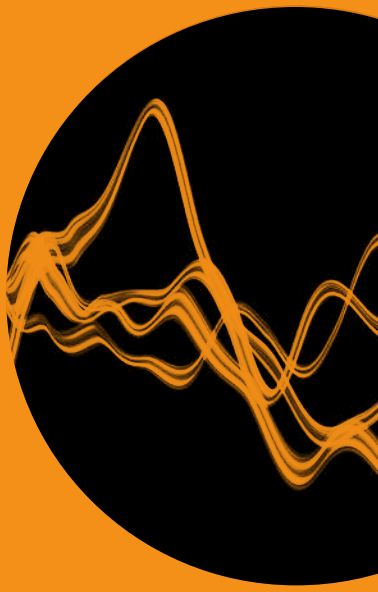
CHARTING THE FUTURE OF INNOVATION | #10·2021

ERICSSON TECHNOLOGY Review



AI-ENABLED RAN AUTOMATION

ERICSSON

AI-enabled RAN automation

Communication service providers need a greater degree of RAN automation to cope with the increasingly advanced RAN. Getting there will require an increased use of artificial intelligence and machine-learning techniques.

DIARMUID CORCORAN,
ERIK WESTERBERG,
HÅKAN OLOFSSON,
MATHIAS SINTORN,
PAUL STJERNHOLM,
PER WILLARS,
STEPHEN TERRILL

A significant and growing portion of communication service providers' (CSPs) opex relates to the manual tuning of algorithms in RANs that do not exploit the full potential of the networks in the field. As 5G and cloud-native RAN implementations continue, the skill level needed to operate the RAN will continue to rise. Our AI-centered approach to RAN automation is designed to overcome both of these challenges.

■ The introduction of 5G has made the RAN more advanced, with many aspects that need to be tuned and coordinated. Not only does NR significantly increase the number of band combinations that have to be managed; it also extends the capability of the network from supporting a single mobile broadband data service to supporting multiple data services (slices) with different characteristics. The Industrial Internet of Things [1] is just one example. Further, a cloud-native RAN implementation is expected to provide a high degree of agility and flexibility through instantiation and scaling of microservices. Manual intervention in the management process becomes impossible at this point. RAN automation is therefore essential to operate a network at this level of complexity.

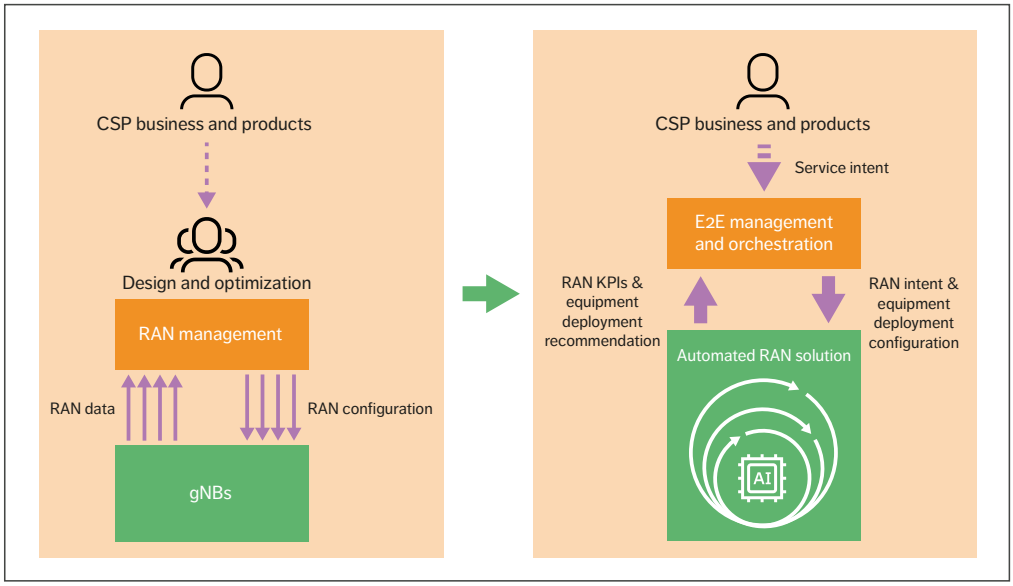


Figure 1 Evolution from manual network operations to automated, intent-based service operations

What is RAN automation?

The objective of RAN automation is to boost RAN performance by replacing the manual work of developing, installing, deploying, managing, optimizing and retiring of RAN functions with automated processes.

RAN automation assists in automating the provisioning and assurance of the RAN part of the consumer and business services that the CSP provides, with the overall objective of maximizing spectrum and energy efficiency. Over time, RAN automation will raise the abstraction level with a machine-and-data-driven approach, where the operator sets goals (also known as intents [2]) for the RAN automation solution instead of configuring detailed parameters of the RAN functions. With intents as input, the RAN automation solution adjusts the resource usage and behavior of the RAN to meet these goals. *Figure 1* visualizes a CSP evolving from manual network operations to

Terms and abbreviations

AI – Artificial Intelligence | **API** – Application Programming Interface | **ARMI** – Automated RAN Management Interface | **ATMI** – Automated Transport Management Interface | **BGP-LS** – Border Gateway Protocol Link-State | **CSP** – Communication Service Provider | **DDD** – Data-Driven Development | **E2E** – End-to-End | **ERAN** – Elastic RAN | **gNB** – gNodeB | **KPI**– Key Performance Indicator | **LCM** – Life-Cycle Management | **ML** – Machine Learning | **MS** – Millisecond | **NG** – Next Generation | **NR** – New Radio | **O-RAN** – O-RAN Alliance | **PCEP** – Path Computation Element Communication Protocol | **RRM** – Radio Resource Management | **SMO** – Service Management and Orchestration | **TM Forum** – TeleManagement Forum

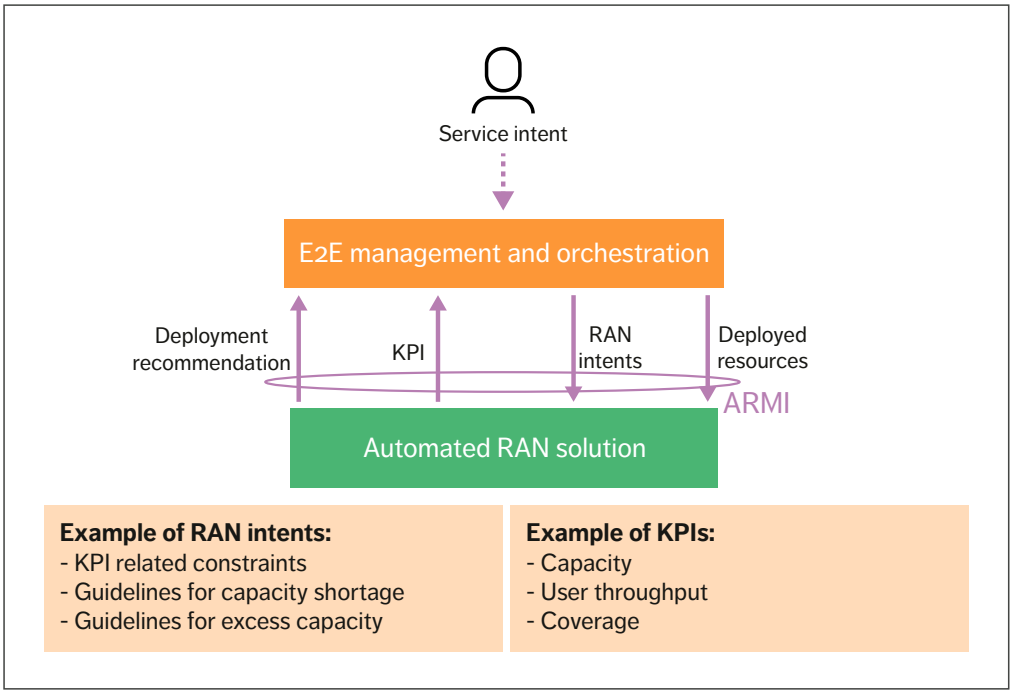


Figure 2 Intent-based management

automated, intent-based service operations with the help of RAN automation.

When automation functionality is added to a RAN solution in the correct way, the abstraction level of the interface between the automated RAN solution and the CSP operations team rises. This approach enables the CSP to define high-level RAN intents as input to the automated RAN solution rather than using detailed configuration parameters of the individual RAN functions. Our approach to RAN automation includes an automated RAN management interface (ARMI) that conveys intents from end-to-end (E2E) management and orchestration to the RAN automation solution, as shown in *Figure 2*.

> OUR APPROACH TO RAN AUTOMATION INCLUDES AN AUTOMATED RAN MANAGEMENT INTERFACE (ARMI)

Intent-based management

The TeleManagement Forum (TM Forum) [3] defines intent as "the formal specification of all expectations including requirements, goals, and constraints given to a technical system." RAN intents are based on overall CSP business intents, (as shown in Figure 1) and priorities with the specific purpose of guiding the RAN automation solution to optimize its behavior given a set of deployed resources (sites, sector carriers, transport capacity, software licenses and so on).

The process of standardizing the language of expressing intents is underway (in the TM Forum [3], for example) but it does not yet contain the expressiveness needed when applying to different application domains. This should be covered by intent extension and intent information models to be specified by other standardization bodies or working groups. For the RAN, it is natural for this to be done by 3GPP SA5 and RAN3 groups, which will ensure that intent extensions allow for coexistence with and evolution of existing interfaces such as the 3GPP slicing interface.

As the RAN intent should guide the RAN automation solution, it is essential that RAN intents define target key performance indicators (KPIs) that are relevant to the RAN, such as user throughput, delay and coverage. The target KPIs should be considered as goals that the RAN automation solution should meet within the possibility of the deployed resources. Each target KPI must be defined in precise detail and based on quantities that the RAN automation solution can measure. This means that both the language for intents as well as the corresponding measurements in the RAN need to be sufficiently standardized. In addition, because of the nature of the RAN, the target KPIs need to be expressed in statistical terms – that is, as a target of a certain percentile of users with a desired consumer experience.

While the target KPIs are required input, they are not sufficient as RAN intents. If the target KPIs are fulfilled by the system and there are still resources available, the system needs additional intents with information about what else it should optimize, such as peak throughput, capacity or energy efficiency. These are rules for how the system will behave in situations when all KPIs are met and there are still free resources in the system (e.g. in periods of low traffic in coverage cells) as well as how to prioritize between KPIs in situations when there are not enough resources to meet all KPIs (in traffic peak situations, for example).

If the system cannot fulfill the target KPIs, it needs a guideline regarding how to prioritize the available resources. Should some services or user groups be prioritized? Should cell edge users be disconnected or deprioritized? Furthermore, in this situation, the RAN automation solution should provide information to the operator about bottlenecks and the need for extra capacity in a given geographic area.

> IT IS ESSENTIAL THAT RAN INTENTS DEFINE TARGET KPIs THAT ARE RELEVANT TO THE RAN, SUCH AS USER THROUGHPUT, DELAY AND COVERAGE

Data-driven development

Through data-driven and continuous software development, the design, deployment and assurance processes can ensure that the functionality is sufficiently adaptive and robust to be used in a variety of environments in the operational networks. New or updated functionality can be brought to market more quickly, enabling rapid response to operator needs. As part of the RAN automation solution, data-driven development (DDD) complements software development, with data driven, machine-learning-based automation. DDD should allow for local adaptations based on the data collected from the field and from digital network twins, which will enable a shift from reactive mitigation to predictive and preventive software management and support and service assurance.

Cross-domain

Beyond the need for E2E cross-domain management and orchestration, it is also important that a RAN automation solution can interact with other domains. Based on the RAN intents received, our RAN automation solution is able to interwork with other network domains through a network automation platform to optimize the RAN performance. For example, it can interwork with the transport domain to request resources for fronthaul.

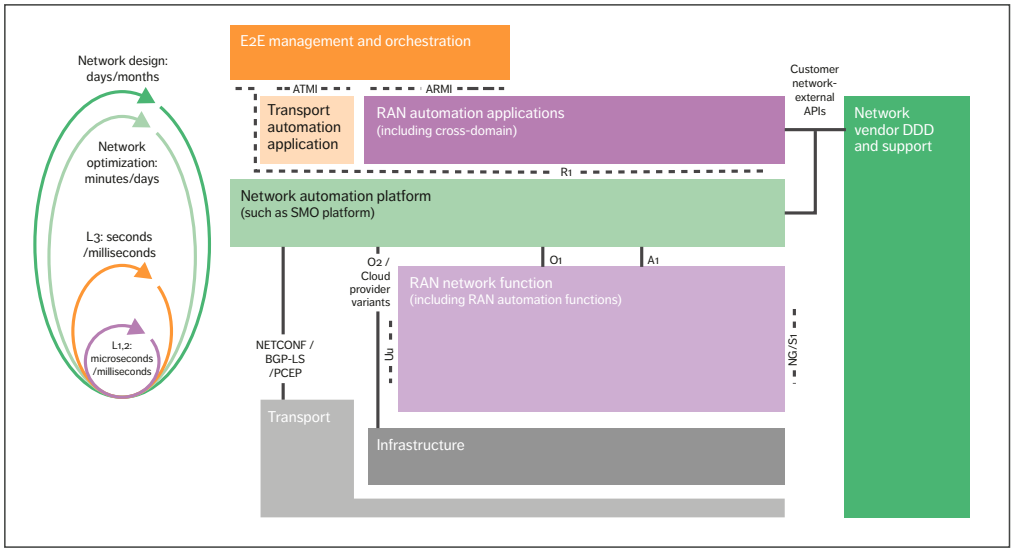


Figure 3 RAN automation architecture

RAN automation architecture

Figure 3 presents a RAN automation architecture based on functional domains and interfaces defined by the O-RAN Alliance (O-RAN) [4], with some proposed additions. The additions include interfaces to a data-driven vendor domain, an interface for the CSP operations team to express the intents (ARMI) and interfaces to other domains (such as transport) with which the RAN automation solution needs to interact. For the architecture to be successful, it will require a long-term, stable industry agreement.

The RAN network function domain (the light-purple box in Figure 3) contains the 3GPP-defined RAN network functions [5] and the Radio Resource Management (RRM) functionality, among others. It uses the standardized and open O1 and A1 interfaces [4] to communicate with the network automation platform domain that is directly above it. The openness of A1 and O1 will allow for third-party automation platform providers. To better support both innovation and openness, ORAN is also standardizing the means of extending the O1, A1 and R1 interfaces to enable a competitive ecosystem and quick time to market of new functionality.

The RAN network function domain provides data collection and distribution services as well as automation support services to higher layers through the R1 application programming interface (API). Examples of such automation services are data management, inventory and topology and services for life-cycle management (LCM) of software in the RAN automation application domain (known as rApps).

The RAN automation applications domain (the dark-purple box in Figure 3) includes some of the intelligence that is used to realize different RAN automation use cases. Consistent with O-RAN terminology, this intelligence is realized with the help of rApps working together with the network automation platform with the objective to optimize the performance of underlying network functions using the R1 interface. The openness of the R1 interface, which provides access to O1, O2 and A1 related services, for example, will allow for the development of rApps from third-party providers. Due to dependencies on RAN features within the network function, closed-loop automation will often work best with rApps from the RAN vendor.

The RAN automation applications domain and network automation platform receive RAN intents from the E2E management and orchestration domain (the orange box in Figure 3) through the ARMI, which guides the actions of the RAN automation functionality.

The bottom of Figure 3 shows the domains that provide resources to the RAN automation solution. For some features – such as Elastic RAN (ERAN) – the RAN automation solution will request resources from the transport domain (light grey) through services exposed by transport automation applications over R1. For a cloud RAN implementation, the infrastructure domain (dark grey) will be essential, as this will provide the compute, storage and local networking resources for the RAN functions on which to execute. When the RAN automation solution requires resources from this domain, it will use the O2 interface.

The right side of Figure 3 illustrates the network vendor's DDD domain (dark green). This domain interacts with the RAN software deployed in the network domain and the RAN automation applications domain by supporting the CI/CD (continuous integration and continuous delivery) flow as well as getting system feedback from live networks into the R&D process. The DDD domain has a data science environment, including AI/ML training infrastructure. This environment enables the design, build, training, testing and deployment of new ML models, used to support the network vendor's product offering.

Fundamental to the architecture but not explicitly shown in the figure is the efficient handling of data within and between the domains through the use of data pipelines [6].

Our RAN automation solution

The left side of Figure 3 illustrates how the task of efficiently operating a RAN to best utilize the deployed resources (base stations or frequencies) can be divided into different control loops acting according to different time scales and with different scopes. A successful RAN automation solution will require the use of AI/ML technologies [6] in all of these control loops to ensure functionality that can work autonomously in different deployments and environments in an optimal way.

The two fastest control loops (purple and orange) are related to traditional RRM. Examples include scheduling and link adaptation in the purple (layer 1 and 2) control loop and bearer management and handover in the orange (layer 3) control loop. Functionality in these control loops has already been autonomous for quite some time, with the decision-making based on internal data for scheduling and handover in a timeframe ranging from milliseconds (ms) to several hundred ms, for example. From an architecture perspective, these control loops are implemented in the RAN network function domain shown in Figure 3.

The slower control loops shown on the left side of Figure 3 represent network design (dark green) and network optimization and assurance (light green). In contrast to the two fast control loops, these slower loops are to a large degree manual at present. Network design covers activities related to the design and deployment of the full RAN, while network automation covers observation and optimization of the deployed functionality. Network optimization and assurance is done by observing the performance of a certain functionality and changing the exposed configuration parameters to alter the behavior of the deployed functionality, so that it assures the intents in the specific environment where it has been deployed. From an architecture perspective, these control loops are implemented in the RAN automation application domain [7].

The green control loops encompass the bulk of the manual work that will disappear as a result of RAN automation, which explains why AI/ML is already being implemented in those loops [8]. It would, however, be a mistake to restrict the RAN automation solution to just the green control loops. AI/ML also makes it possible to enhance the functionality in the purple and orange control loops to make them more adaptive and robust for deployment in different environments. This, in turn, minimizes the amount of configuration optimization that is needed in the light-green control loop.

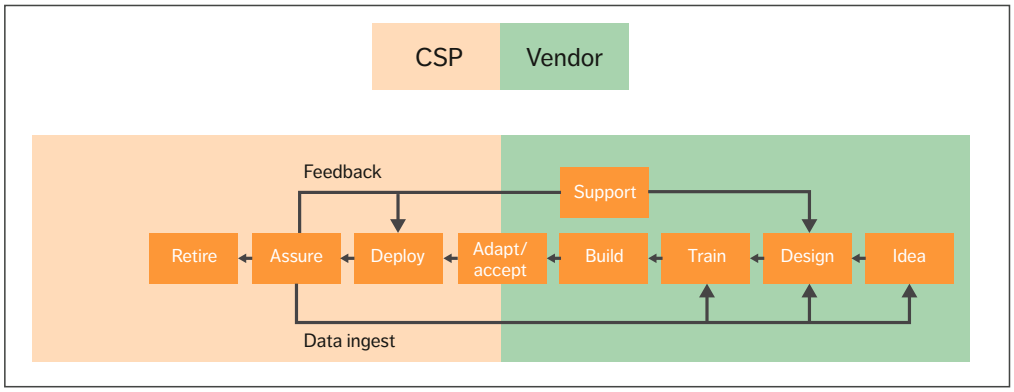


Figure 4 A high-level LCM process

While the control loops in Figure 3 are all internal to the RAN domain, some of the functionality in a robust RAN automation solution will depend on resources from other domains. That functionality would be implemented as part of the RAN automation application domain. The RAN automation platform domain will provide the services required for cross-domain interaction.

One example of RAN automation functionality in the RAN automation application domain is the automated deployment and configuration of ERAN. In ERAN deployments, AI/ML is used to cluster basebands that share radio coverage and therefore should be configured to coordinate functionality such as scheduling [8]. To do this, data from several network functions needs to be clustered to understand which of them share radio coverage. This process requires topology and inventory information that will be made available to the rApps through the services exposed by the network automation platform over R1.

The outcome of the clustering results is a configuration of the basebands that should coordinate as well as a request for resources from the transport domain. This information can also be obtained by services provided by transport automation applications exposing services through the R1 framework. When designing the rApp for clustering, it is beneficial to have detailed knowledge about the implementation of coordination functionality in the RAN network function to understand how the clustering analysis in the rApp should be performed.

An example of RAN automation functionality in the network function domain is AI/ML-based link adaptation, where AI/ML-based functionality optimizes the selection of the modulation and coding scheme for either maximum throughput or minimum delay, removing the block error rate target parameter and thereby the need for configuration-based optimization. Another example is secondary carrier prediction [8], where AI/ML is used to learn coverage relations between different carriers for a certain deployment. Both of these examples use data that is internal to the network function.

Life-cycle management of the RAN automation functionality

As the objective of RAN automation is to replace the manual work of developing, installing, deploying, managing, optimizing and retiring RAN functions, it is certain to have a significant impact on the way that the LCM of RAN software works. Specifically, as AI/ML has proven to be an efficient tool to develop functionality for RAN automation, different options for training and inference of ML models will drive corresponding options for the LCM of software with AI/ML-based functionality.

Figure 4 presents a process view of the LCM of

RAN components, ranging from the initial idea for a RAN component to its eventual retirement. A RAN component is defined as either a pure software entity or a hardware/software (physical network function) entity. As the different steps in the LCM structure include the manual work associated with RAN operations, it is a useful model to describe how RAN automation changes the processes, reduces the manual effort and improves the quality and performance of the RAN.

An important aspect of the LCM is that it represents a structure of responsibility, accountability and ownership among vendors and CSPs. This structure is the baseline for the business model between vendors and CSPs, structuring exactly what is delivered by the vendor in the LCM process.

The light orange and green background colors in Figure 4 highlight the responsibilities of the CSP and the vendor respectively. Software or software/hardware entities are delivered in the adapt/accept step together with support contracts and, in some cases, professional services for integration and deployment.

Using AI/ML models in the RAN automation solution requires the introduction of a model training step to the LCM process. There are four main alternatives for how to add model training to the LCM, each with implications on the responsibility split between the vendor and the CSP.

The first alternative is for the vendor to deliver a global model (that is, the same model for all CSPs) in the form of software entities in the adapt/accept step. A global model can, for some use cases, still allow for consideration of local context and can be very powerful in creating highly flexible automation functionality that can adapt to different deployments. In this case, all training is the responsibility of the vendor and occurs in the train step.

The second alternative is for the vendor to deliver local models in the form of software entities tailored for different uses (CSP-specific or geo-specific, for example) in the adapt/accept step. Local training is the responsibility of the vendor and occurs in the train step. This full model training alternative

> “USING AI/ML MODELS IN THE RAN AUTOMATION SOLUTION REQUIRES THE INTRODUCTION OF A MODEL TRAINING STEP”

requires access to local data, and it is important to be aware that the cost of maintaining different software versions could become substantial. As a result, this alternative is most appropriate for scenarios with centralized inference in a few places per CSP where there is only one or just a few ML models that do not require frequent retraining. In scenarios with distributed inference in thousands of places per CSP that require retraining every other week (for example), this model training would not be the best alternative.

The third alternative is for the vendor to deliver a global model that can be retrained on additional data sets. In the adapt/accept step, the vendor delivers the model in the form of software entities together with information about how to retrain and evaluate it. The CSP is responsible for retraining the model to become a set of local models, which expands the adapt/accept step to include training. In these scenarios, it is unclear how much responsibility the vendor can take for in-field performance and support. Therefore this is not recommended as a direction commercial deployment until responsibilities have been resolved.

The fourth alternative is for the vendor to deliver a base-trained model in the form of software that is designed to be automatically retrained on local data after deployment. We refer to this as embedded training, and the training is transparent to the CSP. In this case, the training is the responsibility of the vendor and occurs both in the train step and autonomously in the deployed software. This is a path toward a fully autonomous system, while keeping the current business relation between vendor and CSP intact.

A cloud RAN implementation will impose

additional changes to the LCM process that go beyond those introduced by AI/ML. A cloud-native, microservice-based architecture will enable the possibility to very dynamically deploy and instantiate functionality in the form of microservices, based on local and temporal changes in the network, such as load. In a network with moving load, this capability should also extend to instantiating/scaling microservices in different parts of the network as load moves around. Because of the dynamics of the changes, these processes need to be automated, meaning that parts of the manual deployment step are automated and governed by functionality provided by the vendor.

As the trend of virtualization and orchestration evolves, it is probable that nearly all deployment, scaling, canary testing and instantiation will happen automatically and highly dynamically. At that point, the CSPs' responsibility will move from the manual deployment of software to monitoring how well the RAN automation solution fulfills the RAN intents.

Conclusion

The near-endless possibilities of 5G RAN and the rising popularity of cloud-native RAN implementations have led to an increasingly urgent need to reduce the manual work involved in developing, installing, deploying, managing, optimizing and retiring RAN functions. To cope with the more and more advanced system, RAN operations and management need to become data driven and automated.

Based on open standards, Ericsson's approach to RAN automation leverages artificial intelligence/machine learning (AI/ML) techniques and the natural dependencies between the functionality in different domains to create an automated RAN solution that is more autonomous and robust for deployment in different environments.

Further reading

- **Ericsson, AI-powered radio access networks, available at:** *https://www.ericsson.com/en/ai/ran*
- **Ericsson Technology Review, Spotlight on the Internet of Things, October 15, 2019, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/spotlight-on-the-internet-of-things*

References

1. **Ericsson Technology Review, Boosting smart manufacturing with 5G wireless connectivity, February 20, 2019, Sachs, J; Wallstedt, K; Alriksson, F; Eneroth, G;** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/boosting-smart-manufacturing-with-5g-wireless-connectivity*
2. **Ericsson Technology Review, Cognitive processes for adaptive intent-based networking, November 11, 2020, Niemöller, J; Mokrushin, L; Mohalik, S.K; Vlachou-Konchylaki, M; Sarmonikas, G, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/adaptive-intent-based-networking*
3. **TM Forum, IG1253A Intent Modeling v1.0.0, available at:** *https://www.tmforum.org/resources/how-to-guide/ig1253a-intent-modeling-v1-0-0/*
4. **O-RAN, Specifications, available at:** *https://www.o-ran.org/specifications*
5. **3GPP, RAN specifications, available at:** *https://www.3gpp.org/specifications-groups/ran-plenary/ran3-iu,-iub,-iur,-s1,-x2-and-utran-e-utran*
6. **Ericsson Technology Review, Data ingestion architecture for telecom applications, March 16, 2021, Rönnberg, AK; Åström, B; Gecer, B, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/data-ingestion-architecture-for-telecom*
7. **Ericsson Technology Review, Artificial intelligence in RAN – a software framework for AI-driven RAN automation, December 8, 2020, Corcoran, D; Ermedahl, A; Granbom, C, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/artificial-intelligence-in-ran*
8. **Ericsson Technology Review, Enhancing RAN performance with AI, January 20, 2020, Calabrese, F.D; Frank, P; Ghadimi, E; Challita, U; Soldati P, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/enhancing-ran-performance-with-ai*

THE AUTHORS

Diarmuid Corcoran

◆ joined Ericsson in 1992. He currently works within Business Area Networks as an expert in software architecture, actively driving and participating in software-related activities across the company. Corcoran holds a B.Eng. in computer engineering from the University of Limerick, Ireland.

Erik Westerberg

◆ is a senior expert in system and network architecture who is responsible for the long-term evolution of the Ericsson RAN architecture. Joining Ericsson in 1996 from Massachusetts Institute of Technology, USA. Westerberg has 25 years of experience from 2G, 3G, 4G and 5G mobile systems, where he holds more than 50 patents. Westerberg also holds a Ph.D. in physics from Stockholm University, Sweden.

Håkan Olofsson

◆ has worked in the mobile industry for 28 years, with a particular focus on RAN. After joining Ericsson in 1994, Olofsson served in several capacities, mostly dealing with strategic technology development and the evolution from 2G to 5G. He is currently head of the System Concept program at Development Unit Networks, focusing on innovative RAN solutions for 5G and 6G. Olofsson holds an M.Sc. in physics engineering from Uppsala University, Sweden.

Mathias Sintorn

◆ is an expert in traffic handling and service performance within Business Area Networks. He joined Ericsson in 1998. In his current role, he defines the long-term evolution of the RAN architecture, specifically in the area of RAN automation. Sintorn holds an M.Sc. in engineering physics from Uppsala University.

Paul Stjernholm

◆ joined Ericsson in 1995. His current work focuses on RAN management strategies and standardization with a recent interest for RAN automation. Stjernholm holds a M.Sc. in applied physics and electrical engineering from Linköping University, Sweden.

Per Willars

◆ is a senior expert in radio network functionality and E2E architecture within Business Area Networks. He joined Ericsson in 1991 and has since worked intensively with 3G, 4G and 5G RAN topics, as well as the interaction between RAN and the core network and service layers. In his current role, he defines the long-term evolution of the RAN architecture. Willars holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm.

Stephen Terrill

◆ is a senior expert and chief architect in automation and management. In recent years, his work has focused on the automation and evolution of operations support systems, and he has been engaged in open source on the Technical Steering Committee of ONAP (Open Network Automation Platform) and as ONAP architecture chair. Terrill holds an M.EngSc. from the University of Melbourne, Australia.

ISSN 0014-0171
284 23- 3363 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000