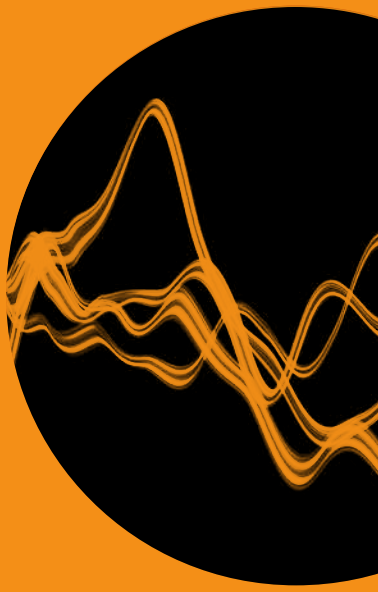
CHARTING THE FUTURE OF INNOVATION | #11·2021

ERICSSON TECHNOLOGY Review



5G SYSTEM NETWORK ROBUSTNESS

ERICSSON

ROBUSTNESS EVOLUTION:

Building robust critical networks with the 5G System

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

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

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

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

The 5G System (5GS) has been designed to provide the robustness required to support the growth of conventional MBB services, while also offering network support to new business segments and use cases with more advanced requirements in terms of reliability, availability and resilience. Consisting of the 5G Core (5GC), the

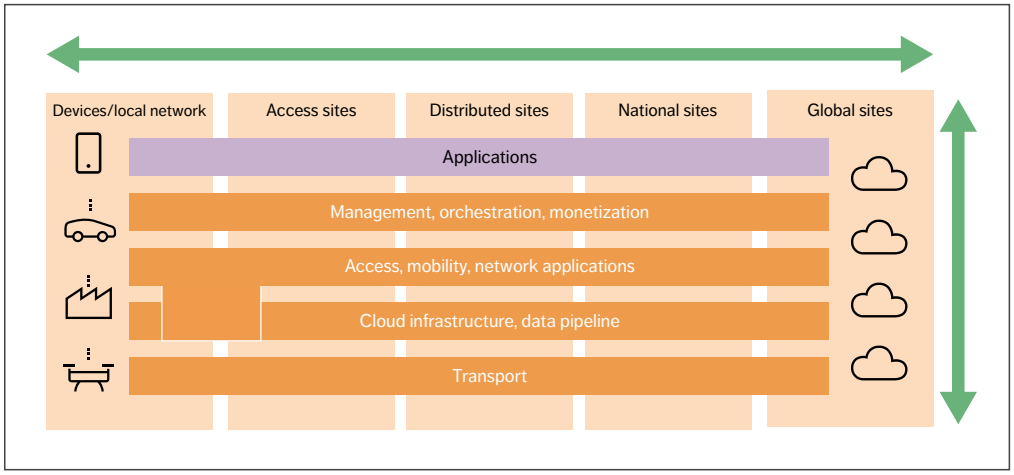


Figure 1 – Aspects impacting network robustness in a typical network

Next-Generation RAN (NG-RAN) and the user equipment (UE), the 5GS delivers new capabilities that enable enterprises with business-critical use cases in segments such as manufacturing, ports and automotive [1, 2] to take a major step forward in their digitalization journeys by replacing older means of communication with the 5GS. These new capabilities are also beneficial for mission-critical networks like national security and public safety deployments that are currently being modernized.

Definition of a robust network

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

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

Aspects impacting network robustness

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

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

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

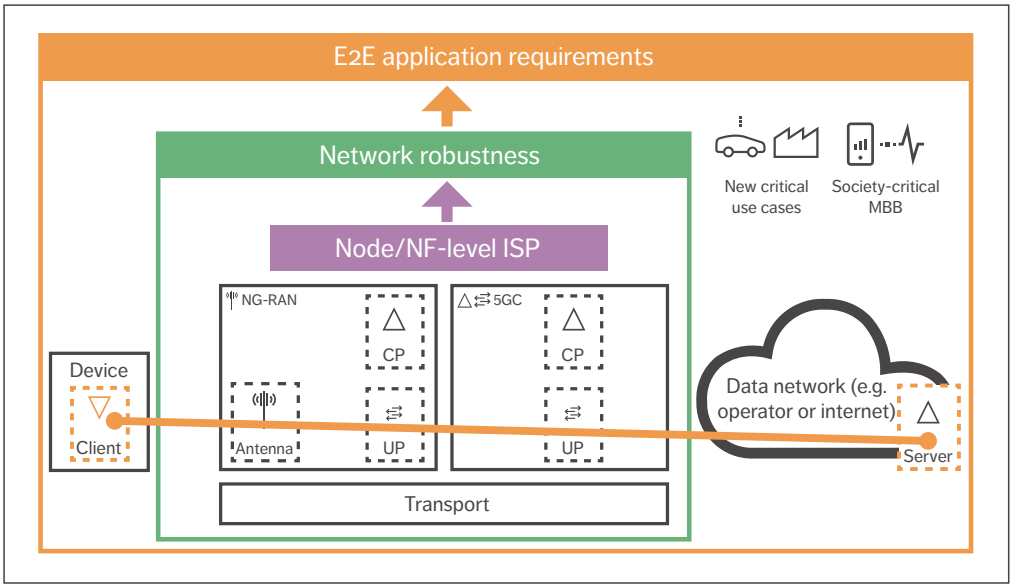


Figure 2 Shifting focus from node/NF-level to network robustness for demanding E2E applications

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

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

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

Improving the overall observability of the network is key to improving network reliability and resilience. Performance management counters within the network are currently used to provide performance visibility to domain management systems in the network. New ways of enabling observability will be needed to monitor the network service characteristics and fulfillment of the new application requirements. Increasing automation in the correction of network failures, predictions and launch of preventive actions must also be considered. In addition, observability related to

network failures is an important area to cover. One example of this would be determining actual network performance in case of failures in different parts of the network and assessing the impact on E2E services.

Robustness mechanisms in existing networks

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

Evolved Packet Core aspects

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

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

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

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

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

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

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

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

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

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

LTE RAN aspects

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

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

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

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

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

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

The 5G System robustness toolbox

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

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

5G Core aspects

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

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

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

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

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

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

Features and mechanisms at the NG-RAN level

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

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

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

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

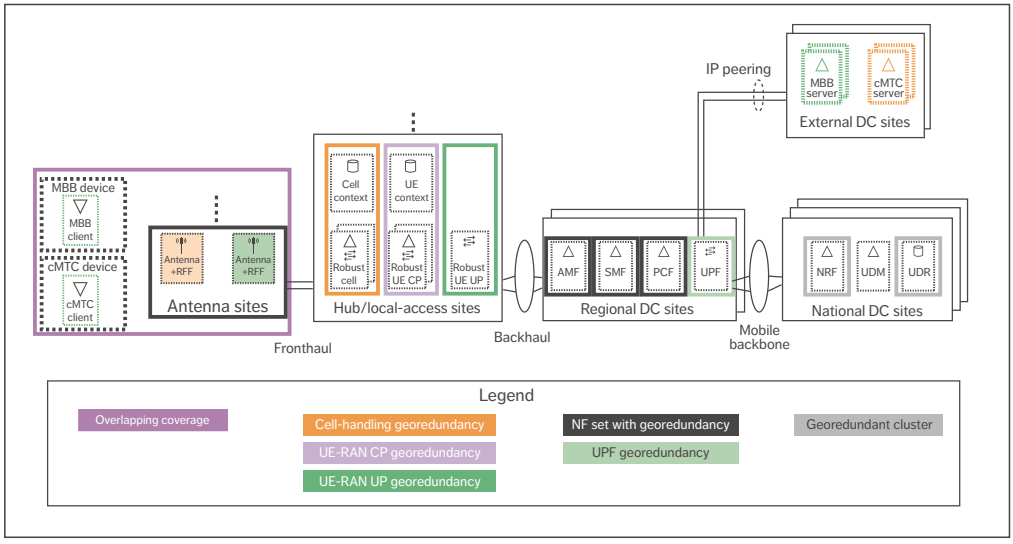


Figure 3 Wide-area network deployment example

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

Solutions at the 5G System level

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

Network architecture deployment examples

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

Wide-area deployments

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

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

In the core network shown in Figure 3, the NF set is used for AMF, SMF and Policy Control Function (PCF) georedundancy. The 5GC UP resilience is

shown as UPF georedundancy. The resilience mechanism for the Network Repository Function (NRF) and the UDR is a georedundant cluster. In the RAN, the new vendor-specific functions for robust UE UP and UE CP, robust RAN resources such as cells and overlapping cells and transmission reception points are shown.

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

Local-area deployments

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

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

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

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

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

Hybrid deployments

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

Terms and abbreviations

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

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

Conclusion

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

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

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

Further reading

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

References

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

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

THE AUTHORS

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

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

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

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

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

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

ISSN 0014-0171
284 23- 3364 | Uen

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