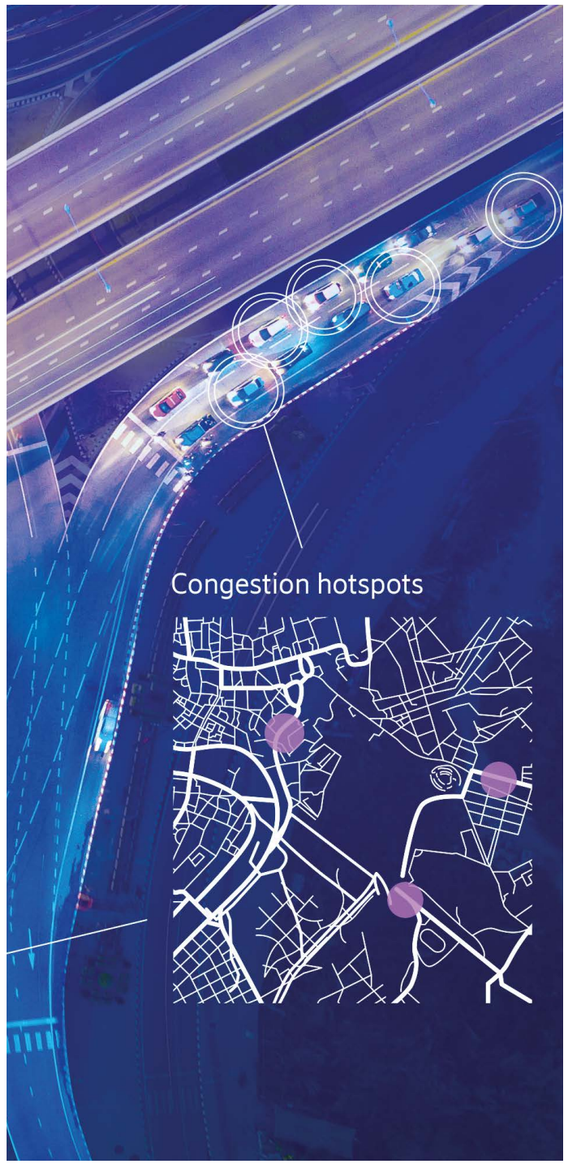
One example is the link adaptation feature. This functionality is designed for situations where there are multiple factors to be considered to find the best signal modulation for optimal transmission.

Multi-objective use cases

These use cases involve seeking an optimal solution under complex constraints and trade-offs. For instance: handling handover choices while maintaining certain network KPIs and maximizing energy savings while improving user experience.

AI is expected to add significant gains in being able to simplify the RAN. Major gains should be expected in high-complexity scenarios that can be also high-demand scenarios with lower spectral efficiency. Moving to AI native features allows to address high complexity decisions derived from multiple objectives with potential trade-offs.

This also means that there are areas where AI will not add significant value such as low complexity scenarios, and where simple heuristics provide satisfactory network performance.



AI evolution phases

Achieving the vision of high performing programmable networks will be an evolution journey executed in four phases:

1. **Rule-based solutions** in RAN software use thresholds to make simple predictions and control assets based on channel conditions, traffic levels and so on. These solutions have been foundational for RAN software through various generations (1G-3G).
2. **AI-enhanced features** include a distinct AI component that provides advanced capabilities, such as making sophisticated predictions based on past patterns. This allows for easy deactivation if desired.
3. **AI native features** integrate AI more deeply, controlling multiple capabilities based on various factors such as traffic patterns, variance, and system capabilities.
4. **Intent-driven and AI native RAN** extends AI's impact broadly in the RAN, for example enabling coordination across different RAN domains like energy efficiency, traffic management, spectral efficiency, and so on.

Moving towards AI-native RAN unlocks significant advantages:

- **Scalability:** AI can handle a growing number of complex intents efficiently, unlike rule-based systems that can´t manage higher-levels of complexity.
- **Optimized decision-making:** AI can make real-time decisions based on vast amounts of data, leading to more efficient and optimal network operations.

AI native will be needed for managing complex intents and unlocking the full potential of high-performing programmable networks. We will elaborate in the next chapters on the building blocks of AI native RAN.

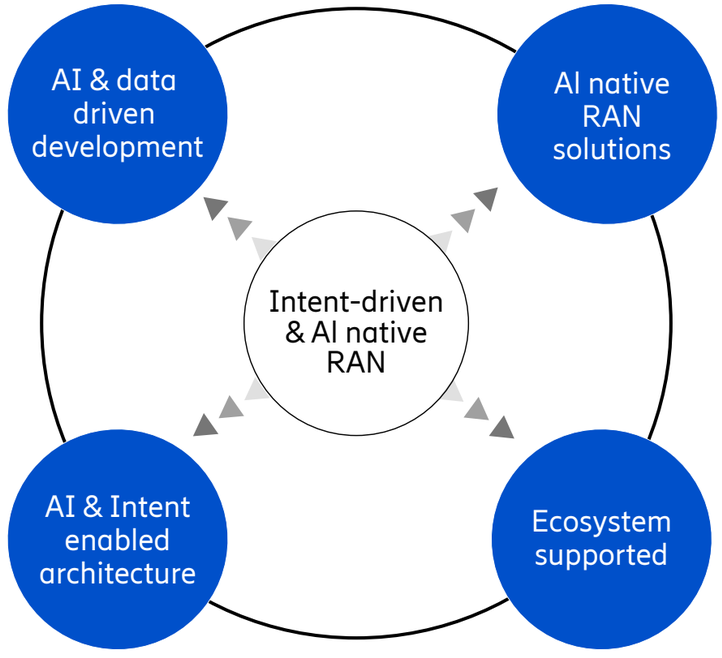


Figure 6: The building blocks of Intent-driven and AI native RAN

AI-enabled architecture and features

AI implementation complexity challenges are:

- Serving RAN data to ML, efficiently and sustainably.
- Understanding DataOps and use cases.
- Resource limitations and latency requirements.
- Data locality, hardware-optimized models, managing the lifecycle of the model.

Based on the above challenges, enabling AI in the RAN follows three principles:

- Minimize complexity
- Evolve and enhance existing networks
- Ensure future AI scalability

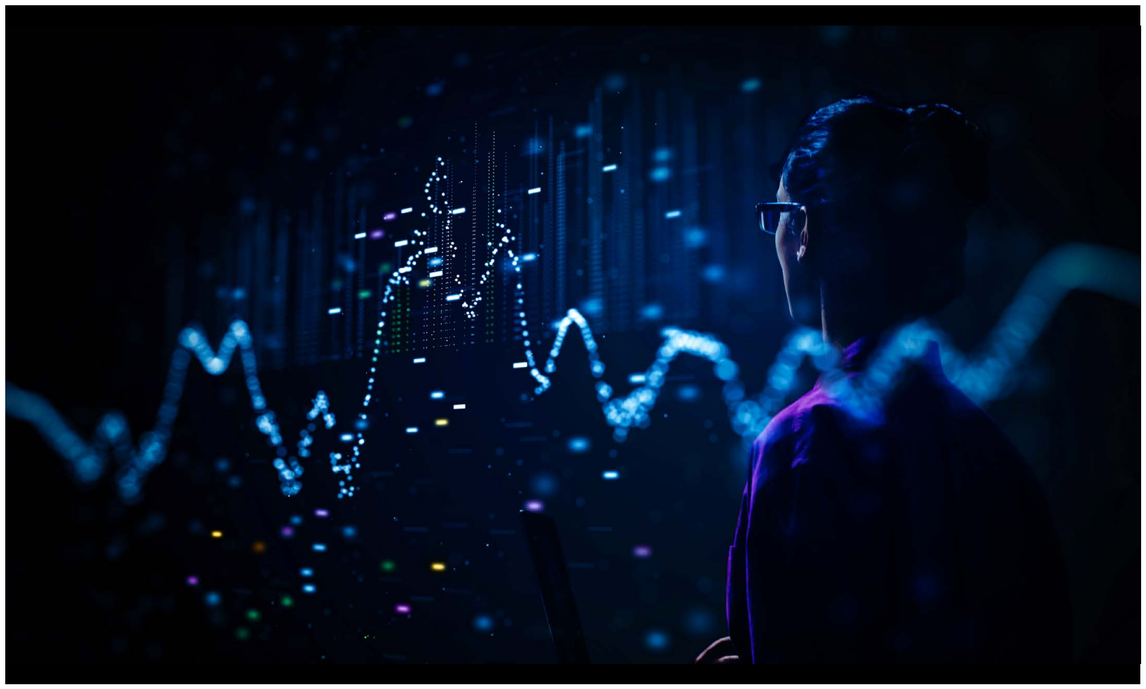
CSPs globally target reduced network complexity and so will not allow solutions that add unnecessary complexity. The principles will ensure that AI is possible and feasible to implement for CSPs.

AI capabilities need to be built with today's networks in mind and must still be robust enough while allowing for dynamic AI development. Addressing AI in RAN includes constructing data pipelines, correlating RAN data serving to the ML algorithm, and managing the output, all with tight latency and high robustness. Developing these solutions takes strong domain expertise and ability to experiment in a large installed base and skills in data operations.

Ensuring the future AI scalability also means ensuring that the data locality and system execution is understood and properly commissioned. Low latency inference is not likely as a candidate for offloading outside of the RAN software stack. Offloading other aspects of the ML lifecycle can optimize data locality and hardware use.

AI-based functionality sensitive to latency (that requires real time) must be implemented and distributed in the radios, while other type of functionality not strongly time-sensitive can be implemented in a centralized location to maximize hardware.

The five required capabilities of the AI system include inference, retraining, assurance, explainability, and experimentation. Ericsson has already deployed and tested such capabilities, which are implemented in the AI model lifecycle. Here, explainable AI helps to understand and interpret the output provided by the AI models. This is relevant to make AI solutions more transparent with regard to reasoning, choices, and impact.



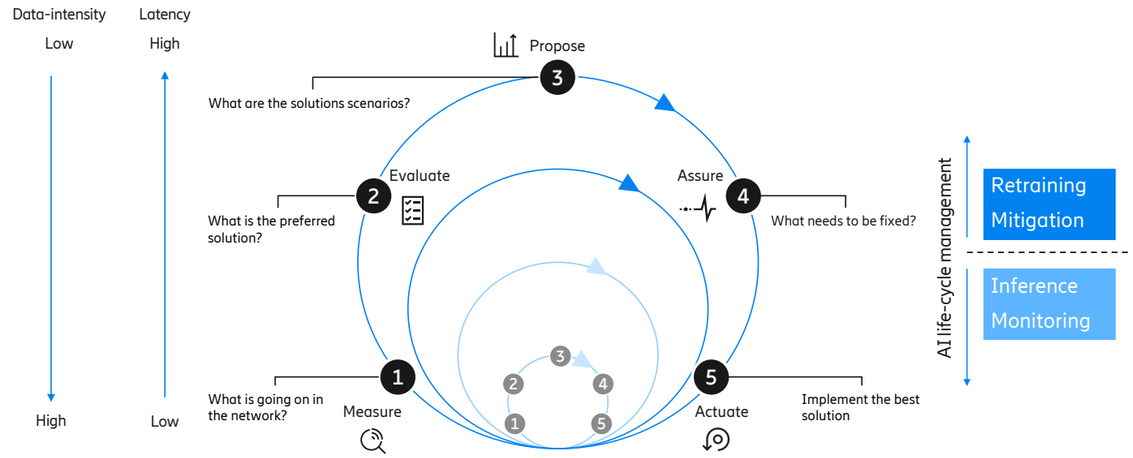


Figure 7: AI lifecycle in RAN

Generative AI

In contrast to discriminative AI, which makes predictions or insights, Generative AI can generate new data in forms like text, image, audio, and video. This is made possible by transformer-based neural networks and attention-based architecture. Generative AI tools can simulate human-like conversation, offering many possibilites in customer support.

As a complementary technology to AI-native, at Ericsson we leverage the recent breakthrough in Generative AI to support CSPs in accessing our product information effortlessly. Generative AI-based product assistants support CSPs to utilize Ericsson's products efficiently and correctly. Product assistants can guide and support the CSPs in various ways such as installing a hardware unit or solving an alarm.

We are also adding the capability to industrialize AI at scale and advanced AI technologies, including Generative AI. Finally, we embrace openness and believe in innovating with the ecosystem and within multivendor setups.

Data is often referred to as the new oil. What's sometimes forgotten in that analogy is that the process of going from crude oil in the ground to fueling a car in motion is a multi-step refinement process that requires the collaboration of an industry, and this applies for data too. By end of 2023, mobile networks carried ~1,6 Exabytes of data daily globally (Ericsson Mobility Report), this does not account for all the computation in RAN and Core to support the traffic. The sheer amount of data requires a multi-step refinement process to be useful for analysis, decisions, development, and deployment of (AI-based) software at quality and scale. Getting AI to work in RAN will require a large amount of collaboration, including data-sharing, use case collaboration and tuning, and adopting best-practices and tools.

Together with pioneering service providers and leading AI experts, we are building the future of AI native networks.

Ecosystem support for AI native networks

Ericsson's is collaborating with leading companies in the AI industry to make the most of cloud hardware and cloud capabilities. This cross industry collaboration helps to share best practices, and improve existing tools to meet the demands for latency, hardware, and robustness. Ericsson is building tools to ensure that AI can be incorporated into RAN, including broad distribution, low latency, high reliability, sustainability, and cost-efficiency.

We are sharing our best practices and driving the standardization in the wider telco eco-system. For instance, the AI-RAN Alliance where Ericsson is a founding member bringing together the technology industry leaders and academic institutions to enable the enhancement of RAN with AI.

Challenges and success factors in the CSP journey for AI in RAN

AI will add a significant value offering new opportunities to CSPs to improve their business. There are solutions and success factors that will help CSPs in the AI adoption, where ICT industry has specific challenges different than other industries.

The challenges in adopting AI in RAN by communication service providers can be roughly divided into two areas: organizational challenges and technology and tool adoption challenges.

Organizational challenges

Research shows that the organizations that successfully leverage AI address the key challenges related to integrating AI in their business and operations. Key activities include creating a coherent strategy with clear and achievable targets supported across the company to ensure proper collaboration, competence, digitalization technologies, and optimal data management. It is crucial that this collaboration includes both intra- and inter-organizational groups.

Competence, collaboration, strategy, and culture that embrace AI and the surrounding data management creates positive feedback loops and set up organizations for exponential development and gains. Organizations that buy in early and emphatically will be the first to reap the rewards and have a significant head-start on laggards.

Ericsson helps CSPs by collaborating closely in a data-driven approach to AI in RAN introduction.

Technology and tool adoption challenges

The main technical challenges for CSPs when adopting AI technology are:

- **Data management**
  Managing a huge amount of data with efficiency and security, especially user data.
- **AI tools for lifecycle management**
  CSPs need to consider future-proof tools in the existing deployments.
- **Deploying new AI capabilities in their systems**
  These are retraining, assurance, explainability, and experimentation, which require optimized AI/ML models and platforms for execution.
- **Optimal hardware considerations**
  There are new hardware requirements to satisfy AI implementation including storage capacity and processing power. Selecting platforms that effectively support these needs will make a difference in hardware expansion and scalability.

Ericsson helps CSPs in these challenges to successfully implement AI in RAN by evolving today's at-need data management to a framework that is simple to manage and flexible in granularity and timing. This ensures that use cases can be individually supported and scaled. An optimal life cycle management solution and a future proof network require an in-depth look at the specific requirements of a RAN network, its current architecture, and deployments.

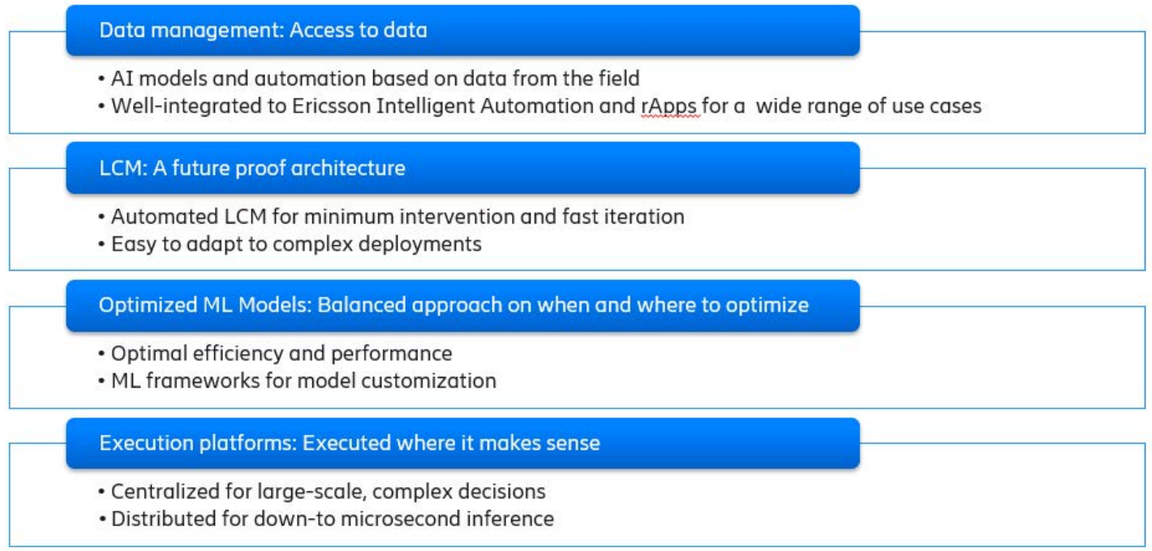


Figure 8: Ericsson´s approach to AI implementation challenges

The data challenge

With 5G, in a typical network deployment, there will be tens of terabytes of data to process from 1000+ distributed sources, opening new opportunities in AI native network optimization, deployment, and healing. This makes data management scalability a key consideration when implementing AI in networks. Either streamed over interfaces or consumed locally, whether aggregated globally or per use case, serving RAN data to ML algorithms in an efficient and sustainable way requires a profound understanding of the data operations and the RAN use cases. At Ericsson, we have been working with our customers over the years, improving our RAN features and providing data analytics services to them.

The hardware challenge

AI industry and current paradigms have seen an exponential growth of hardware needs for today's cutting-edge use cases far beyond the compute growth following Moore's law. Moore's Law is an observation that the number of transistors on a microchip tends to double approximately every two years, reflecting the historical trend of microchip technology. In recent years, maintaining this pace of miniaturization has become increasingly challenging due to physical limitations. The slowdown of Moore's Law has led to increased focus on alternative approaches for improving computing power, such as new chip architectures and AI.

Ericsson is developing AI prepared hardware for the distributed capabilities. Our next-generation RAN Compute portfolio has an architecture that not only enables AI but also provides storage and processing capabilities optimized for AI models execution.

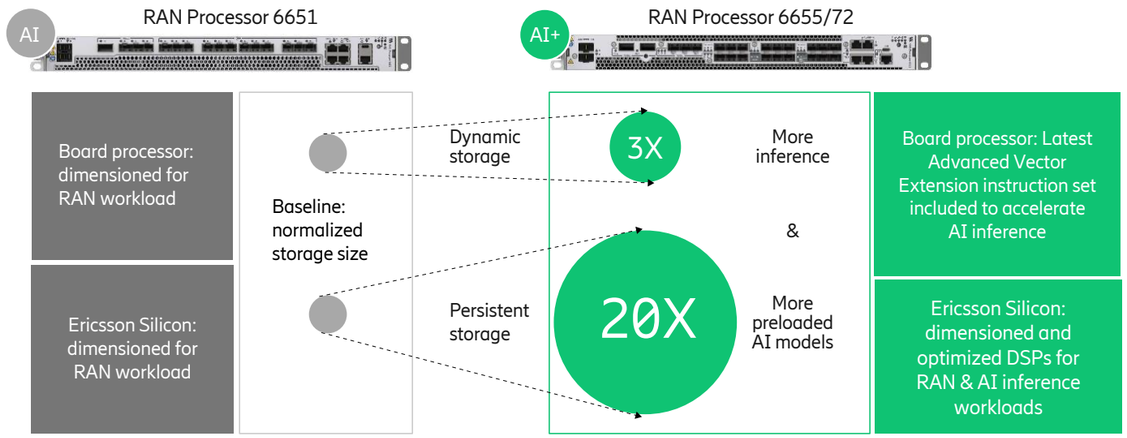


Figure 9: AI enhancements in the latest RAN Compute generation

Building on both existing centralized and distributed capabilities, with room for future evolution and integrating RAN-specific AI lifecycle management, is key to unlocking the potential of future AI-powered networks.

Mobile networks are both distributed (radio sites) and centralized (data centers). This brings the opportunity of executing AI in both locations for different purposes and time requirements (real-time and non-real time). By distributing the time-critical operations in the fast control loops and centralizing non-critical optimization tasks in the slower control loops, it will be possible to effectively utilize and expand today's RAN systems into the AI native RAN systems of the future.

Actionable insights

Modernization of the network towards future-proof AI-native architecture

Network modernization is a key for competitive customer experience and energy efficiency. A future-proof AI-native architecture combines the best out of new hardware and software executing AI-native features seamlessly. AI-native features can increase 11.6% the throughput for heavy users and improve close to 15% power savings in MIMO radios.

Simplifying operations leveraging powerful AI-native features

Breaking the operational complexity curve is made possible with intelligence. Distributed AI monitors and executes faster than the blink of an eye (microseconds) to optimize operations. Meanwhile, centralized AI collects long-term insights allowing for a higher degree of intelligence and further opportunities for automating and simplifying CSP' operations and AI-powered automated solutions.

Data-driven organizational transformation through vendor partnerships

Data-driven organizations leverage data and AI early on in their digital transformation journey. Managing the volume and the quality of data for AI calls for specialized know-how of the network domain, as plug-and-play AI tools fall short for data-driven CSP transformation journey. Vendor partnerships give head start in managing data in a cost-effective and best-use oriented approach.

Recommendations

01 **Embracing AI technology at an early stage will give you a competitive advantage.** Organizations that adopt AI early gain a head start.

02 **Look for AI solutions that are tailored for telco.** Specialized knowledge of the radio network data and radio software is crucial for effective AI implementation.

03 **Speed up network transformation.** Migration to new hardware and software brings gains in spectral, energy, and operational efficiency.

04 **Start your journey with quick wins.** Integrating AI in your existing architecture with a streamlined approach will help in achieving the biggest ROI with the minimal disruption.

05 **Target some big opportunities.** Consider important opportunities like radical improvements to customer experience or a much-simplified network operation.

06 **Plan your data management strategy with AI in mind.** Improve data quality and management by collaborating with vendors.

07 **Evolve your network with a joint distributed and centralized AI strategy.** Leverage centralized and distributed AI to boost further improvements at network level and to maximize hardware use.

08 **Select a future-proof architecture to unlock current and future AI opportunities.** An architecture that follows industry best-practices with efficient data management and life cycle management of the models.

09 **Identify what use cases can benefit the most from AI technology to define your journey.** If you are looking primarily at OPEX reduction or if you are targeting to introduce performance-based business models will determine the evolution of your network.

Authors

Klas Johansson
Head of Automation & Operations
BNEW PAN Automation & Operations

Klas Johansson has over 20 years of experience in the telecommunications industry. Currently, Klas is leading the Automation and Operations team at Ericsson´s within the Product Area Networks. With a rich background in various roles, Klas has been instrumental in driving network evolution, enhancing performance, and shaping AI strategy across different functions; R&D, product management and presales.

Klas Johansson's journey from network optimization to AI strategy and Open RAN exemplifies his dedication to innovation in the telecommunications field. His vision aligns with Ericsson's commitment to creating resilient and intelligent networks for the future.

Christoffer Stuart
Strategic Product Manager
BNEW PAN PL 5G RAN Deployment & Perf

Christoffer Stuart is a Strategic Product Manager at Ericsson working on building smarter networks through implementation of AI for improved RAN performance, energy efficiency, and simplified operations. He has experience from bringing high-value products to maturity and enhancing RAN with tailored MLOps capabilities. Christoffer has helped to set up the vision for intent-based automation and AI-native in RAN and envisions a future where networks are optimized for operator intents, easy to manage, and highly observable.

Melike Erol-Kantarci
Strategic Product Manager AI in RAN
BNEW PAN Automation & Operations

Melike Erol-Kantarci is Strategic Product Manager for AI in RAN at Ericsson within the Product Area Networks unit. Melike plays a pivotal role in developing and implementing forward-thinking data and AI strategies, driving groundbreaking research and products in cloud-native and AI-native RAN. Throughout her 15+ years of career in communications, she has held senior positions in the industry and academia. Melike has received numerous awards and recognitions from technical societies. Most recently she was recognized with a Women in AI North America 2023 award. Melike is a sought-after speaker with 70+ keynotes and invited talks delivered around the globe. She is a highly cited prolific academic, inventor and an influential industry leader in AI in telecom.

Ericsson's high-performing, programmable 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.

ericsson.com