5G RAN slicing

Enabling new services for enterprise and MBB

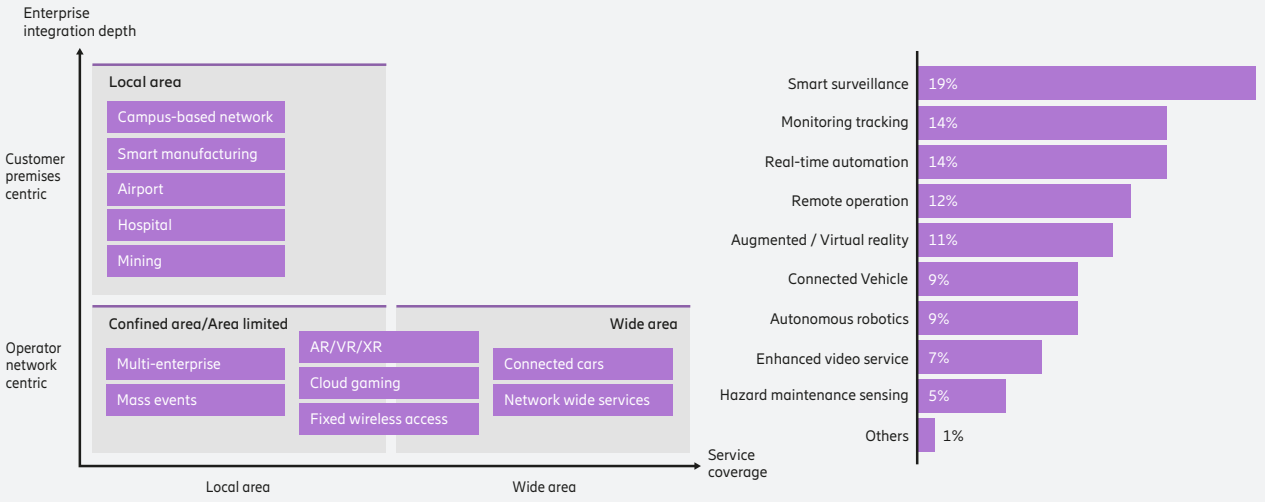
ericsson.com/
5g-access

Network slicing: Game changer for enterprise and MBB services

5G is made for innovation, empowering new services and use cases for consumers, enterprises and society at large. Today's 5G network has been primarily used for mobile broadband (MBB) services for the consumer market. With the combination of 5G and network slicing, service providers can offer new services, such as augmented reality (AR), virtual reality (VR) and cloud gaming, with guaranteed performance to the enterprise and MBB market segment. In doing so, access to potential new sources of revenue, and improved ways to support their customers, will open up.

Few service providers (CSP) have fully explored customized network capabilities and new services to the enterprise and enhanced MBB market segments. 5G can provide access to new sources of revenue in sectors like automotive, healthcare and manufacturing and generate new ways to support customers with enhanced mobile broadband services (eMBB). One way is through end-to-end network slicing. Network slicing facilitates the creation of customized services with automation via logical networks on top of a common shared physical infrastructure. Ericsson is developing a robust 5G end-to-end network slicing solution to enable service providers to offer 'Slicing as a Service' with dynamic orchestration.

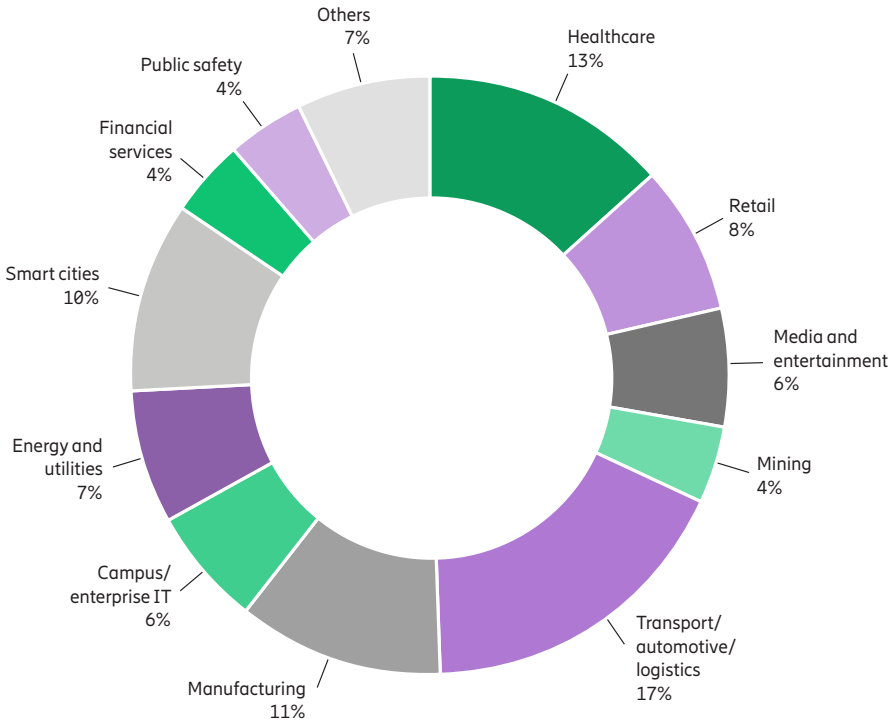
Network slicing use case categories and global service providers' top choices



Service providers are initially focusing on augmented reality (AR), virtual reality (VR), cloud gaming and other MBB-based use cases in the consumer market segment.

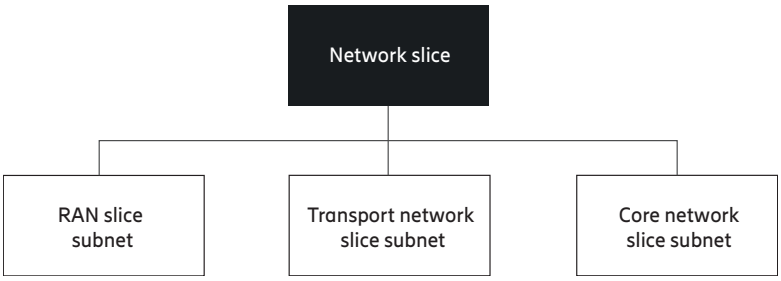
As end-to-end network slicing matures, use cases will continue to grow both in number and complexity. Examples in the enterprise verticals already include smart surveillance, real-time automation and remote operation (as illustrated in the above figure). Verticals may differ from country to country. Strong interest has been observed in tailor-made slices for the financial services sector in certain Asian countries where network coverage varies between locations

The figure to the right provides a snapshot of verticals that service providers expressed interest for network slicing services during a 2020 Ericsson survey of over 35 senior business leaders.

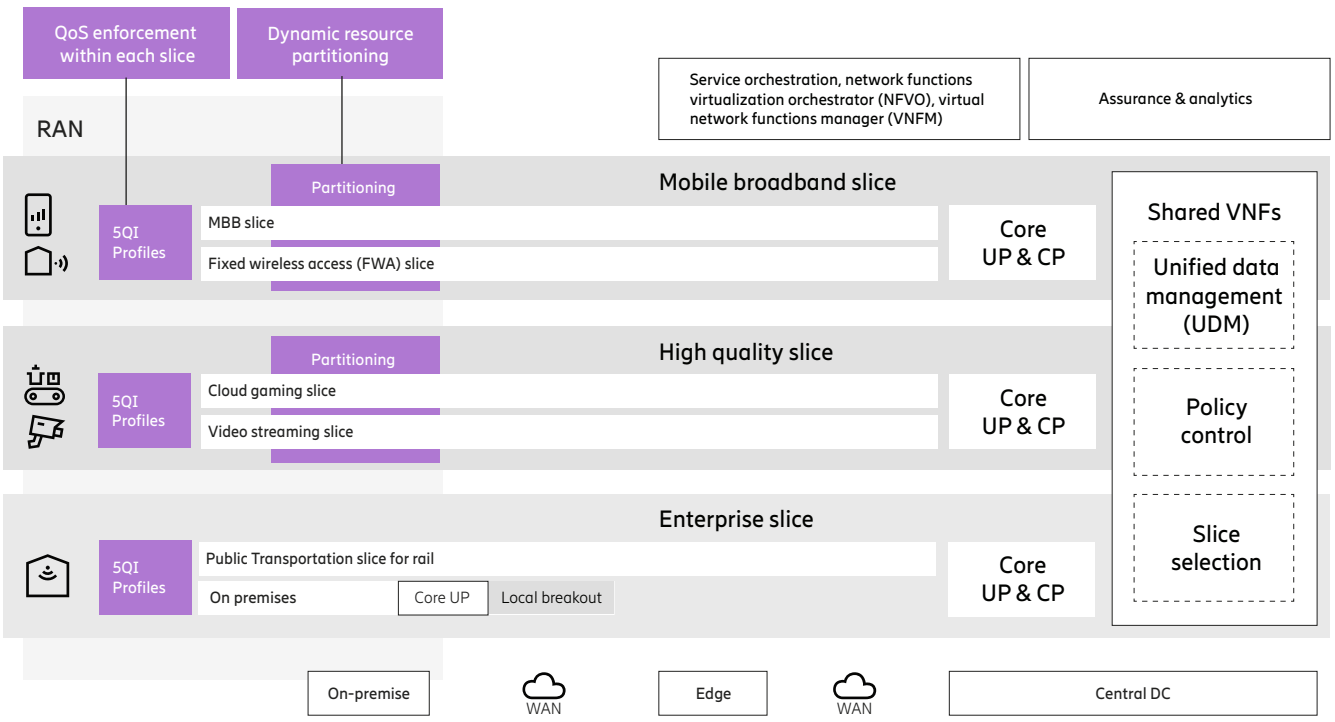


End-to-end network slicing

Network slicing is designed to enable service providers to create logical networks on top of the mobile network by associating user equipment (UE) and communication services with network resources. End-to-end communication service of the network slice is associated with capabilities and Service-level Agreement (SLA) requirements. Service Level Specification (SLS) is related to the SLA between the service provider and their customer. The network slice SLS is further subdivided for the respective slice subnets for radio access network (RAN), transport and core respectively, as shown in the figure to the right.



RAN slicing secures the allocation of limited radio resources to support SLA fulfillment for the associated UEs and services.



Importance of RAN in end-to-end network slicing

Resources need to be carefully managed in the RAN to fulfil SLA of all supported services. As the scale of deployments and number of use cases increases, radio resources become scarcer and the risk of resource starvation increases for services with lower priority. Furthermore, services foreseen to be supported in the 5G system (5GS) are subject to different quality-of-service (QoS) configurations. These include ultra-low latency, high bit rates and improved robustness. Network slicing enables the orchestration of RAN resources, functions and QoS policies in a way that can guarantee fulfillment of SLAs while maintaining access to available resources for different service categories.

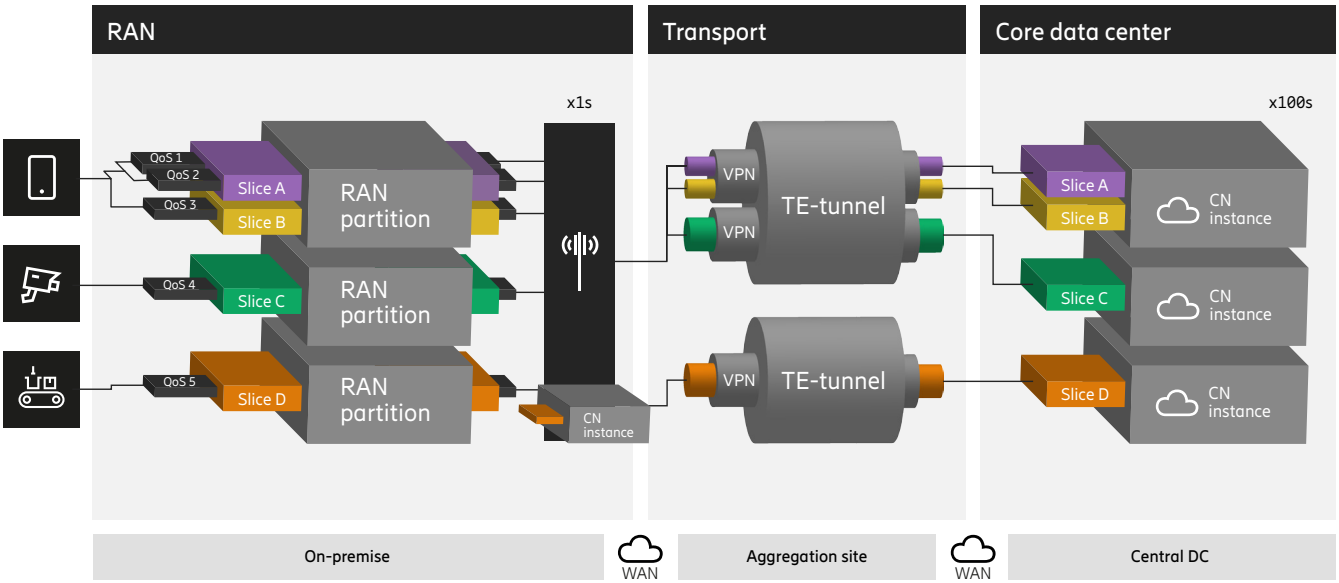
Network slicing is equally important in split RAN deployments where RAN functions can be subject to multiple instantiations and distributed according to QoS demands and infrastructure topologies. A network slice can comprehend RAN functions localized on customer premises to achieve lower traffic latencies, isolation and selection of dedicated RRM policies. They can also consist of centralized RAN functions when resource pooling is deemed beneficial. Network slicing in the RAN also provides a granular level of observability.

It is possible to gather statistics from user terminals and from different RAN functions to reveal the network performance for a given network slice or for the overall RAN.

It is also possible to see the level of resource utilization for each network slice and details of radio coverage and capacity that may influence how services are served per slice. Such detailed observability is crucial for effective network slicing orchestration.

SLA fulfillment

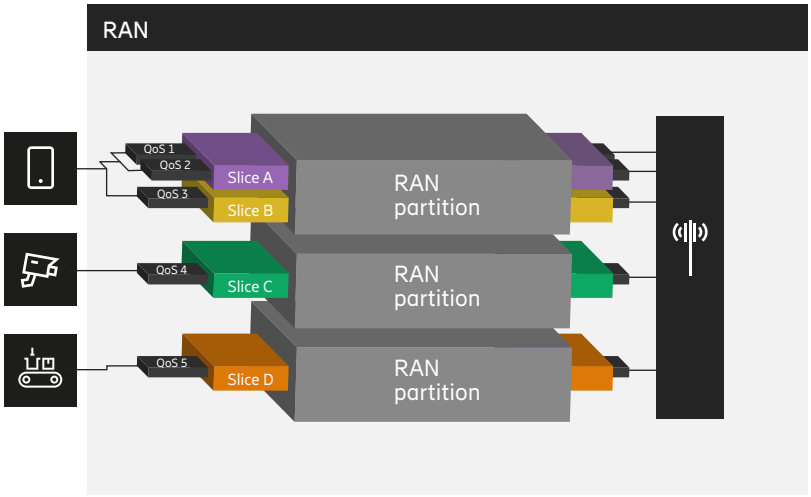
RAN slicing secures the efficient allocation of limited radio resources to support SLA fulfillment.



Ericsson's approach to RAN slicing

From a RAN perspective, end-to-end network slicing provides further means for allocating and prioritizing the limited resources available in RAN. It also offers the possibility of selecting RAN functions in situations with multiple users and groups of users, running multiple services in accordance with the objectives of the operator.

On a high level, one or many end-to-end network slices may be viewed as being represented by a RAN partition, as exemplified in the figure on the right. This is realized by end-to-end network slicing awareness in key functionality areas in RAN, including observability, radio resource management, user plane, transport network traffic management and mobility.



Radio resource management

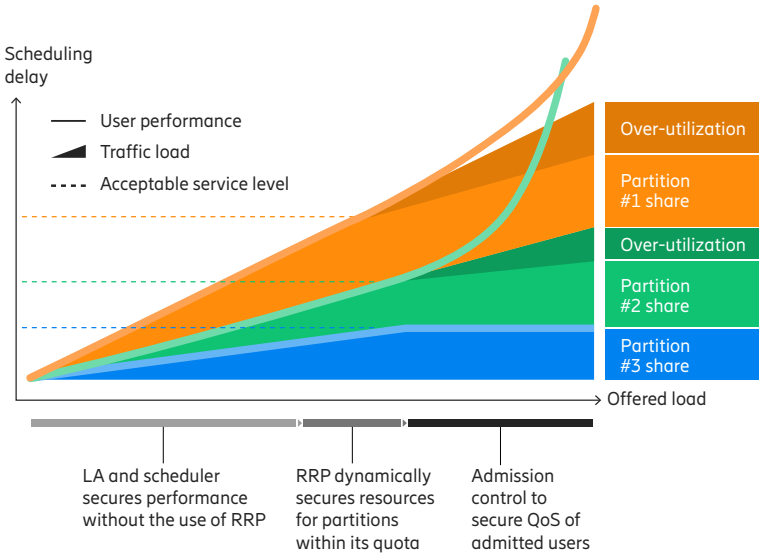
Slice-aware RAN QoS implementation enables service providers to create differentiation in their networks both between subscribers and between services. It also allows them to allocate and supply adequate network resources (including both spectrum and hardware resources) based on subscriber and service requirements.

Important functional areas in the QoS implementation are link adaptation (LA) and scheduler configuration, radio resource partitioning and admission control. With slice awareness in these areas, the available resources can be allocated according to the traffic conditions to meet the service performance requirements for different traffic categories.

The link adaptation and scheduler are fundamental to meeting service requirements with efficient use of spectrum resources. The slicing framework provides full flexibility to configure the desired connection handling, such as scheduling priority and scheduling strategy, for the QoS flows independently for each end-to-end network slice.

At a very high load, admission control provides the scheduler with sufficient resources to secure QoS of all admitted users. Slice awareness in admission control is used to differentiate admitted users based on end-to-end network slicing.

In addition to admission control, radio resource partitioning isolates the performance between different traffic groups and secures that the performance of each traffic group meets SLA requirements. The SLA needs to be described according to the dimensioning of the partition.



Dynamic RRP enables resources to be dynamically shared between different slices without statically reserving them. Any free resources can be used by other slices to avoid performance degradation in unloaded conditions by deploying Dynamic RRP.

Network slicing should ensure service continuity when UE moves between different network domains. The most evident example of such scenario is UE mobility between the 4G and the 5G systems.

It is likely that 4G and 5G coverage areas will be intertwined and network slicing solutions should be able to deliver comparable service QoS during such mobility. In order to achieve this, network slicing solutions can rely on close 4G/5G interworking function of 4G and 5G by which a UE can move between the 4G and the 5G systems while being served by networks functions shared by both systems and without the need to relocate the UE context.

Network slice policies, available on a per S-NSSAI basis in the NG-RAN, can be mapped to equivalent RAN policies in the E-UTRAN, which are identified by means of parameters such as the Subscriber Profile ID for RAT/Frequency priority (SPID) and the PLMN ID. Such mechanism guarantees full UE inter system mobility, while ensuring consistent support of network slice services.

RAN slicing observability

Observability per-slice serves two key purposes. The first is to understand resource utilization and service performance to optimize the RAN slice configuration and operation. The other is to provide service performance indicators, traffic load, and radio conditions per slice to gauge SLA conformity for the respective end-to-end network slices.

Performance management supports observability on a per S-NSSAI basis.

Slicing handling in gNB

Network slicing enables service providers to assign specific policies at user-plane (UP) level based on the network slice serving a given radio bearer.

When a dedicated radio bearer (DRB) is created, the RAN will assign it to a specific network slice as per configuration instructions received from the 5G core. As such, DRB will be able to carry different QoS flows, each with different configurations. Network slicing allows optimal aggregation of QoS flows within the same DRB, to ensure that SLAs of all services served by the DRB are fulfilled while ensuring an efficient usage of RAN resources.

The network slice assigned to a DRB can also influence the selection of RAN functions that will serve such DRB. For example, the gNB-CU-UP which terminates packet data convergence protocol (PDCP) can be selected locally to the antenna site, if local breakout and minimization of traffic delays are to be achieved.

Given that the gNB-CU-UP is also slice aware, specific UP traffic handling policies can be selected on the basis of the network slice assigned to the DRB. Such policies could include specific flow control profiles (more aggressive to maximize throughput, or more conservative to increase reliability), traffic preordering profiles, selection of specific security gateways and more.

Enhanced slicing experience with mobility and traffic management

As part of the network slicing framework, neighboring RAN nodes signal to each other the set of network slices supported for each tracking area. Such information makes it is possible to steer user mobility towards cells where the slices in use by a terminal are supported.

Mobility and traffic management functions take into account the radio conditions of the terminal as well as the candidate target cells for mobility and their supported network slices. A mobility decision that optimizes radio performance while ensuring as much as possible network slice service continuity can be taken by the RAN.

Shared networks and network slicing

Network slicing can be used by any or all involved service providers that are sharing a network in a multi-operator core network (MOCN) and multi-operator RAN (MORAN) configuration. The service providers sharing a network may deploy a different set of services, or similar services with different performance requirements, meaning that the possibility of controlling the performance of the services per service providers and per slice becomes important. In order to control the service performance there are different means, one part is the configuration with the possibility to have separate configuration or parameter settings taking each service providers and slice into account. Another is the need for policy handling when a resource in the system becomes limited. The network always needs to be configured and dimensioned according to the service requirements, but there may still be limited resources.

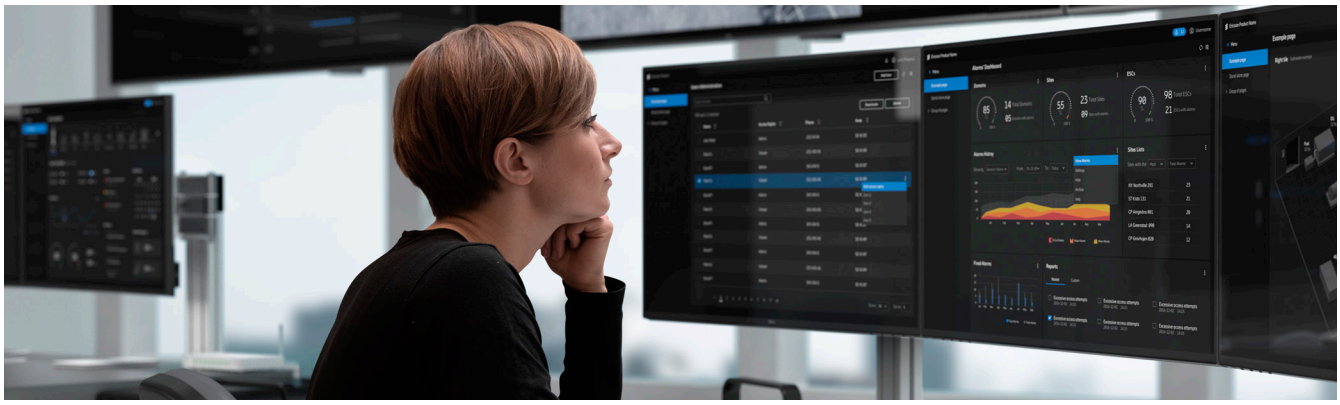
When service providers share spectrum (MOCN configuration) it can be assumed that one of the limited resources will actually be the spectrum, and hence the need of policies for management of the radio resources become important. The policies need to include the service provider view, the network slice view and also the service level view.

When service providers share infrastructure or parts of infrastructure (MORAN configuration) other resources may become limited. This may require policy handling. By having the observability per service provider and per slice, it can be seen if the different service requirements are fulfilled.

RAN transport support for slicing

SLA for slices or a group of slices in the RAN transport can be secured in different ways. One way is to make sure that the transport is over-dimensioned, even though this might not always be feasible or economically justifiable. Dedicated transport service may be required in cases where latency is an issue. Another reason for dedicated transport could be that observability per slice also in transport is required.

In scenarios where dedicated transport is required, transport could be separated logically. Traffic flows for individual or a group of slices can be mapped into separated transport services in the transport network. These transport services should have an SLA that match the required SLA for the end-to-end slice or group of slices. Mapping into transport services could be done in uplink based on VLAN Id, destination IP address or physical port from the RAN node (i.e. baseband node) and in downlink based on source or destination IP address from the data center. If a slice or group of slices have several traffic flows with individual requirement on transport characteristics and the transport service is not over-dimensioned, then it is the QoS marking (DSCP or p-bit) in the packet that is assuring that the individual traffic flows get a proper treatment required for that traffic class. Mapping of 5QI to DSCP and p-bit can be done individually per slice or group of slices.



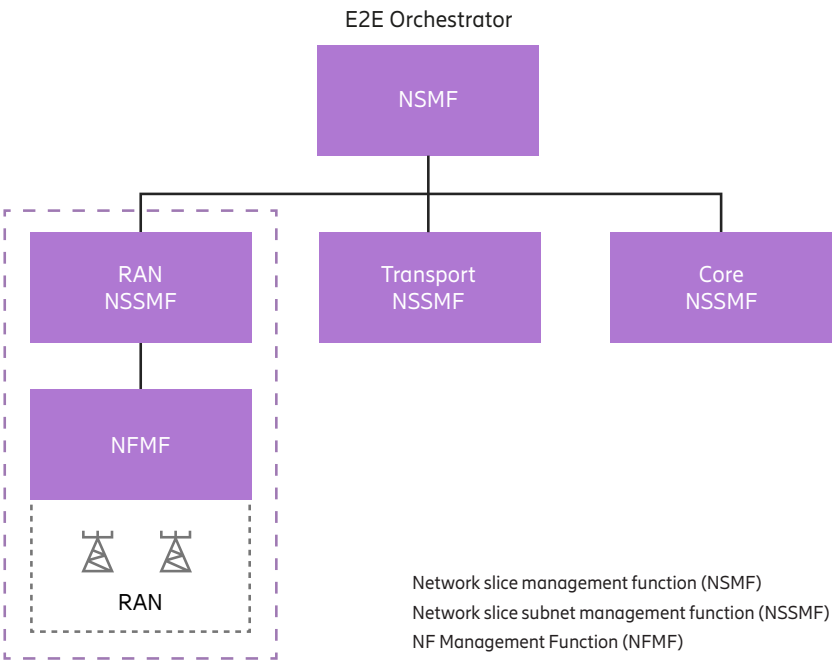
RAN automation with slice orchestration

RAN automation with slice orchestration is designed to enable service providers to carry out slice lifecycle management with slice creation/deletion of various slices.

RAN slice orchestration is part of end-to-end slice orchestration, which enables automation of slice lifecycle management tasks such as slice provisioning, activation, supervision and service assurance.

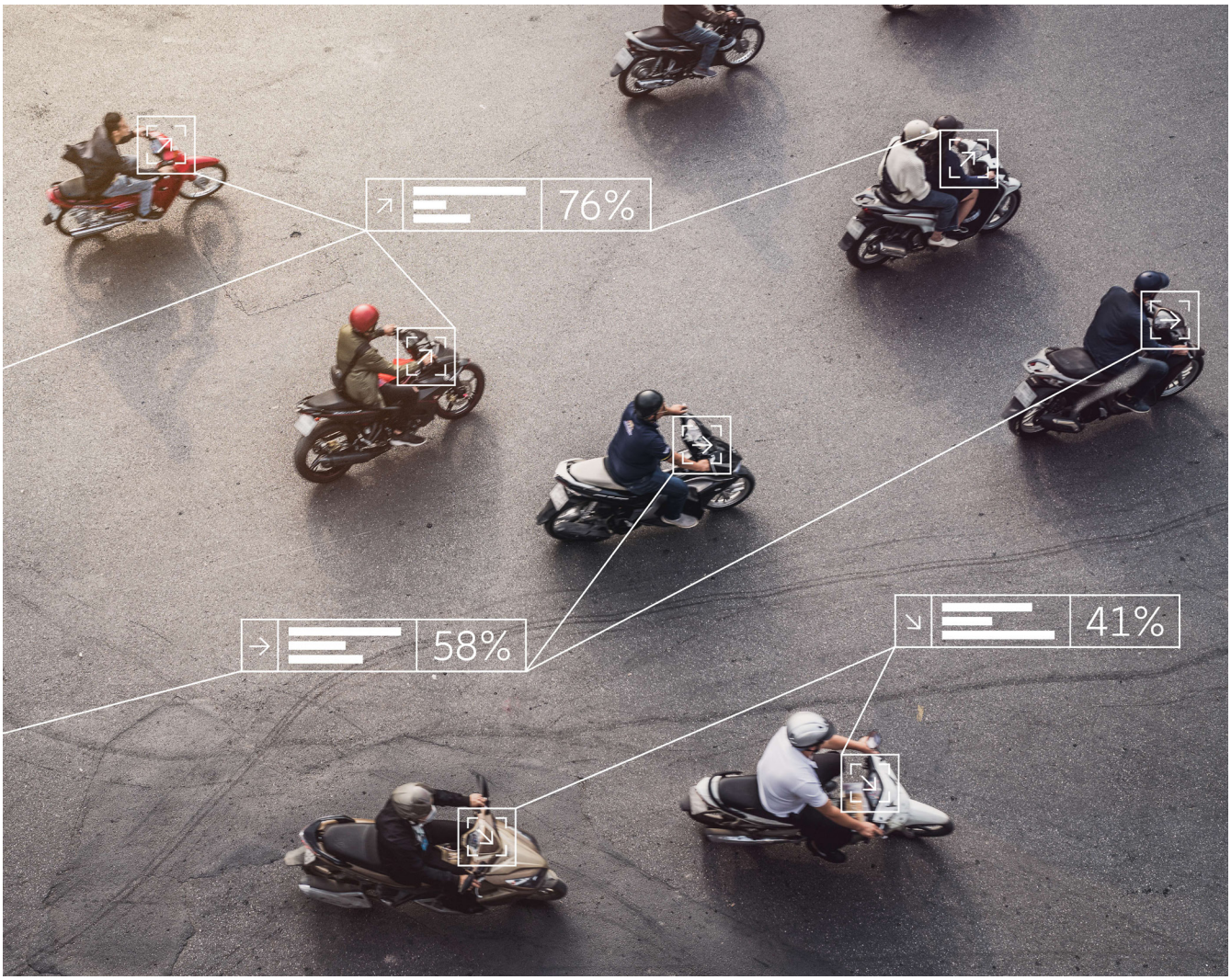
If an enterprise needs service in a dense urban location like a city, RAN slice orchestration can automatically configure all the cell sites that provide coverage in that location. They can also generate the configurations required to meet enterprise service SLA and automatically provision the respective cells by applying the configurations through element manager.

RAN slice orchestration can also monitor slice performance in the RAN network, validate it against the slice SLA, and in future automatically orchestrate configuration changes to help optimize the slice performance in the RAN network to meet assurance policies.



Network slice management function (NSMF)
Network slice subnet management function (NSSMF)
NF Management Function (NFMF)

Use of AI in Ericsson RAN Slicing



Machine learning (ML) is highly relevant for RAN slicing due to the complexity of decision problems that need to be solved. These include the large dimensionality of services and multiple tenants that need to be mapped and taken into account in short time scale in the RAN. It would be very hard – if not impossible – for a human to derive optimized solutions to such problems.

To manage network slicing complexity, service flexibility and deliver service faster, an automation platform is crucial. Enterprises expect highly responsive slices. To meet these demands, a high degree of Automation is needed, and Artificial Intelligence and machine learning influenced advanced policy management is key and named as Intent based AI/ML.

There are three major categories of machine learning solutions: supervised learning, unsupervised learning and reinforcement learning. Each major category offers unique use cases for Ericsson RAN slicing.

Supervised learning is about learning the relationship between selected input and output variables. Ericsson envisions the potential of such learning to augment the Ericsson RAN slicing framework in different technical areas. For observability, the monitoring of current key metrics (e.g., SLAs, resource utilization, performance) can be moved towards forecasting future values of the metrics. This can enable early warning for taking proactive actions to mitigate disruptions of violation of SLAs.

For the QoS framework, machine learning can augment the scheduling strategy and admission control to be data-driven where parameters are adapted dynamically per slice, flow, and/or deployment. Furthermore, root cause analysis techniques can be investigated to support engineers with a rapid slice incident resolution. For RRP, machine learning can guide resource partitioning of slices by providing suggests on demands for each user and/or network partition. Supervised learning can be very well giving us sense of utilization/traffic prediction/forecasting in the RAN. That knowledge can be exploited by slice-aware scheduler.

Unsupervised learning is mainly about automatically grouping (or clustering) similar objects without labelled responses or detecting rate items or events (also referred to as anomaly detection). In the context of RAN slicing, the slice monitoring system can be augmented with KPI degradation detection.

> Artificial intelligence and machine learning influence advanced policy management, that is key to a high level of automation.

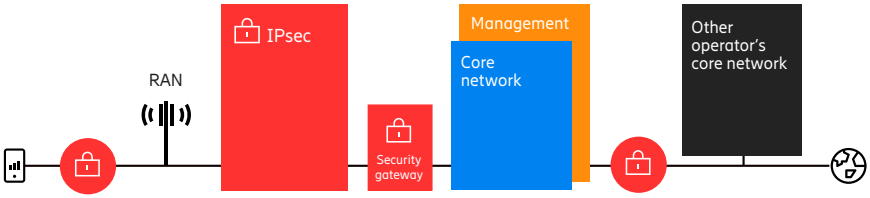
Securing network slicing

A slice can have different security postures based on identified threats and end-customer requirement. One slice shall, for example, not be able to starve another slice. The dynamic sharing of resources mentioned above offers high efficiency of the resources. Equally, in a congested situation a guaranteed bandwidth per slice will prevent from slice starvation and reducing the impact of a potential starvation attack. In addition, QoS with shaping and policing functions are important parameters for preventions of these type of attacks.

A slice can also be exhausted, resulting in denial of service (DoS). Isolation between the slices will here prevent the impact from such attack being spread across the different slices i.e. if one of the slices would be targeted for a DoS attack, the other slices will not be impacted, given that proper isolation is in place. Isolation will also reduce the risks for other attacks, such as side channel attacks, for example for someone who wants to steal data or manipulate with the data. Depending on where a potential adversary would gain access to the network, authentication, encryption and integrity protection of traffic sent over the network is essential. The traffic sent over the air interface is protected by default, while the traffic sent over the transport network can be protected based on service providers' policy and agreements with their customer. Protection is primarily done using IPsec. Different IPsec tunnels can be created for the different slices.

Observability is also an important aspect to consider from a security point of view. Any suspicious activities that indicates that there is an intrusion or breach to happen or on-going shall preferably be detected as quickly as possible. Ericsson Security Manager makes security visible providing fast detection of network anomalies based on security analytics and demonstrates adaptive security by looping back analytics to the policy automation.

In the future, mechanisms for transferring a device from one slice to another dynamically may be supported. This can be interesting in case a device displays suspicious behavior. Affected devices could then be transferred to a quarantine slice, which may be a replica of the original slice with certain modifications. This slice may be configured with more strict security and lower bandwidth. Instead of completely blocking the device, it can be allowed to operate under tighter restrictions reducing the risk of disturbance within the original slice.



Enabling new business opportunities with Ericsson RAN slicing solution

Network slicing is crucial to guaranteeing the performance of the new service to the enterprise and MBB market segment and is an enabler of maximizing the full business potential of 5G investments.

Ericsson is a pioneer of network slicing development. Ericsson has developed a healthy ecosystem with robust specification in 3GPP and product development to support end-to-end network slicing.

Ericsson RAN slicing solution described in this paper refers to the flexible and scalable slicing architecture that dynamically share finite RAN resources to provide differentiated handling for bespoke slices for SLA fulfilment.

Ericsson has the ability to tailor customer service offerings according to varying needs, reduced risks, increased flexibility and agility. Ericsson RAN slicing solution offer this with shorter time-to-market and improved total cost of ownership. Ericsson network slicing solution provide opportunity to monetize CSP 5G investment and open door for new revenue segment from enterprise and MBB segment. The objective is to achieve full dynamic orchestration of end-to-end slicing with optimal automation.

Ericsson RAN slicing solution has the ability to offer customized service offerings according to varying needs, reduced risks, increased flexibility and agility.

Ericsson enables communications service providers to capture the full value of connectivity. The company's portfolio spans Networks, Digital Services, Managed Services, and Emerging Business and is designed to help our customers go digital, increase efficiency and find new revenue streams. Ericsson's investments in innovation have delivered the benefits of telephony and mobile broadband to billions of people around the world. The Ericsson stock is listed on Nasdaq Stockholm and on Nasdaq New York.

Ericsson.com

EN/LZT 1/28701-FGD101489