Ericsson Technology Review

Charting the future of innovation

Volume 112 | # 01 2024

The importance of ICT in society – Ericsson Technology Review and 100 years of innovation

The history of the mobile internet: the technology transformation that changed the lives of billions

Broad beamforming technology in 5G Massive MIMO

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

Editors
Tanis Bestland (Comprend)
Paul Coventry (Comprend)

Editorial Board
Hans Bergström, Elias Blomqvist, Peter Butovitsch, Magnus Ewerbring, Kjell Gustafsson, Ciaran Johnston, Sara Kullman, Johan Lundsjö, Håkan Olofsson, Patrik Roseen and Rohit Singhal

Art Director
Magnus Fredriksson (Comprend)

Project Manager
Susanna O'Grady (Comprend)

Layout
Magnus Fredriksson (Comprend)

Illustrations
Magnus Fredriksson (Comprend)

Subeditors
Ian Nicholson (Comprend)
Paul Eade (Comprend)

ISSN: 0014-0171
Volume: 112, 2024

Ericsson Technology Review – centennial issue

Celebrating a century of transformative innovations

As the publisher of Ericsson Technology Review, I am delighted to release this special centennial issue that celebrates 100 years of making technology insights accessible to the broader public.

In our first issue, published in 1924 as L.M. Ericsson Review, we presented the 500 line selector, an exciting technology innovation designed to enable mass telephony and help people connect with each other more efficiently. The exponential growth in mass telephony in the years and decades that followed became a key enabler of dramatic societal and economic transformation in countries all around the world. Articles from Indonesia, Colombia and Morocco published in early issues make it clear that Ericsson had a strong global presence and perspective right from the start.

Over the decades since, ETR has continued to provide insights about promising technologies and innovations that have led to further societal and economic transformation. In the 1950s, for example, we shared early mobile telephony research in which Ericsson's MTA (mobile telephony system A) was used to make live calls from cars in the field – an essential step on the journey toward the creation of the global mobile internet, the mass-market app economy and the 5G networks we enjoy today.

Ericsson has a long and proud history of invention and innovation in a wide range of areas, including everything from the creation of a single device used for both talking and listening to being a leader in the development of both electrical switches and automatic traffic lights to inventing Bluetooth in the 1990s – and that's just the tip of the iceberg. You can learn more about ETR history – and Ericsson's many innovations over the past 100 years – in two special anniversary articles on pages 4 and 7.

Today, as we look ahead to 2030 and beyond, ETR's primary focus is on exploring the technology solutions that will drive the evolution from 5G to 6G, enabling fully programmable networks that can deliver limitless, global connectivity. In this issue, for example, you can find our proposal for early alignment on 6G network architecture on page 19.

The exponential traffic growth expected with 6G will create a need for highly energy-efficient, high-interconnect bandwidth within the radio access network. This requires the development of a version of co-packaged optics technology that is suitable for radio applications – requirements you can read about on page 25.

Erik Ekudden
Senior Vice President,
Chief Technology Officer

When it comes to improving network efficiency, broad beamforming technology offers several advantages. On page 14 we explain how Ericsson's unique dual-polarized beamforming technique can be used to create broad Synchronization Signal Block beams with excellent power utilization.

Anyone with an interest in the use of 5G network digital twins for industry integration will appreciate the article about Asset Administration Shell on page 37.

As 5G mobile networks become increasingly complex, effective network management is becoming simultaneously more important and more difficult to achieve. The rApp approach to delivering automation, which you can read about on page 32, has great potential to help communication service providers overcome network management challenges.

In any network – 5G, 6G or otherwise – the importance of quality monitoring cannot be overstated. The article on page 44 explains how the exposure of advanced quality-monitoring capabilities through service application programming interfaces enhances the ability to exchange stakeholder information, making it easier to address specific quality-improvement needs.

At Ericsson, we know from experience that what seems impossible today could be the reality of tomorrow. ETR looks forward to serving as your trusted source of world-leading research and thought leadership in the years and decades ahead. We hope you enjoy this extra special issue of our magazine and that you will share our insights with your colleagues and business partners. You can find both PDF and HTML versions of all the articles at: www.ericsson.com/ericsson-technology-review

The importance of ICT in society — Ericsson Technology Review and 100 years of innovation

Authors:
Pernilla Jonsson, Hans Bergström, Peter von Butovitsch

For the past 100 years, Ericsson Technology Review has pioneered new technological frontiers through world-leading research. Now we turn the page on history to explore how these innovations shape modern society.

ISSN 0014-0171 284 23-3419 | Uen

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

Global GDP grew sixfold between 1950 and 1998 [1], largely due to improvements in transport, the emergence of mass production, and significant technological developments across consumer and enterprise markets, including ICT. This unprecedented growth led to an increase in commodity trade that more than tripled the ratio of global exports-to-GDP in the same period [2].

For more than 100 years, ICT has been an important driver of economic and social development, facilitating and enabling progress in countries all around the world. ICT was the first technology to give humankind the ability to transcend physical space and, in doing so, made it possible for us to create a platform that has spurred global economic growth and facilitated the spread of new ideas and technologies. Major innovations within telecommunications and mass telephony in particular over the course of the past century have had a societal and economic impact comparable to other major infrastructural breakthroughs in history, such as roads of the ancient Romans and the railroads of the great industrialists.

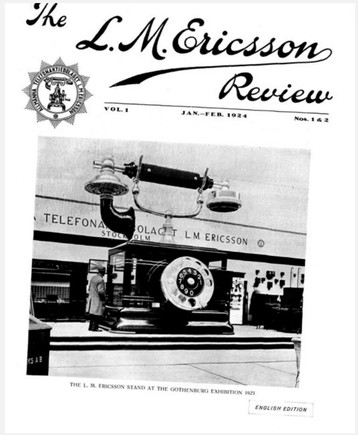
Ericsson Technology Review at 100: innovating a century of socioeconomic progress

Advances in ICT and its impact on modern society are greatly intertwined. As a general-purpose technology, the impact of ICT has not been limited to the sector in which it has been produced, but has spread across all sectors of production and consumption, significantly improving the quality and variety of many products and services that have gone to market.

For 100 years, Ericsson Technology Review (ETR) has been at the forefront of many of these advances. Our first edition [3] in 1924 laid the very foundation to demonstrate technology thought leadership, and that remains the same today.

This publication began at the inception of modern industrialized society, in 1924, when much of what we now take for granted had not yet come into existence. In the same decade, we started using electricity on a large scale to light up our homes, sparking a demand for household appliances. The acceleration of the motorcar industry also began to galvanize the world's strongest economies, driving the build-out of national road networks. The world was innovating and Ericsson too in vastly different areas: from railways [4] to time clocks, and betting systems [5] to radio gramophones [6] and traffic light automation systems [7].

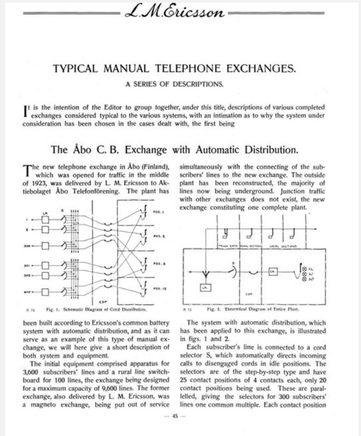
However, when the first ETR was published, under the name LM Ericsson Review, our focus had firmly changed to the build-out of global mass telephony and later ICT in general. Subsequently, one highlight was a feature on the automatic switchboard (with 500 switches), considered a major achievement in bringing telephony to the masses.


The L.M. Ericsson Review

VOL. I JAN.–FEB. 1924 NOS. 1 & 2

THE L. M. ERICSSON STAND AT THE GOTHENBURG EXHIBITION 1923

ENGLISH EDITION


L.M.Ericsson

TYPICAL MANUAL TELEPHONE EXCHANGES.

A SERIES OF DESCRIPTIONS.

It is the intention of the Editor to group together, under this title, descriptions of various completed exchanges considered typical to the various systems, with an intinsation as to why the system under consideration has been chosen in the cases dealt with, the first being

The Åbo C. B. Exchange with Automatic Distribution.

The new telephone exchange in Åbo (Finland), which was opened for traffic in the middle of 1923, was delivered by L. M. Ericsson to Aktiebolaget Åbo Telefonföreening. The plant has been built according to Ericsson's common battery system with automatic distribution, and as it can serve as an example of this type of manual exchange, we will here give a short description of both system and equipment.

The initial equipment comprised apparatus for 3,000 subscribers' lines and a rural line switchboard for 100 lines, the exchange being designed for a maximum capacity of 9,000 lines. The former exchange, also delivered by L. M. Ericsson, was a magneto exchange, being put out of service simultaneously with the connecting of the subscribers' lines to the new exchange. The outside plant has been reconstructed, the majority of lines now being underground. Junction traffic with other exchanges does not exist, the new exchange constituting one complete plant.

Fig. 1. Schematic Diagram of Cord Distribution.

Fig. 2. Theoretical Diagram of Entire Plant.

The system with automatic distribution, which has been applied to this exchange, is illustrated in figs. 1 and 2.

Each subscriber's line is connected to a cord selector S, which automatically directs incoming calls to disengaged cords in idle positions. The selectors are of the step-by-step type and have 25 contact positions of 4 contacts each, only 20 contact positions being used. These are paralleled, giving the selectors for 300 subscribers' lines one common multiple. Each contact position

— 45 —

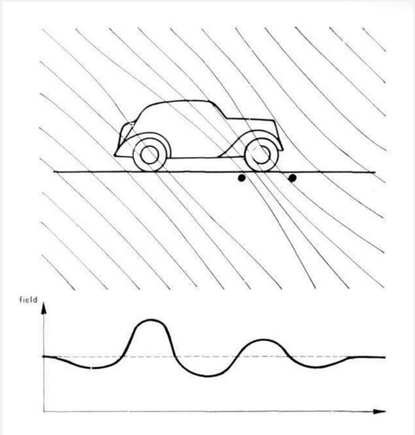


Figure 1: Since its very first issue, the L.M. Ericsson Review (today known as the Ericsson Technology Review) has sought to make technology accessible to a broader public, from in-depth descriptions of telephone exchanges (L.M. Ericsson Review #1 1924) to an early automated traffic signaling system (L.M. Ericsson Review #3 1946).

Driving impact with the digital economy

ETR was also at the forefront when Ericsson helped cut the cord on fixed telephony in the 1980s with 1G, through to the global mobile internet with 3G, and enabling the mass market app economy with 4G.

The smartphone and the enablement of wider societal digitalization have evoked fundamental, structural changes in daily life and business: in 2022, digitally delivered products accounted for a record 54 percent share of global services trade, with a growth rate that has outpaced all trade goods combined over the past two decades, including vehicles, fuels, pharmaceuticals, and more. [8]

Furthermore, the impact on billions of lives has been immeasurable. Mobile devices and app-based models have facilitated the re-invention of services, products and processes. This has made the smartphone our remote control to life, managing everything from email, shopping, music consumption and mobile banking, both in rural and urban areas. [9]

Today, with 5G, communication networks no longer just relay information between people as they did 100 years ago – but between things and machines, allowing entire factories, fleet systems and critical national infrastructure to become digitalized, smarter and more resilient.

As in the very first issue of ETR in 1924, automation remains a hot topic – but now we see a future in which automation and artificial intelligence can autonomously manage the world's networks – predicting and repairing faults, and dynamically distributing network resources in the most optimal way.

The next 100 years of Ericsson Technology Review

In 1924, we set our sights on a world of mass telephony to bring people ever closer together. Today, our engineers are envisioning smart programmable 6G networks [10], delivering limitless, global connectivity.

Similar to yesteryear's science fictive narratives of voice control, mobile devices, and AI, future networks will certainly enable new services and applications that today seem like science fiction.

This includes experiencing the internet with all five of our senses, witnessing connected intelligent machines talking to each other, and making the world much more efficient and sustainable for the progress of humanity. In the coming years, we also envision a blurring of the boundaries between physical and digital worlds, creating a cyber-physical continuum that can transcend time and even predict the future.

In the next 100 years, as we continue our journey through new technological frontiers, ETR will remain an eminent source of world-leading research and thought leadership that pushes the boundaries for innovation and drives critical advances across society. We are ready to take on the challenge and make what seems impossible today, the reality of tomorrow.

The authors

Pernilla Jonsson is the head of Consumer & Industry Lab at Ericsson Research. Pernilla has worked with innovation, market strategy and business development for multinational companies for over 20 years and is a keynote speaker in large global events. She has been part of the core team driving the Ericsson XR strategy towards 2030 and is driving research in XR and the metaverse in a variety of studies, design concepts and demos. Since joining Ericsson in 2015 she has been a key global spokesperson for the company and regularly frequents international media as an expert on the future of connected technology and what it means for people, business and a sustainable society. Pernilla is a board member of RISE – Research Institutes of Sweden, as well as WASP-HS - the Wallenberg AI, Autonomous Systems and Software Program - Humanities and Society research program and is also part of Digital Future's Societal Committee at KTH Royal Institute of Technology in Stockholm, Sweden.

Hans Bergström has close to 35 years of experience in communication services, with extensive global experience in different capacities from customer engagements to architectural and systems design. His background spans several segments including business communications, networks, data centers and wireless communication through 3G, 4G, 5G and 6G. In his current role of Director Architecture Evolution, he drives a special area of High-Performance Networks.

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

References

1. Ericsson research paper, Mobile broadband drives economic development ↗
2. OECD, The World Economy: Volume 1: A Millennial Perspective and Volume 2: Historical Statistics, The World Economy in the Second Half of the Twentieth Century, 2006 ↗
3. L.M. Ericsson Review 1924 ↗
4. L.M. Ericsson Review 1937 ↗
5. L.M. Ericsson Review 1931 ↗
6. L.M. Ericsson Review 1942 ↗
7. L.M. Ericsson Review #3 1946 ↗
8. IMF, IMF Blog, Why Digital Trade Should Remain Open, 2023, Ruta M., Jakubik A. ↗
9. United Nations, UN Chronicle, Mobile Communication and Socio-Economic Development: A Latin American Perspective, 2011, Fernández-Ardèvol, M. ↗
10. "6G network architecture – a proposal for early alignment," Cagenius, T., Mildh, G., Rune G., Vikberg, J., Wahlqvist, M., Willars, P., Ericsson Technology Review, 2023 ↗

Further reading

- Ericsson's commitment to digital inclusion ↗
- Ericsson's commitment to access to education ↗
- Ericsson blog, The impact of technology on education, inclusion and work ↗

Acknowledgements

We would like to acknowledge the contributions made by generations of skilled, innovative and insightful Ericsson colleagues within all domains required to facilitate this monumental paradigm shift.

The history of mobile internet: the technology transformation that changed the lives of billions

Authors:
Joakim Bergström, Peter von Butovitsch, Björn Ekelund, Kjell Gustafsson, Johan Lundsjö

Mobile broadband is the most popular means of internet access for billions of people worldwide. Its breakthrough is the result of several innovations: from the early development of mobile networks in the 1980s, the rapid uptake of internet in the 1990s, and the advances of devices in the 2000s. The move to mobile data and a new app economy secured the unprecedented rise of MBB services throughout the 2010s.

ISSN 0014-0171 284 23-3420 | Uen

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

Just over two decades ago, the first mobile broadband (MBB) services landed across consumer markets. For the first time, this made it possible to connect instantly to the internet from a mobile device almost anywhere in the world.

With this breakthrough, everything became possible. The blueprint for mobile communication was redrawn; the mobile phone was reinvented as a platform beyond voice and messaging, and the rules for communication, streaming, shopping, gaming, commuting and more were all rewritten.

Over the years, Ericsson Technology Review (ETR) [1] [2] has played a central role as a forum for sharing and spreading ideas and visions on many vital parts in this mobile miracle. Many innovations and trends that opened a pathway to MBB can be traced in ETR articles. While the forecasted evolution sometimes differed in detail, the overall direction of travel is clearly mapped.

In this article, we explore the technology story behind mobile internet and uncover how and why it came to be so deeply embedded across our societies and businesses.

1970s-2010s: paving the way to a mobile mass market

The advent of MBB is an innovation journey that stretches over many decades; from early mobile networks in the 1980s and the rapid uptake of the internet in the 1990s, to crucial device breakthroughs in the 2000s and the birth of the app economy in the 2010s.

1G: introduction of the first mobile systems

The foundation for MBB begins with the development of mobile telephony in the late 1970s and early 1980s. Advances in radio technology and the use of software-controlled switches enabled the initial development of mobile networks supporting voice and data services.

These developments were inherent in the first mobile communication systems, NMT (Nordic Mobile Telephony), AMPS (Advanced Mobile Phone System), and TACS (Total Access Communication System), which were based on the use of software switching and analog frequency division multiple access technology with limited capacity.

While the potential of mobile communication was not yet understood, it prompted a lot of learning around network and user devices [3]. Firstly, it quickly became evident that to encourage widespread uptake, user devices had to become more affordable, have good battery properties, and be compact and lightweight enough for mobile use. Secondly, the importance of good network coverage also became paramount, as users quickly expected to use the service everywhere.

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **CSP** – Communication Service Provider | **EDGE** – Enhanced Data Rates for GSM Evolution **GPRS** – General Packet Radio Service | **GSM** – Global System for Mobile Communications | **HSDPA** – High-Speed Downlink Packet Access **HSPA** – High-Speed Packet Access | **LTE** – Long Term Evolution | **MBB** – Mobile Broadband | **OTT** – Over-The-Top | **RAN** – Radio Access Network | **SMS** – Short Message Service | **TCP** – Transmission Control Protocol | **WCDMA** – Wideband Code Division Multiple Access

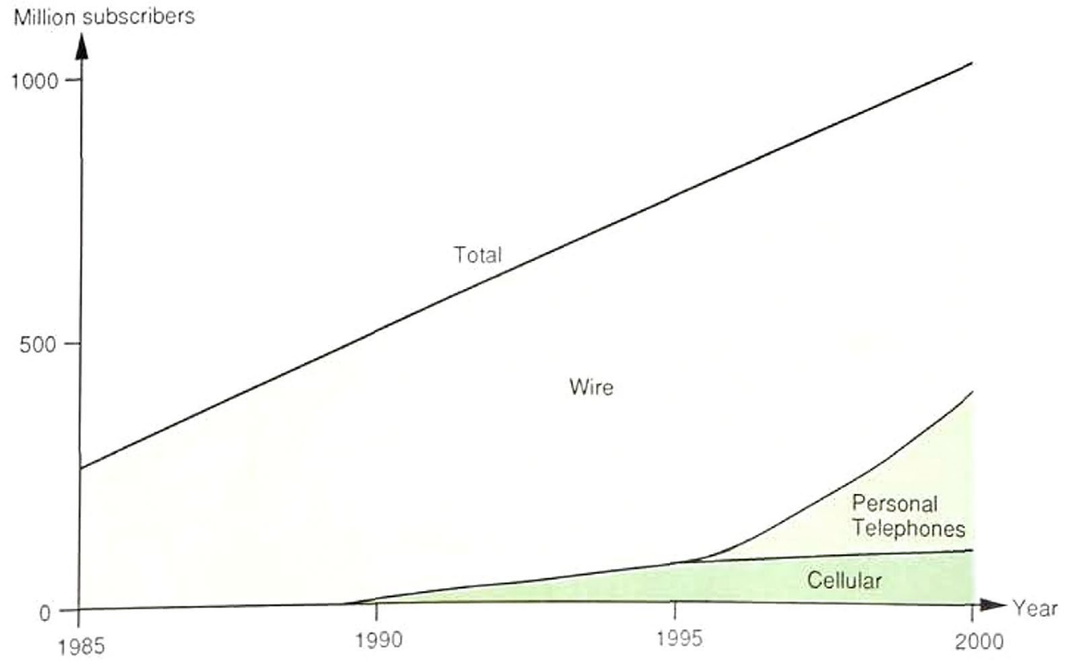


Figure 1: Prediction of subscriber growth in a 1990 ETR article [5]. The predicted success of mobile technology with 100 million mobile phone subscribers by 2000 was considered bullish. In retrospect, the growth was hugely underestimated and reflected the modest uptake of early mobile systems. The first one billion subscribers were reached in 2002, the second billion in 2005 and the growth rate increased even further after that [7].

While early mobile communication systems were well received on the market, high associated costs limited their appeal mainly to business users [4]. This resulted in projections for mobile markets being modest [5].

Even Ericsson underestimated the incredible growth-enabling force created by the availability of more advanced and affordable devices together with improved network capacity and coverage.

As shown in **Figure 1**, a 1990 ETR article [5] predicted 100 million mobile phone subscribers by the millennium; the actual number was more than seven times that amount [6] [7]. Improved device affordability and network performance drove a mass market that spurred even more investment in devices and networks.

Pioneering early pathways to mobile data

As mobile communication systems were being developed, the internet was emerging, together with a rapid uptake in the use of internet services accessed through fixed computers, as shown in **Figure 2**. The internet was built on the transition from circuit-switched to packet-switched data communication networks – a transition that spread across telecommunication domains and dissolved the borders between data and telecommunication. Leonard Kleinrock and Lawrence G Roberts received the L.M. Ericsson Prize in 1982 for their pioneering work in packet-switching theory and practical trials respectively, as elaborated on in their ETR articles [8] and [9].

Early mobile systems were seriously underestimated. By 2000, subscriber rates had outpaced early forecasts by more than 700 percent.

With the launch of Mobitex [11] in the same period came one of the first public mobile networks optimized for data and text communication, using digital packet-switching technology. It was a system separate from the mobile communication systems of the time and designed to replace the private mobile radio systems of taxi operators, blue light services, and haulage operators.

However, designs for application beyond this scope were already being considered, as the author of a 1989 ETR article [11] writes: "The introduction of public mobile data networks is expected to stimulate the development of new applications."

While Mobitex was without doubt a pioneering mobile data system and one of the first of its kind, it became redundant when efficient data communication capabilities were integrated into second generation (2G) systems.

2G: the arrival of SMS and limited data services

The launch of 2G systems like GSM (Global System for Mobile Communication), PDC (Personal Digital Cellular), D-AMPS (Digital Advanced Mobile Phone System) and cdmaOne in the early 1990s made data communication capabilities an integrated, albeit relatively limited part of mobile communication systems. Voice remained the primary service.

Mobile communications now began a very rapid growth period with technology and market maturing from earlier systems, first in Europe and other developed regions, and then spreading globally.

This growth period was propelled by mass production of network equipment and user devices, enabling strong economies of scale, which helped to gradually lower subscription and device prices for users. For the first time, this meant that users were not constrained by fixed telephony and could opt to go mobile.

While voice services remained dominant, the switch from analog to digital technology enabled support for SMS (Short Message Service) and 9.6 kbit/s circuit-switched data

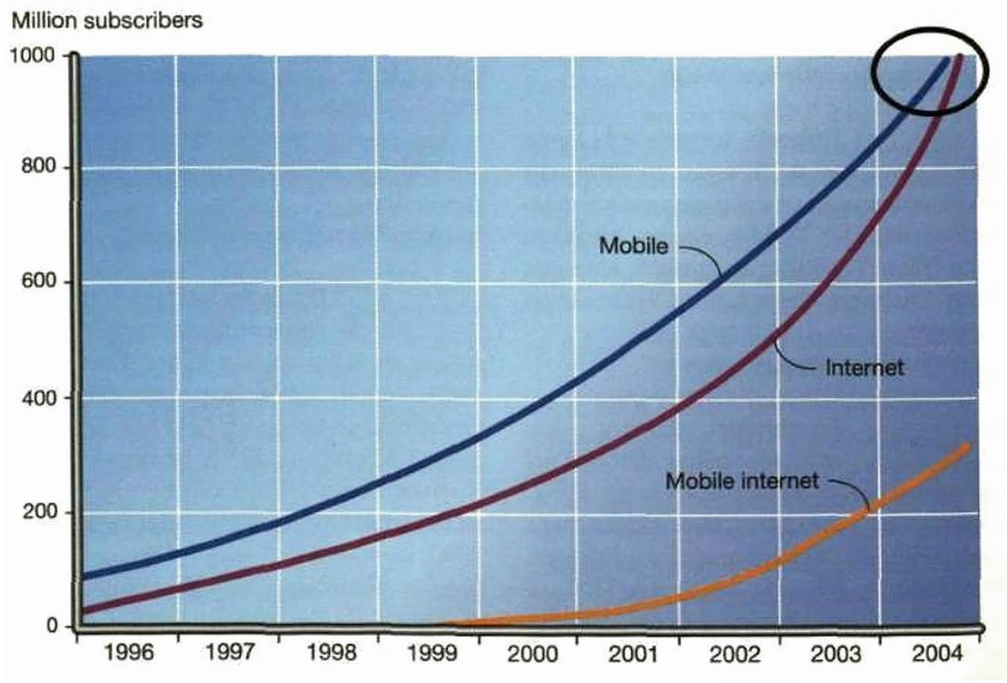


Figure 2: Forecast growth (subscribers) of fixed and mobile networks and the internet [10]. Even in the late 1990s, the number of mobile subscribers remained largely underestimated.

service. Although the data rate was limited, it was enough to enable a viable mobile internet connection for the first time on commercial mobile devices [12].

With GPRS (General Packet Radio Service) [13] came the introduction of packet-based services and network architecture in 2G systems. This evolution continued alongside the emergence of 3G, with the introduction of evolved GPRS and higher 2G data rates through EDGE (Enhanced Data Rates for GSM Evolution) [14].

3G: the concept of mobile internet is born

As 2G systems matured, the mobile industry realized a possible next step was to make the internet mobile. To prepare the groundwork for 3G, the European Commission and other administrative bodies began research projects to outline the technical and service criteria that 3G systems would require [15]. This drew the following conclusions: the peak rate of mobile networks must be comparable to fixed network connections; the mobile system must be capable of handling large amounts of data.

Both criteria were necessary to facilitate access to internet services available at that time and prepare the networks for a considerably richer service offering in the future.

The greatest need for development, and the industry's focus at that time, was to increase radio access network (RAN) capabilities. There were also important developments to the core network [13], user devices [16] [17] and business models.

In the 1990s, collective ambitions for a mobile internet prompted a step change in efforts to evolve mobile networks. Yet, these efforts were still largely guided and governed on a per-region basis, through many different and competing standardization bodies. The 3GPP (3rd Generation Partnership Project) was therefore formed to unify standards and facilitate a global, interoperable market with associated economies of scale benefits across both network and device domains. Through the 3GPP, most of Europe and Asia united around a commonly developed 3G standard [10] [18] [19].

The first release of the 3GPP's air interface standard, WCDMA (Wideband Code Division Multiple Access), enabled peak rates of 384kbps. This was, however, well below the 2Mbps target that the ITU (International Telecommunication Union) stipulated in the early 1990s, and efforts to improve this continued.

The first commercial 3G systems were launched in 2002 but the uptake of 3G users was initially slow. Peak rates of 384kbps improved on the speeds achieved by GSM and other 2G networks. The system was still based on circuit-switched technology inherited from the voice-based systems, making it difficult to reach high peak rates and high network capacity.

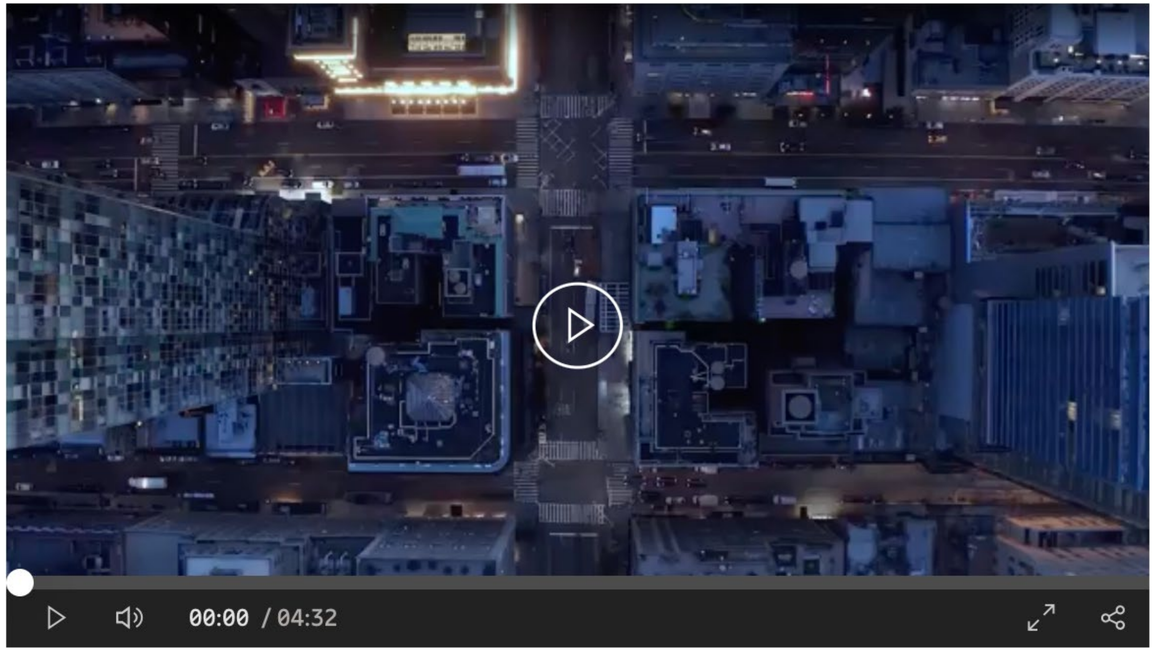
HSDPA (High-Speed Downlink Packet Access) [20] was introduced as the first step toward a true packet-based system, and it was soon able to increase peak rates and the network capacity by more than one order of magnitude, to several megabits per second in both the downlink and the uplink. It is worth noting that packet-based concepts had already been introduced in 2G systems in the late 1990s through GPRS, Enhanced GPRS and EDGE [13] [14].

In addition to higher peak rates, the performance characteristics of TCP/IP (Transmission Control Protocol/ Internet Protocol) and the importance of low latency in the network were also identified as key areas of focus [20]. Specifically, these related to the handling of TCP slow start and its impacts on the service performance of the consumer. WCDMA using HSPA (High-Speed Packet Access) was now technically good enough to enable a revolution on the service side [21].

Lower price indexes led to a mobile boom in the 1990s. For the first time, mobile entered the mainstream.

Breakthrough of feature phones and early smartphones

Until the early 2000s, voice and messaging still dominated mobile communications. The evolution of mobile data was advancing steadily, spurred by pre-smartphone feature phones that enabled email access and media sharing.



Video: Erik Dahlman, Senior Expert Radio Access, rolls back the years on mobile evolution from 1G to 6G ↗

However, mobile data applications, consumer devices and the mobile data business model had not reached a satisfactory stage of maturity for widespread adoption. There was consensus in the industry that communication service providers (CSPs) should own such services.

Regarding devices, other challenges were emerging. In the early 2000s, mobile phones made rapid advances, now supporting other media beyond voice, such as photos, videos and music. Media communication was also enabled through evolved messaging services like MMS (Multimedia Messaging Service). However, this posed a challenge: network data support was not yet sufficiently omnipresent and affordable to allow for heavy data usage, such as that required to stream video or music.

Until that time came, most media were handled locally on the phone and then uploaded/downloaded to a PC through a wired connection or Bluetooth. As network data support improved in later years, the barriers for mobile media streaming were lowered, yet applications for managing the media were often brand specific, which in turn created walled gardens that did not bring mass scale.

The app model gains a foothold in the market

In this period, there were several attempts to create a business model with CSPs at the center of service offerings [22] [23] [24] [25] [26] [27]. A very early initiative, already introduced in 2G, was the specification of WAP (wireless application protocol) [22] [23]. While these solutions were able to penetrate the market, their success was limited. By late 1999, however, mobile internet adoption was gaining momentum, and the early shift into a new paradigm was taking place [18]. Early commercial success of solutions like i-mode in Japan were gaining traction, and expectations on the future success of mobile internet could be envisioned. This sentiment was echoed in the Ericsson Review in the same year: "The pieces of the puzzle are falling into place, opening the way for the mobile internet. Being 'always connected, always online,' users will, simply by clicking or tapping, be able to manage their business and private affairs at any time and from any place." [18]

During the transition to mobile data, many ideas were proposed and developed but did not materialize. The toughest challenge was to find a business model that could facilitate large scale growth of services.

The industry struggled to agree on a concept to facilitate scale and usability, while simultaneously giving CSPs a strong stake in the value chain. The possibility to provide services over-the-top (OTT) was explored, but not initially preferred. There was a strong belief that much business value would come in "owning" the subscriber, and the common mindset was to help the CSPs capture that ownership.

However, pivotal ideals such as simplicity and technical viability soon shifted the momentum in favor of the OTT solution. It may not have been the obvious solution initially, but the creation of the app concept comprised several aspects that proved vital for the success of mobile data [28]. Through apps, a plethora of services could be made available to the user, all delivered over an MBB connection.

The introduction of advanced email services in Blackberry and eventually the smartphone (e.g. iPhone) marked a sea change in commercial mobile devices, paving the way to new app-based business models through an attractive device interface. This helped spur the introduction of numerous creative third-party companies, which then set off an explosive evolution of lucrative mobile data services.

The "app model" allowed the global community of stakeholders and developers to access mobile communication. Through apps, each developer with their own interests and profit incentives could drive their own solution independently. Making the app concept device-centric enabled an economy of scale that would have been unlikely through CSP-centric models. The apps could be found in most cases through an intuitive, touch-based user interface that did not require manuals, reading or training on how to find, download or access relevant apps. This model then became attractive for all ages.

It is worth noting that the app model was developed as a mobile-first concept and became an indicative marker of a wider shift that was beginning across the industry, moving from wireless to fixed solutions.

4G LTE: completing the shift to global mobile internet services

WCDMA-based 3G carried many legacy features from earlier generations. This prompted the industry to develop a new truly packet-based system to support increasingly diverse

and demanding MBB services [29]. Much of this work would form the backbone of the upcoming 4G Long Term Evolution (LTE) standard.

The packet bearer in LTE had many similarities with HSPA, albeit with one distinct and crucial difference: LTE was built only for packet data communication and could therefore enable far lower latency than its predecessor.

LTE became the first truly global standard, receiving widespread buy-in from North American stakeholders, further increasing the potential for economies of scale, particularly in innovation-rich US markets. The launch and evolution [30] of LTE firmly established the global success of MBB, and the paradigm shift that entailed. This evolution has continued into 5G and will most likely be a key aspect of all foreseeable future generations.

Summary: mobile broadband as a mobile evolution driver through the decades

Many defining technology developments and strategic decisions that drove mobile network evolution in previous decades occurred with the explicit goal of delivering MBB opportunities to the masses. These mainly comprised RAN developments, which had to solve numerous problems to increase capacity by several orders of magnitude per decade [31].

In this period, there were also significant developments towards an evolved packet core network [32]. Among the more fundamental steps, the transition from circuit-switched communication, which was initially used both for voice and data, to a packet-based solution, HSPA, ranks as one of the most pivotal breakthroughs to MBB. This same concept constituted the basis for LTE, which delivered a notable step change for MBB opportunities.

Today, the transition to MBB has proved to be a true paradigm shift.

- MBB has revolutionized how we communicate.
- MBB is available to a majority of people on the planet.
- MBB is the preferred internet access for most people and the only internet access for billions.
- MBB is the preferred means of providing private and public services and often the only way of accessing the service.
- Voice calls are no longer as common; other means of communication have reduced the need for voice, which is now also carried over other apps available on the internet.

The paradigm shift to mobile data has changed how we communicate in every country, social group and business. As many public and private services now assume the availability of smartphones, MBB is expected to be the dominant method for private communication, businesses, and service provisioning for a long time to come.

During this transition period, Ericsson has been a key contributor in shaping mobile internet and ETR has played an important role in sharing the development of technologies, services and business development [33]. A retrospective look back into the ETR archives reveals a fascinating journey of innovation leading up to the tremendous success of the mobile internet.

THE ERICSSON MOBILE PHONE STORY

Ericsson's mobile phone adventure started in Lund in 1983. Initially working from inside rented trailers, the company rode on the transition to digital standards [34] and became a household brand and global leader in the 1990s.

Ericsson ventured successfully into all mobile phone segments, from small [16] to rugged [17] as well as early smartphones [35]. The rapidly growing market for accessories such as earphones and in-car media also inspired the invention of Bluetooth [36]. Market consolidation, together with the prospect of a growing mobile phone market, drove the creation of the Sony Ericsson Mobile Communications (SEMC) joint venture, as well as the chipset company Ericsson Mobile Platforms (EMP), in 2001.

The following decade was one of success for both SEMC and EMP. SEMC utilized its holding companies' assets in mobile communications and entertainment to grow its product portfolio as well as pioneering the Symbian smartphone segment. EMP was equally successful as the world's largest supplier of chipsets for 3G phones in the mid-2000s [37].
In line with a strong market consolidation in mobile phone chipsets, Ericsson brought EMP into a joint venture with STMicroelectronics, ST-Ericsson, in 2009. In 2012, Sony acquired Ericsson's stake in Sony Ericsson, forming Sony Mobile Communications and only two years later, in late 2014, the majority of the former EMP engineers were transferred to Ericsson's Networks division, marking the end of Ericsson's 31-year history in mobile phones.

The authors

Joakim Bergström joined Ericsson in 1998 and is a senior expert in RAN standardization at Business Area Networks. He has more than 20 years' experience in RAN-related research and standardization work and has been deeply involved in the creation of 3G, 4G and 5G radio access technologies. With a focus on RAN, he has worked in the areas of operations and maintenance, transport, spectrum regulation and open source. Bergström holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

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

Björn Ekelund joined Ericsson in 1987 to architect Ericsson's first mobile phone for the GSM standard, and over the next three decades has held leading roles in Ericsson's mobile phone and chipset businesses. He now leads electronics, electromagnetics and robotics research at Ericsson Research. Ekelund holds an M.Sc. in electrical engineering and a Ph.L. in telecom microelectronics from Lund University in Sweden. He is a fellow of the Royal Swedish Academy of Engineering Sciences and a delegate of the Royal Swedish Academy of Sciences.

Kjell Gustafsson joined Ericsson in 1994 to build up a research group focusing on mobile phone technology. From research he transitioned into product development and system management and has held leading roles in these areas since then. He is currently Head of Standards & Technology at Ericsson in Lund. Gustafsson holds an M.Sc. in electrical engineering and a Ph.D. in automatic control from Lund University.

Johan Lundsjö joined Ericsson in 1996 where his initial focus was research, design and standardization of 3G radio interface protocols and network architectures. He has held various technical and people leader positions during the course of research and early development of 3G, 4G and 5G mobile systems, as well as in research on related network and cloud technologies. He is now Director of Communication at Ericsson Research, where current focus is on future 6G systems. Lundsjö holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology.

References

1. Ericsson Technology Review. ↗
2. Ericsson Review. ↗
3. Ericsson Review, 1991, Vol. 68, No. 3, pp. 66-71, "Trends in Mobile Communications", Hellström, K., Lundkvist, Å., ↗
4. Ericsson Review, 1988, Vol. 65, No. 3, pp. 82-92, "Ericsson Strategies and Technologies for the 1990s", Ramqvist, L., ↗
5. Ericsson Review, 1990, Vol. 67, No. 1, pp. 42-52, "The Future of Cellular Telephony", Jansson, H., Swerup, J., Wallinder, S., ↗
6. International Telecommunication Union, ICT Data and Statistics Division, 2015, "ICT Facts & Figures – The World in 2015", Sanou, B., ↗
7. World Bank, ITU/ICT Indicators Database, "Mobile Cellular Subscriptions". ↗
8. Ericsson Review, 1982, Vol. 59, No. 3, pp 121-124, "Packet Switching Principles", Kleinrock, L., ↗
9. Ericsson Review, 1982, Vol. 59, No. 3, pp 125-128, "Packet Switching Economics", Roberts, L. G., ↗
10. Ericsson Review, 1999, Vol. 76, No. 3, pp. 110-121, "Third-generation radio access standards", Nilsson, M., ↗
11. Ericsson Review, 1989, Vol. 66, No. 1, pp. 33-39, "Mobitex –a New Network for Mobile Data Communications" , Berntson, G., ↗
12. Ericsson Review, 1997, Vol. 74, No. 3, pp. 98-103, "Internet and intranet connections over GSM", Källström, O., ↗
13. Ericsson Review, 1999, Vol. 76, No. 2, pp. 82-88, "GPRS –General packet radio service", Granbohm, H., Wiklund, J., ↗
14. Ericsson Review, 1999, Vol. 76, No. 1, pp. 28-37, "EDGE –Enhanced data rates for GSM and TDMA/136 evolution", Furuskär, A., Näslund, J., Olofsson, H., ↗
15. Ericsson Review, 1993, Vol. 70, No. 3, pp. 82-92, "Radio Access Technology Evolution", Lindell, F., Sköld, J., Willars, P., Nilsson, E., ↗
16. Ericsson Review, 1994, Vol. 71, No. 4, pp. 93-99, "New-generation True Pocket Phones", Lindoff, M., ↗
17. Ericsson Review, 1999, Vol. 76, No. 1, pp. 8-13, "Ericsson's Pro products – Adapting mass-market technology to fit specialized needs", Gratorp, A., Nilsson, P., Smedman, T., ↗
18. Ericsson Review, 1999, Vol. 76, No. 4, pp. 206-213, Mobile Internet–An industry-wide paradigm shift?, Andersson, C., Svensson, P., ↗
19. Ericsson Review 1999, Vol. 76, No. 3, pp 122-131, "Toward third-generation mobile multimedia communication", Nilsson, T., ↗
20. Ericsson Review, 2003, Vol. 80, No. 2, pp. 56-65, "WCDMA evolved–High-speed packet-data services", Parkvall, S., Englund E., Malm, P., Hedberg, T., Persson, M., Peisa, J., ↗

21. Ericsson Review, 2005, Vol. 82, No. 1, pp. 14-23, “Broadband data performance of third-generation mobile systems”, Sköld, J., Lundevall, M., Parkvall, S., Sundelin, M., ↗
22. Ericsson Review, 1998, Vol. 75, No. 4, pp. 150-153, “WAP—The wireless application protocol”, Erlandson, C., Ocklind, P., ↗
23. Ericsson Review, 2000, Vol. 77, No. 1, pp. 14-19, “WAP—The catalyst of the mobile Internet”, Pehrson, S., ↗
24. Ericsson Review, 2007, Vol. 84, No. 1, pp. 44-49, “Multimedia telephony for IMS —Interoperable VoIP with multimedia support”, Enström, D., Nohlgren, A., Olofsson, H., Peisa, J., Synnergren, P., ↗
25. Ericsson Review, 2007, Vol. 84, No. 1, pp. 50-59, “Ericsson IMS Client Platform”, Kessler, P., ↗
26. Ericsson Review, 2008, Vol. 85, No. 1, pp. 8-13, “Communication Services —The key to IMS service growth”, Olsson, U., Stille, M., ↗
27. Ericsson Review, 2010, Vol. 87, No. 1, pp. 34-39, “Ericsson Business Communication Suite”, Olrog, C., Olsson, U., ↗
28. Ericsson Review, 2010, Vol. 87, No. 1, pp. 16-21, “The app store and beyond”, Eliasson, J., Olander, J., Sporre, P., Wilhelmsson, J., ↗
29. Ericsson Technology Review, 2008, Vol. 85, No. 2, pp. 77-80, “Key features of the LTE radio interface”, Dahlman E., Furuskär A., Jading, Y., Lindström, M., Parkvall S., ↗
30. Ericsson Review, 2010, Vol. 87, No. 2, pp. 22-28, “Next generation LTE, LTE-Advanced Next generation LTE, LTE Advanced”, Dahlman E., Parkvall S., Furuskär A., ↗
31. Ericsson Review, 1987, Vol. 64, No. 3, pp. 160-168, “Digital Cellular Radio for the Future”, Lindell, F., Swerup, J., Uddenfeldt, J., ↗
32. Ericsson Review, 2007, Vol. 84, No. 3, pp. 98-104, “LTE-SAE architecture and performance”, Beming, P., Frid, L., Hall, G., Malm, P., Noren, T., Olsson, M., Rune, G., ↗
33. Ericsson Review, 2009, Vol. 86, No. 1, pp. 27-30, “Ericsson Research—10 years of shaping change”, Uddenfeldt, J., Eriksson, H., Wahlberg, U., Färjh, J., ↗
34. Ericsson Review, 1993, Vol. 70, No. 4, pp. 140-155, “Universal Personal Telecommunication (UPT) — Concept and Standardization”, Sundborg, J., ↗
35. Ericsson Review, 2001, Vol. 78, No. 1, pp. 44-48, “The R380s—The first smartphone from the Ericsson-Symbian partnership”, Bridges, S., ↗
36. Ericsson Review, 1998, Vol. 75, No. 3, pp. 110-117, “Bluetooth—The universal radio interface for ad hoc, wireless connectivity”, Haartsen, J.,↗
37. Ericsson Review, 2005, Vol. 82, No. 1, pp. 32-43, “The EMP Story”, Kornby, M., ↗

Further reading

- The evolution of mobile standards from 1G to 6G ↗
- Ericsson feature series, The rise of telephony ↗
- Ericsson research paper, Mobile broadband drives economic development ↗
- Ericsson Technology Review library, including early editions ↗

Acknowledgements

We would like to acknowledge the contributions made by generations of skilled, innovative and insightful Ericsson colleagues within all domains required to facilitate this monumental paradigm shift.

Broad beamforming technology in 5G Massive MIMO

Authors:
Maksym Girnyk, Henrik Jidhage, Sebastian Faxér

In contrast to the most common approach to Massive MIMO (multiple input, multiple output) in use today, the innovative dual-polarized beamforming technique developed at Ericsson offers the important advantage of creating broad radiation patterns, while avoiding the underutilization of power resources.

Massive MIMO is a key 5G technology that is used in most mid-band time-division duplex (TDD) deployments to achieve better coverage, higher user bitrates and increased network capacity [1].

With Massive MIMO, throughput and network capacity can be increased by enabling user-specific beamforming of the data channel, forming narrow beams with high antenna gain pointed at a certain user. Cell-specific transmission is still needed for broadcast and control signaling, however. A common approach to achieve this is to utilize Synchronization Signal Block (SSB) sweeping, wherein multiple narrow beams carrying control information are transmitted in sequence over the intended cell area. The downside of this approach is that it leads to additional overhead, resulting in lower capacity and peak rate. Ericsson offers an alternative approach that enables an efficient realization of cell-specific transmission through the construction of a single broad SSB beam for Massive MIMO.

Ericsson's dual-polarized beamforming (DPBF) technique – also known as array-size invariant beamforming [2,3] – is already widely used in Ericsson radios for various purposes, including creating broad SSB beams. The technique is applicable to both single-SSB and multi-SSB scenarios, and it provides the ability to design radiation patterns to match nearly any cell shapes of interest.

Beamforming basics

To use spectrum as effectively as possible in 5G mid-band TDD deployments, most communication service providers (CSPs) install Massive MIMO radios at the base stations (BSs). A Massive MIMO antenna array provides the BS with beamforming capabilities, which are realized by forming a radiation pattern that amplifies power in certain directions, while muting it in others. This is achieved by applying complex-valued beamforming weights to the radiating elements of the antenna array that define how their individual radiation patterns are combined in the far-field region. As **Figure 1** illustrates, the radiation pattern of an antenna array incorporates two effects: the radiation pattern of a single antenna element, and the so-called array factor due to the superposition of the radiated electric fields of all the array elements. The radiation pattern of an array is a product of both the element pattern and the array factor.

A popular choice of the beamforming weights is based on discrete Fourier transform (DFT) vectors that ensure narrow beams with maximum possible array gain. In this way, the BS can direct the transmitted energy toward a user mobile terminal (user equipment (UE)), which can increase

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **BS** – Base Station | **CSP** – Communication Service Provider | **dBi** – Decibels Relative to Isotropic | **DFT** – Discrete Fourier Transform | **DPBF** – Dual-Polarized Beamforming | **HPBW** – Half-Power Beamwidth | **LTE** – Long Term Evolution | **MIMO** – Multiple-Input, Multiple-Output | **NR** – New Radio | **PA** – Power Amplifier | **PBCH** – Physical Broadcast Channel | **PDCCH** – Physical Downlink Control Channel | **PDSCH** – Physical Downlink Shared Channel | **RSRP** – Reference Signal Received Power | **SINR** – Signal-to-Interference-plus-Noise Ratio | **SSB** – Synchronization Signal Block | **TDD** – Time-Division Duplex | **UE** – User Equipment

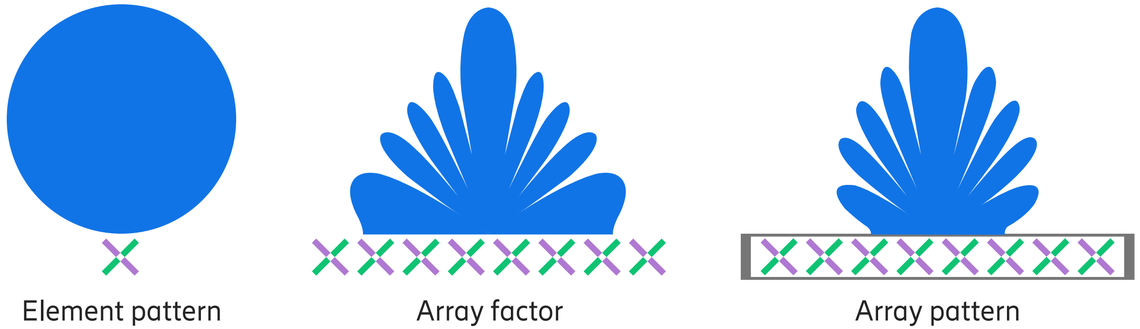


Figure 1: Pattern multiplication property

the received dedicated signal power at the latter and consequently improve its data rate. The larger the antenna array at a BS, the narrower its radiation pattern can become, and the larger the gain of such user-specific beamforming. Corresponding beams are often referred to as traffic beams (or user data beams) and they are essential for transmitting data.

Meanwhile, there are situations where it is beneficial to transmit signals to many UEs at once or when the channel state information to a UE is outdated or unknown. Such transmissions could be related to synchronization, initial access or mobility signaling. This may be particularly relevant for control and broadcast channels, such as the physical downlink control channel (PDCCH) and the physical broadcast channel (PBCH), as well as cell-specific signals for mobility, such as SSB. Such channels are characterized by the broadcast nature of the information and low bitrate per user. For such cell-specific transmission, a broad beam (or a limited number thereof) is needed to cover the entire macro sector with a roughly even radiation level.

Unfortunately, for large antenna arrays, creating a broad beam is not straightforward. Direct enlargement of the array aperture leads to the immediate shrinkage of the radiation pattern. There are, however, ways to overcome this challenge. The simplest approach to create a broad beam is to transmit with a single antenna element, which typically has a broad radiation pattern. Unfortunately, though, this approach results in hugely underutilized power resources, which leads to significantly reduced coverage.

Various beam-shape optimization algorithms have been used in industry and academia to obtain broader beams with certain levels of imperfection. Some of these tune amplitudes of the beamforming weights of certain antenna elements in the array to obtain a broad beam. Others exploit only phases of the beamforming weights to broaden the beam shape. Both approaches have advantages and disadvantages. Most notably, amplitude tapering produces spatially broad beam patterns at a cost of reduced total transmit power, while phase tapering preserves full power utilization at a cost of spatial ripples in the beam shape [4].

An alternative method for covering a sector relies on performing a sweep over a set of narrow beams. Such sweeps constitute a common approach used for beam management procedures in 5G New Radio (NR). For example, instead of transmitting a single SSB utilizing one broad beam, multiple SSBs are sequentially transmitted in what is known as a “synchronization signal burst” containing a set of narrow beams spanning the angular directions of the sector. This solution provides high antenna gains for the entire sector coverage. However, it leads to increased overhead and complexity, which increases with the number of SSB beams to sweep over. The solution also leads to increased UE battery usage because the UE needs to actively listen for an SSB during the entire sweep procedure.

Academic papers have proven that, for a given polarization, transmission from a single element is the only possible solution to produce a broad beam with a spatially flat array factor [5]. These findings led to the belief that the creation of broad beams with full power utilization might not be possible and that SSB sweep was needed to achieve good SSB coverage. Based on this, some experts advocated for 3GPP (3rd Generation Partnership Project) to mandate SSB sweep for beam management in 5G NR but after careful consideration, 3GPP opted for flexible control channel configuration instead. As a result of that decision, our mid-band radios are able to employ a single broad SSB beam without any beam sweep. Thanks to an Ericsson innovation, broad beams can be created without compromising on power amplifier (PA) utilization.

Ericsson’s dual-polarized beamforming technique

In 5G NR, the downlink data transmission over the physical downlink shared channel (PDSCH) is decoupled from that of the SSB. Hence, there is no direct correlation between the reference signal received power (RSRP) and the signal-to-interference-plus-noise ratio (SINR) measured on the SSB to the data transmission performance. The latter is addressed by the traffic beams during the data transmission phase. Note that this is different from 4G LTE (Long Term Evolution), where the RSRP/SINR measured on the cell-specific reference signal is tightly coupled with the bitrate performance. Meanwhile, in 5G NR, the SSB is mostly used for providing coarse synchronization, radio link monitoring, open-loop uplink power adjustments, PBCH decoding in initial access and cell selection/detection as part of mobility procedures.

For large antenna arrays, creating a broad beam is not straightforward.

PBCH is a very robust channel and can be decoded at very low SINR conditions; typically, other channels in the initial access procedure have lower link budgets. Thus, PBCH does not constitute a coverage bottleneck and improving SSB RSPR/SINR through beam sweeping may therefore not result in higher effective coverage or robustness. Considering mobility, the relative differences in RSRP/SINR between SSBs of different cells is what matters for proper cell association. An increase in antenna gain of the SSB will therefore not lead to better mobility decisions. In fact, SSB beam sweeping might even be detrimental to mobility performance when using RSRP/SINR-based

mobility, as interference conditions are not captured accurately.

Given these observations, a trade-off can be formulated between the increased antenna gain and the increased overhead and latency introduced by a beam sweep. Thus, excessively increasing the number of SSB beams beyond the optimal trade-off point will not lead to increased NR coverage in practice but may only reduce system performance due to the increased overhead [6].

The DPBF technique provides a means of creating broad beams with a spatially flat array factor.

The optimal number of SSB beams depends on the frequency band and the size of the antenna array. For sub-4GHz, for example, a single SSB beam is deemed sufficient, while for millimeter wave frequencies, 12 SSB beams are typically used for macro deployments. The upcoming spectrum between 6-15GHz may need a number in between, depending on the antenna size. Regardless of how many SSB beams are deployed, a power-efficient method is needed to synthesize beams with the desired radiation properties.

To design broad beams with excellent power utilization, Ericsson researchers developed the DPBF technique. The approach is based on the fact that modern antenna systems are naturally built to exploit a pair of orthogonal polarizations. This dual-polarized nature provides an additional degree of freedom to design a radiation pattern purely by means of phase-only techniques, removing the limitations of the amplitude-tapering methods.

The DPBF technique makes it possible to achieve a range of beamwidths with large antenna arrays, while guaranteeing efficient PA utilization. This can be used for the purpose of cell shaping, where the beam shapes and pointing directions are adapted in a coordinated manner across cells for cell-defining signaling, such as SSB in 5G NR. This implicitly determines which cells serve the UE, providing a means for reducing interference and load balancing among the cells.

To realize a broad radiation pattern from a dual-polarized antenna array, it is necessary to excite the radiation elements in two polarizations by a pair of Golay complementary arrays [7]. These pairs were discovered by the Swiss mathematician and physicist Marcel Golay in 1949 within the field of multi-slit spectrometry [8]. Their distinctive property is complementarity, which means that the power spectral densities of the two arrays complement each other and add up to a constant for all frequencies. Because we know that the (power-domain) array factor of a beam is related to the power spectral density [2], applying complementary beamforming weights results in per-polarization radiation patterns that, although they are not broad themselves, add up to a broad-beam pattern with an omnidirectional array factor.

The principle is illustrated in **Figure 2**, where the per-polarization beam patterns compensate for each other's nulls with corresponding peaks. The green area indicates a beam pattern in one polarization, while the purple area shows the pattern in the other polarization. Dual-polarized UE observes the total radiation pattern shown in light blue. The total obtained radiation pattern is broad, covering the entire cell with control signaling. UE equipped with two antennas adjusted for two orthogonal polarizations can pick up power from both polarizations and hence observe this broad pattern. Over-the-air trials in a macro-cell scenario [9] have successfully validated NR SSB beams designed using the DPBF approach.

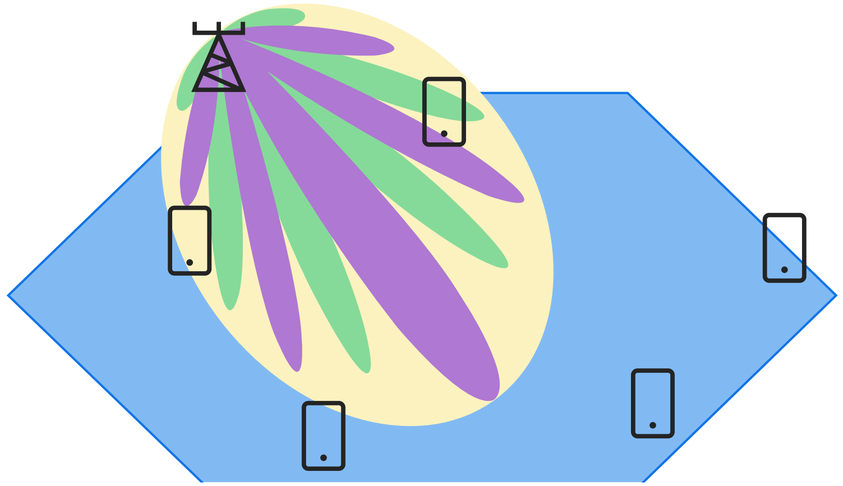


Figure 2: Forming a broad radiation pattern using a dual-polarized antenna array

Dual-polarized beamforming use cases

The DPBF technique provides a means of creating broad beams with a spatially flat array factor. However, most antenna systems are not designed with truly omnidirectional coverage in mind. Instead, they are built to cover an angular sector corresponding to a cell. For example, in a conventional three-sector deployment, a cell has the angular width of only 120°, which means there is a need to design an array factor that is narrower than the omnidirectional one.

Violating the complementarity of the per-polarization beamforming matrices by means of distorting the phases of some of the beamforming weights in a controlled way makes it possible to achieve any beamwidth, ranging from a spatially flat array factor to a narrow DFT beam [10]. This can be used to optimize the radiation pattern to fit the cell shape, use case and deployment scenario of interest. At Ericsson, we design the beamforming weights by combining the complementary phase-only beamforming with a small amplitude taper – with loss of 0.5dB at most, for example – to obtain optimized SSB beam shapes.

The graphs in **Figure 3** are based on a theoretical model representing an active antenna system with eight columns, which is the typical configuration for mid-band products with 32 or 64 branches. Each radiating element has two orthogonal polarizations (slanted +45° and -45°) and a half-power beamwidth (HPBW) of 90°. The figure shows the horizontal farfield patterns of SSB beams designed using the DPBF technique. The target cell shape for the examples in the figure a 120°-wide sector for a typical three-sector deployment. As demonstrated, the DPBF method is useful for creating both broad single-beam designs, such as NR single-SSB, as well as multi-beam designs, such as NR multi-SSB.

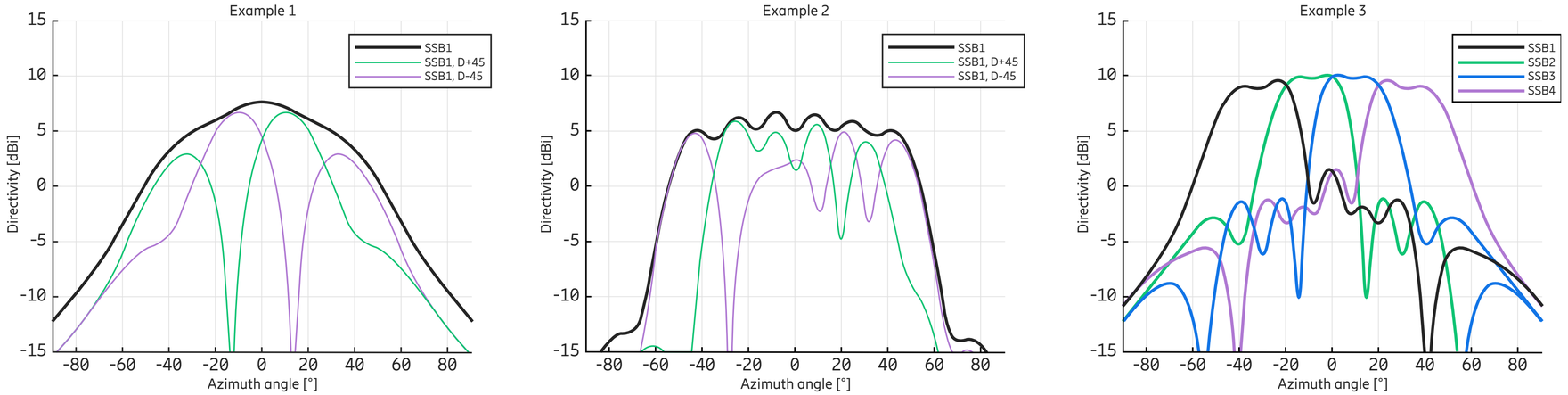


Figure 3 – SSB broad-beam designs using the DPBF technique

The left section of Figure 3 shows an example of a broad-beam design that provides the same cell shape as a Gaussian beam with 65° HPBW. The power radiation pattern (blue) consists of the sum of the +45° and -45° polarized radiation patterns (D+45 and D-45). This beam shape is appropriate for covering a 120°-wide sector with broadcast signaling. However, one potential drawback of this type of Gaussian broad-beam design is that the resulting radiation pattern will most likely not describe the true capabilities of antenna systems with respect to the utilized traffic beams. The data beam capability, which will follow the average embedded subarray patterns, is typically wider than the Gaussian 65°-wide beam.

The middle section of Figure 3 shows an alternative broad-beam design that is also based on the DPBF approach, in which a single SSB beam is optimized to provide improved tracking between SSB and traffic beams. The design target in this case is the envelope of the traffic beams used in the applicable sector. This example shows the results of beam optimization for a sector with a width of 120°, but the method can be used for any sector width. The power radiation pattern (blue) consists of the sum of the two individual polarizations (D+45 and D-45).

The right section of Figure 3 presents a multi-SSB example with four SSB beams, showing the power radiation patterns without individual per-polarization patterns. In this example, the four SSB beams span the desired sector with an increased peak directivity and a reduced ripple in the main beam region.

The three examples in Figure 3 demonstrate the application of the DPBF beam design in the horizontal dimension. The method, however, can be applied simultaneously in the horizontal and vertical dimensions, if desired. Typically, a 2D broad-beam design for antenna systems with many branches can lead to a large and time-consuming optimization. However, oftentimes the desired 2D beamforming weight matrices possess separability properties, and hence the DPBF optimization can be reduced into two simpler optimization problems: one for the horizontal dimension and the other for the vertical dimension.

The DPBF concept also defines how to create a second broad beam with polarization orthogonal to a first broad beam. This can be useful for certain use cases, such as uplink optimization and channel state information reference signal mappings, which may require broad beams in two different polarizations.

Conclusion

Massive MIMO (multiple input, multiple output) is an essential technology to enhance the capacity of 5G networks, most commonly by using narrow traffic beams to improve user data rates. There is, however, also a need for broad-beam coverage that is suitable for broadcast and control signaling, such as 5G New Radio Synchronization Signal Block (SSB). In Ericsson 5G products, SSB beams are created using a dual-polarized beamforming (DPBF) technique that is based on the mathematical concept of Golay array pairs. This innovative technique enables the construction of broad-beam shapes for large antenna arrays without underutilizing power resources. Using this approach, it is possible to design a radiation pattern of an array that defines appropriate cell shapes in relation to user and usage distribution, as well as mobility, which is the essence of a high-capacity cellular system. The DPBF technique can also be used to create broad beams with polarization diversity.

The authors

Maksym Girnyk joined Ericsson in 2014 and currently works as a radio study driver within Product Engineering Unit Radio, leading pre-development activities for the product development of Massive MIMO radio units. In previous roles, he worked with the development and standardization of Massive MIMO technologies for 5G NR, as well as establishing the vision, use cases and technology components for 6G. In 2015, Girnyk received the Best Conference Paper Award issued by the IEEE VT/COMM/IT Sweden Chapter Board. He holds an M.Sc. in radio communication systems from CentraleSupélec, Gif-sur-Yvette, France, and a Ph.D. in telecommunications from KTH Royal Institute of Technology, Stockholm, Sweden.

Henrik Jidhage joined Ericsson in 1996 and is currently working as senior specialist on antenna systems for network performance within Product Engineering Unit Radio, where he is responsible for early phase antenna systems studies and technology roadmap. He has been responsible for implementation studies of dual-polarized beamforming for current Massive MIMO products. Jidhage holds an M.Sc. in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden.

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

Acknowledgements

The authors would like to thank their former colleague Sven Petersson for the initial discovery of the broad-beam concept.

References

1. Ericsson, Massive MIMO handbook ↗
2. IEEE, IEEE Transactions on Communications (Volume: 69, Issue: 12, pp. 8429-8442), Efficient Cell-Specific Beamforming for Large Antenna Arrays, December 2021, Girnyk, M. A.; Petersson, S. O. ↗
3. IEEE, IEEE Transactions on Vehicular Technology (Volume: 71, Issue: 11, pp. 11772-11785), Energy-Efficient Design of Broad Beams for Massive MIMO Systems, November 2022, Petersson, S. O.; Girnyk, M. A. ↗
4. 3GPP R1-1700772, On forming wide beams, January 2017, Ericsson ↗
5. IEEE, IEEE Transactions on Signal Processing (Volume: 64, Issue: 9, pp. 2365-2374), Broadbeam for Massive MIMO Systems, May 2016, Qiao, D.; Qian, H.; Li, G. Y. ↗
6. Ericsson blog, Benchmark measurements in 5G networks, August 21, 2020, Lazarevic, Z. ↗
7. Springer Nature, Designs, Codes and Cryptography (Volume: 44, pp. 209-216), Golay complementary array pairs, July 2007, Jedwab, J.; Parker M. G. ↗
8. OSA, Journal of the Optical Society of America, Multi-Slit Spectrometry (Volume: 39, Issue: 6, pp. 437-444), June 1949, Golay, M. J. E. ↗
9. IEEE, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Dual Polarization Beamforming Coverage Demonstrated with 5G NR SSB, 2021, Simonsson, A.; Petersson, S. O.; Widell, G. ↗
10. IEEE, 2020 IEEE Wireless Communications and Networking Conference (WCNC), A Simple Cell-Specific Beamforming Technique for Multi-Antenna Wireless Communications, 2020, Girnyk, M. A.; Petersson, S. O. ↗

Further reading

- IEEE, 2000 European Conference on Antennas and Propagation (EuCAP), Power-Efficient Beam Pattern Synthesis via Dual Polarization Beamforming, 2020, Petersson, S. O. ↗
- IEEE, 2021 IEEE Vehicular Technology Conference (VTC-Spring), Massive MIMO Muting using Dual-polarized and Array-size Invariant Beamforming, 2021, Frenger, P.; Wang Helmersson, K. ↗
- IRE, IRE Transactions on Information Theory (Volume: 7, Issue: 2, pp. 82-87), Complementary series, April 1961, Golay, M.J.E. ↗
- IEEE, IEEE Transactions on Information Theory (Volume: 24, Issue: 5, pp. 546-552), Multiphase Complementary Codes, September 1978, Sivaswamy, R. ↗
- IEEE, IEEE Transactions on Information Theory (Volume: 26, Issue: 6, pp. 641-647), Polyphase complementary codes, November 1980, Frank, R. ↗
- Ericsson, Massive MIMO ↗

6G network architecture – a proposal for early alignment

Authors:
Torbjörn Cagenius, Gunnar Mildh, Göran Rune,
Jari Vikberg, Mattias Wahlqvist, Per Willars

Network evolution to support the 6G vision, use cases and requirements within the 2030 time frame is a topic of great interest in the telecommunications industry today. To ensure the smooth introduction of 6G, and the ability to monetize on it from day one, Ericsson advocates early alignment on a common set of principles that will lead toward a more focused ecosystem.

6G research is well underway both in the telecom industry and in academia, targeting commercial deployment of the technology around 2030. Based on the substantial work done so far, the industry has the opportunity to agree on some key architecture principles.

The vision for 6G [1,2] is built on the desire to create a seamless merging of the digital and physical worlds. This seamless reality of the future will provide new ways of meeting and interacting with other people, new possibilities to work from anywhere, and new ways to experience faraway places and cultures. This will facilitate further digitalization of industries and management of smart cities, enabling, for example, improved personal safety applications, less waste and greater sustainability.

6G will improve network performance by meeting high demands on traditional performance indicators such as capacity, coverage, bit rates and short latency, as well as new performance indicators related to service availability, service assurance and predictability, network resilience, trustworthiness, energy performance and sustainability. These performance indicators need to be met while simultaneously ensuring cost-effective deployments and a smooth introduction into existing networks.

At Ericsson, we believe the best way to support the 6G vision and requirements is to standardize a 6G architecture that allows for the smooth introduction of 6G capabilities into future public and private networks. We foresee that the 6G architecture will build on the ongoing trend of network horizontalization [3], enabling the 6G radio-access network (RAN) and core network (CN) functions to benefit from the fast evolution of cloudification, IT frameworks, automation, open interfaces and artificial intelligence (AI)/machine learning (ML).

Technology trends impacting future networks

Figure 1 illustrates the most significant technology trends that will impact the overall network architecture in the 6G time frame of 2030. They fall into four main categories:

1. Enablers of monetization and exposure
2. Automation of network operations
3. Cloud-native design and deployment
4. Network architecture evolution.

Enablers of monetization and exposure are critical to the development of 6G. Monetizing 5G capabilities is already a top priority for communication service providers today [4] and it is obvious that it will continue to be important in the 6G time frame as well. 6G networks must be able to reuse and expand on the evolution of 5G exposure and monetization functionality from day one.

Terms and abbreviations

5GC – 5G Core | **AI** – Artificial Intelligence | **CN** – Core Network | **CSP** – Communication Service Provider | **E2E** – End-to-End | **LCM** – Life-Cycle Management | **LLS** – Lower-Layer Split | **MBB** – Mobile Broadband | **ML** – Machine Learning | **NF** – Network Function | **NR** – New Radio | **NSA** – Non-Standalone | **RAN** – Radio-Access Network | **RAT** – Radio-Access Technology | **RNA** – Radio Network Area | **RU** – Radio Unit | **SA** – Standalone | **SBA** – Service-Based Architecture | **SBI** – Service-Based Interface | **SMO** – Service Management and Orchestration | **UE** – User Equipment

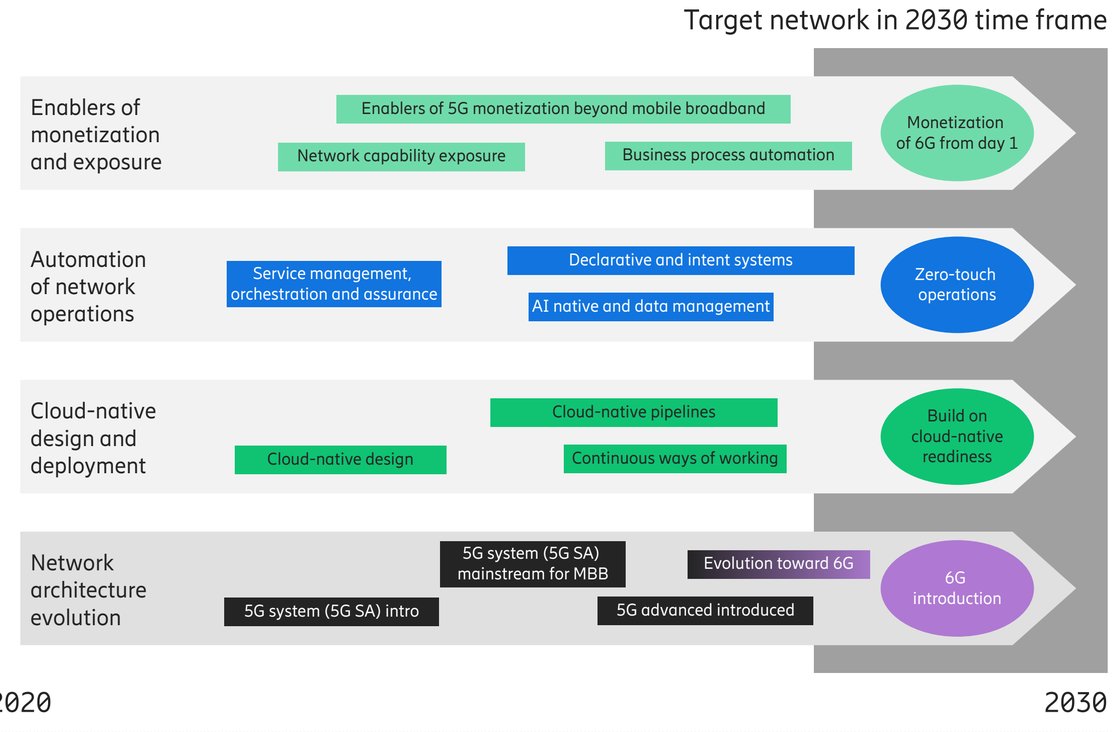


Figure 1: Multiple trends impacting future networks in the 6G time frame

The automation of network operations is another important trend impacting 6G development. As network complexity increases – with more radio-access technologies (RATs), new band combinations, more network slices, different functionalities and so on – it will become ever more difficult to find good optimization points manually. AI/ML functionality that replaces the manual work of developing, deploying, managing and optimizing the mobile network, including intent-based management [5], must, therefore, be a core component of 6G.

Cloud-native design and deployment is the third major technology trend impacting the development of future networks. A cloud-native design with containerized deployments enables an efficient separation of software from hardware. This form of disaggregation creates a possibility to separate evolution and innovation in the cloud-native applications running as workloads and the underlying cloud infrastructure and associated tooling. Additionally, it affects processes in the operation of the networks, such as tools for life-cycle management (LCM) and the automation of integration and deployment. Greater use of AI/ML technology will simplify network operation and optimization and is likely to have wider implications in terms of design and deployment.

The fourth and final trend shown in Figure 1 is network architecture evolution, which is the primary focus of this article. In our view, there are three main aspects of network architecture evolution that require early industry alignment:

1. Migration and spectrum aggregation
2. Radio-access network architecture evolution
3. Core network architecture evolution.

Migration and spectrum aggregation

Since carrier aggregation was introduced in 4G, any new RAT must be able to aggregate more spectrum than its predecessor to improve performance. Further, because the new spectrum that comes with new RATs is typically on higher bands, there is a greater need to combine with lower bands to achieve sufficient uplink performance and coverage.

5G was specified with multiple standardized connectivity options for how to combine 4G and 5G RAT. To avoid market fragmentation [6], the ecosystem settled for only two: non-standalone (NSA) New Radio (NR) and standalone (SA) NR. Even this has proven challenging, splitting the industry focus and pushing some communication service providers (CSPs) to launch 5G twice. The interworking between 4G RAN and 5G RAN for NSA has also proven technically complex for the whole ecosystem, with a large impact on networks and devices. The split control of the user equipment (UE) connection between the gNodeB and the eNodeB implies a tight coupling of the nodes, given that the RATs share a common set of UE capabilities.

In addition, for spectrum aggregation within 5G, two methods were specified (carrier aggregation and dual connectivity), driving extra complexity for interoperability between UE and networks.

5G also provided another solution to the migration problem, by allowing 5G to share spectrum on legacy bands with 4G using dynamic spectrum sharing. Although this provided a paradigm shift for how to migrate RATs, it brought challenges, specifically for overhead, due to 4G having many always-on signals.

Based on experiences from 5G, we believe it would be best to avoid specifying multiple connectivity options for migration to 6G and multiple spectrum aggregation methods within 6G. This decision will reduce complexity and help the industry to focus on a common track.

Radio-access network architecture evolution

Apart from the work done in the traditional standardization bodies, it is important to note that the industry move toward Cloud RAN opens a new multi-vendor environment with a separation of the software application from the cloud infrastructure.

Successful deployment of a standardized multi-vendor interface requires both significant business value and separation of concern to ensure that optimizations in one part of the system can be introduced without affecting other parts. The RAN-CN interface, with multi-vendor deployments all over the globe, is an excellent example of this. Other examples like RAN-UE and RAN-RAN mobility (X2, Xn) require more work because of more technical dependencies but the clear business value motivates the interoperability testing and integration costs.

In addition to the widely deployed multi-vendor interfaces in use today, a lower-layer split (LLS)/fronthaul interface has been identified as an important candidate for 6G and is currently being standardized for 5G. While the business value of an LLS/fronthaul interface seems to be increasing, there is a challenge with respect to the separation of concern, as small inefficiencies may directly lead to performance losses if not addressed properly.

Further evolution of the architecture should aim to minimize complexity.

Core network architecture evolution

Cloud-native design and deployment has provided new implementation technologies and improved ways of working such as software LCM. This development inspired the introduction of the Service-Based Architecture (SBA) of the 5G Core (5GC) network in the 3GPP (3rd Generation Partnership Project), which has made the functional architecture more suitable for cloud deployment by, for example, adopting cloud-friendly protocols. In the SBA, the network functions (NFs) expose services through Service-Based Interfaces (SBIs) instead of using point-to-point protocols, as in previous generations. The purpose of this change was also to create a more flexible and extensible architecture.

The extensibility of the 5GC that the SBA has enabled is demonstrated by the continuous increase in the number of NFs in the 5GC, which has risen from 22 in 3GPP release 15 to 45 in release 17. The number of SBIs has increased at almost twice the rate as the NFs, reaching more than 110 in release 17. The increase of NFs and NF services is strongly related to the introduction of new features and functions in the 5GC.

The flexibility of the 5GC also drives complexity in standardization, development and operational deployment in commercial networks, however. For example, the Service Communication Proxy introduced in release 16 added several modes of operation of the inter-NF communication [7] that need to be applied per SBI and NF service consumer. With different NFs supporting different modes of operation, this leads to a configuration complexity for multi-vendor SBIs of networks in operation.

While continuing to take advantage of the flexibility of the 5GC, further evolution of the architecture should aim to minimize complexity at system level when introducing new functionalities. The ability to manage the complexity in all the steps from design and development to operations will require greater reliance on automation techniques.

By combining existing 5GC extensibility with the potential improvements to the 5GC architecture proposed above, we believe that an evolved 5GC could support a new 6G RAT.

Key assumptions for the 6G architecture

The 6G architecture needs to support the expected new use cases and service requirements for the 2030 time frame and beyond, including enhanced support for immersive communication and new capabilities such as network sensing and zero-energy devices [1].

There is a need for an aligned industry view of a single 6G architecture, avoiding the multiple architecture options

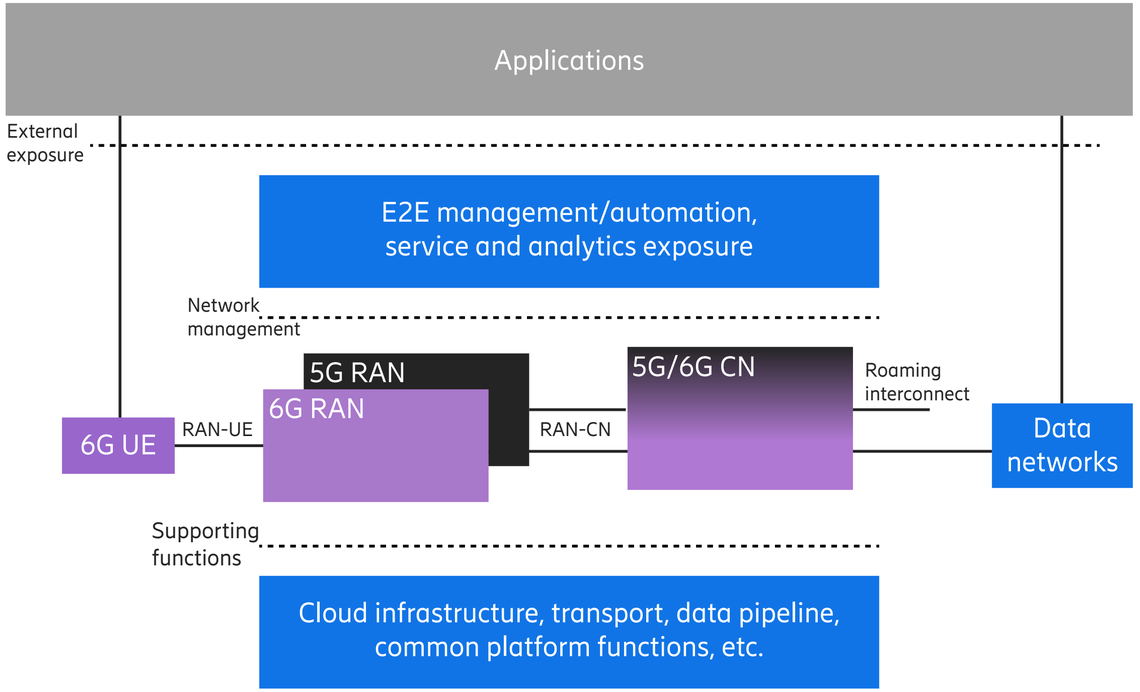


Figure 2: Ericsson's proposal for the 6G architecture, with key open inter-domain interfaces

defined in 5G that caused delayed availability of 5G system network capabilities. This will help to reduce the overall complexity of the 6G standard. Reducing complexity is important, as it will cut down the time to market for new 6G features and the costs for integration and testing in mobile networks.

In addition to the technology trends presented in Figure 1, the 6G architecture will be impacted by the need to support existing and evolving telecom-specific deployment, service, mobility and regulatory requirements. For example, it is expected that 6G will need to provide full support for telephony services and emergency calls, support seamless inter-RAT/system mobility, and reuse existing sites and transport networks.

Based on our understanding of the requirements, the technology trends and experience of 5G, as well as our belief in the need for an evolutionary approach, we have designed the 6G architecture proposal shown in **Figure 2**,

which includes the key open interfaces between domains. The architecture is based on the principle of horizontal separation of the NFs from the underlying platform and overlying end-to-end (E2E) management and exposure.

6G radio-access network as a new standalone radio-access technology

Based on experiences from the 5G migration, Ericsson believes that the 6G RAT should be specified in standalone mode only, with UE that is connected to 6G alone. This would provide an aligned industry focus, avoiding the need to launch multiple versions of 6G. It also simplifies the architecture to have only one control point for the UE connection in the 6G RAN.

6G-connected UE should perform better than 5G UE in the same location. This means that 6G must be able to use potential new centimeter-wave bands together with legacy frequency division duplex and time division duplex bands. An efficient, dynamic spectrum-sharing mechanism therefore needs to be standardized from the start, allowing 5G and 6G UE to share a common pool of resources. Given that 5G has significantly less need for always-on reference signals than 4G, there are opportunities to significantly improve efficiency compared with 4G-5G spectrum sharing.

With 6G deployed on a mix of legacy and new spectrum, a good spectrum aggregation solution will be of critical importance. To allow full RAN optimization, and avoid complex interactions, this should be based on a single instance in the network to control a given UE, with mechanisms for fast adaptation regarding the spectrum that is used at any point in time. This means the full set of radio resources used for a UE connection is decided in one place, considering the full capability of the UE for different band combinations.

To meet increased demands on efficiency using deployed resources (spectrum and radio sites) and increased energy efficiency, a larger degree of elasticity and pooling of radio resources is needed. This includes using multiple radio sites and spectrum resources for connected mode transmission (distributed MIMO (multiple-input, multiple-output) [8], for example) when needed, but also the power down of radio sites or spectrum resources when not needed. Rather than using the concept of nodes represented by a physical base station at one radio site as the basis for optimizing RAN performance, in 6G it should be possible to optimize RAN performance per geographical area, using the radio sites and spectrum in that area according to current needs.

Ericsson believes that 6G should be specified in standalone mode only.

In such a RAN system, the functional dependencies between different parts of the system controlling the resources will be even higher than today. The industry must be very careful to select the key interfaces to be standardized for potential multi-vendor integration. One possible standardized 6G RAN architecture is illustrated in **Figure 3**. The LLS interface is included as a key interface natively supported in 6G RAN, splitting the RAN into two logical network functions – the radio unit (RU) function and the radio network area (RNA) function. The complex control of resources across radio sites and spectrum is therefore contained within the RNA. This architecture

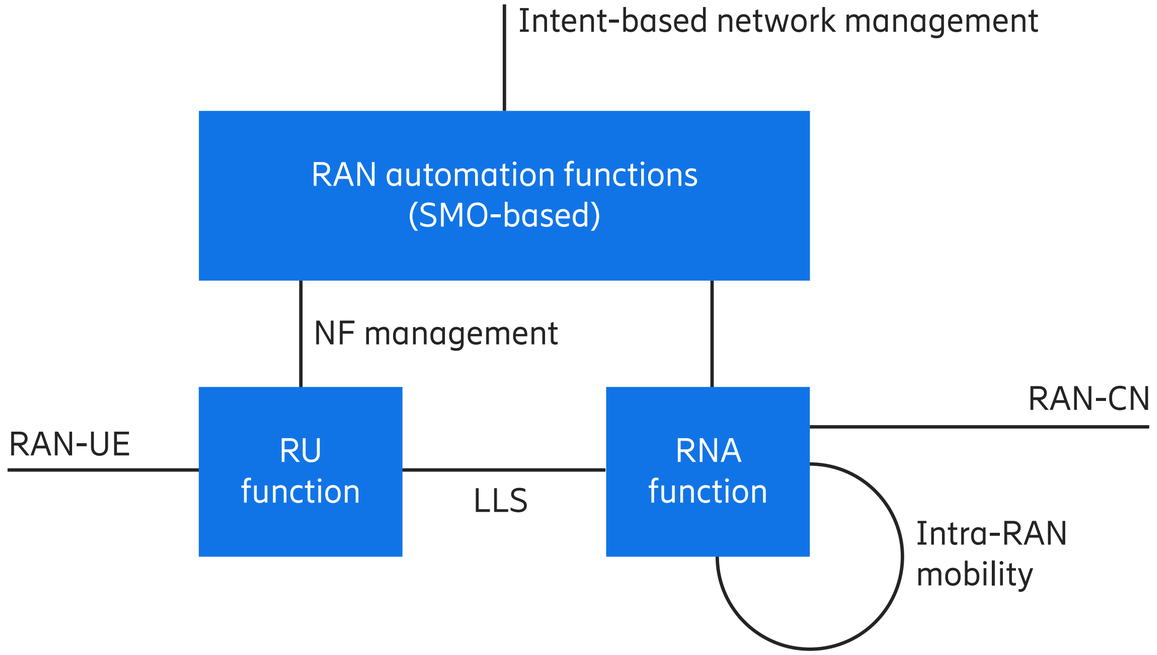


Figure 3: 6G RAN architecture recommended by Ericsson

opens the door for a RAN to optimize across a geographic area by connecting multiple radio sites with LLS to one RNA function, while maintaining the LLS as a key standardized interface.

Another key interface is an intra-RAN mobility interface between two RNA instances. The focus here is on the transfer of the UE control from a source RNA to a target RNA, and not on a complex interface to share control for the same UE (for dual connectivity, for example).

From a management point of view, the 6G RAN will evolve to be increasingly automated and expose a standardized intent-based management interface through the service management and orchestration (SMO)-based RAN automation functions [5]. Automation will occur both in the NFs themselves and in RAN automation functions (rApps) based on evolving AI/ML technologies. There will also be standardized NF management between the RAN automation functions and the RU function and RNA functions respectively.

Evolving the 5G Core to support the 6G radio-access network and new use cases

The strongest argument for building on the industry's previous investments in the 5GC network, rather than starting from scratch, is that it will enable a smoother introduction of 6G and help position CSPs to monetize on it from day one. The most important 5GC features to build on and evolve into 6G include the granular quality-of-service framework, comprehensive support of network slicing and support for time-sensitive and reliable communication, as well as the exposure of these network capabilities to address new service opportunities.

The extensibility of the 5GC architecture proves it can evolve to support a new 6G RAT.

The evolution of the 5GC also includes possibilities for optimizations and simplification of the SBA by, for example, reducing the dependencies across NFs or removing unnecessary flexibility to make the standardized architecture future proof. Reducing the pace of increase in the number of NFs and SBIs by bundling new features and functions with existing NFs (based on the main consumer of the function, for example) is an idea worth considering.

The extensibility of the 5GC architecture proves it can evolve to support a new 6G RAT. However, the 6G use case requirements will make it necessary to evolve some existing functionality and/or introduce new functionality into the CN. While 5GC is expected to evolve with NFs that are common to 5G and 6G, 6G-only NFs cannot be excluded.

Other aspects that need further exploration include the area of energy efficiency in CN deployments. This can partly be addressed in the functional architecture to support new RAN efficiency methods. But the major opportunity is to enable power savings in the implementation architecture and the underlying cloud infrastructure, rather than in the 3GPP standardized architecture.

A cloud-native approach with automation and artificial intelligence/machine learning from start

The 6G network should be designed to enable it to take advantage of the operational benefits that cloud-native design and deployment and network automation provide. This is particularly relevant with respect to the implementation domain and the adoption of relevant automation tools in operations of the network. At Ericsson, we believe that 6G standardization should focus on standardizing the functionality and interfaces needed for multi-vendor interoperability and avoid standardizing functionality that can better be handled in implementation and deployment.

At the same time, we think that the functional architecture standardized by the 3GPP should facilitate the cloud-native and automated deployment model where it makes sense. This includes considering functional dependencies of the underlying cloud infrastructure capabilities at a level that is high enough to allow the underlying technologies to evolve more independently to make it possible to benefit from the rapid evolution of cloud-native technologies. Other aspects include considering data pipeline and management that is more integrated into the architecture shown in Figure 2 and further defined in the AI-native definition [9]. This enables and enhances support of continuous service assurance monitoring, to handle the growing complexity in managing and optimizing the network automation transition toward zero-touch network operation.

Conclusion

At Ericsson, we believe that the whole telecommunications industry will benefit from early alignment on the key architectural principles of the future 6G network architecture. It is our position that the migration to 6G should be standardized as a single step, based on a new standalone 6G radio-access technology that is available on all needed spectrum bands and connected to a core network that is based on an evolution of the 5G Core network. This single step will ensure a strong industry focus and build on existing investments, while evolving the network capabilities needed for 6G.

We strongly recommend that 6G standardization focus on interfaces, network functions and services that are relevant for multi-vendor deployments to provide openness where it matters and enable increased focus in the standardization process. In addition, 6G solutions must be able to take advantage of the rapid evolution that is occurring outside of mobile industry standardization in areas such as cloud-native, artificial intelligence, machine learning, automation and related technologies, which should be included in the 6G system as a part of the implementation architecture and associated operational processes.

We are confident that early agreement on these principles will lead to both simplification and focus for 6G introduction, and ultimately enable early monetization by communication service providers.

The authors

Torbjörn Cagenius is a senior expert in network architecture at Business Area Cloud Software and Services. He joined Ericsson in 1990 and has worked in a variety of technology areas such as fiber-to-the-home, main-remote radio base stations, fixed-mobile convergence, IPTV and network architecture evolution. In his current role, he focuses on 5G and network architecture evolution toward 6G. Cagenius holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Gunnar Mildh is a senior expert in radio network architecture at Research Area Networks. He joined Ericsson in 2000 and has since been working on standardization and concept development for GSM/EDGE, HSPA, LTE, 5G NR and 6G. His focus areas are radio network architecture and protocols. Mildh received his M.Sc. in electrical engineering from KTH Royal Institute of Technology.

Göran Rune is a senior expert in core network architecture at Research Area Networks. He joined Ericsson in 1989 and has worked with standardization, systems design and research for digital cellular standards since 2G, starting with RAN and, while working on 4G, moving on to work with the core network. In his current role, he focuses on long-term core network architecture evolution, including 6G. Rune holds a Lic.Eng. in solid state physics and an M.Sc. in applied physics and electrical engineering, both from the Institute of Technology at Linköping University, Sweden.

Jari Vikberg is a senior expert in network architecture and the chief network architect at CTO office. He joined Ericsson in 1993 and has both broad and deep technology competence covering network architectures for all generations of radio-access and Packet Core networks. Vikberg holds an M.Sc. in computer science from the University of Helsinki, Finland.

Mattias Wahlqvist is a RAN architecture program manager at Business Area Networks. He joined Ericsson in 1998 and has since been working in standardization and concept development for 3G to 6G RAN systems. In his current role, he leads the business area work on long-term RAN architecture evolution, including 6G. Wahlqvist holds a Lic.Eng. in signal processing and an M.Sc. in computer science, both from Luleå University of Technology, Sweden.

Per Willars is a senior expert in radio network functionality and E2E architecture at Business Area Networks. He joined Ericsson in 1991 and has since worked mainly with RAN issues, but also mobile system architecture and service exposure. In his current role, he focuses on long-term RAN architecture evolution, including 6G. Willars holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology.

References

1. Ericsson white paper, 6G – Connecting a cyber-physical world, February 2022 ↗
2. NGMN, 6G Requirements and Design Considerations, February 14, 2023 ↗
3. Ericsson, Future Network Architecture, April 4, 2023 ↗
4. Ericsson, 5G monetization to improve top line revenue capture ↗
5. Ericsson Technology Review, Autonomous networks with multi-layer, intent-based operation, August 31, 2023, Niemöller, J; Silvander, J; Stjernholm, P; Angelin, L; Eriksson, U ↗
6. Ericsson Technology Review, Simplifying the 5G ecosystem by reducing architecture options, November 30, 2018, Cagenius, T; Ryde, A; Vikberg, J; Willars, P ↗
7. Ericsson white paper, Indirect communication for service-based architecture in 5G core, August 2021 ↗
8. Next G Alliance, ATIS NGA, 6G Technologies (pdf), June 2022 ↗
9. Ericsson white paper, Defining AI native: A key enabler for advanced intelligent telecom networks, February 2023 ↗

Further reading

- Ericsson, What is 6G? ↗
- Ericsson, Introduction to 6G ↗
- Ericsson white paper, A research outlook towards 6G ↗
- Ericsson blog, 6G spectrum: Why it's fundamental ↗

Co-packaged optics opportunities in radio-access networks

Authors:
Fabio Cavaliere, Antonio Tartaglia, Agneta Ljungbro, Alessandra Bigongiari, Stephane Lessard, Luca Giorgi, Anna Tavemark, Ulf Parkholm, Alfredo Palagi, Stefano Stracca, Antonio D'Errico, Nicke Svee

Co-packaged optics is an emerging technology with the potential to play a key role in 6G radio-access networks, due to its ability to enable high capacity at low energy consumption. Creating a version of the technology that is suitable for radio applications will, however, require some dedicated development to address network characteristics.

It is becoming increasingly difficult for communication service providers to use traditional copper interconnects to achieve sufficient bandwidth and distance in 5G radio-access networks (RANs) in an energy-efficient manner. The massive traffic growth expected with the introduction of 6G – estimated to be as much as 360 exabytes per month by 2027 [1] – will create an even greater need for high interconnect bandwidth with minimal energy consumption in 6G RANs.

The ability to enable high capacity with low energy consumption in RANs requires radical innovation in areas including integrated circuits (IC) architecture, digital signal processing, optical communications and packaging. Originally developed for use in data centers, co-packaged optics (CPO) technology makes it possible to improve both capacity and energy efficiency by integrating optics and silicon into a single packaged component. Widely recognized as a key enabler of future-proof cloud infrastructure, CPO also has great potential for use in 6G RANs. Much of the work around CPO that has been done for data centers can be reused, but some key aspects of the RAN require dedicated developments, including site cabling and the ability to operate outdoors in extreme temperatures.

Optics definitions and standards

Figure 1 provides a comparison between CPO and three other approaches to optical integration: on-board optics (OBO), near-packaged optics (NPO) and small form-factor pluggable (SFP) optics.

In CPO, at the top, an optical transceiver (TRX) is integrated into the same package as the IC. In this arrangement, the integrated optical TRX – also referred to as a chiplet – is placed directly onto the IC substrate and all electrical high-speed signals occur within the package substrate. Fiber attachment must be done from the package, which places high requirements on tolerances, manufacturability and compatibility with surface-mount technology. Fiber or waveguide alignments can be done from the reverse side of the package or with a fiber attachment on top of it.

In OBO – also known as board-mounted optics – an optical TRX module is placed on the same printed circuit board (PCB) as the packaged IC. High-speed electrical signals from the IC package to the PCB and up to the optical TRX package are transmitted by solder balls.

In NPO, the same type of optical TRX module used in OBO is placed on an extra interposer or substrate together with the packaged IC. The two components can be tested and

Terms and abbreviations

AOC – Active Optical Cable | **ASIC** – Application-Specific Integrated Circuit | **BER** – Bit Error Rate | **BGA** – Ball Grid Array | **CPO** – Co-Packaged Optics | **CRAN** – Centralized Radio-Access Network | **DAC** – Direct Attach Copper | **DRAN** – Distributed Radio-Access Network | **DU** – Distributed Unit | **ELS** – External Laser Source | **ELSFP** – External Laser Small Form-Factor Pluggable | **FEC** – Forward Error Correction | **I-Temp** – Industrial Temperature Range | **IA** – Implementation Agreement | **IC** – Integrated Circuits | **MCM** – Multi-Chip Module | **ns** – nanosecond | **NPO** – Near-Packaged Optics | **OBO** – On-Board Optics | **OIF** – Optical Internetworking Forum | **PCB** – Printed Circuit Board | **PMF** – Polarization Maintaining Fiber | **pJ** – picojoule | **RAN** – Radio-Access Network | **RRU** – Remote Radio Unit | **SFP** – Small Form-factor Pluggable | **TRX** – Transceiver | **XSR** – Extra-Short Reach

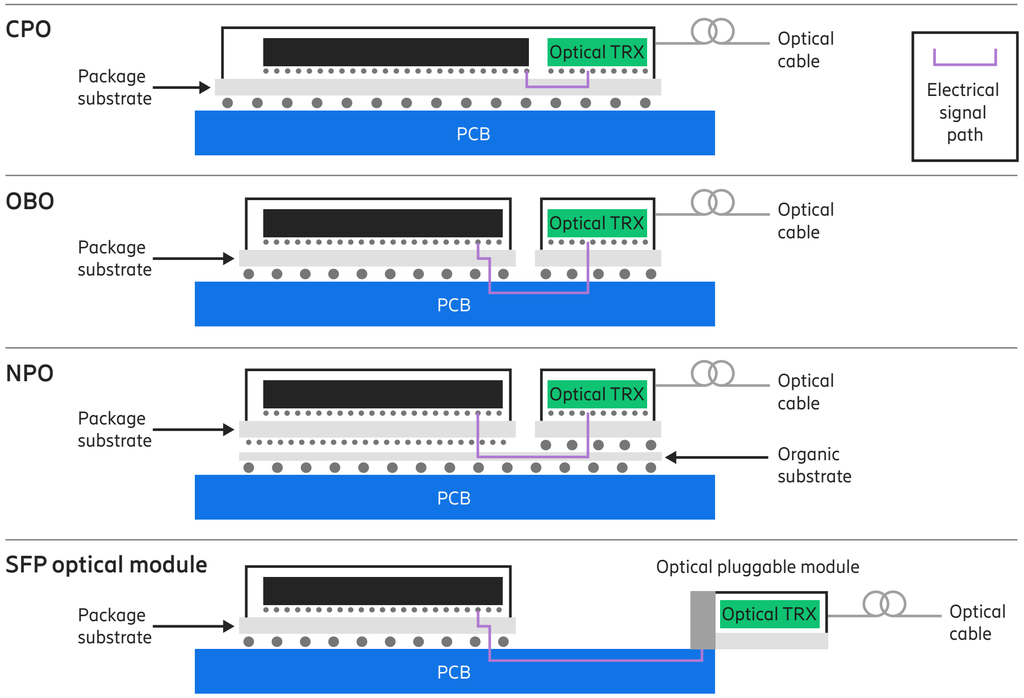


Figure 1: Four approaches to integrating optics

packaged separately. High-speed electrical signals are transmitted from the IC package to the common substrate using solder balls. As in CPO, fiber attachment must be done from the package.

SFP, which is also known as pluggable optics, refers to in-field optical TRX modules that can be plugged into a cage on the edge of the PCB, which is used to transmit high-speed electrical signals. Because SFPs can be easily added and removed, they make it easy to flexibly scale the capacity of the equipment and replace failed units. The downside of using SFPs is that they have a larger footprint than the other options in Figure 1, as well as consuming more power.

Of all the available options, CPO is the most beneficial in terms of minimizing both footprint and power consumption. It should be noted, however, that many of the considerations described in this article for CPO also apply to NPO and OBO – particularly those that pertain to the electrical interface and the external laser sources.

The optics requirements of RANs

The Optical Internetworking Forum (OIF) is developing Implementation Agreements (IAs) for the definition of the application space and the specifications of multi-vendor interoperable CPO modules [2]. The 3.2T Co-Packaged Optical Module IA defines interoperability requirements for a 3.2 terabit/s CPO module utilizing 100G electrical lanes. Both internal and external laser source (ELS) configurations are supported. The module is compatible with the IEEE (Institute of Electrical and Electronics Engineers) 802.3 Ethernet standards 400GBASE-DR4 and 400GBASE-FR4 to enable interoperation with pluggable optical modules, thus facilitating a broader market. The External Laser Small Form-Factor Pluggable (ELSFP) IA defines an ELS form factor for CPO modules. The ELSFP is a front panel pluggable module with a "blind mate" optical connector located at the rear of the module for eye-safety requirements.

A key RAN requirement is for the optical TRXs to have long life spans that cover the full duration of the network service.

Although most of the specifications developed by the OIF for CPO in data centers also apply to RAN, there are some additional requirements to address. The table in **Figure 2** compares the high-level optics requirements for data centers, distributed RAN (DRAN) fronthaul, centralized RAN (CRAN) fronthaul and backhaul [3].

A key RAN requirement is for the optical TRXs to have long life spans that cover the full duration of the network service (typically 15 years) to avoid costly visits to remote sites and tower climbs to do repairs. This requirement is especially important for CPO, as there is no option to perform the kind of field replacements that are possible with SFP optical TRXs.

Another important RAN requirement is the ability for the optical TRXs to tolerate the broad range of environmental conditions at outdoor antenna locations. When antennas are located outdoors on high towers, using heavy heat sinks or other advanced heat dissipation mechanisms such as liquid cooling is problematic. These locations require low-power optical devices, including lasers at the transmitter that can operate at high temperatures that may reach close to 100°C in the most extreme conditions.

Co-packaged optics use cases in RANs

Figure 3 presents the three main use cases for CPO in the RAN: site connectivity, intra-site connectivity and on-board connectivity.

Use case #1: site connectivity

The site-connectivity use case covers the domain of today's pluggable optics – that is, optical cabling at radio sites. A link distance of 500m fits most DRAN scenarios, with the distributed unit (DU) placed in an enclosure close to the tower. In a CRAN, this use case applies to the site cabling between the remote radio units (RRUs) and the antenna site switch. The longer distances (15-20km) to the central office where the DUs are located will still be covered by pluggable optics hosted in the antenna site switch. As the CPO in the RRUs may have to interoperate with pluggable optics used in the switch, compliance with the applicable Ethernet optical standard is mandatory.

	Data center (intra-office)	Fronthaul (DRAN)	Fronthaul (CRAN)	Backhaul
Fiber distance	≤ 2km	≤ 2km (98% ≤ 500m)	≤ 15km (20km in extreme cases)	≤ 40km (80km in extreme cases)
Fiber supply	Abundant	Abundant	Scarce	Scarce
Data rates	100G, 400G	25G, 50G,100G+	25G, 50G, 100G+	25G, 100G
Sync	Not critical	Very critical	Very critical	Critical
Lifetime	3-5 years	15 years	15 years	10-15 years
Environment	Indoor, temperature controlled (commercial temperature range)	Outdoor, industrial temperature range (I-temp)	Outdoor, I-temp	Outdoor, I-temp
Service cost	Low	Very high	Very high	High
Deployment model	Greenfield	Brownfield	Brownfield	Brownfield

Figure 2: High-level optics requirements for data centers versus RANs

The targeted energy consumption of 10pJ/bit in Figure 3 is half that of the current best-of-breed generation of pluggable optics. It is quite challenging to achieve this with a retimed electrical interface, but it may be possible with a linear electrical interface. Depending on the loss budget and the equalizers in the application-specific integrated circuit (ASIC), the CPO attachment on the substrate will use a socket connector or ball grid array (BGA) soldering.

There is an ongoing debate in the CPO developer community about external versus internal laser sources. From the RAN perspective, ELS solutions are much better positioned, as the main driver we see for CPO adoption in radio units is removing optics from designs with critical thermal issues. Radio units are trending toward smaller sizes at feature parity, and the power consumption of several components is not reducing at the same pace as the unit volume. The power consumption of pluggable optics is actually increasing with the bit rate, due to the re-timers required to drive the long chip-to-module electrical channels. By making it possible to disaggregate the laser and ensure its reliability by putting it in a location with a milder environment, CPO modules can survive at junction temperatures similar to those of digital ASICs (105°C-110°C).

Use case #2: intra-site connectivity

The intra-site connectivity use case covers the domain of today's Direct Attach Copper (DAC) and Active Optical Cables (AOCs). DAC and AOCs are used for optical cabling between the units in a telecom rack or inside its radio outdoor equivalent (the rail mounting systems) where new 100G or 200G per-lane DAC generations cannot support the targeted distance of 3m shown in Figure 3. For this use case, compliance with the Ethernet optical standards is not strictly necessary, which opens up opportunities to further optimize energy consumption using novel technology approaches. The focus on energy efficiency in this case also requires the use of "linear" electrical interfaces between the ASIC and the CPO module.

Use case #3: on-board connectivity

The on-board connectivity use case covers the domain of today's high-speed copper interconnects such as PCB tracks and flyover cables. As the signaling speed goes up, it is becoming increasingly expensive to enforce signal integrity for high-speed copper interconnects, both in terms of high-frequency PCB materials and energy spent in the ASICs to compensate for impairments. The targeted distance of 2m in Figure 3 originates from the needs of massive MIMO (multiple-input, multiple-output) radio units with integrated antennas that can be large, particularly those in the sub-6GHz range. The nature of the interconnect is proprietary, and all the metrics need to be more like today's copper links than today's pluggable optics.

Implications of co-packaged optics on RAN site deployment

To minimize the potential disruptive impact of CPO introduction in the site-building process, some additional RAN-specific requirements must be addressed at macro mobile sites and street sites.

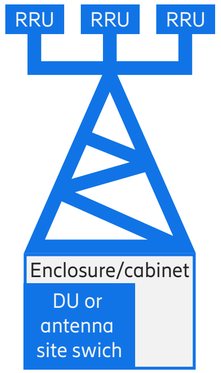
Using co-packaged optics at macro mobile sites

A macro site can be schematized as an antenna tower with radio units mounted on it and a ground-level cabinet to host traffic units plus all the necessary site-support functions including battery packs, power converters and site controllers. The cabinet can generally be regarded as an indoor environment. The traffic units inside the cabinet can be DUs (in DRAN deployments) or transport nodes such as antenna site switches (in CRAN deployments). The radio units on the antenna tower must be able to survive in the harsh outdoor environment of the mast. The short reach (up to 500m) optical connectivity between the traffic node and the radio units can be served by CPO.

The question of where to place the ELS feeding the radio units is important.

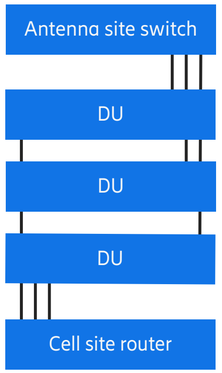
While the traffic unit could use an ELSP placed on the front panel of the indoor unit with a blind mate optical connector, the question of where to place the ELS feeding the radio units is important. The presence of a cabinet makes it possible to integrate an ELS into a dedicated box, placed within the enclosure and operating indoors. This solution allows the reuse of ELSFP defined for datacom use and avoids expensive tower climbs. The typical distance between the cabinet and the radio units is less than 500m,

Site connectivity



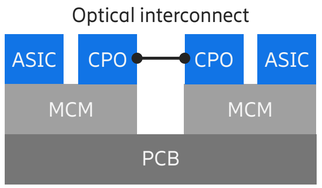
Distance <500m
BER<10^{-15} post FEC
Latency<200ns
Latency determinism <0.2ns
Energy consumption 10pJ/bit

Intra-site connectivity



Distance <3m
BER<10^{-15} post FEC
Latency<200ns
Latency determinism <0.05ns
Energy consumption <6pJ/bit

On-board connectivity



Distance <2m
BER<10^{-15} (possibly without FEC)
Latency<10ns
Latency determinism <0.05ns
Energy consumption <3pJ/bit

Figure 3: The three main use cases for CPO in the RAN

however, and such long strands of polarization-maintaining fiber (PMF) are expensive. Long strands of PMF also raise concerns regarding the attainable polarization extinction ratio at the modulator input. Further, the use of PMF makes it necessary to introduce new materials to current site deployment models.

The use of long strands of PMF can be avoided by placing the ELS that is feeding the radios in the enclosure of a macro site. Several approaches have been proposed for this and tested in experiments, with the "polarization-agnostic laser source" alternative [4] showing particular promise. Moving away from the blind-mate optical connector paradigm also requires careful consideration of eye-safety aspects: depending on the optical output power levels, additional circuitry may be needed in the ELSFP to detect fiber breaks and shut down the laser sources quickly enough to keep the hazard level below 1m, as defined in IEC (International Electrotechnical Commission) 60825-2. If ELSs serving the radios could be integrated directly onto the front panel of radio units with blind mate optical connectors, there would be no changes to the deployment model. However, the challenge would then shift to the ELSFP, which must operate at a high case temperature.

Using co-packaged optics at street sites

A street site is leaner than a macro site – in many cases there are no cabinets or supporting infrastructure at all. For example, a street site can be as simple as a lighting pole for which permission has been granted to mount networking equipment. The use case for CPO at a street site is when a traffic node – a DU for a DRAN or a small antenna switch for a CRAN – is deployed together with the radio units. In these cases, CPO could be used to deliver short reach optical connectivity between the traffic node and the radio units.

ELSs improve system reliability by addressing issues of serviceability and thermal stress.

The absence of site-supporting functions removes the possibility of integrating the ELS into a dedicated box and operating in an indoor environment, but the ELS serving the radios could be integrated into dedicated ports of the traffic node with short additional PMF strands of just a few meters. Alternatively, the ELSs could be plugged into the front panel of radio units with blind mate optical connectors. It is, however, not possible to use ELSFPs designed for datacom in either case. Overcoming this challenge requires a careful redefinition of ELSFP output optical power and reliability figures in relation to the case operating temperature.

Enabling technologies

Two technologies are essential to enable CPO-based RAN deployments: ELSs that operate over standard optical fibers and energy-efficient electrical interfaces.

External laser sources

ELSs improve overall system reliability by addressing the issues of serviceability and thermal stress. This is especially beneficial in hardware-dense boards, where hot spots can easily reach temperatures higher than 100°C. ELSs make it possible to replace failed lasers an unlimited number of times. They also have a longer lifetime than internal laser sources, as they can be placed in a more favorable environment than the CPO module.

There are some restrictions on the use of ELSs, however, due to eye safety requirements that limit maximum power transmission and light polarization. In a system with a single ELS feeding eight CPO modules, for example, the output power of the laser is 21-23dBm with typical insertion loss figures on the optical transmit and receive paths, which is close to the eye safety limits defined by the IEC 60825-2 standard. Solutions can be mechanical (blind mate connectors) or software based (automatic power shutdown mechanisms). Alternatively, it is possible to consider solutions with multiple low-power sources. The example in **Figure 4** shows the case of eight ELSs: as the insertion loss is 8-11dB in this scenario, the required laser power for each ELS lowers to 11-13dBm.

The issue of the polarization arises from the strong polarization dependency of the modulators in the CPO

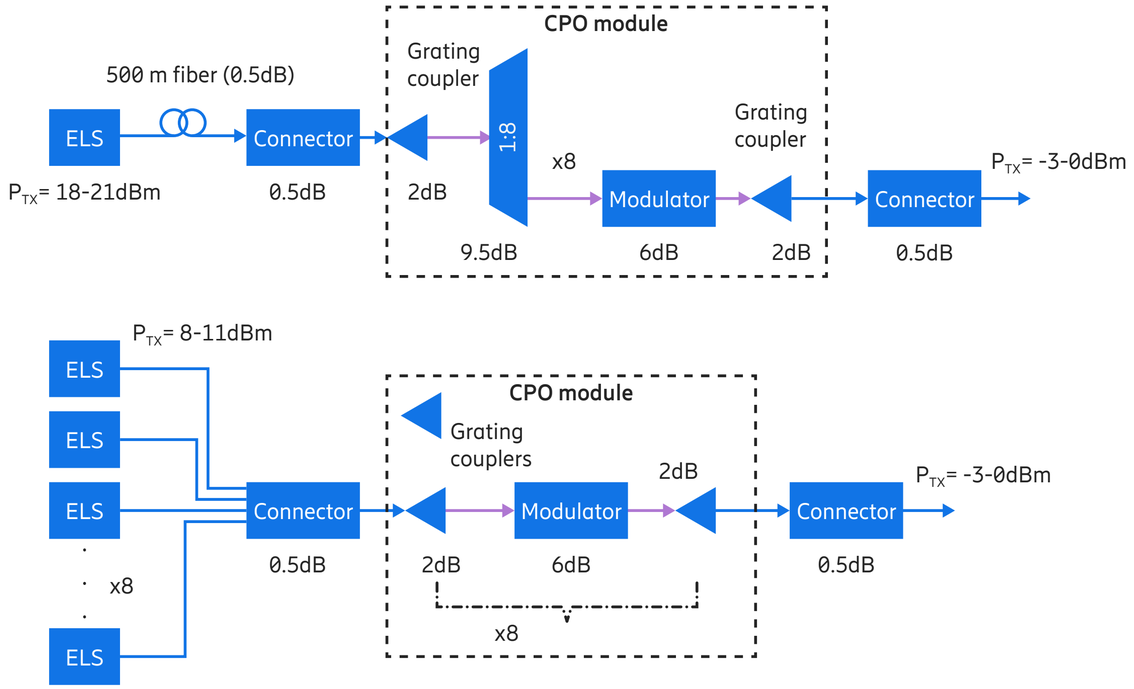


Figure 4: CPO example with eight ELSs

module. When the laser is separated from the module, the polarization may change along the fiber and lead to power fading. While current ELSs often use PMF to connect equipment located in the same site (in a lab, for example), it is less suitable for use over longer distances for cost and performance reasons.

For longer links, beyond 100m, alternative solutions such as polarization controllers and polarization-agnostic laser sources must be considered. Polarization controllers rotate the polarization based on feedback control and need an "endless algorithm" that avoids gaps in the control when the control variable reaches its practical bounds. Polarization-agnostic laser sources do not require complex feedback control; they deliver constant power output at the transmitter by assuring, on average, constant power on any polarization axis.

Experiments have demonstrated that it is possible to combine two laser sources at different wavelengths in such a way that their polarization is orthogonal and any rotation in the polarization that occurs along the fiber does not change the power available at each polarization axis [4]. More recently, novel approaches based on on-chip LEDs (light-emitting diodes) or quantum dot lasers have been proposed as thermally robust internal laser source alternatives to ELSs. However, while they do show promise, these technologies are at an early stage of development.

Energy-efficient electrical interfaces

CPO is expected to provide significant power savings compared with equivalent-capacity electrical interconnects because the placement of optical chiplets close to a digital ASIC minimizes the loss and the impedance discontinuity of the electrical interface between the two components. CPO may rely on several electrical interface standards such as those specified by the OIF or on parallel interfaces such as Advanced Interface Bus or Universal Chiplet Interconnect express. The OIF CPO IA [2] uses the CEI (Common Electrical Interface) 112G-XSR-PAM4 Extra Short-Reach (XSR) Interface, which is a digital retimed electrical interface.

Another electrical interface under consideration is a linear amplified interface, where the electrical re-timer function in the chiplet is removed to reduce the optical TRX power consumption. To avoid distorting the signal, this approach requires a linear modulator driver. However, with a linear interface, the ASIC equalizer must compensate for the propagation impairments introduced by both the electrical and the optical links, which are no longer decoupled by the re-timer. This results either in more complex equalizers, eroding the gain in energy efficiency, or in a shorter distance, keeping the same equalizer. Finding the best compromise between optical TRX power consumption and ASIC power consumption is a challenge for hardware designers.

Based on a survey we carried out among several CPO vendors, a linear interface results in power savings of about 50 percent in the CPO module and about 30 percent at system level, including the ASIC. Considering that the distance between the ASIC and the optical engine spans from about 50mm with XSR to 5mm with a direct drive interface [2], we estimate that there is enough distance to allow for disaggregation of ASIC and CPO, thereby alleviating the problem of hot spots.

Conclusion

Most of the technologies developed for co-packaged optics (CPO) in data centers have strong reuse potential in radio-access networks (RANs) because they are based on cost-effective silicon photonics, meet Ethernet standards and use external laser sources. This makes it possible to leverage on an already established ecosystem, with obvious benefits in terms of costs. However, in comparison to data centers, RANs have stricter requirements in terms of operating temperature, power consumption and site-building practices that make the use of polarization-insensitive ELSs and energy-efficient linear electrical interfaces mandatory. Moreover, most of the current CPO implementations are proprietary, creating a significant barrier to the large-scale deployment of CPO in RANs and making the definition of standardized CPO solutions for RANs an urgent need.

The authors

Fabio Cavaliere is an expert in photonic systems and technologies who joined Ericsson in 2005. He also serves as Rapporteur of ITU-T (The International Telecommunication Union Telecommunication Standardization Sector) Question 6/15 (optical transport systems) and is the author of more than 130 filed patent applications, more than 100 publications on optical networks and the book Photonics Applications for Radio Systems and Networks. Cavaliere holds a M.Sc. in telecommunications engineering from the University of Pisa in Italy.

Antonio Tartaglia joined Ericsson in 2006 He is a system manager and expert in photonics, focusing on optical solutions for RANs and RAN transport networks. Tartaglia has worked with optics in a variety of roles, from production engineering to hardware and optical systems design. He holds an M.Sc. in electronics engineering from the University of Naples Federico II in Italy.

Agneta Ljungbro works as a senior specialist in electronic packaging within Business Area Networks, focusing on heterogenous packaging technologies and related interconnect technologies. Since joining Ericsson in 1993 she has played a key role in the introduction of fine pitch BGA technology for mobile phones and µ-via technology for PCB. Ljungbro holds an M.Sc. in applied physics and electrical engineering from Linköping University in Sweden.

Alessandra Bigongiari joined Ericsson in 2017. She is a senior researcher with a background in material science whose work at Ericsson primarily focuses on integrated photonics and optical technology. Bigongiari is the author of more than 20 filed patent applications and more than 30 publications in scientific journals. She holds a Ph.D. in physics from Ecole Polytechnique in Paris, France.

Stephane Lessard is a senior specialist in photonic system architecture who joined Ericsson in 2007. His work focuses on the use of photonic technologies to revolutionize system architecture and interconnections. He has more than 15 patents in the field of photonics and has co-authored more than 30 journal and conference articles. Lessard holds an M.Sc. in theoretical physics from the Université de Sherbrooke in Canada.

Luca Giorgi is a master researcher who is responsible for Ericsson's optical transmission laboratory. Since joining Ericsson in 2005, his research activity has encompassed RANs, fiber access, high-speed optical transmission and integrated photonics. Giorgi is the author of more than 50 filed patent applications and more than 30 publications on optical networks. He holds a M.Sc. in telecommunications engineering from the University of Pisa.

Anna Tavemark joined Ericsson in 1995 and works as a technical coordinator, focusing on the requirements and control of optical pluggable SFPs. In addition, she also serves as a predevelopment leader or team member for evaluations of and research studies into new optical solutions. Tavemark holds an M.Sc. in electrophysical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Ulf Parkholm joined Ericsson in 2009 and is the lead architect within Ericsson's Silicon Ethernet ASIC Intellectual Property development group. His work focuses on Ethernet and time synchronization system solutions for mobile fronthaul applications. He has served as a standardization delegate for the Ethernet Technology in IEEE 802.3 and IEEE 802.1 working groups since 2020. During his years at Ericsson, Parkholm has filed more than 10 patent applications. He holds an M.Sc. in electrical engineering from Uppsala University, Sweden.

The authors (continued)

Alfredo Palagi joined Ericsson in 2006 and he is a senior specialist in optical technologies for RAN and part of the PEU Transport Optical Solutions and Fronthaul System design team. He has spent more than 25 years in the telecom industry with a focus on optical transport, holding different roles in hardware and system design, as well as in customer support. Palagi holds an M.Sc. in electronics engineering from the University of Pisa.

Stefano Stracca has worked at Ericsson since 1990 and is currently a master researcher with extensive experience in telecommunication networks and nodes, digital system engineering and digital integrated circuits design. He is the author of 17 patent filings and several scientific publications. Stracca holds an M.Eng. in electronics engineering from Sapienza Università di Roma in Italy.

Antonio D'Errico is a senior specialist in optical communications, fiber networks and integrated photonic technologies who joined Ericsson in 2009. He is the author of about 50 filed patent applications, more than 100 publications on optical networks and the book Photonics Applications for Radio Systems and Networks. D'Errico is currently researching photonic applications for radio systems and networks toward 6G. He holds a Ph.D. with honors in optical telecommunication systems from Scuola Superiore Sant'Anna in Pisa, Italy.

Nicke Svee is an expert in high-speed interface technologies. He joined Ericsson Research in 1999 and has been working on high-speed serial channels, precision phase locked loops, system-level noise budgeting, signal integrity and timing. He is currently responsible for predevelopment activities targeting high-speed serial interfaces and timing and synchronization. Since 2022, he has been Ericsson's delegate in the OIF. He holds a B.Sc. in electrical engineering from KTH Royal Institute of Technology.

References

1. Ericsson Mobility Report, June 2023 ↗
2. OIF, Co-Packaging Framework Document, February 3, 2022 ↗
3. Ericsson white paper, Optimized Optical Solutions – small form pluggable, March 2022 ↗
4. Electronics Letters, Experimental evaluation of silicon photonics transceiver operating at 120 °C for 5G antenna array systems, vol. 54, no. 24, pp. 1391–1393, Oct. 2018, Testa, F; Giorgi, L; Bigongiari, A and Bianchi, A ↗

Further reading

- Ericsson, MOPA: Pluggable optics solutions to support 5G rollouts ↗
- Ericsson, Photonic applications for radio systems and networks ↗
- Ericsson blog, Photonic integration becomes a reality ↗
- IEEE, Journal of Optical Communications and Networking, Optical transport for Industry 4.0 [Invited], vol. 12, no. 8, pp. 264-276 ↗
- International Conference on Transparent Optical Networks 2023 (Bucharest, Romania), Perspectives for Co-Packaged Optics in Radio Access Networks ↗
- Ericsson, What is 6G? ↗
- Ericsson, 5G RAN ↗

rApps: Transforming network management with intelligent automation apps

Authors:
Ryan Fitzgerald, Ciaran Johnston

Intelligent automation using rApps has the potential to revolutionize the world of network management and orchestration, augmenting or replacing existing manual processes in order to improve network performance and reduce costs.

Open radio-access network (O-RAN) architecture is maturing, while the network management paradigm is shifting emphasis from vendor-specific, hands-on network management and operations toward an open ecosystem of intelligent automation provided by vendors and communication service providers (CSPs) alike.

Today, on top of having access to more network data than ever before, CSPs also have fine-grained controls to manipulate the network topology and a high degree of configuration flexibility. These powerful tools drive complexity and limit the CSP's ability to optimize performance and reduce operational costs without a significant increase in the use of automation focused on the CSP's specific business needs. A benefit of this type of increased automation in network management is the enabling of higher levels of abstraction and simplification through intents, supporting a more dynamic business with rapidly evolving goals.

The network automation landscape

Mobile networks are rapidly evolving toward increased levels of heterogeneity and sophistication. Equipment and technologies from different vendors must deliver better performing and more differentiated services than ever before. Cloud technologies and software-defined networks promise new opportunities for CSPs to utilize their commodity hardware and services to achieve their business goals. In addition, network services continue to become more central to the day-to-day workings of modern societies, and even transient service degradations are therefore becoming less acceptable. Banking, e-commerce, entertainment, transportation, logistics and emergency services are growing ever more reliant on guaranteed, high-quality network services. These realities, coupled with increased legislative, environmental and energy-efficiency considerations pose an increasing challenge for CSPs.

Higher levels of heterogeneity and sophistication in networks are leading to greater complexity. Infrastructure can change independently of the software running on it, which means that performance can change over time. A higher number of open interfaces and network functions creates a more

WHAT IS THE O-RAN ALLIANCE?

The O-RAN Alliance defines specifications in areas of radio-access network (RAN) automation, cloudification and disaggregation. The ambition of the O-RAN Alliance is to enable an open RAN by creating a multi-supplier RAN solution that allows for the separation – or disaggregation – of hardware and software with open interfaces and virtualization, hosting software that controls and updates networks in the cloud [1].

In O-RAN, the term rApp refers to an app that has been designed to work on the non-real-time RAN intelligent controller targeted toward the open RAN. The rApp concept can, however, be applied to other domains as well.

Terms and abbreviations

AI – Artificial Intelligence | **API** – Application Programming Interface | **CNF** – Cloud-Native Network Function | **CSP** – Communication Service Provider | **ML** – Machine Learning | **Non-RT RIC** – Non-Real-Time RAN Intelligent Controller | **O&M** – Operations and Maintenance | **O-RAN** – Open Radio-Access Network | **PNF** – Physical Network Function | **RAN** – Radio-Access Network | **rApp** – Non-RT RIC Application | **SMO** – Service Management and Orchestration

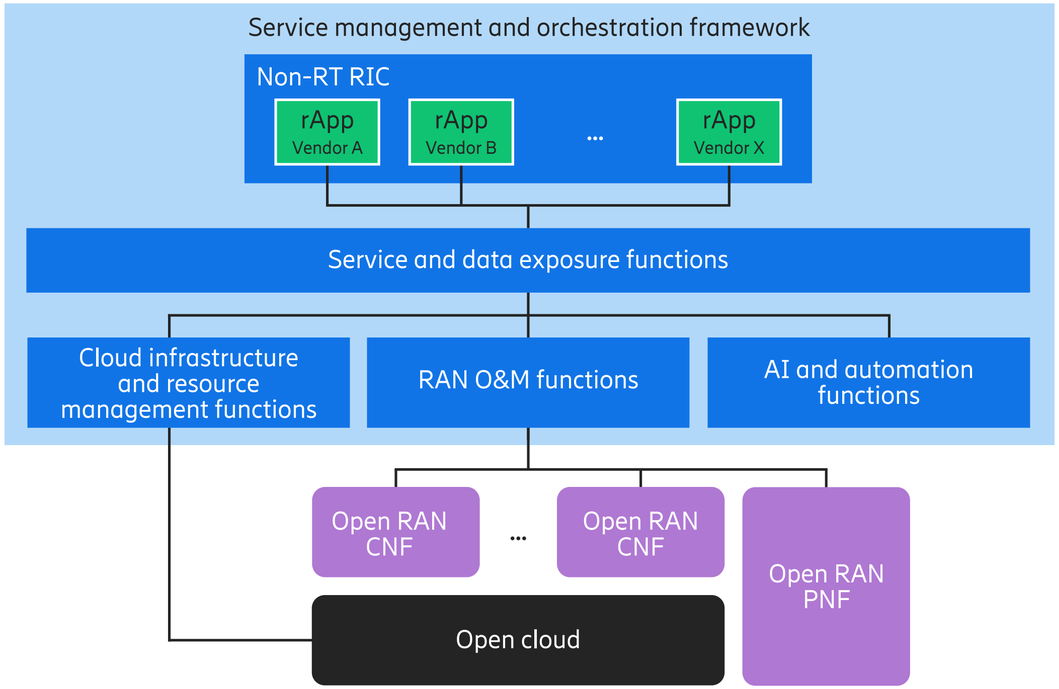


Figure 1: Simplified view of the SMO in the O-RAN context

complex network topology. These network functions are themselves comprised of many small microservices that can be deployed and scaled in different ways.

An increase in the number of network features needed to enable high-quality communication services results in more parameters for vendors and CSPs to configure based on radio network infrastructure performance. More cells and user equipment, together with an increase in dynamic weather conditions and extreme environments (military, industrial and disaster recovery) all lead to an increase in the dynamic nature of the network environment.

Intelligent automation will be required to manage all the complexity and meet expectations while maintaining or decreasing operating costs and capital expenditure. Traditional ways of managing network optimization through static configuration planning and human-centric rollout are no longer sufficient to deliver the network performance required to meet more demanding Service Level Agreements.

To date, the industry has partially addressed these challenges by introducing proprietary SON (self-organizing network) systems performing centralized and closed-loop assurance on top of traditional management systems, with some degree of success.

Intelligent automation will be required to manage all the complexity and meet expectations.

However, such systems are typically heavily reliant on extensive system integration, and expensive to maintain and evolve, as well as often being difficult to manage in coexistence with manual processes or scripted automation. This is especially true given that individual CSPs have different operational needs and are more or less willing to adopt varying levels of automation based on their trust in the technology and operational requirements. CSPs want to be in control of – and in many cases even the creators of – the automation running their network, without the responsibility of managing connectivity, security, conflicts and interoperability issues between vendors.

Modern IT industry best practices have defined a set of architectural patterns and principles to accelerate the development of advanced automation. Standardized security through the OAuth 2.0 protocol and related standards, common deployment patterns using containers, application programming interface (API)-first development using the RESTful (representational state transfer) web service and event-driven APIs with well-defined contracts can all be applied to the creation of well-specified automation enablers. Open-source development in the Cloud Native Computing Foundation, Open Network Automation Platform and O-RAN Software Community provides concrete working code to realize those functions.

rApps and the service management and orchestration framework

The O-RAN Alliance has defined the service management and orchestration (SMO) framework as a key function within the O-RAN architecture. The SMO offers capabilities to orchestrate the deployment of network functions into the open cloud infrastructure and perform fault, configuration, performance and security management for them, while the non-real-time RAN Intelligent Controller (non-RT RIC), a sub-function of the SMO, offers capabilities to enable the implementation of intelligent RAN automation and optimization use cases.

The O-RAN Community and its various working groups are currently developing and agreeing on the specifications that define the SMO in detail. These include the various interfaces and functions exposed within the SMO and the non-RT RIC. **Figure 1** shows a simplified view of the SMO architecture.

The O-RAN Alliance specifies the rApp as a modular application within the non-RT RIC that utilizes the capabilities of the SMO to realize value-adding RAN automation use cases. Because an rApp can be developed, delivered and life-cycled independently of the SMO, and because it utilizes open and standardized interfaces, it can be sourced from different software vendors and interwork with different SMO implementations.

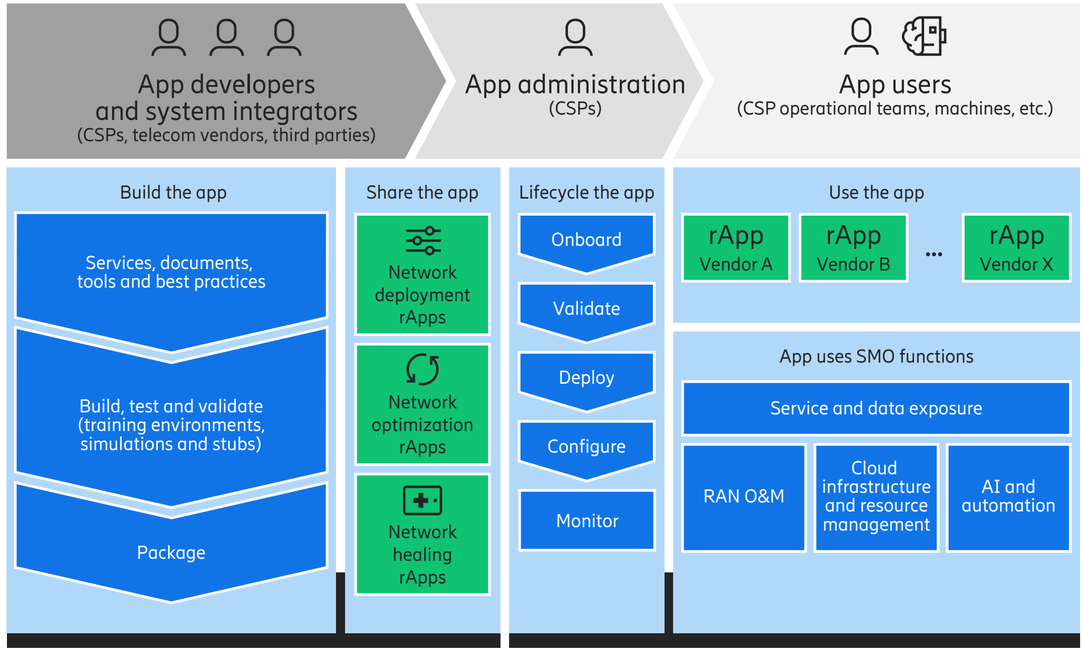


Figure 2: rApp development life cycle

The types of network automation functionality that rApps can implement is extensive. Advanced workload placement and deployment, software canary testing, automated network healing and outage compensation, network performance anomaly detection, configuration validation and optimization are all examples of the valuable automation use cases that have already been proposed.

To achieve these aspirations, rApps will depend on a rich set of SMO capabilities to observe and control the network. Key examples of these capabilities are network inventory and topology discovery, configuration querying and modification, performance and fault data exposure and analysis, and cloud infrastructure and resource management. Furthermore, to help accelerate rApps in delivering smart and intelligent automation use cases, the SMO is likely to provide additional advanced automation supports. Examples of these include artificial intelligence/machine learning (AI/ML), policy, workflow, and intent management frameworks.

The SMO, along with the rApp-based extensibility of the non-RT RIC, represents a paradigm shift in how network automation is achieved. Up until now, automation was built either in an ad-hoc fashion or as separate and costly add-ons to existing proprietary element management systems. Furthermore, automation was difficult to implement because it required specialized knowledge of proprietary interfaces and because each vendor's network equipment (and management infrastructure) presented different technical challenges to overcome.

With the SMO, and the open ecosystem around it, CSPs have much more power to define and control how their networks are managed and optimized. Automation use cases and automation support capabilities are no longer considered as add-ons but are instead a native, built-in capability of the network management system itself, benefiting from significantly better standardization of the observability and control interfaces to the network. Even in brownfield scenarios, existing equipment can be integrated into an SMO environment by applying a consistent set of management principles and adding adapters for the non-standard interfaces.

The SMO also represents a paradigm shift in network operations. As the automation realized by the rApps and the SMO becomes more advanced, the role of the CSP will transform from directly managing the network to managing the automation that enables and supports the network. The measure of success for a management system will shift from how well it assists the CSP in managing the network to how well it runs the network according to the goals and limits that the CSP sets.

While this approach to automation is being standardized by the O-RAN Alliance, it is not specific to the RAN domain; apps can also be built to deliver automation across all the network domains.

The automation development ecosystem

To fully realize the automation potential, it is essential that the rApp development life cycle – from ideation, implementation, onboarding and operation to eventual retirement – is as efficient as possible. The needs of rApp developers, rApp administrators and operations staff using rApps must all be considered. A high-level view of the rApp development life cycle is illustrated in **Figure 2**.

CSPs have much more power to define and control how their networks are managed and optimized.

A broad range of developer and rApp types must be supported. Some rApps will be produced by full-time software developers and have a relatively long lifespan with multiple feature additions and an extended maintenance period. In contrast, other rApps may be created by part-time or "citizen" developers with limited coding expertise, with a focus on automating a particular operational task. Furthermore, some rApps may be long-running and support scaling to handle large workloads, while others may be quite short-lived and only execute periodically to fulfill a particular task or job.

With such a variety of developer and rApp types, the appropriate developer, administrator and operations supports must be in place. This requires achieving a careful balance between enabling freedom and flexibility for the

advanced developer to create sophisticated rApps and providing ease of use and safe guardrails for the citizen developer, who needs to achieve fast results with a much simpler approach (a script, for example). In all cases, security must be of central concern to thwart nefarious actors who may wish to introduce malicious code into the network.

Developers therefore need a rich set of development resources, ranging from easy-to-follow documentation and tutorials to sample code, sandbox test environments and runnable examples. Build and verification pipelines, as well as secure packaging and a "marketplace" to publish, share and distribute them, are also necessary.

The transition from current processes and tools must be carried out in a stepwise manner.

Once rApps are published in the marketplace, administrators who want to utilize them need to have oversight over which of them are to be deployed in their systems, with clear visibility and control over their requirements and dependencies. Administrators need to be able to monitor and control how rApps use system APIs and resources, as well as to easily observe their performance and output.

Network operations staff need to be able to quickly discover, understand and build trust in the rApp automation applicable to their role. They need to know how to smoothly integrate it into their existing management practices and how to replace their existing practices with the rApp automation when necessary. Further, they need to be able to easily observe how automation impacts their network as well as a way to interact with the automation when required.

Automation and conflict management

As the level of rApp automation grows and more rApps coexist in the same environment, there is an increased risk that the automated network operations they perform will conflict with each other. For example, different rApps may attempt to adjust the cell or antenna configuration based on competing goals (such as throughput versus energy efficiency). While this is not a new problem, the flexibility of the SMO architecture makes this issue more important to address. The consequences can range from some rApps being rendered ineffectual, to race conditions and ping-ponging changes that impact network performance and stability.

Some rApps will be developed to coordinate and communicate with each other, thus avoiding such conflicts, but many will not. Those rApps that cannot coordinate and communicate with each other will rely on the SMO's automation capabilities to mitigate or prevent conflicts. In light of this, the SMO architecture should include three types of conflict management functionality:

- Coordination and control
- Conflict detection
- Conflict intervention and resolution.

The coordination and control capabilities will ensure that system administrators and operators can determine where two rApps may be consuming the same network data (the same performance management counters, for example) or modifying the same configuration (the same configuration management parameter, for example) in the same part or in related parts of the network (the same or adjacent nodes). Some of these capabilities may also be exposed to the rApps themselves, so that developers can code their rApp to make use of this data in runtime to avoid conflicts.

The conflict detection functionality will make it possible for the system to observe the control actions that the rApps propose and determine whether they will (or are likely to) result in a conflict. When a conflict is detected, the operator is alerted and given the opportunity to disable or reconfigure the rApps. Since what constitutes a conflict can be highly context-dependent and differ from one SMO deployment or CSP network to another, the decision-making capabilities of the conflict detection functionality must be highly adaptable. A policy-based approach is therefore essential, and AI/ML techniques show great promise in this area.

The conflict intervention and resolution functionality will enable the system to make decisions about whether requested control actions should be blocked or permitted. It can also be used to determine if restorative or repair actions are required. These decisions can be based on a broader range of factors including the impact of the conflict, the relative priorities of the conflicting rApps, operator preferences on guard periods for sensitive configuration management parameters, current network state and other criteria.

With these three conflict management functionality types in place, CSPs can have confidence in the smooth operation of the automation in their network as they build it out.

Transitioning to intelligent automation

It will take time for CSPs and vendors to fully realize the potential of an open ecosystem of rApp-based automation. The transition from current processes and tools must be carried out in a stepwise manner. Firstly, in defining the scope of an automation platform to run multi-vendor rApps, it is important to focus on a small number of prioritized APIs rather than trying to cover every need immediately. Vendors can build trust by delivering on a small set of important use cases and ensuring that the required capabilities are both robust and performant. Secondly, building up an extensive marketplace of rApps will take time and require strong engagements between standardization delegates in the O-RAN Alliance, open-source communities building reference capabilities, CSPs and network equipment vendors across the industry. Last but by no means least, a change management process to help network operations staff adjust to a higher degree of automation will be required.

Conclusion

The rApp approach to automation is a key component of the service management and orchestration architecture in open radio-access networks. Using rApps, communication service providers will be able to overcome a wide variety of network management challenges, delivering significant benefits in terms of operational cost, network service performance and resource utilization. Realizing the full vision will, however, require a deliberative and stepwise approach. Ericsson's comprehensive strategy for the use of rApps is based on our deep understanding of the challenges posed by heterogeneous automation use cases running simultaneously and on our years at the forefront of the development of an open automation platform architecture.

The authors

Ryan Fitzgerald joined Ericsson in 1998 and currently works as a master engineer focusing on operations support systems (OSS) and network automation. He is currently exploring how OSS and network management products can deliver enhanced automation for Ericsson customers. Fitzgerald holds a B.Eng. in computer engineering from the University of Limerick, Ireland, and an M.Sc. in software engineering from the Athlone Institute of Technology, Ireland.

Ciaran Johnston is a senior expert in OSS and programmable network architecture, and he is the chief architect of Ericsson's network management product portfolio. He joined Ericsson in 2000 and has over 20 years' experience in software development and architecture in the OSS domain. Johnston holds a B.Sc. in pure and applied physics from the University of Manchester Institute of Science and Technology in the UK.

References

1. Ericsson – A leader in the O-RAN Alliance ↗

Further reading

- Ericsson whitepaper, An intelligent platform: The use of O-RAN's SMO as the enabler for openness and innovation in the RAN domain ↗
- Ericsson, rApps ↗
- Ericsson, Intelligent Automation Platform ↗

Asset Administration Shell: Enabling 5G network digital twins for industry integration

Authors:
Ahmet Cihat Baktir, Elham Dehghan Biyar, Gergely Seres, Paul Stjernholm, Merve Saimler, Mehmet Karaca, Sultan Ertas, Deniz Cokuslu, Hubert Przybysz, Yunus Donmez

To ease the adoption of 5G technology in industrial applications, Ericsson is exploring the use of tools and processes that are well known in the operational technology industry. Our research shows that the Asset Administration Shell is a tool with great potential to deliver seamless 5G network integration.

ISSN 0014-0171 284 23-3403 | Uen

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

The integration of 5G into future factories and information technology (IT) / operational technology (OT) processes presents an opportunity to establish an ecosystem that includes functionalities to address the various network requirements of smart manufacturing [1].

From an OT operator perspective, the industrial 5G network is understood as an enabler of improved efficiency and productivity. OT operators do not, however, want to focus on the internal structure of the network (NW) and management systems. To enable 5G integration and maximize the benefit for automation processes, the telecommunications industry must simplify the use of 5G for the OT industry and factory operators.

The Asset Administration Shell (AAS) is a widely adopted solution in the industrial domain that enables communication among heterogeneous systems and components within the Industry 4.0 (I4.0) architecture. Its digital-twin-like capabilities and the standardized communication mechanisms make it possible to create a virtual representation of the 5G system (5GS) and seamlessly integrate it into a large industrial ecosystem. The AAS can thereby serve as a proxy for the 5GS and a variety of other complex systems on the factory floor. AAS principles enable the implementation of the relevant interfaces and integration points, ensuring smooth integration at multiple layers.

INDUSTRY 4.0

I4.0 is a revolution in manufacturing that combines digitalization and advanced technologies to create the factories of the future. At its core, I4.0 utilizes digital technologies such as the Internet of Things (IoT), artificial intelligence (AI) and robotics to enable the seamless integration of physical and digital systems. I4.0 aims to enhance production efficiency and flexibility by utilizing 5GS features, such as always-available wireless connectivity with high bandwidth, capacity and low latency, across all stages of processes and assets.

The AAS can serve as an excellent tool for the OT systems to interface with the 5GS, without the need for OT-focused staff to build up competence in cellular networks. Process engineers and system integrators building OT automation systems can perform their jobs using their existing OT processes and tools with the help of the AAS. Translation to the telco-specific terms and technologies is done by the AAS, which relies on the network exposure application programming interfaces (APIs) of the 5GS.

5G integration into Industry 4.0

There are various complex systems and applications on a factory floor, with different vocabularies and interaction

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **5G-ACIA** – 5G Alliance for Connected Industries and Automation | **5GS** – 5G System | **AAS** – Asset Administration Shell | **AI** – Artificial Intelligence | **API** – Application Programming Interface | **DT** – Digital Twin | **I4.0** – Industry 4.0 | **IoT** – Internet of Things | **IT** – Information Technology | **NDT** – Network DT | **NW** – Network | **OPC UA** – Open Platform Communications Unified Architecture | **OT** – Operational Technology | **QoS** – Quality of Service | **RAN** – Radio Access Network | **SEAL** – Service Enabler Architecture Layer | **TSN** – Time-Sensitive Networking | **UE** – User Equipment

models [2]. As the number of separate but related subsystems increases, the complexity of the integration and interoperability increases as well. Furthermore, the new revolution in the industrial domain envisions the creation of digital twins (DTs) of the systems that play a pivotal role in accelerating the digitalization of enterprises.

DTs are virtual replicas of physical assets (machines, robots and automated guided vehicles, for example), processes (in manufacturing and product development, for example) and/or systems that enable real-time monitoring, analysis and optimization of processes. By maintaining the synchronization at regular intervals between digital and physical representations, DT technology differentiates itself from typical simulators [3]. DTs enable accelerated digitalization by offering several key benefits including anomaly detection, proactive maintenance, process optimization and quality improvements. To maximize the benefits, DTs must be integrated into the larger industry ecosystem.

Interoperable interaction among subsystems will be required to realize the vision of the factories of the future, which will create a larger scale system with greater complexity. The most efficient way to deal with this complexity and ensure maximum benefit from 5G investments is to integrate the 5G NW into OT processes as an additional subsystem.

The requirements of a proxy to enable 5G-OT integration

Industrial systems demand optimized operation, which requires always-on connectivity. Cellular systems provide the efficient coverage, Quality of Service (QoS), mobility of field devices, security and flexibility needed to realize use cases with diversified requirements. 5G plays a pivotal role in enabling smart factories by providing wireless connectivity with ultra-reliable and low-latency communication capabilities, and its integration with time-sensitive networking (TSN) offers seamless connectivity for diverse industrial use cases.

To maximize the benefits, DTs must be integrated into the larger industry ecosystem.

TSN is an extension of the widely adopted Ethernet standard. It defines a range of traffic-shaping and redundancy methods that enable time-bounded communication over Ethernet with high reliability for smart manufacturing applications. 5G and TSN technologies are designed to grant converged communication on a common network infrastructure for an extensive range of services [4].

End-to-end 5G-OT integration must be supported at the OT operator's application layer. Delivering 5GS capabilities using exposure interfaces to the OT/IT applications enhances the automation degree throughout the factory floor for process automation, production IT and logistics. Communication service monitoring and network management capabilities are also essential for industrial applications to achieve automation. To support them, the Service Enabler Architecture Layer (SEAL) [5] specified by the 3GPP (3rd Generation Partnership Project) aims to support vertical applications. It provides access to the 5G communication services through RESTful (representational state transfer) web service APIs that comply with the 3GPP Common API Framework to provide flexibility in integrating with vertical applications [6].

The SEAL framework supports various management services and defines reference points for communication between the functional model components. To test the idea of a simplified and standardized API for enterprise access, leading technology company ABB and Ericsson developed a proof of concept, in which a prototype implementation of 5G exposure is integrated [7]. The main motivation of this exposure framework is to simplify 5G NW usage for verticals and industries through a common API. The results have proven that the 5G exposure interface enables OT enterprises to use 5G as a part of their system infrastructure, which increases production flexibility and allows scaling up to a large number of 5G-connected devices in an organized and secure manner.

Interoperability among devices, sensors, applications, networks and any heterogeneous systems is important to realize the vision of I4.0 – that is, developing smart factories equipped with intelligent human-to-machine and machine-to-machine cooperation. It requires augmenting the data from diverse heterogeneous resources under real-time conditions. This requirement brings challenges in creating an efficient and reliable information management infrastructure for both 5G and OT systems.

It is important to arrange complex and partially competing standards on a multitude of communication levels such as device integration, event processing, data analytics and cloud operations. For example, the Open Platform Communications Unified Architecture (OPC UA) [8] information model provides the essential infrastructure for interoperability across the enterprise, from machine-to-machine, machine-to-enterprise and everything in between. This facilitates a level of plug-and-play between applications from various vendors by identifying a standard information model for the cases where simply standardizing the message format is insufficient.

Ultimately, to be successful, a 5G-OT integration solution needs to be widely adopted by the industrial domain and satisfy three key requirements:

1. Enable standardized interaction for interoperability
2. Provide a common interface for 5G data and capability exposures
3. Deliver seamless integration of different systems.

The Asset Administration Shell as an integration technology

The AAS is one of the promising frameworks in I4.0 deployments that provides the relevant capabilities to act as a bridge between various systems on the factory floor using different technologies, such as the 5G NW and industrial processes. The AAS supports interoperability among industrial devices and provides a unified management interface for OT operators.

The AAS was initially proposed by Plattform Industrie 4.0 [9]. To facilitate digitalization for enterprises, Plattform Industrie 4.0 introduced the three-dimensional map called Reference Architectural Model Industrie 4.0 (RAMI 4.0). This is a service-oriented architecture that provides a structured view over I4.0 deployment. In this architecture, the AAS plays a key role in integrating assets (machines, products and documents, for example) into the digital world, fostering a common understanding and viewpoint among all participants involved in industrial processes.

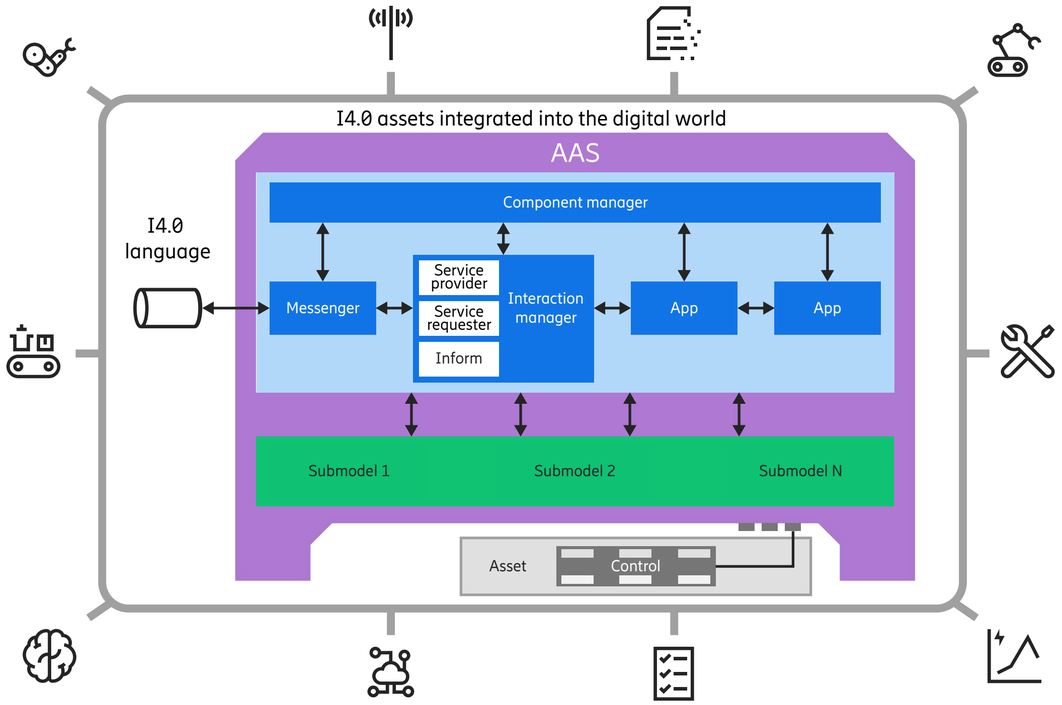


Figure 1: High-level view of the AAS and its components

The AAS enables the representation of functionalities, features, characteristics and other relevant information related to an asset in a digital form. This information is structured into various submodels within the passive component of the AAS. The AAS can also interact with other shells through its active part. While the passive role of the AAS refers to providing responses to the external access requests (read, write and modify submodels, for example), the active part incorporates decision-making functions and interaction mechanisms that enable peer-to-peer interaction.

Figure 1 illustrates the internal components of the AAS and how it connects different subsystems, with the messenger and interaction manager modules being responsible for implementing protocols and interfaces for message exchange. AAS interaction is established through the I4.0 language specified by VDI/VDE (The Association of German Engineers/Association for Electrical, Electronic & Information Technologies) 2193 [10]. Standardizing the I4.0 language not only provides an interaction mechanism but also enables interoperability between different systems, fulfilling an important requirement for an integration mechanism.

While there are multiple approaches that offer uniform data access and interoperability (including OPC UA), standardization in the I4.0 domain positions the AAS as a promising solution. Notably, recent standardization projects in the IEC (International Electrotechnical Commission) have been initiated to define the structure, use cases, information meta-model and security provisions of the AAS.

With the aim of creating digital representations of industrial assets for OT operators, the 5G Alliance for Connected Industries and Automation (5G-ACIA) utilizes AAS principles to describe 5G networks [11]. Recognizing the 5G NW and 5G user equipment (UE) as assets to be integrated into industrial automation systems, the 5G-ACIA proposes the utilization of two AAS types: 5G NW AAS and 5G UE AAS, representing the DTs of these complex subsystems. By utilizing the AAS principles, it is possible to achieve seamless integration of the 5GS into IT/OT processes in the industrial sector.

Given the complexity of existing subsystems with diverse objectives and characteristics, the AAS can serve as a proxy that interfaces various subsystems (network, production, and management and control, for example), enabling the creation of a larger system capable of exchanging information for process automation.

Network management, orchestration and service assurance processes can be customized and enhanced to meet industrial requirements. In this context, the AAS can serve two roles. Firstly, it can act as a DT of the 5G NW and UE, implementing decision-making and optimization algorithms. This enables the simulation of different configurations and the automation of management processes. For instance, the 5GS already offers positioning services for 5G-capable devices, utilizing various positioning techniques for different scenarios. Exchanging the positioning information with other industrial applications through the AAS may therefore extend the scope of other services using this information. The AAS may also be responsible for fusing positioning and industrial process data to propose configurations in the network in a way that the 5GS becomes more efficient in meeting industrial requirements.

Figure 2 illustrates the role the AAS plays in integrating different subsystems within an enterprise (the inner ring), along with the capabilities gained through this integration (the outer ring).

The role of the AAS in factory system automation

The AAS can play an important role in enabling the automation of industrial network management. This is because the inherent properties of the AAS allow interoperability among different subsystems in the OT domain and accommodate the DT-like capabilities.

In a recent article, Ericsson demonstrated that a network DT (NDT) has the ability to reuse existing 5GS management functions to enhance the performance of an industrial network [3]. An NDT that incorporates different analysis tools and behavior models can interact with other DTs — including factory DTs — to further enhance the performance of industrial cellular networks, increase the resolution of management functions and benefit from the information available in other domains.

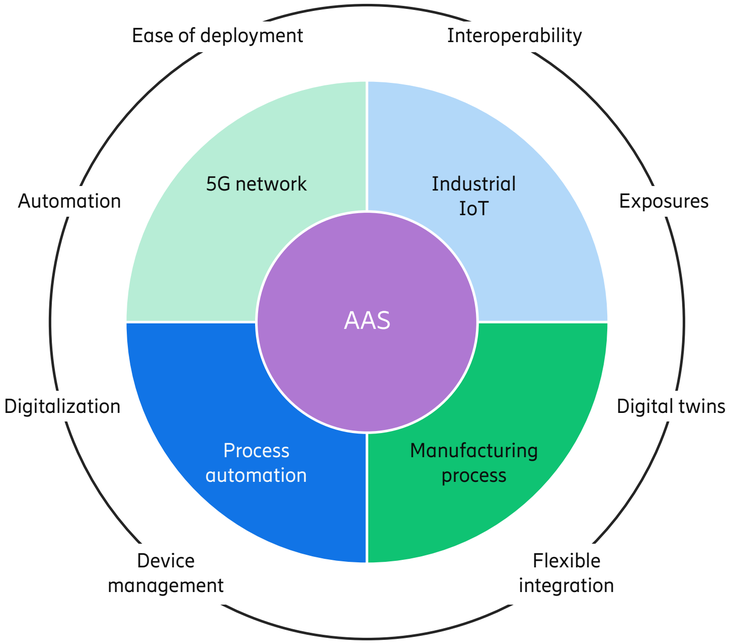


Figure 2: The role of the AAS in system integration and the resulting capabilities

The O-RAN Alliance's QoS-based resource optimization use case [12], for example, introduces a solution that largely depends on resource configuration and performance measurements. The idea of benefiting from enrichment information provided by external applications for improved performance is also presented in this use case. In this context, the AAS can aggregate information from any industrial application, including production scheduling. The ability to preprocess the information collected from different enterprise applications enables the AAS to propose relevant radio access network (RAN) configurations or policies.

Use case: Service assurance for collaborating robots

The use case presented below is part of a use-case family named Collaborating Robots that has been widely explored in various studies [13]. In this version, the AAS is used to integrate the 5GS with industrial applications that are responsible for managing field devices and manufacturing processes. The main goal is to apply a solution for configuration of the industrial 5G NW by utilizing the AAS principles.

In a factory environment, individual robots are not usually capable of communicating with each other, which can be an obstacle for certain tasks that require collaboration. The operation of these robots may depend on some external entities, such as a common machine vision. A vision system, which is connected to the network, can provide a set of relevant functions such as positioning and asset tracking. The system that provides vision-based functions and services can be a part of the wired infrastructure.

We can use the AAS to integrate the 5GS and industrial Ethernet systems.

All factory automation is built on a system-of-systems concept in which several OT automation systems and OT/IT/cellular network systems collaborate to execute the manufacturing process. The AAS can be used to integrate the control of all these systems seamlessly. For example, considering a use case of vision-assisted collaborating robots, we can use the AAS to integrate the 5GS and industrial Ethernet systems. The 5G-ACIA envisions a customized implementation of a similar integration, in which integration of two systems (5G and TSN) is established by adding relevant submodels in the AAS [11].

One potential way to overcome the difficulty of achieving direct communication among machines provided by different vendors is by adding relevant submodels into the AAS representations of various devices (5G UE AAS) and other industrial applications. Necessary submodel elements related to devices and the task can be communicated and negotiated among participating AAS entities. Alternatively, the active part of AAS instances can implement any relevant decision-making system to coordinate the collaboration and configure the assets.

Consider a scenario in which the factory operator initiates a new collaboration between two robots. There is a requirement to reconfigure the 5G NW to support the connectivity and performance demanded by these robots. Assuming that the initial configuration management and QoS parameters are handled in the engineering phase when deploying the 5G subsystem, this new task requires modification of the connectivity, including QoS definitions (maximum latency, minimum reliability and so on) for the corresponding devices.

This network configuration is possible thanks to AAS capabilities that enable standardized interaction for interoperability, provide a common interface for 5G data and capability exposures, and deliver seamless integration of different systems. Typically, 5G configuration/reconfiguration can be performed by any authenticated application using 5G exposure interfaces. However, 5G NW AAS can define the new set of QoS parameters based on the aggregated information provided by different sources including industrial processes (requirements, for example), the current state of the network and device characteristics.

Figure 3 illustrates what happens when a factory operator triggers a new collaboration task by invoking the related functions provided by the industrial application (step 1, at top left). Based on the task requirements, the industrial application collects relevant information from the devices

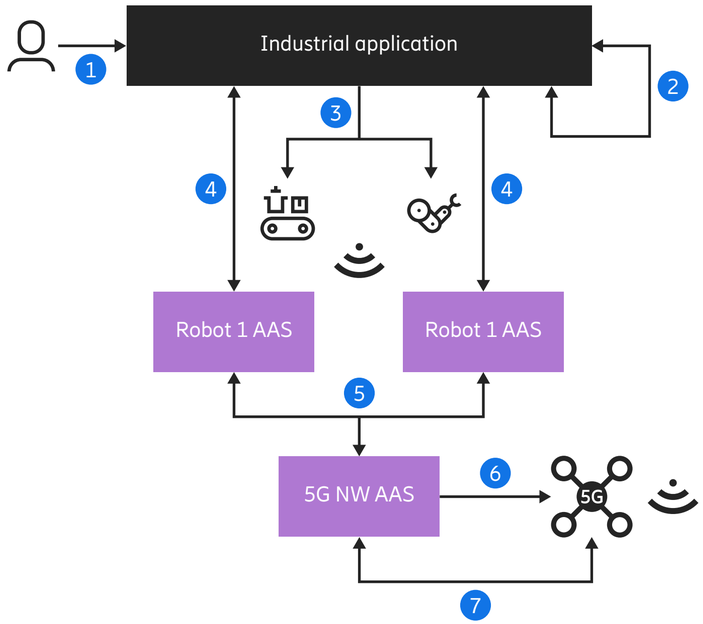


Figure 3: How the AAS automates the network reconfiguration for QoS assurance

such as their location, status and capabilities. Step 2 identifies the devices that are the most appropriate for this task. The devices are provisioned in step 3, and they seamlessly establish a secure connection with their AAS representations. In step 4, the AAS instances of the robots retrieve necessary information from the industrial application. The AAS instances are also able to exchange device capabilities with each other.

In step 5, the 5G NW AAS receives information that the industrial domain has initiated a new task that may require changes to the network configuration to ensure provision of the desired QoS. The 5G NW AAS retrieves the device capabilities and requirements of the task using standardized AAS interaction. In step 6, 5G NW AAS executes DT-like capabilities – such as running what-if scenarios to find out the best possible configuration – to define the QoS parameters for each robot. These parameters are then sent as proposed solutions to the 5G NW through the exposed API (proprietary, network exposure function or SEAL). As a conditional case, if the defined QoS parameters are not satisfactory (which can be inferred through the monitored data) or cannot be applied due to an issue (resource capacity, for example), new parameters can be proposed in step 7. By continuously monitoring the network and the factory floor, the AAS can evaluate whether the requirements of network and industrial process are satisfied or not.

Conclusion

The most efficient way for operational technology (OT) enterprises to adopt 5G technology is to ensure seamless integration of 5G with OT solutions using existing tools and processes. The Asset Administration Shell (AAS) is a widely adopted solution in the industrial domain that enables communication among heterogeneous systems and components within the Industry 4.0 architecture. The digital-twin-like capabilities of the AAS, together with its standardized communication mechanisms, make it possible to create a virtual representation of the 5G system (5GS) that is fully integrated into a large industrial ecosystem.

In addition to the interoperability the AAS brings, the integrated decision-making functionalities in the AAS also make it a powerful framework for automation. By using AAS principles to represent the 5GS in the digital world, factory operators have the opportunity to achieve high levels of both integration and automation, and thereby make significant progress toward the vision of the factories of the future.

The authors

Ahmet Cihat Baktir is a senior researcher who joined Ericsson in 2020. His current research activities and interests focus on network automation for industries, DTs and intent-based management. Baktir holds a Ph.D. in computer engineering from Boğaziçi University in Istanbul, Turkey.

Elham Dehghan Biyar is an experienced researcher at Ericsson Research whose work focuses on network automation and non-public networks. She joined the company in 2019. Biyar holds an M.Sc. (cum laude) in computer networks from Istanbul Technical University in Turkey.

Gergely Seres is an expert and chief architect of software technology and application architecture at Business Area Cloud Software and Services. His current focus is on 5G, 6G and the IoT. Since joining Ericsson in 1998, he has held several research, technical and managerial positions. Seres holds a Ph.D. in electrical engineering from the Budapest University of Technology and Economics, Hungary.

Paul Stjernholm joined Ericsson in 1995 and has been dedicated to the mobile telecom industry since 1G. In his current role as concepts researcher, his work focuses on RAN management strategies, automation and standardization. Stjernholm holds an M.Sc. in applied physics and electrical engineering from Linköping University, Sweden.

Merve Saimler joined Ericsson Research in 2020 and currently works as a senior researcher. Her work encompasses various areas, including AI, AI as a service, network slicing, Service Level Agreement management, Industry 4.0 and the AAS. Saimler holds a Ph.D. from Koç University in Istanbul, Turkey.

Mehmet Karaca is a senior researcher at Ericsson Research, where he focuses on intent-based automation and the design of cognitive networks using machine learning. He joined Ericsson in 2021. Karaca holds a Ph.D. in electronics engineering from Sabancı University in Istanbul, Turkey.

Sultan Ertas is an experienced researcher who joined Ericsson in 2021. Her research interest is centered around slicing, intent-based automation, service management and DTs. Ertas holds a B.Sc. in electronics and communications engineering from Yıldız Teknik University in Istanbul, Turkey.

Deniz Cokuslu joined Ericsson in 2018 as a solution architect. He has been working as a senior researcher at Ericsson Research since 2022. Cokuslu received his Ph.D. in 2012 from Paul Sabatier University in Toulouse, France.

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 and exposure of 5GS capabilities. Przybysz holds an M.Sc. in telecommunications from the Warsaw University of Technology in Poland.

Yunus Donmez joined Ericsson in 2018. In his current role as research leader at Ericsson Research Networks, he focuses on network management and automation, industrial integration, and automation of non-public networks in smart manufacturing and network digital representation. Donmez holds a Ph.D. in computer engineering from Boğaziçi University.

References

1. 5G-SMART Deliverable D1.1 Forward Looking Smart Manufacturing Use Cases, Requirements and KPIs, June 30, 2020 ↗
2. Ericsson Blog, What is semantic interoperability in IoT and why is it important?, July 23, 2020, Widell, N; Keränen, A; Badrinath, R ↗
3. Ericsson Technology Review, Network digital twins – outlook and opportunities, December 15, 2022, Öhlén, P; Johnston, C; Olofsson, H; Terrill, S; Chernogorov, F ↗
4. Ericsson Technology Review, 5G-TSN integration meets networking requirements for industrial automation, August 27, 2019, Farkas, J; Varga, B; Miklós, G; Sachs, J ↗
5. 3rd Generation Partnership Project (3GPP), Service Enabler Architecture Layer for Verticals (SEAL); Functional architecture and information flows, Sophia Antipolis, France, TS 23.434 ↗
6. 3rd Generation Partnership Project (3GPP), Common API Framework for 3GPP Northbound APIs, Sophia Antipolis, France, TS 23.222 ↗
7. Ericsson Technology Review, Creating programmable 5G systems for the Industrial IoT, October 27, 2022, Seres, G; Schulz, D; Dobrijevic, O; Karaağaç, A; Przybysz, H; Nazari, A; Chen, P; Mikecz, M; Szabó, Á ↗
8. OPC Foundation – Unified Architecture ↗
9. Plattform Industrie 4.0, Details of the Asset Administration Shell – Part 1, May 30, 2022 ↗
10. VDI Standards, VDI/VDE 2193 Blatt 1 – Language for I4.0 Components – Structure of Messages, April 2020 ↗
11. 5G-ACIA, Using Digital Twins to Integrate 5G into Production Networks ↗
12. O-RAN Alliance – Use Cases Detailed Specification 11.0 ↗
13. 5G-SMART Deliverable D2.3 Validation of 5G Capabilities for Industrial Robotics, May 31, 2022 ↗

Further reading

- Plattform Industrie 4.0, Specification – Details of the Asset Administration Shell, November 2020 ↗
- Eclipse Foundation, BaSyx / Documentation / AssetAdministrationShell ↗
- Ericsson, Industry 4.0 ↗

Acknowledgements

The authors would like to thank Kurt Essigmann, Stephen Terrill and Torbjörn Cagenius for their contributions to this article.

Service quality monitoring – an essential tool in the digital economy

Authors:
Elisabeth Müller, Malgorzata Svensson, Máté Walthier, Christer Gustafsson, Attila Báder

Successful execution of a new business use case in the digital economy requires the ability to consistently deliver a good user experience. This, in turn, requires the ability to prove that the service delivered in the value chain is in line with the service and application characteristics agreed between all of the stakeholders. Service quality monitoring is a key capability to make such assessments.

ISSN 0014-0171 284 23-3404 | Uen

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

The value chain in the digital economy is comprised of multiple stakeholders including application service providers (ASPs), application developers, aggregators, communication service providers (CSPs) and customers both in the enterprise and consumer segments. Each of these stakeholder groups has an important role to play in experience management.

In the application developer ecosystem, developers create applications for enterprises in areas ranging from critical machine-type communication to health care, public safety and manufacturing industries. Each application has well-defined characteristics that lead to specific quality of service (QoS) requirements on connectivity that must be fulfilled to achieve good user experience. **Figure 1** provides an overview of all the stakeholders that play significant roles in end-user experience management in the digital economy. It also shows the major information flows for managing service quality.

The role of the ASPs is to offer the applications to the enterprises. The enterprises that want to use the applications require connectivity services facilitated by the CSPs that meet the QoS requirements of the various applications. Because ASPs have relationships with multiple CSPs, aggregators are frequently involved in facilitating those relationships [1].

SERVICE QUALITY MONITORING: KEY TERMS

Quality of experience describes the service quality that is perceived by a consumer. Examples of QoE metrics include video resolution and frames per second.

Device-connection quality is the observed quality of traffic generated by applications running on a single user device. The device can have one or multiple sessions active, where one application can use one or many sessions (see Figure 3). Traffic quality is defined by network metrics such as throughput and latency.

Connectivity-service quality is the observed quality of a service that multiple enterprise devices use (see Figure 3). Access to the service results from the contract agreement with strictly defined SLAs between the CSP and the customer.

The quality of experience (QoE) that a user perceives when running an application largely depends on the quality of the device connection service in the CSP network, which is defined by QoS. Assurance processes use QoS insights to improve QoE and take action in the case of service degradation or Service Level Agreement (SLA) violation.

By enabling the exchange of correct and relevant information between the stakeholders in the value chain, service quality monitoring helps to ensure optimal experience management in the digital economy.

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **AI** – Artificial Intelligence | **API** – Application Programming Interface | **ASP** – Application Service Provider | **CSP** – Communication Service Provider | **E2E** – End-to-End | **IMS** – IP Multimedia Subsystem | **KPI** – Key Performance Indicator | **ML** – Machine Learning | **NG** – Next Generation | **QoE** – Quality of Experience | **QoS** – Quality of Service | **RAN** – Radio Access Network | **SLA** – Service Level Agreement | **VoLTE** – Voice over Long Term Evolution | **VoNR** – Voice over New Radio

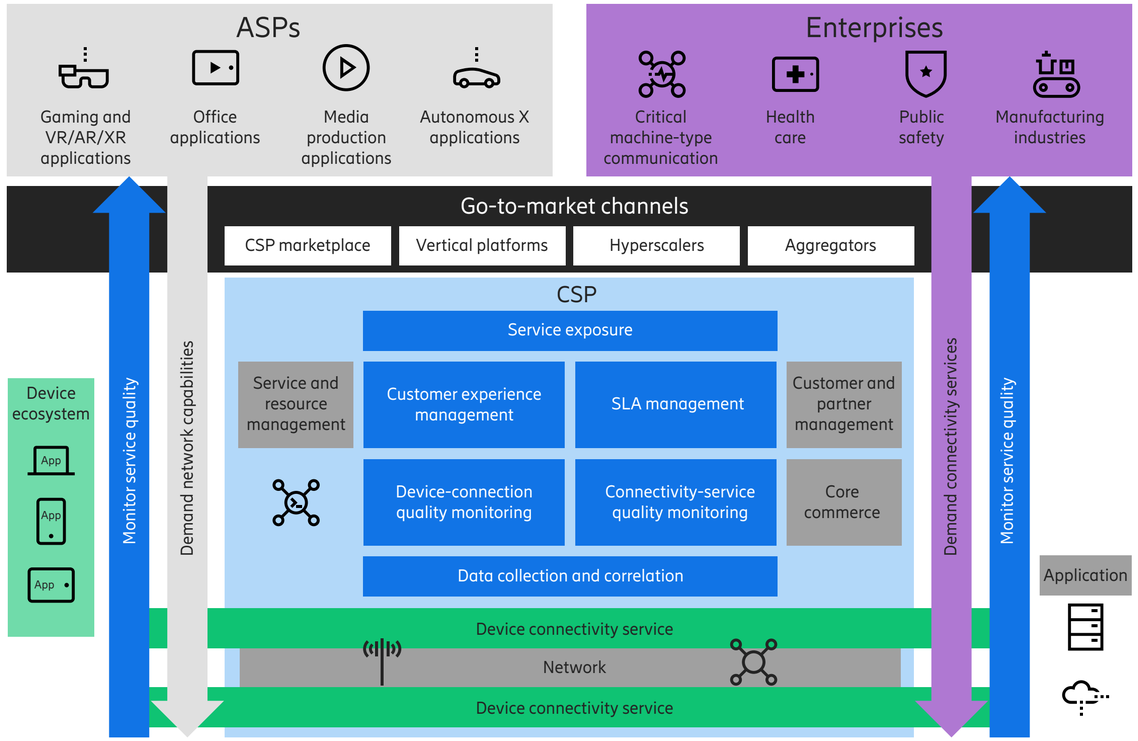


Figure 1: Experience management in the digital economy

The role of service quality monitoring in value creation

Service quality monitoring contributes to value creation by:

1. Assuring SLAs
2. Enabling smoother interaction between applications and networks
3. Facilitating new business models.

Assuring Service Level Agreements

There is a growing market for CSP wireless communication services and other assets to be made available for enterprises' value production in segments ranging from manufacturing to distribution, entertainment, health care, defense, railways, public safety and government, and beyond. Internet of Things applications are playing a key role in driving this development, along with more traditional drivers such as the desire for cost reductions, time to market/customer gains, improved mobility, quality and customer experience, as well as the ability to create new services and/or enhance existing ones.

In enterprise use cases, the communication services delivered by CSPs become an integrated part of an enterprise's production, which means they must live up to the contracted service qualities or the production will be impaired. Service reliability and availability are vital ingredients for the production. CSPs assume the supplier role in relation to the enterprises: they offer SLAs that include service-quality expectations and SLA-violation consequences to back up their offers and price structures. SLAs are based on service-specific quality parameters, which must be monitored and assured.

The service-connectivity quality information is fundamental for the CSP to drive business with enterprises whose applications and production chains depend on the service quality and availability of the connectivity. Therefore, enterprise connectivity contracts are always paired with detailed SLAs. The SLAs specify the expectations on service quality in terms of target values for key performance indicators (KPIs) and quality indicators determined for the individual connections of the devices of the enterprise. Thus, SLA monitoring and service assurance for the enterprise connectivity require the detailed insights derived in device-connection quality monitoring. The CSP can address individual KPI violations by tailored actions applied to individual network functions or the whole connectivity service offered to the enterprise. Prominent examples are traffic scheduling features in a radio-access network (RAN), associated configuration optimizations or intent-driven zero-touch automation of the network.

Enabling smoother interaction between applications and networks

Applications have various behavior capabilities, as well as in-service and performance characteristics, that place demands on networks. Most importantly, networks must support the required traffic mix and patterns to meet the performance requirements, functional behavior and other characteristics of the applications [2]. The applications are either capable of adapting to the network conditions – by adjusting frame rates or postponing certain operations to a later point in time, for example – or they will request that the network adapt its performance. The interaction between the applications and the network is done through application programming interfaces (APIs) that are initiated in one of three ways – by the network, by the application server or by the application itself.

Consider the example of a gaming application in the consumer segment and a collaboration application in the enterprise segment reacting to information about insufficient service quality. Both have high traffic demands and it is obvious that performance degradation in either case would lead to a negative customer experience. Service quality monitoring (and potentially even prediction) would ensure a consistently good user experience for both applications by making it possible for the network to react to information about insufficient service quality. This could be done by using APIs to boost performance or by moving sessions to other more suitable network connections.

Information about service quality and network congestion can be communicated in different ways. The most elaborate method is by exposing the monitoring results of quality at various granularity levels such as application-flow level, device-session level and customer-device level.

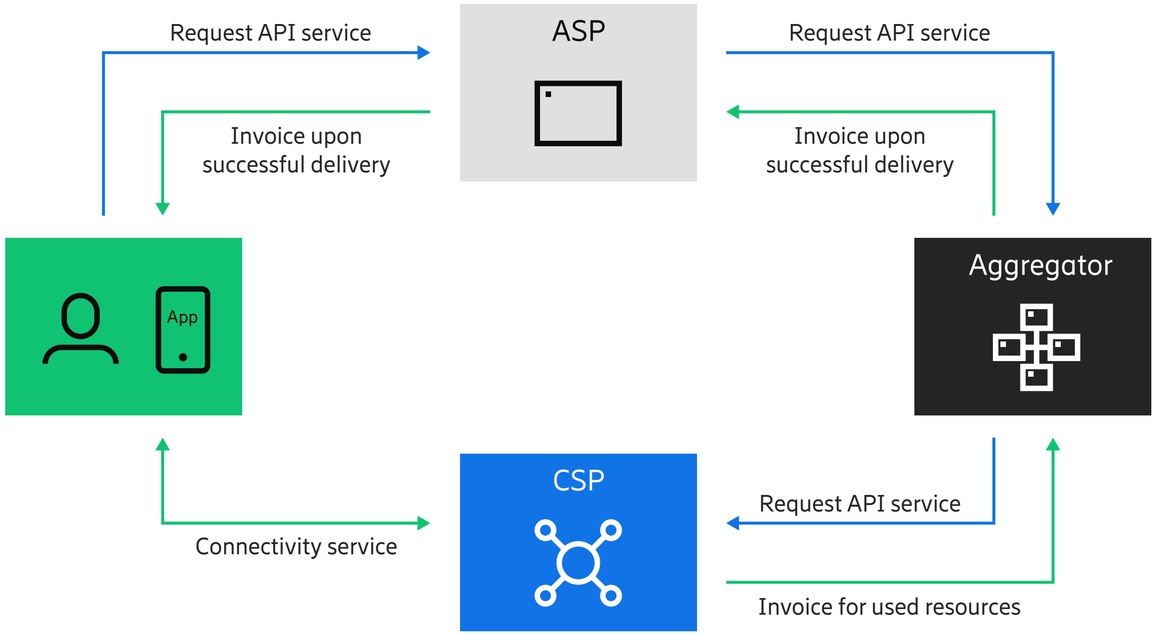


Figure 2: Interactions between stakeholders in value creation

Application developers can design applications to influence and react to service-quality information exposed by the CSP's network. Applications can adjust the amount of data sent – if the application logic permits – or request a quality boost applied to the data traffic originating from the application. Another option is to use the quality insights on different connectivity services to direct certain data traffic to these connectivity services, for example, by using traffic policies implemented in data routers or using User Equipment Route Selection Policy technology on consumer devices.

Facilitating new business models

CSPs are increasingly making use of business partners such as aggregators and hyperscalers to reach developers and ASPs in the network exposure business. One of the most interesting network services here is the ability to dynamically influence the service quality applied to data traffic.

When an ASP requests an API service such as an increase in service quality on behalf of a user through an aggregator, it needs to prove that the API service has been delivered successfully and that the request had the desired effect before it can invoice the user. The detailed insights from service quality monitoring, available on device and even session level, make it possible to compare the service quality delivered against the service quality requested and thus provide the required proof. **Figure 2** shows the various interactions between the stakeholders in this scenario that require service quality monitoring.

Comprehensive service quality monitoring

The scope of service quality monitoring is to retrieve knowledge about service quality from the data sources that comprise a CSP network, including data metrics that originate from single network functions, various network and cloud infrastructure domains, and device and application domains. Service quality can be monitored for a single application, for application groups and/or for a specific device. Alternatively, it can target all the traffic a network function or domain handles over a certain period.

Metrics from the various data sources must be collected and filtered to ensure that only relevant data is processed through correlation to form a solid information base. The amount and variety of the data produced by these various data sources presents a major challenge to make service quality monitoring effective and economical. A huge amount of detailed input data must be processed, consolidated and correlated across different domains to derive meaningful input for the internal and external consumers of this information.

As network functions and entire network domains are the CSP's responsibility, these data sources are easy to access. On the other hand, devices, application data and last-mile connectivity from the CSP's network to the application servers is much more difficult to access.

Service quality monitoring information is useful for network healing, troubleshooting, admission control and adjusting throughput in a RAN on the cell level up to fully autonomous networks driven by intents. When service quality monitoring evolves toward service quality prediction, the value increases significantly, but so does the challenge.

Service quality monitoring already provides machine learning (ML) models for two traffic types – classic mobile broadband traffic, and low-latency video and voice traffic originating from popular conferencing applications – and therefore supports application-specific quality determination. Other traffic types can be supported by training other ML models.

Monitoring device connection quality before and during service execution is a key capability.

Connectivity-service and device-connection quality monitoring

Figure 3 illustrates the key components of service quality monitoring. Connectivity service quality is calculated based on an average of the quality of all the device connections, measured for all devices sending traffic on a particular connectivity service. On a lower precision level, this can be determined without end-to-end (E2E) awareness of individual data sessions.

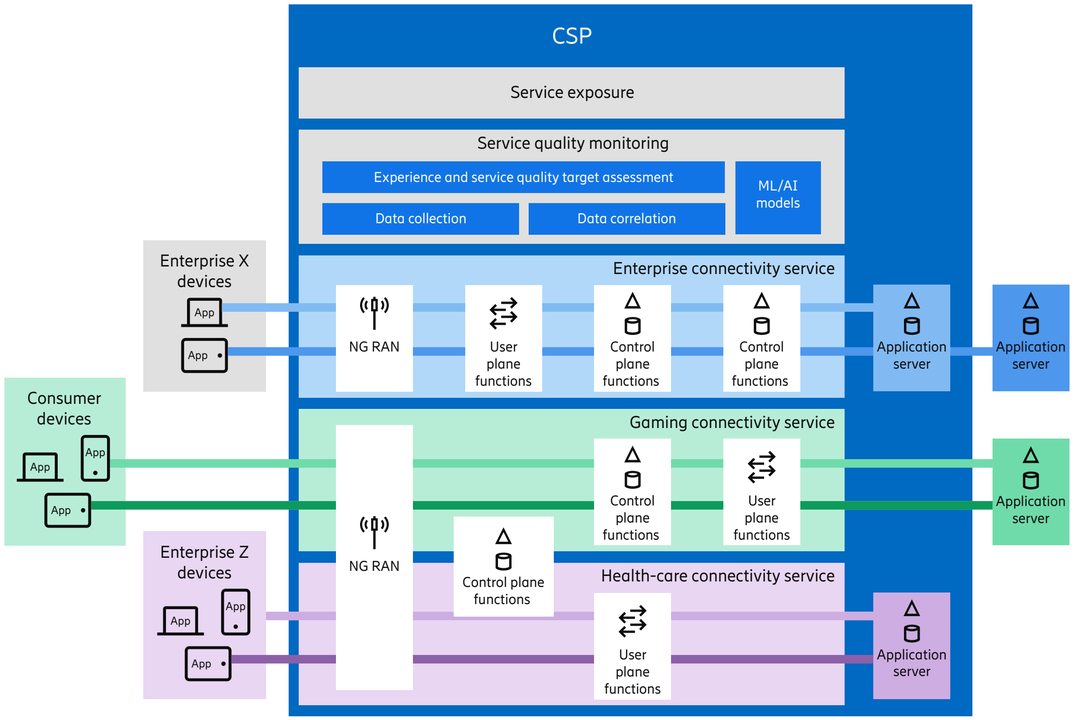


Figure 3: The key components of customer experience management

Device-connection quality monitoring refers to service quality monitoring at the device level. It uses measurement inputs from the different network functions that contribute to delivering the device connectivity service shown in Figure 3, correlates and aggregates the measurements and derives quality information for monitoring time frames. The 3GPP (3rd Generation Partnership Project) has standardized some aspects of this functionality in a network data analytics function.

Service quality monitoring at device level and even more at data-session level is based on individually reported events collected from the network functions. The key metrics needed for a proper evaluation are uplink and downlink bit rates, packet error rates, packet inter-arrival time, burst metrics, delay and jitter, potentially combined with metrics from the RAN related to signal strength on the downlink and uplink. These metrics are used to run objective E2E service quality analysis and also to estimate the QoE of a user running a specific application. Network-wide monitoring at session level requires high-volume event processing and correlation.

Service quality monitoring can be done in either prediction or evaluation mode. Service quality monitoring in prediction mode takes place prior to a network API service request originating from a CSP, a business partner or an application, and results in as accurate guidance as possible for the target function consuming the information.

Service quality monitoring in the evaluation mode takes place either during the API service or after it has been delivered. The purpose is to quantify the impact of application- or system-initiated actions on the delivered service quality, with the aim of providing information that the CSP can use as the basis for accounting and invoicing. The basis for the evaluation is provided by the ASP or the aggregator in the form of a quality target that reflects the original expectation of the ASP or user.

The four cornerstones of service quality monitoring

As highlighted in the top section of Figure 3, the four cornerstones of service quality monitoring are:

1. Data collection
2. Data correlation
3. Experience and service quality target assessment
4. ML/AI models.

Service quality monitoring consumes event streams from network functions. These event interfaces are usually proprietary and require adapters for every network function vendor. A data collection control function governs the process by dynamically configuring event sources to admit only the events required. All data sources must support appropriate filtering criteria that allow only the selection of data about specific subscribers.

Data correlation matches monitoring events from different sources for each subscriber in the network and calculates metrics for them. The processing and storage of monitoring events must only be done for those subscribers that have given their consent to be monitored. The data may be stored in a database or be streamed to a message bus.

Quality target assessment accumulates the partial results for the ongoing monitoring sessions and assesses whether or not the given service quality target has been reached.

The QoE assessment function uses artificial intelligence (AI) and ML to derive user QoE for a specific application based on device-connection-specific measurements retrieved from network functions belonging to the RAN as well as the transport domain. The traffic measurements are fed into an ML model that is responsible for estimating the QoE. This function is a collection of multiple traffic-pattern-specific ML models that must be trained upfront and – at evaluation time – applied to the data traffic types for which they have been trained. It is not possible to derive meaningful QoE estimates with a single ML model, due to the different traffic patterns and resulting effects on user experience in the case of quality degradation in different parts of the network.

The ML models are trained using different data sources representing the input and the expected outcome of an analysis. The data sources representing the input are the measurements collected from the RAN and the transport network for application-specific data traffic both in good conditions and in various failure or congestion scenarios.

The expected output of the modes is objective service-specific metrics such as video frame rate or resolution, and the estimation of subjective user scores. Model training requires consistent input and output data sets. Service-specific metrics – such as WebRTC (real-time communication) traffic metrics – are relatively easy to collect in lab environments, but the data needed for validation of quality estimation models must be collected in live networks that cover the network-wide scenarios for a call. This is usually done through mobility tests. User surveys are even more expensive. During normal use (model inference), when service-specific metrics or consumer feedback is not available for the CSP, the ML model estimates these QoE metrics based on the network service metrics. The service-specific ML models require training, retraining and monitoring during operation.

The evaluation challenge can be addressed by measuring the traffic burst bitrate.

Challenges and solutions

To be both effective and efficient, service quality monitoring must overcome several challenges, particularly with respect to data collection and quality assessment, due to factors such as the lack of standards, the amount of data to be processed and the lack of data about last-mile connectivity to the application backend.

Unfortunately, performance monitoring counters at network or service level are usually unsuitable for characterizing the performance of individual user connections. Because of this, monitoring events at individual session level are used from different domains such as IMS (IP Multimedia Subsystem), Packet Core and RAN. These events, usually in a proprietary format, must be related to user sessions and correlated to get a complete picture of the user session in the CSP network. Collecting measurements may interfere with service performance if performed in real time. Data pipelines based on a harmonized data ingestion architecture will address this challenge by facilitating access to high-resolution data and boosting efficiency on data collection [2].

The sheer volume of data that is available for and relevant to service quality monitoring requires the implementation of intelligent filter functions in various parts of the system, potentially including the data source itself. Constant and complete network monitoring is possible [2] but expensive, hence the need for a dynamic spotlighting function that can monitor parts of the network or a subset of the subscribers to limit the footprint of the solution. Filtering at subscriber level depends on the availability of subscriber information and may therefore only be possible after event correlation.

Another significant challenge to overcome in service quality monitoring is the fact that a CSP does not usually have access to quality metrics from the user equipment, applications and ASPs. This means the CSP can accurately measure the performance of the connectivity service it provides, but it can only estimate the user's perceived quality. The data gap can be closed by sending consumer service quality information from the ASP to the CSP through the application function and the network exposure function that are supported by the 3GPP architecture. It is expected that this functionality will be more commonly used in the future to help CSPs estimate actual E2E service quality and ensure SLAs.

Meanwhile, the quality-on-demand API service designed by CAMARA, the industry alliance driving the standardization of services for exposure, is expected to help overcome the quality target challenge by providing information about minimum expected bitrates for certain quality profiles requested by an application. These bitrates or throughput-based quality targets are complex to configure and evaluate, as there may be multiple root causes in a case where, for example, the CSP-observed bitrate value is lower than the desired target bitrate for a device. It could be that the device or application is not generating sufficient traffic, or that the application data network has a bottleneck, or that the CSP network caused the degradation. The evaluation challenge can be addressed by measuring the traffic burst bitrate and thus considering only cases of significant traffic injected into the network.

Finally, while CSPs cannot measure QoE metrics such as conversational quality for over-the-top services, they must have the ability to estimate them. Because today's user plane data is fully encrypted, it is not possible to derive service-specific parameters directly by network probing (that is, observing packet content, frame structure and so on). Even for the CSP-provided VoLTE (Voice over Long Term Evolution) or VoNR (Voice over New Radio) services the client-side information is limited: only the application server (IMS) data is available to the CSPs. ML models are the most reasonable approach for estimating these technical parameters and the resulting QoE. Training these models for external applications that are not provided by the CSP is an even more challenging task, as the traffic patterns of such applications are not known to the CSP.

Conclusion

Successful customer experience management in the digital economy requires the ability to understand the application-specific quality of experience (QoE) delivered to the users, so that appropriate actions can be initiated by the applications themselves or by the communication service provider (CSP) to improve quality when needed. Machine learning models make it possible to derive the QoE delivered to users by correlating application traffic patterns with device-level, session-level or even data-flow-specific key performance indicators (KPIs).

New services for device-connection quality monitoring and connectivity-service quality monitoring can deliver KPI information that is specific to individual devices and their data connections, as well as providing aggregated information for devices operated by a single enterprise. These insights allow CSPs, their business partners and the application developers to become active in customer experience management by applying a new set of tools for service quality management that are much more specific to the quality improvement needs of the digital economy. The capabilities of these new tools extend far beyond the well-known legacy toolset of throttling or rejecting service requests, which leads to the over-dimensioning of the system. The powerful device-connection and connectivity-service quality monitoring capabilities exposed through service APIs are key enablers for both information exchange between stakeholders and value creation in the digital economy.

The authors

Elisabeth Müller joined Ericsson in 2006. Since then, she has taken on many roles including system design, system management and solution architecture in all BSS areas. Mueller holds various patents within BSS and serves as a senior expert for monetization, partner and customer management, focusing most recently on service exposure architecture for the digital economy. She holds an M.Sc. in mathematics and business economics from Johannes Gutenberg University Mainz, Germany.

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

Máté Walthier joined Ericsson in 2004. He is a system architect with experience in real-time operating systems and Java Virtual Machine (JVM) development, distributed systems, database management systems, analytics, cloud, continuous integration/continuous delivery and user experience management. Walthier holds an M.Sc. in information computer science from the Budapest University of Technology and Economics in Hungary.

Christer Gustafsson joined Ericsson in 1994 and currently works as a principal software developer for RAN systems, focusing on function and performance development in the areas of services and QoS. He has more than 10 years of experience in developing the service performance in VoLTE and VoNR, both in RANs and Evolved Packet Core/5G Core. Gustafsson holds an M.Sc. in technical physics and electrical engineering from the Institute of Technology at Linköping University in Sweden.

Attila Báder joined Ericsson in 2001 and currently works as a solution architect. He is an expert on network analytics, focusing on AI/ML methods for service quality monitoring and assurance. Báder holds a Ph.D. in physics from Lajos Kossuth University in Debrecen, Hungary.

References

1. Ericsson Technology Review, Monetizing API exposure for enterprises with evolved BSS, January 12, 2023, Friman, J, et al. ↗
2. Ericsson Technology Review, Data ingestion architecture for telecom applications, March 16, 2021, Rönnberg, A-K, et al. ↗

Further reading

- Ericsson Technology Review, Network evolution to support extended reality applications ↗
- Ericsson Technology Review, Future network requirements for extended reality applications ↗
- Ericsson Technology Review, Autonomous networks with multi-layer, intent-based operation ↗
- Ericsson white paper, Cognitive reasoning for 5G network lifecycle management ↗
- Ericsson blog, The innovation potential of non-real-time RAN intelligent controllers ↗

Acknowledgements

The authors would like to thank Mikael Klein, Jan Friman, Stephen Terrill and Amandeep Kumar for their contributions to this article.

ISSN 0014-0171
284 23-3408 | Uen

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