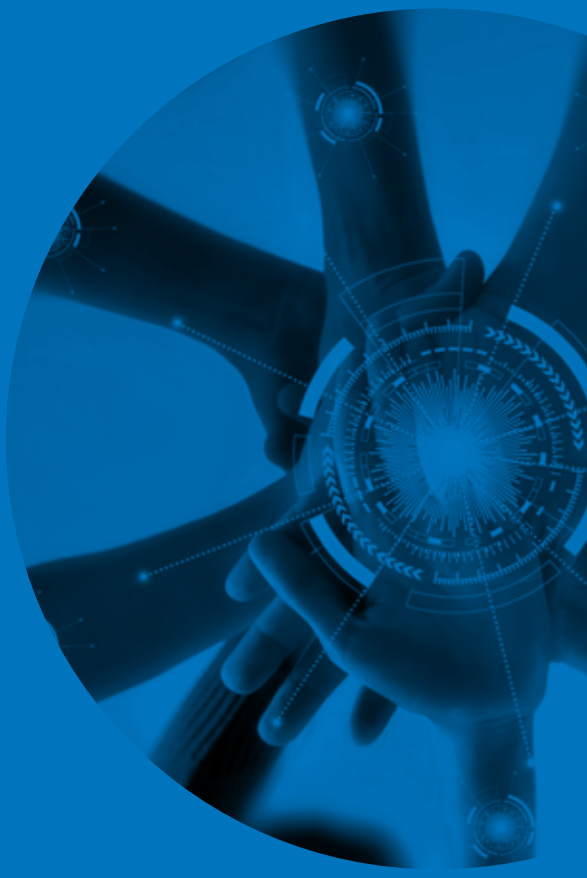
CHARTING THE FUTURE OF INNOVATION | VOLUME 105 I 2021

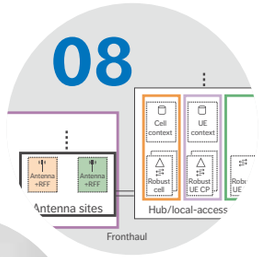
ERICSSON TECHNOLOGY Review

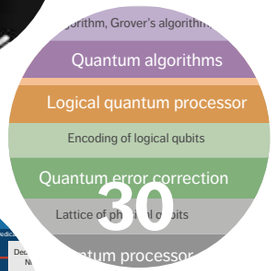
SPOTLIGHT ON

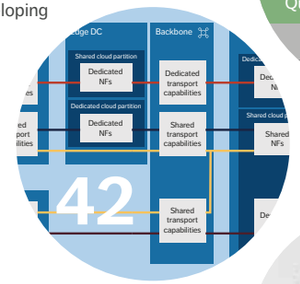
HIGH-PERFORMANCE NETWORKS

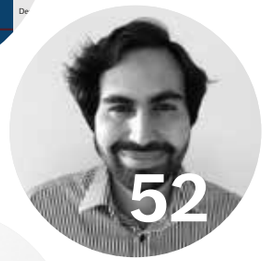


ERICSSON









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

ADDRESS
Ericsson
SE -164 83 Stockholm, Sweden
Phone: +46 8 719 00 00

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

PUBLISHER
Erik Ekudden

EDITOR
Tanis Bestland (Nordic Morning)

EDITORIAL BOARD
Magnus Buhrgard, Magnus Ewerbring, Dan Fahrman, John Fornehed, Kjell Gustafsson, Jonas Högberg, Sara Kullman Johan Lundsjö, Cecilia Nyström, Håkan Olofsson, Patrik Roseen, Anders Rosengren, Robert Skog and Ben Wilmot

ART DIRECTOR
Carola Pilarz (Nordic Morning)

PROJECT MANAGER
Susanna O'Grady (Nordic Morning)

LAYOUT
Carola Pilarz (Nordic Morning)

ILLUSTRATIONS
Jenny Andersén (Nordic Morning)

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

ISSN: 0014-0171
Volume: 105, 2021

SECURING THE FUTURE WITH **HIGH-PERFORMANCE NETWORKS**

ALREADY AT THIS EARLY PHASE in the era of 5G high-performance networks, advanced edge and cloud technologies are opening up massive opportunities for communication service providers (CSPs) across industry segments ranging from manufacturing to retail, banking, health care and travel. Resilient 5G systems that are always available, perform as expected and deliver uncompromised information can and will play a critical role in securing assets and enabling wider economic growth and market competitiveness in the years ahead.

As part of the wider digitalization trend, society is increasingly dependent on workloads deployed on the 5G network platform. Regardless of whether these workloads are mission critical, business critical or best effort, they are all rapidly becoming society critical. Citizen IDs, digital payments, medical applications and educational tools are real-world examples of workloads that have become essential aspects of everyday life for many people today.

A big part of our work at Ericsson right now is using 5G technologies to reinvent the building blocks of enterprise connectivity and compute – delivering the highest possible performance standards and security assurance of mission- and business-critical processes such as factory automation, remote control of assets and more. This new enterprise platform will be characterized by highly resilient and interoperable systems in terms of reliability, availability, robustness, security and privacy.

5G IS DESIGNED TO MEET EVEN THE MOST CHALLENGING REQUIREMENTS ON RESILIENCE, AVAILABILITY AND RELIABILITY

To help our stakeholders better understand the challenges and opportunities that lie ahead in this area, we decided to put together this special issue of Ericsson Technology Review. It addresses the topic of 5G high-performance networks from six different angles: zero trust architecture, robustness in critical networks, network slicing scenarios, extended reality, post-quantum security, and compliance auditing in the cloud.

Since the transition toward zero trust architecture in the latest 5G specifications represents such a major step change for the telecom industry, I think the article on page 20 is particularly worthy of your attention. By combining traditional security principles with policy-based authorization decisions in runtime, a zero trust architecture enables CSPs to both minimize the risk of security breaches and limit damage by preventing lateral movement inside the network if a security breach does occur.

The article about network robustness on page 8 addresses the question of how to cope with the widely varying requirements with regard to resilience, availability and reliability that are cropping up as a result of all the new device categories, use cases and ecosystems encompassed by the digital transformation. Mobile network operators need to have the ability to serve up the required level of network robustness on a case-by-case basis. Since 5G is designed to meet even the most challenging requirements on resilience, availability and reliability, resilient 5G systems make it easy to deliver optimal robustness for mobile networks that are increasingly critical to business and society.

The third article I would like to draw your attention to is the one about the potential threat presented by quantum computers to mobile network security. Although crypto-breaking quantum computers do not yet exist, the possibility of their future existence has prompted action on the part of standards-developing organizations such as the US National Institute of Standards and Technology, the Internet Engineering Task Force and the 3GPP. You can find out about the new post-quantum algorithms that are currently in the final stages of standardization by checking out the article on page 30.

We hope this special issue of our magazine helps you deepen your understanding of the concept of 5G high-performance networks and their enormous potential to transform business and society in the years ahead. Please share it with your colleagues and business partners to help us spread the message as widely as possible. You can find both PDF and HTML versions of all the articles at: www.ericsson.com/ericsson-technology-review

ERIK EKUDDEN
SENIOR VICE PRESIDENT,
CHIEF TECHNOLOGY OFFICER AND
HEAD OF GROUP FUNCTION TECHNOLOGY

ROBUSTNESS EVOLUTION:

Building robust critical networks with the 5G System

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

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

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

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

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

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

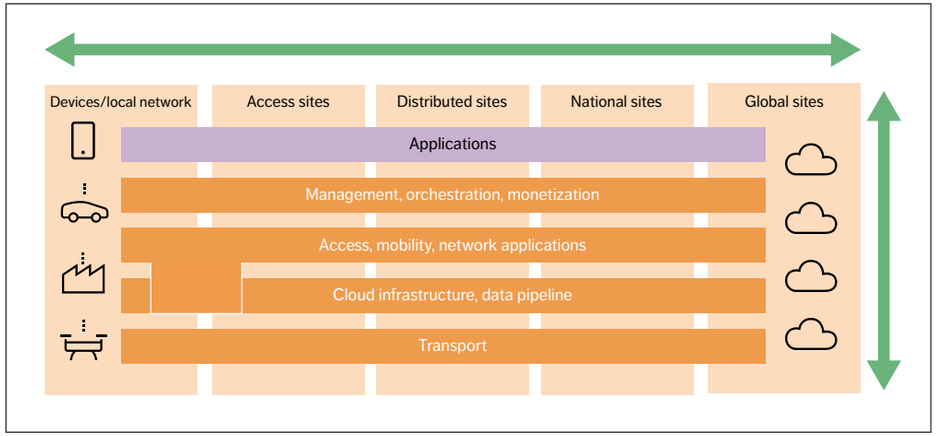


Figure 1 – Aspects impacting network robustness in a typical network

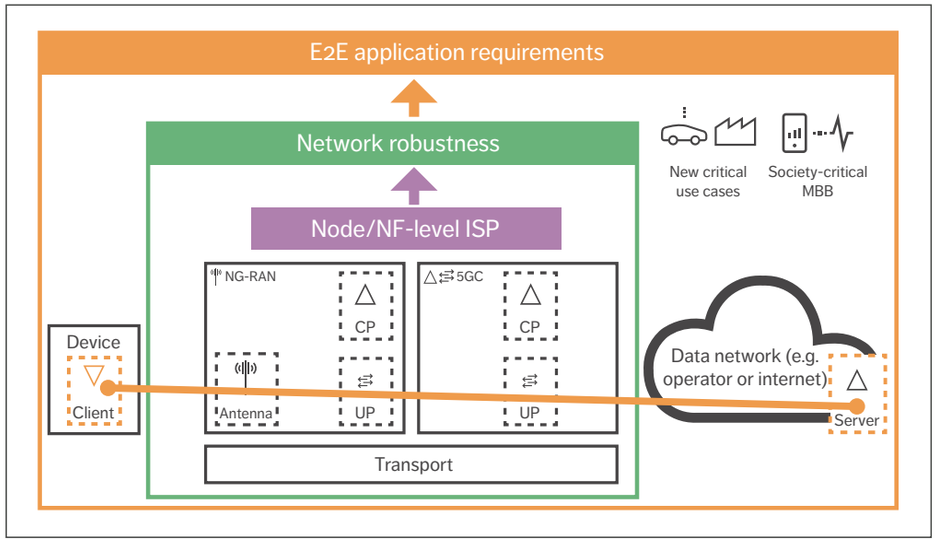


Figure 2 Shifting focus from node/NF-level to network robustness for demanding E2E applications

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

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

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

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

Robustness mechanisms in existing networks

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

Evolved Packet Core aspects

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

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

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

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

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

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

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

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

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

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

LTE RAN aspects

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

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

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

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

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

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

The 5G System robustness toolbox

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

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

5G Core aspects

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

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

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

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

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

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

Features and mechanisms at the NG-RAN level

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

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

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

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

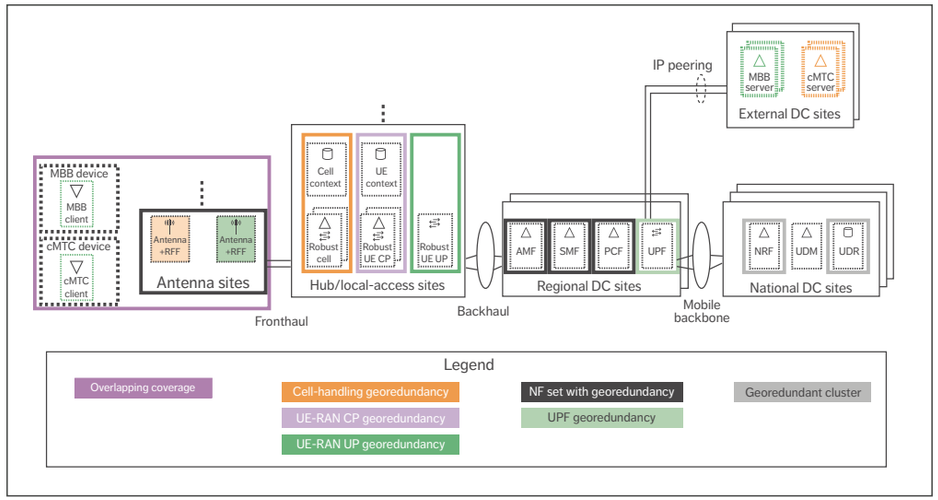


Figure 3 Wide-area network deployment example

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

Solutions at the 5G System level

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

Network architecture deployment examples

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

Wide-area deployments

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

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

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

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

Local-area deployments

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

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

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

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

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

Hybrid deployments

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

Terms and abbreviations

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

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

Conclusion

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

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

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

Further reading

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

References

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

THE AUTHORS

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

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

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

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

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

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

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

5G zero trust

– A ZERO-TRUST ARCHITECTURE FOR TELECOM

The heterogeneous nature of modern telecommunications infrastructure is making it increasingly difficult to protect network resources with conventional perimeter-oriented approaches to network security. By starting from the assumption that the attacker is already inside the network, the zero trust model enhances security by both blocking unauthorized access to network resources and preventing internal lateral movement by an attacker.

JONATHAN OLSSON, ANDREY SHOROV, LOAY ABDELRAZEK, JORDEN WHITEFIELD

The primary aim of any approach to network security is to protect the communication infrastructure so that it can provide services with the expected level of quality, free of disruption. By significantly mitigating risks inside the network perimeter, the zero trust model makes it easier for communication service providers (CSPs) to live up to their security commitments.

■ The perimeter security model operates on the basis of inherent trust, assuming that everything on the inside of a network is trustworthy. As long as the attacker is outside the network and the outer perimeter defenses are strong enough to completely prevent breaches, this approach can work well. But if a breach does occur and an attacker gets inside the network, the perimeter security model allows the attacker to move laterally between systems within the network.

The zero trust (ZT) security model resolves this issue by never making any assumptions about trustworthiness. It first emerged more than a decade ago in the enterprise space, which means the telecommunications sector benefits from the enterprise sector's findings and best practices.

A zero trust architecture (ZTA) works by facilitating secure network access to resources (data, devices and services) that is limited only to subjects (users, devices and services) that are authorized and approved. It is built on an identity-centric approach based on the execution of policy-based authorization decisions in runtime combined with traditional defense-in-depth security principles. When implemented correctly, a ZTA mitigates both the risk of an external attacker getting a foothold in the network as well as the risk of lateral movement, in the case of a security breach.

Ericsson's approach to zero trust architecture applies the ZT principles [1, 2] to telecommunications networks. We have chosen to use the terminology and tenets defined by the US National Institute of Standards and Technology (NIST) SP 800-207 [3] (see highlight box). Several other government bodies and organizations are, however, in the process of publishing ZTA guidance or requirements. The National Cyber Security Centre in the United Kingdom currently has its own ZTA design principles [4]. The NSA and US Cybersecurity and Infrastructure Security Agency's Trusted Internet Connections initiative [5, 6, 7] also aligns with ZT principles.

Built-in support for zero trust architecture in 5G

The 3GPP 5G standards define relevant network security features supporting a zero trust approach in the three domains: network access security, network domain security and service-based architecture (SBA) domain security.

Seven tenets for zero trust architecture

The US National Institute of Standards and Technology has defined seven tenets for zero trust architecture [3]:

T1. All data sources and computing services are considered resources. Devices in a network are heterogeneous and they all interact with the network and software services.

T2. All communication is secured regardless of network location. Trust of a device based on where it is located in a network is not enough. All communication should be secure – that is, confidentiality and integrity must be maintained.

T3. Access to individual [operator] resources is granted on a per-session basis. Trust of devices and services is evaluated prior to granting access. Access is ephemeral and only the minimal set of privileges required are granted for the session.

T4. Access to resources is determined by dynamic policy – including the observable state of client identity, application/service, and the requesting asset – and may include other behavioral and environmental attributes. Clients accessing resources are granted access and permissions based on the client's ascertained state and access rules defined in policies.

T5. The [operator] monitors and measures the integrity and security posture of all owned and associated assets. Trust nothing, verify everything. When a request to access a resource appears, the asset is evaluated. The evaluation of assets is continuous, so as to have an accurate assessment of the threat landscape and risks.

T6. All resource authentication and authorization are dynamic and strictly enforced before access is allowed. Authentication and authorization are always required before accessing any resource for a limited time period and this is continual – that is, reauthentication and reauthorization occurs throughout all transactions where required.

T7. The [operator] collects as much information as possible about the current state of assets, network infrastructure and communications and uses it to improve its security posture. Collected data provide context and insights about where security improvements are needed, such as evaluating access requests, optimizing policy creation and enforcement.

> THE ABILITY TO SECURELY PROVISION, STORE, ACCESS, USE AND REVOKE CREDENTIALS IMPACTS TRUSTWORTHINESS

The network access security features provide users with secure access to services through the device (mobile phone or connected IoT device) and protect against attacks on the air interface between the device and the radio node.

Network domain security includes features that enable nodes to securely exchange signaling data and user data, for example, between radio and core network functions (NFs) [8].

The 5G SBA is built on web technology and web protocols to enable flexible and scalable deployments using virtualization and container technologies and cloud-based processing platforms. SBA domain security specifies the mechanism for secure communication between NFs within the serving network domain and with other network domains.

Key 5G security features that enable zero trust architecture

In our assessment, there are four key security features in 5G that are of most significance in terms of enabling zero trust architectures: secure digital identities, secure transport, policy frameworks and security monitoring.

Secure digital identities

Identities are the new perimeter to defend in ZT security, as they are the primary factor that determines whether access to resources is granted. Secure digital identities consist of two parts. The first part is the identifier (username, fully qualified domain name, serial number) that uniquely identifies a subject or resource. The second part is the credential (password, private key, token) that is secret data used to verify the authenticity of the subject or resource. The use of secure digital identities must be complemented with processes and technologies that enable the secure management of identities and credentials.

In 5G, each and every subject (subscriber or gNodeB, for example) and resource (such as an SBA NF) is uniquely identifiable. Secure digital identities play a fundamental role in building trust and securing communication between entities across security domains. Examples include the digital identity in SIM cards used to authenticate subscribers and network access control, digital identities based on X.509 certificates used for mutual authentication of network devices and NFs, and management user identities for management access control. Secure digital identities enable the creation of an inventory of network assets and are critical to enabling the authentication of subjects and resources to satisfy NIST tenets T2, T3, T4 and T6.

Confidence in the trustworthiness of an identity is determined by the ability to authenticate the asserted identity and the ability to ascertain the integrity of the device being authenticated. The ability to securely provision, store, access, use, renew and revoke credentials impacts trustworthiness.

Identity life cycle management is more challenging for virtual network functions (VNFs) than it is for network appliances. Firstly, the dynamic nature of virtualized deployments – where NFs are instantiated and removed depending on demand – requires secure provisioning, removal and revocation of digital identities in multi-tenant environments. Secondly, consistent exposure and availability of secure hardware across cloud platforms is needed for secure storage and limited access to key material. Secure hardware is used to protect against theft and misuse of secrets, particularly in multi-tenant environments. Further development and maturity in cloud deployments will be required to protect digital identities and attest the system integrity in 5G networks.

Secure transport

The T2 requirement that all communications must be secured is aligned with 3GPP 5G standards that are developed under the presumption of open networks in which all links could be intercepted. Industry-standard mechanisms are used to secure the communication of user and signaling data across 3GPP interfaces.

Data between the user equipment and the radio base station is secured with cryptographic algorithms, providing confidentiality and integrity protection. Additional improvements are introduced in 5G with the Subscription Concealed Identifier (SUCI) to further enhance protection of subscriber privacy against conventional attacks such as passive eavesdropping or active probing of permanent and temporary identifiers.

Communication in transport networks, and between NFs and interconnect networks, is secured with industry-standard security protocols such as (D)TLS 1.2 and 1.3, IPsec, and MACsec, all of which support mutual authentication.

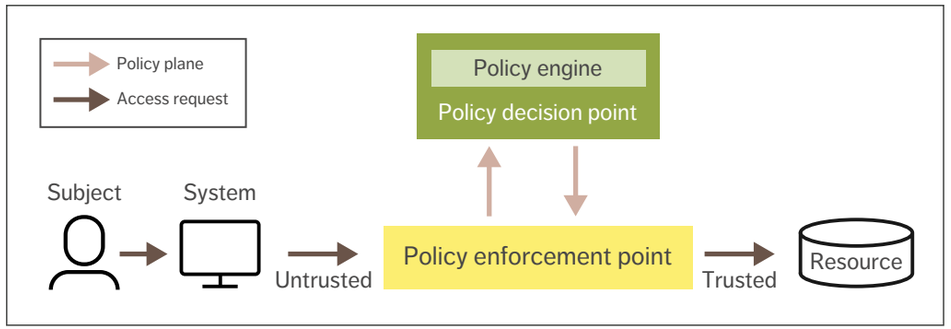


Figure 1 The logical components of the policy framework

Policy frameworks

The relationships and interactions between the many logical and physical entities in telecom networks must be managed to ensure that resources are only accessed by authorized subjects. Policies capture the access rules and requirements to determine the eligibility of a request. These policies are managed, distributed and enforced by a policy framework [3, 9]. This enables the enforcement of micro-perimeters with fine-grained access control based on roles, credentials and environmental attributes.

Figure 1 presents the logical components of a policy framework. The essential logical entities are the policy decision point (PDP) and policy enforcement point (PEP). To access the specific resource, a subject requests permission from the PDP and provides the information needed to perform authentication and authorization.

Policies are created to reflect an organization's processes and acceptable level of risk as well as the sensitivity of the targeted asset. A policy specifies the required level of protection for an object, privileges of a subject and environmental conditions that can change the allowed behavior of the subject toward the object.

The policy engine is part of the PDP. It runs a trust evaluation algorithm to calculate a subject's trust score, which is used to determine whether the subject is allowed to access a resource. The trust algorithm may only use information provided by the subject or it may utilize additional metadata (geographic location of the subject, historical resource usage and behavior).

The PEP is a component that is responsible for setting up a micro-perimeter to protect a resource. Where possible, the PEP is integrated into the resource or placed as close as possible to it, and it forms a logical demarcation point between security zones. The PEP provides access control of connections between the subject and resource based on access control decisions from the PDP.

Policy frameworks are employed in 3GPP-based systems to manage access to resources in different security domains. For example, to gain access to the 5G network services (T1), the user equipment (UE) contacts an Access and Mobility Management Function (AMF) that takes a PEP role. A PDP role can be represented by multiple NFs where Unified Data Management (UDM) and the Policy Control Function (PCF) may be highlighted, among others.

> THE SECURITY POSTURE OF THE REQUESTING ENTITY MUST BE EVALUATED BY DYNAMIC ACCESS CONTROL POLICIES

The AMF transmits the UE's access request to the UDM to validate the UE's identity and trigger authentication and authorization procedures to establish a secure channel (T2, T6). The PCF feeds the AMF with access and mobility policies that may affect UE authorization to access 5G network resources due to, for example, mobility restrictions (T4) [10, 11, 12, 13].

Another example describes how ZT principles apply in 5G SBA. In reference to T1, the SBA identifies NF service consumers and NF service producers. Communication security between core NFs has improved significantly in comparison with previous generations of mobile networks. SBA security specification requires the performance of Transport Layer Security (TLS) based mutual authentication and OAuth 2.0 token-based authorization for any NF that wants to communicate with another NF (T2, T6). The network repository function (NRF) takes the role of authorization server, which makes the NRF act as the PDP.

The introduction of the service communication proxy (SCP) allows indirect communication between NFs. The SCP can take the role of the PEP and provide access control functionality by requesting authorization decisions from the NRF. This makes it possible to implement the zero trust model in the 5G Core, where an NF service consumer (subject) requests access to an NF service producer (resource) through the SCP (PEP), and the NRF (PDP) grants or denies access [10, 13, 14]. With regard to T4, to support decision-making about requested access to resources, the NRF can store additional information, defining the actions allowed for an NF service consumer to specific NF producers [13].

Security monitoring

Security monitoring supports the detection of threats and measuring the security posture of network assets and compliance with security policies. Monitoring and evaluation of subjects, resources compliance, trustworthiness and state are important when deciding whether to permit access to resources.

The European Telecommunications Standards Institute (ETSI) defines security and trust guidance for NFs [15]. With guidelines emphasizing that compliance and state measurements must be continually monitored to effectively evaluate the level of trust of an NF, ETSI's guidance adheres with the principles of zero trust design.

In line with T3 and T4, the security posture of the requesting entity must be evaluated by dynamic access control policies before access is granted to the requested resource. Additionally, to satisfy T5, all owned assets in a telecom network should be monitored and their security posture should be evaluated continuously. These assets include, but are not limited to, devices accessing the network, RAN NFs, core NFs and management functions.

There are different ways to implement a trust evaluation algorithm. Identifying which trust algorithm implementation to adopt depends on two characteristics:

1. How different parameters are evaluated (as binary decisions or as weighted parts of a whole score or confidence level)
2. How requests are evaluated in relation to other historical requests by the same subject.

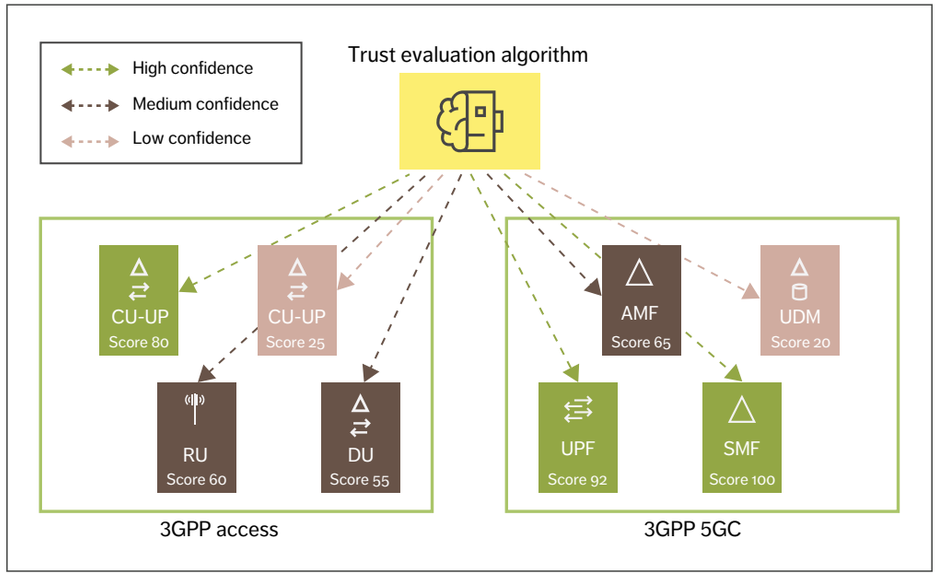


Figure 2 NF trust evaluation

Parameters can be evaluated either based on criteria or score [3]. Score-based evaluation computes a confidence level based on values from every data source, recognizing that there may be various levels of trust between different subjects. Criteria-based evaluation relies on a set of statically configured attributes that must be met before access is granted to a resource or an action is allowed. Moreover, requests can be evaluated either singularly or contextually. Singular evaluation treats each request individually, which risks that an attack can go undetected. Unlike singular evaluations, contextual evaluation takes the subject's history into consideration when evaluating access requests.

The implementation of a trust evaluation algorithm that combines contextual, score-based characteristics would make it possible to offer dynamic and granular access control, since the score provides a confidence level for the requesting account and adapts to changing factors more quickly than static policies.

With respect to T7, there are multiple parameters [15] that can be taken into consideration for evaluating trust that are relevant for telecom networks. Examples include geographical location, NF location, software capabilities (such as patch level, software versions), execution history of an instance, configuration compliance and the appropriate use of encryption techniques.

Future telecom networks should not only consider how to handle trust in subjects, but also trust in resources – particularly in multi-vendor deployments or in cloud where services are provided by a third party.

Figure 2 illustrates how each NF instance may have different trust scores based parameters such as on their current configured state, image version and compliance level. The trust evaluation algorithm in Figure 2 assigns scores in three different ranges on a scale of 0-100: low confidence (<50), medium confidence (51-79) and high confidence (>=80). Based on the evaluation at a certain point in time,

> THE TRANSITION TOWARD ZERO TRUST REPRESENTS A MAJOR STEP CHANGE FOR THE TELECOM INDUSTRY

a score is provided to the NF. If the score falls below a certain threshold, further actions should be taken, such as terminating the NF and replacing it.

Next steps

Telecommunications standards have already evolved the telecom security model by adopting ZT principles that better reflect the security reality facing CSPs. While the latest standards provide improved management of security risks for robust and reliable networks and services, a CSP's ability to fully implement a zero trust architecture will also depend on additional technologies, prioritizations and processes.

Consequently the journey towards zero trust will be gradual with methodical decisions on when, where and how to deploy new security technologies and processes. One of the first important decisions a CSP needs to make is whether the transition should include existing infrastructure and, if so, how to include it with minimal operational and security risk. For example, traditional controls should not be decommissioned until careful evaluation and testing of the new security controls has been completed.

The gradual introduction of zero trust principles, process changes and technology solutions should be driven by risk-based decisions about when and where a CSP wants to invest in modernizing its technologies and business processes. Future challenges include the need to manage the risks of both the infrastructure that has migrated to ZTA and the infrastructure that has not.

A successful implementation of ZT builds on the foundation of effective information security and resiliency practices. While a ZTA can help focus security efforts, it is not by itself sufficient to realize a secure architecture. Rather, a ZTA serves as a cornerstone of a holistic active defense strategy for managing risk, complementing established state-of-the-art information security practices.

Today's concept of zero trust, which focuses on network security, will need to evolve in the years ahead. It will need to expand to tackle the issue of how to address vertical trust from the application, the execution environment and device hardware in cloud environments. This includes measuring the system when instantiating network functions and determining the integrity and origin of software.

Additionally, confidential computing technologies to protect software and data will be critical to protect sensitive assets in shared and distributed environments. Hardware rooted security will be essential to establish a verifiable chain of trust from the hardware to the applications that run on it, as well as protecting data in transit, at rest and in use, to address the risks introduced by hardware and software disaggregation and multivendor deployments.

All of these various technical challenges require further research, development and standardization to fully realize the potential of ZTA for the telecom industry.

Conclusion

The transition toward zero trust represents a major step change for the telecom industry. Ericsson is committed to delivering solutions that enable communication service providers (CSPs) to make that transition as smooth as possible. Fortunately, the new requirements and functionality introduced in the 5G specifications are already aligned with many of the zero trust tenets. It is already evident, however, that further technology development, standardization and implementation are needed in areas such as policy frameworks, security monitoring and trust evaluation to support the adoption of zero trust architecture in new telecom environments that are distributed, open, multi-vendor and/or virtualized.

While various technologies can support organizations in adhering to the guiding principles of zero trust as part of their total active defense strategy, it is important to remember that technology alone will never be sufficient to realize the full potential of zero trust. Successful implementation of a network based on zero trust principles requires the concurrent implementation of information security processes, policies and best practices, as well as the presence of knowledgeable security staff. Regardless of where a CSP is in its transition toward a zero trust architecture, the three pillars of people, processes and technology will continue to be the foundation of a robust security architecture.

> TECHNOLOGY ALONE WILL NEVER BE SUFFICIENT TO REALIZE THE FULL POTENTIAL OF ZERO TRUST

Terms and abbreviations

AMF – Access and Mobility Management Function | **CSP** – Communication Service Provider | **CU-UP** – Central Unit User Plane | **DU** – Distributed Unit | **ETSI** – European Telecommunications Standards Institute | **NF** – Network Function | **NIST** – National Institute of Standards and Technology | **NRF** – Network Repository Function | **PCF** – Policy Control Function | **PDP** – Policy Decision Point | **PEP** – Policy Enforcement Point | **RU** – Radio Unit | **SBA** – Service-Based Architecture | **SCP** – Service Communication Proxy | **SMF** – Session Management Function | **SUCI** – Subscription Concealed Identifier | **TLS** – Transport Layer Security | **UDM** – Unified Data Management | **UE** – User Equipment | **UPF** – User Plane Function | **VNF** – Virtual Network Function | **ZT** – Zero Trust | **ZTA** – Zero Trust Architecture

Further reading

- **Ericsson, 3GPP 5G security overview, available at:** *https://www.ericsson.com/en/blog/2019/7/3gpp-5g-security-overview*
- **Ericsson white paper, Building trustworthiness into future mobile networks, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/building-trustworthiness-into-future-mobile-networks*
- **Ericsson, Future network security, available at:** *https://www.ericsson.com/en/future-technologies/future-network-security*
- **Ericsson, Security, available at:** *https://www.ericsson.com/en/security*

References

1. **O'Reilly Media, Inc., Zero Trust Networks: Building Secure Systems in Untrusted Networks, first edition, 2017, Gilman, E; Barth, D**
2. **Gartner, Market Guide for Zero Trust Network Access, June 8, 2020 (retrieved December 1, 2020) Riley, S; MacDonald, N; Orans, L, available from:** *https://www.gartner.com/en/documents/3986053/market-guide-for-zero-trust-network-access*
3. **National Institute of Standards and Technology, NIST SP 800-207 Zero Trust Architecture, August 11, 2020 (retrieved December 1, 2020), Rose, S; Borchert, O; Mitchell, S; Connelly, S; available from:** *https://csrc.nist.gov/publications/detail/sp/800-207/final*
4. **UK NCSC, Zero trust architecture design principles, available at:** *https://github.com/ukncsc/zero-trust-architecture/*
5. **Cybersecurity and Infrastructure Security Agency, Trusted Internet Connections 3.0 – TIC Core Guidance Volume 2: Reference Architecture, July 2020, available at:** *https://www.cisa.gov/sites/default/files/publications/CISA_TIC%203.0%20Vol.%202%20Reference%20Architecture.pdf*
6. **Cybersecurity and Infrastructure Security Agency, Trusted Internet Connections 3.0 – TIC Core Guidance Volume 3: Security Capabilities Catalog, July 2020, available at:** *https://www.cisa.gov/sites/default/files/publications/CISA_TIC%203.0%20Vol.%203%20Security%20Capabilities%20Catalog.pdf*
7. **National Security Agency Central Security Service, Zero Trust Security Model, February 26, 2021, available at:** *https://media.defense.gov/2021/Feb/25/2002588479/-1/-1/0/CSI_EMBRACING_ZT_SECURITY_MODEL_UOO115131-21.PDF*
8. **Ericsson, An overview of the 3GPP 5G security standard, July 17, 2019, Ben Henda, N; Wifvesson, M; Jost, C, available at:** *https://www.ericsson.com/en/blog/2019/7/3gpp-5g-security-overview*
9. **National Institute of Standards and Technology, NIST special publication 800-162, Guide to Attribute Based Access Control (ABAC) Definition and Considerations, (updated) August 2, 2019 (retrieved December 1, 2020), Hu, C. T; Ferraiolo, D.F; Kuhn, D.R; Schnitzer, A; Sandlin, K; Miller, R; Scarfone, K, available from:** *https://www.nist.gov/publications/guide-attribute-based-access-control-abac-definition-and-considerations-0*
10. **3GPP TS 23.501: System Architecture for the 5G System, Release 16, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144*
11. **3GPP TS 23.502: Procedures for the 5G System, Release 16, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145*
12. **3GPP TS 23.503: Policy and charging control framework for the 5G System, Release 16, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3334*
13. **3GPP TS 33.501: Security architecture and procedures for the 5G system, Release 16, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3169*
14. **Ericsson, Security for 5G Service-Based Architecture: What you need to know, August 21, 2020, Jost, C; Smeets, B, available at:** *https://www.ericsson.com/en/blog/2020/8/security-for-5g-service-based-architecture*
15. **ETSI GS NFV-SEC 003, Network Functions Virtualisation (NFV); NFV Security; Security and Trust Guidance, December 2014, available at:** *https://www.etsi.org/deliver/etsi_gs/NFV-SEC/001_099/003/01.01.01_60/gs_NFV-SEC003v010101p.pdf*

The authors would like to thank Ari Pietikäinen, János Köver, Patrik Teppo, Ilhan Gurel, Mathias Weibull and Antti Jaakkola for their valuable contributions to this article.

THE AUTHORS

Loay Abdelrazek

◆ joined Ericsson in 2019 as a researcher focusing on security concepts in RAN. His research explores new security concepts and technology for RAN, including topics such as air interface security, cloud RAN security and systems security. He holds an M.S. in cybersecurity from Nile University, Giza, Egypt.

Jonathan Olsson

◆ joined Ericsson in 2004 as a researcher for fixed access networks. Since then, he has held roles including standardization coordinator, strategic product manager and security leader in the CTO office. In his current role as RAN security leader, Olsson drives RAN security technology strategy activities in areas such as cloud security, intrusion detection and response, Internet of Things security and trust technologies. He has a B.Sc. in computer science from Uppsala University, Sweden, and is a certified information systems security professional.

Andrey Shorov

◆ is a specialist in security technology at Ericsson Network Security who joined the company in 2019. He identifies key security technologies for the 5G network infrastructure and network slicing. Shorov holds a Ph.D. in computer science from the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences.

Jorden Whitefield

◆ has served as a security technology specialist at Ericsson Network Security since 2019. As an ethical hacker, he performs product security testing on emerging 5G mobile network products, with a focus on platform and operating system security. Whitefield holds a Ph.D. in computer science from the University of Surrey in Guildford in the UK. His doctoral thesis was on the subject of formal verification of security protocols.

Quantum technology and its impact on security in mobile networks

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

JOHN PREUß MATTSSON,
BEN SMEETS,
ERIK THORMARKER

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

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

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

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

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

Risks presented by quantum technology

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

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

Timeline for public-key cryptography and quantum computers

1976 – Diffie-Hellman key exchange

1977 – RSA cryptosystem

1978 – Code-based cryptography

1979 – Hash-based cryptography

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

1984 – Quantum key distribution

1985 – Elliptic curve cryptography

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

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

1996 – Multivariate-quadratic-equations cryptography

1998 – Lattice-based cryptography

1998 – Quantum computer with two physical qubits

2001 – First quantum key distribution network

2011 – Supersingular elliptic curve isogeny cryptography

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

2017 – The NIST announces the PQC standardization program

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

2019 – Quantum computer with 53 physical qubits

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

2022-23 – Target date for draft NIST PQC standards

2024 – Target date for final NIST PQC standards

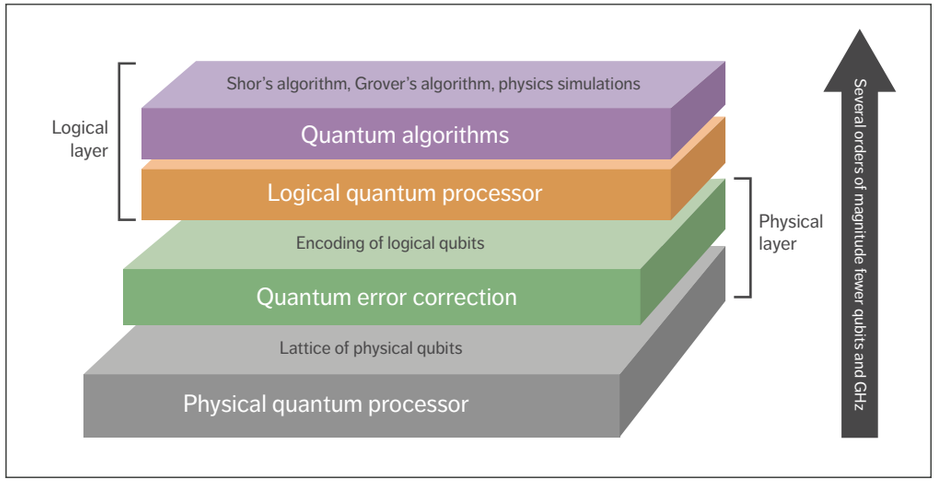


Figure 1 Envisioned structure of future quantum computers

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

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

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

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

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

Migration toward post-quantum cryptography

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

Lattice-based algorithms

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

Potential key encapsulation mechanism and digital signature candidates

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

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

Ericsson's role

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

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

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

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

Post-quantum cryptography algorithm deployment

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

Table A

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

Table B

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

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

Quantum impact on symmetric cryptography

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

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

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

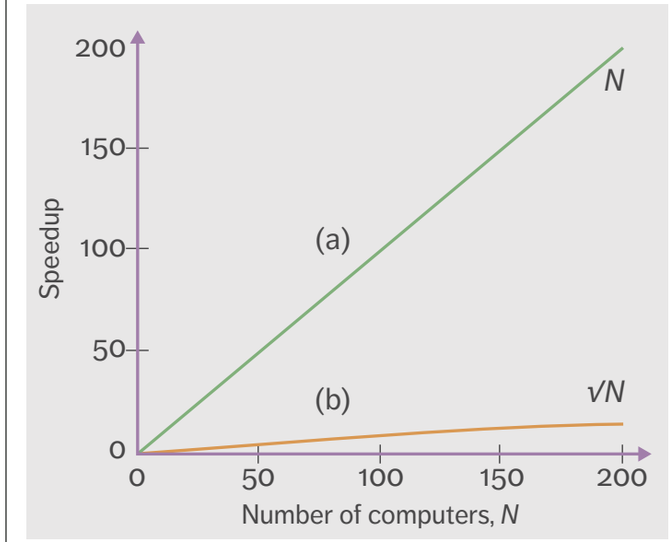


Figure 3 Parallelization of key search using (a) ordinary computers and (b) quantum computers and Grover's algorithm

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

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

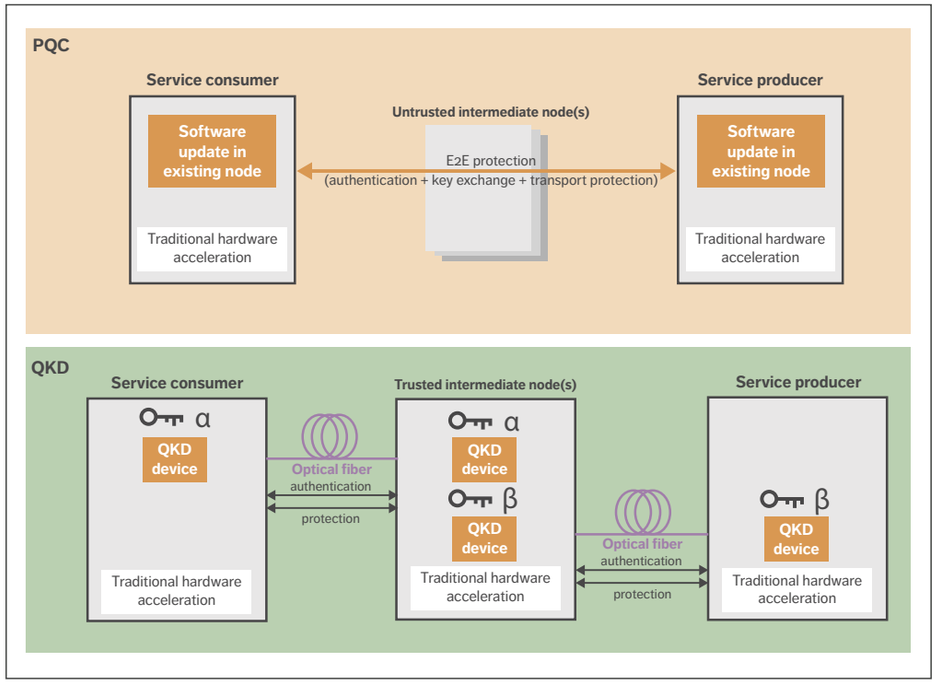


Figure 4 Differences between PCQ and QKD when applied to network infrastructure

Quantum cryptography

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

Quantum key distribution

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

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

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

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

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

Quantum random number generators

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

Terms and abbreviations

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

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

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

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

Conclusion

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

References

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

THE AUTHORS

John Preuß Mattsson

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

Ben Smeets

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

Erik Thormarker

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

Further reading

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

Applied network slicing scenarios in 5G

Network slicing enables new business opportunities across a wide range of use cases and sectors by making it possible to create fit-for-purpose virtual networks with varying degrees of independence. However, the diversity of new commercial and technical requirements has significant implications on how networks are built and managed.

HENRIK BASILIER, JAN LEMARK, ANGELO CENTONZA, THOMAS ÅSBERG

Private 5G networks – the new business offerings that network slicing enables – deliver functionality that extends well beyond that of the current offerings that are typically based on existing public network services. For example, network slicing makes it possible to create a private 5G network with specific service characteristics as well as varying degrees of security/isolation, exposure, self-management, and so on.

There are three main approaches to offering and delivering private 5G networks: the standalone approach, the virtual approach and the hybrid approach. Network slicing enables the customization of system behavior and the isolation of resources/functions for specific services in all three of them.

Standalone private 5G networks are independent, on-premises deployments that have limited interoperability with public networks. They may be sold through a mobile network operator, managed by a customer or provided as a managed service. Network slicing is used to customize the behavior for different use cases/traffic types and to provide isolation between them. A good example of a standalone private 5G network is a dedicated solution deployed at a customer premises such as a manufacturing plant or airport. Although these networks use 5G technologies, they are fully independent and isolated from the public 5G infrastructure.

Virtual private 5G networks, on the other hand, are provided on top of an infrastructure layer that is shared with public services. Network slicing is used to meet customization and isolation needs per use case/traffic type as well as per enterprise customer. Public-safety and connected-car services that make use of the public 5G infrastructure are examples of virtual private 5G networks.

Hybrid private 5G networks are provided by combining infrastructure adopted for public services with infrastructure at a customer's premises. In these networks, network slicing is used to customize and isolate slices per enterprise customer and use case/traffic type. The hybrid approach enables a more flexible distribution of functionality, more efficient use of infrastructure and improved mobility in and out of the customer premises. A good example of a hybrid private 5G network is a manufacturing plant or airport where dedicated on-premises hardware is integrated with and reuses the public infrastructure to improve service and cost-efficiency.

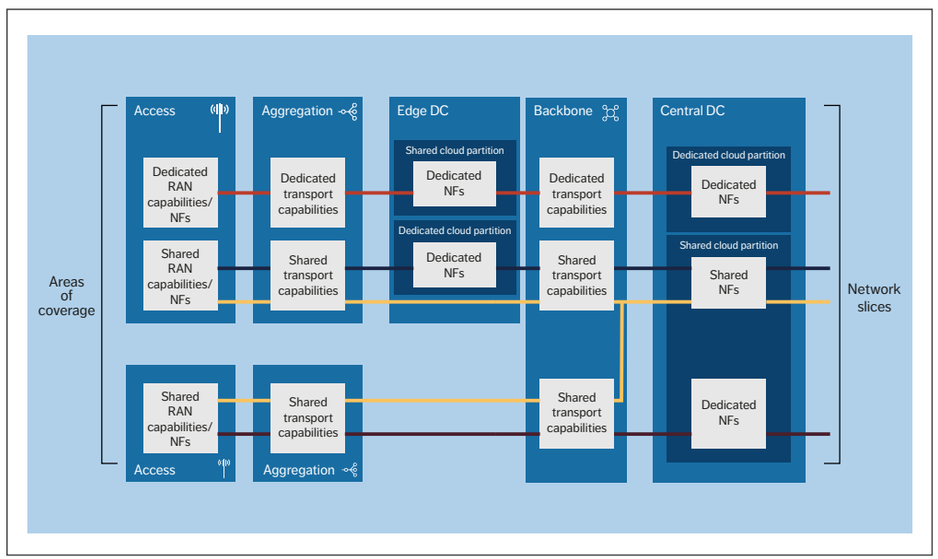


Figure 1 Deployment overview

The enablers of network slicing

Figure 1 provides a deployment overview that illustrates the different ways of composing private 5G networks (network slices). To provide the best level of isolation, resources assigned to a network slice are ideally dedicated. Assuming that it is acceptable, some slices may share resources to reduce cost. Distribution and coverage are considered per slice. Some slices are local, while others may be wider in reach. Some slices require

Terms and abbreviations

5GC – 5G Core | **AMF** – Access and Mobility Management Function | **BSS** – Business Support Systems | **DC** – Data Center | **DNN** – Data Network Name | **DRB** – Dedicated Radio Bearer | **E2E** – End-to-End | **EPC** – Evolved Packet Core | **MBB** – Mobile Broadband | **NF** – Network Function | **OSS** – Operations Support Systems | **PDU** – Protocol Data Unit | **RRM** – Radio Resource Management | **RRP** – Radio Resource Partitioning | **SLA** – Service Level Agreement | **SMF** – Session Management Function | **S-NSSAI** – Single Network Slice Selection Assistance Information | **UP** – User Plane | **UPF** – UP Function | **WAN** – Wide Area Network

APPLIED NETWORK SLICING ✱

> DIFFERENT COMBINATIONS OF ENABLERS WILL BE REQUIRED TO ENGINEER THE APPROPRIATE NETWORK SLICE(S)

local network functions (NFs) – for latency reasons, for example – while others do not.

There is no one-size-fits-all multi-tool for network slicing. In fact, the ability to engineer network slices depends on an evolving toolbox of versatile enablers in five areas: cloud infrastructure, RAN, core, transport, and operations support systems/business support systems (OSS/BSS). Depending on the scenario, different combinations of enablers will be required to engineer the appropriate network slice(s).

Cloud infrastructure

Cloud infrastructure provides great versatility with multiple enablers. The physical infrastructure can be managed, allowing servers, for example, to be allocated to a network slice. Furthermore, the virtualized infrastructure manager enables a set of infrastructure resources to be shared by making use of traditional virtualization. Container-as-a-service provides a more granular and dynamic approach to sharing resources (by using Kubernetes, for example).

The cloud infrastructure is distributed from central data centers (DCs) to the customer on premises and does not have direct awareness of network slices. However, processes can ensure that a cloud platform can fulfill the requirements associated to slices using identifiers of resources. Network slices will have different needs in terms of isolation, distribution and resource guarantees. The cloud infrastructure can provide highly dedicated resources to slices that need it, while other slices share resources. The resources can be optimally used without unnecessary tradeoffs, controlled through orchestration and policies to satisfy the demands of network slices.

RAN

The 3GPP has defined enablers that can be used within a RAN to select appropriate functions and capabilities (such as policies) for network slices. However, the selection of such functions and definition of capabilities relies on implementation. The most important enabler defined in the 3GPP is that of associating each protocol data unit (PDU) session to a slice identifier known as a single network slice selection assistance information (S-NSSAI) as soon as a UE (user equipment) context is created. The RAN reduces a PDU session into dedicated radio bearers (DRBs), which allows the RAN to associate an S-NSSAI to each DRB and to select NFs and capabilities to serve DRB traffic. As an example of capability, a specific next-generation node B central unit user plane (gNB-CU-UP) – hosting PDCP (Packet Data Convergence Protocol) – may be selected for a given S-NSSAI to fulfill delay and security requirements.

Specific layer 1/layer 2 configurations can be tailored for slice policies. A framework for Radio Resource Management (RRM) policies is used to allocate resources and assign QoS levels per slice. For example, the RRM function may use the Radio Resource Partitioning (RRP) capability to allocate a specific partition to a DRB associated to a slice, according to its requirements. Such partitioning may vary depending on slice requirements. Hard partitioning restricts resource usage to a specific slice; soft partitioning allows resources to be used by any slice when they are not utilized by the slice that is nominally assigned to them; shared resources can also be defined for resources accessible by all slices, on demand.

Further, prioritization between DRB traffic can be achieved by means of QoS policies. Such policies allow for the differentiation of DRB traffic within a slice or between slices when shared resources are used.

While it may seem logical to define RRP for each slice supported at the RAN, this is, in practice, suboptimal. There is a tradeoff in performance between the gain of dedicating resources to specific slice services and the overhead in maintaining numerous resource partitions. The balance is to keep sufficient RRPs to guarantee resource isolation per slice where needed, while not impacting radio performance due to excessive partitioning.

Network slice mobility is also enabled by means of signaling between RAN nodes of slice support per tracking area. A mobility function can also take handover decisions on the basis of slice support at the target RAN to achieve radio efficiency while maintaining service continuity.

Core

Core has several enablers, mainly defined by the 3GPP. These enablers make it possible to define dedicated (or shared) user-plane, control-plane or even data-plane NFs at design time that steer the orchestration of the requested network slice at instantiation time. Further, the enablers make it possible to dynamically decide to use dedicated (or shared) user-plane and control-plane NFs based on policy at attach or session requests.

Decisions at design and instantiation times set the context in which the main parameter S-NSSAI (network slice selection assistance information) and secondary parameter Data Network Name (DNN) together with user identities will steer which dedicated or shared NFs that will be used at attach and session requests.

The user-plane function (UPF) is the most valuable NF to dedicate, not only because of the general independency values (optimal redundancy level and no risk of interruption from other services), but also because it ensures low latency through distributed deployment, which makes it possible for the user-data traffic to stay close to customer premises. A dedicated UPF also ensures that established sessions can survive for a period of time, even when the connection to control-plane NFs is lost.

The control-plane NFs are the second-most valuable NFs to dedicate, starting with the session management function (SMF) and followed by the access and mobility management function (AMF). With dedicated distributed SMF and AMF, it is possible to make changes to established sessions and establish new sessions for a period of time, even if connection to data-plane NFs is lost. The final step is to distribute data-plane NFs such as unified data management (UDM) and unified data repository (UDR) to achieve full independence.

Transport

Traffic flows from one network slice (or a group of them) should be mapped into transport resources that match the required Service Level Agreement (SLA) for the slice or group of slices. It is important to take the capabilities and capacity of the transport infrastructure into consideration when selecting which enablers to use.

There are multiple enablers in the transport domains to support network slicing use cases. For many use cases, it is sufficient to rely on QoS mechanisms (compare differentiated services) and IP ranges as the entry point. This involves configuring end points in the RAN and core to map the traffic into preexisting transport capabilities. Additional marking, such as IPv6 flow labels, can be used to enable the observability and mapping of individual slices. The transport network can be configured further to map the slices to transport VPNs with traffic-engineered characteristics. Dedicated transport VPNs can be used to satisfy slices with highly specific requirements.

Coordinated management is essential between the RAN/core and the transport domain to ensure the end-to-end (E2E) SLAs, which may include cross-domain orchestration. The transport VPNs can be tuned and weighed by the assurance capabilities of the OSS. Transport and mobile network capabilities should be harmonized to ensure that mobile network capabilities are not compromised by limitations in the transport network.

Operations support systems/business support systems

The enablers within the OSS/BSS area relate to the management of service characteristics specified by the SLAs included in contracts. This results in a requirement for assurance and analytics capabilities based on business policies, SLA fulfillment and operations. As both the specific KPIs and the approach to monitoring them will differ between slices, there will be a growing need for customization.

Orchestration enables the automation of manual tasks. For example, if an enterprise requires services in a city location, orchestration automatically configures all the cell sites that provide coverage in that location, generates the configurations required to meet enterprise service SLAs and automatically provisions the respective cells by applying the configurations. This process can adapt policies and NF selection based on slice load and the number of slices supported to deliver a solution that can react to operational conditions while fulfilling SLAs. Orchestration provides faster service provisioning across all the nodes through the automation of every step, including instantiation and provisioning of all necessary network functions.

As the use cases, deployment models and business models are diversified, it must be possible to customize and repeat the actions of the OSS/BSS layer, which drives the adoption of a model- and intent-driven approach, where templates and policies dictate the actions. These templates reflect the SLAs and are used to orchestrate the deployment of NFs, the system's capabilities, configuration and policies. This is essential for speed and cost-efficiency.

Monetization and the timely delivery of services are vital. These, together with a need for cost-efficiency, drive demand for automation and flexibility across the OSS/BSS layer. Exposure enablers are required to allow customers to influence and monitor the service.

Network slicing categories

The main differences between network slices have to do with the geographical area within which their services can be reached and their specific service characteristics, particularly in terms of service coverage and enterprise integration depth. Network slices can be loosely grouped into three slicing categories: campus-based, wide-area and limited-area.

> NETWORK SLICES CAN BE LOOSELY GROUPED INTO THREE SLICING CATEGORIES: CAMPUS-BASED, WIDE-AREA AND LIMITED-AREA

In campus-based scenarios, services are mainly consumed locally. In the case of a slice for smart manufacturing plant services, for example, only the base stations covering the plant need to support the plant slices. Core NFs can be deployed in edge DCs within or tightly integrated with the plant, typically together with RAN functions. Depending on the complexity of the particular scenario, either a standalone or hybrid private 5G network is likely to be the best option in these cases. Though not optimal, a virtual private 5G network would also be a possibility.

In wide-area scenarios, services are consumed in a large part of the network. A typical example of this scenario would be services in the transportation/automobile sector that require wide-area coverage. From a radio perspective, this requires RAN slicing to be set up across the entire network. From a core network perspective, it may require core NFs to be deployed in strategically placed edge DCs, potentially together with RAN functions, throughout the network to guarantee the characteristics and performance of critical functions. In these scenarios, a virtual private 5G network is the ideal choice.

In limited-area scenarios, services are consumed within a geographically limited area, such as a sports arena or massive event. All base stations within this area need to be configured to support the slice, and there needs to be a DC close by for critical functions for the core network and potentially for the RAN. A virtual private 5G network is a good fit in these scenarios.

Campus-based scenarios

Campus-based services require tightly integrated solutions such as indoor coverage and local edge DCs to support low latency, data isolation and service assurance during normal operation and fault situations.

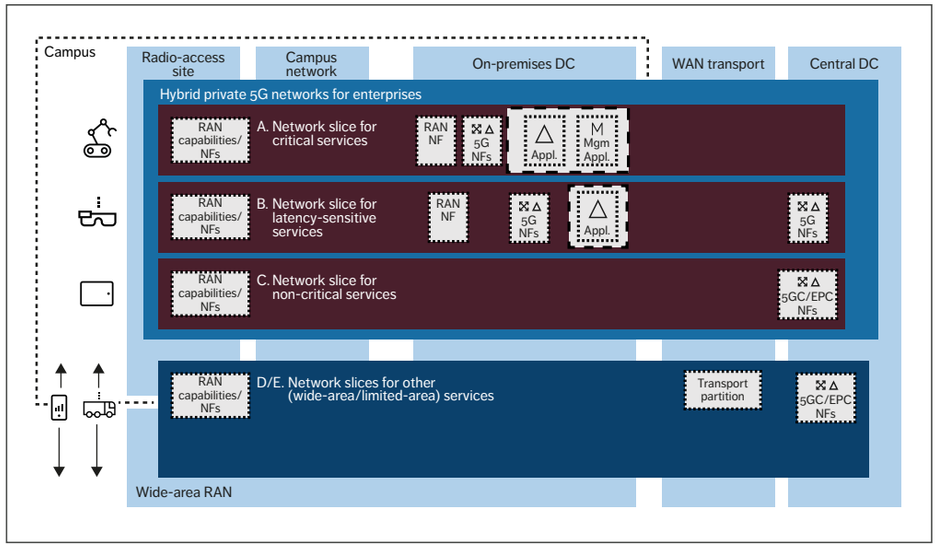


Figure 2 Example of network slicing in a campus deployment (hybrid private 5G network)

Figure 2 illustrates a campus deployment with a hybrid private 5G network that addresses several use cases. A critical slice (A) supports robotics with strong requirements on both latency and survivability. A latency-sensitive slice (B) supports augmented reality/virtual reality application, possibly with lower requirement on survivability. A non-critical slice (C) supports people working at the site (using handheld devices, for example), including mobility across the enterprise sites.

The different types of use cases are supported by one slice each, and each slice is assigned a unique S-NSSAI and DNN. These slices can be assigned dedicated RAN capabilities/NFs to satisfy their unique requirements. For example, services within each slice could use different QoS profiles.

Within the critical service slice (A), hard priority is given (through QoS assignment) to services essential to the smart plant operation. The RAN will then prioritize such essential services in a resource shortage situation. Similarly, RRP in the critical service slice case outlines strict policies for protecting resource utilization for critical services. To ensure survivability, the critical service slice relies on multiple instantiations of RAN and CN functions to increase redundancy. The network slice for critical services will have a complete on-premises core installation allowing services to survive.

The latency-sensitive slice (B) uses QoS profiles, enabling each service to be served even during high load. This will avoid long latencies and service interruptions, possibly at the expense of throughput per service. Most of the functions handling RAN UP traffic are collapsed into one node, while functions governing control plane traffic are more centralized. The core UPF will be deployed locally to ensure that any delays that occur are minimal.

The non-critical slice (C) can have the complete core installed in the central DC. The RRP covers several different services, whose allocation of resources will be regulated by their QoS profile. Such partitions are provided with "soft" borders, meaning that if resources are not used by the slice services, they are reused by other services.

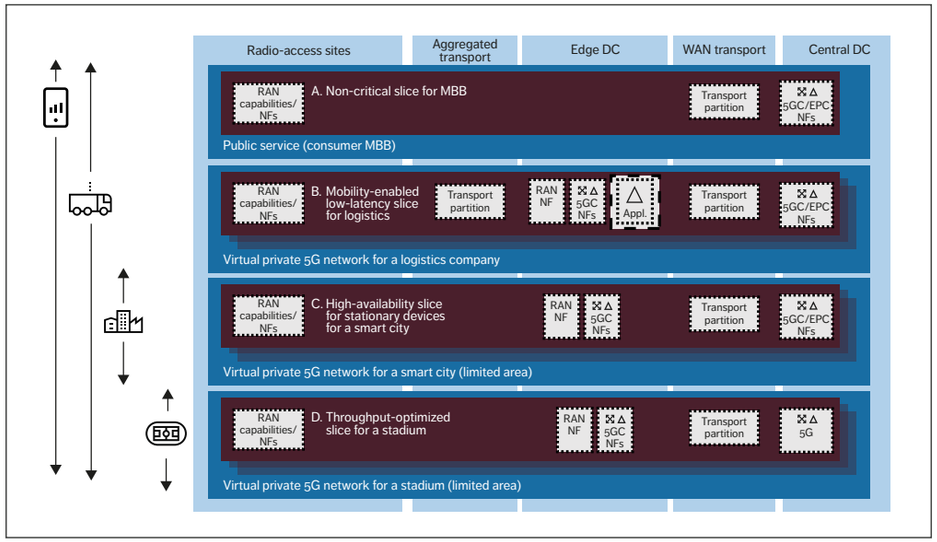


Figure 3 Slicing in wide-area/limited-area deployment (supporting virtual private 5G networks)

As there can be at least as many slices as there are customers, the OSS quickly need to scale to manage a large number of instances. Critical and latency-sensitive traffic would typically stay within the campus, which implies little need to depend on transport slicing for these use cases.

Figure 2 also shows other customers being supported at the site. A slice supporting regular mobile broadband (MBB) service reachable from within the campus (D) allows workers as well as visitors to access the service with full mobility in/out of the campus. Another slice (E) supports virtual private 5G network services for the same enterprise or a different one (such as a logistics company), accessible from within the campus and with mobility in/out of campus.

A specific set of S-NSSAIs and DNNs are used for the MBB slice (D). The core and potentially some RAN functions are deployed in a central DC to provide ubiquitous access. For the wide-area virtual private network (E), a specific set of S-NSSAI and unique DNNs are used to direct to where the enterprise's core is installed, either centralized or on-premises. These slices are configured to allow mobility in and out of the campus. To enable this configuration, slice-based mobility features need to be activated, where appropriate target cell selection ensures slice service continuity. The transport network is critical to ensure the right E2E characteristics for these services.

It is preferable to introduce the various enablers in phases, in response to business needs. One approach could be to start with few campuses and then scale up gradually. It is prudent to start with just a few, static slices and then increase the number of slices and the level of "dynamicity" in terms of orchestration and adaptability later on. It is usually wise to focus on less critical or advanced services first, before moving to more critical ones.

Wide-area and limited-area scenarios

The example in ***Figure 3*** shows a combination of wide area network and limited-area use cases. In this context, network slicing enables virtual private 5G networks. These offerings can be realized as one or several network slices targeting different use cases.

The four use cases shown in Figure 3 – stadium, smart city, logistics company and consumer MBB – each have different requirements in terms of geographical reach and mobility. Network slicing enablers provide each use case with its required functionality and characteristics supporting different levels of SLAs.

A non-critical slice (A) is used for traditional MBB services, providing wide-area service with strong requirements on mobility, interworking with earlier standards, roaming and so on. A mobility-enabled low-latency slice (B) serves the wide-area use case represented by the logistics company in the form of a virtual private 5G network. A high-availability slice (C) is used for stationary devices, serving use cases in a limited area for the smart city use case, as well as in a virtual private 5G network. Finally, a throughput-optimized slice (D) serves use cases in a hotspot for virtual private 5G networks for the sports stadium.

Different use cases and customers may need to be assigned different network slices across the network. The non-critical slice for MBB (A) uses RRP to guarantee bandwidth. Slice-specific QoS regulates resource access among different services. The core network needs to serve full mobility across the coverage but does not require additional distribution.

The mobility-enabled low-latency slice (B) uses RRP available throughout the wide area. It has a dedicated core network with distributed UP and high-availability configuration. Mobility is enabled with robust policies. The high-availability slice for stationary devices (C) uses RRP in a limited area with policies for fairness. The core network is deployed locally without mobility considerations. The throughput-optimized slice (D) also uses RRP in dedicated sites. The core network is optimized for high throughput.

The main challenge for OSS relates to topology. This is due to the fact that even if there are fewer network slices in comparison to the campus case, there may be a large number of sites and areas.

From a phasing perspective, it may make sense to start with a few slices with wide-area coverage and a high degree of sharing, and then evolve to more dedicated, isolated and limited-area ones. It is advisable to start with non-critical use cases, and then move on to latency-sensitive ones later.

> NETWORK SLICING SCENARIOS VARY CONSIDERABLY DUE TO THE DIVERSITY OF USE CASES AND CUSTOMER REQUIREMENTS

Conclusion

Network slicing scenarios vary considerably due to the diversity of use cases and customer requirements. Engineering appropriate slices for each case requires a solid understanding of an evolving set of enablers in cloud infrastructure, the RAN, the core and transport networks and in OSS/BSS. In light of this, we at Ericsson believe that the starting point for pursuing network slicing should be the business needs and use cases rather than the technology behind them. To fully harness the power of network slicing, operators will need to embrace a new and transformational approach to building and operating networks.

THE AUTHORS

Angelo Centonza

◆ joined Ericsson in 2011 after working for several tier-1 telecom vendors in the areas of IEEE/3GPP standardization, telecommunication and defense systems. He is a principal researcher, focused on the areas of RAN automation, network slicing, network architecture and interface design. He also serves as a 3GPP standardization delegate. Centonza holds an M.Sc. in electrical engineering from the Politecnico di Bari, Italy, and a Ph.D. in hybrid broadcast/telecommunication networks from Brunel University London in the UK.

Henrik Basilier

◆ joined Ericsson in 1991. He is an expert in network architecture evolution, focusing on 5G networks and applications and how network slicing can act as a key enabler. He has more than 25 years of experience in the telecom industry across a wide range of technology areas and positions, including packet core networks, cloud technologies and OSS. Basilier holds a M.Sc. in computer science and technology from Linköping University, Sweden.

Jan Lemark

◆ joined Ericsson in 1994 and has worked in technology areas including packet core, user data management, IMS and platforms. He currently serves as a developer in the area of packet core architecture and technology, with a focus on automated orchestration of 5G networks and how to use the possibilities of network slicing. Lemark holds an M.Sc. in electrical engineering from Chalmers University of Technology in Gothenburg, Sweden.

Thomas Åsberg

◆ joined Ericsson in 1987 and currently serves as an expert implementation architect, operation and manitenance (O&M), within the Technology Management department at the Ericsson CTO office. He has held a variety of roles – developer, lead architect, team leader, project manager and line manager – in the areas of hardware and software R&D with systems and architecure evolution, with a focus on the O&M area and value for the user.

Further reading

- **Ericsson, 5G for business, available at:** *https://www.ericsson.com/5g/business*
- **Ericsson, Network slicing, available at:** *https://www.ericsson.com/network-slicing*
- **Ericsson, 5G RAN slicing, available at:** *www.ericsson.com/ran-slicing*
- **Ericsson, What is 5G?, available at:** *https://www.ericsson.com/en/5g/what-is-5g*
- **Ericsson Technology Review, Critical IoT connectivity: Ideal for time-critical communications, June 2, 2020, Fredrik Alriksson, Lisa Boström, Joachim Sachs, Y.-P. Eric Wang, Ali Zaidi, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/critical-iot-connectivity*
- **Ericsson white paper, Critical capabilities for private 5G networks, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/private-5g-networks*
- **GSMA, An Introduction to Network Slicing, available at:** *https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf*

XR and 5G: Extended reality at scale with time-critical communication

The value of 5G extends far beyond the enhanced mobile broadband that more than 1 billion people already have access to through upwards of 150 communication service providers around the world [1]. The time-critical communication capabilities in 5G networks will enable major breakthroughs in a wide range of application areas, including extended reality (XR).

FREDRIK ALRIKSSON, DU HO KANG, CHRIS PHILLIPS, JOSE LUIS PRADAS, ALI ZAIDI

In contrast to enhanced mobile broadband (MBB), the majority of the emerging 5G value-adding applications are time critical in nature with demanding requirements on reliable low latency. Time-critical use cases can be broken down into four categories: real-time media, remote control, industrial control and mobility automation [1, 2].

■ The real-time media category includes innovative extended reality (XR) applications that are of growing interest for consumers, enterprises and public institutions alike. The emergence of time-critical communications over 5G networks makes it possible to offload parts of the XR processing and functionality to the edge cloud and enhance the user experience with lightweight and cost-efficient head-mounted displays (HMDs).

XR use cases

XR is an umbrella term that covers immersive technologies ranging from virtual reality (VR) to mixed reality (MR) and augmented reality (AR). In VR, users are totally immersed in a simulated digital environment or a digital replica of reality. MR includes all variants where virtual and real environments are mixed. AR is one such variant, where digital information is overlaid on images of reality viewed through a device. The level of augmentation can vary from a simple information display to the addition of virtual objects and even complete augmentation of the real world. MR can also include variants where real objects are included in the virtual world.

XR is expected to improve productivity and convenience for consumers, enterprises and public institutions in a wide variety of application areas such as entertainment, training, education, remote support, remote control, communications and virtual meetings. It can be used in virtually all industry segments, including health care, real estate, shopping, transportation and manufacturing. VR is already used for gaming both at home and at dedicated venues, for virtual tours in the context of real estate, for education and training purposes and for remote participation at live events such as concerts and sports.

While VR holds great promise, AR and MR use cases have even greater transformational potential. In VR, the headsets cut users off from their physical surroundings and restrict mobility [3]. With AR, users are present in reality and free to move even when using HMDs. Many smartphone users have already experienced basic forms of AR, through games like Pokémon Go and apps that enable shoppers to visualize new furniture in their homes before making a purchase. AR technology becomes much more powerful, though, when it is used with HMDs. By freeing up the user's hands, AR HMDs transform the user interaction. The ability to have information overlaid on the real world while simultaneously having your hands free has been shown to increase worker efficiency dramatically [4].

“WHILE VR HOLDS GREAT PROMISE, AR AND MR USE CASES HAVE EVEN GREATER TRANSFORMATIONAL POTENTIAL”

XR edge-processing architectures

VR HMDs are already available at scale on the market today, but they are rapidly evolving. Because VR applications are often processing-intensive, enabling high-end VR requires connecting the HMDs to high-end processing units – typically powerful PCs or gaming consoles. VR HMDs with local processing capabilities are also starting to become available, but these are relatively large and heavy and cannot provide the same experience as when off-device processing is used.

AR HMDs are also available on the market today, predominantly for enterprise use [5]. Mass-market adoption will require further progress on ease of use, attractive appearance and content availability [6]. Building fashionable, small form factor HMDs to meet the demands for XR is challenging due to limited processing power, storage, battery life and heat dissipation. We believe that the best way to address these challenges is by offloading parts of XR processing to the mobile network edge.

Terms and abbreviations

5GC – 5G Core | **AAS** – Advanced Antenna System | **AR** – Augmented Reality | **CG** – Configured Grant | **CoMP** – Coordinated Multi-Point | **DAPS** – Dual Active Protocol Stack | **DC** – Data Center | **DL** – Downlink | **E2E** – End-to-End | **L4S** – Low Latency, Low Loss, Scalable Throughput | **HMD** – Head-Mounted Display | **MBB** – Mobile Broadband | **MR** – Mixed Reality | **SPS** – Semi-Persistent Scheduling | **TRP** – Transmission Reception Point | **TTI** – Time Transmission Interval | **UE** – User Equipment | **UL** – Uplink | **VR** – Virtual Reality | **XR** – Extended Reality

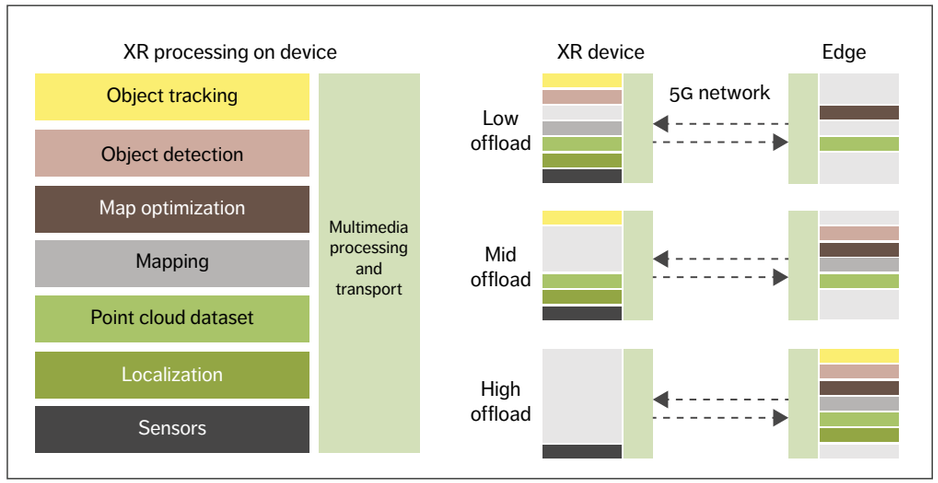


Figure 1 Split architecture options with 5G connectivity

Figure 1 shows the eight main types of XR functionality and how they can be split up between an XR device and the network edge. The major components in XR processing include SLAM (simultaneous localization, mapping and map optimization) [7] with point cloud datasets [8], hand gesture and pose estimation [9], object detection and tracking [10] and multimedia processing and transport. Examples of multimedia processing and transport are rendering, asynchronous time warp [11] and video, audio and sensor encoding.

Figure 1 also illustrates three architectures for splitting the XR processing: low offload, mid offload and high offload. Our internal studies indicate that low offload reduces device energy consumption by threefold while mid offload reduces it by fourfold. High offload reduces device energy consumption by more than sevenfold.

> OUR INTERNAL STUDIES INDICATE THAT HIGH OFFLOAD REDUCES DEVICE ENERGY CONSUMPTION BY MORE THAN SEVENFOLD

In the low-offload architecture, almost all processing is done on the device. The point cloud dataset, spatial map generation and localization occur on the device. As portions of the point cloud and spatial map are built, the point cloud is compressed through multimedia processing and transmitted over the 5G network to the edge, where uploaded spatial map data is created and merged into existing global spatial map data. The object detection and tracking are performed on the device. The rendering can occur locally on the device or at the network edge.

In the mid-offload case, the localization and object tracking functions are performed on the device. The spatial map and point cloud datasets are generated in the network. Key image frames are compressed using a video codec and sent to the network edge to generate the spatial map and point cloud datasets, and to perform object detection. At the network edge, the point cloud datasets and spatial map are created and merged with the global point cloud datasets and spatial map. The overlay rendering occurs at the network edge. The edge-rendered video and audio content is encoded into video and audio streams and transmitted to the device along with the rendering mesh data.

> THE XR CONNECTIVITY REQUIREMENTS DEPEND ON THE LEVEL OF SPLIT ARCHITECTURE AND THE TARGETED QoE

In the high-offload case, only sensor data is sent over the uplink. Sensor data includes camera data. Many AR/VR devices have multiple cameras, including infrared and RGB (red, green and blue) ones. Sensor data also covers sensors such as LiDAR (light detection and ranging) and IMU (inertial measurement unit). The image data is encoded using video compression such as Motion Pictures Experts Group (MPEG) High Efficiency Video Coding (HEVC) or Versatile Video Coding (VVC).

Several compression techniques have been proposed for IMU data, such as delta encoding, linear extrapolation, second- to fifth-order polynomial regression and spline extrapolation [12]. There are numerous techniques for compressing 3D point cloud generated by LiDAR sensors. ISO/MPEG currently has two tracks for point cloud compression standardization under development. These are Video-based Point Cloud Compression (V-PCC) and Geometry Based Point Cloud Compression (G-PCC). LiDAR is covered in the MPEG G-PCC track [13]. The edge-rendered video and audio content is encoded into video and audio streams and transmitted to the device along with the rendering mesh data.

XR traffic characteristics and connectivity requirements

XR traffic is characterized by a mixture of pose and video from/to the same XR device, varying video frame size over time and quasi-periodic packet arrival with application jitter after IP segmentation. Traffic arrival time to the RAN is periodic with non-negligible jitter due to application-processing-time uncertainty. Video frame sizes are an order of magnitude larger and, at the same time, not fixed over time compared with packets in voice or industrial control communication. The segmentation of each frame is expected, which implies that packets arrive in bursts that must be handled together to meet stringent bounded latency requirements.

XR connectivity requirements depend on the level of split architecture and the targeted QoE, leading to a wide range of bit rates and bounded latency requirements. *Figure 2* presents the 5G connectivity requirements for AR, VR and cloud gaming based on the developments in the ecosystem, including the 3GPP [14]. The requirements assume local processing techniques in the split architecture to

Use cases	DL bitrates (Mbps)	UL bitrates (Mbps)	One-way latency (ms)	Frame realibility (%)
Cloud gaming	8-30	~0.3	10-30	≥99
VR	30-100	< 2	5-20	≥99
AR	2-60	2-20	5-50	≥99

Figure 2 Use-case requirements for 5G networks

mitigate consumer latency requirements [15]. Note that latency and reliability requirements are on a video frame (or file) level excluding application error and delay.

For downlink (DL) video traffic, VR typically needs higher bit rates than cloud gaming to support retinal resolution when using HMD and lower compression efficiency due to low-latency encoding. Some AR applications for conversational services can have DL video traffic as VR. However, they potentially have lower resolution to render a video on only part of display, leading to a lower bit rate than VR DL video. In addition, they can have uplink (UL) video streaming for the object detection and tracking, but the bit rate requirements can be lower compared with the DL video. All cloud gaming, AR and VR include pose traffic in the UL, which has much lower bit rates than video traffic, but AR and VR will have higher bit rate requirements than cloud gaming to convey more pose information, such as six degrees of freedom.

AR and VR require more stringent end-to-end (E2E) bounded latencies than cloud gaming since a human is more sensitive to the discrepancy in 3D virtual environments. For instance, it is widely accepted that rendering motion to photon latency greater than 20ms starts to cause nausea when a human wears a VR HMD. Processing techniques such as asynchronous time warp [11] relax the latency requirement to some extent, making VR feasible over 5G networks. For AR, object detection can be performed at the network edge, and object tracking can be done on the device as shown in the low- and mid-offload cases in Figure 1. By performing object detection in the network and tracking on the device, the bounded latency requirement for a 5G network can be relaxed up to 50ms.

Network architecture for time-critical communications

Time-critical communications is an emerging 5G concept for enabling services with reliable low latency requirements such as XR [2]. The aim is to secure data delivery within specific latency bounds (X ms) with the desired reliability level (Y percent). Depending on the user requirements, X ranges from tens of milliseconds to 1 millisecond latency and Y ranges from 99 percent to 99.999 percent reliability. To ensure bounded latencies, the system may have to compromise on capacity, throughput, energy efficiency or coverage.

The 5G RAN, the 5G Core (5GC) and the transport network together with the device contribute to the E2E reliability and latency. E2E latency is the sum of individual latency contributions from every component. E2E reliability cannot be better than the reliability of the weakest link.

Edge deployment of 5GC and applications is key to reducing the transport latency between the application and the RAN. If an application is hosted in a central national data center (DC), the transport network round-trip latency can be in the order of 10-40ms, depending on the distance to the DC and how well the transport network is built out. The transport latency can be reduced to 5-20ms by moving applications to a regional DC or even to 1-5ms for edge sites. For local network deployments with networking functions and applications hosted on-premises, transport latencies become negligible [2].

Achievable RAN latency/reliability performance depends on general deployment factors (such as frequency band, bandwidth, inter-site distances, numerology, duplexing schemes and TDD configuration), RAN and user equipment (UE) capabilities (in terms of hardware and software features) and traffic characteristics (such as data rate and packet size).

5G toolbox to achieve time-critical communications

To ensure that XR applications work well over 5G networks, it is important to separate the XR traffic from best-effort MBB traffic using the 5G QoS framework with optimized QoS flows, as illustrated at the top of *Figure 3*. This enables optimized treatment of XR throughout the mobile network and specifically in the RAN to mitigate the different sources of delay.

The provision of bounded latency for XR and other services requires the prevention of delays and interruptions. The bottom half of Figure 3 presents the comprehensive 5G toolbox for realizing time-critical communications that makes it possible to overcome the following five main causes of delays and interruptions:

1. Congestion
2. Dynamic radio environment
3. Standards/protocols
4. Mobility
5. Device power saving.

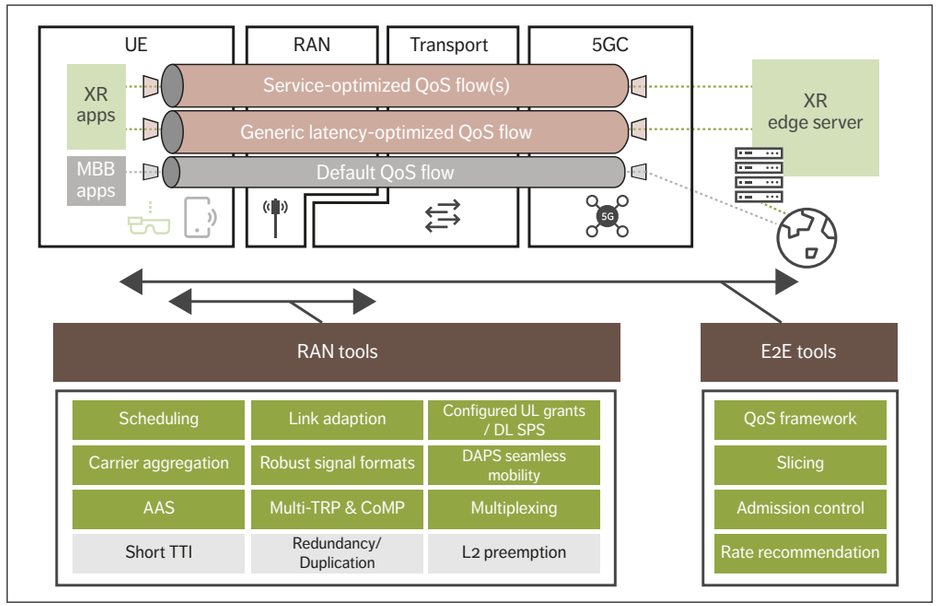


Figure 3 5G toolbox to realize time-critical communications – the tools that are important for XR are marked in green

Congestion

There are several ways to mitigate the congestion-related delays that can occur when end hosts transmit at a higher bitrate than the network can sustain. Fast application rate adaptation is key to avoid congestion for rate-adaptive traffic such as XR and cloud gaming. Low Latency, Low Loss, Scalable throughput (L4S) is an existing method that provides fast indication of congestion from networks that can be applied in the RAN, enabling RAN rate recommendation and forming the basis for a common framework for rate adaptation over 5G [16]. Network slicing and radio resource partitioning together with admission control and latency-optimized scheduling are important tools for reserving network resources for time-critical services and avoiding congestion-related delays to provide a minimum guaranteed bit rate, for example. Network slicing is also beneficial for protecting MBB and other services from resource-hungry, time-critical services.

Dynamic radio environment

Advanced scheduling together with robust link adaptation and signal transmission formats are important tools to combat delays related to the

> THE 5G QoS FRAMEWORK MAKES IT POSSIBLE TO ESTABLISH QoS FLOWS THAT PROVIDE OPTIMIZED NETWORK TREATMENT FOR SPECIFIC FLOWS

dynamic radio environment such as fading/blocking and interference. High-rate XR traffic places new demands on these tools to deliver very high spectral efficiency, while ensuring reliable and timely video frame delivery. Advanced antenna systems (AASs) have tremendous potential to improve the link budget, reduce interference, and increase spatial multiplexing, ultimately leading to more radio system capacity for XR.

Standards/protocols

Features that minimize delays associated with standards/protocols include tools like prescheduling, UL configured grants (CGs) and DL semi-persistent scheduling (SPS). As an example, a combination of periodic CGs and dynamic grants can be used to reduce latency for UL AR traffic consisting of periodic frames of varying size. XR traffic arrival time characteristics call for further optimizations of UL CGs and DL SPS to avoid unnecessary waiting times. Protocol enhancements throughout the different protocol layers, from the SDAP (Service Data Adaptation Protocol) to the physical layer, including the PDCP (Packet Data Convergence Protocol), the RLC (Radio Link Control) and the MAC (Medium Access Control) protocols, can be important for improving the capacity of XR applications as well. For example, control signaling efficiency could be improved to provide grants for multiple radio resource allocations needed for a large XR video frame.

Mobility

Time-critical services place much more stringent requirements on mobility performance than MBB services. For XR services, mobility interruptions need to be well below the inter-frame arrival time (which is typically between 15-50ms) to pass unnoticed. This will require smarter and faster network algorithms as well as stricter processing requirements at the device. 5G New Radio provides a few options for supporting seamless and more robust mobility, including multiple transmission reception points (multi-TRPs), dual active protocol stack (DAPS) handover and conditional handover.

Device power saving

Device power saving is important to support low-power XR devices, and XR traffic arrival time characteristics also call for further optimizations of discontinuous reception.

Traffic awareness in the RAN

One interesting area of future standardization work in the 3GPP is to further optimize 5G performance and capacity by improving XR traffic awareness in the RAN, especially for very short-term traffic variation that may be difficult for an application layer to handle. If, for example, the RAN had the ability to know which IP packets are associated with the same application frame, it could potentially use that knowledge to optimize radio resource allocation, scheduling, link adaptation, packet discarding and other features to increase capacity.

5G QoS approaches for extended reality

The 5G QoS framework makes it possible to establish QoS flows that provide optimized network treatment for specific traffic flows, in addition to the default QoS flow used for MBB. Such additional QoS flows can be established either using 5GC QoS-exposure application programming interfaces to communicate service requirements, or by traffic detection together with pre-provisioned service requirements, such as relying on standardized 5G QoS identifier characteristics. Two complementary 5G QoS approaches can be implemented for XR: a generic latency-optimized QoS flow and service-optimized QoS flows.

RAN deployment	Frequency allocation	Min. bit rate	Max. latency	Frame reliability
Wide area	Mid-band FDD 2x20MHz@2GHz	DL: 8-30Mbps UL: 2-10Mbps	DL: 10-30ms UL: 10-30ms	99%
	Mid-band TDD 100MHz@3.5GHz			
Indoor	Mid-band TDD 100MHz@3.5GHz	DL: 30-60Mbps UL: 10-20Mbps	DL: 10-30ms UL: 10-30ms	99%
	mmWave 800MHz@30GHz		DL: 5-10ms UL: 5-10ms	

Figure 4 Simulation assumptions for 5G RAN

Generic latency-optimized QoS flow

A generic latency-optimized QoS flow that can be implemented in any network is an important tool to enable a large ecosystem of high-rate, rate-adaptive, time-critical applications and services including XR to emerge and hopefully flourish in the same way as MBB and smartphones have done [16]. Implementing such a generic QoS flow enables networks to provide L4S-based early RAN congestion detection and fast rate recommendation to applications as well as packet treatment optimized for low latency and jitter rather than throughput. By not relying on knowledge of specific service requirements it avoids a tight coupling between application and network.

Service-optimized QoS flows

Service-optimized QoS flows can be established for specific XR services with known service requirements in terms of packet delay budgets, packet error rates, minimum guaranteed bit rates and so on, in cases where there is a need to increase coverage or capacity while still providing a good minimum QoE, for example. This requires a tighter coupling between application and network, which adds complexity to the ecosystem but provides QoE beyond what the generic latency-optimized QoS flow can enable and may be required for the most demanding XR applications.

Deployment strategy

To develop insights regarding the 5G network deployment strategy for various XR applications, we have carried out simulation studies for a wide-area deployment and an indoor enterprise deployment. For the wide-area scenario, we evaluated capacity for low- to mid-range XR applications in terms of requirements. For the indoor deployment, we considered high-end applications as well. The simulation assumptions are summarized in *Figure 4*. The 99 percent frame reliability implies less than 0.1 percent packet error rate if each frame is segmented into more than 10 IP packets.

The wide-area scenario is based on a macro deployment in central London with an inter-site distance of approximately 450m. For the mid-band, wide-area deployments, we include an AAS with 32 elements and 16 elements per polarization for 3.5GHz and 2GHz, respectively. The indoor deployment also assumes eight elements and 32 elements per polarization of an AAS for 3.5GHz and 30GHz, respectively. Devices with four receiver

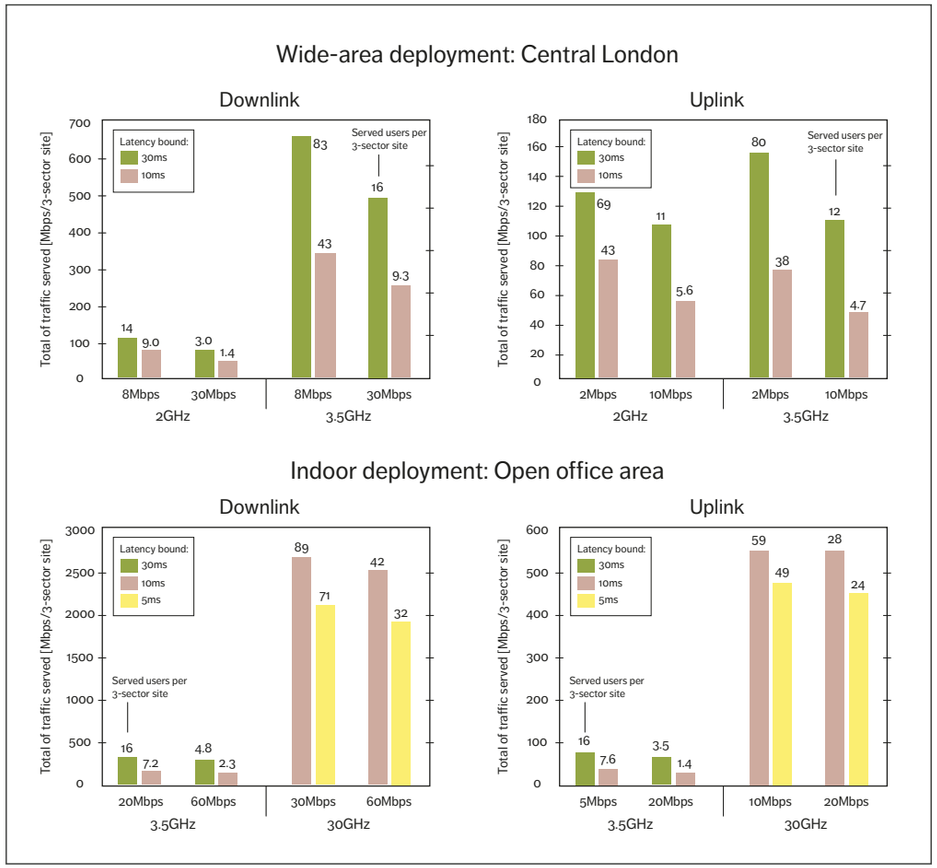


Figure 5 5G RAN-served traffic for various requirements

branches were used in the evaluation. The indoor deployment assumes a 120m by 50m open office area, where four pico sites with three sectors each are placed on the ceiling. For the TDD bands, we have assumed 4:1 DL and UL configuration.

Figure 5 shows the average served traffic per cell (with different combinations of the minimum bit rate and maximum latency requirements per user) for the wide-area and the indoor deployment scenarios assuming 90 percent and 95 percent coverage availability, respectively. From the cell throughput and the minimum bit rate per user, it is possible to derive the maximum number of users served simultaneously per cell, which is shown next to each bar in Figure 5. When the cell capacity is underutilized, the applications can increase the bit rate for superior QoE by rate adaptation.

We can observe that the cell capacity in the traffic served decreases by increasing the minimum user bit rate requirements or by reducing the maximum latency target. Increasing the user bit rate requirement at a given latency budget creates more interference and resource utilization, leading to a smaller amount of total traffic being served. Similarly, when reducing the latency requirement for a given bit rate requirement, a more stringent latency target implies that the resources are available on a shorter time scale, which leads to fewer users being served.

These results show that communication service providers can start to address cloud gaming and low-end AR use cases (for consumers and enterprises) in a wide area with the ongoing 5G network rollout utilizing mid bands, and gradually evolve capabilities and densify deployments to address greater coverage and more demanding requirements in terms of throughput and bounded latency. For users with coverage needs in a small geographic area such as a theme park, industry campus or office, there is an opportunity to support high-end XR with indoor deployments and address more stringent latency and higher bit rate requirements with a local breakout of the core network.

Conclusion

Although extended reality (XR) has the potential to be transformational for both business and society, widespread adoption has previously been hindered by issues such as heat generation and the limited processing power, storage and battery life of small form factor head-mounted devices. The time-critical communication capabilities in 5G make it possible to overcome these challenges by offloading XR processing to the mobile network edge. By capitalizing on their ongoing 5G rollouts, mobile network operators are in an excellent position to enable the realization of XR on a large scale.

Further reading

- **Switch on a better 5G network, available at:** *https://www.ericsson.com/en/5g/5g-networks*
- **5G by Ericsson, available at:** *https://www.ericsson.com/en/5g*
- **How 5G and Edge Computing can enhance virtual reality, available at:** *https://www.ericsson.com/en/blog/2020/4/how-5g-and-edge-computing-can-enhance-virtual-reality*
- **A technical overview of time-critical communication with 5G NR, available at:** *https://www.ericsson.com/en/blog/2021/2/time-critical-communication--5g-nr*

References

1. **Ericsson Mobility Report, November 2020, available at:** *https://www.ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf*
2. **Critical IoT connectivity: Ideal for time-critical communications, Ericsson Technology Review, June 2, 2020, Alriksson, F; Boström, L; Sachs, J; Eric Wang, Y.-P; Zaidi, A, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/critical-iot-connectivity*
3. **Merged Reality: Understanding how virtual and augmented realities could transform everyday reality, Ericsson ConsumerLab, June 2017, available at:** *https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/merged-reality*
4. **Augmented Reality Is Already Improving Worker Performance, Harvard Business Review, March 13, 2017, Abraham, M; Annunziata, M, available at:** *https://hbr.org/2017/03/augmented-reality-is-already-improving-worker-performance*
5. **Augmented Reality Headsets Market, Industry Report, 2019-2025, Grand View Research, November 2019, available at:** *https://www.grandviewresearch.com/industry-analysis/augmented-reality-ar-headsets-market*
6. **Is 2021 finally the year for smart glasses? Here's why some experts still say no, CNBC Evolve, January 23, 2021, Subin, S, available at:** *https://www.cnbc.com/2021/01/23/why-experts-dont-expect-smart-glasses-to-surge-in-2021.html*
7. **AR Products Enhance its Accuracy Through SLAM Technology, PS Marketing Intelligence, February 24, 2021, Shrivastava, D, available at:** *https://psmarketing.home.blog/2021/02/24/ar-products-enhance-its-accuracy-through-slam-technology/*
8. **Point clouds and VR: The future of point cloud visualisation, Vercator blog, April 21, 2020, Thomson, C, available at:** *https://info.vercator.com/blog/point-clouds-and-vr-the-future-of-point-cloud-visualisation*
9. **Pose Estimation Guide, Fritz AI, 2021, available at:** *https://www.fritz.ai/pose-estimation/*
10. **Object Detection and Pose Tracking for Augmented Reality: Recent Approaches, HAL, 18th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), February 2012, Uchiyama, H; Marchand, E, available at:** *https://hal.inria.fr/hal-00751704/document*
11. **The asynchronous time warp for virtual reality on consumer hardware, Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology (pp. 37-46), 2016, van Waveren, J.M.P, available at:** *https://dl.acm.org/doi/pdf/10.1145/2993369.2993375*
12. **Lossless Compression of Human Movement IMU Signals, Sensors (Basel), October 2020, Chiasson, D; Xu, J; Shull, P, available at:** *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7590134/*
13. **3D Point Cloud Compression: A Survey, Web3D '19, Los Angeles, CA, USA, July 26–28, 2019, Chao Cao, C; Preda, M; Zaharia, T, available at:** *https://dl.acm.org/doi/pdf/10.1145/3329714.3338130*
14. **Study on XR Evaluations for NR (RP-193241), 3GPP TSG RAN Meeting #86, 2019, available at:** *https://www.3gpp.org/ftp/TSG_RAN/TSG_RAN/TSGR_86/Docs/RP-193241.zip*
15. **Extended Reality (XR) in 5G (Release 16) TR 26.928, V16.1.0, 3GPP Technical Specification Group Services and System Aspects, December 23, 2020, available at:** *https://www.3gpp.org/ftp/Specs/archive/26_series/26.928/26928-g10.zip*
16. **Enabling time-critical applications over 5G with rate adaptation, Ericsson white paper, 2021, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/enabling-time-critical-applications-over-5g-with-rate-adaptation*

The authors would like to thank Henrik Ronkainen, Christer Östberg, Emma Wittenmark, Oskar Drugge, Yashar Nezami, Lisa Englund, Jonas Ericsson, Per Willars, Kjell Gustafsson, Eric Blomquist, Ben Wilmot, Ann-Christine Eriksson, Håkan Olofsson, Johnny Karlsen, Reiner Ludwig, David Lindero, Ingemar Johansson and Marie Hogan, Héctor Caltenco and José Araújo for their contributions to this article.

THE AUTHORS

Fredrik Alriksson

◆ is a researcher at Development Unit Networks, where he leads strategic technology and concept development within IoT & New Industries. He joined Ericsson in 1999 and has worked in R&D with architecture evolution, covering a broad set of technology areas including RAN, Core, IMS and VoLTE. Alriksson holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Du Ho Kang

◆ is a senior specialist at Ericsson Research who joined the company in 2014. His expertise is in 5G-and-beyond concept developments and performance evaluation toward diverse international standardization and spectrum regulation bodies including 3GPP RAN, ETSI BRAN (European Telecommunications Standards Institute Broadband Radio Access Networks) and the ITU-R (International Telecommunication Union – Radiocommunication Sector). His current interest is developing future RAN concepts for emerging services. Kang holds a Ph.D. in wireless infrastructure and deployment from KTH Royal Institute of Technology.

Chris Phillips

◆ is a master researcher at Ericsson Research and the technical lead for the company's internal XR research project. He has been with Ericsson since 2007. His primary area of expertise is video processing and transport optimization. His latest focus has been in foveated remote rendering for VR, 360 video, cloud gaming and processing/transport optimization for distributed spatial mapping with point cloud datasets. He is also active in the 3GPP SA4, VRIF, SVA and OpenXR organizations. Phillips holds an M.Sc. in computer science from the University of Georgia, USA.

Jose Luis Pradas

◆ is a master researcher at Ericsson Research whose current focus is on RAN enhancements to support XR services in 5G networks. He joined Ericsson in 2007 and has worked in research, performing concept development in architecture and RAN protocols as well as driving 3GPP standardization. Pradas hold an M.Sc. in telecommunication from Universitat Politècnica de València, Spain, as well as an M.Sc. in communications from Helsinki University of Technology, Finland.

Ali Zaidi

◆ is a strategic product manager for Cellular IoT at Ericsson and also serves as the company's head of IoT Competence. Since joining Ericsson in 2014, Zaidi has been working with technology and business development of 4G and 5G radio access. He is currently responsible for Ericsson's radio products for time-critical communication, industrial automation, XR and automotive. Zaidi holds an M.Sc. in innovation management and a Ph.D. in telecommunications from KTH Royal Institute of Technology.

Securing the cloud

WITH COMPLIANCE AUDITING

More and more companies are moving their applications and data to the cloud, and many have started offering cloud services to their customers as well. But how can they ensure that their cloud solutions are secure?

YOSR JARRAYA, GIOVANNI ZANETTI, ARI PIETIKÄINEN, CHIADI OBI, JUKKA YLITALO, SATYAKAM NANDA, MADS BECKER JORGENSEN, MAKAN POURZANDI

Security compliance auditing is an assessment of the extent to which a subject (a cloud services provider or CSP, in this case) conforms to security-related requirements. At a minimum, a CSP must be able to deploy tenants' applications, store their data securely and ensure compliance with multiple regulations and standards.

Many industry sectors – healthcare and utilities, for example – are highly regulated and have to meet stringent data privacy and protection requirements. To serve these types of companies, cloud providers must be able to prove their alignment with the latest standards and regulations such as the Health Insurance Portability and Accountability Act (HIPAA), the Payment Card Industry Data Security Standard (PCI DSS) and the Federal Risk and Authorization Management Program (FedRAMP). Without the right set of tools in place, cloud characteristics such as elasticity, dynamicity and multi-tenancy make proving compliance with such standards both challenging and costly.

Regulations such as HIPAA and PCI DSS define auditing and proving compliance with industry standards and regulations as shared responsibilities. To address users' compliance-related needs, cloud providers must demonstrate evidence of compliance with regulatory requirements across industry segments.

Figure 1 illustrates the cloud security compliance landscape. Providers that can offer tenants credible, trustworthy compliance information on relevant requirements at any time, in a cost-efficient manner, stand to gain a significant competitive advantage.

Auditing security compliance typically involves the manual inspection of regularly generated audit reports and logs, and possibly dynamic tests conducted at runtime. However, applying such techniques in the cloud would be time consuming and costly owing to cloud characteristics.

For instance, to prove network isolation, all layers such as cloud management as well as the virtual network, overlay network, real network (non-virtual), and physical network have to be verified. The results of each verification process on the layers are correlated to avoid any gaps. Current practices such as design document verification, network traffic injection and penetration testing don't work in an environment where tenants share resources, and network parameters change quickly and dynamically.

Operators and cloud providers therefore need a new set of automated tools and techniques that can manage security and compliance, protect consumers' assets, and enable security-related services – in a continuous and cost-effective fashion. In telecom context, the European Telecommunications Standards Institute (ETSI) has proposed an architecture for continuous security monitoring and lifecycle management for network function virtualization to satisfy security requirements at both the operator and consumer level [1].

The ways in which evidence of compliance is provided in the cloud marketplace vary widely at present. It is problematic for a tenant to evaluate cloud providers' capabilities and to understand which party is responsible for what from a compliance perspective. Trust between tenants and their providers is often based on legal texts and disclaimers that can be difficult to comprehend. There is clearly room for improvement, as evidenced by the European Union's call for closer adherence to privacy regulations by global CSPs.

> CLOUD PROVIDERS MUST DEMONSTRATE EVIDENCE OF COMPLIANCE WITH REGULATORY REQUIREMENTS ACROSS INDUSTRY SEGMENTS

Compliance standards in the cloud

To ensure compliance with different security frameworks in the cloud, there are two main types of standards: vertical and horizontal. Horizontal standards are generic standards that are applicable to many industries. Vertical standards are applicable to specific industries. Several standards (horizontal and vertical) have been supplemented

Terms and abbreviations
AICPA – American Institute of Certified Public Accountants | **AWS** – Amazon Web Services | **CCM** – Cloud Controls Matrix | **CCS** – Control Compliance Suite/Services | **CSA** – Cloud Security Alliance | **ETSI** – European Telecommunications Standards Institute | **FedRAMP** – Federal Risk and Authorization Management Program | **GRC** – governance, risk management and compliance | **HIPAA** – Health Insurance Portability and Accountability Act of 1996 | **IaaS** – infrastructure as a service | **ISO 27001** – specification for an Information Security Management System (ISMS) | **ISO 27018** – code of practice for protection of personal data | **NIST** – Network Information Security & Technology | **NIST SP** – Network Information Security & Technology Special Publication | **NoSQL** – not only Structured Query Language | **PaaS** – platform as a service | **PCI DSS** – Payment Card Industry Data Security Standard | **SaaS** – software as a service | **SIEM** – security information and event management | **SOC 1, 2, 3** – Service Organization Controls type 1, 2, 3 report | **SQL** – Structured Query Language | **V&V** – verification and validation | **VM** – virtual machine

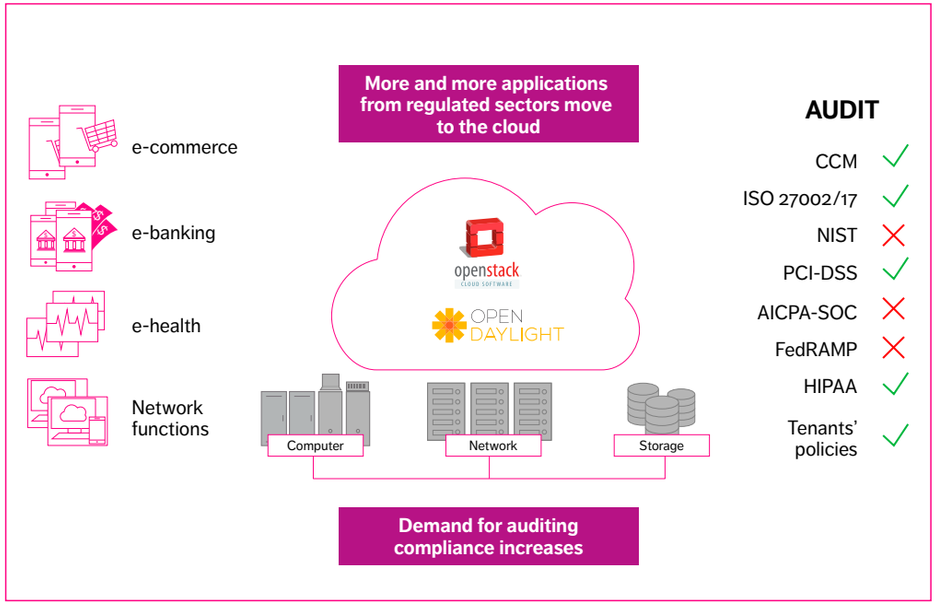


Figure 1
Cloud security compliance landscape with OpenStack as the cloud infrastructure management system and OpenDaylight as the network controller

to guide certification handling in the cloud computing domain.

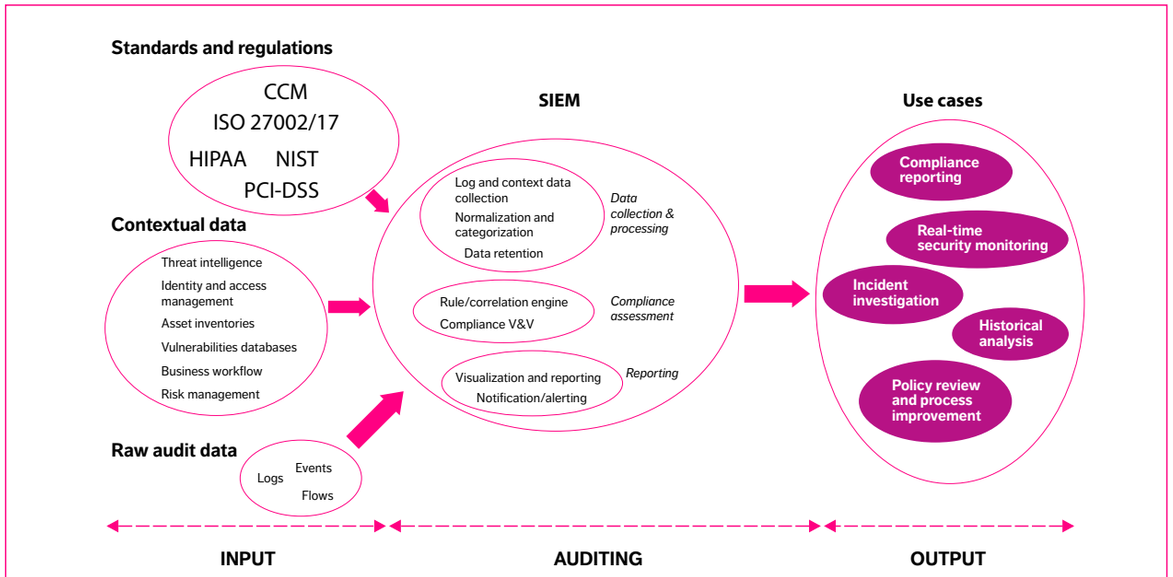
Besides the establishment of horizontal and vertical standards by standardization bodies, other organizations and informal groups such as the Cloud Security Alliance (CSA) address standardization issues related to cloud computing and work on promoting best practices and reaching a consensus on ways to provide security assurance in the cloud. For example, the CSA's cloud security governance, risk management and compliance (GRC) stack [2] supports cloud tenants and cloud providers to increase their mutual trust and demonstrate compliance capabilities.

Current auditing tools

The auditee – in this case the cloud provider or consumer – is required to produce compliance reports to prove that their security measures are protecting their assets from being compromised. Additionally, regulatory bodies require the auditee to retain log data for long periods of time, making it possible for auditors to analyze audit trails and logs. To this end, the auditee can use different types of tools to manage and maintain a holistic view of the security of its environment.

Several open source and commercial tools, including security information and event management (SIEM) and GRC tools, that enable generation of compliance reports on a periodic and/or on-demand basis, exist in the market. *Figure 2* illustrates the main input, output and functionality of an SIEM tool.

In addition to SIEM functionality, GRC [3] tools deliver the core assessment technologies to enable security and compliance programs and support IT operations in the data center.



They enable information security managers to address IT governance, risk and compliance issues by helping them to prevent and respond to non-compliance of security controls while taking into account tolerated risk.

Enterprise class tools

With the advent of the cloud, the makers of several enterprise class tools have proposed integration of their solutions into the cloud environment. While many enterprise-class SIEM engines rely exclusively on correlation to analyze audit data, a new generation of cloud-specific tools includes log search engines and advanced analytics to process the large amount of data and gain security intelligence and knowledge. Nonetheless, most of these tools have been designed to work in enterprise environments whose characteristics differ significantly from the cloud.

Open source projects

Due to the increasing importance of auditing and monitoring in the cloud, open source projects have been created as part of existing cloud management software. For example, OpenStack Congress aims to offer governance and compliance assurance by providing policy as a service. It targets IaaS and does not cover any PaaS or SaaS deployment. Specifically, it allows declare, audit and enforce policies in heterogeneous cloud environments.

A drawback of OpenStack Congress is that it does not allow a full verification through all layers – verification is limited to the information provided by OpenStack services. An elastic stack based on open source tools is another option. This alternative consists of a data search stack that encompasses several components, namely: Kibana for data visualization; Elasticsearch for searching, analyzing and storing data; as well as Beats and

Figure 2
Summary of main SIEM input, output and functionality

Logstash for data collection from various sources in different formats.

When comparing commercial tools with these open source projects, a notable benefit of commercial tools is that they have most of the audit process ready to use out-of-the-box.

Cloud-based services offered by cloud providers

Some cloud IaaS providers are currently proposing partial solutions to help consumers verify that their applications are handled in conformance with their security policies. For instance, AWS offers dynamic customizable compliance checking of cloud resources using AWS configuration rules. Other tools have also been proposed, such as Inspector by Amazon, which is an automated security assessment service that finds security or compliance issues on applications launched within AWS instances.

Cloud-specific tools

Cloud-specific tools such as Catbird Secure offer policy compliance automation and monitoring solutions for private and hybrid cloud environments and focus on software-defined security. Another example is RiskVision Continuous Compliance Service (CCS), an on-demand service allowing providers to gain visibility into their cloud risk exposure and to manage compliance.

Challenges and implementation gaps

A number of challenges make techniques for auditing conventional IT systems unsuitable for use in a cloud environment without significant adaptation. While several common concerns arise when auditing in both domains, a cloud security audit must address unique problems.

Compliance responsibility

Cloud applications run in different deployment models (IaaS, PaaS, or SaaS) and on different types of cloud (public, private or hybrid). This rich set of combinations leads to a complex control dependency and complicates the responsibilities of different actors. The reliance on the CSP varies according to the deployment and type of cloud. For example, in a public IaaS, the hardware and virtual layers are managed by the CSP while the application layer is managed by the tenant. Therefore, there is limited reliance on CSP in IaaS, but most reliance on CSP in SaaS. Thus, it is necessary to define a clear model for the shared responsibility of compliance management.

The massive size of the cloud

The large scale of cloud environments – with the increased number of virtual resources and sources of data – has a direct impact on the size of audit trails and logs. Given the huge amount of data held in them, efficient collection, manipulation and storage techniques are required. Conventional tools were not conceived for large-scale data – they use off-the-shelf fixed-schema SQL databases with a centralized system for the analysis of audit trails. The scale and performance limitations of this type of architecture represent a single point of failure. Auditing and compliance verification tools for the cloud must be designed from scratch to process a very large quantity of data while meeting performance requirements.

The rapidity and dynamicity of cloud services

The speed of events and operations in the cloud constantly changes logs and configuration data. For example, each time a new virtual machine (VM) is created or migrated, new data is generated that may change the compliance status. This is becoming more complex as cloud providers are moving toward more real-time programmable controls by using software-defined networks and NFV in their cloud data centers. One of the major issues in conventional solutions is that they are conceived to execute in a quasi-static environment where auditing is generally performed periodically and remains valid until the next period. They mainly verify a snapshot of the security state at the time of the audit. This is not sufficient in the cloud, where audit and compliance assurance is required each time the infrastructure changes to assess whether these changes give rise to security gaps or infrastructure misuse.

If an audit and compliance assessment tool cannot cope with the high rate of configuration changes for large data centers, it is not fit for the task. Changes in the cloud require the ability to automatically collect data to present near-real-time visibility about compliance to tenants and auditors alike.

Multi-tenancy in the cloud

Audit trails and logs are currently being generated for different actors (tenants, users, cloud provider and so on) on shared physical and virtual layers without a clear separation between them. This approach cannot address all the needs rising from complex use cases such as when a cloud broker leases virtual resources to a third party. Furthermore, it may not be possible for auditing tools to monitor the full stack from the hardware layer up to the application layer because of potential compromise of the privacy of other tenants and of the confidentiality of sensitive information concerning the cloud infrastructure. This is why some providers (particularly SaaS ones) restrict vulnerability assessments and penetration testing, while others limit availability of audit logs and activity monitoring. Most conventional tools are simply not designed to support multi-tenant environments. Therefore, different accessibility schemas must be put in place to give the right access to the common logs for different tenants based on the roles and privileges of different actors.

Privacy protection and GRC support

A CSP with a multi-tenant environment is forbidden to reveal details or metadata that would compromise tenants' privacy or security. Nor is it allowed to disclose any sensitive information to a third party and it must protect against attackers accessing any significant information about the tenants. At the same time, mandated auditors need to access useful and complete information to provide evidence of compliance. In addition, tenants need to receive the right assurances from the CSP and the auditors or perform their own compliance audit of their setting in the cloud, independently of the cloud provider. Therefore, auditing tools should allow for securely outsourcing anonymized logs and audit trails to different interested entities without sacrificing privacy and sensitive information for an evidence-based audit and GRC approach in the cloud.

“ AUDITING AND COMPLIANCE VERIFICATION TOOLS FOR THE CLOUD MUST BE DESIGNED FROM SCRATCH TO PROCESS A VERY LARGE QUANTITY OF DATA WHILE MEETING PERFORMANCE REQUIREMENTS ”

Trust and integrity of audit data

Audited data is often considered to be inherently reliable. But before being presented to the auditor, the original pieces of data will have been passed from the source to the presentation layer via communication interfaces and processed by dynamic software instances. The degree of trust in such a chain is hard to evaluate. Many cloud solutions enable an assessment of the trustworthiness of the hardware platform and bootstrapping of the virtual machines, and safeguard the integrity of log files at rest and in transit. However, audit data would not necessarily be approved as evidence in court if the data integrity had been compromised during any step of the process. The integrity of the audit data source, of the data collector and of the log server should be attestable, assuming that appropriate controls are in place for securing the audit data itself and that there is proof of mutual authentication between the processing elements with an accepted security strength.

> A CONTINUOUS COMPLIANCE VERIFICATION MODEL PROVIDING TENANTS WITH COMPLETE COMPLIANCE VISIBILITY IS KEY TO REDUCING AND LIMITING EXPOSURE TO RISKS

Achieving truly effective auditing in the cloud

In light of the challenges to creating an effective auditing approach in the cloud using the conventional techniques, it is useful to highlight some of the key characteristics of an effective cloud auditing solution.

Continuous monitoring and high automation for compliance

As the cloud is inherently elastic and dynamic, an effective auditing framework must be augmented by continuous compliance and monitoring features [1]. This is not only necessary to maintain compliance but also to improve overall security. It must also provide a high level of automation to cope with quick and transparent changes in collaboration with the cloud management system. Automation is necessary to collect the right information in near real-time and from the right source. Additionally, to enforce an evidence-based compliance verification in a multi-tenant environment, the CSPs should expose information gathered from trusted monitored sources in an open standard format while protecting tenants' privacy by using, for example, anonymization of traces and audit trails for the auditors' and tenants' benefit. Therefore, moving towards a continuous automated compliance verification model that provides complete compliance visibility to the tenants is key to reducing and limiting exposure to risks related to compliance and security breaches.

Building auditing capabilities into the cloud infrastructure

It is much more effective and cost-efficient to build intrinsic auditing capabilities into the cloud infrastructure than to attempt to retrofit existing auditing approaches to the cloud environment. To provide various actors with the necessary audit trails without violating user and tenant privacy, the cloud infrastructure could implement labeling mechanisms to trace the logs to their target tenants. Tackling logs and audit trails in the cloud as opposed to a classical centralized log server in an enterprise environment requires a distributed log collection and retrieval mechanism. Building accountability and traceability into the cloud infrastructure is the best way to provide an efficient and effective auditing solution.

Using analytics for compliance verification

While conventional audit systems specialize in detecting known threats, providing support for identifying unknown threats is a new trend in auditing that is highly relevant to the cloud. Owing to the great quantity of audit data and logs in large data centers, the use of big data analytics based on data mining, machine learning and behavioral monitoring techniques for cloud auditing tools and SIEMs is increasing. In the same vein, storing raw audit data requires new database architecture and technology (such as NoSQL) or support of flat file databases. For the sake of scalability, new deployment options are being considered to move from centralized audit analyses to distributed ones. Analytics must be further explored and improved to tackle cloud-specific characteristics and their actual potential must be investigated in real-world deployments.

Modular compliance approach

Many cloud applications are deployed for highly regulated industries with different compliance needs such as PCI/DSS, HIPAA, ISO 27017 and ISO 27001. These compliance frameworks correspond to different security requirements, which in turn necessitate a large set of controls that must be put in place in the cloud infrastructure.

There are, however, many commonalities between the requirements of all these frameworks in terms of data storage obfuscation, data storage integrity and access control, for example. Therefore, a baseline security requirement needs to be defined to cover the major common requirements. This baseline should be augmented dynamically in the cloud to provide support for different compliance frameworks. Consequently, an efficient auditing approach should be modular, supporting the common denominator requirements as a baseline security requirement and adding different control modules to support specific security frameworks. The CSA CCM compliance matrix is a good starting point for aggregating the major common security requirements.

Application to 5G

5G networks are expected to play a central role in providing a common backbone for information exchange between various applications that belong to different industry segments, which would mean that the security of these applications would depend on the security of the 5G network [4]. This would result in the need to certify 5G networks against all (or at least parts of) the security standards that are related to the served verticals. Implementing isolated network slices for different types of applications would ease compliance assurance by confining certification efforts to each single slice against the appropriate subset of the security requirements. *Figure 3* shows one way this could be accomplished.

Conclusion

The cloud has become a standard in modern computing, and companies in many industry verticals are moving their data to it. Therefore, security assurance, auditing and compliance in the cloud is gaining momentum. Unfortunately, several challenges related to the particular specificities of cloud are limiting the potential benefit of applying current auditing practices and tools.

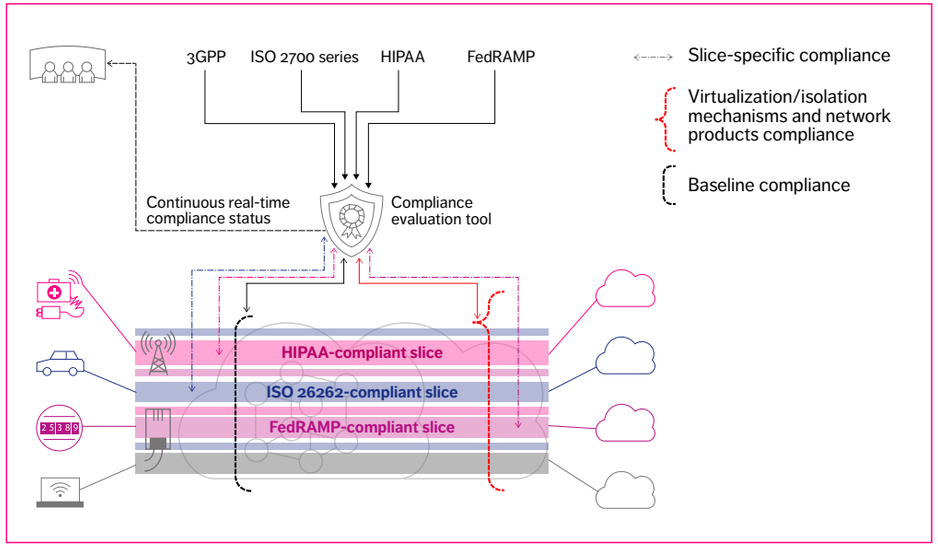


Figure 3
Application to 5G security compliance auditing

Moving toward a continuous automated compliance verification model that provides tenants with complete compliance visibility is key to reducing and limiting exposure to security-related risk. An effective and efficient cloud auditing solution must:

- support large-scale cloud environments
- offer a high level of automation
- allow for near-real-time compliance visibility without compromising stakeholders' privacy and the confidentiality of sensitive data
- fully support multi-tenancy
- provide modular compliance verification to address several standards.

In light of these requirements, new auditing solutions adapted to the cloud environment must be proposed. ❁

Further reading

- **Y. Wang, T. Madi, S. Majumdar, Y. Jarraya, A. Alimohammadifar, M. Pourzandi, L. Wang and M. Debbabi, TenantGuard: Scalable Runtime Verification of Cloud-Wide VM-Level Network Isolation, Network and Distributed System Security Symposium (NDSS 2017), San Diego, USA, February 26 - March 1, 2017, available at:** *https://www.internetsociety.org/sites/default/files/ndss2017_06A-4_Wang_paper.pdf*

- **S. Majumdar, Y. Jarraya, T. Madi, A. Alimohammadifar, M. Pourzandi, L. Wang and M. Debbabi, Proactive Verification of Security Compliance for Clouds through Pre-Computation: Application to OpenStack, 21st European Symposium on Research in Computer Security (ESORICS 2016), Heraklion, Greece, September 28-30, 2016, available at:** *https://link.springer.com/chapter/10.1007/978-3-319-45744-4_3*

References:

1. **ETSI Network Functions Virtualisation (NFV); Security; Security Management and Monitoring specification [Release 3], ETSI NFV-SEC V3.1.1 (2017-02), 2017, available at:** *http://www.etsi.org/deliver/etsi_gs/NFV-SEC/001_099/013/03.01.01_60/gs_nfv-sec013v030101p.pdf*
2. **CSA, CSA Governance Risk and Compliance Stack (V2.0), 2011, available at:** *http://megaslides.com/doc/159998/the-grc-stack---cloud-security-alliance*
3. **David Cau, Deloitte, Governance, Risk and Compliance (GRC) Software Business Needs and Market Trends, 05 02 2014**
4. **Ericsson, 5G Security: Scenarios and Solutions, Ericsson White Paper Uen 284 23-3269, 2016, available at:** *https://www.ericsson.com/res/docs/whitepapers/wp-5g-security.pdf*

THE AUTHORS

Yosr Jarraya

◆ joined Ericsson in 2016 as a security researcher after a two-year postdoctoral fellowship with the company. She holds a Ph.D. in electrical and computer engineering from Concordia University in Montreal, Canada. In the past six years she has produced more than 25 research papers on topics including SDN, security, software and the cloud.

Ari Pietikäinen

◆ is a senior security specialist. He joined Ericsson in 1990 and has worked in the security domain since 2003, most recently with cloud, NFV and IoT security topics. He holds an M.Sc. from Helsinki University of Technology in Espoo, Finland.

Giovanni Zanetti

◆ joined Ericsson in 2010 as a senior security consultant in the IT & Cloud regional unit. His work focuses on security compliance design. He holds an M.Sc. in industrial engineering from Milan University, Italy, as well as CISSP and ISO 27001-22301 Lead Auditor certifications.

Chiadi Obi

◆ joined Ericsson in 2015 as a principal consultant in global IT and cloud services. He has over 19 years of experience centering around information security, the cloud as well as adjacent platforms such as the IoT, with a keen focus on strategy, compliance, governance and privacy aspects. He holds an M.Sc. in information security from Colorado University in the USA as well as industry-driven designations such as the CISSP, CISM and CRISC. He has also authored white papers on cloud and IoT security.

Jukka Ylitalo

◆ is a chief security architect who joined Ericsson in 2001. He has contributed to security standardization work and published several scientific articles during his career. He holds an M.Sc. and a D.Sc. Tech. from Helsinki University of Technology in Espoo, Finland.

Satyakam Nanda

◆ joined Ericsson in 2010 where he worked as a principal consultant in global IT & cloud services until 2017. Over the past two decades, he has served in various leadership roles in consulting, product design, operations and product management driving security strategy and execution for critical infrastructure protection. He holds dual masters' degrees in software engineering and business management from the University of Texas in Dallas, USA.

Mads Becker Jorgensen

◆ is a strategic product manager whose work focuses on the cloud and data platforms area. He has more than 15 years of experience as an information security professional in both the public and private sectors. His current research interests are within secure identity and holistic security.

Makan Pourzandi

◆ is a researcher who joined Ericsson in 1999. He holds a Ph.D. in computer science from the University of Lyon, France. An inventor with 28 US patents granted or pending, he has also produced more than 50 research papers.

ISSN 0014-0171
284 23-3370 | Uen

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