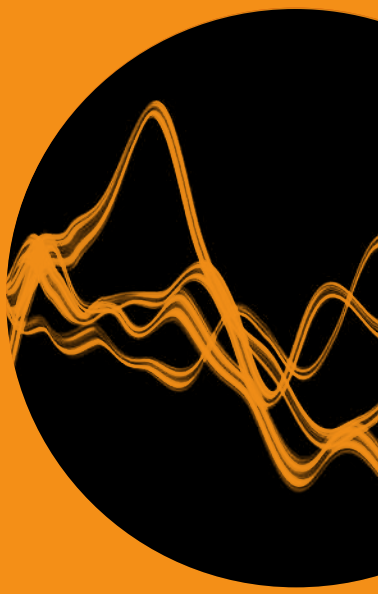
CHARTING THE FUTURE OF INNOVATION | VOLUME 105 I 2022-01

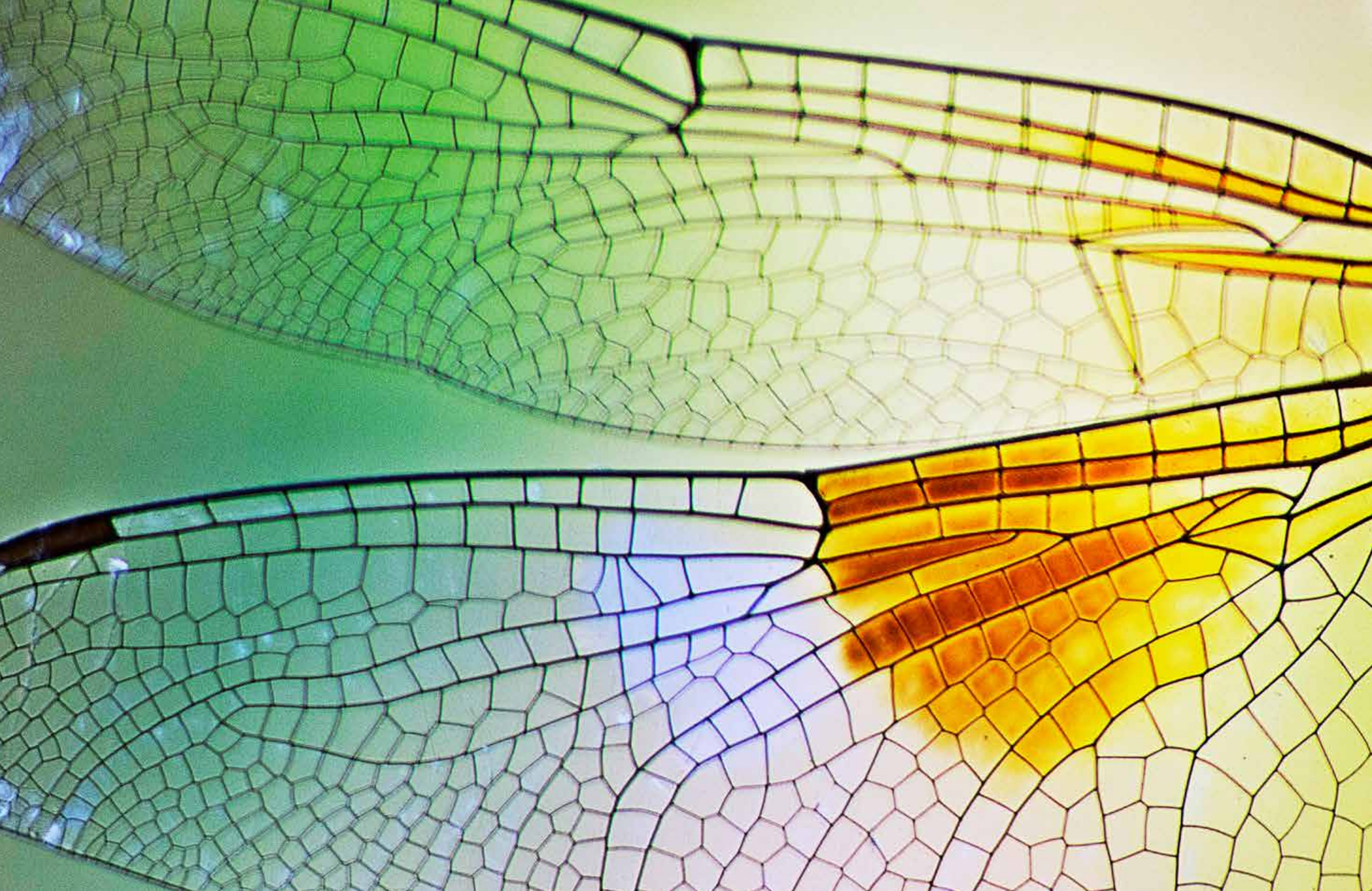
Ericsson Technology Review



BUILDING ROBUST
CRITICAL NETWORKS
WITH THE 5G SYSTEM

5G EVOLUTION
TOWARD
5G ADVANCED

END-TO-END
NETWORK SLICING
ORCHESTRATION



Ericsson Technology Review brings you insights into some of the key emerging innovations that are shaping the future of ICT. Our aim is to encourage an open discussion about the potential, practicalities and benefits of a wide range of technical developments, and provide insight into what the future has to offer.

ADDRESS
Ericsson
SE -164 83 Stockholm, Sweden
Phone: +46 8 719 0000

PUBLISHING
All material and articles are published on the Ericsson Technology Review website:
www.ericsson.com/ericsson-technology-review

PUBLISHER
Erik Ekudden

EDITOR
Tanis Bestland (Nordic Morning)

EDITORIAL BOARD
Hans Bergström, Magnus Buhrgard, Torbjörn Cagenius, Magnus Ewerbring, Dan Fahrman, John Fornehed, Kjell Gustafsson, Jonas Högberg, Sara Kullman, Johan Lundsjö, Cecilia Nyström, Håkan Olofsson, Patrik Roseen and Robert Skog

ART DIRECTOR
Carola Pilarz (Nordic Morning)

PROJECT MANAGER
Susanna O'Grady (Nordic Morning)

LAYOUT
Carola Pilarz (Nordic Morning)

ILLUSTRATIONS
Jenny Andersén (Nordic Morning)

SUBEDITORS
Ian Nicholson (Nordic Morning)
Paul Eade (Nordic Morning)

ISSN: 0014-0171
Volume: 106, 2022

LEVERAGING **5G INNOVATIONS** FOR A BRIGHTER TOMORROW

■ **RESILENT,** high-performance 5G networks will play an essential role in boosting economic growth, creating a host of new opportunities and improving market competitiveness in the years ahead. Achieving the high degree of interoperability needed to deliver the full financial and technological potential of this powerful innovation platform requires that stakeholders across industries, tech ecosystems and governments work more closely together than ever before.

This issue of our magazine includes seven articles that I believe will be of interest to anyone who wants to gain greater insight into the role that 5G networks play in the accelerating digital transformation. For example, readers who want to know about 5G service automation should go directly to page 44, where our experts explain the critical role of service exposure and automated life-cycle management in scenarios that involve private and public clouds provided by hyperscale cloud providers (HCPs). The authors also explain how CSPs can create offerings for enterprises and application developers using service management APIs, network services and assets provided by HCPs.

On page 78, readers will find a fascinating article about an end-to-end (E2E) slice orchestrator that uses transport awareness to guarantee a wide range of QoS levels, including the very high ones that are necessary to meet the most stringent requirements. By automatically matching the particular service requirements of an industry-vertical use case to its specific deployment areas, the transport-aware network slicing orchestration solution they have created ensures E2E QoS without over-provisioning.

Those who want to learn more about network robustness should turn to page 20, where our experts on the topic present what they call the '5G System robustness toolbox', which consists of both standardized and vendor-specific network features and mechanisms. The great thing about this toolbox is that network operators have full flexibility to activate the features and mechanisms that are most appropriate for each particular use case and/or deployment variant and can even choose different mechanisms for different user equipment within a single network.

In the next-level RAN automation article on page 08, the authors present Ericsson's data-driven approach to RAN automation, which is grounded in our comprehensive understanding of the internal workings of RAN network functions. By leveraging the natural dependencies between functionality in different domains and using artificial intelligence to solve complex automation tasks, our researchers have succeeded in creating highly autonomous RAN network functions that can be deployed in a wide variety of environments.

Readers who want to know more about what's coming in 3GPP releases 17 and 18 should definitely check out the article on page 56, which highlights the most notable enhancements and new features. The final two articles in this issue explore the benefits of Massive MIMO (page 68) and the potential risks of quantum technology (page 32) – I highly recommend them both.

We hope you enjoy this issue of our magazine. Feel free to spread the word by sharing it with your colleagues and business partners. You can find both PDF and HTML versions of all the articles at:
www.ericsson.com/ericsson-technology-review

> "STAKEHOLDERS ACROSS INDUSTRIES, TECH ECOSYSTEMS AND GOVERNMENTS WORK MORE CLOSELY TOGETHER THAN EVER BEFORE"

ERIK EKUDDEN
CTO 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.

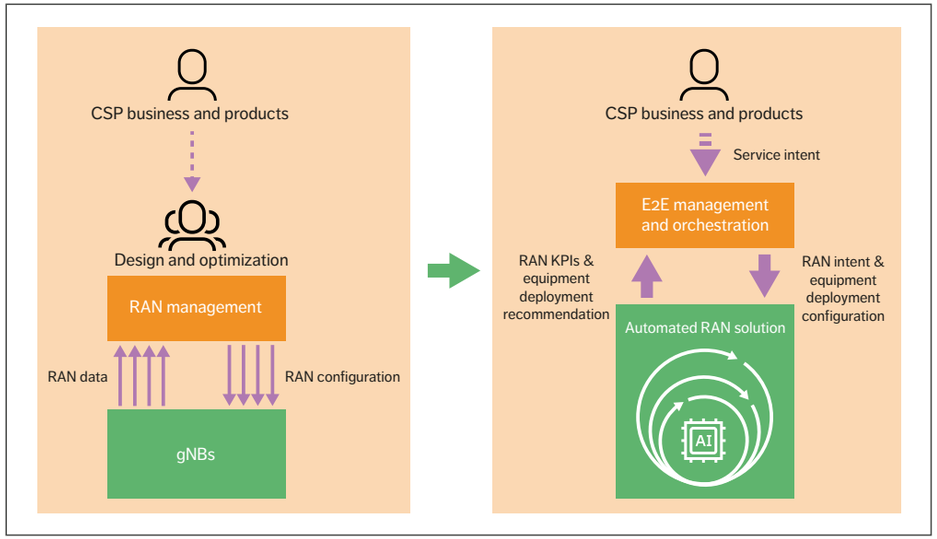


Figure 1 Evolution from manual network operations to automated, intent-based service operations

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

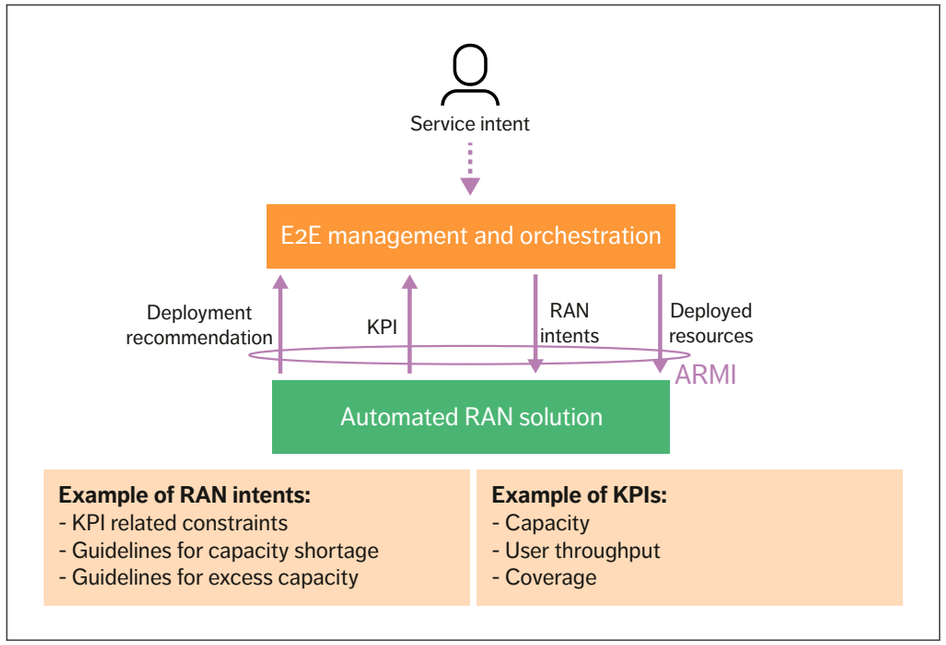


Figure 2 Intent-based management

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.

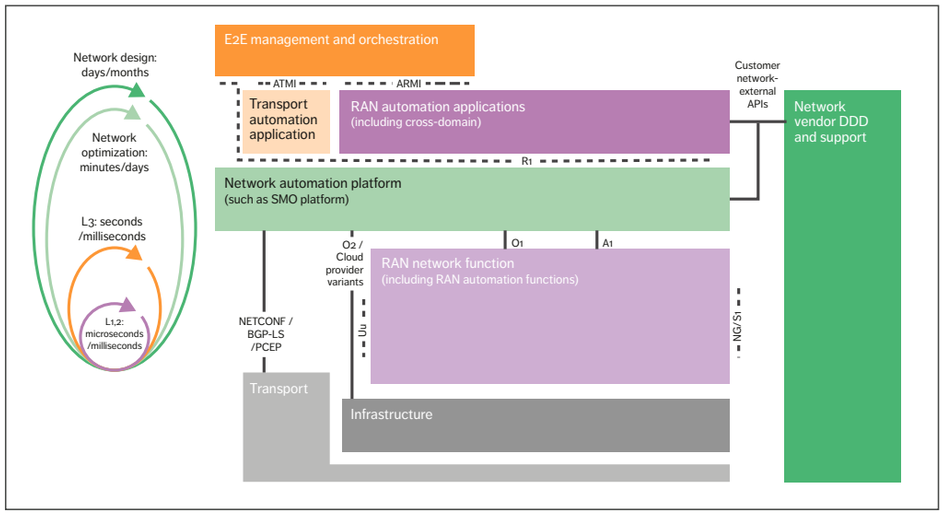


Figure 3 RAN automation architecture

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.

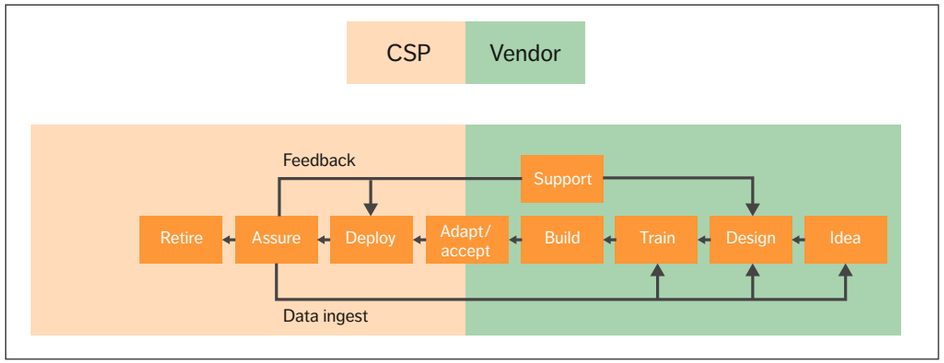


Figure 4 A high-level LCM process

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.

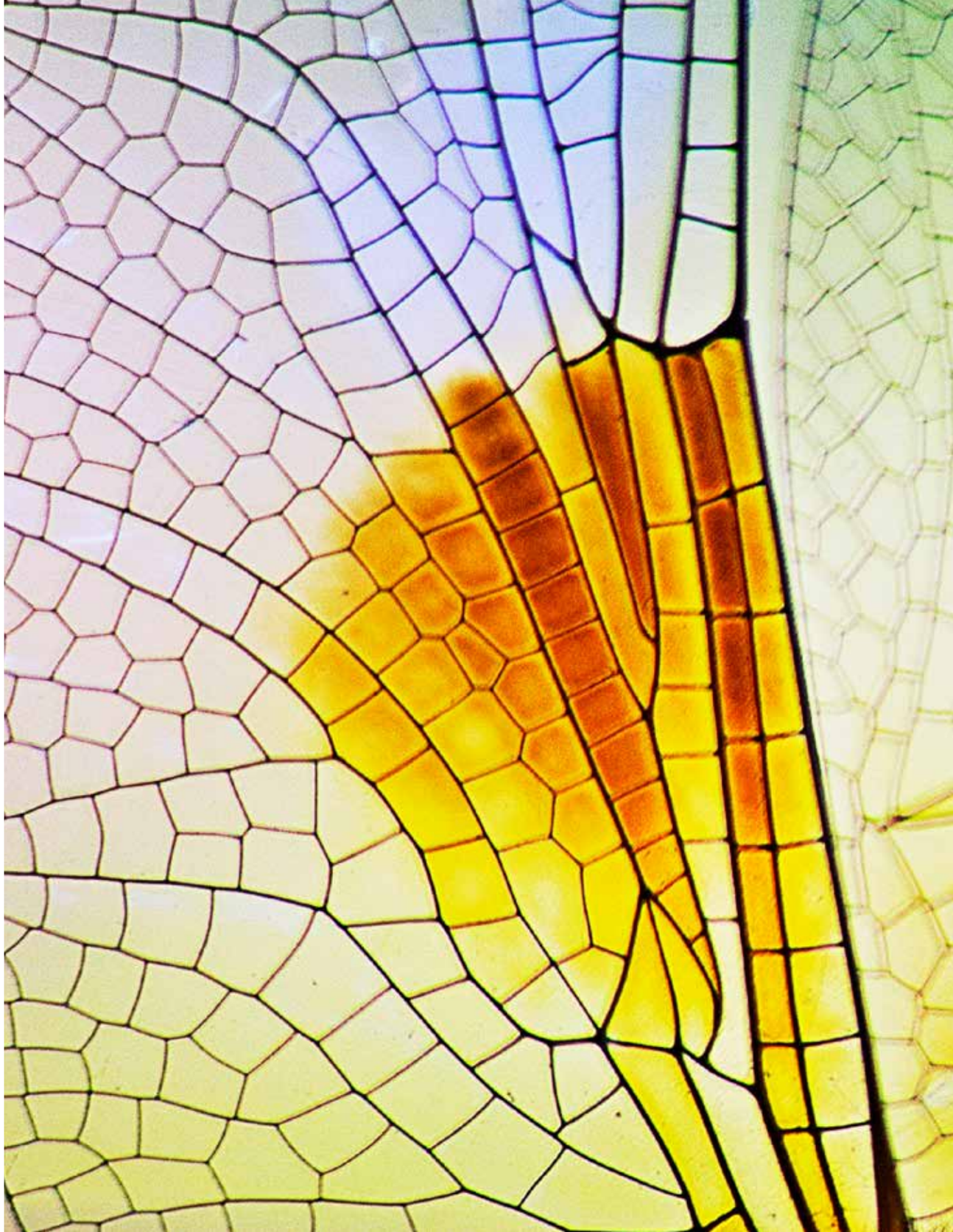
Håkan Olofsson
◆ has worked in the mobile industry for 28 years, with a particular focus on RAN. After joining Ericsson in 1994, Olofsson served in several capacities, mostly dealing with strategic technology development and the evolution from 2G to 5G. He is currently head of the System Concept program 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.



ROBUSTNESS EVOLUTION:

Building robust critical networks with the 5G System

Mobile broadband has become a society-critical service in recent years, with enterprises, governments and private citizens alike relying on its availability, reliability and resilience around the clock. Living up to continuously rising expectations while simultaneously evolving networks to meet the requirements of emerging use cases beyond MBB will require the ability to deliver increasingly higher levels of network robustness.

JARI VIKBERG, GÖRAN HALL, TORBJÖRN CAGENIUS, RICHARD WANG, JOHAN SCHULTZ

The concept of network robustness – a combination of reliability, availability and resilience – is a longstanding cornerstone in the design and development of mobile networks. Among other benefits, network robustness ensures a high level of performance for mobile broadband (MBB), including voice service.

■ As user dependence on apps and mobile services increases, the need for robust networks continues to grow and expand into new areas. Recent examples include the replacement of fixed residential subscriptions for voice and emergency calls with mobile subscriptions, and the increased dependency on smartphone apps for everything including community service, public health care, instant news updates, electronic airplane tickets, mobile banking and payments for both consumers and enterprises.

The 5G System (5GS) has been designed to provide the robustness required to support the growth of conventional MBB services, while also offering network support to new business segments and use cases with more advanced requirements in terms of reliability, availability and resilience. Consisting of the 5G Core (5GC), the Next-Generation RAN (NG-RAN) and the user equipment (UE), the 5GS delivers new capabilities that enable enterprises with business-critical use cases in segments such as manufacturing, ports and automotive [1, 2] to take a major step forward in their digitalization journeys by replacing older means of communication with the 5GS. These new capabilities are also beneficial for mission-critical networks like national security and public safety deployments that are currently being modernized.

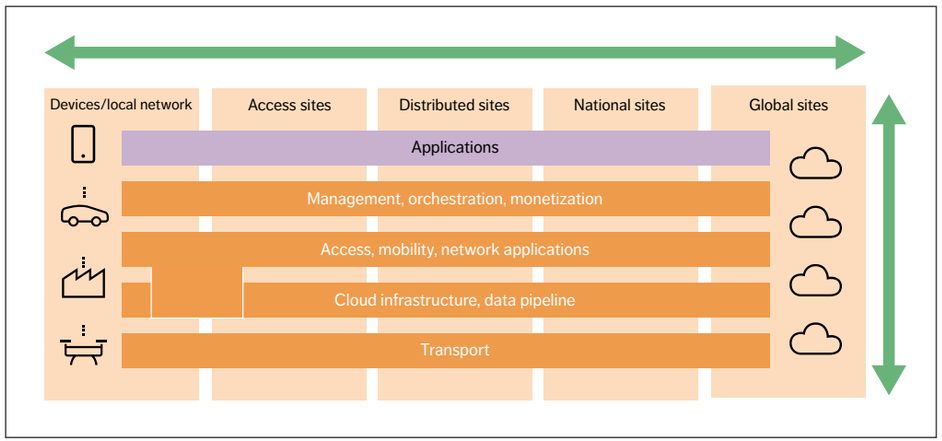


Figure 1 – Aspects impacting network robustness in a typical network

Definition of a robust network

A robust network is a network that delivers the required levels of network availability, reliability and resilience. Network availability refers to a network's ability to accept new traffic. Network reliability refers to a network's ability to support its traffic according to the established use-case-specific requirements – for example, its ability to provide the required use-case-specific QoS for the duration of communication. Network resilience is the ability to provide and maintain an acceptable service level in case of faults, disruptions and other events affecting normal system operation.

The 5GS includes an extensive toolbox of mechanisms and features that can be used in the network design and deployment processes to enhance network robustness.

Aspects impacting network robustness

Figure 1 illustrates the wide range of aspects that impact network robustness, both in the horizontal end-to-end (E2E) and vertical top-to-bottom dimensions, as highlighted by the green arrows.

The functional architecture in the horizontal dimension is the primary focus of this article. It includes UEs and devices, RAN control plane (CP) and user plane (UP), packet core CP and UP, different communication service provider (CSP) network sites including fronthaul and backhaul transport nodes and links/networks between these sites, connectivity to external networks and services, and the actual placement of the application servers. The vertical dimension includes (cloud) infrastructure, automation and orchestration, where in particular the interplay between network function (NF) applications and the infrastructure is important for robust networks. Good security mechanisms are also a prerequisite for robust networks.

The mobile industry continues to measure availability – also known as In-Service Performance

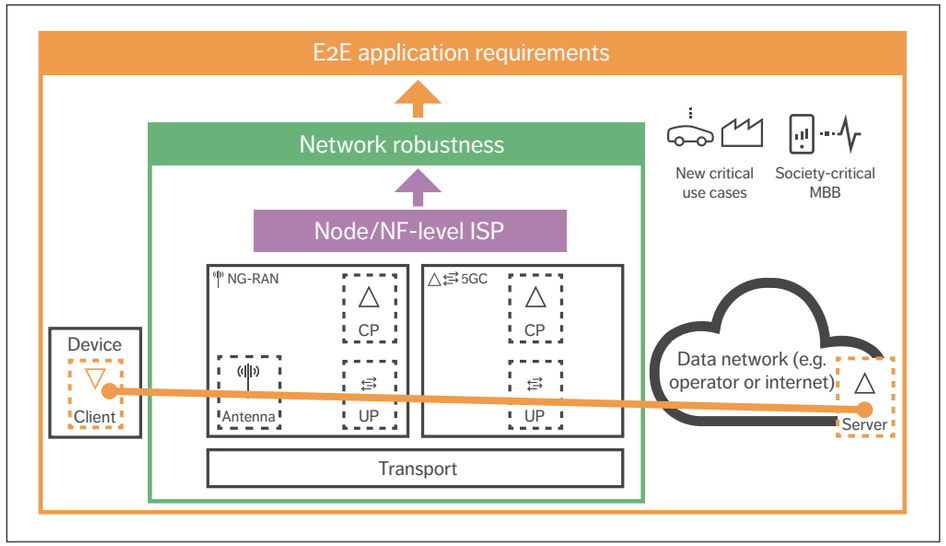


Figure 2 Shifting focus from node/NF-level to network robustness for demanding E2E applications

(ISP) – at individual node and network function (NF) level, which is represented by the purple box in *Figure 2*. However, because the limits for network service characteristics are set by the weakest link in the E2E chain, network robustness (shown in green) requires a broader approach that considers all parts of the network, in order to handle both sunny-day scenarios and different disaster/failure cases.

The large orange-framed section of Figure 2 represents both new critical use cases and society-critical MBB with tougher requirements. The orange line between the application client and the server highlights the significance of that connection in E2E applications. The requirements of new use cases are very specific to their particular characteristics. One of the most challenging reliability-related requirements is survival time as found for time-critical services and defined in 3GPP TS 22.261 [3] as "the time that an application consuming a communication service may continue without an anticipated message."

Applications tend to have different requirements on reliability and (upper) bounded latency. Survival time provides an additional safety margin by, for example, allowing the loss of a very small number of application messages, as long as the survival time limit is not exceeded [1]. In the reliable network context, the important question to answer is how much interruption time is allowed on the UP in different failure cases. The known survival time requirements vary from zero to tens or hundreds of milliseconds, all the way up to seconds.

Improving the overall observability of the network is key to improving network reliability and resilience. Performance management counters within the network are currently used to provide performance visibility to domain management systems in the network. New ways of enabling observability will be needed to monitor the network service characteristics and fulfillment of the new application requirements. Increasing automation in the correction of network failures, predictions and launch of preventive actions must also be considered. In addition, observability related to network failures is an important area to cover. One example of this would be determining actual network performance in case of failures in different parts of the network and assessing the impact on E2E services.

Robustness mechanisms in existing networks

4G networks were primarily specified to deliver MBB and voice services. Considering the large number of consumers that would be impacted by the failure of an Evolved Packet Core (EPC) node (such as the Mobility Management Entity (MME) and packet data network gateway (PGW)), the requirements in 4G extended beyond availability and reliability to also include resilience, which is especially important for voice service continuity.

Evolved Packet Core aspects

The EPC is designed for millions of simultaneously attached users (SAU) and packet data network connections primarily through keeping the control plane (geo)redundant. The purpose is to support continuous EPC services through providing site redundancy and to avoid signaling storms at failures that otherwise can propagate failures from the failing function to other functions. The UP has seldom been redundant in early EPC deployments, but the focus on UP redundancy in deployments is increasing.

The overall requirement on EPC function availability is to have an ISP of 99.999 percent. In other words, the unplanned out-of-service time for each function respectively cannot exceed approximately five minutes per year. The EPC functions have several internal mechanisms to support the ISP requirement, such as failover between different software components and various restart mechanisms from individual connection level to node level. "Fail fast, recover fast" is the governing principle for smaller entities that affect just one or a small number of users, with stepwise escalation to larger (internal) entities if needed.

Ericsson has developed mechanisms beyond the EPC standard for better resilience and redundancy. One example is the georedundancy between MMEs extending the standardized MME pool mechanism. Within the MME pool, one MME has a backup of parts of each UE context stored on another MME in the pool, making it possible to let another MME in the pool take over the UE without a need for reattach.

The PGW/serving gateway also supports several Ericsson-developed mechanisms for redundancy, such as georedundancy solutions for the CP and the UP, both separate and combined. The implementation of Control and User Plane Separation (CUPS) led to the addition of a few more georedundancy solutions for the UP.

While several CSPs view georedundancy deployments as essential – in particular to protect sensitive Access Point Names (APNs) – many of the CSPs with PGWs configured to handle both MBB and VoLTE APNs decided that the hardware costs of georedundancy were too high due to EPC nodes that included both the CP and the UP. However, as a result of the separation of the CP and UP in CUPS (as well as in 5GC) and the increase in subscribers using VoLTE (which requires high reliability), there has been a significant increase in interest in georedundancy for both the CP and VoLTE UP. CSPs may decide to leave the MBB UP without georedundancy for cost-efficiency reasons.

In vendor-specific implementations such as Ericsson's, the policy and charging rules function (PCRF) typically provides a georedundant solution with two PCRFs in either active-active or active-standby deployment. Both the PCRFs in the redundancy solution are connected through a replication channel responsible for the synchronization of the data between the elements.

User Data Convergence includes Home Subscriber Server (HSS) front-ends (FEs) and database back-ends (BEs). The HSS FEs are typically deployed with redundancy, where several FEs can share the load of a failing FE, while the Centralized User Database BEs are typically deployed as georedundant clusters of 1+1 or 1+1+1.

The EPC supports load and overload control in the form of protocol-specific mechanisms in non-access stratum (NAS) congestion control, GPRS

> THE 5G SYSTEM INCLUDES A FLEXIBLE TOOLBOX OF NETWORK FEATURES AND MECHANISMS

Tunneling Protocol Control (GTP-C), Diameter and Packet Flow Control Protocol (PFCP) as well as NF-specific overload-protection mechanisms. The overload-control mechanisms to detect overload and protect the EPC network are largely concentrated to the MME.

LTE RAN aspects

The radio interface in the LTE standard is designed for robustness in aspects such as interference handling, link adaptation, fading/blocking and low-density modulation. Access control and barring solutions exist to protect the network, and to enable high-priority users to access the network in certain situations.

The standard has some inherent Single Points of Failure (SPOFs). For example, in the area of UE-RAN CP, one SPOF is the whole UE on Radio Resource Control (RRC) level. Losing the UE-RAN CP connection leads to a restart of the UP. Another SPOF is the primary cell (pCell) for the UE. Losing the pCell leads to radio link failure, even if additional cells are available.

The LTE RAN is a collection of purpose-built products that perform the required functions, with baseband and radio unit products being the most important. The hardware of the products typically supports telco-grade quality, meaning that it has very high availability, even in the harsh environments where antenna sites are placed around the globe.

Since each product serves only one or a few cells, the effect of one node failing and then restarting was deemed acceptable for MBB services, due to the limited number of affected users, the fast restart mechanism and the very high likelihood of restoring the product. Overall consumer acceptance of short outages is also an important factor. As a result of these factors, the approach to MBB in LTE has been "fail fast, recover fast" at a box level (baseband unit or radio unit level, for example).

When the cost and complexity of designing a more elaborate scheme to increase availability is weighed against the relatively small effect of a failing unit and the temporary loss of a few cells at the most, there tends to be limited interest in increasing availability for MBB. The goal of having 99.999 percent ISP or better availability on the individual products is still considered sufficient.

Ideally, the UE has overlapping coverage from more than one antenna point (overlapping cells or frequency layers, for example) and a failure of the equipment handling one of these antenna points is not catastrophic, as in the worst case it leads to the UE reselecting to a working antenna point.

The 5G System robustness toolbox

The 5GS includes a flexible toolbox of network features and mechanisms that make it easier for CSPs to meet growing requirements on robustness. Some of the tools are standardized, while others are vendor specific. Decisions about which robustness features and mechanisms to use in a specific deployment should be based on the use case(s) it is designed to support. Careful consideration of network design and deployment aspects is essential to the creation of robust networks.

Beyond offering the flexibility of using different tools for different deployments, the 5GS robustness toolbox will also give CSPs the flexibility to activate different tools for different UEs in the same network. The 3GPP standards for the 5GS also include support for ultra-reliable low-latency communication (URLLC), which is essential for use cases that require connectivity with both high reliability and bounded latency.

5G Core aspects

The 5GC has been designed to support millions of SAU and Protocol Data Unit (PDU) sessions for MBB and voice services, while also being scalable for small deployments such as enterprise use cases.

5GC NFs also have the internal mechanisms to tolerate failover at software-component level. Cloud-native implementations of 5GC make it easier than ever to support "fail fast, recover fast" and ensure internal resilience between software components. At network level, the 5GC focuses on standard session resilience support instead of vendor-specific georedundancy solutions. The general ISP requirement on NF availability for MBB service is also 99.999 percent as for EPC, but the requirement for 5G-critical services (such as industrial manufacturing) is even higher, up to zero tolerance of failure interruption.

The 3GPP introduces the generic NF set concept for 5GC control plane NFs to support E2E session resilience at network level, which is not defined in the EPC standard. With the NF set concept, the NF can be deployed so that several NF instances are part of an NF set to provide (geo)redundancy and scalability together. In an NF set, the equivalent NFs share the same context data, which allows an NF instance to be replaced by an alternative NF instance within the same NF set in a failure scenario.

Even though the NF set is a generic mechanism, it does not necessarily apply to all 5GC NFs. For example, the user data repository (UDR) with its internal database has been implemented with resilience based on the georedundant cluster solution derived from the EPC before the NF set was introduced by the 3GPP standard. As a result, the NFs surrounding the UDR already support UDR failure reselection in the UDR georedundant cluster based on product implementation.

5GC supports load (re-)balancing, overload control and NAS-level congestion control to ensure that the NFs are operating under nominal capacity for providing connectivity and necessary services to the UEs. In the 5GC, load and overload control over a service-based interface are the generic mechanisms for all 5GC control plane NFs. Both the Access and Mobility Management Function (AMF) and the Session Management Function (SMF) are in focus of the overload detection and protection for the 5GC network, as both have the protocols to control UE and RAN access.

There is no standardized session resilience solution for the 5GC UP function (UPF). For less critical services (such as MBB), the 5GC CP can recover UP traffic through a restoration procedure after detecting the UPF failure. However, restoring all the traffic takes time and depends on the number of UEs affected. For critical services, a vendor-specific resilience solution is usually required to maintain UP traffic when UPF failover happens. Ericsson has developed mechanisms for both session resilience and georedundancy deployment for the UPF.

Features and mechanisms at the NG-RAN level

On top of the challenging requirements from new use cases, new requirements from RAN centralization and cloud-native evolution also necessitate new network robustness mechanisms and features. As RAN centralization leads to a higher number of UEs being served by a unit that may fail, the "fail fast, recover fast" principle will apply to smaller modules than box level, as in LTE.

The NG-RAN standards still contain similar SPOFs as LTE for the whole UE on RRC-level and pCell for the UE. A new SPOF is also introduced: the UE-RAN UP on PDU session level. In addition, there are functions to support bounded latency and higher reliability, as well as unified access control.

The available robustness features and mechanisms will include a combination of both standardized and vendor-specific functionality. The current understanding is that vendor implementations can solve the above SPOFs, at least

> ERICSSON HAS DEVELOPED MECHANISMS FOR BOTH SESSION RESILIENCE AND GEOREDUNDANCY DEPLOYMENT FOR THE UPF

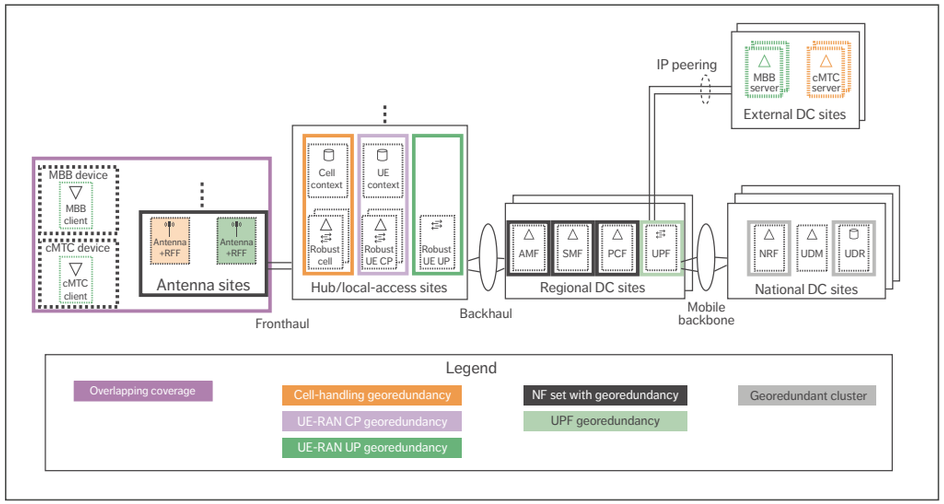


Figure 3 Wide-area network deployment example

partly based on cloud-native principles. These features and mechanisms will cover robust UE UP and UE CP on the RAN side, robust RAN resources like cell availability, radio link and air-interface redundancy mechanisms, as well as fronthaul transport.

Solutions at the 5G System level

5G introduces the support of URLLC services for industrial use cases. To support the high reliability of URLLC services, one standardized solution enables the UE to establish two redundant PDU sessions over the 5G network. In this case the 5GS sets up separate UP paths for the two redundant PDU sessions. Note, however, that RAN SPOFs like UE-RAN CP are not addressed by this solution. To avoid those SPOFs, dual UEs and dual network partitions is another possible solution.

Network architecture deployment examples

Another important aspect to consider when planning the deployment of a robust network is the desired service coverage area. There are three options: wide-area, confined wide-area and local-area deployments [1]. While robustness is important for all three, the most challenging requirements are currently considered to be in the last two categories.

Wide-area deployments

Figure 3 presents an example of a wide-area deployment in which three different national sites serve the whole country and two regional sites serve a specific region. The number of hub/local access sites and antenna sites varies in different networks, ranging from hundreds of hub/local access sites to thousands of antenna sites.

Wide-area deployments include a subset of the features and functionality to support use cases that include MBB and critical machine-type communication (cMTC), for example. Different features and functionality can be activated for UEs supporting these use cases, depending on the actual use-case requirements.

In the core network shown in Figure 3, the NF set is used for AMF, SMF and Policy Control Function (PCF) georedundancy. The 5GC UP resilience is shown as UPF georedundancy. The resilience mechanism for the Network Repository Function (NRF) and the UDR is a georedundant cluster. In the RAN, the new vendor-specific functions for robust UE UP and UE CP, robust RAN resources such as cells and overlapping cells and transmission reception points are shown.

Transport-level redundancy that covers mobile backbone, backhaul and fronthaul is just one example of another important consideration that is necessary to ensure robustness, particularly with respect to network topology and site design.

Local-area deployments

The most challenging use-case requirements for robustness are seen in local-area deployments at sites such as manufacturing premises [4]. These requirements include:

- stringent survival time
- local survivability (no events occurring outside the local area can have an impact on the local deployment)
- local data (no production-related information can leave the local premises).

A local standalone 5GS (including core network, RAN and local management of the connectivity, as well as all other aspects such as local transport and site solutions) is necessary to meet these requirements. In addition, integration with the rest of the local production system needs to be supported through network exposure functionality. A key to success is scaling down the 5G network while also maintaining the required robustness characteristics.

> HYBRID DEPLOYMENTS MAKE IT POSSIBLE TO RELAX THE ROBUSTNESS REQUIREMENTS ON THE LOCAL-AREA DEPLOYMENT

A local standalone 5GS uses most of the same robustness features and functionality that are used in a wide-area network deployment. In addition, it is possible to implement a redundancy solution in which every (industrial) device is equipped with two UEs that are connected either to a single robust network or to two parallel sets of local network partitions without any common failure points.

Hybrid deployments

Some use cases require support for both local-area and wide-area connectivity. In these cases, the local deployment is connected to a CSP network that

Terms and abbreviations

5GS – 5G System | **AMF** – Access and Mobility Management Function | **APN** – Access Point Name | **BE** – Back-End | **cMTC** – Critical Machine-Type Communication | **CP** – Control Plane | **CSP** – Communication Service Provider | **CUPS** – Control and User Plane Separation (of EPC nodes) | **DC** – Data Center | **E2E** – End-to-End | **EPC**– Evolved Packet Core | **FE** – Front-End | **HSS** – Home Subscriber Server | **ISP** – In-Service Performance | **MBB** – Mobile Broadband | **MME** – Mobility Management Entity | **NAS** – Non-Access Stratum | **NF** – Network Function | **NG-RAN** – Next-Generation RAN | **NRF** – Network Repository Function | **pCell** – Primary Cell | **PCF** – Policy Control Function | **PCRF** – Policy and Charging Rules Function | **PDU** – Protocol Data Unit | **PGW** – Packet Data Network Gateway | **RFF** – Radio Frequency Function | **RRC** – Radio Resource Control | **SAU** – Simultaneously Attached Users | **SMF** – Session Management Function | **SPOF** – Single Point of Failure | **UDM** – User Data Management | **UDR** – User Data Repository | **UE** – User Equipment | **UP** – User Plane | **UPF** – User Plane Function | **URLLC** – Ultra-Reliable Low-Latency Communication

supports wide-area connectivity. Hybrid deployments make it possible to relax the robustness requirements on the local-area deployment by making use of the CSP's wide-area network robustness functionality instead. It is important to note, however, that this advantage comes at the expense of losing the ability to support local survivability and local data.

Conclusion

Mobile broadband services have become critically important to the functioning of contemporary society and business. While both 4G and 5G are able to provide the high level of robustness required to deliver those services today, new and emerging use cases require the addition of new features and mechanisms in the network robustness toolbox.

The 5G System (5GS) has been designed to meet even the most challenging network robustness requirements. Ensuring the robustness of future networks requires a shift in focus from node level to network level, as well as consideration of all the different failure cases and a solid understanding of the needs of the most demanding applications. Beyond that, the creation of robust networks also requires careful network planning and deployment.

The 5GS robustness toolbox consists of both standardized and vendor-specific network features and mechanisms. Highly flexible, it gives communication service providers (CSPs) the power to activate the most appropriate mechanisms depending on the use cases and the deployment variants.

Further reading

- **Ericsson white paper, Enabling time-critical applications over 5G with rate adaptation, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/enabling-time-critical-applications-over-5g-with-rate-adaptation*
- **Ericsson white paper, 5G spectrum for local industrial networks, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/5g-spectrum-for-local-industrial-networks*
- **Ericsson white paper, Critical capabilities for private 5G networks, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/private-5g-networks*
- **Ericsson blog, This is the key to mobility robustness in 5G networks, available at:** *https://www.ericsson.com/en/blog/2020/5/the-key-to-mobility-robustness-5g-networks*
- **Ericsson blog, How can network operations make 5G systems resilient? , available at:** *https://www.ericsson.com/en/blog/2021/9/5g-resilient-system-network-operations*
- **Ericsson, 5G network for business growth, available at:** *https://www.ericsson.com/en/5g/5g-networks*

References

1. **Ericsson Technology Review, Critical IoT connectivity: Ideal for time-critical communications, June 2, 2020, Alriksson, F; Boström, L; Sachs, J; Wang, Y.-P. Eric; Zaidi, A, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/critical-iot-connectivity*
2. **Ericsson Technology Review, 5G-TSN integration meets networking requirements for industrial automation, August 27, 2019, Farkas, J; Varga, B; Miklós, G; Sachs; J, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/5g-tsn-integration-for-industrial-automation*
3. **3GPP TS 22.261, Service requirements for the 5G system, Release 18, 2021, available at:** *https://www.3gpp.org/ftp/Specs/archive/22_series/22.261/22261-i40.zip*
4. **Ericsson Technology Review, Boosting smart manufacturing with 5G wireless connectivity, February 20, 2019, Sachs, J; Wallstedt, K; Alriksson, F; Eneroth, G, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/boosting-smart-manufacturing-with-5g-wireless-connectivity*

THE AUTHORS

Jari Vikberg
◆ is a senior expert in network architecture and the chief network architect at CTO office. He joined Ericsson in 1993 and has both wide and deep technology competence covering network architectures for all generations of radio access and packet core networks. He is also skilled in the application layer and other domains, as well as in the impact and relation that they have to mobile networks. Vikberg holds an M.Sc. in computer science from the University of Helsinki, Finland.

Göran Hall
◆ is an expert in network architecture evolution at the CTO office. He joined Ericsson in 1991 to work on development and standardization, primarily within the area of packet core network architecture, which has so far included GPRS, WCDMA, PDC, EPC and 5GC. He has been chief network architect for the Packet Core domain in his previous assignment, including responsibility for the functional requirements

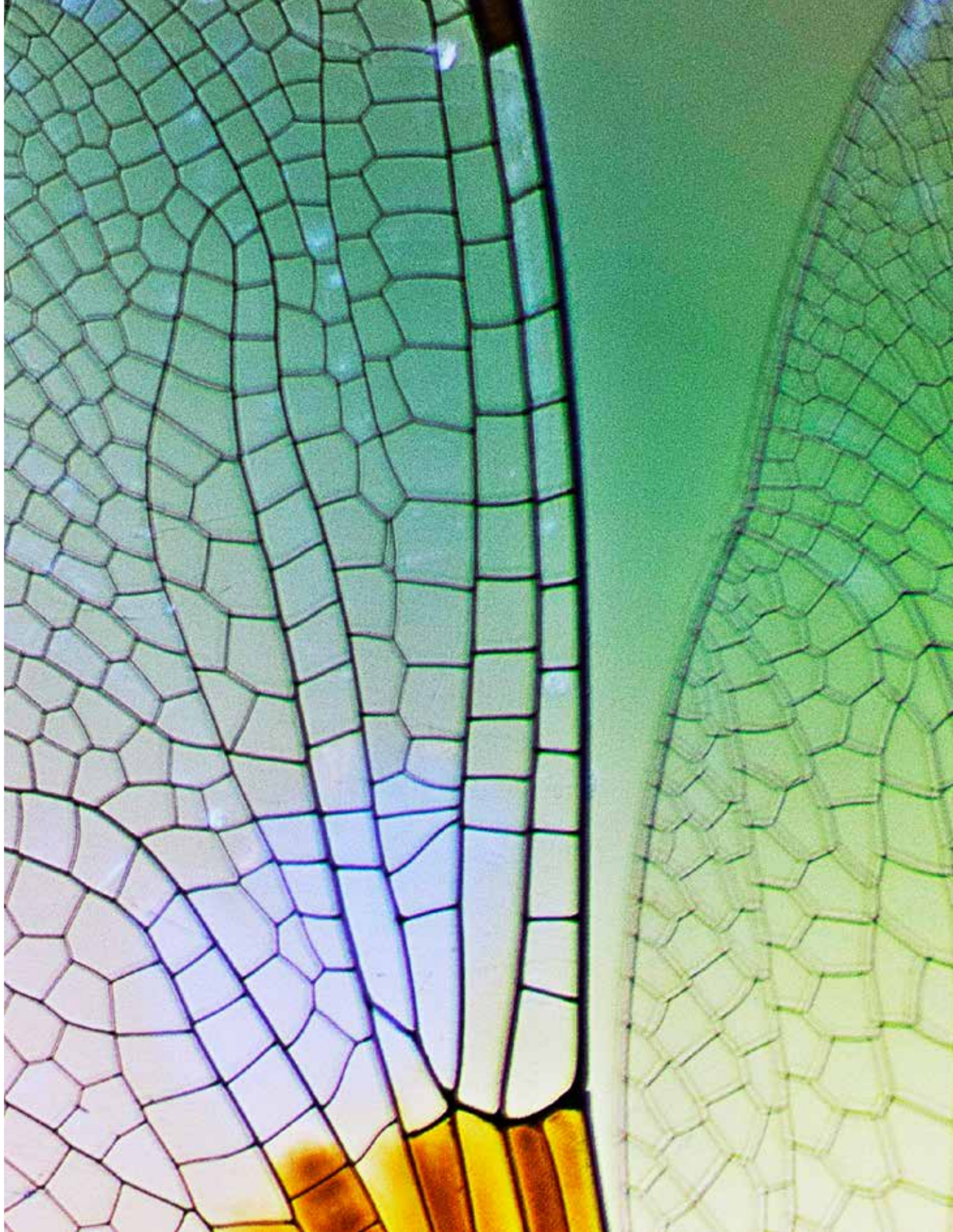
and architecture for the 5G Core network. Hall holds an M.Sc. in electrical engineering from Chalmers University of Technology in Gothenburg, Sweden.

Torbjörn Cagenius
◆ is a senior expert in network architecture at Business Area Digital Services. He joined Ericsson in 1990 and has worked in a variety of technology areas such as fiber-to-the-home, main-remote radio base station, fixed-mobile convergence, IPTV, network architecture evolution, software-defined networking and Network Functions Virtualization. In his current role, he focuses on 5G and associated network architecture evolution. Cagenius holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Richard Wang
◆ is an expert in core network robustness. He joined Ericsson in 2009 and has worked in different technology areas such as mobile-services switching centers, evolved packet core, voice-over-Wi-Fi, virtual EPC, non-3GPP access and 5GC evolution, as well as 5GC network robustness. In his current role, he focuses on the study of 5GC network-level robustness and evolution. He holds a Ph.D. in control theory and control engineering from Shanghai Jiao Tong University, China.

Johan Schultz
◆ is an expert in radio access network systems architecture design. He joined Ericsson in 1989 and has worked in various areas, mostly in or related to RAN, but also with transport and cloud hardware platforms. In his current role, he focuses on 5G RAN architecture and is also a volunteer in Ericsson Response. Schultz studied applied physics and electrical engineering at Linköping University, Sweden.

The authors would like to thank Fredrik Alriksson, Anna Larmo, Joachim Sachs, Robert Drincic, Gunnar Mildh, Torbjörn Keisu, Ben Wilmot, Krister Boman and Johan Torsner for their contributions to this article.



Quantum technology and its impact on security in mobile networks

While today's systems will remain secure against crypto-breaking quantum computers for many years to come, they do present a serious potential risk further into the future. To address this risk, new post-quantum algorithms that can easily be added to existing equipment and protocols are already in the final stages of standardization.

JOHN PREUß MATTSSON, BEN SMEETS, ERIK THORMARKER

Over the last 50 years, cryptography has evolved from its military and diplomatic origins to become a rich and widely-used tool to create complex cryptographic solutions for a multitude of applications. In the ICT industry, for example, an efficient combination of symmetric and public-key (asymmetric) cryptography is critical to the security of virtually every product, service and interface in use today.

■ Modern critical infrastructure such as 5G is implemented with zero trust principles where cryptography is used for confidentiality, integrity protection, and authentication on many of the logical layers of the network stack, often all the way from device to software in the cloud [1]. The cryptographic solutions in use today are based on well-understood primitives, provably secure protocols and state-of-the-art implementations that are secure against a variety of side-channel attacks.

The first signs of a serious quantum challenge to modern cryptography arose in 1994, when the mathematician Peter Shor proved that quantum computers can efficiently factor large integers and solve the discrete logarithm problem, which is believed to be intractable on ordinary computers. Unfortunately, Shor's result also showed that if sufficiently large and robust quantum computers can be built, then today's public-key cryptography – which relies on the intractability of these problems – will be broken.

There are multiple public engagements in industry and academia to build quantum computers at present, but the gap between today's quantum computers and ones that could threaten current public-key cryptography is huge. It is believed that the ability to break today's public-key cryptography with Shor's algorithm would require millions of so-called qubits – the quantum equivalents of bits in ordinary computers. Today's quantum computers typically have a maximum of about 100 qubits and they are not as robust as they would need to be to execute Shor's algorithm.

While the future progress of robust quantum computers is complex and uncertain, it should not be judged on simple metrics such as qubit-count alone. Assuming a Moore's law type of growth in qubit count, the scaling from 100 qubits to millions of qubits would take 25-30 years. Recent claims of researchers reaching quantum supremacy do not tell us anything substantial about the speed at which the gap is closing between today's quantum computers and the hypothetical machines that could threaten public-key cryptography.

Risks presented by quantum technology

Nobody knows if large-scale, robust quantum computers capable of attacking public-key cryptography – sometimes called Cryptographically Relevant Quantum Computers (CRQCs) – will ever be built. A 2019 estimate by a committee of experts said that the emergence of a CRQC during the next decade would be highly unexpected [2]. The committee also pointed out that there are no known applications for the intermediate medium-scale quantum computers that may appear in the coming years.

For most types of problem solving, quantum computers are much slower than ordinary computers, as the quantum error correction decimates the clock speed and number of usable qubits with several orders of magnitude, as shown in

Timeline for public-key cryptography and quantum computers

1976 – Diffie-Hellman key exchange

1977 – RSA cryptosystem

1978 – Code-based cryptography

1979 – Hash-based cryptography

1980 – Realization that a quantum computer can simulate things a classical computer cannot

1984 – Quantum key distribution

1985 – Elliptic curve cryptography

1986 – Grover's quantum algorithm inverts any function using only √N evaluations of the function

1994 – Shor's quantum algorithm introduces integer factorization in polynomial time instead of sub-exponential

1996 – Multivariate-quadratic-equations cryptography

1998 – Lattice-based cryptography

1998 – Quantum computer with two physical qubits

2001 – First quantum key distribution network

2011 – Supersingular elliptic curve isogeny cryptography

2015 – US government (NSA) announces it is planning to transition "in the not too distant future" from Suite B/CNSA to a new suite that is resistant to quantum attacks

2017 – The NIST announces the PQC standardization program

2018 – Standardization of stateful hash-based signatures (XMSS and LMS) by the IRTF Crypto Forum Research Group and the NIST

2019 – Quantum computer with 53 physical qubits

2022 – Target date for NIST to announce the first set of PQC algorithms for standardization and for the NSA to update the CNSA suite with PQC

2022-23 – Target date for draft NIST PQC standards

2024 – Target date for final NIST PQC standards

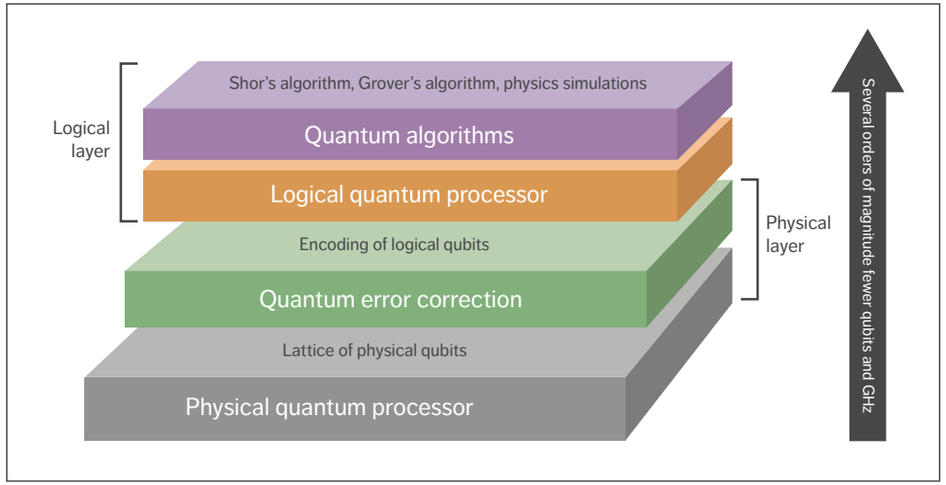


Figure 1 Envisioned structure of future quantum computers

Figure 1. As a result, quantum computers are not general-purpose super computers, but rather potential special-purpose machines for physics simulations and certain problems that require clever quantum algorithms.

Some commentators have argued that the development of quantum computing could lose momentum due to a lack of short-term applications or if its progress is too slow [3]. Nonetheless, as the consequences of success would be so severe from a security point of view, anyone who uses public-key cryptography such as RSA and elliptic curve cryptography (ECC) should start preparing now for the possibility that such large-scale machines could someday be built.

> STATEFUL HASH-BASED SIGNATURES HAVE WELL-UNDERSTOOD SECURITY, AND HAVE ALREADY BEEN STANDARDIZED

After all, a quantum attacker could not only decrypt communication, but also forge certificates and install fraudulent firmware updates. This would completely break the security of most consumer electronics, enterprise networks, the industrial Internet of Things and critical infrastructure. Even worse, information encrypted using public-key cryptography today could be recorded by attackers and used for attacks in the future when large-scale robust quantum computers potentially exist.

Fortunately, an alternative is already available for very long-lived signature keys such as those used in firmware updates. Stateful hash-based signatures have well-understood security, and have already been standardized by the Internet Engineering Task Force (IETF) and the US National Institute of Standards and Technology (NIST) [4]. There is a serious limitation to stateful hash-based signatures, however. Because they are stateful, they are only suitable for very specific applications.

Migration toward post-quantum cryptography

The NIST's post-quantum cryptography (PQC) standardization [5] is the most important ongoing project aimed at securing public-key cryptography against the threat of quantum computers. The purpose of the project is to standardize new algorithms that are believed to be secure against quantum computers. When standardized, these new primitives can replace today's public-key cryptography used for key exchange, public-key encryption and digital signatures. The new algorithms are typically as fast as today's ECC, but with significantly larger public keys, key encapsulations and signatures. The NIST aims to release draft standards for the first new PQC algorithms in 2022-23.

Lattice-based algorithms

The most important new class of post-quantum algorithms is lattice-based. These have public keys, key encapsulations and signatures starting in the 600-900 byte range. The corresponding quantities when using the current ECC are typically 32-64 bytes. There have been no new significant attacks against the lattice-based proposals during the standardization process, and the related mathematical problems have been studied extensively for the past two decades. Lattice-based proposals such as Kyber/Dilithium [6] offer a good middle way for PQC with efficient running times and average-sized communication overhead.

Potential key encapsulation mechanism and digital signature candidates

The two tables in *Figure 2* list performance and communication overhead for some of the key encapsulation mechanism (KEM) and digital signature candidates (finalists and alternates) in the NIST PQC standardization at their smallest parameter set [7, 8, 9]. The LMS algorithm is a stateful hash-based signature scheme with slow key generation, and signing and verification take at most a few milliseconds on a comparable platform to those used by the other algorithms in the table. Being stateful, LMS is not in scope in the NIST PCQ standardization. We have included it in the tables for comparison purposes, along with today's most important public-key cryptography algorithms.

> ERICSSON IS ENGAGING IN THE NIST PQC STANDARDIZATION AND THE PQC DISCUSSIONS IN THE IETF, 3GPP AND ETSI

Ericsson's role

Ericsson is engaging in the NIST PQC standardization and the PQC discussions in the IETF, 3GPP and ETSI, and will remain active when standards used in 5G such as TLS (Transport Layer Security), IKEv2 (Internet Key Exchange version 2), X.509, JOSE (JavaScript Object Signing & Encryption) and 5G SUCI (Subscription Concealed Identifier) are updated with the finalized NIST algorithms. While standards may be updated to support the new NIST PQC algorithms, it remains to be seen at what speed our current public-key cryptography is deprecated. This may, in part, depend on the progress in building quantum computers in the coming years. There is a balance between prudent preparations for switching to PQC and making sure that the investment in implementing PQC will be a long-term secure and good choice.

One way in which we are preparing Ericsson's products is by aligning with practices in the NIST Migration to Post-Quantum Cryptography project [10]. One key is crypto agility – the ability to upgrade cryptography and be prepared for the larger public keys used in PQC, for example. The US National Security Agency's (NSA's) Commercial National Security Algorithm (CNSA) cryptography suite is used to protect information in national security systems (NSSs) [11]. The CNSA suite is still not quantum-resistant, and information in NSSs may need protection for decades. This indicates that the NSA feels confident that large-scale robust quantum computers will not be a threat for decades to come.

For the most part, standardization organizations, governments and industries are

waiting for the final outcome of the NIST PQC standardization before they take action. The NSA became the exception recently when it announced its plans to add support in the CNSA suite for some of the lattice-based proposals at the end of the third round of the NIST standardization, planned for early 2022.

Post-quantum cryptography algorithm deployment

The initial deployment of the new PQC algorithms may be done in combination with current public-key cryptography so that, for example, an attacker would need to break both conventional elliptic curve Diffie-Hellman KEMs and one of the new PQC KEMs to learn an established session key in a communication protocol. For the most part, the migration to PQC is an algorithm update just like the previous updates from DES (Data Encryption Standard) to AES (Advanced Encryption Standard) and SHA (Secure Hashing Algorithm)-1 to SHA-2, but the larger sizes and slightly limited properties may require changes in protocols and application programming interfaces. The communication overhead of the new algorithms could lead to packet fragmentation in network communication, for example.

Table A

KEM algorithm	Generate key	Encaps.	Decaps.	Public key size	Encaps. size
NTRU (lattice-based PQC)	0.048ms	0.0073ms	0.012ms	699B	699B
Kyber (lattice-based PQC)	0.0070ms	0.011ms	0.0084ms	800B	768B
SABER (lattice-based PQC)	0.012ms	0.016ms	0.016ms	672B	736B
Classic McEliece (code-based PQC)	14ms	0.011ms	0.036ms	261120B	128B
SIKE (isogeny-based PQC)	3.0ms	4.4ms	3.3ms	197B	236B
ECDH (X25519) (non-PQC)	0.038ms	0.044ms	0.044ms	32B	32B
ECDH (P-256) (non-PQC)	0.074ms	0.18ms	0.18ms	32B	32B
RSA-3072 (non-PQC)	400ms	0.027ms	2.6ms	384B	384B

Table B

Signature algorithm	Generate key	Sign	Verify	Public key size	Signature size
Falcon (lattice-based PQC)	5.9ms	0.23ms	0.029ms	897B	666B
Dilithium (lattice-based PQC)	0.015ms	0.041ms	0.019ms	1312B	2420B
Rainbow (multivariate-based PQC)	2.7ms	0.017ms	0.0087ms	161600B	64B
SPHINCS+ (stateless hash-based PQC)	27ms	210ms	0.28ms	32B	7856B
LMS (limited to 220 messages – stateful hash-based PQC)	-	-	-	56B	2828B
Ed25519 (non-PQC)	0.014ms	0.015ms	0.050ms	32B	64B
ECDSA (P256) (non-PQC)	0.029ms	0.041ms	0.086ms	32B	64B
RSA-3072 (non-PQC)	400ms	2.6ms	0.027ms	384B	384B

Figure 2 Tables showing performance and communication overhead for some of the KEM and digital signature candidates in the NIST standardization

Quantum impact on symmetric cryptography

In 1996, Shor's result was complemented by an algorithm developed by the computer scientist Lov Grover, which showed that quantum computers could search through the possible inputs to a black-box function to find an input that gives a sought output. While Grover's algorithm can do this in much fewer evaluations of the black-box function than any ordinary algorithm, it is still very slow compared with Shor's quantum algorithm. (The meaning of black box in this context is that Grover's algorithm does not rely on any internal structure of the function – it is a generic method.)

In theory, an attacker with a quantum computer can use Grover's algorithm to break the symmetric cipher AES-128 through a quantum computation that consists of 2^{64} serial AES-128 encryptions. Each such AES-128 encryption in turn consists of approximately 2^{11} serial quantum gates. This gives a total serial computation of length 2^{75} quantum gates. However, the quantum gates can introduce errors, and further overhead piles up from quantum error-correction. What all this means in practice is that the attacker must split up the computation over multiple quantum computers. Since Grover's algorithm does not parallelize efficiently, as illustrated in *Figure 3*, the use of 100 quantum computers would only speed up the computation by a factor of 10.

Considering all this, Grover's algorithm does not pose any apparent threat to symmetric cryptography. Some years ago, there was a common conception that Grover's algorithm required symmetric key sizes to be doubled – requiring use of AES-256 instead of AES-128. This is today considered a misconception – NIST, for example, now states that AES-128 will likely remain secure for decades to come, despite Grover's algorithm [5].

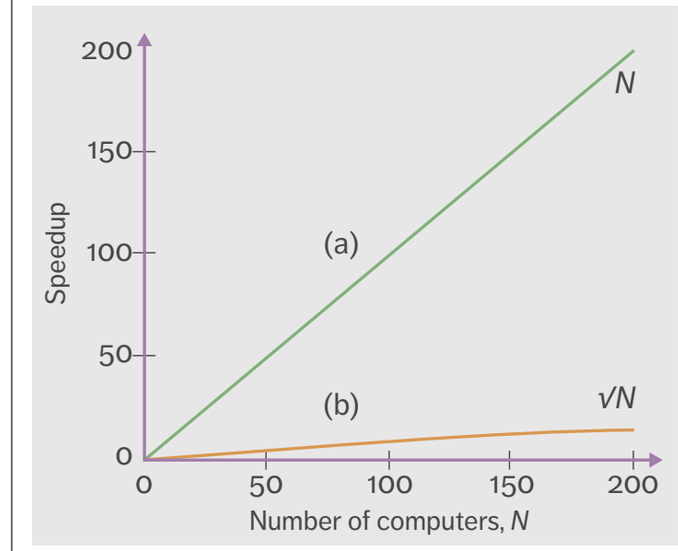


Figure 3 Parallelization of key search using (a) ordinary computers and (b) quantum computers and Grover's algorithm

In fact, one of the security levels in the NIST PQC standardization is equivalent to that of AES-128. This means that NIST thinks it is relevant to standardize parameters for PQC that are as strong under quantum attacks as AES-128. There could, of course, be other reasons why a longer key is needed, such as compliance, and using a longer key only has a marginal effect on performance.

In summary, our most important symmetric cryptographic tools (AES, SNOW 3G, SHA2, SHA3 and so on) remain secure against quantum computers as they are. This also applies to the authentication, key generation, encryption and integrity in 3G, 4G and 5G that rely purely on symmetric cryptography.

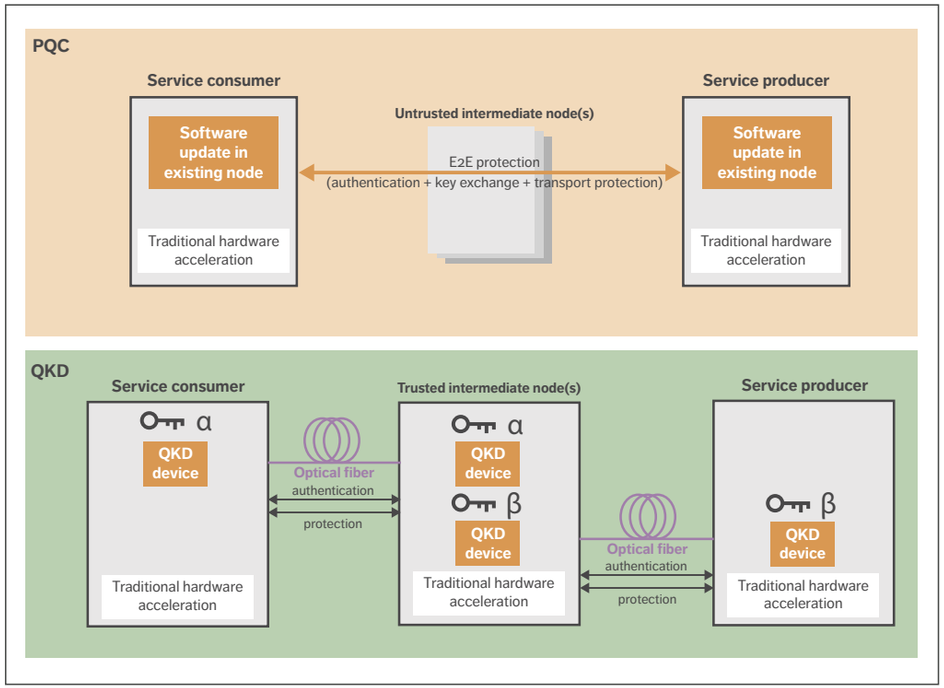


Figure 4 Differences between PCQ and QKD when applied to network infrastructure

Quantum cryptography

The idea of quantum cryptography is to leverage quantum mechanics to build cryptography. This is very different from, for example, the post-quantum cryptography that is being standardized by NIST, which can run completely in software like any other conventional cryptography. While quantum cryptography is an exciting academic research topic, its practical security applications are as yet uncertain. So far, quantum key distribution (QKD) and quantum random number generators (QRNGs) are the two types of quantum cryptography that have sparked the most interest. However, current implementations still have a long way to go before they are hardened and certified for practical use.

Quantum key distribution

QKD is a quantum-resistant mechanism for key distribution in which two parties agree on a secret key by sending photons between them with the help of a second (ordinary) authenticated communication channel, as shown in the bottom half of *Figure 4*. An idealized mathematical abstraction of QKD is famously unconditionally secure. While security proofs for theoretical constructions are an important building block in conventional cryptography as well, it is important to understand that the most important threat surface of cryptography is consistently found to be in the implementation details. The main principle for managing this threat in conventional cryptography is to use well-reviewed implementations that build on collective implementation knowledge that has been gained over decades.

In contrast to conventional cryptography and PQC, the security of QKD is inherently tied to the physical layer, which makes the threat surfaces of QKD and conventional cryptography quite different. QKD implementations have already been subjected to publicized attacks [12] and the NSA notes that the risk profile of conventional cryptography is better understood [13]. The fact that conventional cryptography and PQC are implemented at a higher layer than the physical one means PQC can be used to securely send protected information through untrusted relays, as illustrated in the top half of Figure 4. This is in stark contrast with QKD, which relies on hop-by-hop security between intermediate trusted nodes. The PQC approach is better aligned with the modern technology environment, in which more applications are moving toward end-to-end security and zero-trust principles. It is also important to note that while PQC can be deployed as a software update, QKD requires new hardware.

Regarding QKD implementation details, the NSA states that communication needs and security requirements physically conflict in QKD and that the engineering required to balance them has extremely low tolerance for error. While conventional cryptography can be implemented in hardware in some cases for performance or other reasons, QKD is inherently tied to hardware. The NSA points out that this makes QKD less flexible with regard to upgrades or security patches. As QKD is fundamentally a point-to-point protocol, the NSA also notes that QKD networks often require the use of trusted relays, which increases the security risk from insider threats.

> PQC CAN BE USED TO SECURELY SEND PROTECTED INFORMATION THROUGH UNTRUSTED RELAYS

As QKD requires external authentication through conventional cryptography, the UK's National Cyber Security Centre cautions against sole reliance on it, especially in critical national infrastructure sectors, and suggests that PQC as standardized by the NIST is a better solution [14]. Meanwhile, the National Cybersecurity Agency of France has decided that QKD could be considered as a defense-in-depth measure complementing conventional cryptography, as long as the cost incurred does not adversely affect the mitigation of current threats to IT systems [15].

Quantum random number generators

Secure randomness is critical in cryptography – if the quality of randomness generators is poor, numerous cryptographic protocols will fail to deliver security. Although conventional hardware randomness generator technology is robust and

Terms and abbreviations

AES – Advanced Encryption Standard | **CNSA** – Commercial National Security Algorithm | **CRQC** – Cryptographically Relevant Quantum Computer | **ECC** – Elliptic Curve Cryptography | **ECDH** – Elliptic Curve Diffie–Hellman | **ECDSA** – Elliptic Curve Digital Signature Algorithm | **IRTF** – Internet Research Task Force | **KEM** – Key Encapsulation Mechanism | **LMS** – Leighton-Micali Signature | **NIST** – National Institute of Standards and Technology (US) | **NSA** – National Security Agency (US) | **NSS** – National Security System (US) | **NTRU** – N-th degree Truncated polynomial Ring | **PQC** – Post-Quantum Cryptography | **QKD** – Quantum Key Distribution | **QRNG**– Quantum Random Number Generator | **RSA** – Rivest-Shamir-Adleman | **SHA** – Secure Hashing Algorithm | **SIKE** – Supersingular Isogeny Key Encapsulation | **XMSS** – eXtended Merkle Signature Scheme

secure against quantum computers, QRNGs have nonetheless attracted some attention in recent years. QRNGs work according to a physical realization of a quantum model, instead of the other physical processes used in conventional hardware randomness generators.

QRNGs are sometimes advertised as generating perfect unbiased random bits in contrast to the biased bits that come from conventional generators. In reality, though, any bias in the bits output by conventional generators is smoothed out in post-processing through the application of pseudo-random number generators, which work according to the same mechanism that enables a single 128-bit AES key to produce many gigabytes of random-looking encrypted data.

If QRNG technology becomes as well understood in the future as our current hardware randomness generator technology, then it could, in principle, be certified, validated and evaluated on the same grounds.

Conclusion

While we do not expect quantum computers with the ability to attack current cryptography to emerge for many years to come, we strongly encourage communication service providers to start planning the process of migrating to post-quantum cryptography. With the support of vendors including Ericsson, standards-developing organizations such as the US National Institute of Standards and Technology, the Internet Engineering Task Force and the 3GPP are working on new, post-quantum algorithms and updated protocols that can easily be added to existing equipment and interfaces. Currently in the final stages of standardization, these algorithms will be available in the next couple of years to help our industry mitigate potential future threats against mobile infrastructure and services.

References

1. **Ericsson Technology Review, Zero trust and 5G – Realizing zero trust in networks, May 2021, Olsson, J.; Shorov, A.; Abdelrazek, L.; Whitefield, J., available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/zero-trust-and-5g*
2. **NAP, Quantum Computing: Progress and Prospects, 2019, available at:** *https://www.nap.edu/catalog/25196/quantum-computing-progress-and-prospects*
3. **IEEE Spectrum, The case against quantum computing, November 15, 2018, Dyakonov, M, available at:** *https://spectrum.ieee.org/the-case-against-quantum-computing*
4. **NIST, SP 800-208, Recommendation for Stateful Hash-Based Signature Schemes, October March 2020, available at:** *https://csrc.nist.gov/publications/detail/sp/800-208/final*
5. **NIST, Post-Quantum Cryptography, available at:** *https://csrc.nist.gov/projects/post-quantum-cryptography*
6. **CRYSTALS Cryptographic Suite for Algebraic Lattices, available at:** *https://pq-crystals.org/index.shtml*
7. **eBACS: ECRYPT Benchmarking of Cryptographic Systems (r24000 machine), available at:** *https://bench.cr.yp.to/supercop.html*
8. **SIKE, Supersingular Isogeny Key Encapsulation, October 1, 2020, Jao, D et al., available at:** *https://sike.org/files/SIDH-spec.pdf*
9. **SPHINCS+: Submission to the NIST post-quantum project, v.3, October 1, 2020, Aumasson, J-P, et al., available at:** *https://sphincs.org/data/sphincs+-round3-specification.pdf*
10. **NIST, Migration to Post-Quantum Cryptography, August 2021, Barker, W; Souppaya, M; Newhouse, W, available at:** *https://csrc.nist.gov/publications/detail/white-paper/2021/08/04/migration-to-post-quantum-cryptography/final*
11. **NSA, Commercial National Security Algorithm Suite, available at:** *https://apps.nsa.gov/iaarchive/programs/iad-initiatives/cnsa-suite.cfm*
12. **Physical Review A 78, Experimental demonstration of time-shift attack against practical quantum key distribution systems, October 28, 2008, Zhao, Y.; Fung, C.; Qi, B.; Chen, C.; Lo, H., available at:** *https://journals.aps.org/pra/abstract/10.1103/PhysRevA.78.042333*
13. **NSA, Post-Quantum Cybersecurity Resources, available at:** *https://www.nsa.gov/Cybersecurity/Post-Quantum-Cybersecurity-Resources/*
14. **National Cyber Security Centre, Quantum security technologies, March 24, 2020, available at:** *https://www.ncsc.gov.uk/whitepaper/quantum-security-technologies*
15. **ANNSI, Should quantum key distribution be used for secure communications?, May 2020, available at:** *https://www.ssi.gouv.fr/uploads/2020/05/anssi-technical_position_papers-qkd.pdf*

THE AUTHORS

John Preuß Mattsson

◆ is a senior specialist in internet security protocols. He joined Ericsson in 2007 and has been active in many standardization organizations such as the 3GPP, the GSMA, the IETF, the IRTF (Internet Research Task Force) and the NIST. His work focuses primarily on cryptography, security protocols, the Internet of Things and trade compliance. Mattsson holds an M.Sc. in engineering physics from KTH Royal Institute of Technology, Stockholm, Sweden, and an M.Sc. in business administration and economics from Stockholm University.

Ben Smeets

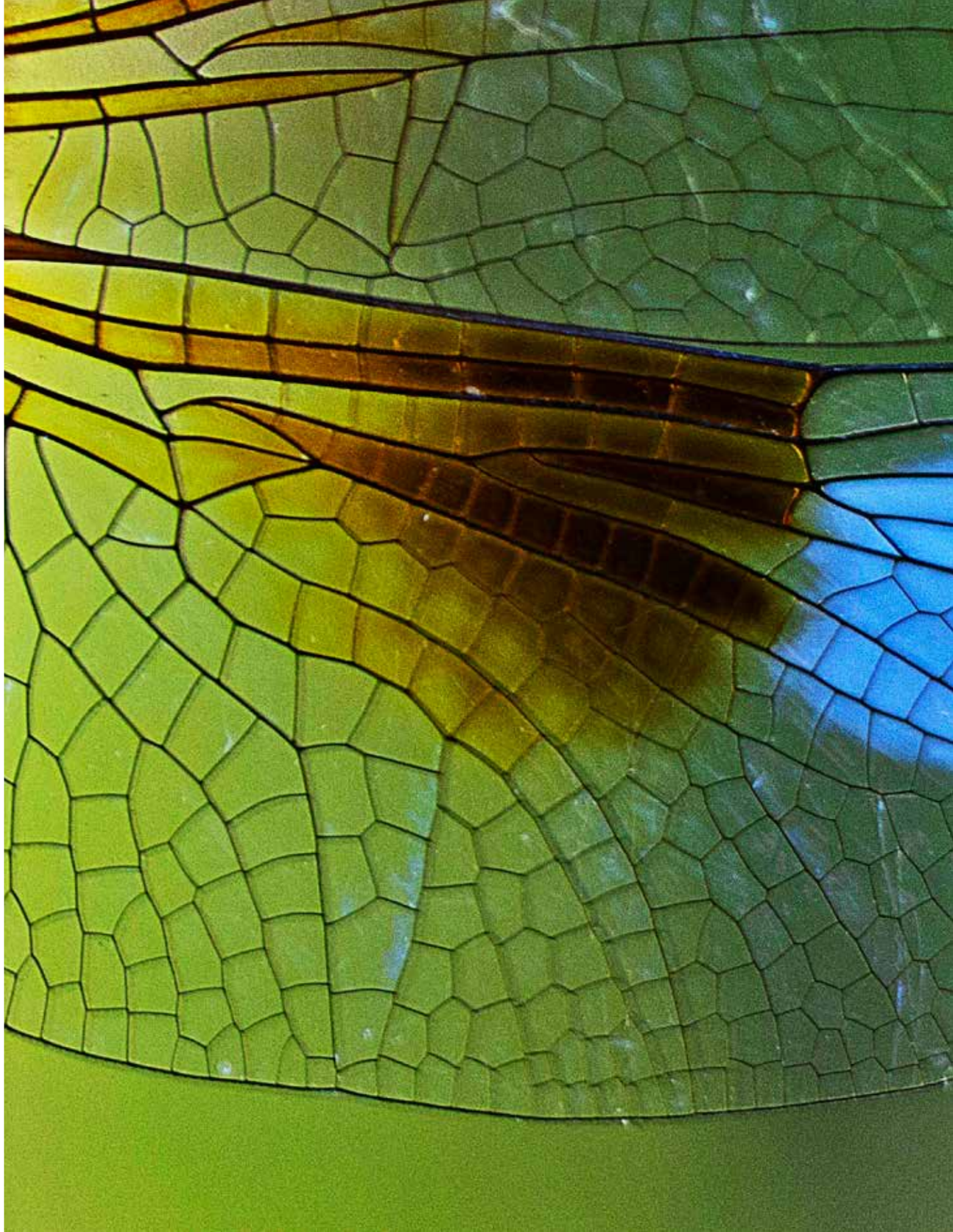
◆ is a senior expert in trusted computing at Ericsson Research. He joined Ericsson in 1998 and started out working on security solutions for mobile phone platforms. He is currently working on trusted computing technologies in connection with containers and secure enclaves. Smeets holds a Ph.D. in information theory from Lund University, Sweden, where he also serves as a professor.

Erik Thormarker

◆ joined Ericsson in 2018 as an experienced researcher. His research interests include post-quantum cryptography, cryptographic protocols and cryptanalysis. Thormarker holds an M.Sc. from the joint master's program in mathematics at KTH Royal Institute of Technology and Stockholm University.

Further reading

- **Ericsson blog, The evolution of cryptography in mobile networks and how to secure them in the future, June 29, 2021, Preuß Mattsson, J; Çomak, P; Karakoç, F, available at:** *https://www.ericsson.com/en/blog/2021/6/evolution-of-cryptographic-algorithms*
- **DOI, The security implications of quantum cryptography and quantum computing, September 2020, Cavaliere, F; Preuß Mattsson, J; Smeets, B, available at:** *https://doi.org/10.1016/S1353-4858(20)30105-7*
- **Ericsson blog, An introduction to quantum computer technology, July 25, 2019, Vall-llosera, G; Awan, A. J.; Sefidcon, A, available at:** *https://www.ericsson.com/en/blog/2019/7/introduction-to-quantum-computer-technology*



Service exposure and automated life-cycle management:

THE KEY ENABLERS FOR 5G SERVICES

Service exposure and automated life-cycle management enable communication service providers to offer a variety of innovative services to enterprises and application developers, while simultaneously establishing new revenue streams through relationships with hyperscale cloud providers.

MALGORZATA SVENSSON, BENEDEK KOVÁCS, ELISABETH MUELLER, MASSIMILIANO MAGGIARI, RÓBERT SZABÓ

The digitalization of society and the growing popularity of 5G-enabled use cases are creating new business opportunities for communication service providers (CSPs) to utilize 3GPP and cloud-based technologies in the enterprise domain.

■ CSPs that want to pursue new business opportunities in the enterprise domain must be able to frame their service offerings to fit the individual needs and desired use cases of enterprises and their partners. Three capabilities are required to achieve this. Firstly, CSPs need to create an expanded service portfolio that combines their connectivity offerings with the cloud and edge platform offerings of hyperscale cloud providers (HCPs). Secondly, they need a network that can serve as a programmable platform with the ability to expose services to developers. Thirdly, they need the ability to onboard new applications into the network with optimized runtime support for traffic routing toward applications.

Being able to offer connectivity services in combination with HCPs' cloud and edge platform offerings will make enterprise applications available to users close to their office locations in an efficient and scalable way, and in compliance with security requirements and local regulations. A growing number of CSPs have already started establishing relationships with HCPs to deploy applications and network functions in HCP environments. Expanding these relationships to offer the HCP's environment in combined offerings for enterprises, partners and application developers is a logical next step. Service-level-agreement (SLA) driven automation is a critical system capability to enable such combined offerings.

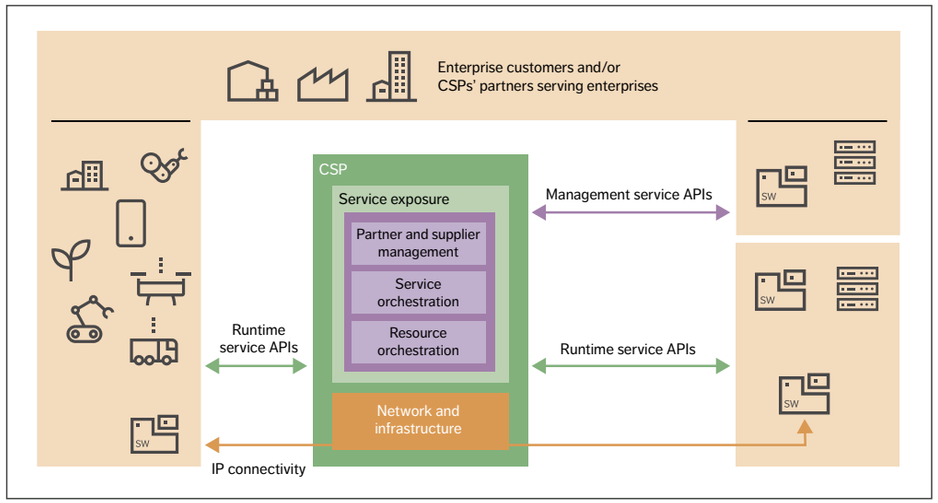


Figure 1 Service exposure and automated life-cycle management

Exposing the network as a programmable connectivity platform enables enterprises and their partners to build applications that can observe the network, influence it, and optimize the traffic flow to enable use cases like unmanned aerial vehicles that demand well-defined response times. As a consequence, developer communities require easy access to exposed service application programming interfaces (APIs).

In the later stages, once applications have been built, they need to be onboarded and integrated into a CSP's environment, so that they can seamlessly be made available on the new services platform, bringing connectivity and applications together. Mechanisms for monetization of combined offerings and service exposure need to be put in place, where the latter is powered by centralized API hubs that monitor API usage and provide input to the various business models.

Service exposure and automated life-cycle management

The term "service exposure and automated life-cycle management" refers to a set of capabilities that targets increased efficiency in CSPs' operational processes through a higher level of automation and scalability in service orchestration architectures. As shown in *Figure 1*, service exposure and automated life-cycle management combine various service management and runtime connectivity services that

Terms and abbreviations

API – Application Programming Interface | **CSP** – Communication Service Provider | **DNS** – Domain Name System | **E2E** – End-to-End | **HCP** – Hyperscale Cloud Provider | **I-WLP** – Intelligent workload placement | **NF** – Network Function | **OTT** – Over-the-Top | **SLA** – Service Level Agreement | **SW** – Software | **UE** – User Equipment | **URSP** – User Equipment Route Selection Policy

are targeted to enterprises, partners and application developers. APIs and orchestration are used to enable the simple and efficient creation and launch of new services and provide value-added functionality in runtime on top of the basic connectivity.

Service exposure provides service and network abstraction layers on top of network APIs, available for usage and integration by enterprise customers and partners to support various business models, where the abstraction layer hides the network complexity.

Partner and supplier management (the top purple box in Figure 1) helps CSPs manage agreements with all types of business partners, including HCPs. These agreements regulate service requirements and the infrastructure capacity required for own or resell usage.

Service orchestration (the middle purple box in Figure 1), together with inventory and assurance, are used for service design, creation and activation, as well as to ensure that the SLA is always fulfilled. These capabilities include interaction with service management on the partner side to onboard services provided by partners and intelligent workload placement that enables data-driven service and network design. The intelligent workload placement fulfills requirements for geographical availability with the distributed target environments and must be integrated with inventory, topology, and service assurance to ensure that decisions are based on the real state of the network. Service orchestration must be tightly integrated to service order management processes to automate the service life-cycle management.

Resource orchestration (the bottom purple box in Figure 1) provides an abstraction layer to various technology domains that result from different business agreements, such as those between CSPs and HCPs. Resource orchestration offers efficient, flexible and automated life-cycle management of CSP software, realizing the network functions (NFs) as well as the consumer and enterprise applications. Initial deployments, service allocation, updates and upgrades of the software can be orchestrated in target environments owned by the CSP or through partner and/or supplier agreements. Multi-cloud capacity management is supported to enable service and resource orchestration.

> PARTNER AND SUPPLIER MANAGEMENT HELPS CSPs MANAGE AGREEMENTS WITH ALL TYPES OF BUSINESS PARTNERS

Service exposure (the light green box in Figure 1) enables management service APIs (represented by the purple arrow) to manage connectivity offerings and runtime service APIs (represented by the green arrows) to influence network behavior to meet the desired service characteristics. The management service APIs enable the simple and efficient creation and launch of new services, offering traffic steering for applications at edges. The service APIs are well defined and compliant with the relevant standards, including those of the 3GPP and TM Forum, to ensure interoperability and ease of use.

Partner and supplier management

On top of managing the SLAs between CSPs, HCPs and other digital partners, the partner and supplier management stack also serves as the connection between CSPs' and partners' catalogs, making it possible to compose offerings targeted at enterprise customers and application providers. The partner and supplier management stack controls how multi-tenancy is used by both service and resource orchestration, keeping track of which services and resources are used by whom.

CSPs can use these new capabilities to pursue new business opportunities through building relationships with HCPs. HCPs can become the CSPs' partners, visible to enterprise customers and application providers in combined exposed services offered by the CSP.

Partnerships with HCPs enable CSPs to broaden their end-to-end (E2E) connectivity capabilities for enterprises with application platform offerings and thereby facilitate other kinds of business relationships as well. Relationships between CSPs and industry verticals is just one example.

CSPs may also choose to extend their relationships with the HCPs, so that the HCPs become their suppliers. In the supplier role, the HCP provides a full-service offering comprised of critical infrastructure along with the cloud and edge stack for consumer applications and telecommunication service deployment.

When CSPs have their own cloud platform offerings, they typically create them by leveraging the services in their partner catalogs that are provided by HCPs. In the case of telecommunication services, applications and NFs are deployed in the cloud, and the cloud resources for service deployment are shielded by services owned and supplied by an HCP. This means that telecommunication services must refer to services defined in the CSP's partner catalog.

Services supplied by HCPs are subject to SLAs between CSPs and HCPs and must be monitored accurately. It is the HCP's responsibility to provide the service assurance for these services in compliance with the service level objectives agreed between the CSP and the HCP. The supplier SLA sets boundaries for any customer-facing SLAs agreed between a CSP and its enterprise customers when delivering combined offerings.

The service-ordering process is driven by connectivity and application platform offering definitions stored in the service catalog. The ordered combined services are decomposed into various domain-level services, and the instantiation of these services is triggered toward service orchestration.

New monetization opportunities arise from flexible business-to-business-to-anything business models, which can consider multiple parties (customers and partners) during charging and billing business processes.

Service and resource orchestration

Automated service life-cycle management aims to simplify the CSP's operational processes and increase efficiency. Service and resource orchestration plays an important role in enabling the simple and efficient creation and launch of new services, as well as providing value-added functionality in runtime on top of the basic connectivity.

> NEW MONETIZATION OPPORTUNITIES ARISE FROM FLEXIBLE BUSINESS-TO-BUSINESS-TO-ANYTHING BUSINESS MODELS

The service orchestration stack includes capabilities to manage services across multiple technologies and business domains in complex operator environments. It also makes it possible to onboard partner services and associated deployment rules in an automated way. These partner services are used as building blocks in the service creation process of services offered by the CSP to enterprise customers. It is also possible to orchestrate various services including network connectivity, enterprise connectivity services and applications, along with orchestration of their deployments in HCP-provided clouds.

The introduction of HCP cloud infrastructure in the telecommunication world is happening gradually and at different speeds, depending on the business objectives of individual CSPs. To address the market and technology situations, full, partner and limited edge [1] have been identified as the go-to-market scenarios. In the full edge scenario, the CSP operates the infrastructure for telco workloads, third-party and over-the-top (OTT) workloads. The infrastructure provider can be any private infrastructure provider.

In the partner edge scenario, the HCP operates the infrastructure for third-party and OTT workloads, but it is deployed in the CSP or enterprise premises. The infrastructure provided for third-party and OTT workloads is used exclusively for workloads controlled by the CSP. The infrastructure for telco workloads can be operated by the CSP or the HCP, deployed in the CSP or in the enterprise

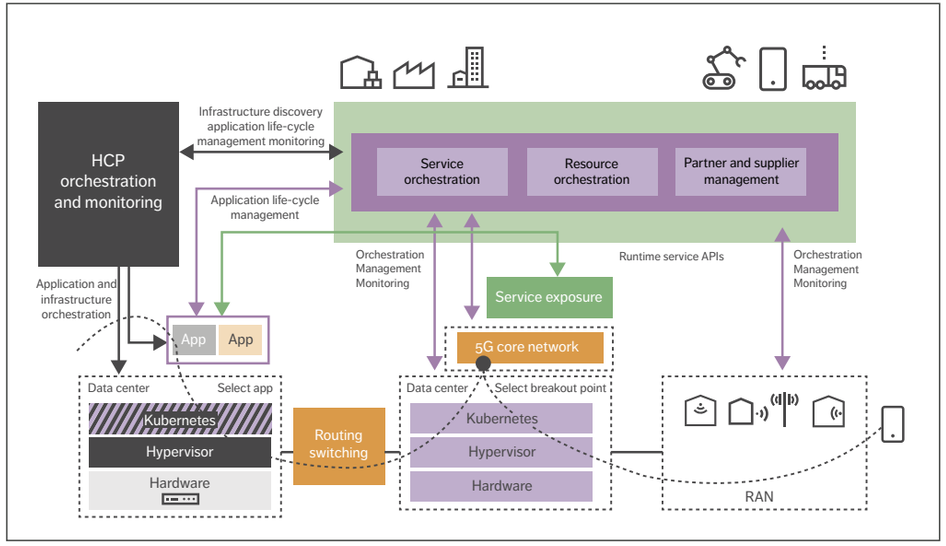


Figure 2 Operations support systems architecture for dual stack

premises and for the exclusive use of the CSP.

In the limited edge scenario, the HCP operates the infrastructure for third-party and OTT workloads. It can be deployed in the CSP or enterprise premises. The HCP will use and control the infrastructure provided for third-party and OTT workloads. The infrastructure for telco workloads can be operated by the CSP, as in the full edge deployment scenario, or operated by the HCP, deployed in the CSP premises or in the enterprise premises and for the exclusive use of the CSP.

These scenarios are not isolated from each other and we can already see the full, partner and limited edge variants happening at the same time, which has led to the dual-stack architecture depicted in *Figure 2*. Telecommunication workloads are deployed in a private telco cloud, while enterprise and consumer applications are deployed in HCP edge zones. In the initial phase, where both the full and partner edge go-to-market scenarios apply, the E2E services, the applications' connectivity awareness, and the operational processes automation across the two cloud stacks are essential to fulfill the service requirements.

At a later stage (or in the short term for more advanced cases) the integrated-stack architecture shown in *Figure 3* will be introduced, in which telecommunication applications are deployed together with enterprise and consumer applications in the cloud infrastructure supplied by the HCP for the exclusive use of the CSP.

In both the dual and the integrated-stack architecture scenarios, the CSP's service and resource orchestration stack performs the life-cycle management of the telecommunication applications and combined services through direct access to the requested Kubernetes clusters.

The main objective of the service and resource orchestration functionality is to minimize the impact of public and/or private cloud deployments on telecommunication services and applications. It is essential to provide the same operational experience to CSPs and enable the commercialization of the services for enterprises irrespective of the variations of underlying cloud architecture resulting from the various business models and go-to-market tactics. In the full edge scenario, the service and resource orchestration stack handles the life-cycle management of the cloud infrastructure, including Kubernetes clusters. In the partner and the limited edge scenarios, on the other hand, the orchestration stack interworks with the HCP orchestration tools to enable the life-cycle management and monitoring of the cloud infrastructure.

Dual-stack architecture

Service and resource orchestration supports business models in which CSPs own the telco cloud infrastructure and the enterprise cloud infrastructure is provided by HCP suppliers. These business models result in the dual-stack architecture shown in Figure 2, where the CSPs use their own infrastructure (represented by the light purple boxes under the 5G core network in Figure 2) to deploy and life-cycle-manage NFs, and enterprise applications are hosted in the infrastructure provided by the HCP (represented by the purple-black striped box in Figure 2). The two cloud infrastructures are deployed side by side, with E2E automated processes realized by the orchestration stack.

The dual-stack architecture supports the life-cycle management and monitoring of NFs as well as the life-cycle management of enterprise applications that are part of the CSP-managed catalog. It also enables the life-cycle management of CSP-owned cloud infrastructure. The architecture assumes that the integration with the HCP's orchestration and assurance tools and APIs is done through the CSP's service exposure, which is used to delegate application deployment in HCP-based managed cloud infrastructure (such as container as a service and platform as a service). The CSP's service exposure is used to collect metrics to monitor enterprise applications and it is the CSP orchestration stack that provides an E2E service control and configuration point.

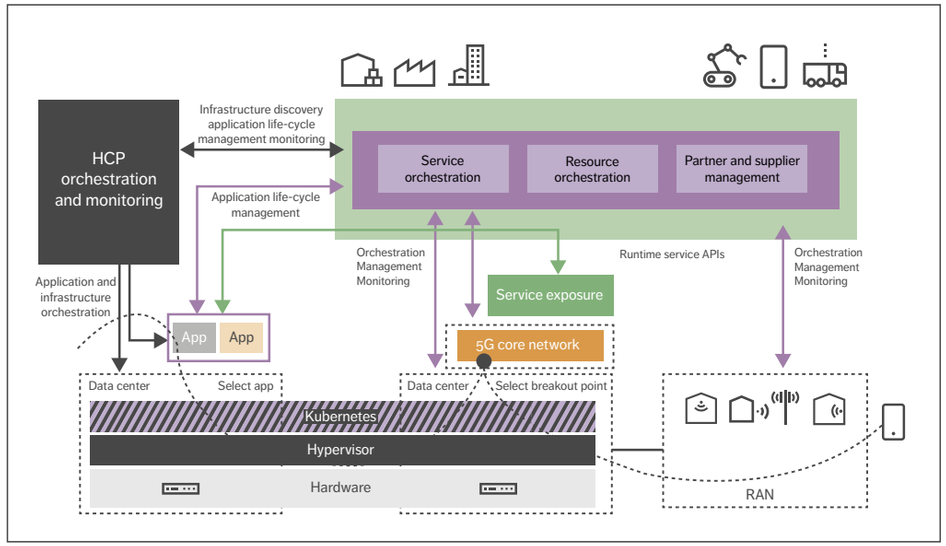


Figure 3 Operations support systems architecture for integrated stack

THE MULTI-CLOUD ABSTRACTION LAYER HARMONIZES HCP-SPECIFIC BEHAVIORS AND UNIFIES CSP OPERATIONAL EXPERIENCE TO AVOID VENDOR LOCK-IN

The runtime service APIs (represented by the green arrow in Figure 2) can be used by applications to influence network behavior as URSP rules, network slicing policy rules, QoS and traffic steering.

Integrated-stack architecture

In scenarios where the NFs run in the HCP's cloud, service and resource orchestration provides the capabilities to orchestrate and manage the integrated cloud stack. Based on agreements with CSPs, HCP-supplied infrastructure (represented by the purple-black striped box in Figure 3) is used to run NFs along with enterprise and partner applications. The architecture supports management capabilities for the NFs and applications if they are part of the CSP-managed catalog. The management capabilities comprise orchestration, life-cycle management, and monitoring, as well as the integration with the HCP's orchestration and monitoring through service exposure.

Both the dual-stack and integrated-stack architectures utilize the same set of APIs provided by HCPs – that is, infrastructure automation, infrastructure monitoring and Kubernetes APIs. These APIs are consumed by the CSP operations support systems stack. Both architectures also support capabilities such as third-party infrastructure and service discovery, and topology and inventory management, enabling data-driven intelligent service design in mixed cloud environments.

Service and resource orchestration also offers service configuration to allow application traffic breakout points at desired network edges (represented by the purple arrows in Figure 3). Furthermore, it supports the configuration of user equipment route selection policies (URSPs) to support edge application deployments, as well as the enablement of service exposure APIs for application providers (represented by the green arrows in Figure 3).

The multi-cloud abstraction layer harmonizes HCP-specific behaviors and unifies CSP operational experience to avoid vendor lock-in. The layer includes the HCP-agnostic information model and the south-bound adapters for HCP APIs.

Advanced multi-tenancy provides the level of isolation required to manage the HCP's resource allocations. This is essential to enable visibility and troubleshooting of the HCP's availability zones, as well as the onboarded enterprises' and partners' users across the network.

Intelligent workload placement and data-driven network design

Optimal placement of applications and dynamic traffic routing are key differentiators that automated service life-cycle management provides. The key enabler for the optimal placement is topology-aware orchestration, where the knowledge about how to connect 5G connectivity to enterprise and consumer applications resides. Enhanced topology discovery and unified topology models are the most significant capabilities of the orchestration.

An intelligent workload placement (I-WLP) service is characterized by four key features. First and foremost, it must be driven by the CSP's business intents. Secondly, it must respect the onboarding status and availability, SLA objectives, and resource and hardware availability of the components. Thirdly, it must be able to work under continuously changing circumstances, including onboarding changes, software upgrades, new resources and capabilities, changes to the business intents or CSP operational policies, and changes to the costs and availabilities of HCP cloud locations. And lastly, it needs to have knowledge of the operational state of the distributed system – that is, which service instances are running where with what dependencies. Inventories can provide running, planned and pending reservation states of services, while monitoring systems can provide actual and historical resource usage and SLA metrics.

From the algorithm perspective, I-WLP is a multi-constraint, multi-objective graph allocation problem, which cannot be solved in polynomial time. Therefore, we have created methods based on greedy heuristics, which are very flexible in terms of supporting service constraints such as latency bounds, QoS classes, vendor preferences, locations and anti-affinities. They can also consider combinations of different operational policies relating to cost, utilization, performance and resilience optimizations, for example.

An inventory of all resources, services and capabilities is central to I-WLP. Both the network and the cloud are structured into several layers, each of which represents services, resources and capabilities to the layers above. I-WLP incorporates models for every layer of the CSP's offering. Each layer is life-cycle-managed in the stack, which means that if a higher layer requires reconfiguration of a lower layer the reconfiguration steps are included in a technology-agnostic homing and assignment recommendation. For example, if a service requires additional bandwidth between two datacenters then a VPN connection between the two datacenters can be reconfigured to support those needs.

I-WLP is also invoked by the service orchestrator during service instance design, resulting in a homing and assignment recommendation that is processed by a workflow engine with various southbound adapters.

I-WLP integrates HCP domains into CSP resource and capability pools. It can design and assign network services and applications for co-location, such as dual-stack deployments. Since I-WLP automates the homing and assignment process, the information that would be required for manual homing and assignment is unnecessary. This makes it possible to keep the exposure, APIs and interaction with consumers at a higher level. Intent-based service orchestration, where homing and assignment is derived from service-level requirements and constraints, is an excellent example of this.

Looking ahead, we are investigating how to extend intent-based operation [2] to the full stack and how to provide I-WLP for the network compute fabric [3].

Service exposure

Key components such as partner and supplier management, service and resource orchestration, network and infrastructure leverage exposure capabilities to provide services to customers, partners and application developers. Service exposure provides a uniform way to handle the service APIs (represented by the green and purple arrows in Figures 1, 2 and 3) and user access rights, as well as enabling the delivery of network services in combined offerings to enterprises, partners and application developers.

Additionally, the exposed services are used to influence network connectivity characteristics such as QoS parameters, charging conditions, security settings, network-slicing selection policy rules and traffic routing to selected edge sites to support deployments at the specific location of latency-sensitive applications. In the latter case, we encourage the use of the network slicing signaling to user equipment (UE) based on the URSP paradigm, which introduces rule-based mechanism to separate the Protocol Data Unit (PDU) session traffic toward different edge or non-edge sites.

Industry alignment on the required UE exposure capabilities is important to enable the more

SERVICE EXPOSURE PROVIDES A UNIFORM WAY TO HANDLE THE SERVICE APIs AND USER ACCESS RIGHTS

sophisticated routing policies needed for more advanced edge and network slicing use cases.

Managing QoS parameters in mobile networks is important from multiple angles. Edge computing applications (such as automotive and gaming) are sensitive to latency and therefore often require specific QoS conditions at given locations. User mobility and uneven radio conditions at different physical locations may also affect application performance. The 3GPP defines different QoS classes and parameters to serve different kinds of traffic [4, 5]. The enforcement of these rules is executed on the user plane and controlled through the 5G system control plane.

Rather than using URSP-based routing, most of the enterprise use cases have the IP anchor point for all traffic on a selected edge. The location of the IP anchor point is based on the principle that the mobile network selects the breakout point for the mobile device. The application layer selects the application server in the edge cloud, especially in the case of applications running in a third-party cloud, as shown in Figure 2 and Figure 3. The application layer must locate the mobile device and request the network to select the correct breakout point – that is, the point where the application traffic leaves the mobile network administrative domain. After leaving the network domain, the application traffic must be directed to the optimal application server. In the case of data centers provided by HCPs, this process is governed by internet rules and generally executed by HCPs' Domain Name System (DNS) service (Amazon Route 53 or Google DNS, for example).

For the partner-edge go-to-market scenario, service and resource orchestration offers services that enable application developers to take advantage of HCPs' standardized development environments and CSP-provided services to build applications that can utilize network insights and drive network optimization in compliance with enterprise and CSP requirements and use cases. It also offers enterprises, partners and integrators access to services in a simplified way without the need to understand complex 3GPP APIs.

Conclusion

Service exposure is essential for communication service providers (CSPs) to enable close collaboration with enterprises and application developers. By combining service exposure with automated life-cycle management, Ericsson has made it possible for CSPs to establish new revenue streams through mutually beneficial relationships with hyperscale cloud providers. Powerful service and resource orchestration functionality is particularly critical to the automation of 5G services in hybrid, public and private clouds. With the help of service exposure and automated life-cycle management capabilities, CSPs can significantly enhance their ability to capitalize on the rapid digitalization of business and society.

Further reading

- **Ericsson Technology Review, The future of cloud computing: Highly distributed with heterogeneous hardware, May 12, 2020, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/the-future-of-cloud-computing*
- **Ericsson Technology Review, Creating the next-generation edge-cloud ecosystem, February 18, 2020, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/next-generation-cloud-edge-ecosystems*
- **Ericsson Technology Review, Service exposure – a critical capability in a 5G world, May 7, 2019, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/service-exposure-a-critical-capability-in-a-5g-world*
- **Ericsson Technology Review, Distributed cloud – a key enabler of automotive and industry 4.0 use cases, November 20, 2018, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/distributed-cloud*
- **Ericsson, Service automation, available at:** *https://www.ericsson.com/en/service-orchestration/service-automation*
- **Ericsson, Service orchestration, available at:** *https://www.ericsson.com/en/service-orchestration*
- **Ericsson, Network automation, available at:** *https://www.ericsson.com/en/network-automation*

References

1. **Ericsson white paper, Edge computing and deployment strategies for communication service providers, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/edge-computing-and-deployment-strategies-for-communication-service-providers*
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. **Ericsson Technology Review, The network compute fabric – advancing digital transformation with ever-present service continuity, June 30, 2021, Sefidcon, A; John, W; Opsenica, M; Skubic, B, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/network-compute-fabric*
4. **3GPP, Specification 29.122, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3239*
5. **3GPP, Specification 23.502, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145*

THE AUTHORS

Malgorzata Svensson
◆ is an expert in operations support systems (OSS). She joined Ericsson in 1996 and has worked in various areas within research and development. Svensson has broad experience in business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an M.Sc. in technology from the Silesian University of Technology in Gliwice, Poland.

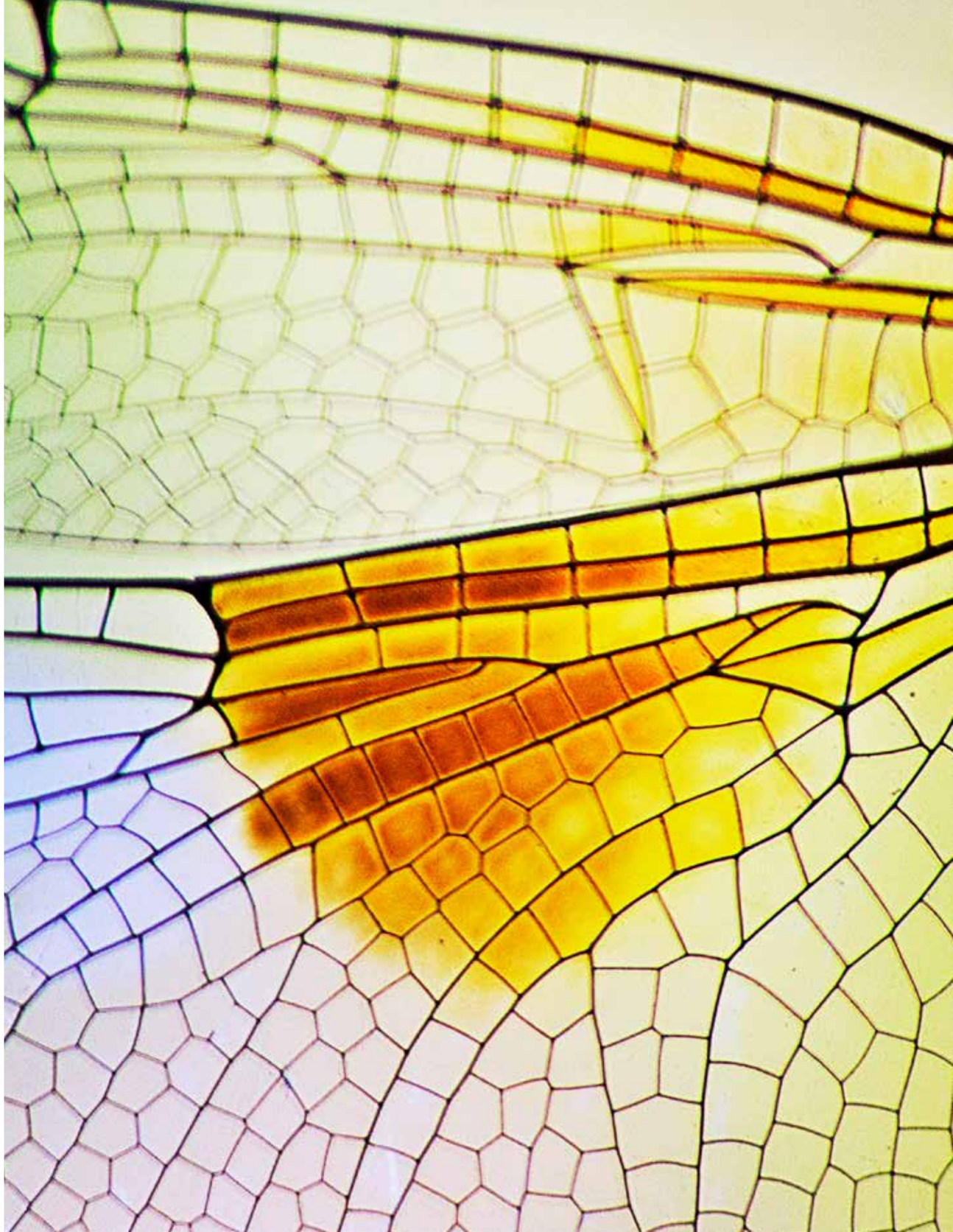
Benedek Kovács
◆ joined Ericsson in 2005 as a software developer and tester. Today his work focuses on 5G networks and distributed cloud, as well as coordinating global engineering projects. Kovács holds an M.Sc. in information engineering and a Ph.D. in mathematics from the Budapest University of Technology and Economics (BME) in Hungary.

Elisabeth Mueller
◆ joined Ericsson in 2006 and has held many business support systems (BSS) related roles in areas including system design, system management and solution architecture. She is currently an expert for BSS E2E systems focusing on 5G/Internet of Things BSS architecture. Mueller holds several patents as well as an M.Sc. in mathematics from Johannes Gutenberg University in Mainz, Germany.

Massimiliano Maggiari
◆ is an Ericsson fellow and a senior expert in OSS architecture. Since joining the company in 2006, he has held many roles across product development and product management in the OSS domain. Maggiari holds numerous patents related to OSS and control plane-based networking, as well as an M.Sc. in electronic engineering from the University of Genoa, Italy.

Róbert Szabó
◆ joined Ericsson in 2013 and currently works as a principal researcher at Research Area Cloud Systems and Platform, where he focuses on distributed cloud, zero-touch automation and Network Functions Virtualization. Before joining Ericsson, he worked as an associate professor at BME in Hungary. Szabó holds both a Ph.D. in electrical engineering and an MBA from BME.

The authors would like to thank the following people for their contributions to this article: Carlos Bravo, Ignacio Más, Stephen Terrill, Magnus Buhrgard, Patrick Maguire, Christer Jakobsson and Jan Backman.



5G EVOLUTION TOWARD 5G ADVANCED:

An overview of 3GPP releases 17 and 18

Together with enhancements aimed at existing use cases such as mobile broadband, industrial automation and vehicle-to-everything, 3GPP release 17 introduces support for new ones including public safety, non-terrestrial networks and non-public networks. Meanwhile, the early planning of release 18 indicates that it will significantly evolve 5G in the areas of artificial intelligence and extended reality.

IMADUR RAHMAN, SARA MODARRES RAZAVI, OLOF LIBERG, CHRISTIAN HOYMANN, HENNING WIEMANN, CLAES TIDESTAV, PAUL SCHLIWA-BERTLING, PATRIK PERSSON, DIRK GERSTENBERGER

The 3GPP has passed the midpoint in its work on its release 17 (Rel-17) specifications, with plans to publish them at the end of the first quarter of 2022. Meanwhile, the discussions on the scope of Rel-18 are well underway. In fact, 3GPP has already announced its decision to recognize Rel-18 as the first release of 5G Advanced to highlight the significant evolution of the 5G System (5GS) that it represents.

■ Several of the features in Rel-17 are intended to enhance network performance for existing services and use cases, while others address new use cases and deployment options. 5G Advanced will build on Rel-17, providing intelligent network solutions and covering numerous new use cases in addition to previously defined use cases and deployment options. *Figure 1* shows Ericsson's view on 3GPP's tentative time plan for releases up until 2028.

One key component of 5G Advanced is the use of artificial intelligence (AI) based on machine learning (ML) techniques. AI/ML is expected to trigger a paradigm shift in future wireless networks. AI/ML-based solutions will be used to introduce intelligent network management and solve multi-dimensional optimization issues with respect to real-time and non-real-time network operation.

AI/ML will also be used to improve the radio interface by further optimizing the performance of complex multi-antenna systems, for example. New use cases such as extended reality (XR) communication will use wireless networks to provide immersive experiences in cyber-physical environments and enable human-machine interactions using wireless devices and wearables.

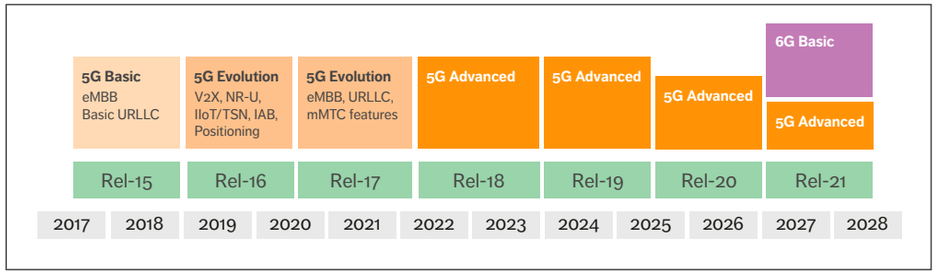


Figure 1 3GPP's 5G evolution tentative time plan

Enhancements in 3GPP release 17

The path toward 5G Advanced begins with Rel-17, which includes significant enhancements to several radio access network (RAN) functionalities that are already deployed in live New Radio (NR) networks.

Beamforming and multiple-input, multiple-output (MIMO)

As shown in *Figure 2*, Rel-17 MIMO enhancements address four areas: beam management; multiple transmission and reception point (mTRP) for ultra-reliable, low-latency communication (URLLC); mTRP for enhanced mobile broadband (eMBB); and TDD and FDD reciprocity.

The multi-beam enhancements are intended to improve performance at high mobility by

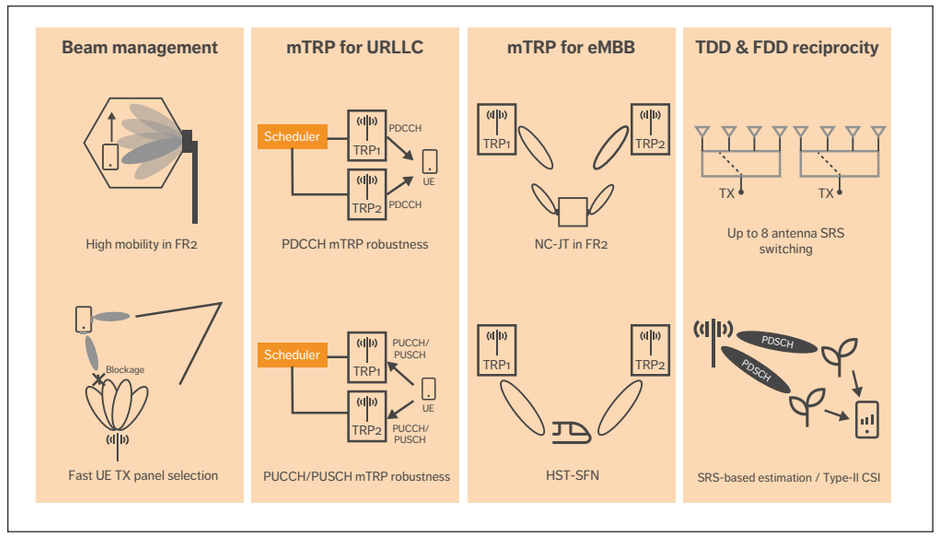


Figure 2 Rel-17 NR MIMO enhancement areas

streamlining signaling and to optimize performance for user equipment (UE) with multiple antenna panels. The mTRP enhancements increase robustness for the physical downlink control channel (PDCCH), physical uplink shared channel (PUSCH) and physical uplink control channel (PUCCH). They also enable richer channel state information (CSI) feedback for non-coherent joint transmission (NC-JT) and optimize performance for high-speed-train (HST) communication scenarios.

Finally, the enhancements to reciprocity-based operation include new codebooks with reduced feedback overhead, where partial channel knowledge is available at gNodeB (gNB), as well as improvements to the Sounding Reference Signals (SRSs).

Dynamic spectrum sharing

The dynamic spectrum sharing (DSS) included in Rel-15 already makes it possible to deploy an LTE cell and an NR cell on the same base station using shared spectrum, which enables an operator to provide 5G services by initiating a migration of spectrum from LTE to NR. Rel-16 primarily improved the capacity of the NR physical downlink shared channel (PDSCH). Enhancements in Rel-17 make it easier for operators to overcome PDCCH resource shortages in the NR cell, which can occur as the number of NR UEs increases. From Rel-17 onwards, cross-carrier scheduling allows for the data channels to be scheduled on the shared primary cell using the PDCCH of a downlink secondary cell.

User equipment power savings

Rel-17 includes power-saving enhancements for UEs in Radio Resource Control (RRC) connected, idle and inactive modes. Power-efficiency improvements are specified both for eMBB UEs and reduced-capability (RedCap) devices. The list of power-saving enhancements includes relaxed radio resource monitoring for devices operating at low mobility or in very good radio conditions, extended discontinuous reception (eDRX) for latency-tolerant devices, reduced PDCCH monitoring during active time, and power-efficient paging reception.

Positioning

NR has supported positioning since Rel-15 through the use of LTE positioning (for non-standalone deployments) and radio-access technology (RAT) independent positioning (Bluetooth, wireless LAN, pressure sensors and so on). Rel-16 introduced time-based positioning methods for NR standalone deployments (multi-round-trip time (RTT), Downlink and Uplink Time Difference of Arrival), as well as an angle-of-arrival and angle-of-departure-based positioning measurements, which can be used in combination with timing-based solutions to achieve higher accuracy.

In Rel-17, NR positioning is further improved for specific use cases such as factory automation by targeting 20-30cm location accuracy for certain deployments. Rel-17 also introduces further enhancements to latency reduction to enable positioning in time-critical use cases such as remote-control applications.

Aside from high-positioning accuracy, industrial Internet of Things (IIoT) and automotive use cases also demand integrity protection of the location information. From a higher layer point of view, Rel-17 introduces key performance indicators to indicate the reliability/integrity of the measurement report limited to the global navigation satellite system (GNSS) positioning procedure.

Ultra-reliable, low-latency communication

URLLC has been a key enabler for the 5GS to enter various verticals. Rel-15 established a solid foundation, and Rel-16 introduced further enhancements by the 3GPP's System Architecture (SA) and RAN groups to better serve various industry verticals such as factory automation, the transport industry and electrical power distribution. These enhancements included various user-plane redundancy schemes as well as enhancements to improve reliability, reduce latency and support time-sensitive communication (TSC).

The enhancements in Rel-17 aim to improve spectral efficiency and system capacity, support URLLC in unlicensed spectrum environments and strengthen the framework to support TSC. They include Hybrid Automatic Repeat Request-Acknowledgement (HARQ-ACK) enhancements, CSI enhancements, intra-UE multiplexing, time-synchronization enhancements and service survival time as an extension to TSC assistance information.

> IN REL-17, NR POSITIONING IS FURTHER IMPROVED FOR SPECIFIC USE CASES SUCH AS FACTORY AUTOMATION

NR coverage

The direct impact that coverage has on service quality, opex and capex makes it a key factor for both commercialization and competition. In Rel-17, the 3GPP has identified the PUSCH as a potential coverage bottleneck. To improve PUSCH coverage, the 3GPP is considering mechanisms for repetition and support for transport block processing over multiple slots. Moreover, Rel-17 specifies mechanisms to support demodulation reference signal (DMRS) bundling across PUSCH repetitions and signaling support for dynamic PUCCH repetition factor indication.

Small data transmission

To support power-efficient connection establishment, the existing NR RRC inactive mode enables a UE to resume a previously established RRC connection. To further enhance the UE power consumption at system access, Rel-17 specifies support for data transmission in RRC inactive mode. Not having to resume an RRC connection reduces the control plane signaling overhead, which is especially relevant for low-power devices that support traffic characterized by infrequent and small data transmissions.

Non-public networks

In Rel-16, the 3GPP specified support for non-public networks (NPNs), which provide access that is limited to a certain group of users such as the devices belonging to a given factory. To provide full support for industrial verticals, the 3GPP specified support for two NPN deployment options. The first, known as public-network-integrated NPNs, allows public operators to support NPNs by associating them directly to their networks. The second deployment

The 3GPP and 5G

The 3GPP published the first versions of the 5G standard in 2018. The work and specifications are divided into three main areas: System Architecture, Core and Terminal, and RAN. The 5G RAN is also known as NR (New Radio) and is part of the 5th Generation System.

The 3GPP organizes its work in releases with a continuous numbering scheme. The first version of the 5G specifications surfaced in 3GPP Rel-15 in 2018 and provided the base functionality as well as a large set of optional features. In subsequent releases, the 3GPP has added new functionality to the existing baseline. This is done with backwards compatibility, so that older terminals can still function in upgraded networks and vice versa.

The 3GPP adds functionality that is required to satisfy increasing demands on existing services (higher data rates for mobile broadband, for example) or to satisfy requirements of new services, use cases and deployment options (such as public safety applications and relaying). However, features are typically specified in a service- and use-case agnostic manner, which means that it is up to vendors and operators to decide how to use and combine the specified features.

> A KEY ASPECT OF 5G NR IS THE CONTINUOUS DRIVE TO SUPPORT NEW VERTICALS AND DEPLOYMENT SCENARIOS

option is known as standalone NPN (SNPN). Broadly speaking, an SNPN has the same functionality and characteristics as a regular public network.

The 3GPP provides further enhancements for SNPNs in Rel-17. These enhancements include support for a UE accessing an SNPN using external credentials (such as those from a public network or those belonging to another SNPN), SNPN UE onboarding (to provision the UE with new NPN credentials and/or subscription parameters, for example) and support for emergency services.

Edge computing

Edge computing, which enables operator- and third-party services to be hosted close to the UE's access point of attachment, was supported in the initial 3GPP Rel-15 of 5GS. The baseline architecture enables efficient service delivery by reducing end-to-end latency and load on the transport network.

Rel-17 introduces mechanisms to discover edge application servers. For example, it defines an Edge Application Server Discovery Function (EASDF) primarily to support the session breakout connectivity model. The EASDF acts as a Domain Name System (DNS) resolver to the UE and can complement the DNS queries with UE location-related information. This enables the DNS system to resolve to application servers close to the UE location.

Rel-17 also clarifies and enhances the use of UE route-selection policy (URSP) rules for edge computing for the distributed anchor and multiple Protocol Data Unit (PDU) session connectivity models. The URSP rules configuration in the UE can take specific application server information into account. This in turn provides the UE with the ability to dynamically establish PDU sessions for specific application servers, eliminating the need to deploy support for complex session breakout solutions.

Furthermore, Rel-17 defines enhanced support for the relocation of the application server in case of UE mobility and includes new mechanisms to expose QoS monitoring results.

Data networks analytics

Several architectural enhancements and newly defined types of analytics in Rel-17 increase the scope and usability of network data analytics. Support for the aggregation of analytics enables use cases where a Network Data and Analytics Function (NWDAF) is able to collect data and analytic reports from other localized NWDAFs. The NWDAF has been disaggregated into two separate logical entities, which enables multiple NWDAF (analytics logical function) in the network to produce analytic reports according to a model distributed from the NWDAF model training logical function (MTLF).

Rel-17 also optimizes some procedures by including new network functions (NFs) that make it possible to process data closer to the data sources and enable lower signaling. The new data collection coordination function enables a single collection of data from 5G Core NFs, with the data being distributed by a non-standardized message bus. The analytics data repository function can store massive amounts of both data and analytic reports. Rel-17 also enhances the input to analytic reports by enabling the addition of information originating from UE applications.

New features in 3GPP release 17

A key aspect of 5G NR is the continuous drive to support new verticals and deployment scenarios. Rel-17 strengthens 5G support for new use cases primarily through new development in five areas: RedCap UE, non-terrestrial networks, frequency bands beyond 52GHz, and the multicast and broadcast service (MBS).

Reduced-capability user equipment

To further widen the range of use cases for NR, Rel-17 introduces support for RedCap UE. RedCap UE will fulfill service requirements somewhere in between the relaxed massive machine-type communication (mMTC) requirements and highly stringent URLLC requirements, as shown in *Figure 3*. RedCap UE provides performance comparable to Rel-8 LTE UE but with additional benefits such as improved latency and the capability to operate in NR frequency bands ranging all the way up to 52GHz.

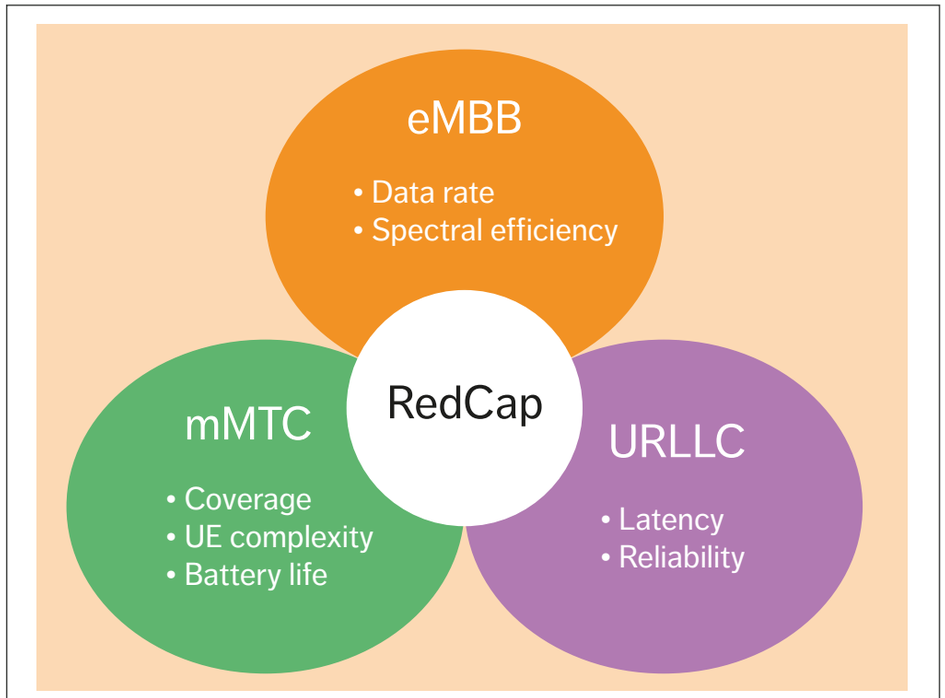


Figure 3 Rel-17 RedCap targets the requirement space between eMBB, mMTC and URLLC

RedCap UEs are significantly less complex than regular NR UEs. This is thanks to a reduced number of radio receiver (RX) antenna branches, reduced RX and radio transmitter (TX) bandwidth and half-duplex operation, meaning that the UE is not required to transmit and receive at the same time. The reduced complexity is anticipated to result in a reduced device price point that will support the use of NR in new applications such as industrial sensor networks. The support of a single antenna branch will facilitate more compact device form factors, which is critical in popular wearable applications such as smart watches.

Non-terrestrial networks (NTN)

The NTN work in Rel-17 introduces new network topologies into the 3GPP specifications. These topologies are based on high-altitude platforms and low Earth orbit (LEO) and geosynchronous orbit satellites. NTN complements terrestrial networks with network coverage in remote areas over sea and

> THE MBS SUPPORT IN REL-17 REQUIRES SIGNIFICANTLY LESS OPERATIONS, ADMINISTRATION AND MAINTENANCE EFFORT

land where terrestrial coverage is absent. The work done by the 3GPP addresses NR, Narrowband-Internet of Things (NB-IoT) and LTE for Machine Type Communication (LTE-M), and it will thereby facilitate 3GPP NTN-based MBB and massive IoT services from Rel-17 onwards.

Rel-17 work builds on earlier studies performed in Rel-15 and Rel-16, where NTN channel models and necessary adaptations of the NR technology to support NTN were identified. The main challenges identified in Rel-16 and addressed in Rel-17 are related to the mobility and orbital height of the satellite. The height causes a high path loss and a large RTT. The mobility of an LEO satellite introduces a very high Doppler offset on the radio link, and it also inevitably requires all devices to frequently change their serving nodes. Rel-17 establishes basic mechanisms to manage these challenges and provides a first set of specifications to support NTNs based on NR, NB-IoT and LTE-M.

NR beyond 52.6GHz

Rel-16 supports operation in frequency range (FR) 1 and 2 covering the ranges 410MHz–7.125GHz and 24.25GHz–52.6GHz, respectively. In Rel-17, FR2 is extended beyond 52.6GHz all the way up to 71GHz using the existing NR downlink/uplink waveforms with the purpose of encompassing new licensed and unlicensed frequency bands in this range.

Operation in these bands does, however, affect several parts of the NR radio. It impacts the signal phase noise characteristics, the transmitter linearity, power efficiency and the receiver noise figure, among other things. However, the 3GPP has concluded that the use of new, advanced phase noise cancellation algorithms will make the Rel-15 physical layer (that is, the existing phase tracking reference signal and sub-carrier spacing of 120kHz) sufficiently robust to support this frequency range. Increased sub-carrier spacing of up to 960 kHz is still specified to allow the 3GPP to exploit even wider carriers of up to 2GHz and thereby unlock a new range of data rates.

Multicast and broadcast service

The MBS support in Rel-17 requires significantly less operations, administration and maintenance effort than its 4G predecessor, Evolved Multimedia Broadcast Multicast Service, as well as improving resource efficiency. 5G MBS is primarily intended to support important uses cases for public safety such as mission-critical push-to-talk, as well as enabling features like over-the-air software updates and live TV, video delivery and IoT solutions.

The 5G QoS framework is also applicable to 5G MBS traffic. It enables differentiated packet forwarding, which is crucial at high traffic load in the context of applications in the public safety domain.

Rel-17 also enables multicast sessions to UEs in RRC connected state, as well as broadcast sessions to UEs in RRC connected, inactive and idle states. The broadcast support for UEs in inactive and idle states is important to support maximum capacity for the broadcast service. Part of the feature is the support for group scheduling, mobility for service continuity and configurable feedback for reliability when needed.

Aside from those enhancements, to expedite the time to market, the MBS is facilitated by features and functionality that have already been specified. Implementation and configuration in a way that is transparent to the UEs is expected to enable the creation of single-frequency networks (SFNs).

3GPP release 18 – introducing 5G Advanced

The 3GPP RAN standardization team began discussing the scope of Rel-18 in June 2021 at the 3GPP RAN Rel-18 Workshop and aims for approval of the detailed scope by December 2021. Of the more than 500 proposals that were submitted to the workshop, Ericsson has identified what we consider to be the most important highlights and placed them in three categories.

Key enhancements for e-MBB use cases

Three of the most notable Rel-18 additions for eMBB use cases are beamforming/MIMO, mobility enhancements and network power savings.

Advanced antenna systems (AASs) are the main driver for increasing spectral efficiency of wireless networks, and they will continue to evolve due to factors such as enabling layer 1/layer 2 mobility, further improvements of uplink MIMO and enhancements related to fixed-wireless access applications.

DSS is extremely useful when transiting from 4G to 5G and many commercial networks already rely on it. To increase network efficiency during that transition, further enhancements are envisioned, such as improved NR performance when the number of LTE UEs decreases gradually, and reduced impact on NR performance due to interference from LTE broadcast signals.

Rel-18 also includes efforts to explore opportunities to further reduce network energy consumption.

Key enhancements for non-eMBB use cases

The most notable enhancements for non-eMBB applications (such as new or existing verticals) include RedCap, XR and national security and public safety (NSPS).

RedCap UEs are expected to play a significant role in many future applications. Based on Rel-17, Rel-18 RedCap solutions will further reduce device cost and power consumption. Solutions enabling energy harvesting, such as energy-efficient wake-up radios, will be investigated.

In Rel-17, the 3GPP RAN standardization team is studying various forms of augmented reality and virtual reality services and assessing their performance when operating through 5G. The main challenge is to simultaneously provide a very high data rate and low/bounded latency. In Rel-18, the 3GPP RAN group will look into traffic management

Terms and abbreviations

5GS – 5G System | **AAS** – Advanced Antenna System | **AI** – Artificial Intelligence | **CSI** – Channel State Information | **DNS** – Domain Name System | **DSS** – Dynamic Spectrum Sharing | **EASDF** – Edge Application Server Discovery Function | **eMBB** – Enhanced Mobile Broadband | **FR** – Frequency Range | **gNB** – gNodeB | **HST**– High-Speed Train | **IAB** – Integrated Access and Backhaul | **IIoT** – Industrial Internet of Things | **IoT** – Internet of Things | **LEO** – Low Earth Orbit | **LTE-M** – LTE for Machine Type Communication | **MBB** – Mobile Broadband | **MBS** – Multicast and Broadcast Service | **MIMO** – Multiple-Input, Multiple-Output | **ML** – Machine Learning | **MMTC** – Massive Machine-Type Communication | **mTRP** – multiple Transmission and Reception Point | **NB-IoT** – Narrowband-IoT | **NC-JT** – Non-Coherent Joint Transmission | **NF** – Network Function | **NPN** – Non-Public Network | **NR** – New Radio | **NR-U** – NR-Unlicensed | **NSPS** – National Security and Public Safety | **NTN** – Non-Terrestrial Networks | **NWDAF** – Network Data and Analytics Function | **PDCCH** – Physical Downlink Control Channel | **PDSCH** – Physical Downlink Shared Channel | **PDU** – Protocol Data Unit | **PHY** – Physical Layer | **PUCCH** – Physical Uplink Control Channel | **PUSCH** – Physical Uplink Shared Channel | **RAN** – Radio Access Network | **RAT** – Radio-Access Technology | **RedCap** – Reduced Capability | **RRC** – Radio Resource Control | **RTT** –Round-Trip Time | **RX** – Radio Receiver | **SFN** – Single-Frequency Network | **SNPN** – Standalone NPN | **SRS** – Sounding Reference Signal | **TRP** – Transmission and Reception Point | **TSC** – Time-Sensitive Communication | **TSN** – Time-Sensitive Networks | **TX** – Radio Transmitter | **UE** – User Equipment | **URLLC** – Ultra-Reliable, Low-Latency Communication | **URSP** – UE Route-Selection Policy | **V2X** – Vehicle-to-Everything | **XR** – Extended Reality

> 5G ADVANCED WILL ALSO INTRODUCE MORE INTELLIGENCE INTO WIRELESS NETWORKS

for resource-efficient and low-latency radio resource allocation, mobility support with consistent data rates, UE energy-efficient operation compatible with XR traffic and latency requirements.

Aside from automotive and industrial use cases, NSPS is the most prominent new vertical using 5GS. RAN enhancements for the remote control of drones and rogue drone detection are being considered to improve the situational awareness of first responders. Rel-18 will also further improve 5G's support for out-of-coverage scenarios by means of techniques such as UE-to-UE relaying.

Cross-domain functionalities for both MBB and non-MBB use cases

We also want to highlight three cross-domain functionalities that target both MBB and non-MBB use cases: AI/ML for physical layer (PHY) enhancements, AI/ML for RAN enhancements, and full duplex.

It is generally expected that AI/ML can significantly improve PHY performance. The RAN standardization will therefore explore the opportunities by setting up a general framework for AI/ML-related PHY enhancements, including proper AI/ML modeling, evaluation methodologies and performance requirements/testing. A first area for concrete AI/ML enhancement could be on beam management or channel estimation/prediction.

In Rel-17, one of the study items is to identify suitable use cases and corresponding AI/ML-based solutions for RAN. In Rel-18, enhancement for selective use cases from Rel-17 will be taken into the normative phase – that is, efficient traffic steering and load balancing. The focus will be on enhancements to current interfaces in the existing architecture. To incentivize vendor competitiveness, one goal is to ensure that AI models remain implementation-specific.

Despite the practical challenges and unclear performance potential, there is a proposal to study the feasibility of full duplex, where gNBs transmit and receive simultaneously on TDD frequency bands. The study will investigate the achievable gains and their dependency on cross-link interference and self-interference mitigation.

Conclusion

3GPP release 17 builds on previous releases with the aim of improving 5G System performance, supporting new use cases and verticals, and providing ubiquitous connectivity in different deployment conditions and scenarios. In the next phase, release 18 will create 5G Advanced, which will include new solutions and technology components that continue to boost network performance for mobile broadband and verticals.

5G Advanced will also introduce more intelligence into wireless networks by including suitable machine-learning-based techniques in different levels of the network. Future enhancements will also cover a wide variety of new verticals and use cases powered by artificial intelligence/machine learning technologies based on a single platform. As the work progresses, we are committed to ensuring that like 5G, 5G Advanced has the ability to support all use cases from one system design, focusing on forward compatibility and diverse configurability while ensuring maximum simplicity.

Further reading

- **Ericsson blog, Non-standalone and Standalone: two standards-based paths to 5G, July 11, 2019, Ekström, H, available at:** *https://www.ericsson.com/en/blog/2019/7/standalone-and-non-standalone-5g-nr-two-5g-tracks*
- **Ericsson blog, A technical overview of time-critical communication with 5G NR, February 25, 2021, Dudda, T; Shapin, A, available at:** *https://www.ericsson.com/en/blog/2021/2/time-critical-communication--5g-nr*
- **Ericsson blog, 5G positioning: What you need to know, December 18, 2020, Dwivedi, S; Nygren, J; Munier, F; Gunnarsson, F, available at:** *https://www.ericsson.com/en/blog/2020/12/5g-positioning--what-you-need-to-know*
- **Ericsson blog, What is reduced capability (RedCap) NR and what will it achieve?, February 11, 2021, Eric Wang, Y.P; Veedu, SNK; Bergman, J; Höglund, A, available at:** *https://www.ericsson.com/en/blog/2021/2/reduced-cap-nr*
- **Ericsson Technology Review, 5G NR evolution, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/5g-nr-evolution*
- **Ericsson, 5G standardization, available at:** *https://www.ericsson.com/en/standardization/5g*
- **Ericsson, Assessing 5G technology leadership, available at:** *https://www.ericsson.com/en/standardization/leadership*
- **Ericsson, Enable far-reaching performance fast with 5G standalone, available at:** *https://www.ericsson.com/en/ran/5g-sa*

THE AUTHORS

Imadur Rahman
◆ joined Ericsson in 2008. He is a master researcher in Research Area Radio at Ericsson Research in Stockholm, Sweden. He is currently co-manager of the 5G Advanced standardization research project at Ericsson Research. Rahman holds a Ph.D. in wireless communications from Aalborg University in Denmark.

Olof Liberg
◆ joined Ericsson in 2008 and currently leads the company's 3GPP RAN standardization team. He has an M.Sc. in engineering physics from Uppsala University, Sweden.

Christian Hoymann
◆ joined Ericsson Research in 2007 and currently leads a research group at Ericsson Eurolab in Herzogenrath near Aachen, Germany. Hoymann holds a Ph.D. in electrical engineering from RWTH Aachen University.

Sara Modarres Razavi
◆ joined Ericsson in 2014, where she is currently a senior researcher and co-manager of the 5G Advanced standardization research project at Ericsson Research. She holds a Ph.D. in infra informatics from Linköping University, Sweden.

Henning Wiemann
◆ is a senior expert in radio network protocols. He joined Ericsson Research in 2000 after completing a Dipl-Ing. in electrical engineering at RWTH Aachen University.

Claes Tidestav
◆ joined Ericsson in 1999 and specializes in radio resource management, with a special focus on multi-antenna technologies and millimeter wave systems. He holds a Ph.D. in signal processing from Uppsala University.

Paul Schliwa-Bertling
◆ joined Ericsson in 1996 and currently serves as an expert in mobile networks architecture and signaling at Ericsson Research. He has an M.Sc. in electrical engineering from the University of Duisburg-Essen in Germany.

Patrik Persson
◆ is a principal researcher and joined Ericsson Research in 2007. Currently, he has a position as the program manager for the Ericsson Research program on 5G evolution and 6G. Persson holds a Ph.D. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Dirk Gerstenberger
◆ joined Ericsson in 1997 and is currently a manager at the Standards & Technology department within Business Area Networks at Ericsson, working with the evolution of radio-access standards and radio-network deployments. Gerstenberger has a Dipl-Ing. in electrical engineering from Paderborn University in Germany.

Meeting 5G network requirements with Massive MIMO

5G New Radio (NR) has been designed to fully support Massive MIMO as a native technology from the start. The vastly increased coverage, capacity and user throughput that Massive MIMO provides has quickly made it a natural and essential component of cellular network deployments.

DAVID ASTELY, PETER VON BUTOVITSCH, SEBASTIAN FAXÉR, ERIK LARSSON

Massive MIMO (multiple-input, multiple-output) radios are the leading radio solution for new 5G deployments on mid-band TDD spectrum. The ability to use a larger number of radio chains – 16-64, for example, compared with the 2-8 typically used in conventional radio solutions – makes it possible for communication service providers (CSPs) to benefit significantly from multi-antenna techniques.

■ Although Massive MIMO is still a relatively new technology – the term was first coined in academia about a decade ago with the first commercial solutions hitting the market about five years later – the technology has matured rapidly, enabling the creation of cost-efficient solutions that are both small in size and light in weight. Most of the Massive MIMO solutions on the market today have already gone through multiple hardware generations, often with optimized application-specific integrated circuits for beamforming and physical layer processing. They are available in a multitude of different models optimized for a wide variety of deployment scenarios [1, 2, 3].

Network requirements overview

5G networks are expected to outperform today's 4G networks in terms of capacity and user experience to cater for never-ending traffic growth and rising expectations, not only on mobile broadband services but also on new services such as XR (extended reality).

With this in mind, there are three ways a CSP can improve capacity and user throughput: by improving the spectral efficiency of existing frequency bands, by adding new spectrum and by densifying the network with more sites [3]. The high cost associated with acquiring and maintaining new sites means the decision to densify the network is typically only made when the other two alternatives have been exhausted.

Improving spectral efficiency is typically explored first, as it is associated with the least cost. However, additional spectrum will almost certainly be needed to meet 5G performance requirements. Many countries have released substantial amounts of new spectrum for 5G deployments that have the potential to unlock vast amounts of capacity. However, this spectrum is usually on a higher frequency band, such as 3.5GHz, with more challenging radio propagation compared with the frequencies used for 4G. The only way to efficiently use this spectrum on existing sites is with radio solutions that provide improved coverage.

From a user perspective, the requirements on throughput are often similar in all parts of the network. From a network perspective, however, the cells served by different sites may differ greatly in terms of size and traffic load, implying varying requirements on coverage and capacity. Cells with high, medium or low traffic load can be found in all environments.

Furthermore, there are often considerable variations in traffic load for each site over time. The peak-hour load level for each cell together with the expected traffic growth over time set capacity requirements. A site must handle the expected traffic load over the entire investment cycle, which is typically five to seven years.

> NETWORK REQUIREMENTS IN TERMS OF COVERAGE, CAPACITY AND EASE OF DEPLOYMENT VARY FOR DIFFERENT SITES

In addition to the performance requirements, there are deployment-related requirements to consider for some sites. The most important of these constraints are ease of deployment and cost efficiency. Ease of deployment includes aspects such as size, weight and the visual impact of the equipment. Cost efficiency in terms of both capex (that is, the cost of site equipment) and opex (site rental and energy consumption costs, among others) is always important, as the investments that the CSP makes are expected to provide sufficient value. A commonly used metric for cost efficiency is cost per capacity, which provides a trade-off between the cost of the product itself and the value it offers in terms of network performance.

In short, the network requirements in terms of coverage, capacity and ease of deployment vary for different sites in the network. Using the same radio solutions at all sites would be neither cost-efficient nor feasible, which is why different radio solutions are available.

Multi-antenna technologies

Massive MIMO improves network coverage and capacity through the use of the three multi-antenna

Terms and abbreviations

CSI – Channel-State Information | **CSP** – Communication Service Provider | **DL** – Downlink | **EIRP** – Effective Isotropic Radiated Power | **MIMO** – Multiple-Input, Multiple-Output | **MU-MIMO** – Multi-User MIMO | **NR** – New Radio | **RRU** – Remote Radio Unit | **SINR** – Signal-to-Interference-plus-Noise-Ratio | **SU-MIMO** – Single-User MIMO | **TCO** – Total Cost of Ownership | **UE** – User Equipment | **UL** – Uplink

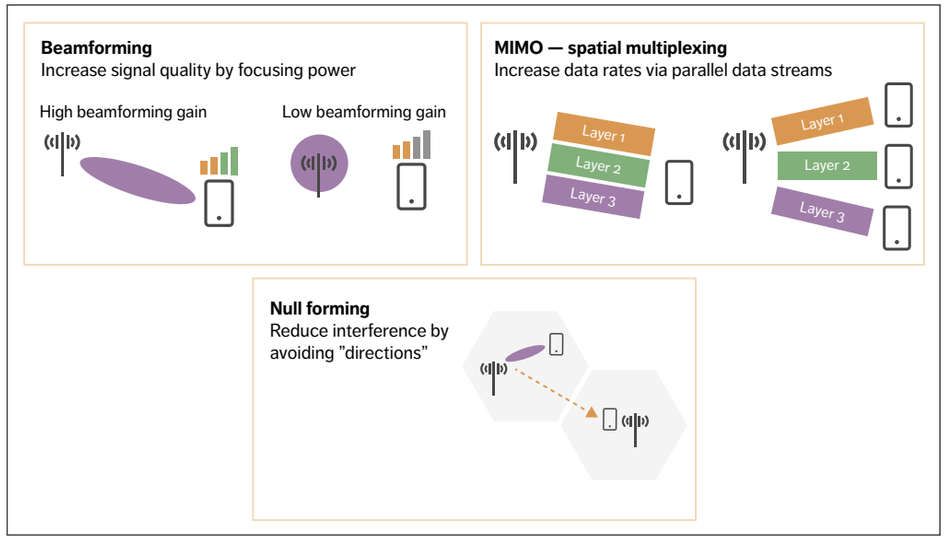


Figure 1 The three main multi-antenna technologies used in Massive MIMO

technologies – beamforming, null forming and spatial multiplexing – shown in *Figure 1*. All three are applicable to both the downlink (DL) and the uplink (UL).

The purpose of beamforming is to amplify transmitted/received signals more in some directions than others. The goal is to achieve a high beamforming gain in the direction of the device of interest to improve link quality in terms of signal-to-interference-plus-noise-ratio (SINR). This translates into higher spectral efficiency and/or better coverage for a single link, which in turn results in better network coverage, capacity and user throughput.

Null forming is a variant of beamforming that strives to lower the beam gain in certain directions or even reduce it to zero. By intentionally creating nulls or lower gain in the directions where the interfered transceivers are, interfering signals can be filtered out, resulting in a lower interference level, higher SINR and higher spectral efficiency.

Spatial multiplexing refers to the technique of multiplexing several data streams, or layers, on the same time-frequency symbol. The multiplexed data streams can all go to the same device or to different devices. Cases in which all the layers belong to the same device are referred to as single-user MIMO (SU-MIMO), while cases that involve spatial multiplexing of multiple devices are called multi-user MIMO (MU-MIMO). Spatial multiplexing can increase the spectral efficiency, which translates into increased user throughput and network capacity.

Beamforming and SU-MIMO are central to Massive MIMO. The ability of beamforming to increase the received signal level while not increasing the average interference level is key to obtaining high performance. Substantial beamforming gains can be achieved in a wide range of situations, regardless of DL/UL, traffic load, or if the user is in a good or bad spot. Coverage, capacity and user throughput are generally improved.

A particularly important strength of beamforming is its ability to increase DL and UL coverage, hence extending the area where users can benefit from TDD mid-band deployments based on reuse of the existing site grid dimensioned for 4G FDD deployments.

Spatial multiplexing with SU-MIMO benefits from high signal levels. Beamforming helps to improve signal levels, which can then be exploited for single-user spatial multiplexing. Particularly in the DL, more than one layer to a specific user can often be supported in large parts of a cell. This contributes to its general applicability.

MU-MIMO improves performance at high traffic loads and in good channel conditions. These are conflicting requirements, as high traffic loads often lead to higher inter-cell interference levels, which means worse channel conditions. Compared with SU-MIMO, there are considerably more requirements on MU-MIMO to reach meaningful performance improvements. MU-MIMO is nevertheless a great capacity enhancement tool for highly loaded cells.

Intentional null forming to selected users serves to reduce interference to those users. It is a key sub-component of MU-MIMO to mitigate intra-cell interference, and it is also commonly used on the receiver side in both the UL and the DL to suppress inter-cell interference.

Massive MIMO features

All Massive MIMO solutions consist of both hardware (one or more Massive MIMO radios) and software (Massive MIMO features). A Massive MIMO feature can be described in terms of three factors [3]:

1. The network requirement(s) that the Massive MIMO feature is intended to meet
2. The available channel knowledge
3. The multi-antenna technique (or combination of techniques) that can be applied using the channel knowledge gathered in #2 to meet the requirement(s) in #1.

Different permutations of these three factors will yield unique Massive MIMO features – potentially with varying trade-offs and applicability to different conditions.

Firstly, it is essential to be clear about the requirement(s) that the feature is intended to meet – should it improve coverage, boost capacity or increase throughput? In some cases, one feature can solve multiple problems, while in others, trade-offs may be necessary. A feature that improves energy efficiency may have a negative effect on capacity, for example. It is therefore essential to assess which performance requirements are most important for a certain cell at a certain time. For instance, during off-peak hours, the capacity demand in a cell may be low, making it acceptable or even desirable to apply a feature that sacrifices capacity to improve energy efficiency.

All Massive MIMO features result from applying a combination of the three basic multi-antenna techniques – beamforming, null forming and spatial multiplexing – to a physical channel or signal, using available channel knowledge to solve a certain problem. This may sound simple, but there are several aspects to consider, resulting in a wide variety of potential features. A central question is how to acquire the channel knowledge required to perform beamforming, null forming or spatial multiplexing. This can be achieved in several ways, but it is important to understand that there is always a cost associated with acquiring channel-state information (CSI). Increased overhead is just one example.

There is also a problem of CSI availability. Different sounding and feedback methods are available in the 3GPP standard, and different user equipment (UE) may have different capabilities and support different CSI feedback and sounding formats. The network must therefore support several Massive MIMO features in parallel. Even if a UE supports a certain

> ALL MASSIVE MIMO SOLUTIONS CONSIST OF BOTH HARDWARE (MASSIVE MIMO RADIOS) AND SOFTWARE (MASSIVE MIMO FEATURES)

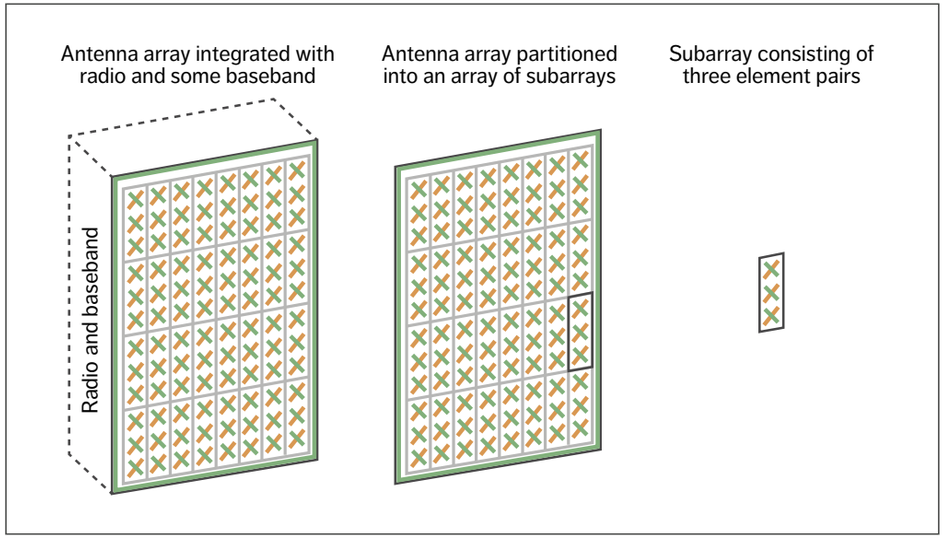


Figure 2 Massive MIMO radio, antenna array and subarray

CSI feedback and sounding formats, that CSI may not be available at a certain instance in time. For example, when a UE first connects to a cell, no channel information is generally available and measurement or sounding configurations will need to be set up, implying that there is a lead time before such CSI is available to the network.

Different sets of MIMO features are needed when limited/no CSI is available, compared with when CSI is available. Massive MIMO features can be classified at a high level as using either feedback- or sounding-based channel information and employing either SU-MIMO or MU-MIMO. In practice, there are many options for how to implement these aspects of a feature, both from what is available in the 3GPP standard and from a proprietary algorithm perspective.

By comparing the Massive MIMO features with respect to the network key performance indicators of interest (coverage, capacity and user throughput) they exhibit different strengths and weaknesses. Feedback-based beamforming has an advantage in coverage over sounding-based beamforming. Similarly, SU-MIMO has a coverage advantage over MU-MIMO. This is because MU-MIMO requires more detailed CSI and because MU-MIMO needs to split the available transmit power between multiple users. To fully utilize the potential of a Massive MIMO solution, it is necessary to dynamically adapt/switch the algorithm so that coverage, capacity and peak rate can be maximized jointly, which is how Massive MIMO solutions are typically designed.

Massive MIMO radios

Unlike conventional solutions in which separate basebands are connected to remote radio units (RRUs) that are connected to separate passive antennas, Massive MIMO radios integrate the radio, the antenna and some baseband functionality in the same unit. The reason for constructing Massive MIMO radios in this way is to avoid the need for them to support very high data rates on the interface between the radio and the baseband. *Figure 2* illustrates a Massive MIMO radio with an integrated antenna array that is partitioned into multiple subarrays.

There are two main characteristics of the antenna that have an impact on the properties of the MIMO radio solution. The first is the total antenna array size: the maximum antenna gain is proportional to the total antenna array size. The second is how the antenna array is partitioned into subarrays. Each subarray is controlled individually using a pair of radio chains. The finer the partition (that is, the smaller the subarrays), the better the steerability. A finer partition also results in higher cost and greater complexity, however, as more radio chains are needed for the same size of array.

A key factor affecting the effectiveness of subarray partitioning is the deployment scenario, and in particular the angular spread of users. As users typically are distributed uniformly in the horizontal domain, it follows that a fine partitioning in the horizontal domain, offering superior horizontal domain beamforming, is beneficial. The spread of users in the vertical domain is, however, highly scenario dependent.

In dense, urban, high-rise deployments, there may be a significant spread of users in the vertical domain – that is, the cells could be almost as tall as they are wide. In these scenarios, vertical-domain beamforming that features short subarrays and many radio chains offers performance benefits. However, in other, more suburban or rural deployments where the spread of users in the vertical domain is smaller, taller subarrays and fewer radio chains are most likely to offer competitive performance, making this solution the more cost-efficient choice.

In general, there are several important design parameters to consider when choosing between different radio solutions:

1. Radio parameters, such as the number of radio chains, output power, bandwidth and the number of frequency bands
2. Antenna array characteristics such as antenna size and subarray structure
3. Cost-efficiency and form-factor parameters such as size and weight.

> A CSP TYPICALLY WANTS TO MAXIMIZE THE USE OF THE AVAILABLE SITES BEFORE ACQUIRING NEW ONES

All of these parameters are essential when selecting a Massive MIMO radio solution with suitable characteristics for the different parts of the network.

Selecting the appropriate radio solution – guiding principles

For many CSPs, the first and most cost-efficient way of evolving their networks to meet network requirements is to deploy all available spectrum, including new 5G mid-band spectrum. This will unlock substantial capacity and provide a superior consumer experience. Performance can then be further differentiated by the choice of radio solution, either Massive MIMO or conventional, along with software features. Coverage of new 5G mid-bands is often more challenging than for existing 4G bands, which gives Massive MIMO, with its superior beamforming capabilities, an advantage over RRUs.

To ensure the selection of the appropriate radio solution, a CSP should begin by reviewing the existing network assets along with the strategies that it has put in place to meet its unique set of business objectives. Existing network assets – spectrum, sites and equipment – are central to determining how to evolve a network. Spectrum is the most valuable asset for a CSP, as it directly affects the achievable network capacity and consumer experience. Radio sites constitute another important asset that are often difficult and expensive to acquire and maintain. Therefore, a CSP typically wants to maximize the use of the available sites before acquiring new ones.

With respect to strategy, a CSP must consider any strategies it has that relate to the question of which services to offer, what QoS to offer, and where in the network to offer these services. The requirements of different services can be mapped to network

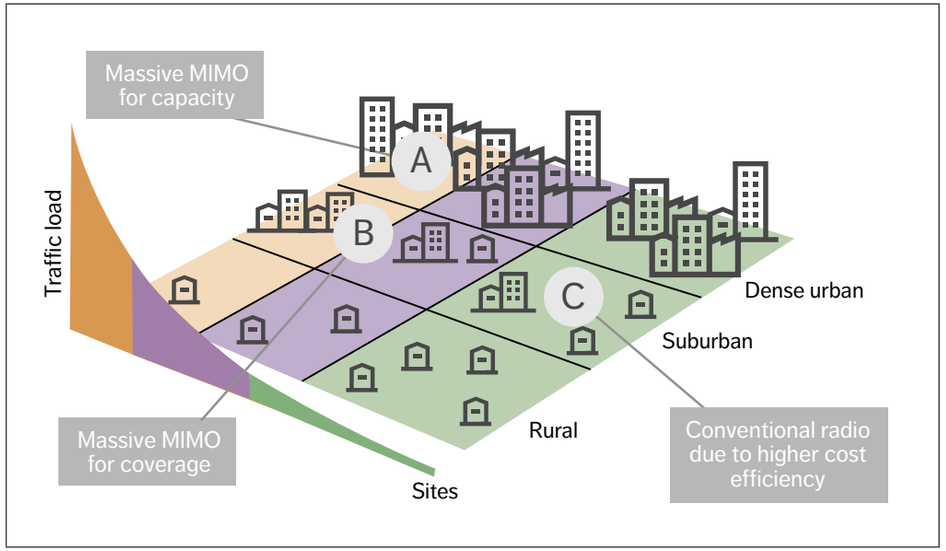


Figure 3 Suitable radio solutions when adding 5G mid-band spectrum at three site locations

requirements in terms of capacity and user throughput.

Once the inputs about existing network assets and strategy have been gathered, the next step is to do an analysis to find the performance requirements and constraints relevant for each site in the network. The answers to the questions about services (which, what and where) can be translated into specific network requirements. The current network traffic load and the predicted traffic growth including old and new services provide input on the maximum traffic volume (capacity) the network must support.

The deployment environment also has a profound impact on which radio (hardware and software) characteristics pay off in terms of network performance. For many sites, there are requirements on physical size, ease of installation, small visual impact and so on. These requirements may be driven by factors such as site building constraints, esthetical requirements, wind load and site accessibility, and in some cases, they are the determining factor in the selection process.

Based on an assessment of the total cost of ownership (TCO) per capacity for the whole investment cycle (typically five to seven years), the final step in the decision-making process involves choosing radio and feature solutions that meet the requirements and constraints for each site over the investment cycle. The radio solution toolbox includes both Massive MIMO and conventional radios. Feature solutions can be implemented more gradually than radio solutions, in response to emerging requirements.

Network evolution example

Figure 3 shows a network evolution example in which 5G mid-band spectrum is added at three different site locations – sites A, B and C. The colors orange, purple and green represent the traffic load levels high, medium and low, respectively.

Site A is in an area with high traffic load and high traffic growth, and there are no deployment restrictions with respect to size, weight and so on. To unlock the full potential of the 5G mid-band spectrum in this scenario, we would recommend the use of a high-end Massive MIMO radio that provides large bandwidth, high effective isotropic radiated power (EIRP) and many radio branches facilitating superior horizontal- and vertical-domain beamforming. Vertical-domain beamforming is motivated, as the UE distribution in the vertical domain is large. From a feature perspective, all available capacity-enhancing features should be deployed.

Site B is in a suburban area with a large inter-site distance, high traffic load and high expectations on traffic growth. As in site A, in this scenario we would recommend a high-end Massive MIMO product that provides large bandwidth and high EIRP to meet coverage and capacity requirements. Unlike site A, however, the UEs in site B are confined to a small angular area in the vertical domain. Therefore, a product supporting less vertical-domain beamforming (that is, fewer radio branches) would be sufficient.

Site C is in a low-traffic suburban area with low traffic growth where ease of deployment is an important factor. The latter point calls for a small radio solution, while the former indicates that a low-end radio offering less capacity would still meet the requirements. A conventional radio solution with few radio chains would therefore be a cost-efficient alternative to a Massive MIMO solution in this scenario.

Conclusion

Massive MIMO (multiple-input, multiple-output) technology boosts spectral efficiency through the use of multi-antenna technologies, which results in significantly increased network coverage, capacity and user throughput. Most 5G mid-band TDD deployments today incorporate Massive MIMO technology to unlock the full potential of new spectrum without the need for site densification. Massive MIMO radios have matured quickly and become competitive in terms of size, weight and cost.

“MASSIVE MIMO UNLOCKS THE FULL POTENTIAL OF NEW SPECTRUM WITHOUT THE NEED FOR SITE DENSIFICATION”

Multiple radio and feature options are available with different characteristics to meet the specific requirements of various types of sites in the network in a cost-efficient way.

Further reading

- **Ericsson blog, How to build high-performing Massive MIMO systems, available at:** *https://www.ericsson.com/en/blog/2021/2/how-to-build-high-performing-massive-mimo-systems*
- **Ericsson, Massive MIMO boosts network throughput, available at:** *https://www.ericsson.com/en/ran/5g-radio-fdd*

References

1. **Elsevier, Advanced Antenna Systems for 5G Network Deployments: Bridging the Gap between Theory and Practice (1st Edition), 2020, Asplund, H, et al., available at:** *https://www.elsevier.com/books/advanced-antenna-systems-for-5g-network-deployments/asplund/978-0-12-820046-9*
2. **Ericsson white paper, Advanced antenna systems for 5G networks, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/advanced-antenna-systems-for-5g-networks*
3. **Ericsson, Massive MIMO handbook, available at:** *https://www.ericsson.com/en/ran/massive-mimo*

The authors would like to thank Henrik Asplund, Thomas Chapman, Mattias Frenne, Christer Friberg, Farshid Ghasemzadeh, Bo Göransson, Billy Hogan, George Jöngren and Jonas Karlsson for their contribution to this article.

THE AUTHORS

David Astely

◆ joined Ericsson in 2001 and has held several positions in both research and product development over the years. He currently serves as a principal researcher at Ericsson Research in the radio area. Astely received his Ph.D. in signal processing from KTH Royal Institute of Technology in Stockholm, Sweden, in 1999.

Peter von Butovitsch

◆ joined Ericsson in 1994 and has held various roles at Ericsson Research and in RAN system design during his time with the company. From 1999 to 2014, he worked for Ericsson in Japan and China. He currently works as a technology manager at Systems & Technology. Butovitsch holds both an M.Sc. in engineering physics and a Ph.D. in signal processing from KTH Royal Institute of Technology. In 2016, he completed an MBA from the University of Leicester in the UK.

Sebastian Faxér

◆ joined Ericsson in 2014 and currently serves as a strategic product manager within Product Line 5G, where he is responsible for Massive MIMO software solutions. He has more than 150 patents and was the recipient of the 2020 Ericsson Inventor of the Year award for his contributions to the design of the 5G NR standard in the Massive MIMO area. He is also coauthor of the book 5G New Radio: A Beam-based Air Interface. Faxér holds an M.Sc. in applied physics and electrical engineering from Linköping University, Sweden.

Erik Larsson

◆ joined Ericsson in 2005. He currently serves as a researcher working with concept development and network performance for 5G with a focus on Massive MIMO. Larsson holds both an M.Sc. in engineering physics and a Ph.D. in electrical engineering, specializing in signal processing, from Uppsala University, Sweden.

End-to-end network slicing orchestration

– A KEY ENABLER FOR INDUSTRY-VERTICAL USE CASES

By automatically matching the particular service requirements of an industry-vertical use case to its specific deployment areas, transport-aware network slicing orchestration makes it possible to ensure end-to-end QoS without over-provisioning.

PAOLA IOVANNA, MALGORZATA SVENSSON, ALEXEY SHAPIN, GIULIO BOTTARI, FABIO UBALDI, FILIPPO PONZINI, MARZIO PULERI

To support use cases with extreme performance requirements in a cost-efficient manner, communication service providers need comprehensive, end-to-end (E2E) network slicing orchestration that covers the full range of network resources – RAN, core, cloud and transport.

■ The acceleration of industrial digitalization and the rise of new applications such as cloud robotics and remote-assisted surgery are leading to a high demand for the new capabilities available in 5G. At the same time, future use cases like massive twinning, immersive telepresence and collaborative robotics are helping to shape the journey toward 6G. There is also a desire to achieve ubiquitous radio access in specific deployment areas, supporting a high density of users and access points to the network. The goal is to achieve all of this without increasing the cost/revenue ratio.

Based on Service Level Agreements (SLAs), new services for industry verticals [1] – including those that demand extreme performance – will require E2E QoS at scale over three main network deployment areas: local, confided wide and general wide. Examples of the local area include indoor locations and campuses, while examples of confined wide areas include ports and railway systems. General wide area refers to larger geographical locations such as cities.

E2E QoS depends on factors such as throughput, latency, availability, reliability and resilience. To ensure service delivery with the required E2E QoS without resorting to over-provisioning of network resources, the corresponding RAN, core and transport resources must be made available in the specific deployment areas. The network must have the ability to support a mix of heterogeneous services in the corresponding deployment areas. Our research shows that by including transport awareness in orchestration and slice operations, it is possible to tailor the network resources to the actual QoS values.

Our end-to-end slice orchestration concept

In a 5G network, there is always at least one default slice and each UE (user equipment) is associated to a slice. The 5G network slicing feature makes it possible to set up independent logical networks on a shared physical and virtual infrastructure. A slice can, for example, ensure ultra-reliable low-latency communication (URLLC) to support a service related to the remote control of robots in a factory. Each slice operates on specific tracking areas (TAs) served by a set of gNodeB base stations along with the Access and Mobility Management Function. This means that each network function can be placed in accordance with both the area and the service conveyed by the related slice.

An industry-vertical service that spans a specific deployment area can be mapped on a slice that includes multiple TAs. The service is characterized by specific E2E QoS requirements that the slice must support within the deployment area. The E2E QoS

“E2E QoS DEPENDS ON FACTORS SUCH AS THROUGHPUT, LATENCY, AVAILABILITY, RELIABILITY AND RESILIENCE”

is defined by the combination of the QoS in the radio layer (RAN/Core Network (CN)) and the QoS in the transport layer. An optimal combination can be achieved by using automatic and dynamic techniques to create smart mapping between the RAN/CN QoS and the transport QoS that takes into account the specific technology of the infrastructure. For example, the 5G QoS Identifier of the RAN could be mapped to the corresponding Differentiated Services Code Point of the transport.

The transport connections are implemented according to the specific transport data-plane technologies such as IP, VPN and VLAN. Because a particular service is available in a specific deployment area, there can be multiple transport connections for it. As a result, it is essential that there is a procedure during service provisioning that automatically maps the E2E QoS parameters (peak rate, guarantee rate, resilience, availability and so on) of the slice on all the available transport connections. A requirement that all the transport connections support the slice peak rate would result in a waste of resources, while splitting the slice peak rate among all the transport connections could adversely affect the service level. An effective solution assigns the

Terms and abbreviations

AC – Admission Control | **AI** – Artificial Intelligence | **AR** – Augmented Reality | **CIR** – Committed Information Rate | **CN** – Core Network | **CNF** – Container Network Function | **E2E** – End-to-End | **eMBB** – Enhanced Mobile Broadband | **HSS** – Home Subscriber Server | **IoT** – Internet of Things | **mMTC** – Massive Machine-Type Communication | **PIR** – Peak Information Rate | **PoP** – Point of Presence | **SLA** – Service Level Agreement | **TA** – Tracking Areas | **URLLC** – Ultra-Reliable Low-Latency Communication | **vEPC** – Virtual Evolved Packet Core | **VNF** – Virtual Network Function

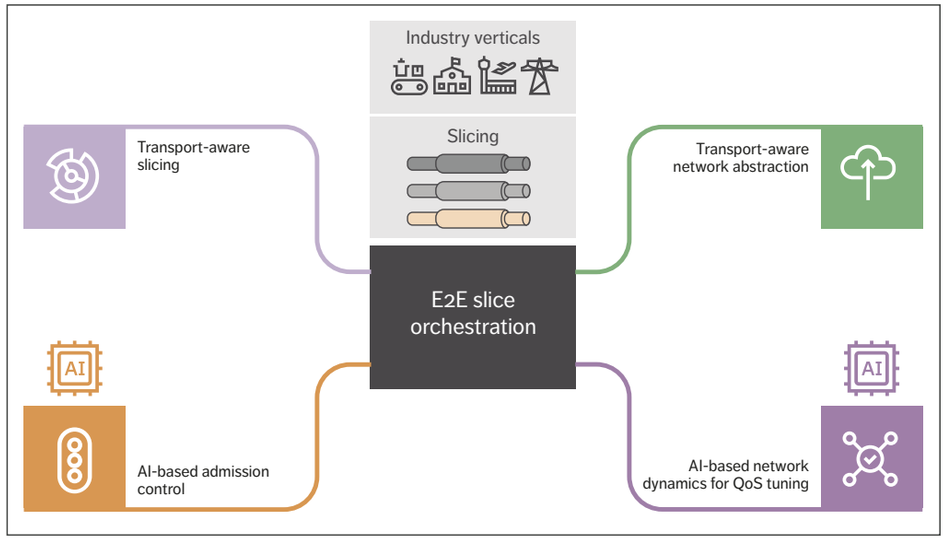


Figure 1 The four core components of Ericsson's E2E network slicing orchestration solution

rate for each connection according to the actual needs.

In fact, the orchestration of all the various infrastructure components should be extended to ensure effective mapping based on actual traffic behavior. The mapping must be fully dynamic and automatic for each specific service, including those that impose a guaranteed and deterministic performance level.

Moreover, each network slice should operate as an isolated E2E network, tailored to fulfill the specific requirements requested by a particular application. The most appropriate type of isolation for a particular slice depends on the transport technology it uses. For example, in the case of optical networks, it is possible to utilize the separation provided by the wavelength channel, while a dedicated scheduler could be used for packet networks. Our E2E orchestration solution bases decisions about isolation on information from the transport domain related to the types of isolation that it can provide.

As shown in *Figure 1*, our transport-aware, E2E network slicing orchestration solution consists of four main components:

- Transport-aware slicing including the interaction between RAN/CN and transport
- Transport-aware network abstraction
- Artificial intelligence (AI) based admission control (AC)
- AI-based network dynamics for QoS tuning.

Transport-aware slicing

One important aspect of network slicing orchestration is to map traffic from a single slice or group of slices to transport resources that match the required E2E QoS for that slice or group of slices. Dedicated transport may also be required when latency is an issue, when there is a need for transport observability per slice, or to guarantee isolation.

Since the transport layer is logically separated from the radio layer and the expected radio needs are known, the standard approach to transport resource allocation in this scenario is to base it on the peak of expected radio needs. The downside of this approach is that it frequently results in the over-provisioning of the transport layer, which may not always be feasible or economically justifiable.

As an alternative to the current approach in which the QoS of the RAN/CN and the QoS of transport are orthogonally associated and independently configured, we have developed an approach that avoids over-provisioning by making the radio layer orchestration aware of the transport resources. The traffic flows for a single slice (or group of slices with heterogeneous SLA needs) are mapped to the most appropriate transport connection in a shared (not dedicated) transport network.

It is important to note that our transport-aware approach requires some changes to current 5G slicing practices. Firstly, it requires that the service and its deployment area are associated to a particular slice. Secondly, it requires that the slice TAs are chosen to cover the deployment areas of the service. Thirdly, the E2E QoS parameters of the service need to be mapped to the corresponding network resources (RAN, CN and transport) associated to the slice by using a RAN/CN and transport abstraction to expose a suitable view of the network resources to the orchestrator.

When all of these conditions are met, a consistent association of the QoS of slices in the RAN/CN and the QOS of transport will be performed automatically, and both layers will be automatically configured.

Transport-aware abstraction

Transport-aware abstraction is a compact description of all network resources (radio, transport and cloud) that expose the corresponding QoS parameters (latency, bandwidth, resilience and so on) to the E2E orchestrator. Abstraction simplifies the resource details (such as quantity, vendors, location of the resource, physical details, real topology and so on) for the E2E orchestrator so that it can consider the essential transport features in a simplified way concurrently with the features of radio and cloud resources.

A network service is constituted by the sequence of virtual network functions (VNFs), physical network functions (PNFs) and container network functions (CNFs). Transport provides the connectivity among them. One of the main challenges is the optimization of resource placement on top of the underlying transport infrastructure. For example, VNFs/CNFs can be connected through a simple point-to-point transport link or, alternatively, through a meshed geographical transport network. These two options imply different latency values or different availability. Knowledge about transport characteristics is particularly relevant in the case of services with critical performance requirements. Transport-aware abstraction provides a flat view of all the network resources, including transport, to facilitate the best resource selection and enable cross-optimization.

By logically separating the service from the infrastructure technologies, the abstraction technique makes independent services from technologies, allowing these two elements to evolve independently. Additionally, the abstraction enables a clear separation of responsibility and roles between the infrastructure provider and the service provider.

Our proposed abstraction models a geographical site as a point of presence (PoP) that is composed of a set of functions. The various PoPs in the network are connected by virtual links that represent the transport connections. In the evolution of 5G toward 6G, some of the RAN functions can be located at geographically distant sites (in the cloud RAN scenario, for example), and some of the CN functions can migrate toward the access site up to the antenna sites. In other words, each site can host a mix of RAN and CN functions. The transport network is abstracted by point-to-point virtual links with associated QoS parameters, including information related to resilience and availability.

Most relevant industry verticals

- Manufacturing
- Energy and utilities
- Transportation
- Health care
- Media and entertainment
- Smart cities
- Governments

> OUR TRANSPORT-AWARE APPROACH REQUIRES SOME CHANGES TO CURRENT 5G SLICING PRACTICES

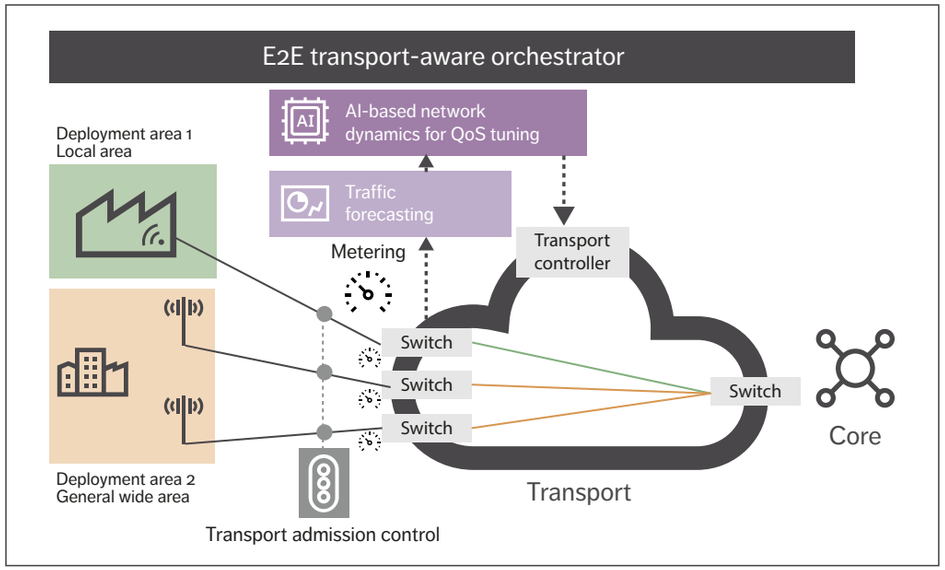


Figure 2 Functional blocks in the reference network

AI-based network dynamics for QoS tuning

Effective QoS tuning is challenging for transport-aware network slicing orchestration. The traffic associated with a specific service will change over the deployment area, both in space and time, which will affect the QoS parameters. Therefore, the assignment of QoS parameters to the transport tunnels supporting a slice should be done dynamically with appropriate mechanisms. In most cases, service traffic is by nature dynamic and at least partially related to predictable situations or historical trends that influence traffic load such as the time of day, weekday versus weekend and the like.

Dynamically tuning the parameters of slices to support current needs is especially useful in cases where the QoS parameters – peak information rate (PIR) and committed information rate (CIR) , for example – that are assigned to the services are not well known in advance and could potentially be overestimated. This plays an important role in the bandwidth partitioning that is needed to support services for industry verticals, by making it possible to do it dynamically according to actual traffic need.

AI-based admission control

The role of AI-based AC is to check whether network resources are available to support the QoS and traffic parameters of an incoming connection. Three types of AC should be considered when supporting services for industry verticals:

- AC related to the radio domain (managed by the Radio Resource Controller)
- AC for the transport domain
- E2E AC that combines the AC from radio and transport and manages the E2E service accordingly.

Reference network

Figure 2 shows the three main functional blocks in our reference network: transport AC, machine learning based or statistical traffic forecasting, and AI-based network dynamics for QoS tuning. The combined use of these three blocks enables optimal dimensioning and operation of the transport network and reduces the level of over-provisioning.

Transport AC is invoked in two main phases. The first is when the slice for the E2E service is created, to dynamically verify the availability of the transport resources before their configuration/placement. If resources are not available, the connection is rejected and a notification is sent back to the originator or requester of the service. The second main phase is during service transmission, to ensure that QoS is in accordance with the SLA. Our proposed E2E orchestrator includes a novel function of transport AC that can be combined with the radio AC.

The traffic forecasting engine, which is either AI-based (machine learning) or statistical, utilizes metering functions on the traffic that enters the transport network nodes. This data is integrated with the current network status and with information related to specific circumstances (time of day, special events and so on). The traffic forecasting engine is responsible for determining traffic trends and their time behaviors over the considered deployment area. It also provides insights on actual radio traffic conditions that could not be observed and understood otherwise.

The QoS tuning functionality in the reference network is responsible for allocating and optimizing performance and network resources for all the admitted services, deciding at runtime the best routing based on a transport snapshot and the trends derived by traffic forecasting. Guided by policy, a certain amount of bandwidth is assigned to each transport tunnel (VLAN/VPN). The QoS parameters (effective bandwidth, PIR, CIR and committed burst size) are tuned according to actual needs, based on traffic prediction and measurements.

Case study: a smart factory

In a smart factory, many use cases are realized indoors in parallel. A smart factory pilot hosted at facilities operated by Comau, an industrial automation company, and TIM, a telecom operator, in Turin, Italy [2, 3] is a good illustration of this. Each use case in the smart factory puts specific, and often challenging, performance requirements on the telecommunication network. Failing to meet those requirements would immediately translate into bottlenecks in the manufacturing process.

The experimental area in the Comau factory is covered with a 5G network (RAN/CN) that is connected to TIM's central office (CO). The pilot includes a shared transport infrastructure, based on optical technologies deployed by Ericsson, which conveys radio traffic with appropriate E2E QoS. Cloud platforms, located on site and at TIM's CO, enable the implementation of Network Functions Virtualization and the support of Comau's applications that are running remotely. *Figure 3* visualizes the pilot architecture, where the orchestration functionality runs in a specific server hosted at TIM's CO (in the bottom-right corner).

Three use cases have been deployed in the Comau pilot, as shown on the left side of Figure 3. The first use case captures the motion of a real robot and, through an ultra-low latency radio link, produces a synchronized digital twin. The movement of the mechanical robot and of the respective virtual renderings are perfectly aligned in time.

The second use case is dedicated to demonstrating real-time monitoring of the industrial assets. Data is captured from a massive number of sensors and sent to an application deployed by Comau that runs in TIM's cloud. This application uses the acquired data to plan predictive maintenance, improve the accuracy of its production planning forecast, and improve quality, among other things.

The third use case demonstrates immersive telepresence for an enhanced remote support scenario in which the maintenance staff are assisted by a remote expert to investigate and solve a failure using augmented reality (AR) and step-by-step digital tutorials.

The Comau pilot features two of the four key components of our E2E transport-aware slice orchestration solution: the transport-aware slicing

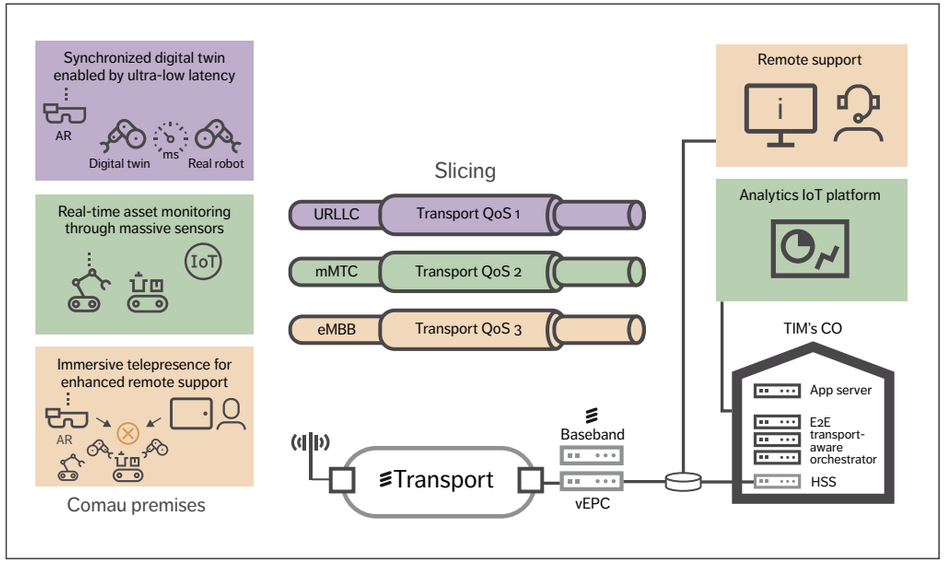


Figure 3 Architecture of the Comau pilot

and the transport-aware network abstraction functionalities shown in Figure 1. As a result, the solution ensures the alignment of all the resources (radio, transport and cloud) involved in the provisioning of the services with the related E2E QoS characteristics. Based on the QoS defined on the radio network, the corresponding requirements on transport are identified by the orchestrator, which then sends the requests to the transport domain. E2E QoS is managed automatically and dynamically through a unique infrastructure composed of radio and transport.

AI-based network dynamics for QoS tuning and AI-based AC – the other two key components of the solution presented in Figure 1 – have been defined and demonstrated as software modules in Ericsson laboratories.

The Comau pilot clearly demonstrates that the transport domain can maintain the properties of the network slice(s) it transports without the need to dimension it to the peak of the radio traffic, which will be a critical enabler for a wide variety of uses cases. Our concept of introducing transport-awareness into the slicing mechanism, and the other concepts demonstrated in the Comau pilot, have already gained significant recognition in our industry [4, 5].

Conclusion

The digital services of the future will demand new capabilities in 5G and beyond, including appropriate end-to-end (E2E) QoS at scale, in specific deployment areas (local, confined wide, general wide), where services vary dynamically in time and space. To meet these requirements, the network infrastructure will need to support a mix of heterogeneous services, including those demanding extreme performance.

Transport-aware network slicing orchestration will serve as a key enabler of services for industry verticals because its ability to manage network resources according to actual needs ensures cost-efficient service support. It does this by mapping the E2E QoS parameters of the service associated to the slice on the corresponding network resources, including the transport connections in the deployment area. The infrastructure domains can automatically maintain the properties of the network slice(s), from the service provisioning phase, without the need to dimension the transport network resources to the peak of the radio traffic. Our solution also features an artificial intelligence method to perform traffic forecasts and dynamically tune the QoS of the transport connections in the service deployment area, with the scope to optimize the usage of resources while guaranteeing the required performance level.

In essence, to support services for industry verticals, our research indicates that 5G slice orchestration should be extended in several aspects to increase automation and awareness of transport, and to maximize the amount of traffic served by reducing over-provisioning.

Further reading

- **Ericsson blog, Unlocking network transport in 5G and 6G networks, available at:** *https://www.ericsson.com/en/blog/2021/12/network-transport-5g-6g*
- **Ericsson blog, Highlights of key end-to-end network slicing capabilities, available at:** *https://www.ericsson.com/en/blog/2019/5/highlights-of-key-end-to-end-network-slicing-capabilities*
- **Ericsson blog, Network slicing orchestration, available at:** *https://www.ericsson.com/en/blog/2018/5/network-slicing-orchestration*
- **Ericsson, service orchestration, available at:** *https://www.ericsson.com/en/service-orchestration*

References

1. **Ericsson, Top 10 network slicing use cases to target, available at:** *https://foryou.ericsson.com/eso-network-slicing-use-cases-report.html*
2. **5GROWTH, 5G-enabled Growth in Vertical Industries, available at:** *https://5growth.eu/*
3. **YouTube, 5G for Industry 4.0: COMAU pilot, available at:** *https://youtu.be/tlyQBmRbNf0*
4. **5GPPP, 5G Infrastructure PPP – Trials and Pilots, December 2020, available at:** *https://5g-ppp.eu/wp-content/uploads/2020/12/5GInfraPPP_10TPs_Brochure2.pdf*
5. **5GPPP, 5G Infrastructure PPP – Trials and Pilots, August 2021, available at:** *https://5g-ppp.eu/wp-content/uploads/2021/10/5GInfraPPP_10TPs_Brochure2021_v1.0.pdf*

THE AUTHORS

Paola Iovanna

◆ joined Ericsson in 2000. She currently serves as a principal researcher driving research activities on transport network and orchestration solutions for next-generation mobile networks (beyond 5G). She has previously led a variety of activities within several EU projects, with responsibility for both pilots and field trials within industry verticals. The author of 70 patents and more than 80 publications, Iovanna holds an M.Sc. in telecommunications engineering from the University of Rome Tor Vergata, Italy.

Malgorzata Svensson

◆ is an expert in operations support systems. She joined Ericsson in 1996 and has worked in various areas within research and development. Svensson has broad experience in business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an M.Sc. in technology from the Silesian University of Technology in Gliwice, Poland.

Alexey Shapin

◆ joined Ericsson in 2017. In his role as senior researcher, he contributes to the radio architecture and protocol design of LTE and 5G New Radio, working closely with 3GPP standardization teams. His research focus is on time-critical communication and ultra-reliable low-latency communication. Shapin holds a Ph.D. in telecommunication from the Siberian State University of Telecommunications and Information Science, Novosibirsk, Russia.

Giulio Bottari

◆ is a master researcher at Ericsson Research in Pisa, Italy. Since joining the company in 2006, his research interests have focused on transport for radio, optical networks and ICT applications for industries. He also served as the innovation manager of the H2020 5G transformer project. Bottari is the author of 80 patents and several articles in publications by the IEEE (Institute of Electrical and Electronics Engineers). He holds an M.Sc. in telecommunications engineering from the University of Pisa.

Fabio Ubaldi

◆ joined Ericsson in 2011. A senior researcher in control plane methods for optical and radio systems, his current research focuses on transport network architecture and orchestration solutions for next-generation radio systems (beyond 5G). He is also active within the frameworks of several EU projects. Ubaldi is the author of 16 filed patent applications and more than 10 publications in the IEEE journal. He holds an M.Sc. in telecommunications engineering from the University of Perugia, Italy.

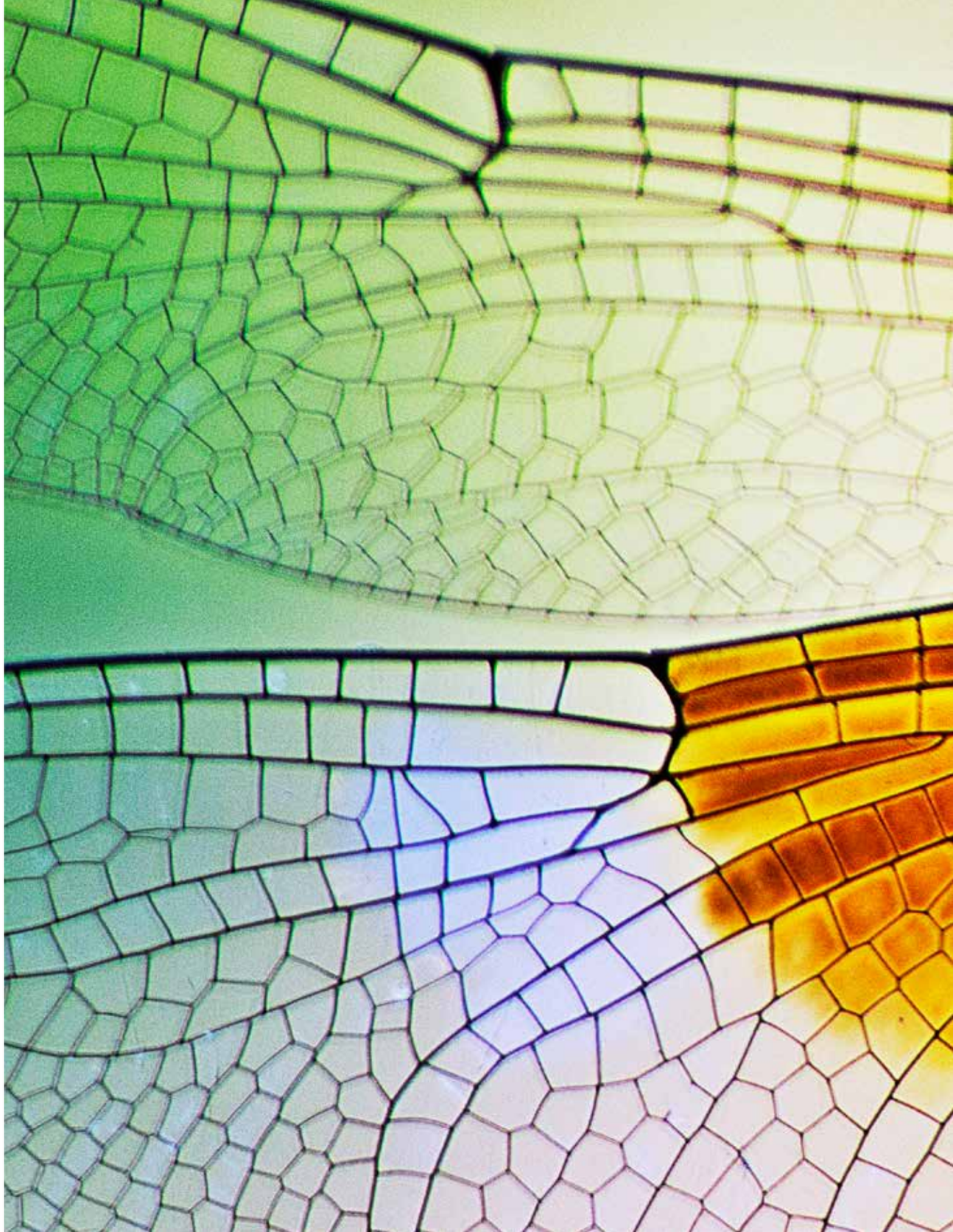
Filippo Ponzini

◆ is a senior researcher who joined Ericsson in 2007. His expertise includes 5G radio systems, optical transport and integrated photonics, and at present his work focuses on the definition of next-generation radio systems (beyond 5G). Ponzini holds an M.Sc. in telecommunications engineering from the University of Parma, Italy, and an MBA from the Sant'Anna School of Advanced Studies in Pisa, where he also serves as a researcher. Ponzini is the author of more than 30 publications and 40 patents.

Marzio Puleri

◆ is a master researcher whose interests include packet networks, intelligent traffic management, robotics, the Internet of Things, artificial intelligence and the support of industrial and logistics services through mobile networks. Puleri also has extensive knowledge of microelectronics and microwave systems. He holds an M.Sc. in electronic engineering from the Sapienza University of Rome and has worked at Ericsson since 1993.

The authors would like to thank to Massimiliano Maggiari and Gunnar Mildh for their contributions to this article.



ISSN 0014-0171
284 23-3385 | Uen

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