



150 years



Optimizing RAN energy efficiency with AI and automation

Executive summary

In this report

- 01 Today's energy challenges
- 02 Rethinking AI and energy: Greater energy efficiency and beyond powered by AI
- 03 Ericsson's energy-efficient approach for AI RAN: Optimizing energy where it matters most
- 04 Swisscom uses AI-driven RAN automation to improve energy efficiency and maintain user experience
- 05 Building the path to AI-native 6G

Energy has become a strategic priority for communication service providers (CSPs) today. As user demand drives mobile network expansions in coverage and capacity, energy demand continues to increase and impact both operational expenditure and environmental commitments.

At the same time, CSPs are expected to deliver superior user experience and continuously improve network performance to support new artificial intelligence (AI)-driven services.

Traditional optimization approaches alone cannot address these competing demands. Static configurations, manual tuning, and isolated energy-saving features cannot keep pace with the complexity and dynamic nature of modern 5G networks. To improve energy performance at scale, networks must continuously adapt to changing traffic conditions and operational objectives.

AI and automation are making this possible. When the right model is deployed in the right place for the right purpose, AI enables networks to use resources more efficiently, reduce unnecessary signaling, optimize radio operations, and automate energy-saving decisions in real time.

While AI introduces additional compute demand, AI-specific services represented less than one percent of total information and communication technology (ICT) electricity consumption in 2023 [1], highlighting the importance of deploying AI as a powerful tool to efficiently optimize overall network energy and enhance the user experience.

Ericsson supports CSPs in this transformation through a full-stack, AI-driven radio access network (RAN) approach. From high-performing, energy-efficient radio hardware —supporting both purpose-built RAN with Ericsson Silicon and Cloud RAN with any compute platform to intelligent software and automation architectures, Ericsson integrates AI across every network layer. This ensures that compute power is applied where it delivers the most value, with the right AI model in the right place, to achieve the best telco-grade AI RAN energy efficiency.

“At Ericsson, we are committed to achieving Net Zero carbon emissions across our value chain by 2040*. A critical part of this commitment is supporting our customers in achieving their own climate ambitions. A well-designed, AI-powered portfolio could drive energy efficiency and enable sustainable network evolution, through collaboration between Ericsson and our customers.”



David Hammarwall
Head of Business Area Networks

*<https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/environment/climate-action>

Key insights

Will AI drive energy demand or energy efficiency?

Both. AI-powered applications increase traffic demands, uplink needs, and new AI workloads, driving higher energy demand. On the other hand, AI and automation offer greater potential to improve energy efficiency, operational efficiency, and network performance, and to secure a pre-defined user experience.

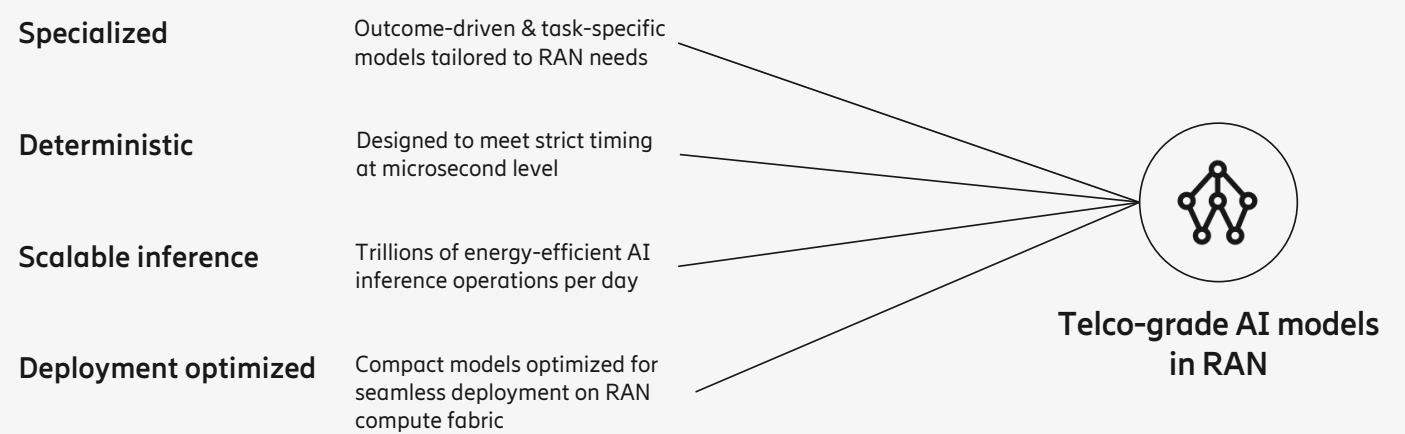
How can AI deliver the greatest value for the telecom industry with minimal additional energy use?

Not all AI models are suitable for the RAN. The greatest value of AI in RAN comes from specialized, deterministic, scalable, and deployment-optimized AI models that are purpose-built for the realities of live radio environments. The proven telco-grade AI model provides better network performance with close-to-zero additional energy consumption. (see figure 1)

How is AI shaping energy-efficient networks today and 6G tomorrow?

AI-driven energy optimization is already delivering measurable results in live networks, with [Swisscom](#) highlighted as one example in this report, and it is shaping the design principles and operational foundation of future AI-native, energy-efficient 6G networks.

Figure 1: Telco-grade AI model



1

Today's energy challenges

Energy is no longer just an operational cost for CSPs. As networks grow, CSPs must balance energy efficiency, network performance, user experience, OPEX, and carbon reduction objectives.

A holistic approach to network evolution and operations is required to align business and sustainability ambitions and support long-term competitiveness. [1]

According to GSMA analysis [2], AI-related traffic could account for 15–60% of mobile data traffic by 2030. Both network infrastructure and data centers are expected to continue growing as AI-driven workloads expand.

For CSPs, this translates into direct business impact. Energy typically accounts for 15 to 20 percent of total operational spend for an average telco [3], and it is one of the few cost categories expected to increase in the coming years. At the same time, sustainability commitments, including Net Zero targets and regulatory requirements, are placing additional pressure on CSPs to reduce carbon emissions while continuing to scale network capacity.

RAN remains the dominant contributor of mobile network energy consumption. For most CSPs, over 80 percent of energy is consumed in the RAN, with the remainder used by the core network, support systems, and owned data centers (see figure 2) [3]. This is driven by the need to support higher throughput, wider coverage, and increased densification, particularly with the rollout of 5G and massive multiple input, multiple output (MIMO) technologies.

In addition to hardware modernization and retiring 2G and 3G systems, CSPs have historically relied on techniques such as static sleep schedules, manual configuration changes, and rule-based automation to reduce energy consumption. While these energy saving software solutions have delivered valuable gains, they were designed for networks that were significantly less complex than today's highly dynamic 5G environments.

As the industry prepares for AI-native 6G networks, energy efficiency must be designed with intelligence and automation embedded across the network from the start.

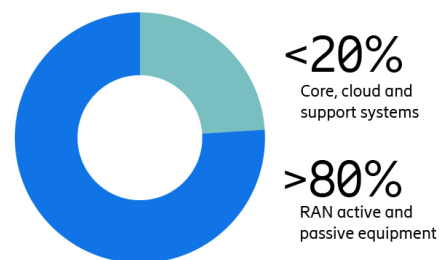
Modern networks must continuously balance competing objectives:

- supporting growing traffic demand
- maintaining network quality and delivering the wanted user experience
- controlling operational costs and meeting sustainability targets

Static approaches struggle to achieve these objectives simultaneously because they cannot adapt fast enough to real-world conditions.

At the same time, concerns about AI energy consumption in mobile networks are rising. An Ericsson white paper shows that, while AI introduces additional compute demand, AI-specific services represented less than 1 percent of total ICT electricity consumption in 2023 [1]. Despite the soaring data demand, the energy use of telco networks has risen only modestly, representing 1% of global electricity use in 2023, up just 12% since 2019 [3].

Figure 2: Breakdown of CSP network energy consumption



Will AI drive energy demand or energy efficiency?

AI is reshaping the connectivity energy landscape from two directions. First, AI-driven applications and use cases such as AI assistants, agentic applications, industrial automation, immersive experiences, and intelligent devices will reshape the connectivity needs on mobile networks and are expected to further increase traffic, leading to growing network energy demand. AI will drive additional energy consumption through data center workloads and advanced compute requirements.

It also changes the type of traffic networks must support; AI-native experiences generate more uplink traffic, require lower latency, and increase real-time inference requirements. These new demands will need additional intelligence in the RAN itself.

The challenge is therefore not only managing network energy consumption, but also delivering AI capabilities with the lowest possible energy cost per inference and per delivered bit.

Second, AI also offers significant potential to elevate network performance, automation, and energy efficiency for best return on investment (ROI), implemented only when needed. By continuously analyzing network conditions, predicting demand, and automating optimization decisions, AI enables a fundamentally different approach to managing performance and energy use. As AI adoption accelerates, intelligent, dynamic energy management aligned with capacity will be needed.

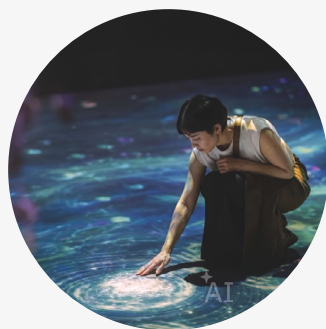
This creates an important inflection point for the industry. The question for CSPs is no longer whether energy optimization is important. The question is how to evolve networks and place compute resources to balance growing demand for capacity, service quality, and AI capabilities. Achieving this balance requires a holistic evaluation of network evolution choices, with energy efficiency and the intelligent use of AI in mind.

Figure 3: AI is reshaping the connectivity energy landscape from two directions

AI can increase energy demand

AI-driven applications reshape connectivity:

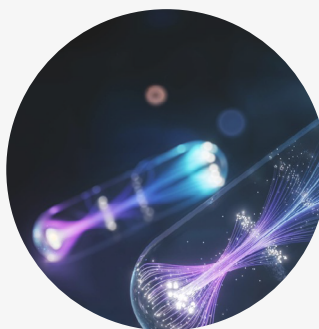
- Increasing traffic demand
- Stronger uplink need
- New AI-driven workloads



AI improves energy efficiency

- Implemented only when needed

- Elevated energy performance
- Enhanced user experience
- Best for operational efficiency, OPEX and ROI



2

Rethinking AI and energy: Greater energy efficiency and beyond powered by AI

The dual impact of AI on energy is clear: AI can increase energy demand through AI-driven workloads while also creating opportunities to improve efficiency through AI-enabled optimization across networks and operations.

For CSPs, the implication is strategic: the question is not whether AI consumes energy, but whether AI is implemented in a way that reduces total system energy while increasing operational efficiency, protecting performance, and user experience. In mobile networks, that system view matters. The key is to implement the right model in the right place with the right compute.

Why inference efficiency matters
Different compute platforms have different energy efficiency profiles. The energy impact of AI depends not only on the model, but also on where and how inference is executed. Large foundation models are typically trained and executed in centralized data centers using high-performance GPU infrastructure. These models require substantial compute resources and consume significant amounts of energy. Telco networks have different requirements.

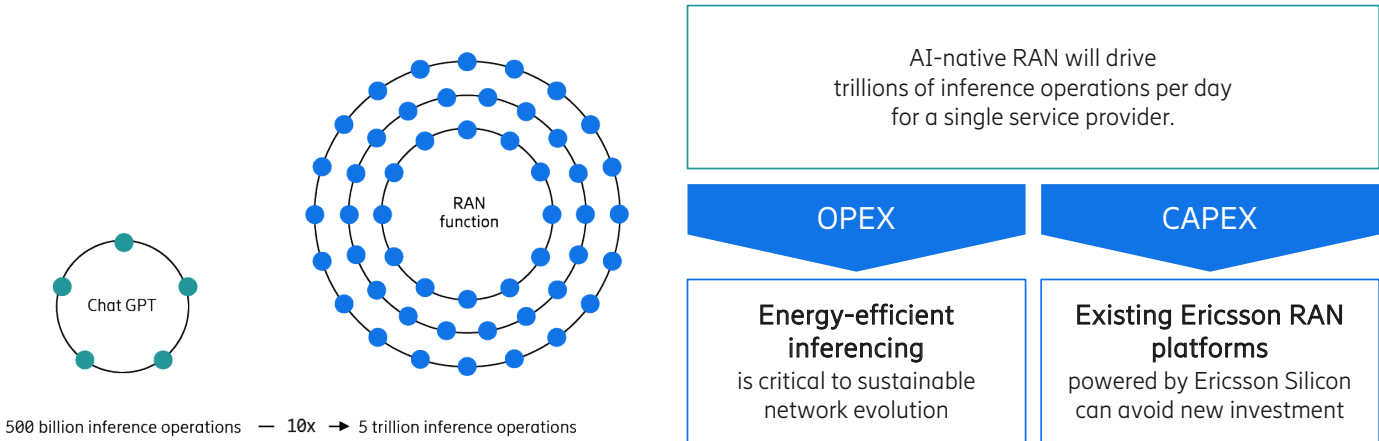
AI-native RAN drives inference volumes into trillions per day for a single CSP [4]. That scale has direct business implications:

- On the OPEX side, energy-efficient AI inference becomes critical to network evolution.
- On the CAPEX side, the ability to run AI on existing Ericsson RAN platforms helps CSPs eliminate avoidable spending.

AI used inside the RAN must operate at large scale, deliver deterministic behavior, and respond within extremely short time intervals. These workloads do not require massive foundation models. Instead, they rely on highly specialized, telco-grade AI models that are optimized for specific network functions.

Because these models are significantly smaller and designed for operational efficiency, they can deliver meaningful network improvements with minimal incremental energy consumption. AI efficiency therefore depends not only on the model itself, but also on where and how inference is performed.

Figure 4: How AI-native RAN drives inference growth



Energy-efficient telco-grade AI RAN

The RAN remains the dominant energy consumer across mobile infrastructure, which means that even modest percentage improvements in RAN energy performance can outweigh the incremental energy needed to run AI models.

Ericsson's approach is to place AI capabilities where they create the greatest value. It is built around the right compute platform, the right AI models, and the right place to achieve the best energy efficiency, optimized for telecom workloads.

Telco-grade AI models outperform highly optimized traditional algorithms and provide clear gains in spectrum efficiency and network capacity, while enabling new services with greater operational and energy efficiency. Best-in-class AI models with optimal data and training strategies will provide higher performance and the lowest energy consumption for AI inference.

CSPs will have near 0 additional energy cost in Ericsson RAN Compute*

*based on Ericsson's field measurements of AI-native Scheduler for Link Adaptation.

Examples include:**AI-native Scheduler for Link Adaptation:**

20% higher downlink throughput and up to 10% improvement in spectral efficiency versus a highly optimized traditional link adaptation algorithm by running the solution as AI-native software on existing Ericsson RAN hardware, with energy-efficient AI inference and no additional hardware required.

AI-managed beamforming:

Running on AI-ready radios with Ericsson Silicon optimized for energy efficiency, supporting twice as much high-traffic and improving 5% spectral efficiency for high load.

AI-powered coverage prediction:

90% to 95% coverage prediction accuracy, improving energy efficiency and user experience through reduced signaling and intelligent resource allocation.

RAN automation for energy efficiency

Energy efficiency in modern networks increasingly depends on real-time, intent-driven RAN automation. Instead of relying on static configurations, CSPs can continuously adapt capacity to traffic demand, prioritize outcomes over features, and avoid energy waste.

Site-level AI delivers fast, local decisions in a distributed manner. But energy performance is also a system-wide orchestration challenge: CSPs must coordinate policies across cells, clusters, and frequency layers while protecting key performance indicators (KPIs). This is where energy optimization rApps become essential by complementing distributed automation to reach energy improvements at scale.

The following Ericsson rApps represent the initial wave of AI-driven energy optimization capabilities, with further innovations expected as adoption grows.

Protecting performance while saving energy:

The Energy Performance Orchestrator rApp analyzes the relationship between network performance and energy-saving features. It monitors affected and neighboring cells to detect degradation early and protect KPIs.

Seeing where energy performance can improve:

The RAN Energy Cockpit rApp provides visibility into energy performance across radio units and the wider network. Site-level visualizations, anomaly detection, recommendations, and metrics such as kWh/GB help CSPs benchmark performance and identify opportunities for improvement.

Optimizing Long Term Evolution (LTE) energy use automatically:

The 4G Energy Optimizer rApp dynamically adjusts LTE cell sleep-mode parameters based on traffic behavior while continuously monitoring KPIs.

Using the right energy source at the right time:

The Site Energy Orchestration rApp uses AI and machine learning to coordinate energy sources and supply site systems with the most suitable source at each moment. This helps CSPs reduce energy OPEX and unlock new revenue opportunities.

These rApps enable controlled, KPI-aware optimization through continuous monitoring, validation, and adjustment, ensuring that energy efficiency gains do not compromise network performance.

AI can add compute load and increase energy consumption, but when it is deployed and executed according to the so-called precision, not overkill principle, it reduces much larger sources of waste in the network. AI is therefore not an additional cost layer, but a mechanism to continuously align capacity with demand, transforming energy optimization from a periodic, manual activity into a real-time, intelligent capability.

3

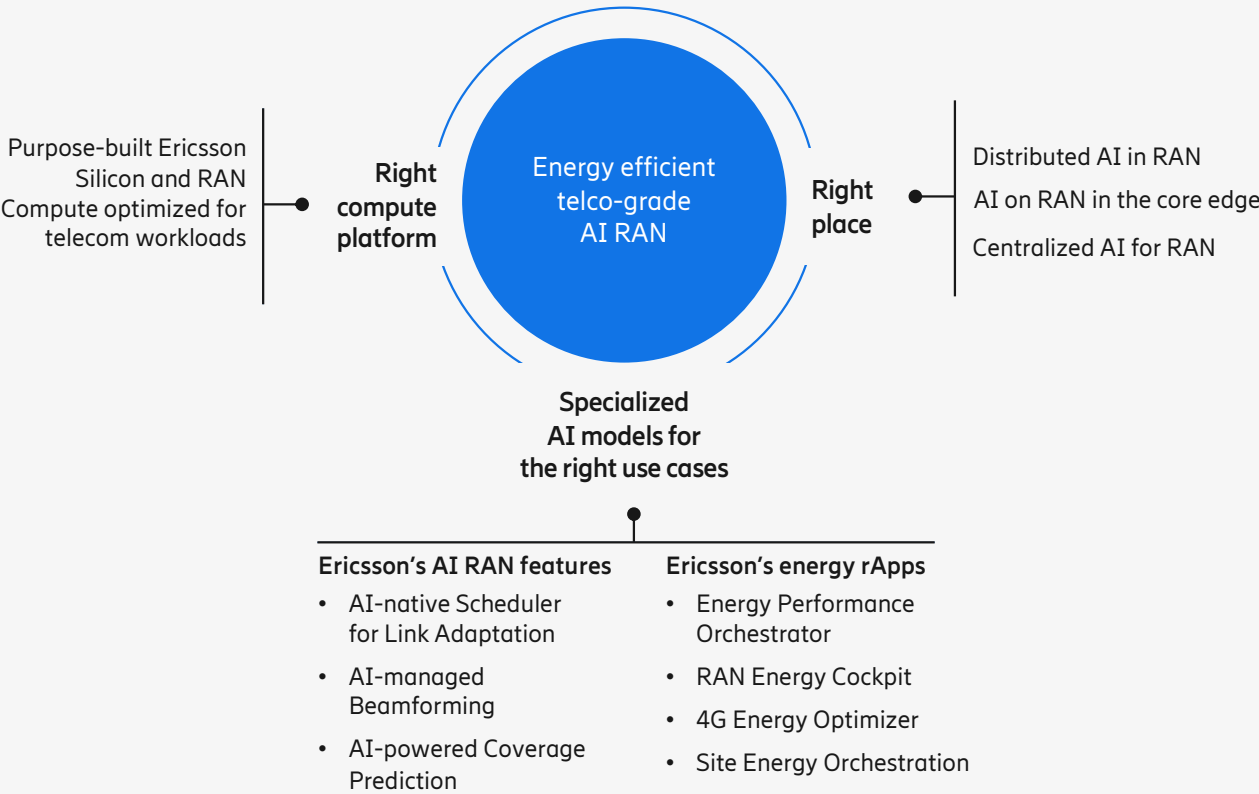
Ericsson’s energy-efficient approach for AI RAN: Optimizing energy where it matters most

Improving energy efficiency in the RAN requires more than isolated features. As networks become more complex, energy efficiency depends on how effectively compute, hardware, software, and automation work together across the entire RAN. This is where Ericsson’s approach differs.

Ericsson focuses on using AI with precision, not overkill. The objective is to maximize network performance, user experience, and business outcomes while minimizing the energy required to deliver them with optimal ROI.

By embedding intelligence across the full RAN stack, the network can dynamically activate, scale, and optimize resources, ensuring that energy is used only where it delivers value: without compromising performance or the user experience.

Figure 5: Energy-efficient telco-grade AI RAN with Ericsson examples



AI-ready full RAN stack: Embedding energy- efficiency across all layers

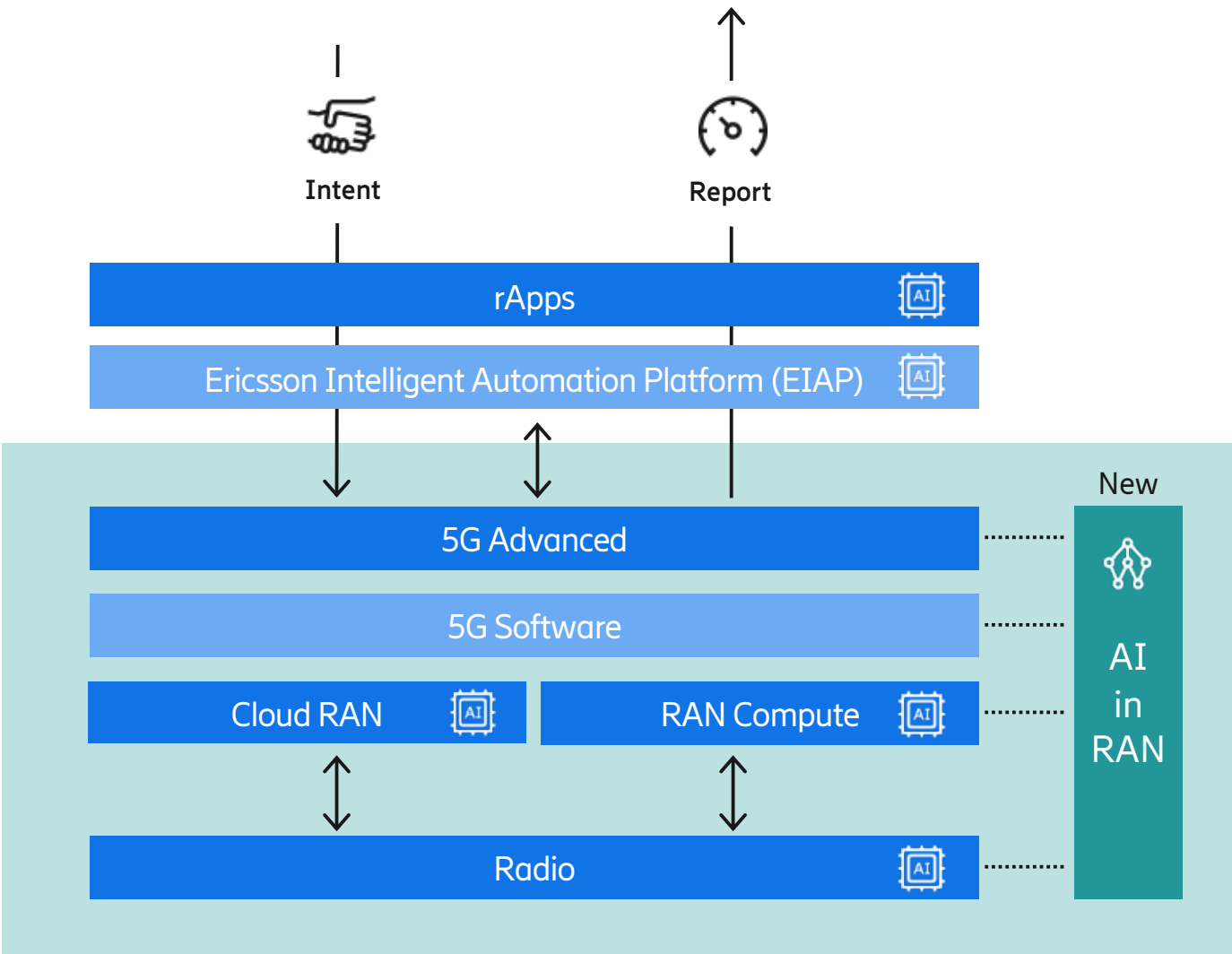
Improving energy efficiency in mobile networks requires coordinated control across the entire RAN.

An AI-ready RAN combines intelligent radios, high-performance RAN compute, advanced software, automation platforms, and network-wide orchestration into a unified system capable of continuously optimizing energy consumption.

- At the radio level, AI makes near-real-time decisions on resource usage, such as whether specific carriers, frequency layers, or antenna configurations are required. These decisions are driven by continuous traffic prediction, allowing the network to dynamically reduce power consumption during low-load periods without affecting the user experience.
- RAN Compute enables this capability at scale by providing the processing infrastructure for AI inference. Because these decisions must run continuously across thousands of sites, compute efficiency is essential. Embedding AI execution within existing RAN Compute enables energy-saving decisions locally, with minimal additional overhead.

- At the software layer, intent-driven capabilities allow CSPs to define energy-efficiency intents directly. The network continuously interprets these intents and automatically adjusts its configuration, ensuring energy consumption stays aligned with both traffic demand and performance targets.
 - AI-enabled, network-wide orchestration ensures that energy optimization is consistent across all layers. Rather than optimizing isolated components, the network behaves as a system, balancing performance, capacity, and energy usage simultaneously.
- By embedding AI across all RAN domains, the network can shift from static energy provisioning to dynamic, demand-driven energy consumption, which is the fundamental driver of long-term efficiency gains.

Figure 6: AI-enabled RAN



Hardware: Purpose-built infrastructure for energy-efficient AI

AI in RAN operates continuously, often making thousands of decisions every second across thousands of sites. These decisions must be executed within strict latency requirements while maintaining carrier-grade reliability. In this environment, every inference’s efficiency matters.

Ericsson addresses this challenge through Ericsson Silicon, a purpose-built architecture developed specifically for radio and RAN compute workloads. It provides specialized processing capabilities optimized for both signal processing and AI inference, enabling high-performance execution with low energy consumption. Because radio processing and AI workloads share similar mathematical operations, such as matrix computations, the same hardware resources can serve both efficiently.

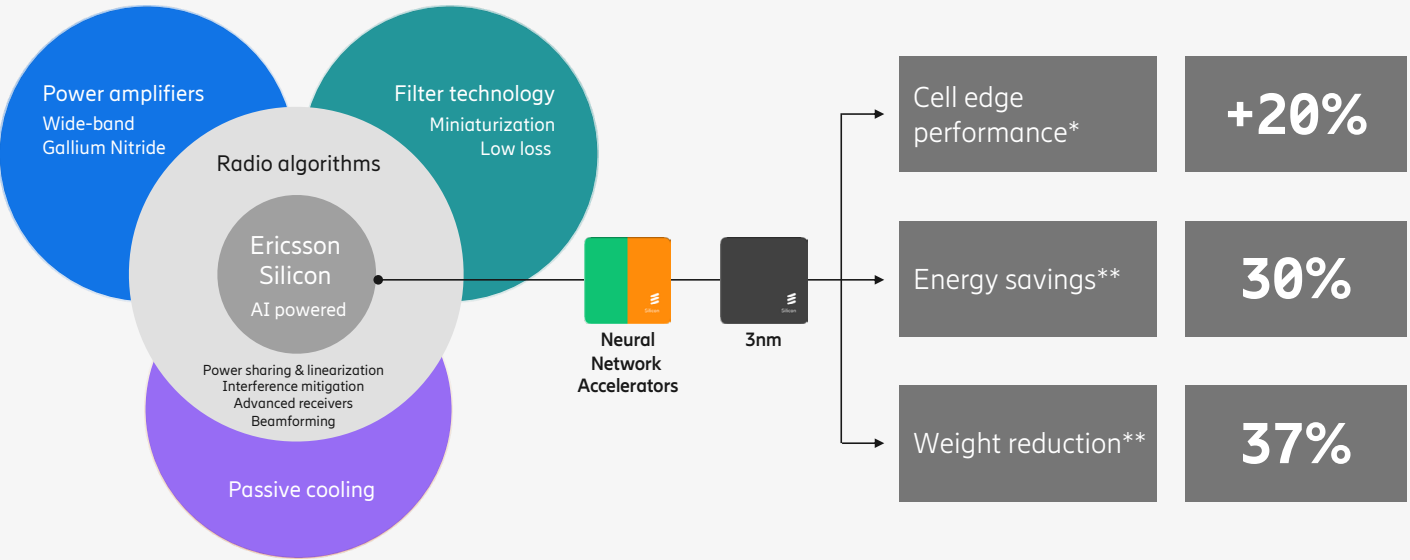
The introduction of neural network accelerators into Ericsson Silicon initially provided massive MIMO radios with up to 10 times higher AI inference compute capability while enabling up to 30 percent lower energy consumption, 37 percent lower weight, and 20 percent better cell-edge performance compared to the previous generation.

Modern RAN compute combines purpose-built application-specific integrated circuits with general-purpose processing and optimized instructions and memory architecture. This enables real-time and AI-based decisions for a large number of users leading to enhanced network performance while maintaining low energy consumption.

AI-ready radios further enhance energy efficiency by enabling local execution of time-critical optimization tasks. For example, AI can dynamically adjust transmission power, reduce interference, and lower signal power. Executing these decisions directly in the radio reduces the need for centralized processing and minimizes data transport, further reducing energy usage.

The combination of energy-efficient Ericsson Silicon-powered radios and RAN compute creates a hardware foundation where AI can deliver sustained energy savings without increasing the overall energy footprint of the network.

Figure 7: Ericsson Silicon enables higher performance with lower energy consumption in the radio



*AI-powered receiver with neural network accelerators.
**Comparing AIR 3266 using the latest Ericsson Silicon with the previous generation

RAN software: Driving energy performance through intent-driven automation

While hardware provides the foundation, it is software that determines how effectively that foundation is used to unlock the full energy-efficiency potential of AI RAN. Historically, network energy optimization has been largely manual. Today's networks contain millions of constantly changing relationships between users, traffic patterns, radio conditions, applications, devices, and business objectives.

The significant shift toward intent-driven automation enables CSPs to manage this complexity at scale. Instead of configuring individual energy-saving features, CSPs can achieve desired outcomes such as:

- maintaining service quality
- minimizing energy consumption
- protecting user experience

The networks then continuously determine how to achieve those objectives in an intent-driven approach. Ericsson's Automated Energy Saver (AES) exemplifies this approach. AES dynamically adjusts sleep conditions based on real-time traffic and user experience metrics, adjusting radio sleep time while maintaining performance. This enables up to 7.4 percent energy savings without compromising KPIs, demonstrating how intent-driven automation can align energy consumption with actual demand [5].

AI-powered RAN software can forecast when specific cells, carriers, or frequency bands will be needed, allowing it to activate resources as network conditions evolve without impacting performance.

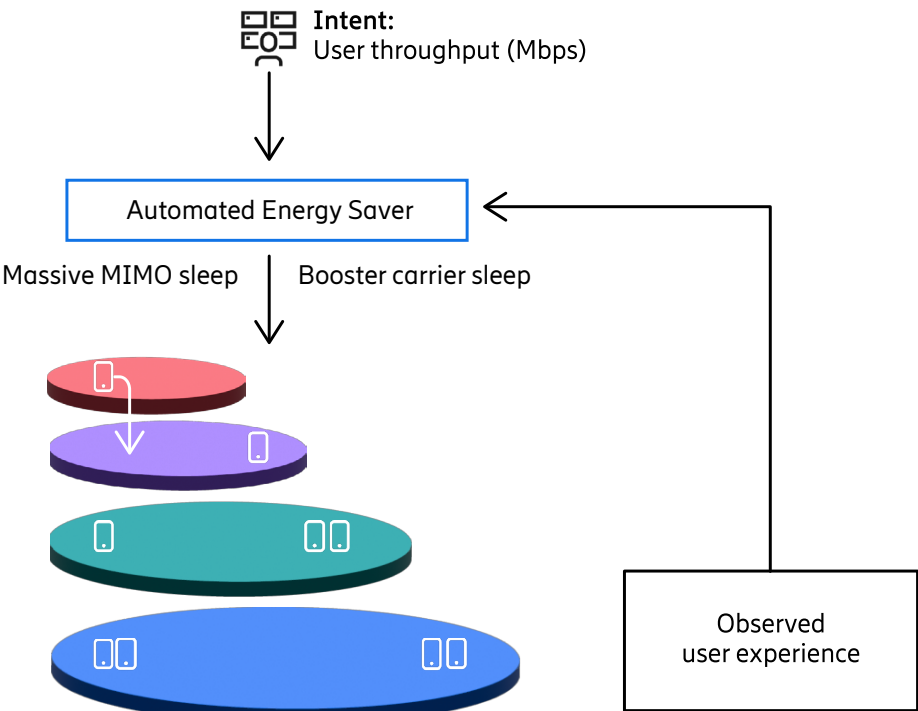
This predictive capability enables dynamic power management at a much finer granularity than previously possible. Instead of reacting to current conditions, the network proactively adjusts its configuration based on expected demand, reducing unnecessary energy consumption.

By combining predictive analytics with continuous closed-loop control, AI-driven software transforms energy optimization from a manual, reactive process into an automated, proactive system that continuously improves over time.

Selecting the right high-quality data for energy-efficient and high-performing AI models

A critical enabler of this approach is a robust data strategy. AI models require high-resolution data to accurately predict network behavior, particularly at short time intervals. Traditional network telemetry, which is often aggregated at coarse intervals, is insufficient for this purpose. Bridging this gap requires collecting and processing data at a much higher frequency and with greater contextual awareness.

Figure 8: Automated energy saving that maximizes energy performance while maintaining user experience



Architecture: Combining distributed and centralized automation for energy optimization

Achieving large-scale energy efficiency gains requires executing AI in a way that balances performance, complexity, and energy overhead. This is enabled through a hybrid architecture that combines distributed and centralized intelligence [6].

Distributed AI in RAN operates directly at the radio site, where it can make real-time decisions with minimal latency. This is particularly important for energy optimization tasks that require immediate response, such as deactivating resources during brief periods of low traffic.

Centralized AI for RAN complements this by providing network-wide visibility and optimization. It operates on longer time scales, using aggregated data to identify patterns, coordinate actions across sites, and refine predictive models.

AI on RAN enables distributed and offloaded AI processing for devices such as smart glasses, robots, and extended reality applications, and avoids unnecessary data transport and centralized processing, reducing overall energy consumption [7].

In this way, intelligence is embedded across all layers, enabling a closed-loop system for real-time energy optimization and energy-efficient performance across the entire network footprint.

Ericsson's strategy to apply AI in the right layer with the right compute elevates network performance and maximizes service providers' ROI. By placing each model where it performs the best, Ericsson delivers the following benefits:

- optimized compute usage
- reduced unnecessary hardware investments
- maximized performance per watt

By matching workloads to the optimal layer and avoiding a one-size-fits-all deployment model, this approach reduces energy consumption while boosting performance and ensuring that AI itself does not become a source of excess energy consumption.

Portfolio differentiator: Delivering energy-efficient AI RAN with maximum ROI and utmost precision

Ericsson's AI RAN differentiates by positioning AI not as an add-on, but as a native capability applied to elevate the network itself. This distinction is fundamental to energy efficiency.

By embedding intelligence directly into RAN operation, the network continuously optimizes how resources are used, ensuring that energy consumption is dynamically aligned with real-time demand rather than statically provisioned.

Key differentiators include:

Telco-grade AI, energy-efficient inference

Specialized, deterministic, and deployment-optimized AI for the RAN, executed through highly scalable, energy-efficient inference to improve performance and efficiency.

Intent-driven automation

AI automates energy optimization at scale with built-in capabilities to deploy energy-saving strategies without compromising performance and user experience.

Structured data and training foundation

Machine-readable RAN data and continuous learning enable ongoing optimization, adapting to evolving traffic patterns as network conditions change.

No additional hardware footprint

AI is deployed without additional hardware and operates within existing power and compute budgets.

4

Swisscom uses AI-driven RAN automation to improve energy efficiency and maintain user experience

For Swisscom, Switzerland's leading CSP, maintaining network leadership and customer satisfaction remains paramount. At the same time, the company is committed to operating its network more sustainably and efficiently. Swisscom has set ambitious climate targets and is on track to achieve Net Zero by 2035.

By combining AI-powered energy-saving features with machine learning-based automation, Swisscom is demonstrating that significant energy savings, approximately 6 percent, can be achieved without significantly compromising network performance or user experience.



The challenge

Swisscom has built its reputation on delivering the best network experience in Switzerland, consistently ranking at the top of independent network quality assessments. These results reflect Swisscom’s long-standing ambition to provide customers with superior coverage, performance, and service quality.

However, maintaining this level of excellence comes with challenges. Like many CSPs worldwide, Swisscom faces increasing operational expenditure pressure from rising energy prices and growing network operating costs. As mobile traffic continues to increase and the 5G rollout expands, energy consumption has become a key concern in network operations.

At the same time, modern RAN sites now support multiple radio technologies, frequency layers, and advanced energy-saving features. While these create new optimization levers, they also add complexity. Determining when and where to activate energy-saving functions without impacting the user experience demands continuous analysis of network conditions, traffic patterns, and performance indicators.

The solution

To address this challenge, Swisscom is combining advanced RAN energy-saving features with machine learning-driven automation.

One key component is Ericsson’s LTE AI-powered MIMO Sleep Mode. MIMO radios deliver strong capacity and performance benefits but do not always require full transmission capability during low-traffic periods.

AI-powered MIMO Sleep Mode automatically identifies when lower capacity is sufficient and places parts of the MIMO system into a low-power state. This helps reduce energy consumption while maintaining performance.

The second component is Swisscom’s own machine learning-based automation framework, designed to orchestrate these energy-saving features at scale. Swisscom had already been using Ericsson RAN features, LTE Cell Sleep, and NR Booster Carrier Sleep, at fixed intervals during low-usage night hours. However, daytime traffic patterns vary significantly from cell to cell, making a one-size-fits-all schedule ineffective.

Swisscom’s machine learning models analyze traffic patterns and customer usage behavior to determine the optimal daytime windows to activate energy-saving functions. Instead of using static schedules or manual configuration, these models identify when traffic is low and customer impact is minimal. Based on these predictions, the system automatically activates:

- LTE Cell Sleep functionality
- 5G NR Booster Carrier Sleep functionality

By dynamically adapting daytime activation windows to actual traffic conditions, Swisscom maximizes energy savings while keeping network performance aligned with customer expectations. To further minimise the risk of affecting the customer experience, Swisscom has limited daytime sleeping windows to a maximum of 8 hours, with specific rules in place for high-priority sites. (see figure 9) The result is an intelligent, data-driven approach that balances energy efficiency and service quality in real time.

The result

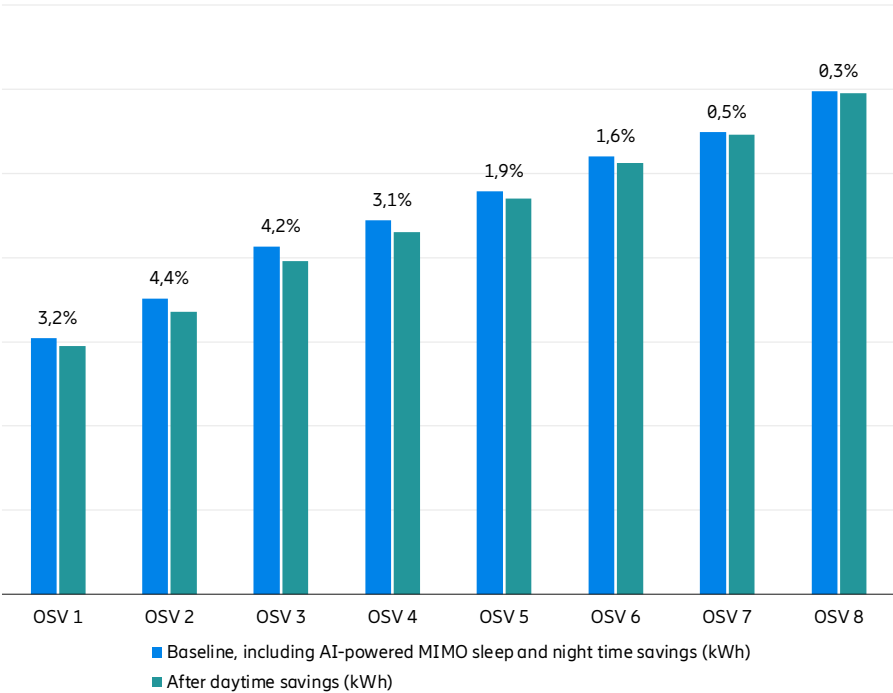
By combining Ericsson’s AI-powered LTE MIMO Sleep Mode with Swisscom’s machine learning-driven RAN automation, Swisscom has built a scalable framework to reduce network energy consumption without compromise in the user experience. AI-powered MIMO Sleep Mode, deployed since early 2023, delivered an estimated 1.7 percent energy savings across RAN in 2025. In addition, LTE Cell Sleep and 5G NR Booster Carrier Sleep provided a further near 4 percent reduction, with the machine learning-based automation layer contributing about 0.4 percent by dynamically extending sleep periods based on real-time traffic conditions and cell importance.

Together, these AI- and ML-driven capabilities enable Swisscom to:

- reduce energy consumption across the RAN by approximately 6 percent through optimized use of advanced energy-saving features
- lower operational complexity via automated decision-making and feature activation
- increase operational efficiency by continuously adapting energy-saving strategies to real traffic patterns

Swisscom’s results demonstrate that energy efficiency and user experience are not competing priorities. By embedding AI and machine learning into RAN operations, energy management becomes a continuous, data-driven optimization process that balances sustainability goals with network performance.

Figure 9: Incremental daytime cell saving with ML automation



The graph compares baseline energy use with daytime cell sleep enabled through Swisscom’s ML automation. Sites are grouped by Operative Site Value (OSV), from OSV 1 (rural, low-traffic) to OSV 8 (dense urban). Daytime cell sleep with ML automation delivers the largest incremental energy savings at low-traffic sites, with benefits gradually decreasing at higher-traffic locations.

5

Building the path to AI-native 6G

The telco industry is entering a new phase where energy efficiency, network performance, and user experience must progress together. As we enter the AI era, mobile network expectations and operation complexity are increasing rapidly, when traditional energy optimization approaches are reaching their limits.

AI and automation fundamentally change how energy is managed. Rather than relying on static configurations or periodic optimization, AI predicts user mobility patterns to optimize both user experience and energy consumption.

However, not all architectures designed to add AI to mobile networks provide the most energy-efficient outcome. The energy impact of AI depends on where intelligence is deployed, how it is executed, and the efficiency of the underlying infrastructure. Optimizing energy efficiency requires a coordinated full-stack approach that combines purpose-built hardware, intent-driven software, and a balanced, intelligent automation architecture. This is where Ericsson's approach stands apart. By applying the right AI model at the right place with the right compute, Ericsson enables CSPs to improve energy performance while minimizing the energy cost of AI itself. Embedded intelligence across all layers ensures that optimization decisions are made where they create the greatest value.

6G outlook: Designing energy efficiency from day one

Building on 5G's lean design, 6G will further reduce always-on signaling, extend sleep opportunities, and enable dynamic resource adaptation. Early 6G design principles, such as lean signaling, separation of concerns, and scalable adaptation, show how significant gains can be achieved by reducing always-on transmissions and enabling deeper sleep states. For example, extending signaling periodicity and optimizing idle-mode behavior can substantially reduce baseline energy consumption.

6G is expected to be AI-native by design and embed energy efficiency from day one, enabling real-time adaptation of resources, traffic, and performance requirements. By proving in 5G that energy consumption can be dynamically aligned with demand, CSPs are building the operational foundation for energy-efficient 6G networks, where capacity can scale without energy use growing at the same pace.

Act now: Accelerate energy-efficient AI RAN evolution with Ericsson

Energy efficiency has long been a CSP priority. AI-powered networks are now taking it to the next level. AI and automation are proven to enable measurable improvements in energy performance, cost efficiency, and user experience across today's networks. Ericsson supports CSPs through a holistic network evolution approach, combining modernization, intelligent operations, and AI-driven automation to break the energy curve and align energy consumption with real-time demand.

By starting with existing networks, modernizing hardware, activating energy-efficient software, and embedding intent-driven automation, CSPs can deliver immediate energy savings while improving performance. This approach applies across all markets, enabling CSPs to scale efficiently, reduce OPEX, and build the foundation for AI-native, energy-efficient 5G Advanced and 6G networks.

Recommendations

Move to AI-driven, intent-based automation, aligning with dynamic traffic patterns.

Use AI and automation to optimize and balance energy efficiency, network performance, operational efficiency, user experience and sustainability targets.

Deploy the right AI model in the right place for the right use cases, optimized for performance, energy efficiency and ROI.

Prepare now for AI-native and energy-efficient 6G networks, accelerating sustainable network growth.

Build an energy-efficient and AI-ready network foundation, with a full stack approach.

References

1. ICT energy evolution: Telecom, data centers, and AI
2. Sustainable networks: The RAN modernization guide
3. GSMA: Energy implications of AI for mobile operators
4. AI in RAN: The inference tsunami is here
5. Sustainable networks: The intelligent operations guide
6. Ericsson Intelligent RAN Automation
7. AI RAN: Smarter connectivity

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