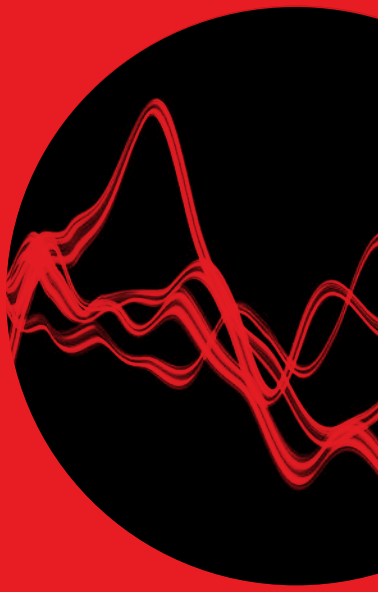
CHARTING THE FUTURE OF INNOVATION | VOLUME 109 I 2023-01

ERICSSON TECHNOLOGY Review



APPROACHING AI-NATIVE RADIO-ACCESS NETWORKS

OPTIMIZING NETWORKS WITH DIGITAL TWINS

PROGRAMMABLE 5G FOR THE INDUSTRIAL IoT

ERICSSON

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 00 00

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, Peter Butovitsch, Magnus Ewerbring, John Fornehed, Kjell Gustafsson, Jonas Högberg, Sara Kullman, Johan Lundsjö, Håkan Olofsson, Patrik Roseen, Robert Skog and Jessica Östergaard

ART DIRECTOR
Carola Pilarz (Nordic Morning)

PROJECT MANAGER
Susanna O'Grady (Nordic Morning)

LAYOUT
Carola Pilarz (Nordic Morning)

ILLUSTRATIONS
Nina Andersson (Nordic Morning)

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

ISSN: 0014-0171
Volume: 109, 2023

EVOLVING NETWORKS FOR A RAPIDLY CHANGING WORLD

■ **ERICSSON IS DETERMINED** to create a universal connectivity platform for innovation, automation and digitalization that has the flexibility to grow and expand to meet the changing needs of individuals, communities and enterprises all around the globe. Fixed wireless access, the Industrial Internet of Things (IIoT), enhanced network automation, digital twins and API (application programming interface) exposure all have key roles to play in the transition.

Mobile broadband has unquestionably become one of the cornerstones of social and economic development in the 21st century. Despite this, a recent report from the UN's International Telecommunication Union found that 37 percent of the global population has never used the internet. At Ericsson, we believe that the best way to tackle this challenge of inclusion in the near term is by leveraging the potential of 5G fixed wireless access (FWA).

5G mmWave extended range is a major technological breakthrough that boosts the capacity of 5G FWA solutions by enabling the use of mmWave spectrum to serve a much larger number of homes than was previously possible. This approach makes it possible to focus mid-band resources on homes at more distant, challenging locations, thereby opening the opportunity to offer high-end "wireless fiber" services to more homes in more sparsely populated suburban, rural and unconnected areas.

Beyond greater reach, the universal connectivity platform that we envision will also require a much higher degree of automation, particularly with respect to BSS/OSS (operations support systems/business support systems) functions and the deployment of software adopting CI/CD (Continuous Integration and Continuous Deployment) practices. Our latest research indicates that the most promising way to achieve this is by taking a declarative, GitOps-based approach to telecom software delivery and operation processes. One of the articles in this issue explains how.

Meanwhile, we can see that digital twins are quickly becoming a reality in a variety of industries. Our latest research indicates that network digital twins (NDTs) have the potential to deliver massive benefits for communication service providers (CSPs) by supporting use cases in areas such as R&D, planning, deployment and operations. As different types of tailored NDTs become more common, they can be combined to create increasingly sophisticated digital representations that can deliver even more powerful functionality for enhanced network operations.

With regard to IIoT solutions, there are signs that a growing number of large enterprises are interested in managing their own connectivity needs and building applications that leverage network characteristics and assets. The availability of properly scoped and easily accessible APIs is essential to support this. As part of our extensive research in this area, we recently completed a proof-of-concept project with operational technology company ABB that demonstrates the benefits of using programmable 5G systems to support IIoT use cases. It is our belief that enterprises in the operational technology industry will soon be able to use this type of standardized exposure service to seamlessly integrate their industrial automation systems with private 5G networks provided by CSPs.

> MOBILE BROADBAND HAS UNQUESTIONABLY BECOME ONE OF THE CORNERSTONES OF SOCIAL AND ECONOMIC DEVELOPMENT

We hope you enjoy this issue of our magazine and that you will share it with your colleagues and business partners. You can find both PDF and HTML versions of all the articles at: http://www.ericsson.com/ericsson-technology-review

ERIK EKUDDEN
SENIOR VICE PRESIDENT,
CHIEF TECHNOLOGY OFFICER

Creating programmable 5G systems for the Industrial IoT

In close collaboration with the operational technology company ABB, we have developed and tested a prototype of a programmable 5G system and successfully integrated it with an ABB automation system. Beyond demonstrating the advantages of using 5G to support industrial automation solutions, the ABB proof of concept highlights the importance of emerging 3GPP standards to address the expectations of industry verticals with regard to system integration.

GERGELY SERES, DIRK SCHULZ, OGNJEN DOBRIJEVIC, ABDULKADIR KARAAĞAÇ, HUBERT PRZYBYSZ, ALA NAZARI, PETER CHEN, MÁRK LÁSZLÓ MIKECZ, ÁRON DÉNES SZABÓ

A steadily growing number of factories, plants, mines and ports around the world are exploring the potential of 5G technology and considering how best to deploy it. This is to be expected, since 5G has been designed with vertical use cases in mind, and industrial automation systems are one of the most promising segments.

■ Private 5G networks [1] are becoming a critical and indispensable tool for enterprises in the operational technology (OT) vertical. The transformation of production environments such as process plants (chemical industry, mining, pharma, food and beverage, and so on) and factories (automotive or electronics manufacturing) driven by Industry 4.0 [2, 3] creates a dynamic environment that necessitates the reconfiguration of the automation system infrastructure and, by extension, the reconfiguration of the supporting 5G network and the continuous monitoring of the wireless connectivity service it provides.

Such flexibility enables the stepwise introduction of industrial applications over a common 5G infrastructure. In most cases today, the reconfiguration and monitoring of private 5G networks is done manually, often with the involvement of the communication service provider (CSP) or other entity that operates the 5G network. In wired automation networks based on technologies such as Industrial Ethernet or fieldbuses, the automated configuration and monitoring from within the automation system is the state of the art, translating the needs of applications into network configuration without lag, effort and quality problems. To use 5G as a part of the automation infrastructure on scale, that same seamless integration is required.

Therefore, the next step is to establish a live connection between private 5G networks and existing OT/IT systems. A private 5G network is expected to act as an integral part of the OT/IT communication infrastructure, seamlessly integrated with existing wired networks and upcoming technologies such as Time-Sensitive Networking (TSN) from the Institute of Electrical and Electronics Engineers (IEEE). Industry verticals expect to perform this system integration relying on their existing OT/IT skills, without the need to acquire additional competence in cellular wireless communication systems.

While 5G technology is designed to be scalable, flexible and extremely versatile regarding performance, these advantages come with a complex approach to building and operating networks that requires expertise commonly not available in OT companies today. The need to understand cellular technology in detail is therefore a significant roadblock to the adoption of 5G in the industry sector.

To overcome this challenge, private networks need to include user-oriented 5G exposure interfaces that are much simpler to use than any of the current telco-oriented exposure interfaces that assume deep knowledge of the internal workings of cellular systems. Such interfaces must offer the adequate level of abstraction that allows factory or plant operators to execute their regular operational tasks without the need for dedicated support from the service (and network) provider. In short, the ability to execute network automation across the organizational boundaries of CSP and OT enterprises needs to be an integral part of an industrial private 5G network offering.

Identifying 5G exposure requirements for industry verticals

With broad participation from the OT/IT and telecommunication industries, including Ericsson and ABB, the 5G Alliance for Connected Industries and Automation (5G-ACIA) collected and documented the requirements on 5G exposure capabilities for the process automation, production IT, logistics and warehousing industry verticals and published them in a white paper [4].

Figure 1 visualizes the concept of 5G exposure interfaces of a 5G private network, as presented in 5G-ACIA's white paper [4]. These interfaces enable Industrial Internet of Things (IIoT) applications to program the 5G network in a variety of ways, such as establishing connections of device-to-device and device-to-enterprise-network types with customized quality of service (QoS).

The 5G-ACIA concept builds on nine key exposure requirements in the area of device management:

1. Device connectivity management
2. Device connectivity monitoring
3. Device group communication management
4. Device provisioning and onboarding
5. Device identity management
6. Device location information
7. Security
8. Time-sensitive networking (TSN) integration
9. Time-sensitive communications.

Private 5G networks and the Industrial Internet of Things

A private 5G network is a deployment of the 5G system for private use. A private 5G network can be either standalone or deployed with the support of a public 5G network. In either case, a standardized exposure service offering is necessary to allow enterprises to customize 5G connectivity to fit the specific communication needs of industrial applications.

The **Industrial Internet of Things** (IIoT) is a subset of IoT applications tailored for advanced industrial automation.

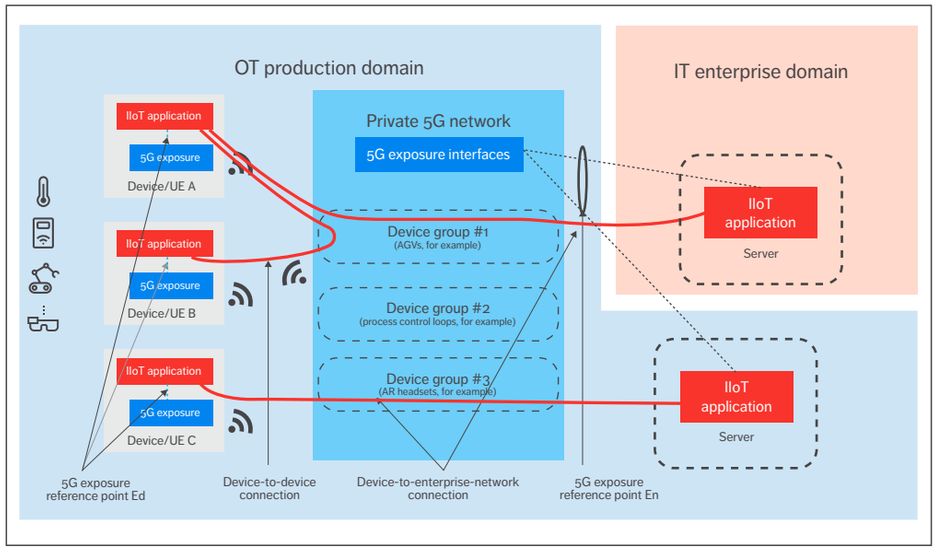


Figure 1 The 5G-ACIA concept of 5G exposure interfaces

To test the concept, we addressed the majority of these requirements in the joint ABB-Ericsson proof of concept.

Device connectivity management

IIoT applications are often time-critical, requiring low bounded latency and reliable communication. Through the 5G exposure interfaces, applications must be able to set up one or more connections per device with customized QoS, including guaranteed and minimum bitrate, latency and packet transmission reliability. These IP or Ethernet connections must support device-to-device or device-to-enterprise-network configurations. The exposure interfaces must hide the underlying realization aspects of the 5G network, such as QoS flows, resilient connections with low interruption time in case of node/link failure or disjoint user plane paths.

Device connectivity monitoring

For business continuity reasons, it is essential that a factory or plant operator can monitor the 5G connectivity service continuously through its OT/IT applications. The 5G exposure interfaces must enable the monitoring of connections of a device or a group of devices, allowing the retrieval of current and historical performance metrics either on demand, periodically, or on an event-triggered basis related to connection bitrate, latency and packet loss, for example.

Device group communication management

Industry verticals expect to be able to isolate the traffic of different use cases and traffic types for the purposes of performance and security management. Traffic segmentation is enabled by 5G group communication, where devices in the same group communicate privately with one another and can also access services in enterprise networks. 5G provides group communication with 5G local area network (5GLAN)-type services for both IP virtual network (VN) groups and 5GLAN virtual local area networks (VLANs). The 5G exposure application programming interface (API) must provide the means for applications to manage device groups, including creating groups and adding and removing group members, as well as creating dynamic VLAN assignment for devices when connecting to the network.

Device provisioning and onboarding

Industry verticals want to be able to add devices to the 5G network in a plug-and-play manner. The 5G exposure interfaces must enable the provisioning of device identifiers and credentials into the 5G network both for individual devices and groups of devices. In the onboarding step, the 5G network must provide the means for the device to establish a user plane connection to the IIoT application and have the ability to notify the application about the newly established connection.

Device identity management

IIoT applications use a wealth of identifiers both in the application layer and in the connectivity layer, depending on the applied technology. In 5G networks, the primary unique identifier of 5G user equipment (UE) is the Generic Public Subscription Identifier (GPSI), which means that this ID must be used by the 5G exposure interfaces. Translation between the IIoT device's application layer (OT/IT, for example) identifiers and the GPSI must be done in the application. The static IP address of the device or the device's media access control (MAC) address may also be used as the device identifier in the 5G exposure interfaces.

Device location information

Use cases such as mobile robots, automated guided vehicles, portable assembly tools, mobile control panels and plant asset management require the positioning of IIoT devices with different levels of accuracy. IIoT applications may request the location information of one or a group of devices over the 5G exposure interfaces. Device tracking is achieved by reporting device location triggered by events such as movements.

> INDUSTRY VERTICALS WANT TO BE ABLE TO ADD DEVICES TO THE 5G NETWORK IN A PLUG-AND-PLAY MANNER

Security

IEC 62443 standards [5] introduce the concepts of "zones" and "conduits" as a way to segment and isolate the various subsystems in a control system. A zone is defined as a grouping of logical or physical assets that share common security requirements based on factors such as criticality and consequence. A conduit consists of the grouping of cyber assets dedicated exclusively to communications within and also external to a zone and which share the same cybersecurity. Device groups (5GLAN VLANs or IP VN groups) combined with secured slicing and application-level security protect the factory zones achieving IEC 62443 Security Levels SL3 and SL4 [5].

Time-sensitive networking integration

OT verticals consider TSN to be the next-generation technology that will bring about convergence in OT networking. When combined with 5G networks, the fully centralized TSN configuration model of IEEE 802.1Qcc must be used. It postulates that a centralized network configuration (CNC) entity configures all the TSN streams in the 5G network, which acts as a TSN bridge. The 5G exposure interfaces must serve as the TSN application function (AF) and provide port and bridge management information. This enables the CNC to determine the allocation of network resources to the streams and configure them in the 5G network through the 5G exposure interfaces.

Time-sensitive communications

5G-native time-sensitive communications (TSC) refers to a time-sensitive communication service that the 5G network offers to 5G devices natively (that is, without integration into a TSN system). 5G exposure

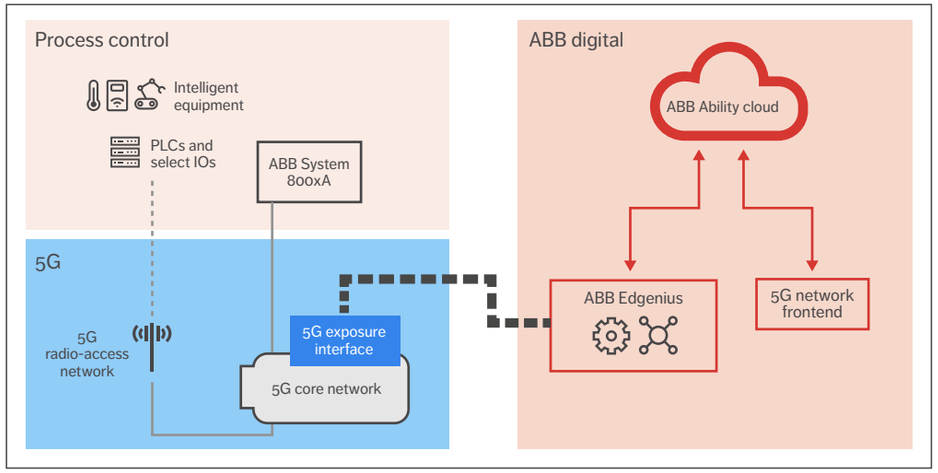


Figure 2 Overview of the demonstrator that integrates an ABB Ability automation system with Ericsson's prototype 5G exposure interfaces

interfaces acting as TSC CNC enable applications such as Centralized User Configuration to discover the availability of resources for a TSC stream and request the creation of a TSC stream with QoS.

Validating the 5G-ACIA exposure concept in partnership with ABB

The proof of concept at ABB integrates Ericsson's pre-standard, prototype 5G exposure interfaces implementation with ABB's cloud-based digital ecosystem (ABB Ability [6] and ABB Ability Edgenius [7]) to achieve 5G network programming by an automation system and thereby validate the exposure capabilities in practice.

The proof of concept offers a plant operator an easy-to-use environment for managing and monitoring the 5G connectivity of networked industrial devices. It also makes it possible to tailor the behavior of the 5G network to the communication requirements of industrial applications and obtain knowledge regarding the status and performance of 5G connections and virtual networks. By means of these capabilities, network-aware applications can interact with an externally operated network infrastructure without the need to know or understand the details of the underlying network technologies.

As illustrated in *Figure 2*, the 5G network frontend allows the control of 5G connectivity for ABB core control devices and intelligent equipment from a web application using the Ability cloud platform and an on-site Edgenius edge module. The Edgenius module interacts with the 5G network's programmable interfaces that expose the management capabilities of the 5G network, thereby offering a unified way of centrally controlling 5G performance (using ABB Ability, in this case).

The 5G network frontend web application provides the means to seamlessly provision and onboard 5G devices, monitor their connectivity performance and create device groups with different QoS attributes on top of the shared 5G infrastructure. Therefore, the developed tool also creates a convenient and scalable solution for configuring and monitoring the 5G network to facilitate the execution of everyday tasks by OT users, from automation engineers to plant operators.

By bringing 5G device management and monitoring data "closer" to the application operation and engineering data, this proof of concept allows the management of several 5G device groups from the tool, mainly based on ABB's digital solutions. This approach would also make it possible to utilize existing tools in ABB's digital portfolio to achieve easier and faster development of both simple and advanced network management solutions for the industrial wireless domain.

As a result of the collaboration with Ericsson, ABB also investigated the feasibility of using the exposure capabilities for flexible 5G network programming. This involved allowing different ABB automation solutions to make use of 5G technology, while at the same time shielding these solutions from 5G network implementation details and complexities. The results show how an OT organization (automation vendor or plant operator) could use one logical cloud instance to manage all 5G networks centrally as they grow, or as new ones are added.

The proof of concept at ABB clearly demonstrates that it is possible to run network automation across the boundaries between OT automation systems and private 5G networks, which is a prerequisite for using 5G as a part of automation solutions.

The critical role of 3GPP in enabling the IIOT

Both ABB and Ericsson are active proponents of standardized technologies such as 3rd Generation Partnership Project (3GPP)-based solutions. Emerging 3GPP standard technologies such as the network exposure function (NEF), the service enabler architecture layer (SEAL) for verticals and the common API framework (CAPIF) have the potential to address the exposure-related requirements of the verticals by offering integration points between automation systems and the 5G network.

3GPP standard exposure technologies hide the complexity of 5G and offer industry verticals a simple, secure, use-case-oriented configuration interface to the 5G system. The exposure interfaces will be invaluable to a multitude of industrial use cases, allowing industry verticals to make use of the key features and performance that 5G has to offer in a simple and straightforward manner.

> "3GPP STANDARD EXPOSURE TECHNOLOGIES OFFER INDUSTRY VERTICALS A SIMPLE, SECURE, USE-CASE-ORIENTED CONFIGURATION INTERFACE"

The 3GPP has already made significant progress toward exposing the capabilities of mobile networks through APIs. While it is well known that the 3GPP core network capabilities are exposed by the NEF, since release 16 the 3GPP has also been standardizing higher-level APIs to address requirements from various vertical applications, with further enhancements specified in release 17, and additional functionality currently under study for release 18.

Figure 3 provides an overview of 3GPP standards that are applicable for IIoT use cases, as defined by the 3GPP SA6 working group. From the bottom to the top, the following three layers of 3GPP exposure are depicted:

1. The network exposure layer, which exposes core network capabilities
2. The SEAL, which exposes common service enablers for verticals
3. The vertical application enabler (VAE) layer, which exposes vertical-specific service enablers.

3GPP network exposure function

The basic 3GPP core network exposure layer consists of the 5G NEF, which offers network capabilities exposure of the 5G Core toward external applications integrated with the 3GPP network. The following subset of NEF APIs [8] are relevant for IIoT use cases:

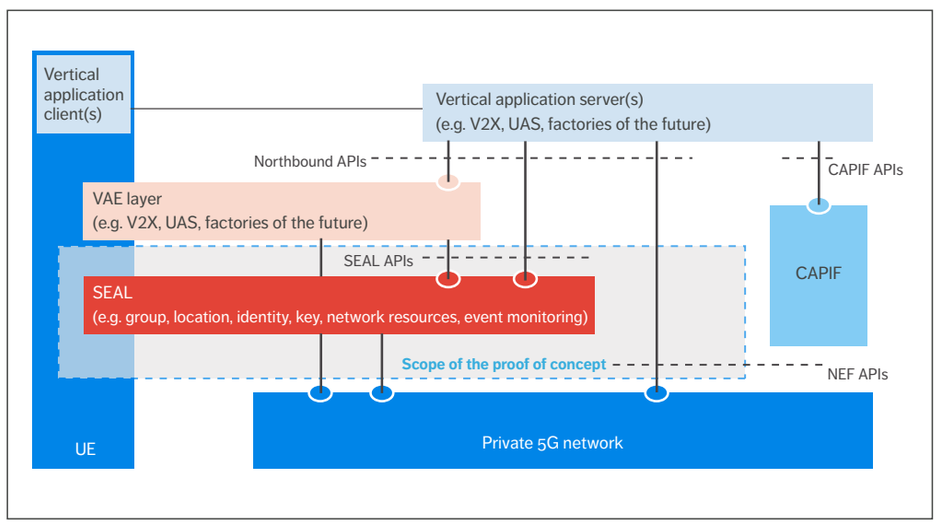


Figure 3 Overview of 3GPP standards applicable for IIoT use cases

- Event monitoring (device location, reachability, connection status)
- AF session with QoS (on-demand QoS for IP and Ethernet connections)
- Analytics exposure
- 5G LAN parameter provisioning (device group management)
- Service parameter provisioning (route selection parameters)
- Time sync exposure
- UE ID retrieval (AF-specific device ID retrieval, such as GPSI)

> THE SEAL ENABLES COMMON SERVICES THAT APPLICATIONS CAN UTILIZE FROM VARIOUS VERTICAL DOMAINS

3GPP service enabler application layer

Because the NEF APIs expose network capabilities in a highly granular manner, application developers that use them must have a good understanding of the underlying network concepts. To simplify application development and deployment, 3GPP has specified a new layer of simplified service enablers. The SEAL [9] consists of service enablers that provide services that are not specific to any vertical – that is, they are common services that applications can utilize from various vertical domains. The APIs currently defined by SEAL are for group management, location management, identity and key management, and network resource management (NRM).

Group management allows the application to create and manage device groups for different purposes such as group communication and location-based groups, while the SEAL ensures that the devices are properly notified and joined into the group. Location management makes it possible to provide device location information from different sources – both 3GPP and non-3GPP (such as Global Navigation Satellite Systems) – to an application either on demand or upon change, and to define location areas of interest for specific use cases. Identity and key management support applications in managing security material used in the authentication and authorization of users and devices.

Lastly, NRM enables application-specific usage and monitoring of network resources used by the devices covering:

- Unicast and multicast connection activation, deactivation and modification including QoS parameters
- Unicast connection QoS monitoring including packet delay, packet loss rate, data rate and traffic volume
- Event monitoring including device mobility, communication, loss of connectivity, location reporting and connection status
- Time-sensitive, deterministic device-to-device and device-to-enterprise-network communication.

Industry vertical requirements (5G-ACIA)	3GPP exposure capabilities
Device connectivity management	• SEAL NRM QoS and TSC APIs • NEF QoS API • NEF time sync exposure API
Device connectivity monitoring	• SEAL NRM event monitoring and QoS monitoring APIs • NEF event monitoring API • NEF QoS API • NEF analytics API
Device group communication management	• SEAL group management APIs • NEF 5G LAN parameter provisioning API
Device provisioning and onboarding	• Provided by the APIs of business support systems
Device identity management	• GPSI is supported as the device identifier • An IP/MAC address can be used to identify a device • SEAL identity and key management APIs can be used as part of the security framework
Device location information	• SEAL location management APIs • NEF event monitoring API (UE location reporting)

Figure 4 Summary of 3GPP exposure capabilities and APIs that satisfy the requirements of IIoT use cases

The SEAL is expected to evolve and grow with additional service enablers in the upcoming 3GPP release 18.

3GPP vertical application enabler layer

In contrast to the SEAL, the VAE layer is tailored to satisfy specific vertical applications. These types of vertical service enablers are currently defined for automotive applications referred to as vehicle-to-anything (V2X) communication and drone applications known as unmanned aerial systems (UAS). VAE for factories of the future will include future enhancements specific to OT verticals.

Meeting IIoT requirements with 3GPP exposure capabilities

Figure 4 shows how the IIoT requirements outlined by 5G-ACIA match up to 3GPP release 17 exposure capabilities and APIs.

Conclusion

Widespread use of private 5G networks in the Industrial Internet of Things (IIoT) ecosystem will require a standards-based exposure solution that makes it possible to flexibly configure the 5G system according to the specific communication requirements of individual production processes. Communication service providers (CSPs) have an excellent opportunity to monetize the IIoT with a service offering that exposes the powerful capabilities of 5G networks to industry verticals. The reduction in manual network configuration tasks allows customer support departments of CSPs to scale up the number of enterprise customers they can serve.

A standards-based 5G IIoT exposure solution will enable industrial enterprises to use 5G as a part of system infrastructure, increasing production flexibility and scaling up to a large number of 5G-connected devices in an organized and secure manner. It will also open the door for IT/OT platform vendors to develop their own products that take advantage of 5G capabilities and enable system integrators to simplify the integration of operational technology applications with the wireless connectivity that 5G systems provide.

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **5G-ACIA** – 5G Alliance for Connected Industries and Automation | **AF** – Application Function | **AGV** – Automated Guided Vehicle | **API** – Application Programming Interface | **AR** – Augmented Reality | **CAPIF** – Common API Framework | **CNC** – Centralized Network Configuration | **CSP** – Communication Service Provider | **GPSI** – Generic Public Subscription Identifier | **IO** – Input/Output | **IoT** – Internet of Things | **IIoT** – Industrial IoT | **LAN** – Local Area Network | **NEF** – Network Exposure Function | **NRM** – Network Resource Management | **OT** – Operational Technology | **PLC** – Programmable Logic Controller | **QoS** – Quality of Service | **SEAL** – Service Enabler Architecture Layer | **TSC** – Time-Sensitive Communications | **TSN** – Time-Sensitive Networking | **UAS** – Unmanned Arial Systems | **UE** – User Equipment | **V2X** – Vehicle to Anything | **VAE** – Vertical Application Enabler | **VLAN** – Virtual Local Area Network | **VN** – Virtual Network

Further reading

- **Ericsson blog, How enterprises can exploit the exposure capabilities of private 5G networks, available at:** *https://www.ericsson.com/en/blog/2020/7/private-5g-network-capabilities-enterprise*
- **Ericsson, Network exposure, available at:** *https://www.ericsson.com/en/service-orchestration/network-exposure*
- **Ericsson blog, Network programmability in 5G, available at:** *https://www.ericsson.com/en/blog/2019/1/network-programmability---in-5g-an-invisible-goldmine-for-service-providers-and-industry*
- **Ericsson, Dedicated networks, available at:** *https://www.ericsson.com/en/portfolio/enterprise-wireless-solutions/dedicated-networks*
- **Ericsson, Industry 4.0, available at:** *https://www.ericsson.com/en/industry4-0*
- **Ericsson, 5G for manufacturing, available at:** *https://www.ericsson.com/en/5g/manufacturing*

References

1. **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*
2. **Ericsson-Hexagon Report, Connected Manufacturing – A guide to Industry 4.0 transformation with private cellular technology, November 2020, available at:** *https://www.ericsson.com/en/enterprise/forms/connected-manufacturing*
3. **German Federal Ministry for Economic Affairs and Energy (BMWi), Fortschreibung der Anwendungsszenarien der Plattform Industrie 4.0 (Continuation of the Application Scenarios of the Plattform Industrie 4.0), October 2016, available at:** *https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/fortschreibung-anwendungsszenarien.pdf*
4. **5G-ACIA white paper, Exposure of 5G Capabilities for Connected Industries and Automation Applications, February 2021, available at:** *https://5g-acia.org/wp-content/uploads/WP_039_Network-Exposure-Interface_single-pages.pdf*
5. **IEC, Understanding IEC 62443, February 2021, available at:** *https://www.iec.ch/blog/understanding-iec-62443*
6. **IEC 62443-3-3, Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels, available at:** *https://webstore.iec.ch/preview/info_iec62443-3-3%7Bed1.0%7Db.pdf*
7. **ABB, ABB Ability, available at:** *https://global.abb/topic/ability/en*
8. **ABB, ABB Ability Edgenius Operations Data Manager, available at:** *https://new.abb.com/process-automation/edgenius*
9. **3GPP Technical Specification 23.502, Procedures for the 5G System (5GS): Stage 2, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145*
10. **3GPP Technical Specification 23.434, Service Enabler Architecture Layer for Verticals (SEAL); Functional architecture and information flows, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3587*

THE AUTHORS

Gergely Seres
◆ is an expert and chief architect of software technology and application architecture at Business Area Cloud Software and Services, currently working with 5G and the Internet of Things (IoT). Since joining Ericsson in 1998, he has held several research, technical and managerial positions. He holds a Ph.D. in electrical engineering from the Budapest University of Technology and Economics, Hungary.

Dirk Schulz
◆ is a senior principal scientist with ABB Research, responsible for the communication architecture of industrial automation systems. He has been with ABB since 2006, working in different scientific and project management roles. He holds a Ph.D. (Dr. rer.-nat.) in communications engineering from the University of Mannheim, Germany.

Ognjen Dobrijevic
◆ is a principal scientist with ABB Corporate Research. He is working on different aspects of future industrial communication systems, with a focus on wireless connectivity and edge computing. Dobrijevic has been with ABB since 2018 and holds a Ph.D. in electrical engineering from the University of Zagreb, Croatia.

Abdulkadir Karaağaç
◆ is a scientist with ABB Corporate Research, working on communication and interoperation solutions for industrial automation systems. Karaağaç holds a Ph.D. in computer science from Ghent University in Belgium, and he has been with ABB since 2020.

Hubert Przybysz
◆ is an expert in core network exposure at Business Area Cloud Software and Services. He joined Ericsson in 1990. His current assignments are focused in the areas of Industrial IoT (IIoT) and exposure of 5G system capabilities. Przybysz holds an M.Sc. in telecommunications from the Warsaw University of Technology in Poland.

Ala Nazari
◆ is an expert in media delivery architecture and in 5G for critical IoT. He joined Ericsson in 1998 and has been working with 3G/4G/5G, broadband access, transport and media delivery. He has also worked as a senior solution architect and engagement director. Nazari holds an M.Sc. in computer science from Uppsala University in Sweden.

Peter Chen
◆ is the system owner of core network exposure at Business Area Cloud Software and Services, where he focuses on technology strategy and evolution in the network exposure area. He has been working in different areas within the core network since he joined Ericsson in 2006, and contributed more than 20 patents within Ericsson. He holds a B.Sc. from Dalian University of Technology, China.

Márk László Mikecz
◆ is an architect of 5G network exposure at Business Area Cloud Software and Services. He joined Ericsson in 2016. His current assignment is focused on the 5G exposure interface proof of concept. Mikecz holds a B.Sc. from the Eötvös Loránd University in Budapest, Hungary.

Áron Dénes Szabó
◆ joined Ericsson in 2021 as a system architect of 5G network exposure at Business Area Cloud Software and Services. His work focuses on standardization and prototyping in 5G IIoT. Szabó holds an M.Sc. in engineering physics and a Ph.D. in electrical engineering from the Budapest University of Technology and Economics.

Closing the digital divide with mmWave extended range for FWA

5G mmWave extended range is a technology breakthrough that redefines millimeter wave as a solution to deliver wide-area 5G coverage. This innovation extends mmWave coverage from a few hundred meters to more than 7km, enabling cost-effective, fixed wireless access services for suburban and rural communities.

ANDERS ERICSSON,
LAETITIA FALCONETTI,
HÅKAN OLOFSSON,
JONAS EDSTAM,
TOMAS DAHLBERG

There is a huge demand for more broadband connectivity all around the world, with more than one billion homes still unconnected and even more underserved. Fixed wireless access (FWA) using 5G New Radio (NR) has emerged as one of the most important ways to close this digital divide, and it is rapidly becoming one of the key use cases for 5G [1].

■ FWA over 5G NR is a technique that delivers high-speed internet access by connecting the 5G network to customer premises equipment (CPE) at a fixed location at the consumer's home or enterprise. FWA has been shown to provide a very attractive and cost-efficient alternative to fixed broadband across a wide range of cases and situations, in particular in areas outside city centers where large distances make fiber deployment expensive [2].

FWA traffic volumes can be 10-50 times higher per subscription than for smartphones, and during the COVID-19 pandemic we have seen an increased use of home broadband and greater digitalization of homes and enterprises – effects that are predicted to drive further data volume increases. Since the average revenue per user of communication service providers (CSPs) is not expected to increase to the same degree, it is critical to design FWA solutions cost-efficiently.

Key methods for efficient fixed wireless access solutions

The large unmet demand for broadband connectivity can be met most cost-efficiently with FWA when it is built on the large installed base and global reach of 3rd Generation Partnership Project (3GPP) mobile technologies (4G LTE and 5G NR). CSPs can maximize the established momentum behind 3GPP technologies by deploying FWA together with mobile broadband (MBB) in existing and new spectrum bands, thanks to the options for ensuring efficient spectrum sharing between the two services. Mid bands using time division duplex (TDD) and frequency division duplex (FDD) low bands are sufficient in many FWA cases. In particular, 3GPP TDD mid band such as n41 (2.5-2.7GHz) and n77 (3.3-4.2GHz) unlocks major MBB+FWA combined opportunities.

High-end offerings in dense suburban areas often require additional capacity. In these cases, "high band" millimeter wave (mmWave) spectrum such as 26GHz and 28GHz can be added, either at the macro sites and/or through densification with new street sites. While mmWave spectrum is often associated with these dense deployments, with each site covering only a few hundred meters, the extended range of mmWave provides CSPs with a golden opportunity to expand the use of mmWave spectrum for FWA to sparser suburbs and semi-rural areas as well.

The principle is illustrated in *Figure 1*, where a macro cell site serving a range of several kilometers has been equipped with 5G NR for both mid band and mmWave radios. As all spectrum assets are available for FWA services in the entire sector, the system will automatically serve homes with mmWave coverage primarily using mmWave (shown in red), while other homes without mmWave coverage will be served by mid band (shown in black). The longer mmWave range needed in these deployments is made possible through a new innovation known as mmWave extended range.

> "CSPs CAN MAXIMIZE THE ESTABLISHED MOMENTUM BEHIND 3GPP TECHNOLOGIES BY DEPLOYING FWA TOGETHER WITH MBB"

How to extend the distance covered by millimeter wave signals

Two aspects must be addressed to operate an FWA network in mmWave spectrum over long distances: maximizing the received signal strength and accommodating long propagation delay.

Maximizing the received signal strength

The wavelengths of mmWave signals are short, as are the antenna elements. Consequently, the energy density of the signal reaching the antenna must be high to get a sufficient signal-to-noise level. Using large antenna arrays on the transmitter side together with beamforming enables the energy to be focused in the desired direction. Beamforming on the receiver side further improves the received signal strength.

Beamforming is a core component of the transmission on mmWave frequencies. It can be applied on both the network and device sides,

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **CAGR** – Compound Annual Growth Rate | **CPE** – Customer Premises Equipment | **CSP** – Communication Service Provider | **DL** – Downlink | **FDD** – Frequency Division Duplex | **FWA** – Fixed Wireless Access | **MBB** – Mobile Broadband | **mmWave** – Millimeter Wave | **NR** – New Radio | **TDD** – Time Division Duplex | **UL** – Uplink | **xDSL** – Arbitrary Digital Subscriber Line

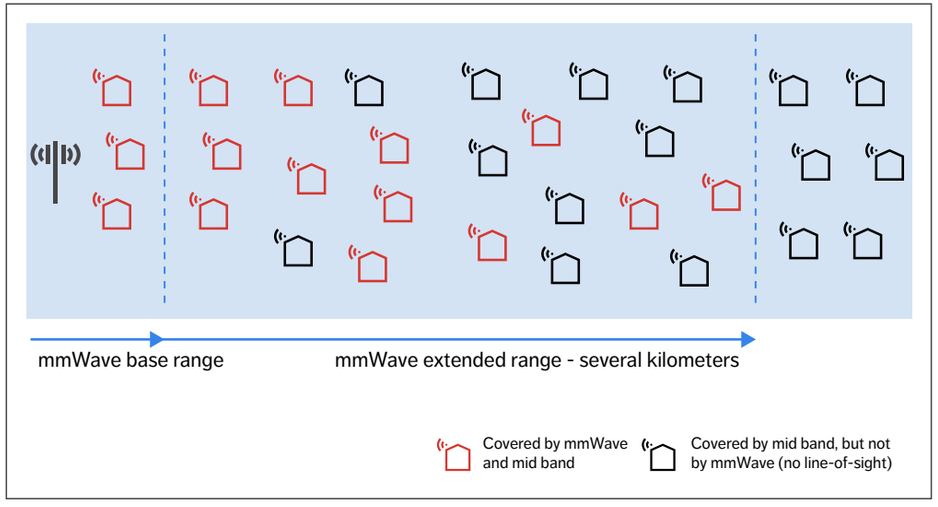


Figure 1 A macro-cell site equipped with 5G NR for both mid band and mmWave, in which well-located homes at a range of several kilometers can be served using mmWave extended range

although the beamforming capability on the network side is more advanced. Beamforming alone is, however, not sufficient to reach distances of many kilometers. Mobile networks operating in the mmWave frequencies today apply beamforming but cannot reach users that are several kilometers away.

With FWA, it is possible to create conditions that are very favorable to long distance coverage. On the network side, existing deployments of high-power radios placed above the main obstacles (by means of macro towers, for example) are already well suited to maximize the downlink (DL) received signal strength.

On the device side, AC-powered CPEs have significantly higher transmit power than battery-powered mobile devices. They can also be placed in the most favorable location for the connection, which is usually outdoors, to avoid wall penetration loss. Both the CPE's transmit power and its location are essential to improve the uplink (UL) received signal strength, which is the factor that typically limits coverage.

Accommodating long propagation delay

The ability to accommodate long propagation delay is a key enabler for long-distance communication. The extended range feature [3] provides this ability. Once sufficiently good signal strength is ensured at a given distance, it is necessary to adapt the communication system to accommodate the propagation delay corresponding to the targeted distance. In mmWave spectrum TDD is applied. With interleaved UL and DL slots, the transceivers need a few microseconds to switch between receiving and transmitting mode. The TDD format therefore includes a short gap period between DL and UL symbols.

The length of this gap period should also accommodate the signal propagation from the transmitter side to the receiver. The longer the distance between these, the larger the gap needs to be. For an MBB-centric deployment where the mmWave cell range does not typically exceed a few hundred meters, mmWave transmissions would only need a very short gap between the DL and UL slots of a TDD pattern.

Figure 2 Testing of the mmWave extended range feature in a semi-rural area at a 6km distance from the macro base station

For longer distances intended for an FWA-capable deployment scenario, the gap needs to be enlarged. This means that a few additional data symbols will be muted to cover for the longer distance. In addition to a larger gap duration, a 3GPP-defined random access preamble format that is favorable to a long-distance scenario is used. Random access preambles are a basic technical component of the 3GPP release 15 specifications and are therefore supported by all devices. On the network side, the detection of random access preambles subject to long propagation delay can be improved by means of an advanced random access receiver algorithm.

The larger gap duration in the TDD pattern required to cover long distances increases the overhead for all devices in larger cells by a few percent and slightly decreases capacity and peak rates. The extended range configuration should therefore be applied only where needed.

Several field testing activities have been carried out to determine how far a radio signal on mmWave frequencies can propagate and what data rates can be expected at this range [4, 5, 6, 7]. *Figure 2* shows the measurement setup for one of the trials that took place in a semi-rural area. The 5G mmWave macro base station 6km away is highlighted with a red circle, and the CPE in the foreground is a Qualcomm 5G Fixed Wireless Access Reference Design. Our field testing showed that in the right conditions, it is possible to achieve DL data rates larger than 1Gbps at a cell range beyond 7km in mmWave frequencies. The feature is now also deployed in commercial networks [8].

Case study: US digital village

The extended range opens up new opportunities to use mmWave spectrum in sparser suburbs and semi-rural areas, which then makes it possible to offload lower frequencies. In such areas, there could be several hundreds of homes per sector. With increasing consumption on both MBB and FWA, the additional capacity that mmWave brings makes such a scenario a sweet spot for combining mmWave and TDD mid band, thereby providing FWA subscribers with high service levels.

In the following simulated scenario, which models

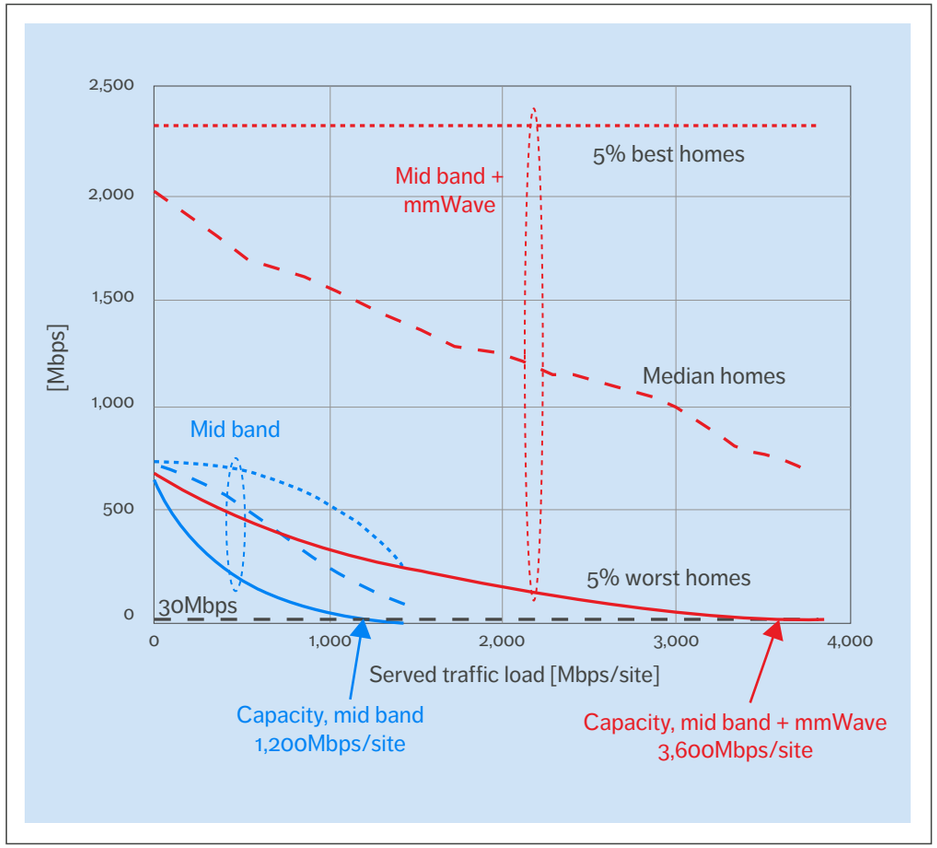


Figure 3 Experienced user data rate – DL user data rate as a function of system load for mid-band-only and mid-band + mmWave deployments

the performance achieved in field trials, we illustrate how mmWave extended range can be used to increase capacity and boost user experience. The case is a version of the digital village case study in the Fixed Wireless Access Handbook [9] that has been adapted to US data consumption patterns.

The original case study includes a stepwise solution and business case analysis showing a return on investment of 22 months. Here, we focus on a comparison of the achievable network capacity, with and without deploying mmWave. The targeted case is a village together with surrounding, more sparsely populated areas where the overall home density is around 150 homes per square kilometer. Current broadband offerings are mainly provided by xDSL or best-effort MBB, but there is no fiber-to-the-home, which makes the area an attractive candidate for FWA.

The existing MBB deployment has as a macro inter-site distance of 3km, and lower FDD bands are used to serve current traffic. Over time, as the MBB traffic grows, it will utilize part of the acquired mid-band spectrum. The excess spectrum can be used for FWA: 100MHz TDD at 3.5GHz and 400MHz in the 28GHz band.

The service targeted by the CSP is FWA with a "fiber-like" experience. This means sold DL data rates of 100-1,000+Mbps without a data cap and with typical DL rates of at least 100Mbps. Combining available spectrum, including lower FDD bands, and mid bands and 28GHz using TDD, the CSP can obtain a combined network deployment catering for both MBB and FWA.

In this analysis, we focus on the mid band and on 28GHz, and we leave out the details on lower bands as well as the performance for the MBB users. However, the suggested approach includes a joint solution for FWA and MBB that also handles the anticipated growth of MBB traffic. Furthermore, as the case is limited by the DL capacity, we leave out the analysis of the UL. To maximize link performance, the case is based on the use of rooftop-placed, high-power CPE that supports mmWave as well as lower bands.

The system is dimensioned to target a minimum DL data rate of 30Mbps for the 5 percent worst located homes, at peak traffic hours, to sustain a fiber-like experience, including multiple HDTV streams per home, also in those worst cases. Regarding data usage, we define a baseline scenario, based on observed current US fixed broadband levels, where the average data consumption per home is expected to be 670GB per month, out of which 90 percent (600GB) is DL traffic [10, 11].

Assuming that 10 percent of the daily traffic occurs during the busiest hour, this corresponds to an average consumption of 2GB per hour at busy hour. We assume an annual growth of 28 percent, partly driven by many homes transitioning from consuming linear TV over satellite or terrestrial broadcast, to using broadband for all media consumption including linear TV and streamed services.

In addition, for comparison we have also defined an all-broadband-media scenario that assumes that all homes have already made this transition. For this case, we assume a consumption rate of 1TB per month per home (900GB per month in the DL) but expect lower annual growth of 10 percent, as the shift to all media consumption over broadband is already completed.

As the capacity needs to grow with an increasing number of customers, as well as with higher average data consumption and speed requirements, it makes sense to gradually increase the capabilities of the network on a needs basis. This means that costs for increased capacity can be taken as late as possible, as opposed to fiber, where a major part of the cost is taken upfront when deploying fiber trunks passing all homes. Furthermore, decisions about capacity enhancements can be made selectively on a sector by sector basis as the numbers of subscribers – and the revenues – increase.

Experienced user data rate

Figure 3 shows the experienced DL user data rate as a function of varying system load for the worst, median and best located homes respectively. The blue curves represent a mid-band-only deployment, while the red ones represent the combined mid-band and mmWave case.

We define the DL capacity as the system load at which the fifth-percentile worst-located homes experience a data rate of 30Mbps (the dashed black line) according to the dimensioning criterion described above. With 100MHz of mid band, the capacity is 1,200Mbps per site, while it is three times higher (3,600Mbps per site) when adding the mmWave spectrum.

Figure 3 demonstrates that, already with a mid-band-only deployment, even the worst-located homes will experience DL rates of 100Mbps or higher at moderate system load, which is most of the time. The peak user rates with mid band alone are in the range of 690-730Mbps depending on the location of the home. After adding 400MHz of mmWave spectrum, the rates of median homes increase drastically, and we also see a significant variation in this range of the peak

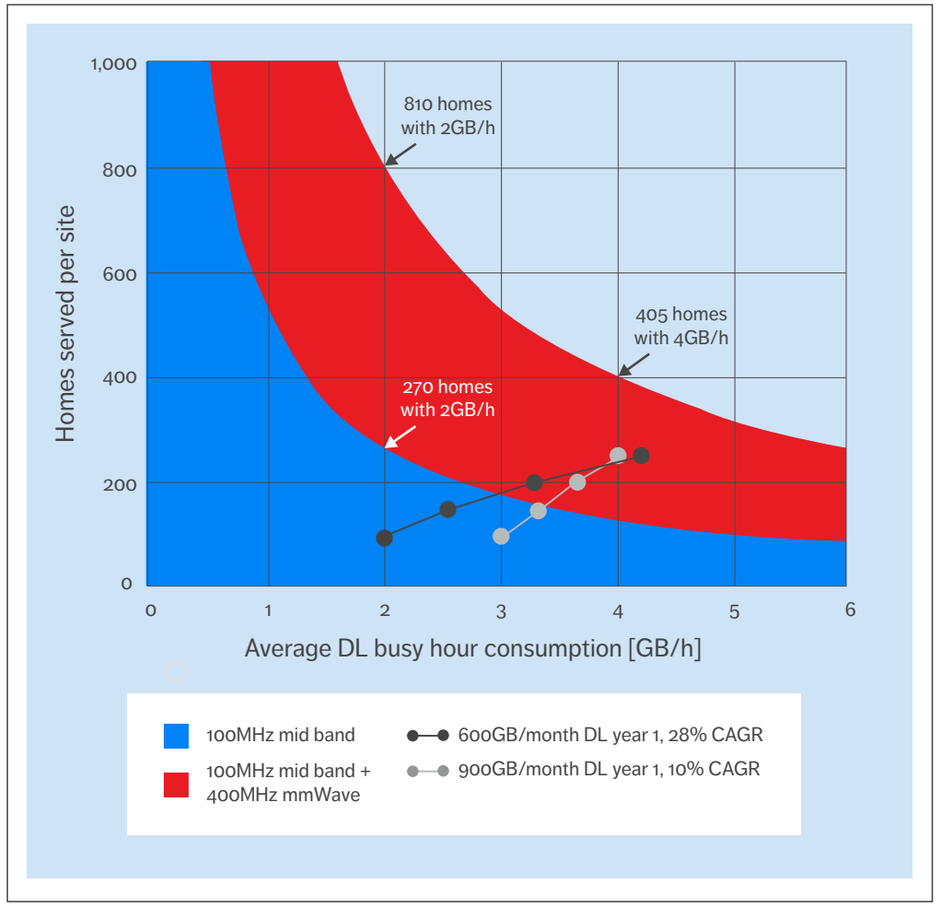


Figure 4 Capacity utilization

rates, which reach up to 2,300Mbps depending on whether or not the home can be served by mmWave.

Capacity utilization

Figure 4 shows how the capacity can be utilized. The colored areas indicate the achievable number of homes served depending on average consumption per home. Blue represents mid band only while red is used for the mid band + mmWave case. As specified in the case study assumptions, the average consumption during peak hours is 2GB in the DL, and then a maximum of 270 homes can be connected per site. With the addition of mmWave spectrum, this number rises to 810, corresponding to 69 percent of all homes if there are 150 homes per square kilometer.

The curves represent two hypothetical growth scenarios, with 100 homes connected in the first year and an additional 50 homes added per year. The dark grey curve represents the baseline scenario starting from 600GB per month and 28 percent annual growth in data consumption, while the light grey curve is the all-broadband-media scenario with higher initial consumption but slower growth. The curves cross at year four, at which point we expect most homes have become all-broadband-media homes, with a corresponding lower annual growth of 10 percent after this point. We clearly see the benefit of adding mmWave spectrum, but also its necessity to handle larger market shares.

The reason for the large capacity improvements that result from adding mmWave extended range is best explained by Figure 1. As mmWave can serve a large number of users close to the base station, it is possible to offload lower bands and use these for homes in less favorable locations instead. The further the reach of mmWave, the larger the offload. Consequently, even though many homes may be unable to use the mmWave spectrum, they will all benefit from the offload from mid band.

The number of homes able to use mmWave will vary by sector size (inter-site distance), foliage and terrain topology, as well as by means to improve the link budget, such as using higher transmitted power and better beamforming capabilities. The propagation conditions of the simulated scenario represent a flat area with a foliage level of around 15 percent, which corresponds to a typical situation in US states such as Illinois and Indiana. In our case, the mmWave coverage was 63 percent at an inter-site distance of 3km.

In addition, homes that can use mmWave can have significantly higher speeds, which opens up for differentiating service offerings with respect to data rates. Homes in favorable locations could be offered higher service levels with subscriptions of 1Gbps or higher to selected homes, using CPE with mmWave capability. Meanwhile, subscribers in less favorable locations, where the mmWave signal is not sufficient, could be offered less expensive mid-band CPE and lower speeds. A dedicated prequalification method would be needed to achieve this.

> “THE DEPLOYMENT OF MMWAVE RADIOS IN ADDITION TO MID-BAND RADIOS RESULTS IN THREE TIMES HIGHER CAPACITY”

Result: Three times higher capacity

To summarize the case study, the deployment of mmWave radios on the macro sites in addition to mid-band radios results in three times higher capacity compared to mid-band-only deployment. It can handle both the baseline scenario and the all-broadband-media scenario with a margin, including realistic growth rates for both MBB and FWA. The network will be able to serve more than 400 homes per site with an average consumption of 4GB per busy hour, corresponding to a monthly DL consumption of 1.2TB. In addition, mmWave spectrum opens up for differentiated speed offerings with 1Gbps+ subscriptions to prequalified homes. Again, it is worth pointing out that the solution is a joint FWA and MBB deployment and that we also utilize the radio resources of legacy FDD bands.

Conclusion

As billions of people continue to wait for reliable fixed broadband connections, fixed wireless access (FWA) is an efficient and scalable alternative with significantly faster time to market. The extended range of millimeter wave (mmWave) spectrum is an innovation that, in combination with mid band, further enhances the comprehensive 5G FWA solution and enables profitable use of mmWave spectrum. With the capacity offload that mmWave enables, mid band can serve a larger number of homes at more distant, challenging locations, making it possible for communication service providers to offer high-end “wireless fiber” services in sparser suburban and semi-rural areas and make substantial progress toward closing the digital divide.

Further reading

- **Ericsson, Fixed Wireless Access, available at:** *https://www.ericsson.com/en/fixed-wireless-access*
- **Ericsson, Leveraging the potential of 5G millimeter wave, available at:** *https://www.ericsson.com/en/reports-and-papers/further-insights/leveraging-the-potential-of-5g-millimeter-wave*
- **Ericsson, 5G RAN, available at:** *https://www.ericsson.com/en/ran*

References

1. **Ericsson Mobility Report, June 2022, available at:** *https://www.ericsson.com/en/mobility-report/ reports/june-2022*
2. **5G Americas white paper, Fixed Wireless Access with 5G Networks, November 2021, available at:** *https://www.5gamericas.org/wp-content/uploads/2021/11/5G-FWA-WP.pdf*
3. **IEEE, 97th ARFTG Microwave Measurement Conference, Extended Range mmWave for Fixed Wireless Applications, June 2021, available at:** *https://ieeexplore.ieee.org/document/9734723*
4. **ZDNet article, NBN approaches 1Gbps using mmWave 5G over distances of 7 kilometres, January 12, 2021, Duckett, C, available at:** *https://www.zdnet.com/article/nbn-approaches-1gbps-using-mmwave-5g-over-distances-of-7-kilometres/*
5. **Ericsson press release, UScellular, Qualcomm, Ericsson, and Inseego Address Digital Divide with Multi-Gigabit Extended-Range 5G Milestone Over mmWave, May 6, 2021, available at:** *https://www.ericsson.com/en/press-releases/6/2021/5/uscellular-qualcomm-ericsson-and-inseego-address-digital-divide-with-multi-gigabit-extended-range-5g-milestone-over-mmwave*
6. **Ericsson, TIM, Ericsson and Qualcomm set record for long-distance speed with 5G mmWave, December 4, 2020, available at:** *https://www.ericsson.com/en/news/2020/12/mmwave-speed-distance-record*
7. **UScellular press release, UScellular, in Collaboration with Qualcomm and Inseego, Launches 5G mmWave High-Speed Internet Service in 10 Cities, April 28, 2022, available at:** *https://newsroom.uscellular.com/uscellular-qualcomm-inseego-launches-5g-mmwave-high-speed-internet-service-in-10-cities/*
8. **Ericsson, Bridging the digital divide: Extended-range millimeter-wave 5G Fixed Wireless Access, available at:** *https://www.ericsson.com/en/cases/2022/bridging-the-digital-divide-with-fwa-uscc*
9. **Ericsson, Fixed Wireless Access Handbook 2021, 4th edition release, available at:** *https://foryou.ericsson.com/fixed-wireless-access-new-handbook-2021.html*
10. **LightReading, Average data consumption eclipses half a terabyte per month – OpenVault, March 1, 2022, Baumgartner, J, available at:** *https://www.lightreading.com/cable-tech/average-data-consumption-eclipses-half-terabyte-per-month---openvault/d/d-id/775689*
11. **OpenVault, OVBI Broadband Insights Report Q4 2021, January 2022, available at:** *https://openvault.com/resources/ovbi/*

The authors would like to thank Ericsson's partners in the numerous mmWave extended range trials, as well as Michael Kühner, John Yazlle and Ali Moradian for their contributions to this article.

THE AUTHORS

Anders Ericsson

◆ joined Ericsson in 1999 and currently works as a system designer at Business Area Networks. During his time with the company, he has worked at Ericsson Research and in system management, as well as heading up the Algorithm and Simulations department at Ericsson Mobile Platforms/ST-Ericsson. Ericsson holds a Lic. Eng. in automatic control and an M. Sc. in applied physics and electrical engineering from Linköping University, Sweden. Ericsson is one of the coauthors of the FWA Handbook.

Laetitia Falconetti

◆ joined Ericsson in 2008 and is a strategic product manager responsible for 5G radio-access network (RAN) software solutions in the areas of coverage and Ericsson Spectrum Sharing. She has been driving Ericsson's mmWave long range initiative from its early days. Previously, she worked as the company's 3GPP

standardization delegate on 4G latency and reliability improvements and at Ericsson Research on innovative 4G software algorithms. Falconetti holds a Ph.D. in electrical engineering from RWTH Aachen University, Germany.

Håkan Olofsson

◆ has served in several capacities since joining Ericsson in 1994, mostly dealing with strategic technology development and the evolution from 2G to 5G. Olofsson is currently head of the System Concept program at Development Unit Networks, focusing on innovative use cases and RAN solutions for 5G and 6G. He holds an M.Sc. in physics engineering from Uppsala University, Sweden. Olofsson is one of the coauthors of the FWA Handbook.

Jonas Edstam

◆ joined Ericsson in 1995 and currently works with portfolio management for 5G RAN, with a focus on FWA. Throughout his career, he has served in various leading roles, working on a wide range of topics. The commercial use and evolution of mmWave applications is his passion. He has more than 25 years of expertise in wireless backhaul. Edstam holds a Ph.D. in physics from Chalmers University of Technology in Gothenburg, Sweden. Edstam is one of the coauthors of the FWA Handbook.

Tomas Dahlberg

◆ joined Ericsson in 1995 and is currently responsible for FWA technical sales. He previously held various management positions within R&D and product management. Dahlberg holds an M.Sc. in computer science and technology from KTH Royal Institute of Technology in Stockholm, Sweden, and a B.Sc. in business administration and economics from Stockholm University. Dahlberg is one of the coauthors of the FWA Handbook.

Network digital twins – outlook and opportunities

Digital twins that are tailored to the requirements of individual use cases have significant potential to create value in telecommunications processes, ranging from R&D to network operations such as deployment, management and site engineering.

PETER ÖHLÉN, CIARAN JOHNSTON, HÅKAN OLOFSSON, STEPHEN TERRILL, FEDOR CHERNOGOROV

A digital twin is a digital representation of a real-world object synchronized at a specified periodicity and fidelity. The original concept emerged decades ago, based on ideas that first arose in science fiction and in the Apollo program [1].

■ The digital twins used today in industries such as aerospace, automotive, energy and manufacturing have been created fit for purpose to improve processes, products and business outcomes by replicating the relevant aspects of reality in a virtual environment. They add value by combining data and knowledge with various analytics and visualization tools, and they are regularly synchronized to keep the real and virtual worlds in sync with each other.

In recent years, there has been a growing interest in applying the digital twin concept on a much broader scale. In the case of mobile networks, digital twins can be applied to a broad set of use cases within both the communication service providers' and the vendors' own organizations, making it possible to both enhance existing capabilities and introduce entirely new functionality.

Characteristics of a digital twin

Both the high interest in digital twins and their diverse applicability have resulted in many definitions of the term [1]. At Ericsson, we have adopted the Digital Twin Consortium's [2] definition for mobile networks. This considers a digital twin to be a virtual representation of real-world network entities and processes, along with their environment and users, which is synchronized at a specified periodicity and fidelity, and which provides added value for specific purposes.

A digital twin is based on having access to accurate, relevant and timely data, and it utilizes advanced analysis, simulation and visualization tools. Interaction with a digital twin results in the creation of a knowledge base related to the twin and its real-world counterpart. Irrespective of the realization, there are a couple of characteristics of digital twins that are essential. First, a suitable data structure is needed to represent the real-world state and knowledge at a level of detail that is appropriate to the scenario, ranging from low-level hardware through to aggregated characteristics of services and networks.

Second, there must be a continuous synchronization between the real world and the twin, and the two must evolve in parallel. Depending on the scenario, the time scale can be from milliseconds to days or longer. In most use cases, the dominant data flow is of measurements and events from the real world asset to the twin, which are needed to characterize performance and behavior. Configurations and control actions can also be pushed back to the real world using appropriate control mechanisms.

Overview of digital-twin use cases in mobile networks

There are several use cases with varying scopes where network digital twins (NDTs) have the potential to create value. A good way to categorize NDTs is to identify the real-world target for the twin, the process to which it applies and its purpose. The considered processes can be diverse, ranging from R&D to network control to network operations such as deployment and site engineering.

In some cases, a real-world NDT target can be the physical environment such as a cell site [3], accurately modeling the site structure, cabling and equipment. This gives the operator a complete view of the site,

“ THERE ARE SEVERAL USE CASES WITH VARYING SCOPES WHERE NDTs HAVE THE POTENTIAL TO CREATE VALUE ”

which can be used to simulate rollout and design processes, and to support troubleshooting and site upgrade processes.

NDTs can also represent multiple cell sites in a geographic environment, creating models for radio propagation based on measurements and physics-based simulations. The use of an NDT to optimize transmit power levels in network operations [4] is a good example. Similar approaches could be applied to other key performance indicators (KPIs) like coverage and throughput, in city districts, factory sites, workplace environments or combinations thereof.

The NDT target could also be the logical network – a single node, a network domain or the end-to-end (E2E) network, depending on the target use case. Here, the NDT can be used to observe and experiment with network configuration to improve network operations. A comprehensive NDT implementation will often incorporate aspects of both the logical network and the physical environment.

NDTs can be a huge help in simulating network expansion processes, to select site locations and model the impact of new technologies, new equipment and radio frequency (RF) expansion, as well as helping to predict upcoming capacity

Terms and abbreviations

AGV – Automated Guided Vehicle | **AI** – Artificial Intelligence | **BSS** – Business Support Systems | **CSI** – Channel State Information | **E2E** – End-to-End | **GPU** – Graphical Processing Unit | **KPI** – Key Performance Indicator | **LA** – Link Adaptation | **MCS** – Modulation and Coding Scheme | **MIMO** – Multiple-Input, Multiple-Output | **NDT** – Network Digital Twin | **NR** – New Radio | **OSS** – Operations Support Systems | **RAN** – Radio-Access Network | **RF** – Radio Frequency | **RRM** – Radio Resource Management | **SINR** – Signal-to-Interference Noise Ratio | **URLLC** – Ultra-Reliable Low-Latency Communication

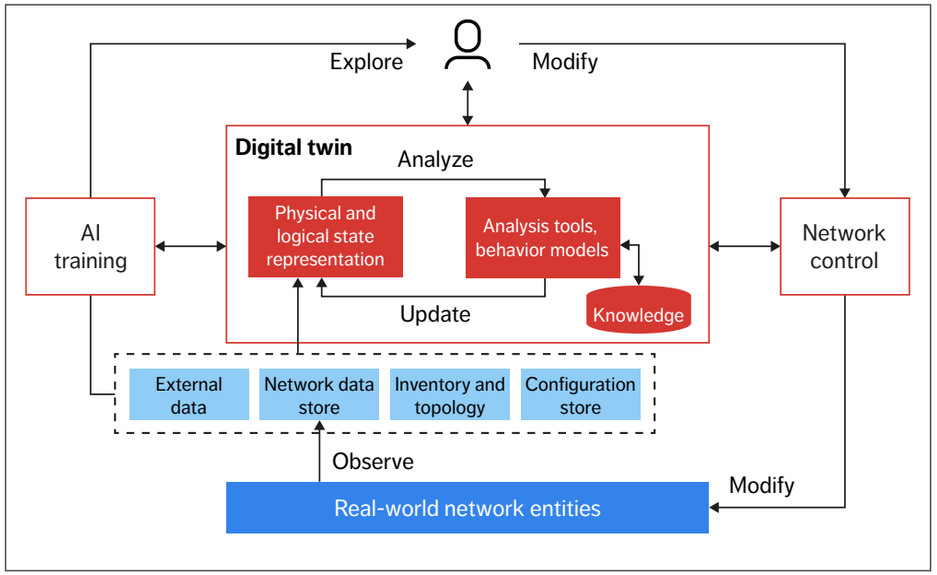


Figure 1 A network digital twin's main elements and environment

bottlenecks, enabling preemptive recommendations for network growth.

Another way that NDTs can add value is by predicting the impact of operator actions – either by humans or by autonomous systems [5] – that might have detrimental side effects. Using an NDT, potential actions could be evaluated before they are implemented. Similarly, new machine-learning algorithms can be safely trained and validated in the twin environment and only introduced in the real network once they have achieved a satisfactory level of performance.

The twin of the E2E network can add value by analyzing changes in consumer services. This type of NDT can evaluate the cost and performance of a new network slice before deployment and assess its impact on existing services. Based on this analysis, workload placement and network configuration can be further optimized to meet the required characteristics.

Yet another type of NDT can add value by evaluating various what-if scenarios such as equipment failures, security attacks or power loss to define contingency plans and network improvements.

NDTs can also address the need to represent traffic and user behavior. In these cases, the NDT needs to be tuned according to its use, from aggregate or detailed traffic characteristics and application flows to behavior in relation to subscription offerings and plans. By using anonymization techniques of sufficient accuracy, it is possible to realize these types of twins while also respecting privacy and safeguarding personal data.

Beyond the many benefits that NDTs can deliver for public networks, they also have great potential to add value in industrial connectivity use cases. To enable the integration of a mobile network in Industry 4.0 environments, the 5G Alliance for Connected Industries and Automation (5G ACIA) has defined the 5G network asset administration shell [6], which extends digital twins for factory management to include representation of industrial 5G.

Realization of a network digital twin

In an ideal world, an NDT would be a perfect copy of its real-world entity. But in reality, some degree of simplification based on domain knowledge is usually needed to make it feasible. At the core of the NDT are the data and knowledge that describe the relevant aspects of the real world, complemented with tools that create insights from the data captured, as depicted in *Figure 1*.

Several real-world entities are represented in an NDT. To capture their state and relations, an NDT aggregates data from multiple sources, including inventories, configuration settings and measurements performed at different levels of the network. Depending on the purpose, additional data from external sources, such as geospatial mapping information, may be required to build the full picture.

The inventory and configuration of the network can be accessed through operations support systems (OSS) that can correlate network application topologies with the topology and connectivity provided by virtualization, hardware and transport systems.

The sheer volume and constant flow of new data about measurements in the network make data management systems [7] essential. Data in different forms, including files, streams or discrete events, is all gathered and made available as a common reusable data asset for different consumers. As new data becomes available, a digital twin updates its models.

To perform the analysis in the twin and extract the necessary knowledge, it is necessary to apply a suitable tool or model that can perform the tasks at hand, fulfilling requirements such as response time and accuracy for the given use case. Analysis can be based on calculation, simulation or an artificial intelligence (AI) algorithm. Typically, this involves

“ AN NDT AGGREGATES DATA FROM MULTIPLE SOURCES, INCLUDING INVENTORIES, CONFIGURATION SETTINGS AND MEASUREMENTS ”

the prediction of system behavior based on the current state.

The role of simulators

Our view of the role of simulators is that they are part of the digital twin's analysis toolbox used for knowledge extraction. One of the key differentiators between conventional simulators and digital twins is that the latter has regular synchronization with the real-world entity.

It is possible to make highly detailed simulations. However, there is a trade-off between the size of the simulated scenario, the time it takes to evaluate and the required computing capacity. Smaller scenarios can be simulated quickly, whereas larger scenarios require a longer execution time. In many use cases, fast response times are important, which leads to the need for simplified models focusing on the most relevant aspects, or to limiting the scenario size to a subset of the network.

New simulation models and the evolution of computing hardware continue to expand the boundary of what is possible, enabling more advanced and realistic simulations. More sophisticated models require the data to be captured in greater detail than it is at present. Often, the selection of data and modeling needs to be defined iteratively, where models are refined and data selection is updated until a satisfactory balance between accuracy, evaluation time and cost has been reached. AI models have a significant role to play, whether they are used for predictions, data generation or creating black-box models based on observed behavior, further driving the need for data availability for training and optimization.

> SIMULATIONS OF NETWORKS AND THEIR USERS IS A KEY METHOD EMPLOYED IN RESEARCHING AND DEVELOPING NEW RAN FUNCTIONALITY

The integration of a digital twin within the network operational processes requires some consideration, both to invoke the digital twin and to take advantage of the resulting insights. For human-centered use cases, visualization is an important part of the twin, whereas other use cases require an application programming interface to integrate the twin within an automated process. Depending on the process for which the digital twin is used, it may be located in the operations support system, the business support system (BSS), the network functions or even in the R&D domain of a vendor.

Two promising use cases

Two of the many NDT use cases that have the potential to deliver significant value are those that can carry out network performance evaluations and those that can handle Radio Resource Management (RRM) in factory environments.

Use case #1: Network performance evaluation

Simulations of networks and their users is a key method employed in researching and developing new radio-access network (RAN) functionality, when evaluating network performance of different RAN products and deployment strategies, and when exploring ideas for "the next G." There is a rich history of radio network simulation models both at Ericsson and in the wider telco industry – from 3GPP (3rd Generation Partnership Project) models and common RAN deployment planning tools to more refined proprietary RF propagation models. They all capture deep physical properties of RF transmission and deliver predictive accuracy.

The evolution from today's radio network simulators to NDTs that utilize decades of investment in 3D gaming and computer-generated imagery technologies is expected to unlock significant benefits. New features and functionality will include:

- High-resolution and complex city or indoor geometry (bridges, tunnels, foliage, indoor to outdoor and more)
- Detailed surface materials that influence RF propagation, such as metallic coatings and metal features in places such as factories
- Representation of user mobility and dynamic scene features such as automotive traffic
- The ability to explore complex models visually
- Internal and external collaboration and data sharing
- Extremely high computational complexity of physically accurate models.

When 3D gaming technology is applied to the development of relevant models for 5G radio networks, the result is an attractive combination of gaming and radio technologies – for example, the use of Graphical Processing Unit (GPU) hardware to model radio propagation [8]. Games deal with incredibly detailed and complex scene geometry, apply fine-grained textures, and have highly evolved "actor" AI that provides scene dynamics, which is appealing for representing a network.

Technological advances have improved our capability in several aspects: visualization, integration, standardization of formats, collaboration and modeling accuracy. Together, these enable leaps in our ability to accurately simulate a radio network in a digital twin.

To illustrate the power of combining these capabilities, *Figure 2* shows an example from a network simulation study, in which we collected data from the real world, imported it and used it to create a set of simulated models. The real-world section on the left side shows a car that is connected to Ericsson's cellular network driving down a Stockholm street. On the right side of the figure, the digital twin dynamically illustrates the resulting massive multiple-input, multiple-output (MIMO)

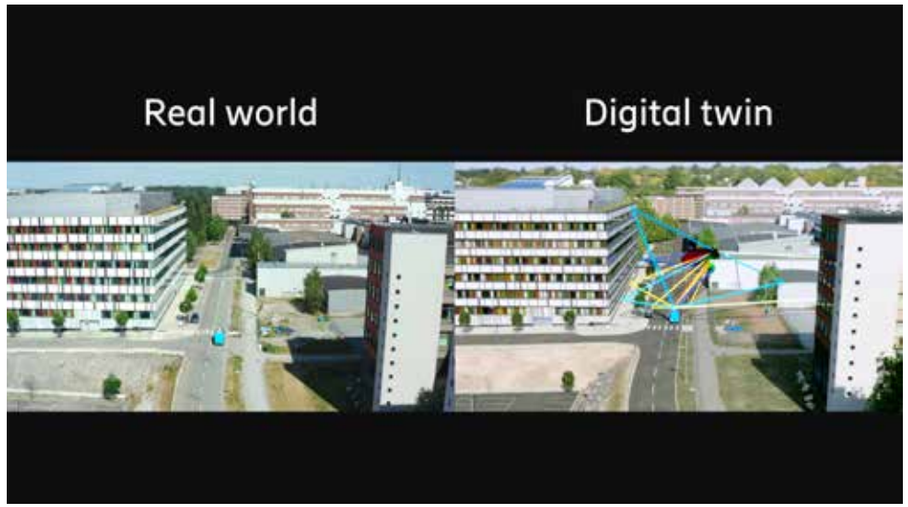


Figure 2 Example of a real-world city environment (left), accurately modeled in a digital twin simulator (right)

antenna and signal propagation paths, thereby making it possible to analyze them.

The data used in the simulation shown in Figure 2 included the city environment details down to the level of building materials, roof shapes and windows. The network deployment included network equipment deployment information as well as the users' mobility pattern. This was combined with detailed ray-tracing models of propagation and models of radios and the 5G New Radio (NR) layer 1-3 as specified in 3GPP to create an NDT representation of the whole system under study, enabling extremely accurate network simulations of the 5G NR system in this "real" environment.

An example of RAN functionality that benefits from this type of detailed evaluation is massive MIMO interference sensing. Using an NDT, we can determine the impact of interference on the beamforming action of the base station. Beyond enabling the extraction of very accurate statistical performance measurements available from ordinary network simulators, the NDT is able to provide such a high level of detail that it is also possible to illustrate concrete beam patterns toward single users, as shown on the right side of Figure 2. Similarly, when an NDT is applied to interference sensing, it can show how the radio network can avoid directing power toward users within its interference range, thereby lowering interference levels and increasing performance.

This simulation capability enables a much more detailed R&D analysis in the digital world, allowing us to develop RAN solutions and products to address the real challenges much faster than before. Further, it enables the early detection of problems that affect product design long before products are deployed, thereby reducing R&D costs, improving time to market and ensuring the reliable performance of products when they are deployed in real networks.

Use case #2: Radio resource management in factories

One important way that NDTs can deliver benefits in factory deployments is by enhancing the

Figure 3 Illustration of a factory floor with 5G-connected stationary robotic arms and moving AGVs

performance of the industrial cellular network. This is achieved through cooperation between an NDT in the RAN system and a factory digital twin with the ability to provide information about stationary and moving factory objects. The NDT uses the information from the factory digital twin to enable better link adaptation (LA) decisions from the 5G base station scheduler.

We used a dynamic system-level simulator to accurately model the factory scenario illustrated in *Figure 3*. The stationary connected robotic arms, colored according to their connecting cell, periodically send data packets to 5G base stations placed at a height of 6m, while metallic-gray automated guided vehicles (AGVs) carrying 2mx4m containers move around the production shop floor. We evaluated the performance of factory network at millimeter wave frequencies where the impact of blockage on radio wave propagation is especially high.

The conventional LA method uses signal-to-interference noise ratio (SINR) information based on measurement history, delivered in a channel state information (CSI) report. In digital-twin-based LA, the SINR information is gathered with the help of an NDT that predicts the path loss, which is impacted by the locations of the blockers, and performs inter-cell interference coordination. The NDT achieves the former by receiving accurate data about the location of major obstacles from an AGV controller or similar entity. To achieve the latter, the NDT controls the time-frequency resource allocations made by 5G schedulers in each cell. These two capabilities enable the NDT to perform a more optimal modulation and coding scheme (MCS) selection than conventional LA, going beyond performance of the well-known inter-site interference coordination-based scheduling.

The table shown in *Figure 4* presents the results of our evaluation of the two LA algorithms in terms of median spectral efficiency, two different latency percentiles and the reliability of packet transmissions. Although the CSI measurement reporting is delayed by a certain number of NR slots, the conventional LA algorithm performs reasonably well, capturing both interference and path-loss variation to a certain extent. Our results show that the achieved reliability and latency of conventional

Link adaptation algorithm	Median spectral efficiency	Latency 90th percentile	Latency 99th percentile	Reliability % (1-packet error rate)
Conventional LA	0.3 bits/s/Hz	0.5ms	20.66ms	99.89%
Digital-twin-based LA	4.2 bits/s/Hz	0.125ms	0.125ms	99.99%

Figure 4 Performance comparison for conventional and digital-twin-based link adaptation algorithms

LA will be sufficient for most non-URLLC (Ultra-Reliable Low-Latency Communication) industrial applications, but not for URLLC services, which require at least 99.99% reliability and under 1ms latency.

The performance of the DT-based LA algorithm is significantly better than that of the conventional LA algorithm and therefore better suited for industrial URLLC services. This is because the DT-based LA algorithm can make more accurate predictions of the SINR, resulting in higher spectral efficiency, lower latency and two orders of magnitude higher reliability. This performance improvement is particularly important for industrial applications that depend on reliable connectivity, indicating that digital twins have the potential to significantly improve functions such as RRM in industrial 5G networks.

In short, the results of our evaluation show that the DT-based algorithm derives a significant performance gain from the integration between the factory digital twin and the NDT.

Standards and industry alignment

Initial efforts to define a framework for NDTs have been made in standards bodies such as the IETF (Internet Engineering Task Force) [9]. Due to the diversity of potential NDT use cases, we believe that the value of standards for NDTs will be in terms of providing alignment on terminology and defining a high-level architectural framework without being so specific as to inhibit innovation. The various types of NDTs each have their own needs in terms of data and characteristics, as well as different starting points. A high degree of flexibility will be required to support innovation both in terms of evolving existing functionality and introducing entirely new functionality.

Conclusion

Network digital twins (NDTs) have the potential to deliver massive benefits to mobile networks by supporting use cases in areas ranging from R&D to planning, deployment and operations. We foresee that there will be many types of NDTs in the not-so-distant future, each designed to fit the process it is intended to support. Some will emerge as an evolution of existing functionality, while others will be new creations that deliver novel capabilities. Depending on their purpose, NDTs may reside with the communication service provider and/or with their vendors.

To ensure that NDT solutions have the greatest possible impact, we believe it is wise to begin by identifying opportunities to reuse existing functionality such as management functions and simulation and analysis tools, while also ensuring that other necessary enablers, such as access to the required data, are in place. As digital twins proliferate, it is likely that individual NDTs will be combined both with each other and with industrial digital twins to create increasingly sophisticated digital representations that can deliver ever more powerful functionality.

Further reading

- **Ericsson, Ericsson's 5G Digital Twin Simulated in NVIDIA Omniverse (video), available at:** *https://www.youtube.com/watch?v=yTbUSXJ8M-8&t=35s*
- **Ericsson blog, Next-generation simulation technology to accelerate the 5G journey, available at:** *https://www.ericsson.com/en/blog/2021/4/5g-simulation-omniverse-platform*
- **Ericsson blog, Using digital twins to be in control of your network assets, available at:** *https://www.ericsson.com/en/blog/2022/5/using-digital-twins-to-be-in-control-of-your-network-assets*
- **Ericsson blog, Digital twins: what are they and how are they enabling future networks, available at:** *https://www.ericsson.com/en/blog/2022/3/what-are-digital-twins-three-real-world-examples*
- **Ericsson blog, The future of digital twins: what will they mean for mobile networks, available at:** *https://www.ericsson.com/en/blog/2021/7/future-digital-twins-in-mobile-networks*
- **Ericsson blog – intelligent deployment and maintenance of network sites , available at:** *https://www.ericsson.com/en/network-services/deployment/intelligent-site-engineering*
- **Future technologies: digital twins – bridging the physical and virtual world, available at:** *https://www.ericsson.com/en/about-us/new-world-of-possibilities/imagine-possible-perspectives/digital-twins*

References

1. **IEEE Access, vol. 7, pp. 167653-167671, A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications, 2019, Barricelli, B.R.; Casiraghi, E; and Fogli, D, available at:** *https://ieeexplore.ieee.org/document/8901113*
2. **Digital Twin Consortium, Digital Twin Consortium Defines Digital Twin, December 3, 2020, Olcott, S; Mullen, C, available at:** *https://blog.digitaltwinconsortium.org/2020/12/digital-twin-consortium-defines-digital-twin.html*
3. **Ericsson blog, Using digital twins to be in control of your network assets, May 3, 2022, Kirac, E; Björkander, P; Garnett, N, available at:** *https://www.ericsson.com/en/blog/2022/5/using-digital-twins-to-be-in-control-of-your-network-assets*
4. **Ericsson blog, Digital twins: what are they and how are they enabling future networks?, March 31, 2022, Muñiz, C, available at:** *https://www.ericsson.com/en/blog/2022/3/what-are-digital-twins-three-real-world-examples/*
5. **Ericsson Technology Review, Creating autonomous networks with intent-based closed loops, April 19, 2022, Niemöller, J; Szabó, R; Zahemszky, A; Roeland, D, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/creating-autonomous-networks-with-intent-based-closed-loops*
6. **5G-ACIA white paper, Using Digital Twins to Integrate 5G into Production Networks, February 2021, available at:** *https://5g-acia.org/whitepapers/using-digital-twins-to-integrate-5g-into-production-networks/*
7. **Ericsson Technology Review, Data ingestion architecture for telecom applications, March 16, 2021, Rönnberg, A-K; Åström, B; Gecer, B, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/data-ingestion-architecture-for-telecom*
8. **Next-generation simulation technology to accelerate the 5G journey** *https://www.ericsson.com/en/blog/2021/4/5g-simulation-omniverse-platform*
9. **IETF Datatracker, Digital Twin Network: Concepts and Reference Architecture (work in progress/internet draft), Zhou, C et al., available at:** *https://datatracker.ietf.org/doc/draft-irtf-nmrg-network-digital-twin-arch/*

THE AUTHORS

Peter Öhlén

◆ is a principal researcher whose work centers on service and network automation. He joined Ericsson in 2005 and has more than 25 years of experience working with fixed and wireless networks, network management and cloud systems. His current focus is on realizing the cognitive network vision based on a foundation of data, AI and a flexible end-to-end architecture. Öhlén holds a Ph.D. in photonics from KTH Royal Institute of Technology in Stockholm, 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 use cases and RAN solutions for 5G and 6G. Olofsson holds an M.Sc. in physics engineering from Uppsala University, Sweden.

Stephen Terrill

◆ is a senior expert and chief architect in automation and management at Ericsson. Since joining the company in 1994, he has worked primarily in telecommunications architecture, implementation and industry engagement. In recent years, his work has focused on the automation and evolution of OSS. Terrill holds an M.Eng.Sc. from the University of Melbourne, Australia.

Fedor Chernogorov

◆ is a master researcher who joined Ericsson in 2018. He has 12 years of experience in applied research in wireless communications and simulation tools. Chernogorov is currently leading a research project dedicated to E2E cellular network solutions for the digitalization of enterprises, industries and societies. His scientific interests are centered around cellular networks for industrial use, ultra-reliable low-latency communication and digital twins. Chernogorov holds a Ph. D. in mathematical information technology from the University of Jyväskylä, Finland.

Realizing 5G smart-port use cases with a digital twin

5G New Radio is the most effective and cost-efficient technology available to meet the connectivity requirements of advanced automation use cases in the smart ports of the future. To accurately dimension and realistically model the performance of a private 5G network in a smart-port environment, we have created a digital twin powered by state-of-the-art graphical processing unit computing.

RONG DU,
AHSAN MAHMOOD,
GUNTHER AUER

5G technology is a cornerstone in the development of the smart ports of the future, supporting the uptake of use cases that involve everything from transmitting video streams recorded by multiple cameras to enabling remote operators to monitor and control connected cranes and automated guided vehicles (AGVs).

■ Most smart ports today rely heavily on wired sensors, cameras and other data-gathering devices, a solution that has delivered significant benefits but which also has obvious limitations. The wireless connectivity that 5G New Radio (NR) can provide enables a range of new smart-port use cases that are impossible to connect with wires, as well as making it possible to remove the wires on existing use cases, thereby offering enhanced flexibility at a lower cost [1]. With all the capabilities of 5G NR, the vision of a wirelessly connected port is significantly closer to becoming an everyday reality.

Smart-port use cases

A recent Ericsson report identified several promising use cases for 5G-connected smart ports [1], including remote-controlled ship-to-shore (STS) cranes, automated rubber-tired gantry (RTG) cranes and AGVs.

Remote-controlled ship-to-shore cranes

The purpose of STS cranes is to move containers between ships and the dock. They are increasingly remote-controlled and supervised by operators sitting in a control hub. Remote crane operation is a commercial solution that vendors such as ABB provide. Control data with strict latency and reliability requirements is exchanged between cranes and a remote hub. Each crane is equipped with 3D sensors and HD video cameras that monitor the surroundings from various angles. Video streams are transmitted continuously, which imposes stringent requirements in terms of bandwidth, latency and reliability.

Device type	Video cameras per device	Bitrate per camera [Mbps]	Number of devices	Total video traffic per device [Mbps]	Remote control data per device [kbps]	Network latency (one way) [ms]
STS crane	up to 20	8-20	10 per km of quayside	200	<600	15-25
RTG crane	up to 20	8-20	3-4 times the number of STS cranes	<200	<600	15-25
AGV	up to 2	8-20	2-3 times the total number of cranes	20	<600	15-25

Figure 1 Requirements to support wireless crane and AGV use cases

Automated rubber-tired gantry cranes

Automated RTG cranes are used to stack containers. The number of RTG cranes is typically three to four times the number of STS cranes. In terms of sensors and cameras, RTG crane equipment is similar to that of STS cranes. However, automated RTG cranes can carry out many of their operations in an autonomous manner, which means that video streams may not be transmitted continuously and that a crane operator may supervise several cranes simultaneously. An automated RTG crane operator only takes over using remote control to resolve extraordinary incidents.

Automated guided vehicles

AGVs are used extensively for container handling in ports today. They navigate through restricted areas in the port guided by a safety controller, which requires a reliable exchange of positional (and other) data gathered by 3D sensors using wireless connectivity. AGVs are also equipped with video cameras for supervision and human intervention, especially for navigating through areas where people are located.

Network requirements of the key use cases

Figure 1 provides a summary of the requirements of wirelessly connecting cranes and AGVs. High-quality video streams and control information are the main types of data to be exchanged. Cranes equipped with multiple HD cameras will generate significant data volumes, which poses a major challenge to the capacity of a wireless network. AGVs, on the other hand, are likely to be equipped with only one or two cameras. Furthermore, both cranes and AGVs exchange control data with a remote hub using standardized communication protocols for automation, such as PROFINET [2]. This requires low latency and high reliability from the wireless network.

Definition of key terms

A **smart port** makes use of automation and innovative technologies including artificial intelligence to enhance both efficiency and safety, while simultaneously lowering costs. In most cases, today's smart ports use wired networks for connectivity.

A **wirelessly connected port** uses wireless technology such as Wi-Fi, 4G or 5G for connection.

In a **5G-connected port**, 5G NR delivers ubiquitous and reliable wireless connectivity with low latency that enables next-level automation and provides port operators with a comprehensive, end-to-end view of their operations, right across the terminal.

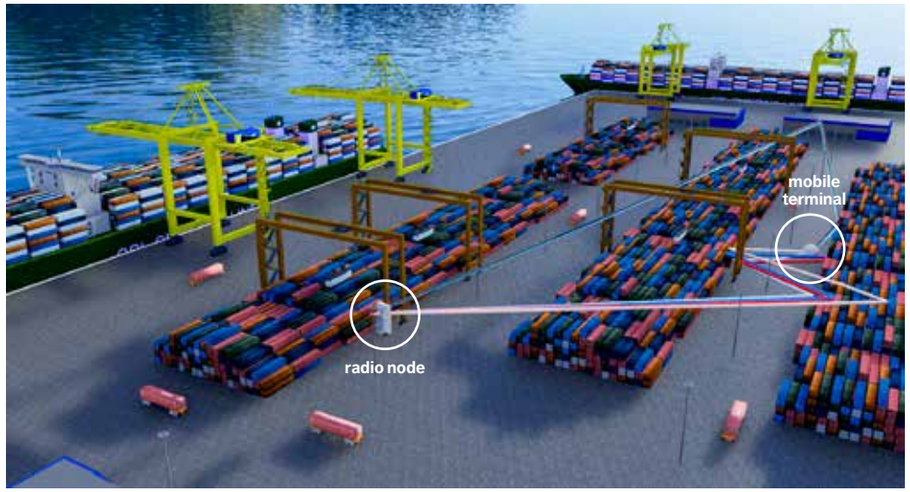


Figure 2 Digital 3D model of a seaport showing radio propagation between a radio node (mounted on a lamppost) and a mobile terminal

Creating a digital twin

The most efficient way to evaluate network performance in a 5G-connected port is to create a digital twin. By serving as a virtual representation of a real-world entity, a digital twin enables accurate assessment of the functionality of that entity without having to actually build it in the physical world. The digital twin of a wireless network is created by combining a digital 3D model with the ability to accurately model propagation effects and emulate radio network procedures and capabilities.

Figure 2 is a snapshot from the digital twin we created using a realistic 3D model of a seaport that we imported from a commercial supplier [3]. It includes photorealistic models of ships, containers, and STS and RTG cranes, and comprises close to 10 million vertices. In its current form, it is 0.35km wide and 1km long, which amounts to an area of 0.35sq km, comprising a small to medium-sized port. The digital 3D seaport model may be reconfigured to resemble a port of any size, allowing for the creation of a digital twin of any physical port.

Graphical processing unit accelerated computing

While conventional simulation techniques would mandate a vast reduction in the level of detail of the digital 3D seaport model, graphical processing unit (GPU) accelerated computing offers unprecedented opportunities and elevates the way digital simulations are conducted to the next level. GPU-based rendering and visualization originally emerged in the movie and gaming industries and have revolutionized 3D graphics. More recently, GPUs have been adopted for a multitude of applications, ranging from machine learning, robotics and computer-aided design to various kinds of artificial intelligence.

In the context of running computer simulations for a wireless network, ray-tracing GPUs have emerged as a disruptive enabler by facilitating complex radio propagation modeling that delivers an unprecedented level of detail compared to traditional central processing units and ray-tracing techniques. However, the application of commercial ray-tracing GPUs for radio propagation modeling is far from straightforward. The difference in wavelength between rays of visual light and radio frequencies is several orders of magnitude, giving rise to different propagation effects, such as diffraction and diffuse scattering.

To realize the potential of GPU-accelerated computing, Ericsson is cooperating with NVIDIA, a leading manufacturer of GPUs, in high-fidelity radio propagation modeling and visualization of 3D graphics [4]. The ambition is to be able to process any advanced 3D graphics model represented by triangles, from a real city, over stadiums and factory halls to a seaport [5].

Radio propagation

The seaport model shown in Figure 2 illustrates the valid radio propagation paths between a radio node mounted on a lamppost at a height of 20m and a mobile terminal positioned at a height of 1.5m. A propagation path may be composed of up to three specular reflections, diffractions and diffuse scattering on rough surfaces. Using the digital twin of the seaport, we are able to test various smart port use cases without a real deployment in a physical port.

Spectrum option	Carrier frequency	Bandwidth	Antenna configuration (VxHxP)	TDD pattern
Mid-band TDD	3.5GHz	100MHz	4x8x2	4:1
High-band TDD	28GHz	800MHz	16x24x2	4:1

Figure 3 5G NR band and antenna configuration

How 5G capabilities meet the use-case requirements

5G NR is the most effective and cost-efficient technology available to meet the connectivity requirements of the smart ports of the future. In most markets around the world, 5G spectrum is allocated on mid band between 3GHz and 5GHz, as well as on high band above 25GHz, which is also known as millimeter wave (mmWave).

Figure 3 illustrates a representative 5G NR band configuration applicable to the majority of markets, which we have used to realize smart port use cases. The antenna configuration is denoted in the form VxHxP, where V, H and P account for the number of radio chains in the vertical, horizontal and polarization domain, respectively [6].

Without resorting to sophisticated radio network simulations it is clear that mid band, with 100MHz bandwidth, lacks the capacity to serve the bitrates generated by STS and RTG cranes. With eight times the bandwidth, high band appears more suitable. On the other hand, mid band provides better coverage, due to the larger radio wavelength, and may therefore be the better option to serve AGVs, which

Terms and abbreviations

AGV – Automated Guided Vehicle | **DL** – Downlink | **GPU** – Graphical Processing Unit | **LoS** – Line-of-Sight | **MIMO** – Multiple-Input, Multiple-Output | **MU-MIMO** – Multi-User MIMO | **NR** – New Radio | **RTG** – Rubber-Tired Gantry | **SINR** – Signal-to-Interference-Plus-Noise Ratio | **STS** – Ship-to-Shore | **SU-MIMO** – Single-User MIMO | **TDD** – Time Division Duplex | **UE** – User Equipment | **UL** – Uplink

roam around the port and may suffer from shadowing in non-line-of-sight (LoS) scenarios.

Challenges

A port presents significant challenges to the capacity of a wireless network that mandate a dense deployment of base stations. The combination of short distances between base stations and the relatively open propagation environment of most ports can cause excessive levels of inter-cell interference.

Stringent requirements on reliability and bounded latency add further challenges to the ability to deliver the required network capacity. Moreover, in a smart-port scenario, as in many other industrial applications, most data is generated on the UL, while most public 5G networks serving mobile broadband services are configured to predominantly serve data in the DL. This is reflected in the DL-heavy time division duplex (TDD) pattern of 4:1, which assigns four times as many resources to the DL than the UL and is being adopted in most public networks globally.

Unfortunately, regulatory requirements often mandate the same DL-heavy 4:1 TDD pattern even for dedicated private networks aimed to serve industrial applications. Coexistence with public networks that often operate in adjacent bands further complicates the adoption of a more balanced TDD pattern for private networks, especially on mid band.

Massive MIMO

Massive multiple-input, multiple-output (MIMO) is a cornerstone of 5G NR, and its ability to boost the capacity of wireless networks is well-documented [6, 7]. It is based on adaptive antenna arrays that employ numerous radio chains. Mid-band antenna arrays typically have a size of 16-64 radio chains, while the array dimension for high band often approaches several hundred. As shown in the antenna configuration column in Figure 3, our work is based on the assumption of antenna arrays with 64 radio chains for mid band and 768 radio chains for high band.

Massive MIMO is composed of spatial multiplexing and beamforming. Spatial multiplexing means superimposing multiple parallel data streams

> MASSIVE MIMO IS A CORNERSTONE OF 5G NR AND ITS ABILITY TO BOOST CAPACITY IS WELL-DOCUMENTED

on the same time-frequency resources, while beamforming adaptively shapes the signal energy in the spatial domain to direct the signal toward the desired destination. Beamforming improves coverage by enhancing the link quality as well as mitigating the detrimental effects of inter-cell interference by reducing the spillage of signal energy elsewhere.

5G NR distinguishes between two types of spatial multiplexing: single-user (SU) and multi-user (MU) MIMO, as well as combinations thereof. For SU-MIMO, all spatial data streams serve one user, while for MU-MIMO spatial streams are directed to different users who are well separated spatially.

It has been demonstrated that the combination of beamforming, SU-MIMO and MU-MIMO enhances the capacity of a wireless network sixfold, both in dense urban and suburban scenarios [6]. Given the 12-times larger array dimension, beamforming gains on high band are significantly larger than for mid band. The fact that beamforming in the user equipment (UE) is a common feature in high band further mitigates inter-cell interference.

Realization of the smart-port use cases

Using our digital twin, we have conducted radio network simulations to assess the ability of 5G NR to serve the wireless crane and AGV use cases depicted in the 3D port model in Figure 2. Base stations are deployed on lampposts 20m above ground. Each site is equipped with three sectors that have a service area of 120 degrees in azimuth. Deployments ranging from four to eight sites are simulated, which equates to 12 to 24 sectors. The equivalent inter-site distance is between 320m and 225m, which is in line with contemporary mobile networks in dense urban environments.

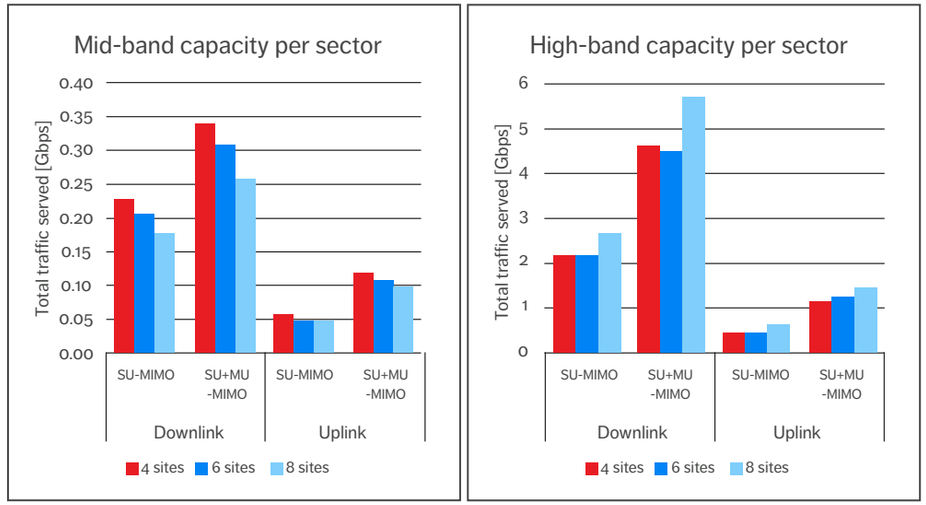


Figure 4 5G NR capacity achieved on mid and high band for both the DL and the UL

Figure 4 shows the capacity of a 5G network in terms of the data volumes served per sector in gigabits per second. The throughput requirement in the DL and UL is set to 100Mbps and 20Mbps, respectively. Owing to the DL-heavy TDD pattern, DL capacity significantly exceeds UL capacity. Likewise, the capacity achieved on high band is significantly superior to mid-band capacity, due to the larger bandwidth.

Interestingly, the capacity per sector decreases for mid band with an increasing number of sectors, while it increases for high band. This means that for mid band, doubling the number of sectors from 12 to 24 boosts capacity by about 60 percent, compared with a capacity gain of 150 percent for high band.

The reason for this is twofold: first, in contrast to mid band, coverage on high band improves when the number of sites increases from four to eight. Second, the larger array dimension in high band produces a narrower beam, mitigating inter-cell interference and giving rise to an improved signal-to-interference-plus-noise ratio (SINR).

In absolute terms, high-band capacity is more than tenfold in many cases, while there is only an eightfold difference in bandwidth. Clearly, the UL on mid band is the bottleneck in meeting the capacity demands of the considered use cases. Utilizing the full UL feature set, including the combined use of SU-MIMO and MU-MIMO, is therefore of paramount importance.

Figure 5 shows the 5G network capacity achieved with 24 sectors for the wirelessly connected crane and AGV use cases. A port of this size typically accommodates 40-50 cranes and 80-150 AGVs that predominantly produce UL traffic. This equates to about two cranes and three to six AGVs per sector that need to be served. A significantly larger number

> THE CAPACITY ACHIEVED ON HIGH BAND IS SIGNIFICANTLY SUPERIOR TO MID-BAND CAPACITY, DUE TO THE LARGER BANDWIDTH

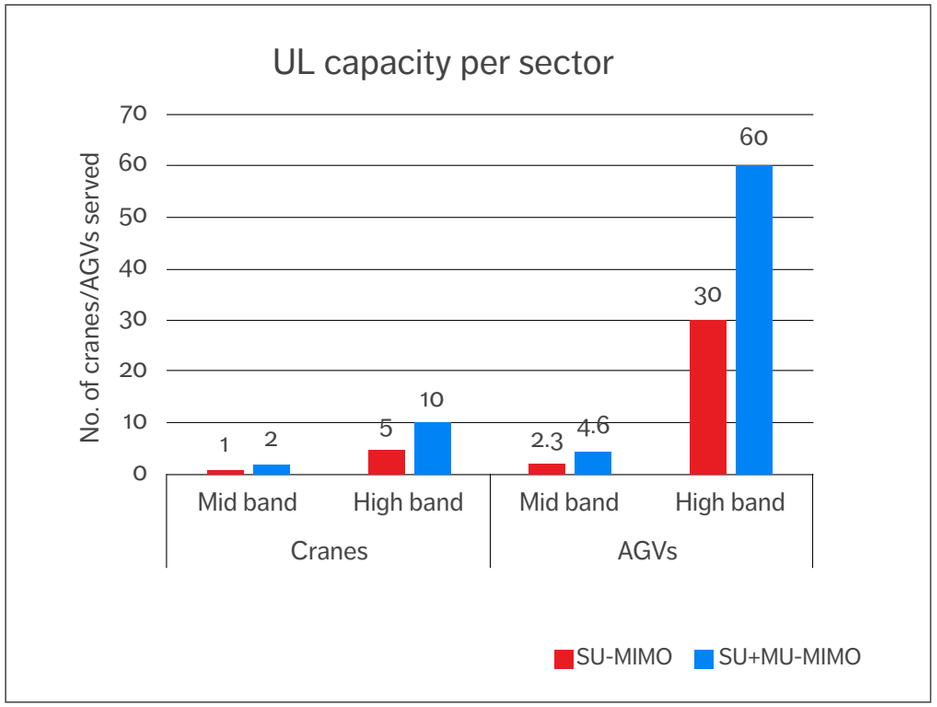


Figure 5 Number of 5G-connected cranes and automated guided vehicles served for mid and high band

of devices can be supported by suspending transmissions or reducing the number of active video streams when an AGV or an automated RTG is in a safe area where unsupervised autonomous operation is possible.

However, cranes and AGVs move around the port, which may lead to an uneven distribution of devices that calls for an overprovisioning of sector capacity to avoid congestion in cells that temporally accommodate many devices.

We assume that each 5G-connected crane produces 200Mbps of data. On mid band the required SINR to serve one or two cranes per sector with SU-MIMO and MU-MIMO respectively is prohibitively high, which effectively rules it out for connecting cranes. In contrast, up to 10 cranes per sector may be served on high band. Since the height of the cranes allows for the placement of UE antennas above container stacks and most other obstacles, thereby ensuring a stable LoS connection to the base station, high band appears suitable for this use case.

Given the throughput requirement of 20Mbps per AGV, the MU-MIMO capacity on mid band of serving up to five AGVs may not be sufficient to meet the target of six AGVs per sector. This is in sharp contrast with high band, which significantly exceeds the AGV requirement. However, unlike cranes, AGVs move around more freely and UE antennas must be mounted at low heights, well below the height of a container, increasing the possibility of blocking the direct LoS propagation path. These shadowing effects may compromise the reliability of the wireless link especially for high band, making mid band the preferred choice for serving AGVs.

Bridging the gap

There are four methods that may help to bridge the gap between the mid-band capacity that is delivered and that which is required. The first method involves using a more UL-heavy TDD pattern that allows for a better match between the supply and demand of UL capacity. Unfortunately, however, regulatory constraints and coexistence with public networks imply that changing the TDD pattern may be difficult, especially for mid band.

The second method involves increasing bandwidth by combining the spectrum assets of dedicated private and public networks. As seaports are restricted areas, public consumption of mobile broadband services is likely to be low in the vicinity of the port. Some of the spare capacity in public networks could therefore be redirected to serve the smart port.

The third method is to configure a dual-band 5G network with carrier aggregation between mid and high band. This approach makes it possible to dynamically distribute traffic between mid band and high band to serve AGVs in LoS and non-LoS conditions, respectively.

The fourth method is to employ adaptive codecs that allow for rate adaptation of the video streams. Morever, the number of simultaneously active video streams per crane may also be adjusted to reduce the bitrate requirement consequently improving the network capacity.

Delivering reliable connectivity at bounded latency

In addition to serving the required traffic volumes, a connected port demands reliable connectivity at bounded latency. Our simulations indicate that high band achieves lower latencies than mid band, due to shorter transmission and reception cycles. The main benefit of mid band, on the other hand, is that ubiquitous connectivity is maintained, whereas high band may suffer from spotty coverage due to shadowing in non-LoS conditions. Both mid and high band are therefore indispensable components in connecting the smart ports of the future.

> BOTH MID AND HIGH BAND ARE INDISPENSABLE COMPONENTS IN CONNECTING THE SMART PORTS OF THE FUTURE

Conclusion

The digital twin of a smart port that we have created at Ericsson makes it possible to realistically model the performance of a 5G network in a port environment. Our research indicates that massive multiple-input, multiple-output (MIMO), including beamforming and multi-user MIMO, will play a key role in fulfilling smart-port requirements.

One of the main challenges in delivering wireless connectivity in a port is the imbalance between the supply and demand of uplink (UL) and downlink (DL) capacity, due to the DL-heavy time division duplex pattern that is predominantly used in 5G networks today. Significant benefits will be gained from using the different frequency bands allocated to 5G to serve different purposes in a port. The high UL capacity needs of wirelessly connected cranes require the use of high-frequency bands in the millimeter wave range, while the mid-band frequencies below 6GHz will provide reliable connectivity for use cases with unrestricted mobility, such as automated guided vehicles.

Further reading

» **Ericsson, Smart ports: At the gateway to a new shipping age, available at:** *https://www.ericsson.com/en/industries/ports*

» **Ericsson, Massive MIMO, available at:** *https://www.ericsson.com/en/portfolio/networks/ericsson-radio-system/radio/macro/massive-mimo*

» **Ericsson blog, IoT condition monitoring in smart ports, available at:** *https://www.ericsson.com/en/blog/2022/8/iot-condition-monitoring-smart-ports*

References

1. **Ericsson, Connected Ports: A guide to making ports smarter with private cellular technology, February 2021, available at:** *https://www.ericsson.com/en/enterprise/forms/connected-ports*
2. **PROFIBUS & PROFINET International, PROFINET, the leading Industrial Ethernet standard, available at:** *https://www.profibus.com/technology/profinet*
3. **TurboSquid, Sea Port 3D model by onurzgen, available at:** *https://www.turbosquid.com/3d-models/port-sea-3d-model-1347032*
4. **NVIDIA, GTC November 2021 Keynote with NVIDIA Founder and CEO Jensen Huang, GPU Technology Conference (GTC) November 2021, [Online from 45:57 to 48:09], available at:** *https://www.nvidia.com/en-us/on-demand/session/gtcfall21-a31660/*
5. **IEEE.tv, Advanced High-Fidelity Channel Modelling and Methodology – Demo – 2021, Brooklyn 6G Summit (B6GS), Asplund, H, et al., available at:** *https://ieeetv.ieee.org/channels/communications/advanced-high-fidelity-channel-modelling-and-methodology-demo-2021-b6gs*
6. **Academic Press, Advanced Antenna Systems for 5G Network Deployments: Bridging the Gap between Theory and Practice, 2020, Asplund, H; Astely, D; Larsson, E, et al., available at:** *https://www.ericsson.com/en/reports-and-papers/books/aas-for-5g-deployments*
7. **Ericsson Technology Review, Meeting 5G network requirements with Massive MIMO, February 16, 2022, Astely, D; von Butovitsch, P; Faxér, S; Larsson, E, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/using-massive-mimo-to-meet-5g-network-requirements*

THE AUTHORS

Rong Du

◆ joined Ericsson in 2020 and he is currently employed as a researcher in the Technology and Strategy Unit where he works on concept development and network performance for 5G with a focus on the Internet of Things and industrial applications of 5G. Du holds a Ph.D. in electrical engineering from the KTH Royal Institute of Technology in Stockholm, Sweden.

Ahsan Mahmood

◆ joined Ericsson in 2018. In his role as a researcher in the Technology and Strategy Unit, he works on radio network deployment and performance evaluation for 5G with a particular focus on cellular Internet of Things and 5G for industries. Mahmood holds a Ph.D. in electrical, electronics and communications engineering from the Polytechnic University of Turin, Italy.

Gunther Auer

◆ joined Ericsson in 2012, and currently works as a researcher in the Technology and Strategy Unit. His research interests include new concepts and performance evaluations of radio-access networks in realistic environments, including heterogeneous networks, indoor environments and high-capacity venues such as stadiums. Auer holds a Ph.D. in electrical engineering from the University of Edinburgh in the UK.

This work has been carried out in collaboration with ABB, whose contribution in defining use cases and providing constructive feedback and suggestions is greatly appreciated. The authors would also like to thank the GPU computing team at Ericsson, including Magnus Lundevall, Leo Frössling and Johan Janzén, who helped in scenario modeling and simulations. We also thank Johan Torsner, Mats Buchmayer, Lisa Boström, and Jan Diekmann for their contributions to this article.

Monetizing API exposure for enterprises with evolved BSS

The ability to expose application programming interfaces opens up the opportunity for communication service providers to strengthen their role in the enterprise ecosystem by enabling new use cases, applications and programmable networks that deliver greater flexibility and automation. Evolved business support systems are essential to commercialize these offerings and support new business models.

JAN FRIMAN,
ELISABETH MUELLER,
BART VAN KAATHOVEN

Regardless of whether an enterprise uses a private (on-premises) or public network, emerging use cases in the enterprise ecosystem require that it has the ability to automate the management of connectivity and connected devices. The only way to achieve this automation is through the exposure of application programming interfaces (APIs).

■ Enterprise use cases increasingly have specific requirements that demand tailored connectivity solutions and flexible, efficient self-service capabilities. In their work to meet these requirements, enterprise application developers prefer to develop applications generically across platforms. Depending on the service or application they are creating – which could be a gaming app, a boost application or an enterprise management application, for example – their service API exposure needs can vary greatly. As a result, the services that need to be exposed range from network capabilities and network information to management and orchestration.

Enterprises expect applications in the communication service provider (CSP) domain to be able to capitalize on unique CSP capabilities to optimize the user experience. To build applications that are tailored to meet these needs, enterprise application developers require easy access to network information, influence over network behavior and the ability to manage and orchestrate enterprise connectivity demands simply and efficiently. The network services that are exposed today are not sufficient in this respect, as they require detailed telecom industry knowledge that is cumbersome to access and differs from one CSP to another. This is a major challenge that CSPs must overcome to meet enterprise use case requirements.

Meeting enterprise requirements

The speed of growth in the service exposure business will depend on how quickly CSPs can meet the key requirements. First and foremost, service APIs must be easy to use, access and consume. To attract developers, the service APIs need to hide telco complexity and should be harmonized as much as possible across CSPs. Second, the service exposure platform must support a variety of channels for API exposure, including direct exposure, communication platforms, marketplaces of platform providers and service aggregators. Finally, the services must have the right scope and be able to address specific enterprise needs.

To engage successfully in the service exposure business, CSPs will need enhanced business support systems (BSS) and operations support systems (OSS) functionality to support the customer and partner management APIs, as well as a range of different business and monetization models. BSS functions play a particularly critical role in the service exposure domain because they are necessary to implement the management and orchestration services that will be exposed to developers.

Beyond that, they are also critical for service enablement and exposure monetization. Both the BSS and OSS need to be addressable through APIs from different exposure channels (multi-country, multi-CSP and so on) depending on the use case. They also need to have the flexibility to support the exposure business models.

To scale effectively, service APIs must fulfill the requirements of simplification and standardization. Developers of any type of application need standardization and harmonization across CSPs, along with the ability to reach other CSPs through a single technical and commercial interface. They benefit from global tools and software development kits (SDKs) that can be reused across CSP APIs. These APIs need to act the same way across all CSPs without the need for CSP-specific information, while at the same time allowing inter-CSP API communication to reach customers of other CSPs or to address multi-country enterprise services.

Before they can go into production in the enterprise ecosystem, service APIs must go through a three-step process: API development, API setup in the CSP network and, finally, customer onboarding, which is when the customer can use the service API from within its own application. At this point, the service APIs can be reached both by applications on devices and by those on the server side. They can be used by both enterprise and consumer applications, made available on marketplaces, or consumed by API aggregators that form new services on top of the CSP's APIs. A CSP's BSS and OSS stack must support all of these different types of API consumers.

Exposing network service through application programming interfaces – technology aspects

By transforming their connectivity network into a platform for flexible service creation and innovation, CSPs can deliver significant value for enterprises that stretches far beyond connectivity alone. As ease of use is such a critical factor for enterprises and developers, there is much to gain from presenting the network to them as a black box that provides APIs that can easily be integrated with applications, managed by either their IT or operational

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **API** – Application Programming Interface | **B2B2X** – Business-to-Business-to-X | **BSS** – Business Support Systems | **CSP** – Communication Service Provider | **eSIM** – Embedded SIM | **iSIM** – Integrated SIM | **KPI** – Key Performance Indicator | **ODA** – Open Digital Architecture | **OSS** – Operations Support Systems | **OT** – Operational Technology | **POS** – Point-of-Sale | **QoS** – Quality of Service | **REST** – Representational State Transfer | **SDK** – Software Development Kit | **SLA** – Service Level Agreement | **SIM** – Subscriber Identity Module

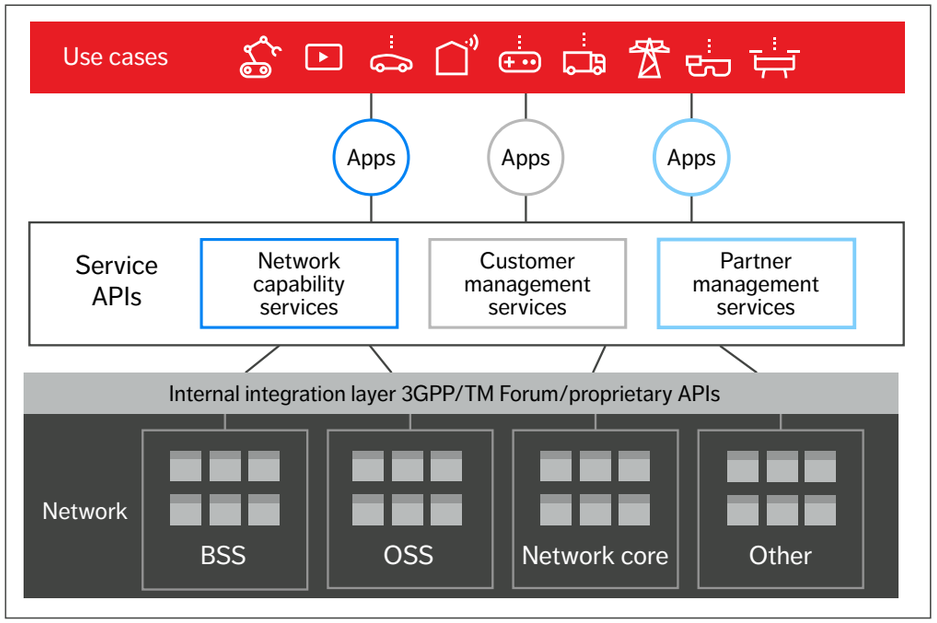


Figure 1 Visualization of the network as a black box that provides easy-to-use service APIs

technology (OT) organizations. *Figure 1* visualizes our approach.

Different APIs enable different use cases, but rather than being specific to a particular vertical use case, many of them will address the needs of the case in terms of service management, connectivity, application management and so on. In addition, all APIs must comply with the same architecture standards when it comes to design, deployment, granting access and allowing for monetization.

For this model to work, the CSP's service exposure architecture must be integrated with BSS that can manage the ecosystem partners and support various business models for monetization.

Figure 2 provides a comprehensive overview of all the CSP IT and network functions that are engaged in service exposure and monetization, in alignment with TM Forum's Open Digital Architecture (ODA). The figure combines the functions for exposing, realizing and monetizing service APIs in one architecture and emphasizes that the functions available in the IT domain and the network play multiple roles.

The service exposure platform must support the automated life-cycle management of services. The service monetization architecture combines the service exposure domain with the business domain. It enables awareness of the API invokers (the applications that call the APIs and the parties operating the applications) and API usage in the core commerce systems as well as monitoring and assuring service operations.

Service implementation consists of service APIs that are logically placed in the engagement layer of the ODA and interact southbound with the functions from the IT domain and the network. The southbound interaction utilizes the standardized APIs (from TM Forum's ODA and the 3GPP

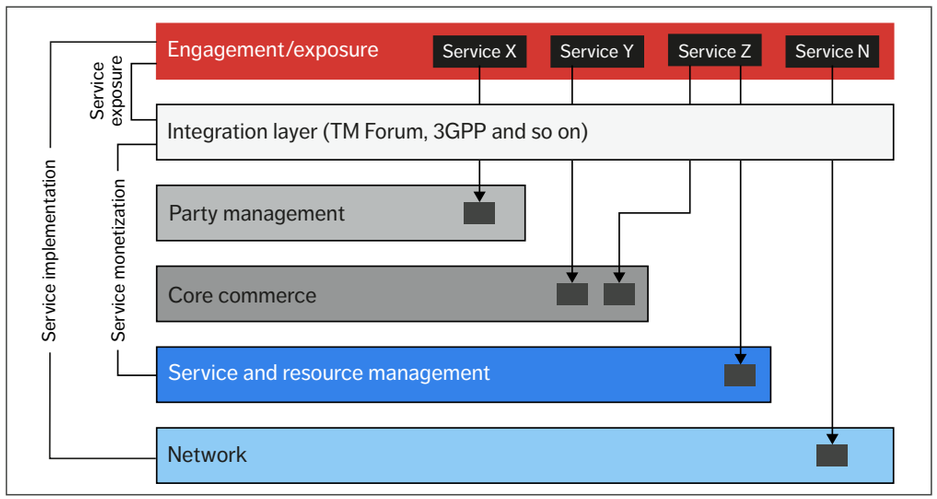


Figure 2 CSP functions engaged in service exposure and monetization

(the 3rd Generation Partnership Project), for example) and proprietary APIs that are exposed from the various functions through the internal exposure layer.

Service exposure

An efficient and scalable service exposure architecture requires a service exposure platform that provides gateway functionality as well as a set of centralized supporting functions responsible for authentication and authorization of API invokers, API management and some functions for developers (such as API catalogs). The service exposure platform belongs to the engagement and exposure domain. It also serves as the provider domain for the new service APIs. It is therefore desirable if the service exposure platform can also provide mechanisms and tools for service creation.

All APIs made available on the service exposure platform (described in an API/technical catalog) need to be modeled as commercial offerings in the commercial product catalog to allow them to be sold and/or to grant access of them to future API consumers. The API invokers are applications, which are operated by parties acting as API customers. The API customers must therefore be onboarded, and all parties that may be engaged in service monetization have to be made known in the CSP's BSS. The service monetization architecture must support a variety of business models depending on the use case, the service API in question, the chain of intermediatory actors and the business rationale behind the service exposure.

APIs are used by various stakeholders that are served by the BSS and OSS. These include the initial application developer who develops the application that invokes APIs to the network and the partner onboarding developer who enables the onboarding of the partner and the partner applications to the CSP's core commerce system and the network. The developer of the customer onboarding functions is another important stakeholder, with responsibility for ensuring that the CSP's customers' applications can connect to the service APIs as a subscription or connectivity management service, for example. When the application is in service, the exposure function is responsible for monitoring key performance indicators (KPIs) and providing this information to the application developer, the party

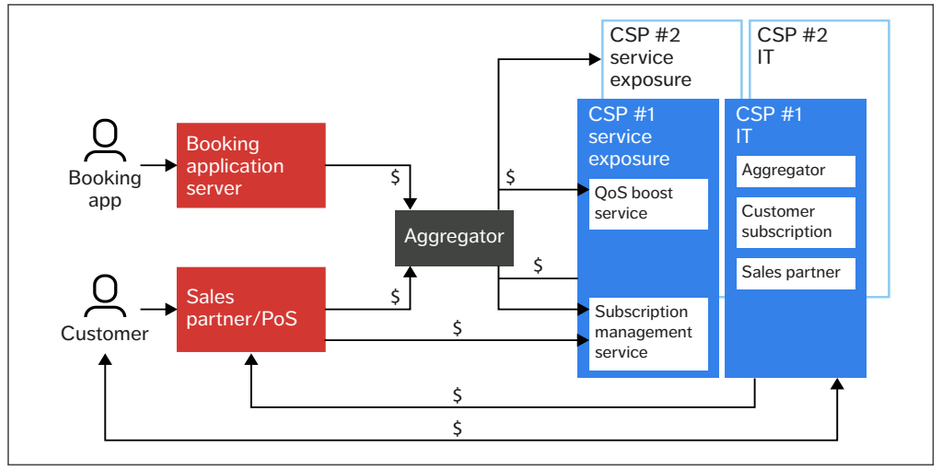


Figure 3 Two use cases for service APIs

operating the application and the service provider operating the service APIs.

Each role in the ecosystem has BSS and OSS needs. For example, the application developer needs to have a relationship with a CSP, a service aggregator or a developer community that includes access to a developer portal and the ability to download the proper documentation, SDK and tools, as well as access to testing and other CSP-based development tools. The partner needs to have a commercial and operational relationship with the CSP based on functionality, volume or any other measurement, to facilitate application usage by its customers. The customer needs to be able to sign up for the service (either an end-service or the API as a service), connect to it and be charged for it. When all the different roles are connected and set up, it is essential that the CSP monitors API exposure and ensures compliance with the agreed access KPIs as well as recording all transactions related to monetization, logging and security.

Service monetization

Several functions in the BSS/OSS domain are responsible for managing the commercial aspects of service exposure, covering a variety of business processes in the areas of product and service design, party management and core commerce, as shown in the middle of Figure 2. Together, these functions enable the onboarding of the service APIs as commercial assets into the product and service catalog. They also manage the parties engaged in API exposure and consumption, register the applications operated by customers or partners as API invokers and bill for the API invocations (or alternatively, for the assets that the API invocations create and life-cycle manage).

The service and resource management functions are responsible for orchestrating the exposed services and for monitoring service operation, which includes providing information about service behavior to Service Level Agreement (SLA) management. The core commerce functions integrated into the service exposure domain are service monetization responsibilities and must support various business models.

Figure 3 illustrates two different use cases that demonstrate the need for maximal flexibility regarding business models and monetization approaches.

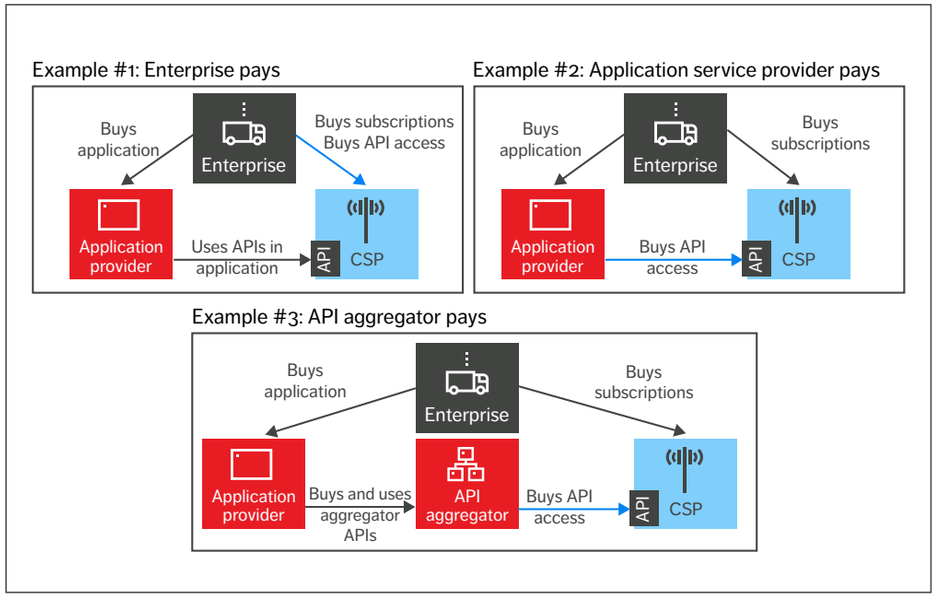


Figure 4 Examples of three business models for API access

In the upper use case, the operator of a booking application wants to improve app responsiveness for small data packages. The operator engages the services of an aggregator to trigger a Quality of Service (QoS) boost API that a CSP provides. The aggregator has API contracts with multiple CSPs and can therefore offer the QoS boost service across several (or all) of them. In this scenario, the CSP only has a business relationship with the aggregator. The API invocation will be monetized based on this relationship alone.

The lower use case in Figure 3 has a slightly more complicated setup. In this case, the sales partner of one or multiple CSPs operates a portal where consumers can buy subscriptions from CSPs. The sales partner interacts with the CSPs directly or through an aggregator. In this scenario, the CSPs must be able to support different monetization models so that, for example, both the sales partner and the aggregator can receive compensation for new subscriptions. It may also be the case that the aggregator or sales partner pays a fee to the CSP to make the subscription management service available and scalable.

Successful service monetization will require a high degree of flexibility in terms of supporting different business models related to API consumption. To illustrate this point in a slightly different way, *Figure 4* provides three examples of how a CSP can monetize service APIs through different kinds of relationships – that is, with the enterprise operating the application, with the application service provider or with an API

> SUCCESSFUL SERVICE MONETIZATION WILL REQUIRE A HIGH DEGREE OF FLEXIBILITY

> THE NETWORK SERVICE FOR CONNECTIVITY SERVICE MANAGEMENT IS INTENDED TO BE USED BY ENTERPRISES AND PARTNERS

aggregator. Another possibility is that application service providers and API aggregators may not pay for API access at all, but rather act as catalyzers of new enterprise and consumer business for CSPs. In this scenario, they would likely strive for a business model in which they receive compensation for facilitating subscriptions or add-on sales.

Service implementation

To meet the requirements of the enterprise ecosystem, service implementation must:

- Comply with open API standards
- Hide telco complexity by focusing on a use case and exposing specialized REST (representational state transfer) resources
- Orchestrate native telco functions southbound
- Comply to cloud-native standards that are micro-service based, modular and configurable to allow adaptation to different customer and market needs
- Use a development toolset that makes it possible to build services that can be natively integrated into the service exposure platform
- Support stateful and stateless services
- Provide duplicate detection for requests
- Support synchronous and asynchronous communication patterns towards the API invoker
- Have the ability to scale from handling low-traffic, high-value operations (such as a customer onboarding or refill request) to handling high-volume traffic (such as event streaming).

Exposed services must also seamlessly integrate with supporting functions such as access control, consent management, reporting, monetization and so on that service exposure platforms provide.

The list of service APIs that expose CSP network capabilities will expand over time as the need for new management and orchestration services continues to grow. The BSS/OSS domains in particular must support a service scope that addresses the most important management and orchestration needs of enterprises. These include:

- Connectivity service management and device connectivity management
- Edge application hosting, onboarding and registration
- Connectivity service quality monitoring and predictions
- Customer experience quality monitoring and predictions
- Services related to customer management supporting business-to-consumer, business-to-business and business-to-business-to-anything (B2B2X) models
- Services related to partner management
- Services related to carrier billing, cost control and sponsoring
- Advanced services for exposing network insights.

The network service for connectivity service management is intended to be used by enterprises and partners (on behalf of their enterprise customers) to browse for available offerings for connectivity services by indicating their traffic needs as a filter mechanism. They can then order a suitable connectivity service offering, personalize it with some limited parameters and read up on the installed base. The available connectivity service offerings will be based on preconfigured service templates with limited options for parametrization.

The service for connectivity service management needs to work together with a peer service for device connectivity management that will onboard device subscriptions onto the connectivity service sold to an enterprise. After onboarding, the devices will be able to send their traffic on the new connectivity

> SERVICE EXPOSURE REPRESENTS A SIGNIFICANT BUSINESS OPPORTUNITY FOR CSPs TO GENERATE VALUE BEYOND CONNECTIVITY

service. Onboarding will occur as the result of the sale of an access offering that refers to the target connectivity service.

Edge capabilities that optimize latency, security and reliability make it possible to bring services closer to the consumer. BSS and OSS need to support both CSP-owned and edge locations, so that services can be placed at the optimal location, connected to the customer, monitored and charged throughout the service life cycle. The CSP needs to expose capabilities for the application developer to upload their application or request infrastructure as a service with the requested SLA, to monitor resources and to influence their behavior if needed.

The importance of quality monitoring the delivered connectivity services and associated device connectivity sold to enterprises cannot be overstated. This information must be exposed to the customers to allow for service-quality verification as well as life-cycle management. Service-quality predictions are needed for service planning, including runtime distribution of applications.

With regard to services related to customer management, in the example of private networks, party management and resource management are involved in the B2B2X processes. These enterprises must be given the rights to life-cycle-manage their own subscribers. Enterprises that are using a private network also need to have the management of resources such as SIM/eSIM/iSIM cards delegated to them.

Services related to partner management, cost control and sponsoring are also important to consider. Partner management services make it easy to establish partnerships and manage partner offerings, which will enrich the CSP's portfolio. Sponsoring scenarios are popular for venues, sports sites, campus sites and company premises. A cost control service makes it possible to sponsor the connectivity of a user under certain conditions, such as when the user purchases goods, agrees to receive advertisements or participates in surveys. Carrier billing enables the use of a mobile pre- or post-paid wallet as a new payment mechanism.

Advanced services for exposing network insights can also add significant value for an enterprise. Radio coverage maps and user density maps are good examples of these kinds of services.

Conclusion

Service exposure represents a significant business opportunity for communication service providers (CSPs) to generate value beyond connectivity in the digital ecosystem. The first step toward success in this space is to make it as easy as possible for enterprise application developers to access and consume service application programming interfaces (APIs). The second step is of equal importance: CSPs' service exposure solutions must address the need for monetization across the whole ecosystem. Different stakeholders will be engaged in making the APIs available to the ecosystem, depending on the specific use case involved. In light of this, CSPs must provide flexible support for business models ranging from direct business relationships between the CSP and the API consumer to indirect ones facilitated by aggregators and service platforms.

Ericsson is committed to bringing together the perspectives of stakeholders across the value chain to ensure the delivery of market-leading solutions in this area. It is already clear that service exposure platforms will need to be integrated into the service API monetization architecture and support automation of the different business processes related to API invoker registration, API consumption and API monetization. The approach to service monetization that we have presented in this article addresses the commercial aspects of the whole life cycle of a service API, from the moment it is made available for consumption to the billing and settlement processes.

THE AUTHORS

Jan Friman
◆ is an OSS/BSS expert in the architecture and technology team within Business Area Cloud Software & Services. Since joining Ericsson in 1997, he has held various OSS/BSS-related positions within the company's R&D, system management and strategic product management organizations. Friman holds an M.Sc. in computer science from Linköping University, Sweden.

Elisabeth Mueller
◆ joined Ericsson in 2006 when LHS in Frankfurt, Germany, was acquired to complement the Ericsson BSS offerings with a billing system. She has taken on many roles since then, including system design, system management and solution architecture in all BSS areas. Mueller holds various patents within BSS and currently serves as a senior expert for monetization, partner and customer management, lately focusing on service exposure architecture for the digital economy. She holds an M.Sc. in mathematics and business economics from Johannes Gutenberg University Mainz, Germany.

Bart van Kaathoven
◆ is the software ecosystem lead to industries that have exposure needs from CSPs, working within the Architecture and

Technology team within Business Area Cloud Software & Services. He joined Ericsson in 1995 as a system administrator and webmaster and since then has held various customer and partner-engaging roles in OSS/BSS, value-added services, exposure and device applications, where he combines his strong IT and telecom knowledge.

Further reading

- **Ericsson Technology Review, Evolving BSS to fit the 5G economy, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/evolving-bss-to-fit-the-5g-economy*
- **Ericsson Technology Review, Service exposure: a critical capability in a 5G world, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/service-exposure-a-critical-capability-in-a-5g-world*
- **Ericsson, Digital BSS, available at:** *https://www.ericsson.com/en/portfolio/cloud-software--services/digital-bss*
- **Ericsson, Telecom BSS, available at:** *https://www.ericsson.com/en/telecom-bss*
- **Ericsson, Network exposure, available at:** *https://www.ericsson.com/en/service-orchestration/network-exposure*

Automating telecom software deployment with GitOps

A declarative, GitOps-based approach to software delivery and deployment enables communication service providers to increase automation and speed up the introduction of new features and updates – including security fixes – into their networks.

PETER WÖRNDLE, STEPHEN TERRILL, TORSTEN DINSING

An always-up-to-date software (SW) base will be a huge advantage for communication service providers (CSPs) by making it possible to rapidly introduce new features and updates that open up new business opportunities, improve efficiency and proactively address security threats. To make it happen, greater automation of SW delivery and operational processes will be essential.

■ Current telecom SW delivery and operation processes – in which even the smallest updates must undergo the same semi-frequent, complex processes as large packages that deliver new functionality – are simply too complicated, manual and time-consuming to support an always-up-to-date SW base. To ensure efficient delivery and deployment of network and operations support systems/business support systems (OSS/BSS) functions in the future, CSPs need to evolve their processes and tools to fit the new application architectures and incorporate the use of new technologies.

At Ericsson, we believe that the most effective approach to achieving the necessary automation of telecom SW delivery and operation processes is by adopting continuous integration/continuous deployment (CI/CD) best practices for SW delivery pipelines. Further, we recommend a microservice-based architecture of componentized cloud-native network functions (CNFs), which enables more targeted SW modifications, while also making it possible to maintain a larger number of SW artifacts. Evolving the application architecture and the underlying platform in this way also opens up the opportunity to use cloud-native best practices for life-cycle management (LCM).

With a DevOps mindset as our starting point, the automation efforts we present in this article focus on the implementation of a continuous SW flow and the creation of feedback loops between organizations that develop SW and those that operate it. Our concept separates LCM on the functional and realization domains and introduces new automation on top of the declarative deployment functionality, which makes it possible to separate the concerns between the network function (NF) operations and their realization. This approach is already widely adopted in various IT domains and ecosystems and therefore provides a solid baseline for adoption in the telecom industry.

Terms and abbreviations

API – Application Programming Interface | **BSS** – Business Support Systems | **CaaS** – Containers-as-a-Service | **CI/CD** – Continuous Integration/Continuous Deployment | **CNA** – Cloud-native Application | **CNF** – Cloud-native Network Function | **CSP** – Communication Service Provider | **E2E** – End-to-End | **LCM** – Life-cycle Management | **NF** – Network Function | **OSS** – Operations Support Systems | **SW** – Software

What is GitOps?

GitOps is a way of working that applies best practices in development to software automation. It uses a declarative approach that forms the basis of continuous everything. The main benefits of GitOps are that it:

- Establishes a single source of truth
- Enables the tracking and management of changes
- Provides stable and reproducible rollbacks to any previous version
- Reduces complexity by using cloud-native best practices.

Cloud-native application and life-cycle management: a joint evolution

As an emerging best practice, DevOps is starting to replace existing processes. During the past decade, there has been a technology evolution among CSPs from physical NFs toward virtual NFs, CNFs and cloud-native applications (CNAs). This has brought with it new LCM characteristics that have enabled and driven a coevolution from existing ways of working [1] to new practices. Most notably, development teams are now able to observe SW performance in live systems and quickly update CNAs in the case of unwanted behavior [2].

When CNAs are developed according to the 12-factor application principles [3], microservice in-service SW updates and the LCM of the underlying Kubernetes cluster and infrastructure do not impact the service provided by the running NFs or applications. This is because, in the cloud-native paradigm, a product is composed of microservices that each have their own independent life cycles. As a logical consequence of this, the life cycle of the CNA becomes decoupled from the life cycle of the service it provides. The service will continue while the underlying deployed SW can evolve and change.

Each microservice in a CNA is represented by a number of accompanying artifacts such as the container image, helm charts, Flux manifests and a deployment configuration in the form of values.yaml. These artifacts make it possible to deploy a microservice initially as part of a CNF. The same mechanism can be applied to post-deployment configuration, which is commonly referred to as Day-1/n configuration. The relationships and customizations of these artifacts can be formulated in a declarative description that explains what should be deployed in a Kubernetes cluster.

> IN THE CLOUD-NATIVE PARADIGM, A PRODUCT IS COMPOSED OF MICROSERVICES THAT EACH HAVE THEIR OWN, INDEPENDENT LIFE CYCLES

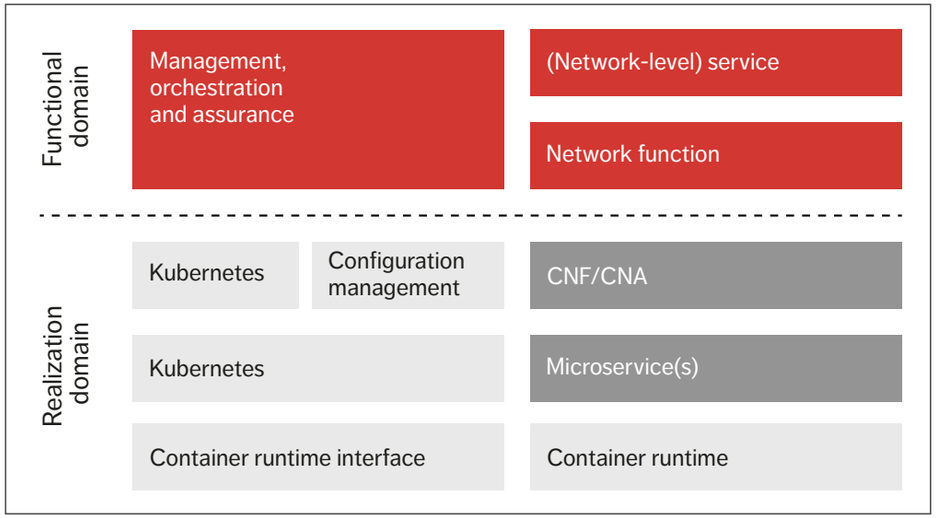


Figure 1 Separation of functional management from realization management

There are two key characteristics that enable the evolution of DevOps best practices and CI/CD pipelines to new GitOps ways of working:

1. Separation into functional and realization domains
2. Declarative description of complex deployments (including version control).

Enhancing the architecture with repositories and GitOps

The separation between the realization domain (how the service is realized) and the functional domain (the service itself) that is illustrated in *Figure 1* makes it possible to decouple their life cycles and therefore also decouple the life-cycle automation between them. Most importantly, this means that the SW life-cycle automation in the realization domain becomes a more generic issue that can be solved with applicable tooling from the cloud-native ecosystem. (Note that this approach is equally applicable to NFs and applications from OSS/BSS, which do not differ on the technology side.)

The functional domain is responsible for modeling and managing the functionality of a network service, determining which functions will be necessary to provide it. The realization domain, on the other hand, is responsible for the LCM of the artifacts that realize the functional components. Examples of artifacts in the realization domain are container images, Helm charts, Kubernetes manifests and configuration artifacts. The realization domain also manages the resources that the artifacts need in terms of compute, memory, storage and networking.

> THE SEPARATION BETWEEN THE REALIZATION DOMAIN AND THE FUNCTIONAL DOMAIN MAKES IT POSSIBLE TO DECOUPLE THEIR LIFE CYCLES

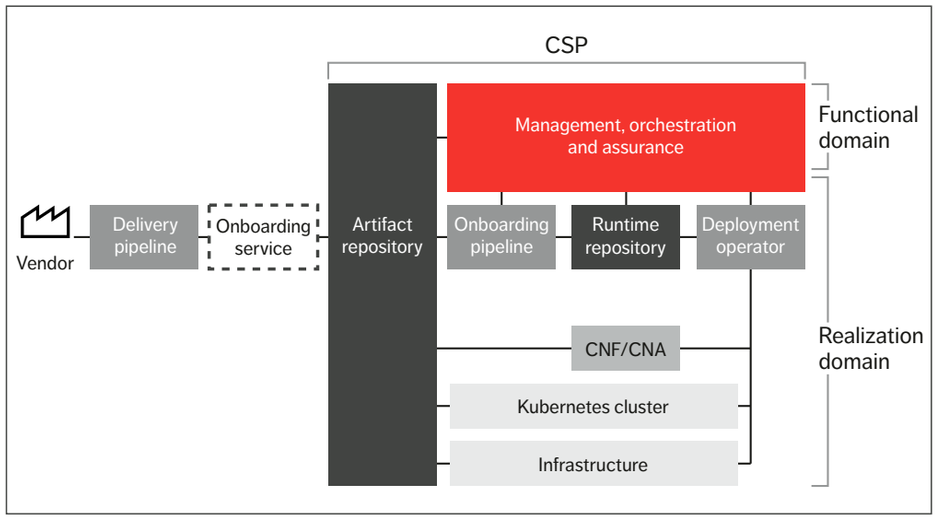


Figure 2 Target architecture for declarative deployments

Figure 2 depicts our proposal for an architecture based on the separation of realization and functional domain concerns. The demarcation between the domains is represented by the runtime repository, which contains the desired system state for the realization domain. The deployment operator ensures that the desired system state is applied to the target system: a Kubernetes cluster, a CNF or any other application. It enables application deployment through automatic reconciliation between the declarative description (desired deployment) and the actual deployment on the infrastructure, so that both are in constant sync.

The role of management, orchestration and assurance tools in this architecture is to determine the optimal system desired state based on the desired state of the entire network supporting the network services, optimization targets and the present network state. Whenever a change of the desired system state is needed, the tools ensure a syntactically and semantically correct representation of the new desired state and store it into the runtime repository.

The architecture also establishes a clear demarcation line between vendor deliverables and the functional and realization domains for operations. The main role of the SW supplier is to establish a constant stream of updated and validated combinations of artifacts into the CSP artifact repositories, enabling the CSP to access the latest NF SW at any time. Delivery pipelines and onboarding services ensure the authenticity and integrity of the onboarded artifacts and allow the service delivery organization(s) to supply complementary artifacts.

The onboarding pipeline is triggered by the availability of new or changed artifacts in the artifact repository and facilitates their processing in the management, orchestration and assurance tools, or alternatively, makes them available in the runtime repository directly.

Determining the desired declarative state with a management and orchestration solution

It is the responsibility of the management and orchestration system to manage the deployed

services and resources over their entire life cycle, ensuring that they meet the necessary service requirements efficiently. As shown in Figure 1, in a declarative management model the functional LCM is separated from the realization LCM. In this scenario, the management and orchestration system is responsible for the functional LCM and provides the realization management with the desired state of the deployed system.

The management and orchestration tools gradually resolve a high-level service model or service intent, which can be declaratively described, until a realizable desired state is reached. This desired state is then pushed into a runtime repository, from which the deployment operator reconciles the state and starts modifying the target systems to reach the desired state. The runtime repository reflects the demarcation between the management aspect of the functional domain and the realization domain.

As deployment operators act on specific target systems with limited context, the management and orchestration tools need to determine context information, such as placement or initial size and size limits of a CNF, and encode this information into the desired state.

The dynamic nature of telecom systems requires constant reevaluation of the desired system state against the key operational characteristics of the network. With the support of the assurance systems, the management and orchestration tools provide the control loop that evaluates and determines the desired state. To support this, the management and orchestration system places policies into the repositories and deployment operator to be notified in cases such as when:

1. The desired state and the actual state are the same
2. The desired state cannot be reached
3. New artifacts appear that may require evaluation.

Understanding the roles of repositories

There are two separate repositories in our proposed architecture: the artifact repository and the runtime repository.

Artifact repositories

An artifact repository is a harmonized landing point for vendor artifacts that continuously receives all the latest releases. Depending on the type of artifacts in question, different repository types are used, including Open Container Initiative-compliant registries (to store container images and Helm charts), Git repositories (to store text-based artifacts) and object stores (to store arbitrary larger binary files).

Vendor-specific delivery pipelines are required to adapt to the vendor's delivery process. A CSP may use different instances of the same artifact repository implementation for different vendors or purposes. Meanwhile, a vendor can make different artifacts versions available to CSPs, such as offering prerelease SW for early trials or testing purposes. Tagging features in the artifact repositories can be used to differentiate between the different versions.

In addition to artifacts released by a vendor's research and development CI/CD process, artifacts created as part of an individual customer project are typically used in a delivery process as well. The artifact repositories can also serve as a landing point for predefined configuration files and customer-specific adaptations, for example. To avoid the addition of complex manual processes, it is possible to automate customization and configuration generation efforts as part of the delivery process, represented by separate CI/CD pipelines.

Runtime repositories

A runtime repository contains the desired state for the systems in the network. The deployment operator continuously monitors the contents of the runtime repositories and triggers actions if the target system does not reflect the desired state. This is called a reconciliation process.

The desired state is encoded in a declarative format – that is, it describes how the systems should be deployed and configured. The desired state can apply to areas including:

1. Infrastructure resources, such as Kubernetes clusters and their configuration
2. Workload resources, such as pods belonging to a CNF workload
3. Configuration for an NF.

The deployment operators consuming these state representations do not have to be specific to the area but are specific to the system on which they actuate. For example, a deployment operator actuating on a Kubernetes cluster can consume information for any of the three areas, given that all the areas can be configured using the Kubernetes application programming interface (API).

The runtime repository is version-controlled to provide a change history for the entire system and to allow for auditing. The declarative nature of the approach makes it possible to simplify recovery use cases by restoring an earlier version and then letting the system reconcile to the state described in the earlier version.

Often, a single repository is used to provide a single source of truth for the entire desired state of the network from a single system. Depending on the operational needs, multiple runtime repositories are sometimes used to separate concerns between the areas. In this case, the collection of all runtime repositories represents the single source of truth.

Git is a prominent implementation for a runtime repository that provides version control and allows for a multiuser workflow. Today's Git solutions such as GitLab or GitHub provide sophisticated workflows on top of the basic Git operations, enabling review, testing and approval flows. Due to its popularity as a runtime repository, a large ecosystem of agents, controllers and other tools has developed around Git. Operations that use Git as a single source of truth in combination with agents and controllers are referred to as GitOps.

The role of the deployment operator

The reconciliation of the information stored in the runtime repository with the target system is a critical function in a declarative system. The deployment operator realizes this task by executing a control loop that compares the information in the runtime repository with the actual system state and by continuously working to modify the system state to match the single source of truth.

There are two reasons why a mismatch between the desired system state in the runtime repository and actual system state can occur:

1. The management, orchestration and/or assurance systems have changed the desired state.
2. A fault in the target system has changed the actual system state.

In either case, the deployment operator will take corrective actions to align the actual system state with the desired state. As the corrective actions are specific to the target system, deployment operators are often designated to a specific target system.

Flux [4] is an example of a deployment operator that operates on one or more Kubernetes clusters as a target system. It can use several types of sources but is mostly used in combination with a Git repository as a runtime repository. Flux can actuate on any object that can be described as a Kubernetes object.

Repository-based software life-cycle management automation flows

The declarative, repository-based management approach provides efficient automation for configuration and SW handling of a target system. In many deployments, the approach is enhanced by further automation capabilities to facilitate review,

Declarative deployment automation in Kubernetes

Kubernetes is a prime example of a management system that follows a declarative model. The Kubernetes API uses the Kubernetes Resource Model [6] to describe any resource accessible through the Kubernetes API server. The declarative resource descriptions are stored in the internal Kubernetes state database.

Kubernetes controllers monitor specific resource types in a continuous loop. Whenever the desired resource state in the database is changed, the controllers act by creating or terminating pods or by modifying network policies. Custom Resources Kubernetes makes it possible to introduce new declarative resource descriptions for arbitrary objects. Custom controllers, or Kubernetes operators, monitor the custom resources and take arbitrary actions based on the information stored in the desired resource state.

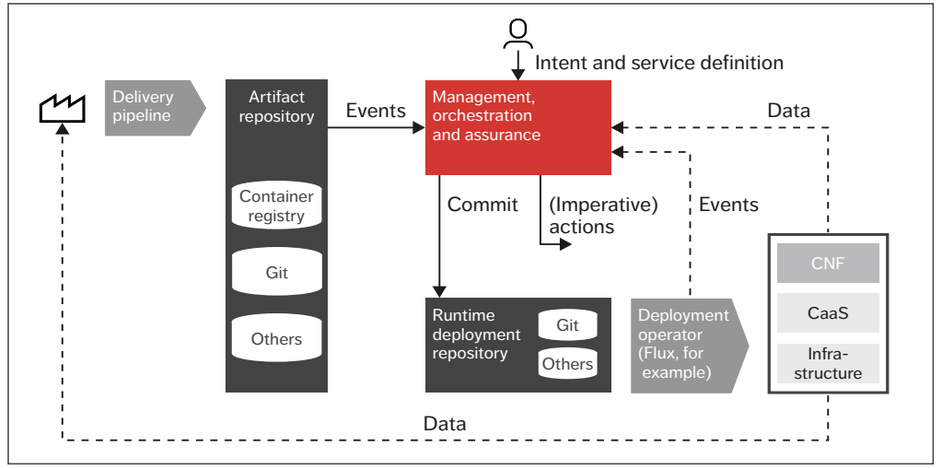


Figure 3 Declarative delivery and deployment automation flow

security scanning, approval processes, health checks and testing at different stages. *Figure 3* illustrates our concept for a declarative delivery and deployment automation flow.

Compared to the current imperative end-to-end (E2E) pipelines that describe long processes, a repository-based management approach enables smaller imperative actions (automation jobs) that can be triggered by different events in the orchestration and SW LCM flow. Some automation jobs will be generic to the target system, such as a security-scanning job that is triggered when new SW artifacts become available. Other automation jobs may be specific to a target system, such as a health-check job for a specific function. Smaller imperative automation jobs specific to a target system can be provided by the software supplier while others specific to the CSP's operational procedures will be developed by them.

Health checks are also used at different stages in the SW life-cycle process, such as checking system health before modifying the desired state and verifying system health after reconciliation. In the first case, the orchestration system issues the event that triggers the health check job prior to modifying the runtime repository. In the second case, the health-check triggering event is issued by the deployment operator that successfully completes the reconciliation.

Events can trigger different types of deployment automation jobs or chains of automation jobs (pipelines). Event-to-action resolution is required to address the CSP's operational needs with respect to flexibility and configurability. CDEvents [5], a common specification that provides a standard way to describe Continuous Delivery events, is a powerful enabler for flexible event systems.

Further input to the management and orchestration system comes from the network and OSS/BSS functions in the form of logs and performance data. This input can be used to further adjust the declarative description of the desired state to meet expectations of the functional domain. This data may potentially be shared with the SW vendor in an anonymized form as part of a DevOps feedback loop.

A common pattern in a repository-based management is to apply a multistage deployment or rollout process. Each stage relies on the same basic architecture with a single source of truth. If the same repository is used for several stages, replicating configurations between environments can become as simple as copying files from one folder to another. Typically, the promotion in between stages is automated and follows a dedicated approval process. Like other DevOps processes, every stage covers different aspects of the rollout, such as testing, for example. It is crucial that the further the process advances, the closer the environment is to the production environment in terms of hardware, surrounding applications, SW version and so on.

How to introduce declarative deployment automation

At present, imperative pipelines are often used to automate either all or parts of delivery and deployment processes. As a first step in transitioning to a declarative approach, we recommend the modularization of existing E2E pipelines into reusable automation jobs in order to gain the flexibility needed to introduce a new deployment mechanism. The second step we recommend is to replace the existing deployment stage with a declarative approach by introducing a GitOps solution. Modularization in combination with the GitOps-based deployment enables a loose coupling between onboarding, deployment and validation automation and allows for the reuse of automation jobs in the different pipeline fragments.

Using the modularization, the onboarding pipeline stages in E2E pipelines are split into two main fragments: the ingestion of artifacts into artifact repositories and the triggering of system-specific onboarding actions. While the former is often specific to the vendor deliverables and packaging formats, the latter is specific to the target management systems. The modularization of imperative E2E pipelines and the triggering of jobs based on events and explicit actions increase the reusability of pipeline fragments and simplify the adaptation to different sets of components of the management system.

> THE INTRODUCTION OF A DECLARATIVE RUNTIME REPOSITORY CREATES A CLEAR DEMARCATION BETWEEN THE DOMAINS

Conclusion

Continuous service improvements and innovation are essential to help communication service providers (CSPs) stay competitive in an evolving landscape. The ever-expanding Cloud Native Computing Foundation ecosystem characterized by cloud-native applications and evolving life-cycle management (LCM) practices/tooling demands a change in the management approach for network services and functions both in terms of tools and operational flows. CSPs must, therefore, move toward continuous deployment, where each updated microservice is validated in each staging step. A split of the management space into functional and realization domains makes it possible to separate the life cycles of the two domains, as well as their paces.

The introduction of a declarative runtime repository creates a clear demarcation between the domains. The separation enables CSPs to adopt new LCM practices, such as GitOps, and tap into the wide ecosystem of generic software life-cycle automation capabilities in the cloud-native realization domain. CSPs can independently build fit-for-purpose automation in line with their business processes and needs in the functional domain, and adopt a life cycle that matches the needs of their businesses, focusing on the key values and characteristics that their networks provide to customers.

The modular pipeline concept, with deployment automation jobs triggered by events from both domains, complements the declarative management approach and allows CSPs to continue to benefit from imperative pipelines matched to their process needs. This approach allows them to evolve their internal business processes and automation independently from the underlying software LCM.

THE AUTHORS

Peter Wörndle

◆ is a senior expert in deployment architectures whose work focuses on the use of cloud and infrastructure technologies in different types of management ecosystems. He joined Ericsson in 2007 while still at university. Since then, he has held several positions in R&D within the area of virtualization and cloud. Wörndle holds a M. Sc. (Dipl.-Ing.) in electrical engineering and information technology from RWTH Aachen University in Germany.

Stephen Terrill

◆ is a senior expert and chief architect in automation and management at Ericsson. Since joining the company in 1994, he has worked primarily in telecommunications architecture, implementation and industry engagement. In recent years, his work has focused on the automation and evolution of OSS. Terrill holds an M.Eng.Sc. from the University of Melbourne, Australia.

Torsten Dinsing

◆ joined Ericsson in 2000 and currently serves as a senior expert in service architecture at Group Function Technology and Strategy. In recent years, his work has focused on the interface between R&D and CSP organizations and the application of new practices such as DevOps, CI/CD and GitOps for LCM and orchestration. Dinsing holds an M.Sc. (Dipl.-Ing.) in electrical engineering from RWTH Aachen University.

Further reading

- **Ericsson, CI/CD:** *Continuous software for continuous change, available at: https://www.ericsson.com/en/ci-cd*
- **Ericsson blog, Your guide to CI/CD in telecom networks – for today and tomorrow, available at:** *https://www.ericsson.com/en/blog/2021/3/your-guide-to-cicd-in-telecom-networks--for-today-and-tomorrow*

References

1. **Ericsson, The cloud-native transformation, available at:** *https://www.ericsson.com/en/core-network/guide/forms/guide-cloud-native-transformation*
2. **Ericsson, CI/CD: Continuous software for continuous change, available at:** *https://www.ericsson.com/en/core-network/guide/forms/guide-ci-cd*
3. **12factor.net, The twelve-factor app, available at:** *https://12factor.net*
4. **Flux, Flux – the GitOps family of projects, available at:** *https://fluxcd.io*
5. **CDEvents, a common specification for Continuous Delivery events, available at:** *https://cdevents.dev/*
6. **GitHub, The Kubernetes Resource Model (KRM), February 20, 2018, Grant, B, available at:** *https://github.com/kubernetes/design-proposals-archive/blob/main/architecture/resource-management.md*

Approaching AI-native RANs through generalization and scalability of learning

"AI native" is a trending concept that is gaining momentum in many industries, and ours is no exception. To maximize the benefits of making radio-access networks more AI-native while simultaneously minimizing complexity and cost, we propose a stepwise approach based on generalization and scalability of learning.

PABLO SOLDATI,
EUHANNA GHADIMI,
BURAK DEMIREL,
YU WANG,
MATHIAS SINTORN,
RAIMUNDAS GAIGALAS

A holistic vision of an AI-native radio-access network (RAN) would be a system designed for artificial intelligence (AI) algorithms, in which a single AI algorithm could learn and govern most networking operations, ranging from the physical layer to Radio Resource Management (RRM).

■ As appealing as a holistic vision of an AI-native RAN might seem, making it a reality would almost certainly break the logical boundaries of traditional communication protocols. It is also likely to be hampered by telecom industry regulations and/or the constraints imposed by the standardized systems – those that are compliant with the 3GPP (3rd Generation Partnership Project) and the IEEE (Institute of Electrical and Electronics Engineers), for example – that are used today to ensure the interoperability of radio communications across national borders and system vendors.

Ericsson has chosen to take a less abrupt approach to creating more AI-native RAN systems by embracing design principles and concepts that enable the integration of AI into RAN systems with minimal disruption. While we agree that AI algorithms should be at the center of future RAN design, we believe the paradigm shift should occur gradually within the traditional boundaries of RAN systems (RAN protocol layers and hierarchy). We think it is also important to recognize that not all RAN design aspects need AI.

Ericsson initially deployed AI for network design and hyperparameters optimization of RAN algorithms a few years ago [1], using AI to fine-tune well-established RAN algorithms whose automation would otherwise be complex and costly. One example is optimizing the block error rate (BLER) target of downlink link adaptation (DLLA) in Ericsson 4G systems.

Academic and industrial research indicates that AI could be helpful for virtually every RAN design aspect [2]. However, most of this research considers specific RAN problems in isolation, with AI tailored to fit specific conditions or environments, such as those in the cell where training data is collected. We refer to such an approach as AI model specialization. While specialized AI models enjoy simple state reconstruction (few model inputs, for example) and less computational complexity, they do so at the cost of not having the robustness to cope with the diversity of RAN deployments and environments.

Our concern is that the adoption of solutions based on specialized AI models in live RANs could lead to an uncontrolled proliferation of AI models per radio feature, as shown on the left side of *Figure 1*, which would increase both RAN

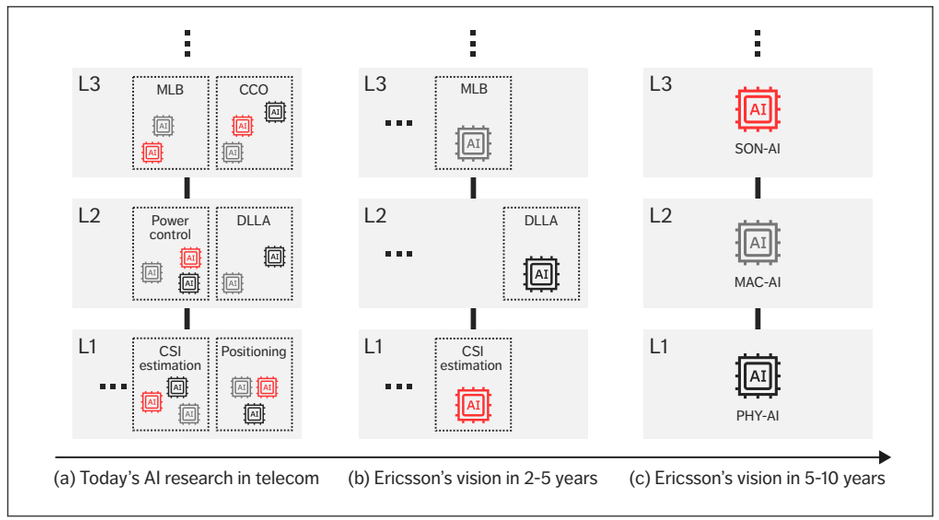


Figure 1 Ericsson's approach to integrating AI into a RAN over the next 10 years

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **AI** – Artificial Intelligence | **BLER** – Block Error Rate | **CCO** – Coverage and Capacity Optimization | **CDF** – Cumulative Distribution Function | **CRAN** – Cloud RAN | **CSI** – Channel State Information | **CSP** – Communication Service Provider | **CU** – Central Unit | **DLLA** – Downlink Link Adaptation | **DQN** – Deep Q-Networks | **DU** – Distributed Unit | **GNN** – Graph Neural Network | **HARQ** – Hybrid Automatic Repeat Request | **KPI** – Key Performance Indicator | **LCM** – Life-cycle Management | **MAC** – Medium Access Control | **MAC-AI** – Medium Access Control based on Artificial Intelligence | **MIMO** – Multiple-input, Multiple-output | **ML** – Machine Learning | **MLB** – Mobility Load Balance | **mMIMO** – Massive MIMO | **MSRBS** – Multi-standard Radio Base Station | **NF** – Network Function | **QoS** – Quality of Service | **PHY** – Physical Layer | **PHY-AI** – Physical Layer based on Artificial Intelligence | **RAN** – Radio-Access Network | **RL** – Reinforcement Learning | **RLE** – RAN Learning Engine | **RRM** – Radio Resource Management | **SEED RL** – Scalable and Efficient Deep RL | **SL** – Supervised Learning | **SON-AI** – Self-Organizing Networks based on Artificial Intelligence | **UE** – User Equipment

> WE HAVE IDENTIFIED TWO KEY ENABLERS: AI MODEL GENERALIZATION AND A SCALABLE AND VERSATILE LEARNING ARCHITECTURE

complexity and operational expenditure. Additionally, the coexistence of multiple AI-driven RAN features that are independently designed and expected to operate at the same timescale (that is, residing in the same protocol layer) and to learn from the same environment (the same cells, for example), would make the learning process difficult and potentially unstable.

The middle section of Figure 1 shows Ericsson's vision for integrating AI in RAN systems in the next 2-5 years. We are currently working to replace legacy RAN algorithms with AI counterparts that are designed to generalize across different radio conditions and environments, network deployments, types of traffic, RAN intents and so on. The intention is for these new AI algorithms to be reusable across the network. In this phase, we are focusing on few key RAN operations that are best suited to capitalize on AI – that is, those with the highest potential to boost performance.

In the longer term, which is illustrated on the right side of Figure 1, we aim to broaden the scope and reach of an individual AI algorithm to jointly solve multiple RAN operations. Initially, this could lead to the creation of a single generalized AI algorithm per RAN protocol layer that is designed to jointly learn and govern the most relevant (but not necessarily all) operations of its protocol layer.

Reliance on the traditional RAN protocol hierarchy ensures the correct separation of AI algorithms based on the timescale of RAN operations. Identifying which RAN operations within a protocol layer are the most relevant to be jointly addressed with AI is instrumental to avoid the risk of increasing the complexity of RAN products for diminishing returns. For instance, a generalized AI algorithm for the medium access control (MAC) layer may be designed to jointly control only the few most critical MAC layer operations, such as scheduling and link adaptation, but not necessarily power control, beamforming, precoding and so on.

Looking further ahead, the integration of AI technology in RAN systems may become closer to the holistic vision of an AI-native RAN where a single AI agent may learn to control several (if not most) RAN operations, possibly bridging across the traditional protocol layers. The feasibility and the blueprint of this vision is uncertain as it relies on the evolution of several technological areas, ranging from AI algorithms to computing, storage and RAN technologies. Furthermore, the telecommunication industry may also need to embrace more advanced and disruptive design concepts, that may lead to revisiting some traditional design pillars that have shaped our industry for more than three decades.

In our efforts to conceptualize a more AI-native RAN in the nearer term, we have identified two key enablers: AI model generalization and a scalable and versatile learning architecture.

Enabler 1: AI model generalization

A fundamental goal of machine learning (ML) is to achieve model generalization – that is, to train a function that can cope with conditions not directly observed during training yet inferable from training data. For AI in RAN systems, we extend this concept into three generalization domains:

1. RAN environment
2. RAN intents
3. RAN control tasks.

While each generalization domain brings us one step closer to a more AI-native RAN, the untapped potential of AI resides in a design that integrates all generalization dimensions, paving the way toward the vision of a single AI algorithm governing and replacing multiple RAN operations at once.

Generalization over the RAN environment

Training AI models to generalize over the RAN environment, rather than specializing to fit specific conditions (such as a cell environment), provides the necessary robustness to deploy a single AI model for a given task across the entire network. This improves scalability and reduces the complexity of the AI functionalities and operations. For example, model life-cycle management (LCM) becomes less complex when there is only one AI model per radio feature or RAN protocol layer (which is ideal).

AI generalization unlocks the potential for a single AI model to learn from the distributed experience of thousands of network entities. Distributing training-data generation over space (with thousands of network entities each contributing a few training samples) and time drastically reduces the impact of the learning process on RAN performance, such as for reinforcement learning (RL) algorithms taking suboptimal actions during training in live networks.

The RAN environment can be characterized by a combination of static and semi-static information that describes the network deployment and configuration, along with more dynamic information about the radio environment, traffic, load, specific user equipment (UE) conditions and so on. At Ericsson, we have investigated two main design approaches for generalizing AI models over the RAN environment. The first relies on traditional feature engineering, where domain knowledge is applied to identify the most critical information to reconstruct the RAN environment for specific AI applications. Techniques such as feature sensitivity analysis make it possible to filter out redundant information (based on correlation metrics, for example) to reduce model complexity.

A more systematic approach is to separately learn an embedding of static and semi-static RAN environment characteristics and combine them with dynamic RAN environment information that is specific to individual use cases. Graph neural networks (GNNs) [3] are a prime candidate for embedding the static and semi-static aspects of the RAN environment. These include network configurations and relations between different RAN entities (such relations between one cell and a set of direct or indirect neighboring cells), which may characterize wider relations between network entities beyond the reach of traditional feature-engineering approaches. Such embedding could be learned and reused across various use cases (from L1 to L3, for example). At Ericsson, this technique has been investigated in the context of L3 functionalities, such as secondary carrier prediction and antenna tilt control for coverage and capacity optimization (CCO).

Generalization over RAN intents

RAN performance cannot be defined by a single key performance indicator (KPI) such as throughput, spectral efficiency, latency, packet loss, reliability or resource utilization. Instead, it is defined using multiple, sometimes competing, KPIs. Therefore, many control problems in RAN, such as RRM, do not offer a single optimum-for-RAN KPI but rather a Pareto frontier of RAN KPIs, each trading off against another.

Ericsson's approach to generalizing the AI model for RAN intents includes the use of multi-objective RL, which enables a single AI model to learn the Pareto frontier of multiple RAN KPIs. In execution, the same AI model could be used to achieve the best KPIs to meet specific requirements of use cases, such as fulfilling different quality-of-service (QoS) requirements (including 5G QoS Identifier values) for different UE traffic types.

Generalization over RAN control tasks

Generalization over RAN/RRM control involves designing an AI algorithm to jointly control multiple transmission parameters by solving multiple RAN/RRM tasks simultaneously. This is instrumental in reducing the number of AI-based control loops operating at the same timescale and ultimately realizing our vision of a single AI algorithm per RAN protocol layer.

At the MAC layer (L2), for instance, a single AI algorithm, rather than two, could jointly solve resource scheduling and link adaptation. However, solving broader multi-decision-making problems

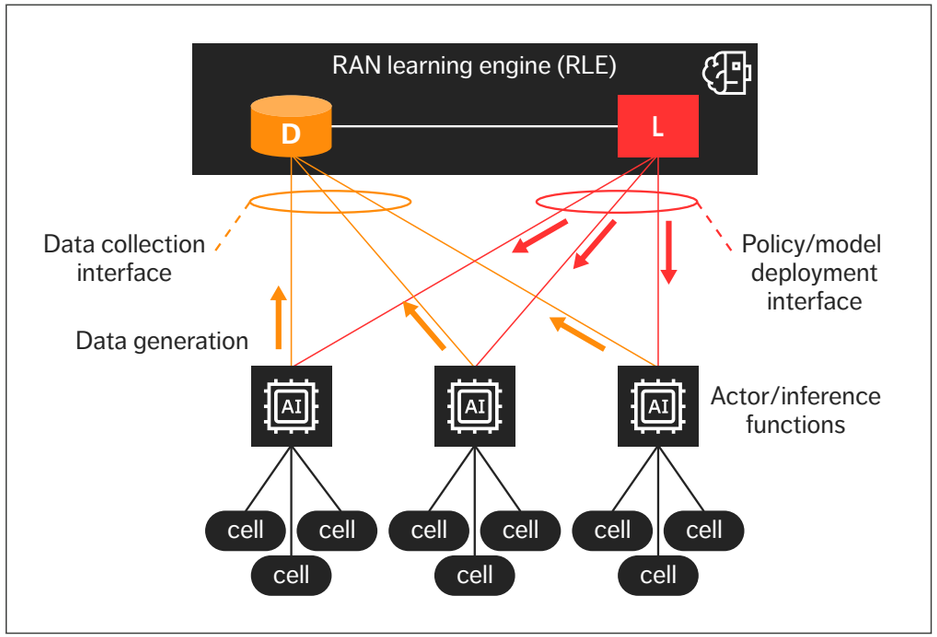


Figure 2 Our vision for the generic functional framework of a RAN learning architecture

with a single AI algorithm is only viable if it leads to a substantial performance boost, as it requires solving a more complex problem within stringent time limitations for the execution of AI models and entails more advanced hardware capabilities.

Enabler 2: A scalable, versatile learning architecture

The second enabler that we consider to be essential to realizing a more AI-native RAN is an scalable, versatile learning architecture that is, possibly, agnostic to the specific type of AI algorithms to be trained, such as RL and supervised learning (SL). Designing AI algorithms to generalize in any of the previously mentioned domains poses new challenges and opportunities. In addition to designing an appropriate AI algorithm to solve the task at hand, a major challenge for AI generalization in RAN is finding a scalable solution for training the AI model, as well as for the generation, collection and management of the necessary training data.

Since the publication of the seminal DQN algorithm [4], several architectural works have gradually addressed the scalability and efficacy of the learning for RL algorithms [5-10]. The Impala architecture [7] achieved the first milestone by exploiting off-policy learning combined with a decoupled distributed actor-learner architecture to capitalize on, rather than suffer from, the lag between when the actors generate actions and when the learner estimates the gradient. A distributed RL architecture advocating a similar actor-learner decoupling for the RAN applications was independently proposed [11]. Such principles have recently been further explored by other RL architectures, including APEX [8], R2D2 [9] and SEED RL [10], which pushed the boundaries of learning scalability and efficacy to new levels.

Despite being referred to as distributed RL architectures, these methods exploit the actor-learner decoupling through a single centralized learning function supporting several distributed actor functions, all of which contribute to learning a single RL policy. (Note that because an RL policy is typically represented by a functional approximator, such as an AI model, the terms RL policy and AI model are often used interchangeably when referring to RL algorithms.)

Actors asynchronously receive a policy update and evaluate the policy, possibly on several parallel environments (in parallel simulations, for example). This massively scales policy evaluation and training data generation, enabling more frequent policy updates, which leads to faster and more efficient learning. For instance, the R2D2 architecture [10] achieved a fivefold to twentyfold performance improvement over traditional DQN with a corresponding two-and-a-half fold to fiftyfold training time reduction, respectively.

Our solution to address the challenges of AI generalization is a RAN learning architecture inspired by the most advanced RL architectures, as illustrated in *Figure 2*. There are two main elements:

1. A centralized RAN learning engine (RLE) function
2. Distributed actor functions (for control purposes, for example) and/or inference functions (for insight generation, for example).

Despite its resemblance to some RL architectures [8], the RAN learning architecture in Figure 2 offers a versatile functional framework with a broader AI algorithmic scope, ranging from RL algorithms for RAN control to SL to provide RAN insights. It also supports unsupervised learning, offline RL and/or transfer learning by relaxing the typical conditions of the RL learning loop. Even federated learning could be supported with some functional modification based on the Gorila architecture [5], for example.

The RLE is a centralized support function that provides AI learning services and data services to different AI-driven RAN features. These "AI features" are executed by distributed actor/inference functions, possibly at different RAN protocol levels and timescales, by means of two interfaces for model/policy deployment and data collection, respectively. A natural implementation of the RLE function is in a cloud environment, for both multi-standard radio base station (MSRBS) systems and cloud RAN (CRAN) systems hosting actor/inference functions.

The deployment of actor/inference functions executing an AI feature in the RAN architecture is use-case specific – for instance, internal to RAN network functions (NFs) like a gNB central unit (CU) for mobility use cases or a distributed unit (DU) for PHY and MAC use cases. On the other hand, the centralization and unification of the AI support functionalities within the RLE makes its deployment in the RAN architecture use-case agnostic, thus allowing the flexibility needed for a lean integration of AI technology in RAN systems, with reduced complexity and operational expenditure.

The RLE function is responsible for the entire learning and data management pipeline. For ease of description, we decouple its functionalities into two main categories: a centralized learning engine (L) and a centralized data engine (D). The centralized learning engine provides AI learning services (including training, validation and testing, as well as AI model LCM management services) for actuation/inference functions deployed in RAN nodes. The learning engine may also control the generation of training data (including exploration) from the underlying distributed actor/inference functions. The centralized data engine provides data services related to the storage and management of training, monitoring and other types of data; the storage and version control of algorithmic objects (including trained models and learning data structure); the storage and version control of unit tests; feature store functionality and so on.

Actor and inference functions are broadly responsible for AI model execution, the generation and collection of training data, the execution of unit tests and some AI model LCM support

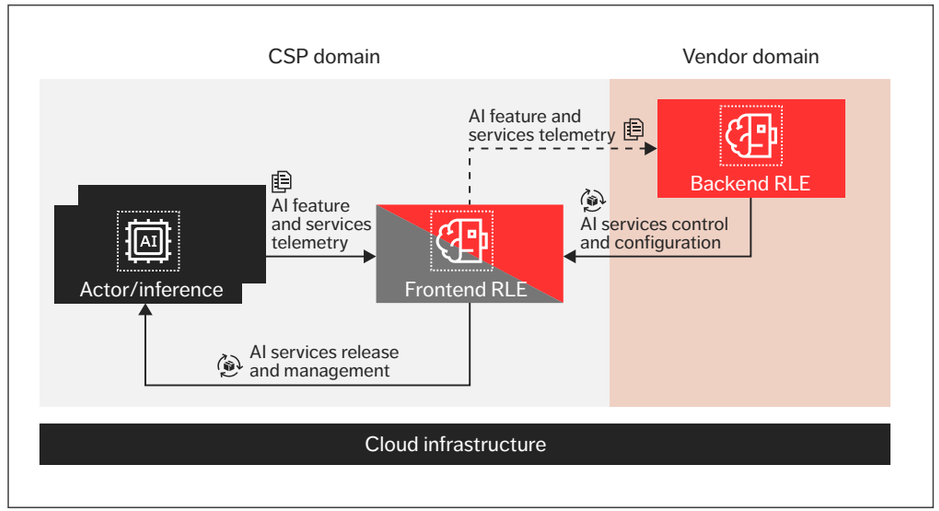


Figure 3 Example of RAN architecture deployment in a cloud infrastructure shared between vendor and CSP

functionalities, such as local performance monitoring, model inaccuracy detection and fallback operations. However, certain AI algorithms will require specific functionalities. An RL actor, for instance, would handle action selection (including exploration) and actuation, reward computation and so on. An SL inference function would need to provide automated data labeling.

Before an AI algorithm is deployed in a communication service provider's (CSP's) network, it is first designed and fully tested in a digital representation of the real network that is known as a network digital twin. The digital twin not only reproduces the deployment and configuration of the network and the radio propagation, traffic mix and user mobility of the environment condition, but also implements the same RAN learning architecture as shown in Figure 2. Following the same architecture enables both the training of performant AI models and the portability of the developed AI algorithm from the digital twin to the CSP's network with minimum overhead, thereby reducing the time to market.

Providing AI services in the RAN

The introduction of a RAN learning architecture that relies on a centralized RLE function makes it possible to efficiently provide the support services that are necessary to integrate AI algorithms in RAN systems.

We broadly distinguish AI services to reflect different phases of the lifecycle of an AI feature. Service type 0 refers to the initial release of an AI feature, which may be delivered in a RAN system as a software package upgrade. An incremental release in RAN nodes can allow safe integration of the new AI feature, which at this stage may be trained either in simulations or with data available to the vendor from proprietary testbeds or field trials, for example. In some cases, a CSP that is introducing an AI feature might want to train it with data from the CSP's own network. This would be supported by a process referred to as service type I. Finally, we refer to the continuous LCM of the AI feature while operational in a CSP network as service type II, which includes lifecycle managing the AI model, along with more advanced functionalities as algorithmic or design upgrades.

This simple and modular structure of AI services enables flexible customization on a per-CSP and per-AI-feature basis. A CSP could initially decide to introduce a new AI feature requiring only a standard service type 0. However, the service agreement may later be modified to include continuous AI services of type I or type II, with either full or partial service functionalities.

While there are numerous ways to deploy the learning architecture in RAN systems, only a few of them meet our criteria for a more AI-native RAN and enable the system vendor to retain responsibility for providing AI services during all phases of an AI feature lifecycle. The most efficient way to provide the continuous and flexible AI services of type I (feature training) and type II (feature LCM) is to maintain the learning functionalities as close as possible to the training data itself. To that end, the deployment of the RAN architecture must support two scenarios:

1. The provision of AI services from within the CSP network – that is, when the CSP data cannot be transferred to the vendor domain
2. The provision of AI services directly from the vendor domain – that is, when the CSP data is made available in the vendor domain.

We achieve such flexibility in design by deploying the two RLE function instantiations illustrated in *Figure 3*: a backend RLE in the vendor domain and a frontend RLE in the CSP network. In the example in Figure 3, the vendor and the CSP share the same cloud infrastructure, although this is not a prerequisite.

The two RLE instantiations are functionally almost identical, but with some crucial differences. The backend RLE is lifecycle managed by the vendor and retains the responsibility to control and configure the AI services for all AI features used in a CSP network. The frontend RLE is deployed and lifecycle managed by the CSP and it is responsible to release and administer the AI services in the CSP's own network, as it has direct access to the network functions hosting the actor/inference functions used by specific AI features.

The execution of type I (feature training) and type II (feature LCM) AI services is handled by the RLE function that has direct access to the CSP data. Therefore, when the CSP data is made available to the vendor domain (the light red area in Figure 3), the backend RLE also executes model training and LCM services within the vendor domain. Further, it instructs the frontend RLE to administer the services in the CSP network (AI model deployment and so on).

If the CSP data is not available in the vendor domain, the frontend RLE executes and administers the AI services within the CSP domain (the gray area in Figure 3) under the supervision and control of the backend RLE. Depending on which RLE instantiation provides continuous AI services of type I or type II, the "AI feature and service telemetry information" transferred from the frontend RLE to the backend RLE may include monitoring information of training and LCM services, AI models trained in the CSP domain, or training data generated in the CSP network.

Concept validation

In the pursuit of AI generalization for RAN features, we implemented the APEX architecture [8] for off-policy RL with actor-learner decoupling and considered DLLA as a use case to develop, validate and evaluate our design principles. We designed an AI algorithm that directly controls the modulation and coding scheme index for individual user transmissions to replace the legacy DLLA procedure. The introduction of a highly scalable learning architecture that capitalizes on distributed policy evaluations from multiple actors, each handling multiple parallel simulations, proved instrumental in boosting learning efficiency. Without altering the algorithm design itself, moving from a single actor to a 15-actor setup led to a twentyfold reduction in the training time and improved performance by 20 percent.

Most importantly, our approach made it possible to generalize our AI-driven DLLA solution over the RAN environment, intents and control parameters, which are more data-hungry to train than a single

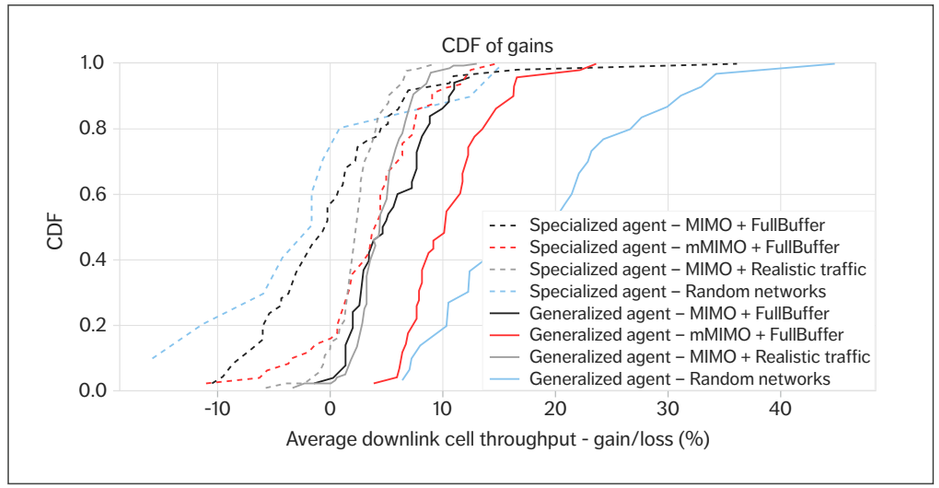


Figure 4 Cumulative distribution function (CDF) of average cell downlink throughput gain/loss over legacy DLLA evaluated for several realizations of the different benchmark scenarios

specialized AI model (but not necessarily compared to training thousands of specialized AI models).

To generalize DLLA over the RAN environment, we must capture characteristics of the RAN deployment surrounding UEs (including static and semi-static deployment and configuration information) and dynamic information characterizing the radio environment and UE state (including channel state information, HARQ (hybrid automatic repeat request) process and UE buffer).

Figure 4 shows that the generalized AI algorithm outperforms the legacy DLLA algorithm (which has a 10 percent BLER target) in relevant benchmark scenarios, comprising conditions, parameters and deployments unseen during training. Average cell throughput gains of well over 20 percent are observed in most benchmark scenarios, with smaller gains observed in scenarios where the legacy DLLA is expected to perform well. Figure 4 also shows that training an AI model to specialize in certain conditions leads to a less robust design that cannot cope with environmental changes.

Conclusion

Our research has identified two key enablers for efficiently integrating and systemizing artificial intelligence (AI) in future radio-access network (RAN) systems. Firstly, AI algorithms must be designed with the goal of generalization across the entire RAN environment, including intents and control tasks. Secondly, RAN systems must be empowered with an advanced and scalable learning architecture that can support a variety of AI algorithms and enable efficient learning at scale from thousands of distributed network entities. This architecture can be achieved with a centralized RAN learning engine that is able to support thousands of distributed actor/inference functions. While the deployment of actor/inference functions in RAN systems is use-case dependent, our analysis indicates that only certain deployments of the RAN learning engine lead to a lean integration of AI in RAN systems, resulting in the best AI performance and services.

Further reading

- **Ericsson Technology Review, AI-enabled RAN automation, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/ai-enabled-ran-automation*
- **Ericsson blog, Don't take AI and ML for granted - understand why they're critical for your RAN automation, available at:** *https://www.ericsson.com/en/blog/2022/4/dont-take-ai-and-ml-for-granted---understand-why-theyre-critical-for-your-ran-automation*
- **Ericsson, AI-powered radio-access networks, available at***: https://www.ericsson.com/en/ai/ran*
- **Ericsson, 5G RAN, available at:** *https://www.ericsson.com/en/ran*

References

1. **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/4ac66f/assets/local/reports-papers/ericsson-technology-review/docs/2020/enhancing-ran-performance-with-ai.pdf*
2. **IEEE Communications Surveys & Tutorials, Vol. 21, issue 4, Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues, June 2019, Sun, Y, et al., available at:** *https://ieeexplore.ieee.org/document/8743390*
3. **IEEE Transactions on Neural Networks 20 (1), 61-80, The Graph Neural Network Model, January 2009, Scarselli, F, et al., available at:** *https://ieeexplore.ieee.org/document/4700287*
4. **NIPS, Playing Atari with Deep Reinforcement Learning, 2013, Mnih, V, et al., available at:** *https://arxiv.org/pdf/1312.5602v1.pdf*
5. **ICLM, Massively Parallel Methods for Deep Reinforcement Learning, January 2015, Nair, A, et al., available at:** *https://arxiv.org/abs/1507.04296*
6. **ICML, Asynchronous Methods for Deep Reinforcement Learning, January 2016, Mnih, V, et al., available at:** *https://arxiv.org/abs/1602.01783*
7. **MLR, IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures, 2018, Espeholt, L, et al., available at:** *https://arxiv.org/abs/1802.01561*
8. **ICLR, Distributed Prioritized Experience Replay, September 2018, Horgan, D, et al., available at:** *https://arxiv.org/abs/1803.00933*
9. **ICLR, Recurrent Experience Replay in Distributed Reinforcement Learning, September 2018, Kapturowski, S, et al., available at:** *https://www.deepmind.com/publications/recurrent-experience-replay-in-distributed-reinforcement-learning*
10. **ICLR, SEED RL: Scalable and Efficient Deep-RL with Accelerated Central Inference, 2020, Espeholt, L, et al., available at:** *https://arxiv.org/abs/1910.06591*
11. **IEEE Communications Magazine 56 (9), 138-145, Learning radio resource management in RANs: Framework, opportunities, and challenges, 2018, Calabrese, F. D. et al., available at:** *https://ieeexplore.ieee.org/document/8466370*

THE AUTHORS

Pablo Soldati
◆ joined Ericsson in 2018 as a standardization and concepts researcher for radio networks and AI. In his current role, he defines solutions and strategy for the integration of AI in radio access networks. He holds a Ph.D. in telecommunications from KTH Royal Institute of Technology in Stockholm, Sweden.

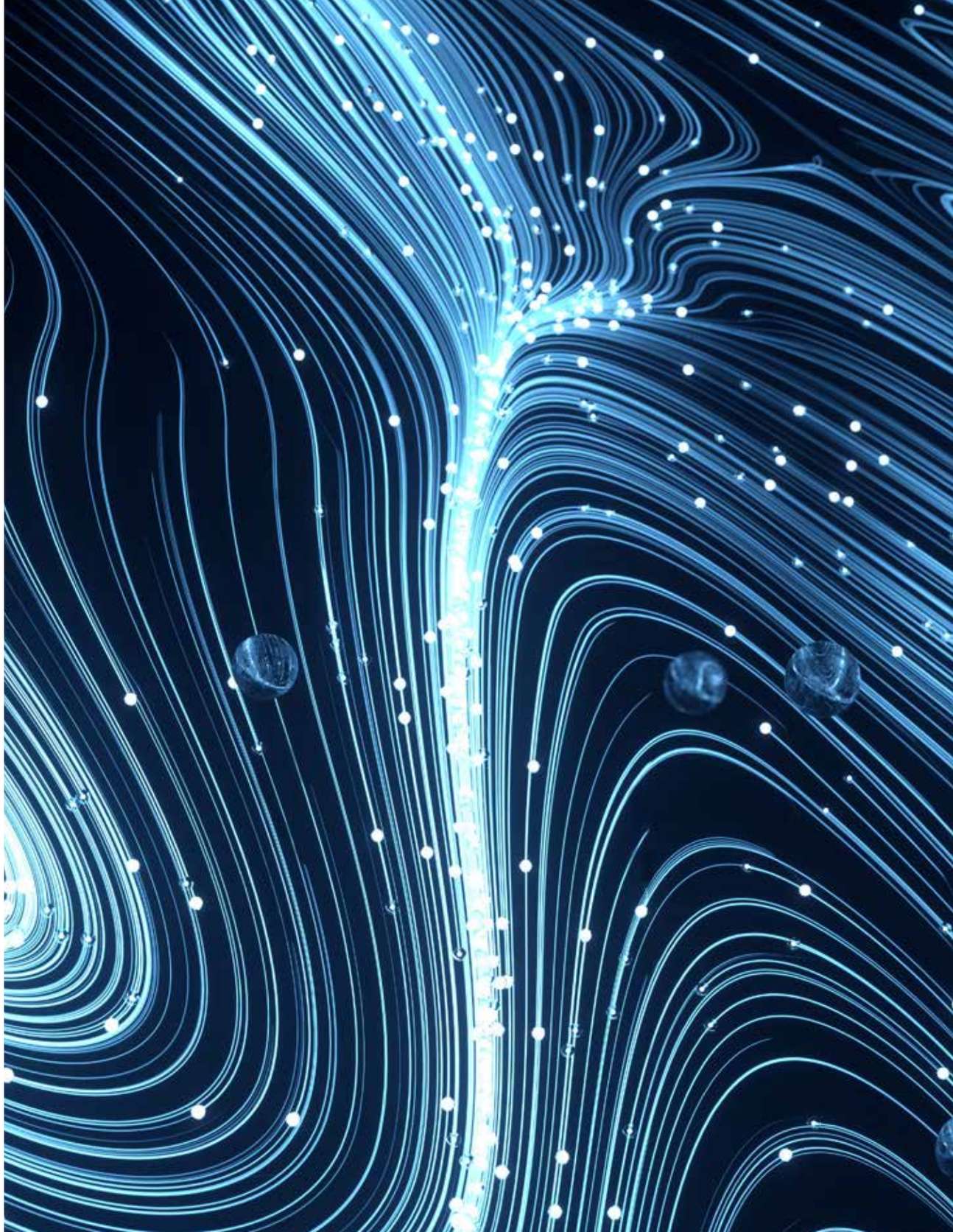
Euhanna Ghadimi
◆ is a RAN data scientist working with AI algorithms and systemization within Business Area Networks. His research interests include AI, optimization theory and wireless networks. Ghadimi joined Ericsson in 2018. He holds a Ph.D. in telecommunications from KTH Royal Institute of Technology.

Burak Demirel
◆ joined Ericsson in 2020 and is currently a senior researcher whose work focuses on AI, RL, control theory and cyber-physical systems. He holds a Ph.D. in automatic control from KTH Royal Institute of Technology.

Yu Wang
◆ joined Ericsson in 2010 to drive research activities in RRM, network management and telecom data analytics. He is currently a concept developer, focusing on the development of AI/ML-based radio network automation solutions. Wang holds a Ph.D. in communication engineering from Aalborg University, Denmark.

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, Sweden.

Raimundas Gaigalas
◆ joined Ericsson in 2005 and has been working with concept and simulator development in RRM Layer 2 and the systemization of commercial features in various RAN products. He currently serves as a developer in Cloud RAN with a focus on AI/ML functions in the distributed unit. Gaigalas holds a Ph.D. in mathematical statistics from Uppsala University.



ISSN 0014-0171
284 23-3400 | Uen

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