# 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

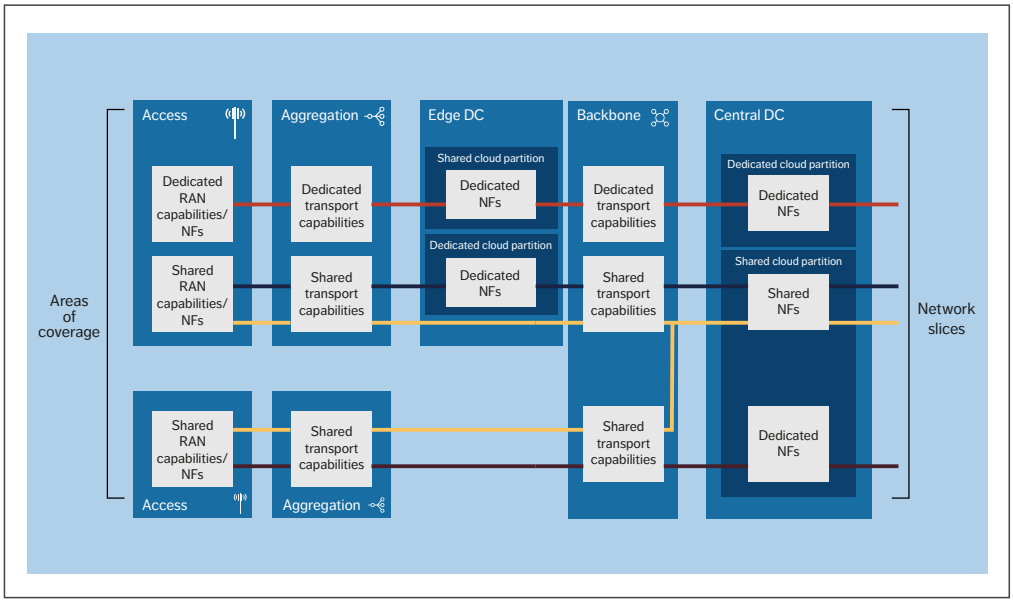


*Figure 1* Deployment overview

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.

## 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

> 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.

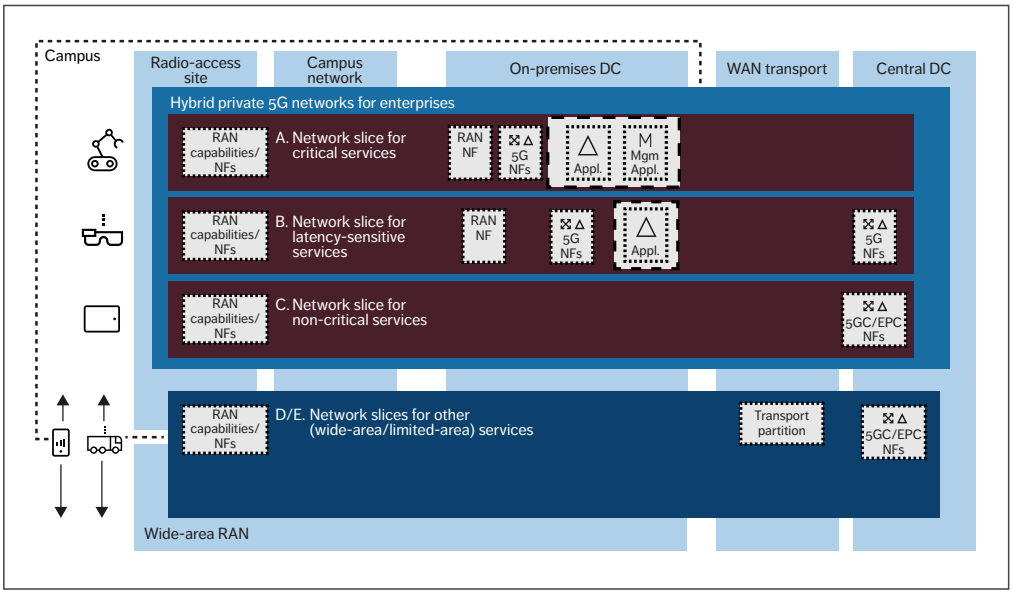


*Figure 2* Example of network slicing in a campus deployment (hybrid private 5G network)

*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.

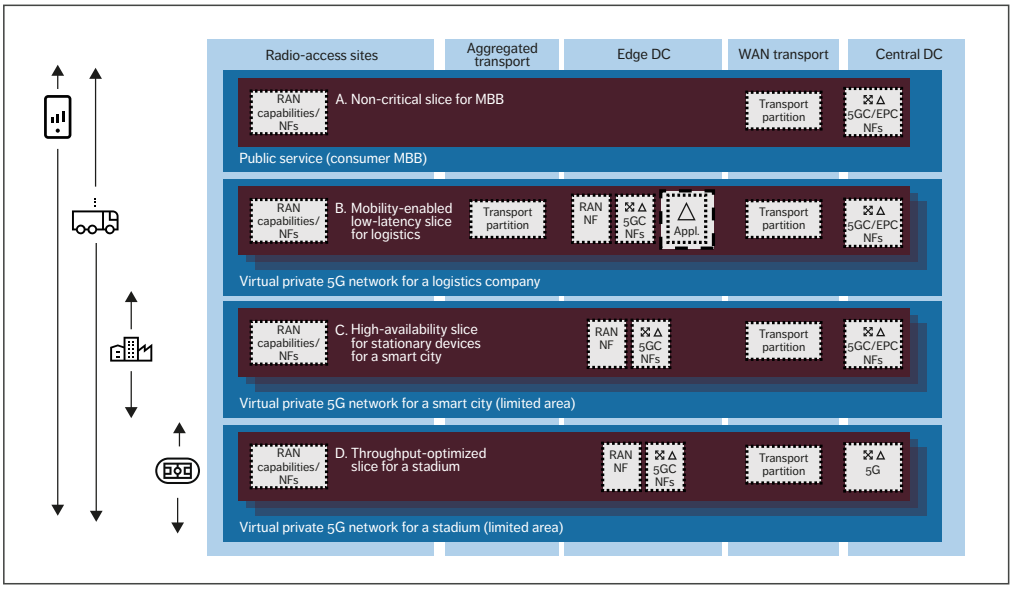


*Figure 3* Slicing in wide-area/limited-area deployment (supporting virtual private 5G networks)

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*