Future Network Architecture

Description

The importance of and dependency to the mobile network is growing as is its versatility to be used to support society-, mission- and business-critical applications. The future network will evolve from today's 5G via 5G Advanced to 6G. It will be based on a horizontal architecture approach and be flexible enough to support a wider range of use cases than before. It has become the largest innovation platform seen.

Disclaimer: The statements shall not be seen as product commitments and can change at any time depending on ongoing standardization, technology evolution and customer feedback.

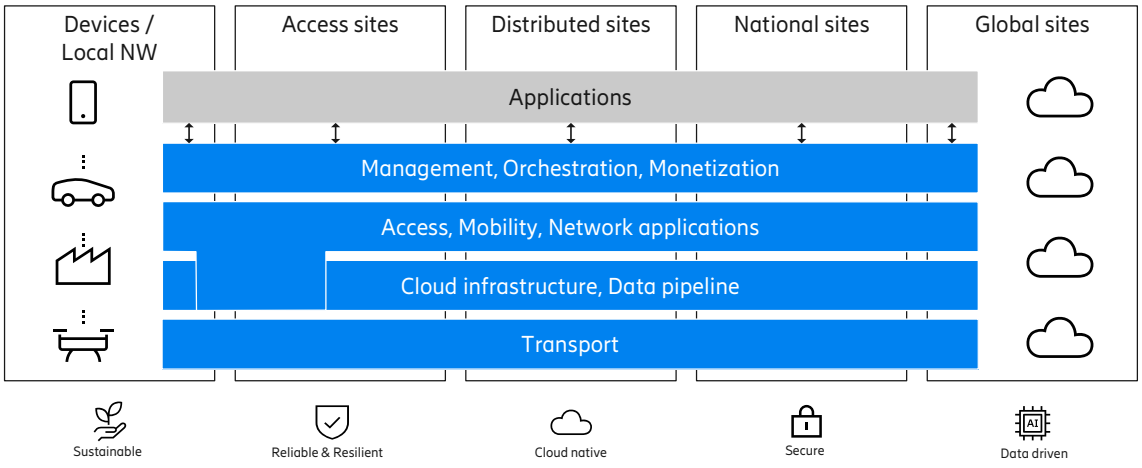


Table of Contents

1 Executive summary

The current 5G mobile networks continue to take a central role in normal users' everyday life while extending into the communication for mission critical as well as business critical communication which both require more available, reliable, and resilient networks. The powerful innovation platform created by those networks, for virtually any industry sector and society, is going to be continued as 6G develops.

Future networks must address the needs of the 2030 society to improve in efficiency, trust, and sustainability aspects all while preparing for advanced services in the cyber-physical future.

Above will require the networks to be more dynamically adaptable. Network functions and applications should be possible to run where and when they are needed optimizing performance, cost, and business agility.

To be able to meet these requirements there is a horizontalization transformation trend from dedicated, well-defined, and vertically integrated nodes to horizontally deployed networks.

Further these dynamic and adaptable networks will require a higher degree of automation in combination with intent-based management. This is especially true for CSPs going beyond MBB which is also the approach taken by Ericsson.

This introduction of automation in operations requires AI support which in turn introduces the area of MLOps which is a set of processes and technology capabilities for building, deploying, and operationalizing Machine Learning (ML) systems which unifies ML system development and ML system operation with DevOps. This presents new challenges but will result in a more expedient handling of artifacts like models, pipelines, datasets, etc.

Several CSPs are currently expanding the use of telecom to explore the exposure of network capabilities, e.g. to address enterprises. By leveraging their networks, enabling ease of consumption of network services can be achieved through a set of simple yet powerful APIs for business management as well as network capability access.

Security is top of mind in the current geopolitical climate and the introduction of a Zero Trust Architecture (ZTA). According to the ZTA principles a network must be seen as untrusted, and each subject or resource must be protected by the correct security measures.

The 6G architecture needs to support the expected new use cases and service requirements as described above. Further there is a need for an aligned industry view of a single 6G architecture, avoiding the multiple architecture options defined in 5G. This will help reducing the overall complexity of the 6G standard and help with faster introduction in the market.

As alluded to 3GPP based networks are increasingly versatile and used in several deployment cases described in the generic examples; General public networks, Wide-area dedicated networks and Local dedicated networks each with options of deployment and a variety of combinations possible.

2 Introduction

5G mobile networks have taken a more central and present role in everyday life for normal users, but also in the role of communication for enterprises as well as for critical communication in society. The networks create a powerful innovation platform for virtually all industries and society. 6G network will continue this trend.

We are experiencing an ongoing transformational trend of horizontalization in how networks are built, operated, and opened for innovation. Instead of dedicated, well-defined, and vertically integrated nodes connected in a static network setup, the networks are evolving towards a more horizontal and dynamically adaptable architecture where Network Functions (NF) and applications are running where and when they are needed to optimize and balance performance, cost, and to support business agility.

The Global Architecture, Figure 1 can be viewed as a top-level view for a start of network architecture discussions but will require further detailing and breakdown.

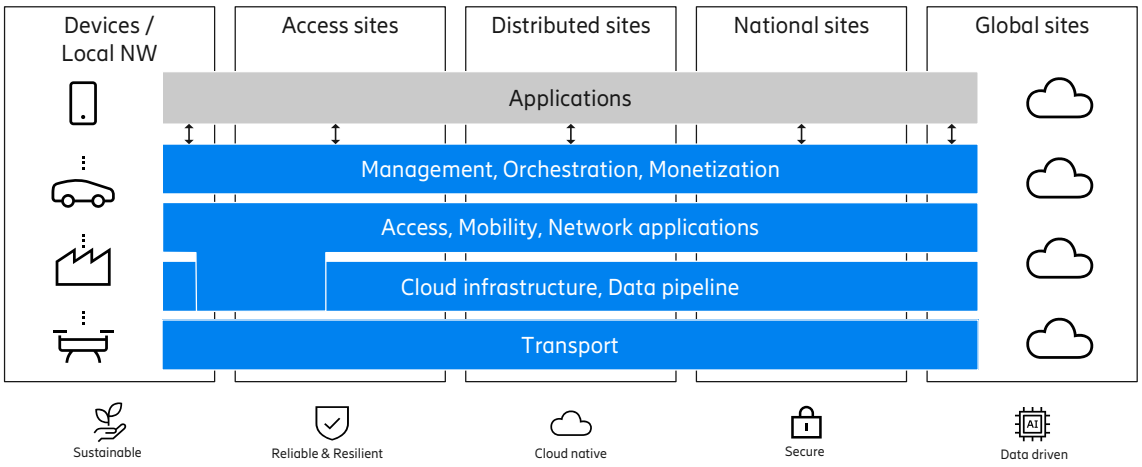


Figure 1 The Global Architecture

This horizontalized network architecture enables a distribution of cloud resources, joint data pipelines and Open APIs for the programmability and flexibility required - both inside the network and towards the outside world opening for new business models through e.g. Service Exposure.

While the network is an essential asset for the future of the CSP, the connectivity services market is mature and displays low growth. Some CSPs have thus started to extend their focus to evolve their network into a platform for innovation and new business.

Below Figure 2 presents on a high level the drivers and implications of the horizontalization of networks.

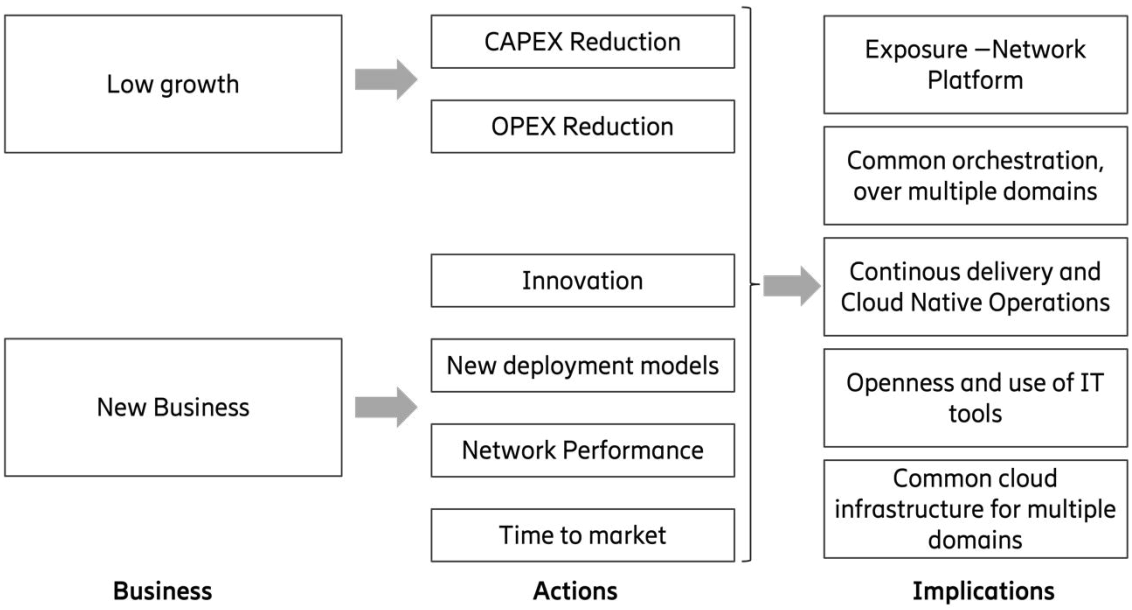


Figure 2 High level horizontal drivers and implications

The two business situations of either Low or High growth leads to different possible actions and implications. An OPEX reduction action could for example result in an implication of higher investments in automation and orchestration. Addressing new business on the other hand may rely on new deployment models using continuous delivery and cloud native operations.

Below is a non-exhaustive list of reasons for the horizontalization:

- Technology evolution and innovation should be focused on horizontal areas to leverage wider ecosystems with openness in each horizontal layer, e.g., AI, L1 HW acceleration, etc.

- Increase business agility using horizontal, external B2B interfaces and APIs to support integration into enterprise and industrial solutions.

- Network automation should be based on business level intents for automated, dynamic and adaptive networks addressing business level needs.

- Leverage the strength of a few business-relevant vertical interfaces and new horizontal interfaces coordinated across relevant standardization bodies, open-source projects, alliances, and partnerships to retain system values.

These opportunities will come at the back of some challenges such as; change of mindset in the industry from today's node centric models to a horizontal architecture, evolved commercial models, new ways of working adapting to patterns from IT, e.g. for Life Cycle Management (LCM), etc.

3 Future Network Outlook

3.1 High-level needs

Future networks must address a wide array of demands targeting the 2030 time- frame such as mission-, business-, and society critical needs through network capabilities, which in turn are delivered by technical solutions. We should therefore start from the needs. Four areas of drivers have been identified, see Figure 3. These drivers indicate what 6G should be designed for. Concretely, the first three drivers can be said to point at networks to improve in efficiency, trust, and sustainability aspects, while the last driver points at networks to be extended to provide novel advanced services for cyber-physical environment.

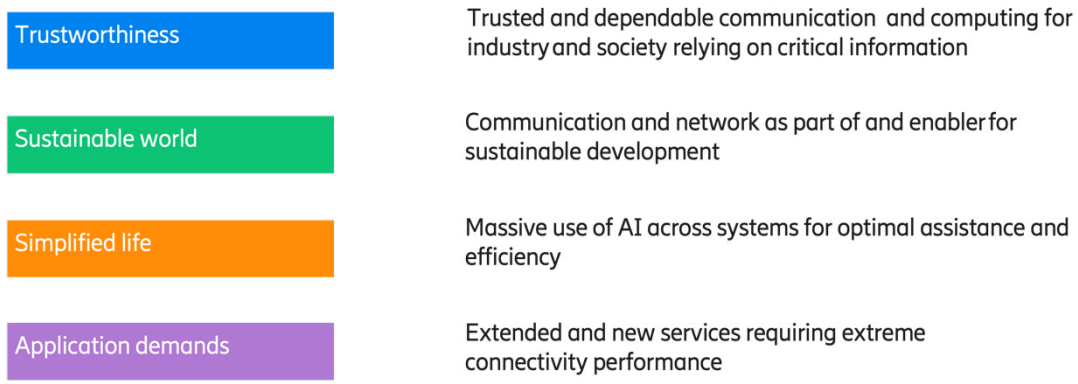


Figure 3 External drivers for 2030 networks

Mobile networks are expected to be gradually upgraded to support 6G technology in the prevailing cycle of mobile generations. 6G will likely be both an improvement and expansion like what 5G was to 4G. The "5G triangle" of eMBB, URLLC, mMTC services, will extend with new capabilities added for delivering networking in the cyber-physical world, as shown in Figure 4. These new ambitions have been agreed in ITU-R [1].

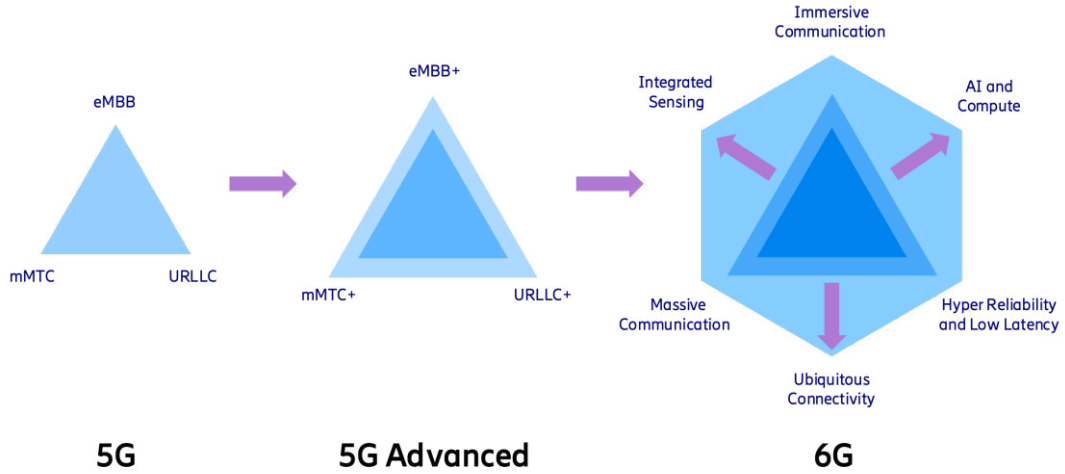


Figure 4 Evolving 5G and the journey to 6G: enhancing and expanding into new services

3.2 Major capabilities and use case trends

A set of "hero use cases", Figure 5 for the 2030 networks are being studied corresponding to the needs and composed of the services mentioned above. They map on a high level to the two improvement directions mentioned: offering new advanced monetizable cyber-physical services and improvements in trust, efficiency, and sustainability. The use cases are broad and will partly overlap. Some imply increasing requirements in uplink (UL) capacity and performance [9], others require good positioning capabilities. It should be noted that the use cases are quite different in scale (i.e., coverage, bandwidth, and latency requirements), requiring flexible deployment solutions as described in Chapter 6 Network Deployment Cases.

Figure 5 Hero use case examples for 6G/2030

Above use case examples will rely on some capabilities from 5G to be further evolved and new to be developed in 6G. These are described below. These examples are selected to showcase new capabilities and ambitions in 6G, but naturally the use cases from 5G, such as FWA and the Enterprise area, will also evolve into 6G.

Networks as platforms for connectivity and beyond
Ericsson would like to see networks broaden their scope and offer more services beyond transport of data, including the ability to process data on different levels and offer the result as services.

Offering more add-on services through APIs means potential for new revenue and a way to grow in the value chain.

A use case illustrating the network platform scenario is Situational awareness, where connected machines (e.g., drones) move about in a city. With the use of the network platform, they are localized and have reliable data links throughout their paths. They can also access updated digital maps of the city where the location of other connected objects (through positioning) as well as unconnected objects (through sensing) is inserted.

Performance differentiation
Today most applications rely on best effort eMBB and going beyond this offering is relatively rare. Ericsson would prefer that applications and users increasingly rely on good performance while being willing to pay for this. Market research supports this willingness [4], though the larger changes are yet to be realized.

To meet the above the network should contain components for flexible exposure of APIs to aggregator platforms, support evaluation of SLAs, be able to turn those SLAs into concrete network configuration settings (perhaps via intents), make predictions of the performance, and collect data at all network levels (observability).

Diverse requirements, diverse deployments
Ericsson expects future networks to have increased heterogeneity, both in terms of radio access technologies as well as in deployments. For example, in the emerging enterprise market, we expect to see private networks in combined deployments in customer premises and public cloud.

Resilient networks
Enabling Mobile Network Operators and Private Network integrators to offer service performance guarantees (SLA) to their business-critical and safety-critical service users will enable new valuable applications and additional revenues.

Data and AI everywhere
The broader trend of AI is also applicable to networking. Future networks are expected to handle larger amounts of traffic in all parts of the spectrum, including sub-THz frequencies, providing services to devices of different capabilities, and covering larger areas. Automation is needed for cost-effective network operation (see chapter 4.1).

Artificial Intelligence (AI) and specifically Machine Learning (ML) can improve network automation by use of trained models, generating insights that in turn benefits automated decision making.

Horizontalization
The trend of horizontalization is increasingly applied also in mobile networks promoting open interfaces which have always been an important building block for the success of mobile networks. Current networks are in different stages with respect to their horizontalization.

Softwarization
Most industry sectors have been or are still undergoing stepwise digitalization of processes, services, and products [8]. This has also been true for the telco industry. It is safe to say that today almost everything is software (SW) and in some way programmable and traditionally physical functionality starts to move into the SW domain, e.g. with digital twins based on real-time data from the physical counterparts.

Some examples of this Softwarization are: Open-Source use for innovation, more frequent SW deployment cycles relying on DevOps and CI/CD, WebAssembly runtimes in service mesh, SmartNICs.

E2E encryption
In the 6G time frame we assume that all traffic will be encrypted end-to-end, both when it comes to the content and the header and control information and applications will have increased ability to choose how they use the network and what information gets revealed

to the network. Consequently, we expect most mobile networks to start applying Zero Trust principles.

Environmental footprint
Sustainability is one of the key external drivers [5], influencing most if not all fields of industry already today and impacting future networks in two ways.

Networks are expected to enable sustainable use cases with potential of societal benefits addressing the environmental, social, and economical sustainability described in the UN Sustainable Development Goals (SDGs) [4]. Secondly the networks themselves should consume only the resources that are necessary both in production and in operations.

Digital Twins everywhere in society
Digital twins (DT) are a centerpiece of the envisioned cyber-physical continuum. Though simple forms have been used in many industries, Ericsson expects a more comprehensive use in the future. A network DT is a digital representation of a live network which permits for more accurate predictions of the impact of actions before they are taken.

Related articles/additional reading:
Ericsson Technology Review, 6G – Connecting a cyber-physical world
NGMN, 6G Requirements and Design Considerations
5G monetization to improve top line revenue capture - Ericsson
Intent-based networks in telecom operations - Ericsson

4 Network Capabilities

4.1 Automation and Intents

As the service needs for MBB increase or new use cases are addressed and supported in a CSP's network, automation is expected to be necessary. For CSPs still pursuing only MBB, efficiency will be the main reason for automation. Going beyond MBB requires a balance between efficiency and flexibility to be supported by automation. Ericsson's architectural approach assumes going beyond MBB.

The progression of automation will move from managing the resources and details to managing services and objectives.

The journey towards achieving a fully autonomous intent-based network, requires its architecture to be prepared by raising the level of abstraction, though maintaining precision. A couple of aspects of this journey are discussed below.

Separation of concerns or centralization, i.e. whether to have a more centralized responsibility domain covering RAN, Core and E2E management or more decentralized responsibility domains. Figure 6 is a prime choice to make. (Note that if the decentralized approach is used, it can still be co-deployed with internal separation of concerns via permissions.)

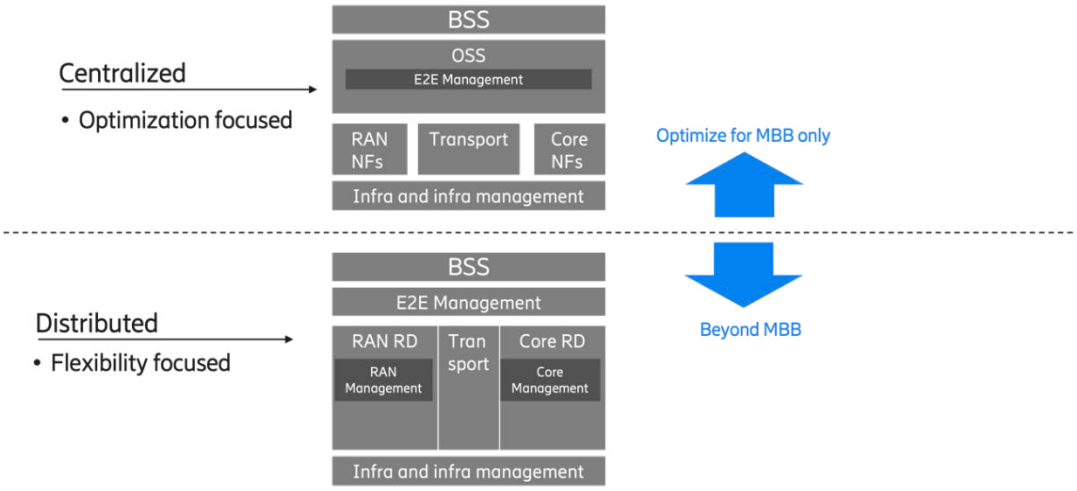


Figure 6 Centralized or distributed Responsibility domains

The trade-off is that a centralized approach offers more optimization and a decentralized approach favors flexibility at the sacrifice of some level of optimization. The flexibility may range from e.g., new service introduction, or flexible offering of services to other domains (e.g., RAN sharing).

Intents can be used to increase the total industry revenue by enabling CSPs to potentially offer multiple SLA-based products and/or to address increasing operational costs simplification of planning, optimization and assurance.

Various industries have previously used the concept of "intents", but with varying meaning and definition. Even some mobile systems, have used intents e.g., as part of Kubernetes in the LCM of software. In this document, if not explicitly stated otherwise, we follow the basic definition of intent, intent owner, and intent handler set by Telecom Management Forum (TMF) [6], with intent API following either TMF or specified as complements to existing interfaces.

For the Introduction of Intents, three steps are required, Figure 7: **1)** Introducing intent technologies within the constraints of the existing interfaces in TMF, **2)** introducing the intent modelling approach and **3)** Introducing the intent interfaces.

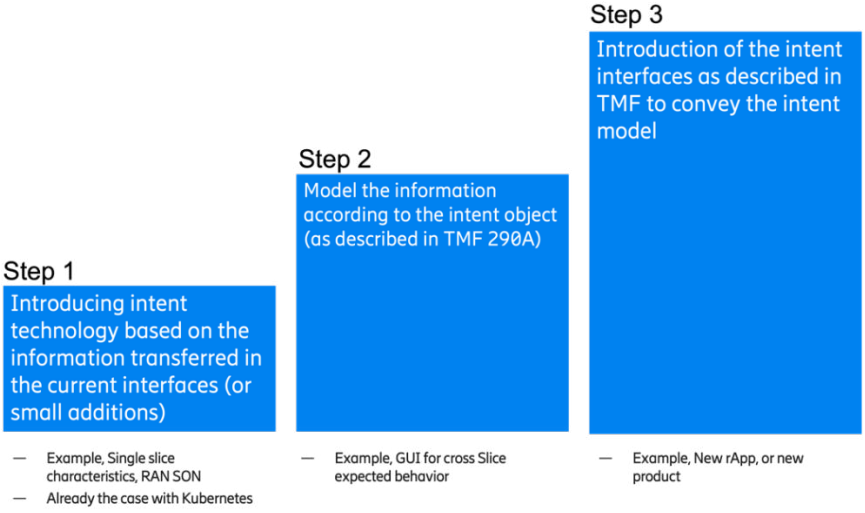


Figure 7 Intent introduction steps

An intent owner uses an API to set an intent to the intent handler. Specific intents point to a set of measurable quantities (meaning measurable by the intent handler) plus targets for the measurable quantities. In most systems the intent handler receives more than one intent. In such systems the intent handler shall prioritize between the intents in situations where not all intents can be met.

The domain-level target network architecture for intent handling is depicted in Figure 8. The domains are functional domains following a typical CSP operational responsibility structure.

The intents enter the architecture in the operations parts of the architecture while the run-time traffic and control interfaces do not use intents. In particular:

The interface between the Business support system and the domain for service- and QoE management is complemented with missing information to support intents e.g. intent-priorities associated with network services.

The interfaces between the domain for service- and QoE management and the network sub-domains "RAN", "Transport", Infrastructure" and "Core network" will be extended with intents and intent-priorities associated with sub-domain network services, e.g. a transport service or a RAN service.

The interface between the domain management system and the network functions will be extended with intents, intent-priorities, and intent reporting (observability) associated with function behavior, gradually migrating away from traditional node configuration to intent-based function configuration.

The target architecture also includes intents for management of software, infrastructure, and transport, where intents are to some extent already used.

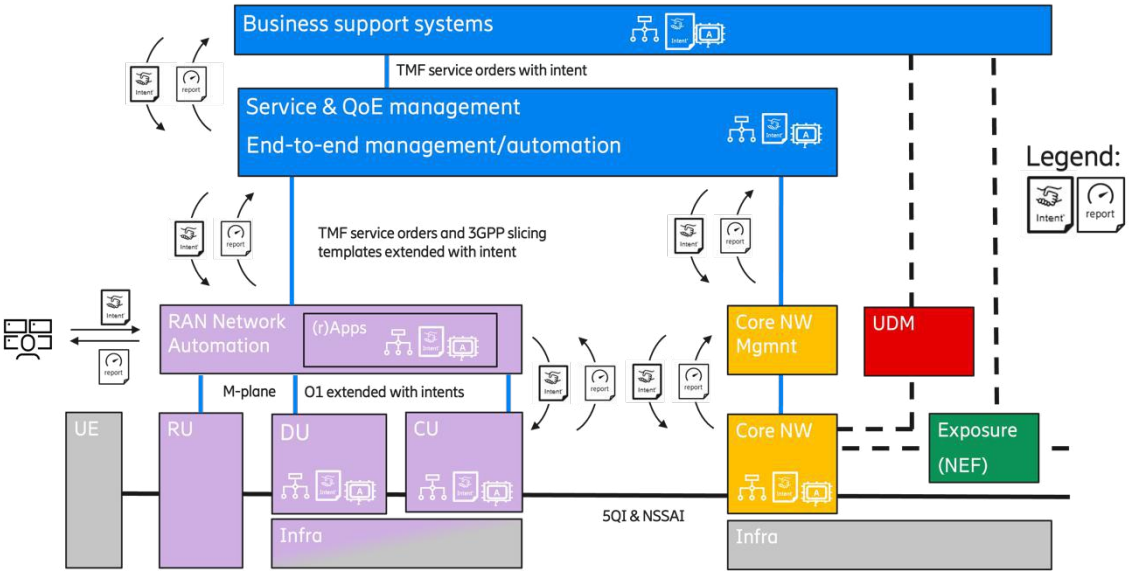


Figure 8 Domain level target network architecture for mobile systems using intents

The ground rule is that traditional configuration parameters have priority over intents meaning they will serve as constraints including SON (Self Organizing Network) functions. Traditional configurations can then form constraints and boundaries for dynamically optimizing a system with intents within these constraints.

This migration will gradually enable inclusion of intents in two parallel main tracks:

1. Intent handlers will take control over the setting of selected configuration parameters, replacing manual configuration and SON functions.

2. Intents will increasingly replace configuration parameters, communicating to sub-systems directly with intents.

These ground rules enable legacy systems to form stable and safe starting points. Intents can start off "small" where existing configuration parameters act as guard rails towards unwanted and unexpected system behavior.

Intents should complement and interwork with 3GPP QoS, Network Slicing, etc.

Finally, intents give an opportunity to improve various aspects of network performance through automation and optimization towards specified goals.

To fully leverage intents, a multivendor eco-system should be maintained supported by standardization in the areas of Autonomous Networks and Intent-driven management already ongoing (in e.g. TMF, 3GPP, ETSI O-RAN). An architecture with wide industry acceptance is desired to address the aspects of automation such as AI, intent-driven management, digital twins, data-driven management, MLOps, and others.

Related articles/additional reading:
Creating autonomous networks with intent-based closed loops
Multi domain orchestration business opportunities
Intent-based networks in telecom operations - Ericsson
Intelligent automation apps - Ericsson
Cognitive framework for better 5G network LCM - Ericsson
SMO enabling intelligent RAN operations - Ericsson

4.2 Artificial Intelligence and MLOps

MLOps is a set of processes and technology capabilities for building, deploying, and operationalizing ML systems, including how data is refined and transformed to serve the ML system, aiming to unify it with DevOps, targeting the introduction of software in a repeatable/reproducible and fault-tolerant workflow.

Thus, MLOps advocates automation and monitoring in all steps of ML system construction and deployment with a main goal to achieve shorter TTM with high confidence level of addressing challenges in the automated processes of development, verification, etc.

Additional challenges of adopting MLOps to highly reliable live telecom networks exist such as the need to handle lifecycle management and automatic re-training of the many instances of ML models.

Adoption of MLOps enables a more expedient handling of artifacts like models, pipelines, datasets, etc. in a uniform way across the different stages of the process.

Targeting products and services, will require MLOps to be able to be deployed for several scenarios, e.g., provided as-a-Service (aaS) or licensed SW/product on customer site, deployed on cloud infrastructure or deployed on dedicated HW, and likely in several more.

CSP realities vary depending on selection of cloud infrastructure where a CSP potentially selects to use a particular HCP or use private cloud which can be done for various reasons, like applications execution, licensed SW, data management and storage, etc.

Being able to apply MLOps across several large HCPs (e.g., AWS, Azure, GCP), with the limited compatibility between their APIs for AI services, requires a certain level of adaption for vendors' products/services to adopt to each of these HCPs. Thus a multi-cloud approach will require certain efforts in a variety of areas depending on each HCP.

A few abstraction layer initiatives exist that could help to provide an abstraction layer for different HCP services. None of these alternatives, however, yet provide a complete solution to the problem.

The telecommunications domain poses challenges not readily addressed by mainstream MLOps solutions. We see a few key requirements on MLOps compared to what third-party solutions offer:

MLOps needs to operate in multiple different environments: services, products, and embedded solutions, including deployments on CSP site. This means that multiple deployment scenarios for MLOps functionality for developing, deploying, and executing AI-models/applications needs to be supported in parallel, such as all local to a CSP site or all at vendor premise or combinations of these.

Products and services may need to be developed to be compatible with multiple HCPs. Each HCP offers different APIs, for deployment as well as for ML services, which are not aligned. To overcome this and enable multi-cloud deployment, there are three main alternatives: HCP specific implementations, Abstraction Layer (e.g., through relying on a 3P framework) or relying on CaaS layer. Since HCPs are currently not converging, the direction is to continue evolving selected tools for training and inference. We will rely on CaaS layer to support hybrid and multi-cloud deployment of ML workloads by norming the execution environment. This approach needs to be constantly re-evaluated if HCPs' services are converging, or other solutions are emerging.

Data - Important data is generated at the network edge, and not all data is owned by the vendor. Solutions are needed to manage data on the CSP side, and solutions may not be able to retain all data. MLOps require iterating on pipelines and models using data, ideally from multiple customers. Data needs to be available where the MLOps environment is executing and needs to be able to be injected to all the different MLOps phases. This data is important for the vendor to develop new and enhance existing features, this is done in the vendors data driven development environment.

Simulators are essential for developing e.g., reinforcement learning use cases, or for estimating the effect of network actions taken based on ML predictions. Data from live networks is an attractive asset. To best support the wide variety of use-cases, there may be a need for domain/use case specific tools to fit the development flow.

The main components in MLOps can be deployed in several sites and combinations depending on product/service needed. In Figure 9 below the deployment architecture is outlined for one of the most complex product cases where development is initially performed in a vendor site, while it also supports deployment of Experimentation, Development, Staging and Execution environments in national or regional sites at the customer. This may be more limited for some products, e.g., only including Execution environment, but it is included to show one end of the scale of complexity. For some products the Execution environment may also be pushed further out in the network.

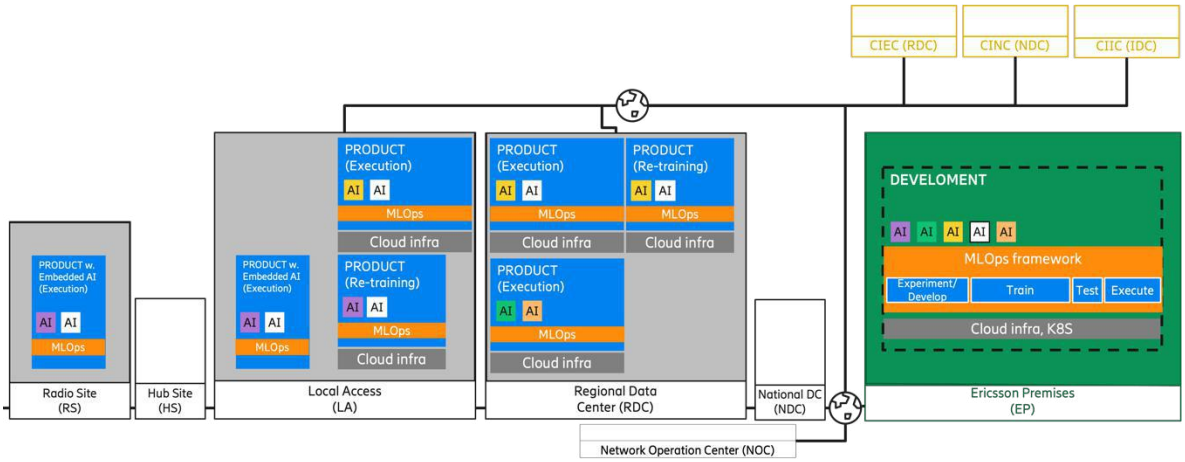


Figure 9 MLOps deployment architecture, scenario with complete MLOps in customer deployment

Related articles/Additional reading:
[Defining AI native: A key enabler for advanced intelligent telecom networks]
[AI-powered RAN: An energy efficiency breakthrough]
[Trustworthy AI - What it means for telecom]

4.3 Network Reliability, Availability and Resilience - NRAR

Enterprises, governments, and private citizens alike rely on its availability, reliability and resilience around the clock. To be able to meet ever-rising expectations on the network, while simultaneously meeting the requirements of emerging use cases beyond MBB, increasingly higher levels of network robustness will be required.

5GS (5G System) 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 NRAR. 5GS delivers new capabilities that enable enterprises with business-critical use cases in segments such as manufacturing, ports, automotive, etc. to take a major step forward in their digitalization journeys by replacing older means of communication with the 5GS. These new capabilities can also be leveraged for mission-critical networks.

As private citizens, we have become increasingly dependent on the availability of MBB networks and rarely leave home without our cellphone. In our daily life we conveniently rely on being able to use the network for services like banking, ticketing, early warning systems, etc.

With MBB networks availability is taken for granted, it may be challenging to monetize NRAR. Some proactiveness is advised to avoid being caught out by the expected regulations emerging due to society's dependency to the network. A more reliable and robust network should be seen as a premium service and it can also attract new subscribers and prevent churn.

Mission-critical networks are already deployed using mobile technology. 5G will provide possibilities to extend into new areas and new use cases such as Public Safety, Utilities or Digital Airspace. It will further enable XR use cases for, e.g. first responders or in other suitable cases across different segments and verticals.

The characteristics of wide area coverage makes mission-critical networks a very compatible case, as a symbiotic extension to the existing mobile network of the CSP, who can present a robust and reliable network.

For business-critical or Enterprise verticals, low latency in 5G may appear to be a killer case. Although valuable, industries express the need for a reliable and predictable network to relate to their processes, machinery, etc.

Delivering new capabilities to enable enterprises with business-critical use cases as well as the modernization and digitalization of public safety, requires that the e2e perspective for a service/application is considered. This will require a trustworthy network while at the same time provide extensive coverage and high reliability.

This is a shift in the view of robustness from the more traditional use of Node Level In Service Performance (ISP) towards Network Robustness to support the top level e2e application requirements.

While both 4G and 5G can provide the high level of robustness required to deliver MBB services of today, the new and emerging use cases will require addition of new features and mechanisms in a network robustness toolbox. Beyond that and as depicted by the examples shown in the blue and red fields in Figure 10 below, it is equally important to consider Planning and Deployment of the Functional and Traffical parts to be able to secure a NRAR capable Network. Network-level design must include consideration of both sunny day scenarios and different disaster/failure cases in all parts of the network.

The green line between the application client and the server, highlights the significance of the E2E perspective.

The 5GS robustness toolbox consists of both standardized and vendor-specific network features and mechanisms. Highly flexible, it gives CSPs the power to activate the most appropriate mechanisms depending on the use cases and the deployment variants. Even if many areas to build robust networks may appear to be using the same or similar principles; network planning, redundant power, solid operations, comprehensive procedures for LCM, etc. they may differ in how they are applied for the above cases.

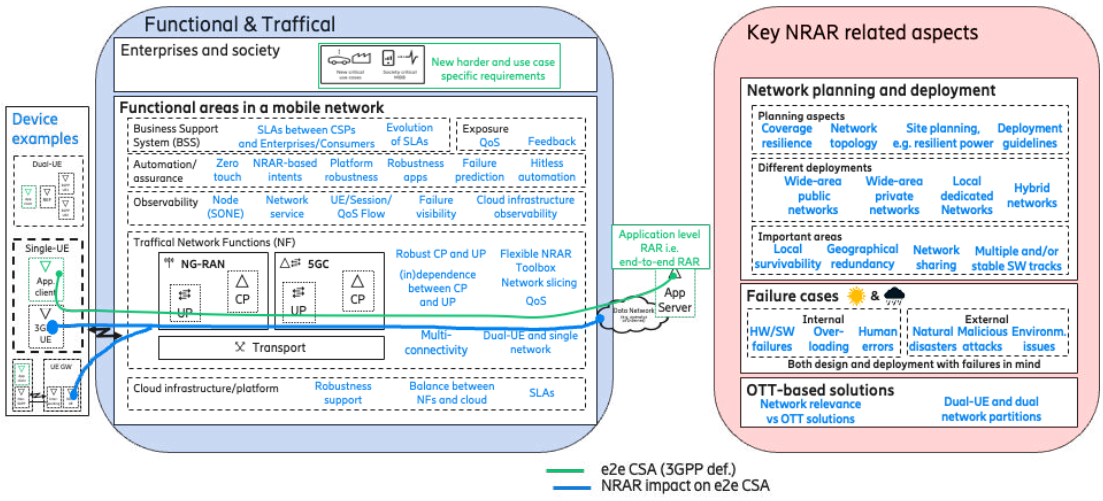


Figure 10 Key aspects for NRAR

NRAR-levels is a methodology introduced by Ericsson as a guidance to network operators and their customers (see Figure 11). The key characteristic for each NRAR-level is the Supported/ Maximum Service Interruption Time. The ASIT (Allowed Service Interruption Time) requirement of an application is used to identify the NRAR-level needed i.e., based on the use cases supported by the network operator. Each NRAR-level defines the NRAR tools and network planning/deployment guidelines needed. Once these tools are activated, and the guidelines followed, the intention is to meet the SIT for the selected NRAR levels at different (single failure) situations.

The number of NRAR levels is based on the combination of existing ASIT requirements and identified NRAR tools, and to some level of future proofness in both these aspects. The 6 NRAR levels are the current best thinking and may change going forward.

A network can support multiple NRAR levels simultaneously for different UEs and applications. This means that different NRAR tools can also be activated to different UEs, especially in the RAN. The finest granular level is a QoS Flow for a specific UE.

Any breaches to the ASIT will impact the Communication Service Availability (CSA) of the concerned application(s) negatively. For applications with high CSA requirements, it is recommended to apply an NRAR-level with a SIT which is lower than the ASIT of the application, to ensure that (single) network restarts/failures will not cause communication failures for this application.

NRAR-6

Supported/Maximum SIT	**0 ms**
For applications with ASIT requirement	0 ms - <20 ms
Deployment Area example	Local Area
Use Case examples	OT Factory – Extreme U.C

NRAR-5

Supported/Maximum SIT	**20 ms**
For applications with ASIT requirement	20 – <50 ms
Deployment Area example	Local Area
Use Case examples	OT Factory

NRAR-4

Supported/Maximum SIT	**50 ms**
For applications with ASIT requirement	50 ms - <200 ms
Deployment Area example	Confined Wide Area
Use Case examples	Rail

NRAR-3

Supported/Maximum SIT	**200 ms**
For applications with ASIT requirement	200 – <500 ms
Deployment Area example	Wide Area
Use Case examples	Public Safety, utilities

NRAR-2

Supported/Maximum SIT	**500 ms**
For applications with ASIT requirement	500 ms - <2 s
Deployment Area example	Wide Area
Use Case examples	Public Safety, utilities

NRAR-1

Supported/Maximum SIT	**2 seconds**
For applications with ASIT requirement	2 – 3 s
Deployment Area example	Wide Area
Use Case examples	Society Critical

Best effort ("*NRAR-0*")

Figure 11 NRAR-level framework

Related articles/Additional reading:
[Ensuring required network robustness with the 5G System - Ericsson]

4.4 Traffic Classification and QoS

There are two main ways for CSPs to monetize QoS in 5G networks. The first alternative is via subscription(s), where the end-user (consumer or enterprise user) pays the operator a premium for differentiated services, optimized for video conferencing, XR, gaming or some other service. This model is also referred to as B2C/B2B model and is currently supported by the category-URSP (UE Route Selection Policy) and enterprise-URSP solutions.

The second alternative is via Network APIs, where an Application Service Provider (ASP) pays the CSPs a fee for granting a specific user a differentiated service for a limited period.

This model is also referred as B2B2X model and is supported by the CAMARA NI-QoS (Network Initiated QoS).

These two monetization options are not in competition with each other. Instead, they can work well together, and we believe that a combination of B2B/B2C and B2B2X models is the best approach to maximize the overall TAM that CSPs can address.

Traffic classification is about mapping of different applications and application flows from a specific UE to different network resources (e.g., network slices, PDU sessions and Radio Bearers) in both uplink (UL) and downlink (DL). These network resources may have different QoS levels associated to them (Figure 12). NI-QoS, URSP and L4S (Low Latency Low Loss Scalable Throughput) are examples of traffic classification mechanisms with different control points that can be used for QoS support in mobile networks. Traffic classification is seen as a dynamic way of handling a single UE connected to multiple network resources. UE classification would be a more static way of controlling that different UEs are connected to different network resources.

Additional network functionality beyond traffic classification will be needed to support QoS. Examples of such functionality include SLA, SLA assurance support with relevant network KPIs and e2e dimensioning and configuration of the different performance levels, incl. RAN, transport and core, based on the traffic mix expected in an area. In addition, a key requirement for any traffic classification solution is that most applications use multiple application flows with different requirements.

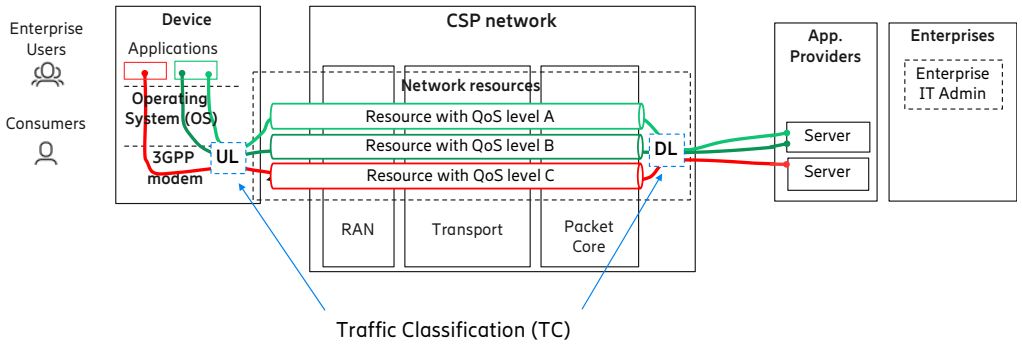


Figure 12 Traffic Classification

Existing 3GPP standards and products designed based on these standards are not fully prepared to support QoS for data applications beyond VoLTE/IMS. As an example, there is only 5 radio bearers available beyond existing applications (default internet, IMS, SMS/MMS/SUPL). This means that any new traffic classification based on URSP and/or NI-QoS will be solved with up to 5 radio bearers (in the smartphone case).

NI-QoS is standardized in 3GPP since long time and is used in commercial operation mainly for VoLTE/IMS. NI-QoS is an excellent basis for Network API as it provides dynamic and fine-granular application-flow-level traffic differentiation. NI-QoS is orthogonal to network slicing but can be used in combination with it. The main hurdle with wider adoption of NI-QoS to any applications have been the complicated integration between CSPs and ASPs, and all initiatives to encrypt traffic at Over-The-Top (OTT) level over the mobile networks e.g. due to end-user privacy needs.

URSP is standardized by 3GPP, since Rel-15, for a UE connected to multiple network slices. The 3GPP standards define multiple URSP Traffic Descriptors (TD) that would allow both application and application flow level mapping to network resources. Device OS

vendors have however taken their own initiatives on interpreting the URSP rules. They have also defined separate URSP solutions for consumers/B2C and enterprises/B2B.

The B2C URSP solution is called **category-URSP** and is based on the concept of Traffic Category (TC) at UE OS-level. The applications specify the TC(s) when setting up the communication, for example "Low Latency" and "High Bandwidth". Device OS vendors have their own Traffic Categories. A generalized concept is still needed to support all OSes, tethering, FWA etc. The CSP defines, controls and monetizes on subscription levels and QoS matching the TCs.

The B2B URSP solution is called as **enterprise-URSP** and is based on the Enterprise IT Admin defining how different enterprise applications are mapped to the connectivity services defined, controlled and monetized by the CSP. This solution likely needs to evolve to the granularity of enterprise application flows.

The more fine-granular URSP TDs defined by 3GPP would be a good basis for Network APIs but are currently being blocked by smartphone OS vendors claiming end-user privacy, NN/OI and technical limitations. We are looking at alternatives to let API-based services (via e.g. CAMARA) control when and how the traffic category solution is to be used. This would allow for activation only when certain applications are needing the service and could be seen as a network slicing compatible alternative to NI-QoS.

L4S is an IETF-defined solution promoted by Ericsson for time critical communications to ensure that latency-critical high-rate apps can detect the available bitrate while keeping a bounded latency. Ericsson drive for L4S is based on enabling fast RAN based end-to-end rate adaptation indication for applications to minimize latency. Traffic classification in L4S is based on NI-QoS. L4S can either be included as a generic network optimization feature available for all applications, and UEs, or support for it can be limited to specific subscriptions and/or traffic characteristics.

Future direction will require a Traffic Classification Toolbox addressing a wide set of needs to be able to handle the ongoing alignment, settlement and potential standardization initiatives existing in the market. Ericsson understands the need for supporting multiple solutions for traffic classification and handling, incl. NI-QoS, L4S and category-URSP and enterprise-URSP for the different needs and use cases, and the related network APIs. We see both the category-URSP and enterprise-URSP solutions as good initial solutions with support on the smartphone ecosystem side. For other use cases, device types and device OSes, we see possibilities in going to both application and application flow level granularity based on network APIs for NI-QoS and more fine-granular URSP TDs, and even combinations of the different solutions.

4.5 Service exposure

Our vision is a network with a high degree of abstraction where consumption of services by an application should not require knowledge of the details of how telco networks work but instead it should be possible to interact using a set of simple yet powerful APIs for business management as well as network capability access.

Currently CSPs are expanding the use of telecom to explore the exposure of network capabilities, e.g. to address enterprises. The network resources exposed must be made easy to consume and shaped to fit the needs and desired use cases of enterprises and their partners.

To be successful, CSPs need to expand their service portfolio and turn their network into a programmable platform with the capability to onboard new applications and aggregators while leveraging their existing connectivity offerings and combine them with cloud and edge offerings.

Exposure can be applied in different places, both in the network and in the device as illustrated in Figure 13 below (based on The Ericsson Global Architecture). Only the most important interfaces are described here.

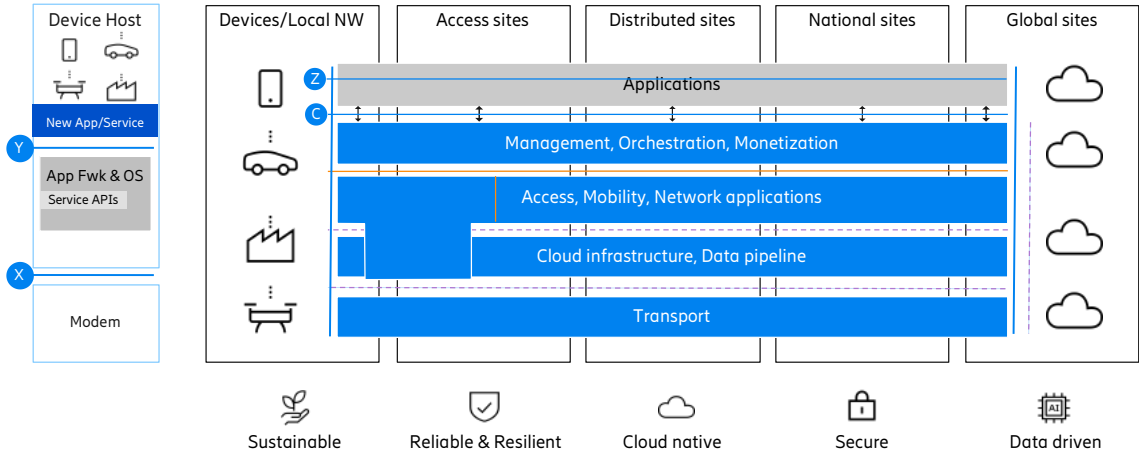


Figure 13 Exposure Interfaces

Z interface layer represents higher level and domain specific abstractions, interfaces, and services, within environments that Developer's trust, encapsulating/wrapping the C layer as needed. Aggregators typically expose APIs on the Z interface.
C interface layer contains a collection of northbound exposed capabilities and services of the network, reachable via Service Exposure Frameworks and its APIs/Protocols/SDKs, covering domains such as BSS, OSS, Packet Core and Communication Services.
Y interface layer is a collection of exposed abstractions of capabilities and services in Z and C from the device side. Exposed capabilities are similar but may also differ.
X interface layer is a collection of network services exposed via the modem / UNI interface, typically AT commands. Many standardized, but a large set of proprietary

Although the Z and C interfaces are expressed as thin lines (see Figure 13), these can contain a set of functions that are common to all exposed services, e.g., discovery, access control, identity management, throttling, etc. This drives a consistent experience towards different consumers of the APIs (developers, integrators, enterprises etc.) enabling scale and eliminates the need to use a proxy through the Management, Orchestration and Monetization layer.

An efficient and scalable service exposure architecture requires a service exposure platform that provides gateway functionality as well as a set of centralized supporting functions responsible for authentication and authorization of API invokers, API management and some functions for developers (such as API catalogues).

An Aggregation Platform, Figure 14 has the purpose of realizing and accelerating the development and availability of network services and APIs, the onboarding of CSPs and exposing APIs to application developers and ASPs (Application Service Providers) as well as to enterprises. Together with service exposure platform, it results in a scalable cloud platform with APIs and communication services including a channel to the developer- and enterprise eco systems.

The Figure 14 below shows the relation between the Aggregation Platform the CSP exposure and exposure towards the Developer ecosystem.

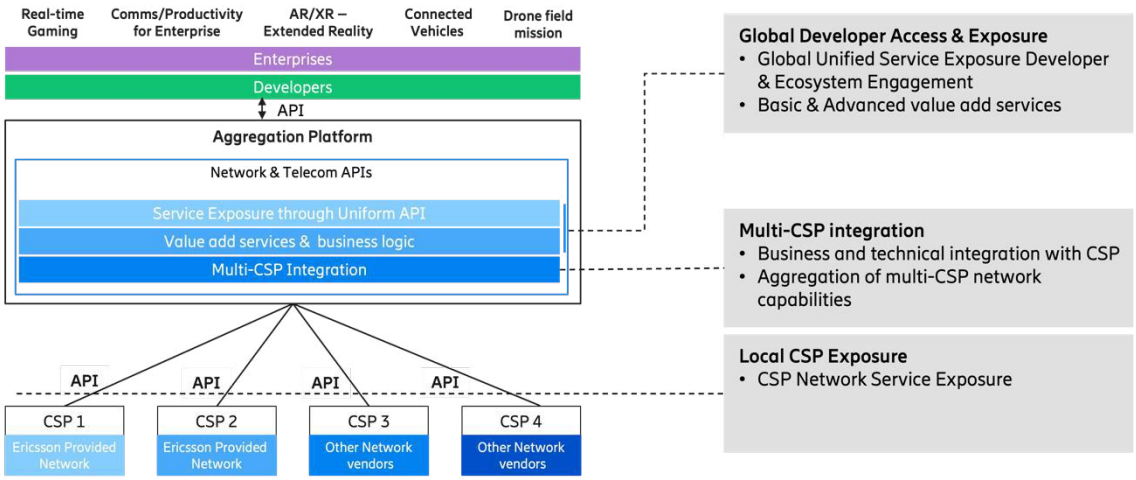


Figure 14 Aggregation Architecture overview

New commercial services for enterprises, will be exposed through the Aggregation platform, utilizing underlying network services exposed by the CSPs, like QoS, location etc.

It will integrate the APIs and services exposed by CSPs, using Ericsson or other vendors' products then aggregated and exposed in a consistent way so that developers, enterprises, and other platforms can consume the APIs in a uniform way on a global scale.

Ericsson always strive for an open ecosystem of different players consuming services from the CSPs. All players, including HCPs and other CPaaS players, will benefit from standardized / industry aligned APIs exposure from the CSPs, through instances like the CAMARA[1] initiative. This has been proven by e.g. the early global standardization of voice and SMS leveraged successfully by e.g. Vonage and other CPaaS players.

Additional value will be continuously added by utilizing the exposed network APIs, evolving into more advanced solutions with increased 5G network deployments.

Related articles/additional reading:
Programmable 5G for the Industrial Internet of Things
Monetizing API exposure for enterprises with evolved BSS

4.6 Zero Trust Architecture – ZTA

Historically (enterprise) networks have been built to keep out an attacker with different perimeter security control. Today this is not enough as we already have the adversary in these networks directly or indirectly which means perimeter protection is insufficient. Today's networks shall rather be seen as untrusted and correct security measures must be deployed to protect the users (subjects) and resources that reside on these networks. NIST published a document in 2020 [2]) that defines and discusses principles for protecting subjects and resources in untrusted networks. The focus of the NIST document is

[1] CAMARA - Telco Global API Alliance, is an open-source project within Linux Foundation

enterprise networks, but most of what NIST states is equally applicable for any network including 3GPP networks.

The term Zero Trust (ZT) is frequently used with as many meanings as there are speakers and writers. The term Zero Trust (ZT) is used in talks, discussions, papers, etc. with as many meanings as there are speakers/writers. ZT as a term has no agreed definition, but NIST has defined ZTA [2] .The ZTA principles are not new, and many have been used for years, but NIST now collected them in one document to describe the principles for security interaction over any network.

Principles of a Zero Trust Architecture (ZTA)

Zero trust is a concept in which digital systems cannot earn trust the way humans do. No network user, packet, interface, or device can be assumed to be trusted.

The implementation of ZT affects all assets, including digital systems, people, and processes. Zero trust has evolved from a concept to a ZTA, defined by NIST[2], with the principle that "there is no implicit trust granted to an asset based upon its ownership, physical location, or network location". This is a paradigm-shift for securing mobile critical infrastructure, particularly 3GPP networks, which traditionally have been within CSP premises and secured with perimeter security under the assumption that the internal network is trusted based upon ownership and location.

An important aspect of ZTA is to, even if new security controls are implemented, follow the ZTA tenets. The perimeter protection shall not be removed. The defense in-depth paradigm still applies and perimeter protection and ZTA security controls should coexist to protect against the threats.

Table 1 ZTA tenets

T1	All data sources and computing services are considered resources
T2	All communication is secured regardless of network location
T3	Access to individual resources is granted on a per-session basis
T4	Access to resources is determined by dynamic policy
T5	The enterprise (operator) monitors and measures the integrity and security posture of all owned and associated assets
T6	All resource authentication and authorization are dynamic and strictly enforced before access is granted
T7	The enterprise (operator) collects information about the current state of assets, network infrastructures and uses it to improve its security posture

Guidance of ZTA in 5G networks

The migration of 5G critical infrastructure to private, hybrid, and public cloud deployments introduces new actors and stakeholders to a shared-responsibility ecosystem. The new security paradigm for 5G infrastructure, including both RAN and core network, is based on a ZTA to protect against external and internal threats.

3GPP has concluded a study on how the NIST zero trust security principles/tenants are fulfilled in 5G core during 2023 and identified set of gaps. A second study is ongoing to look at these gaps and see what new security controls 3GPP can standardize to fulfil a

ZTA. Examples of areas of investigation are monitoring on network functions and possible new interfaces needed.

Security architecture in the evolution of 5G

Current 5G security architecture implements parts of the ZTA tenets in different places of the architecture. For some parts like UE access, most of the ZTA tenets are implemented either based on 3GPP standards or based on vendor implementations, but in other areas improvements are needed to become a ZTA. An important part of ZTA is that in a 5G network ZTA must be addressed on different levels, see Figure 15.

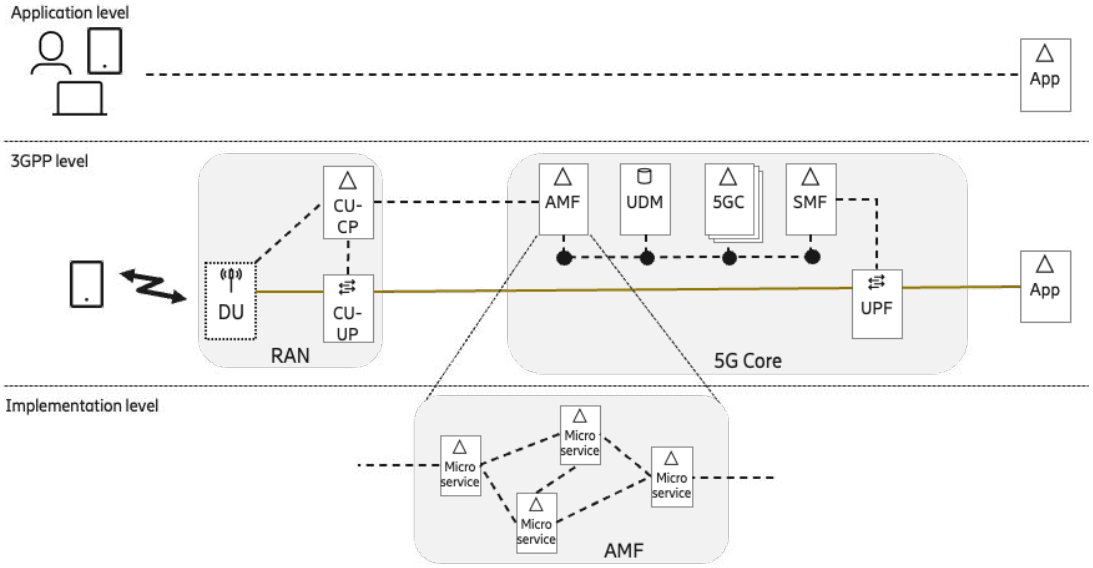


Figure 15 Different levels on which to apply ZTA

Starting from the top level, the user (of a device) is accessing an application (consumer or enterprise) and security here is mostly handled OTT provided by the application provider. Security controls like authentication and access control plus secure communication (confidentiality and integrity protection) is based on what the application provider has implemented. 3GPP has standards that can be utilized by the application provider like GBA(Generic Bootstrapping Architecture) and AKMA (Authentication and Key Management for Applications).

The middle level is where 3GPP comes into play even if some of the ZTA tenets today are outside the scope of 3GPP and rely on vendor implementations. As the scope of 3GPP is about functionality of UE and network functions and interfaces, implementation aspects of how NFs are implemented and deployed are not part of the scope (except in a few work items like Network Equipment Security Assurance Scheme(NESAS)/Security Assurance Specification(SCAS). 3GPP networks have some compliance to ZTA security controls, but it depends on the part/domain of the 3GPP network. In case of 5G Core and SBA, many aspects of ZTA are implemented in relation to the 3GPP scope, but other domains/parts of the network have fewer principles in place. ATIS(Alliance for Telecommunications Industry Solutions) has prepared a report[3] with gaps and proposed mitigations to make 5G a zero-trust architecture. Ericsson was co-chairing this work group.

To consider security aspects, it's important to start this work early in the ongoing 6G architecture discussions rather than becoming an afterthought. The 6G security

architecture should be an evolution of 5G and the ZTA aspects considered already now in the standardization work.

Telco network functions are more and more using cloud native principles for improved scaling, management and deployment options and there we have the bottom level when it comes to ZTA. In cloud native, network functions are built up based on micro-services that from a security perspective introduce new attack surfaces that must be secured. Governments and regulators have new requirements on operators and for example CISA(Cybersecurity & Infrastructure Security Agency) [4] has provided four guideline documents regarding how to secure 5G networks in a cloud infrastructure which are very much in line with Ericsson's security requirements, architecture principles and guidelines used for product development.

5 Network architecture domains

5.1 6G Architecture Direction – The 2030 perspective

The goal of 6G is to improve performance by meeting high demands on traditional performance indicators such as capacity, coverage, bitrates, and short latency, as well as new performance indicators related to service availability and assurance, predictability, network resilience and trustworthiness all while improving energy performance and sustainability. As well as meeting these performance indicators a cost-effective deployment and smooth introduction into existing networks must be ensured.

To support the 6G visions and requirements, it is beneficial to standardize a 6G architecture that allows for a smooth introduction of 6G capabilities into the future public and private networks. We foresee that the 6G architecture will build on the ongoing trend of network horizontalization enabling the 6G RAN and CN functions to benefit from the fast evolution of cloudification, IT frameworks, automation, open interfaces, and AI/ML.

Overall Architecture Direction
The 6G architecture needs to support the expected new use cases and service requirements for the 2030 timeframe and beyond. This will for example include enhanced support for immersive communication and new capabilities like network sensing and zero-energy devices.

There is a need for an aligned industry view of a single 6G architecture, avoiding the multiple architecture options defined in 5G which caused delayed availability of 5GS network capabilities and at the same time reducing the overall complexity of the 6G standard. Drawing on the experience from 5G, the 6G core should be an evolution of 5GC. This will not only reduce time to market for new 6G features but lower the costs for integration and testing in the operators' network.

This seamless reality of the future will provide new ways of meeting and interacting, new possibilities to work from anywhere, and new ways to experience faraway places and cultures making digitalization of industries and management of smart cities, enabling e.g. improved personal safety, less waste and improved sustainability possible.

Below Figure 16 illustrates trends for network towards 6G and the 2030 timeframe.

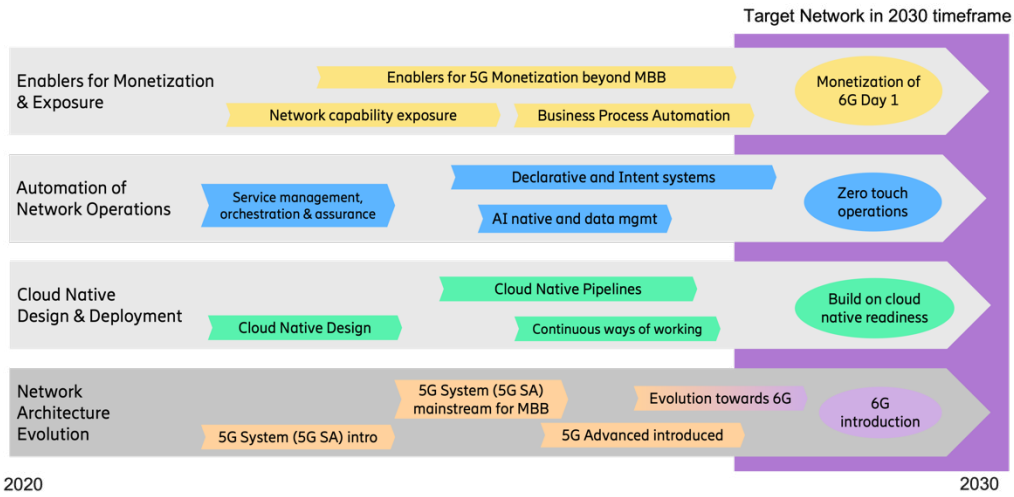


Figure 16 Multiple evolution trends towards the 6G time frame

A top priority for operators today is to monetize 5G capabilities to improve top line revenue capture. This ongoing journey will continue in 6G to make it possible for 6G networks from day one to reuse and expand on the evolution of 5G exposure and monetization functionality.

Automation of business process includes the automation of processes handling the full lifecycle of customers, partners, suppliers, products, orders etc., both from commercial and operational aspects. A successful automation also includes aligned and easy to use APIs, exposing relevant services to external actors, to enable the automation processes to spawn business borders, e.g. application provider – aggregator – connectivity provider.

Another trend is the leveraging of the IT eco-system, realized by the adoption of cloud native design and deployment. The cloud native design with containerized deployments enables an efficient separation of SW from HW.

The final trend is about network architecture evolution itself. It will be key to focus on network architecture being an evolution of the 5G System (5G SA) to 5G Advanced and further to 6G.

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

The 6G architecture is based on the principle of horizontal separation of the network functions from the underlying platform and overlying e2e management and exposure. More detailed aspects of this architecture are further discussed in the following sections.

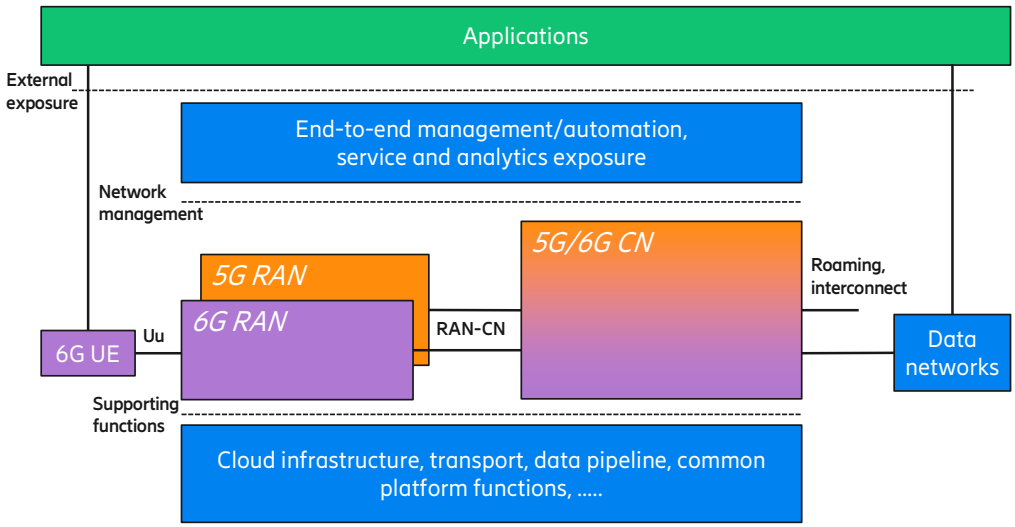


Figure 17 Future 6G Architecture with key open interfaces between domains (open i/f between E2E management and domain Management is not shown)

Related articles/additional reading:
[6G network architecture – a proposal for early alignment]

5.2 RAN migration, spectrum and standardized architecture

The 5G standard specified multiple connectivity options, including Non-Standalone (NSA) 5G (NSA) and Standalone 5G (SA), lead to a fragmentation in the market, and a delay of introducing 5G services. Based on these experiences, the 6G RAT should be specified only in a Standalone mode, whereby the UE is connected only to 6G from the start. This provides an aligned industry focus, avoiding the need for launching multiple versions of 6G whilst simplifying the technical architecture.

In 2030, the legacy spectrum assets (FDD, mid-band TDD and high-band TDD) are expected to be extended with new centimetric range spectrum (7-15 GHz) for 6G, possibly subject to co-existence with incumbent non-3GPP use. A 6G-connected UE should have better performance than a 5G UE in the same location. This means that 6G must be able to use potential new centimeter-wave bands together with legacy FDD and TDD bands, to improve throughput aggregation and uplink performance/coverage.

The 5G standard included two methods to combine with legacy spectrum, namely Multi-RAT Dual Connectivity (= Non-Standalone 5G) and 4G-5G Dynamic Spectrum Sharing. Besides the reasons to avoid NSA stated above, the Dual Connectivity solution also proved to be technically complex for the industry, with large impact on the old RAT and simultaneous uplink transmission in UEs. The split control of the UE connection between the gNB and the eNB implies a tight coupling of the nodes, given that the RATs share a common set of UE capabilities. Hence, an efficient dynamic spectrum sharing mechanism needs to be standardized from the start, allowing 5G and 6G UEs to share a common pool of resources. Given that 5G has significantly less need for always-on reference signals than 4G LTE, there are opportunities to significantly improve efficiency compared to 4G-5G spectrum sharing.

With 6G deployed on a mix of legacy and new spectrum, 6G needs to provide a good spectrum aggregation solution. To allow full RAN optimization and avoid complex interactions such as the Dual Connectivity in 5G, this should be based on a single instance

in the network to control a given UE, with mechanisms for fast adaptation on what spectrum is used at any point in time. Thus the full set of cells used for a UE connection is decided in one place, considering the full capability of the UE for different band combinations.

To meet increasing demands on efficiency using deployed resources (spectrum, radio sites), and increasing requirements of energy efficiency, a higher degree of elasticity and pooling of radio resources is needed. This includes using multiple radio sites and spectrum resources for connected mode transmission (e.g. Distributed-MIMO), when needed, but also the ability to power-down radio sites or spectrum resources when not needed. This dissolves to some degree the concept of a "node" in RAN to be represented by a physical base station at one radio site. Instead, RAN performance will be optimized per geographical areas, using the radio sites and spectrum in that area per current needs.

In such a RAN system, the functional dependencies between different parts of the system controlling the resources will be even higher than today. The industry must be very careful to select the key interfaces to be standardized for potential multi-vendor integration, considering both the business values and the feasibility from a separation of concerns perspective. A possible high-level 6G RAN architecture is illustrated in Figure 18.

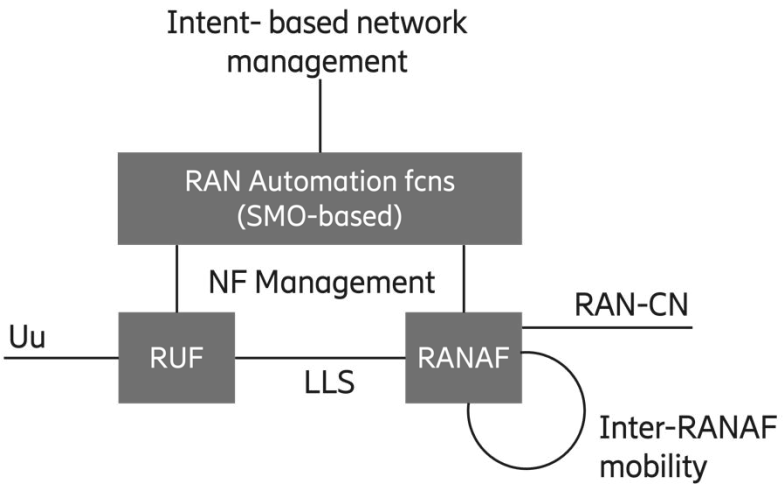


Figure 18 Potential 6G RAN Architecture

The LLS interface is included as a key interface natively supported in 6G RAN, splitting the RAN into two logical network functions – the Radio Unit Function (RUF) , and the RAN Area Function (RANAF) . The complex control of resources across radio sites and spectrum is hence contained within the RANAF, allowing e.g. efficient carrier aggregation across RUFs from different vendors. This architecture opens the door for a RAN to optimize across a geographic area by connecting multiple radio sites with LLS to one RANAF, while maintaining the LLS as a key standardized interface.

Another key interface is an inter-RANAF mobility interface between two RANAF instances. Focus is on transfer of the UE control from a source RANAF to a target RANAF, and not on a complex interface to share control for the same UE (e.g. for Dual Connectivity).

From a management point of view, the 6G RAN evolves to be automated exposing a standardized intent-based management interface via the SMO-based RAN automation functions. Automation will be done in both the Network Functions themselves, and in RAN

automation functions ("rApps"), leveraging evolving AI/ML technologies. There will also be standardized network function management between the RAN automation functions and the RUF and the RANAF.

5.3 RAN implementation optimized for 5G & 6G

By 2030, the RAN needs to optimize for a mix of 5G and 6G, including wide use of 5G-6G spectrum sharing on many legacy bands. Also 4G, especially for m-IOT, needs to be supported, but not optimized for, and hence not described further here.

The vision is that RAN performance is self-optimized by the RAN solution:

- Given available network resources(e.g. , spectrum, sites, transport) and UE population,RAN controls available resources to optimize performance in a geographic area, according to operators' intent
- Performance dimensions: throughput, latency, capacity, energy consumption, NRAR, scalability, TCO, ease of use, security, ...
- Differentiate by service types & user groups / slices
- Intents to balance between performance dimensions, service types, user groups and UEs with different radio conditions

The need to achieve the full optimization potential, aggregation/coordination between frequency bands and between radio sites as discussed above for 6G will apply also to 5G, although naturally limited by 5G specifications and UE capabilities. One possible solution could be a scalable scheduler architecture, that can provide full performance potential to coordinate over a high number of sector-carriers in e.g. Centralized RAN deployments, also possible to be scaled down to a medium or low number of sectorcarriers for traditional Distributed RAN deployments, and everything between.

RAN should leverage possibilities with a multipoint fronthaul, using IP routing for flexibility & resilience, pooling gains by decoupling RU resources from user-plane processing in baseband, reducing needs to peak dimensioning the fronthaul transport network and automated orchestration between RAN and transport domains.

By 2030, a common architecture supporting deployments on Cloud infrastructure, Purpose-built infrastructure or a mix should be available leveraging the continuously evolving solutions from the IT industry, such as cloud infrastructure, data pipelines, automation and AI/ML support functions.

Related articles/additional reading:
Improving energy performance in 5G networks and beyond
The 5G Advanced, an evolution towards 6G - Ericsson
6G spectrum - future mobile life - Ericsson
AI-native RAN – generalization and scalability of learning - Ericsson
Broad beamforming technology in 5G Massive MIMO - Ericsson

5.4 Core Network

The development of 5G provides some learnings worth noting when defining the future network architecture. Since cloud native design and deployment offers new implementation technologies and improved ways of working including software LCM, it inspired the introduction of the Service Based Architecture (SBA). This made the functional architecture of the 5G Core Network (5GC) in 3GPP more suitable for cloud

deployment. The SBA makes the Network Functions (NFs) expose services via Service-based Interfaces (SBIs) instead of point-to-point protocols as in previous generations. The purpose of this change was also to create a more flexible and extensible architecture.

The extensibility is proven by the continuous increase of the number of NFs in the 5GC, from 22[2] in 3GPP Rel. 15 to 45[2] in Rel. 17. The number of SBIs has increased almost at twice the rate as the NFs, reaching more than 110 SBI in Rel. 17. The increase of NFs and NF services is strongly related to the introduction of new features and functions in the 5GC.

However, the flexibility of 5GC also drives complexity in standardization, development, and operational deployment in commercial networks. E.g., the Service Communication Proxy (SCP) introduced in Release 16 added several modes of operation of the inter-NF communication. These modes of operation need to be applied on a per SBI and NF service consumer. With different NFs supporting different modes of operation, this leads to a configuration complexity for multi-vendor SBIs of networks in operation. This intricacy opens for the need of optimizations when evolving the architecture further by, e.g., considering how to minimize additional complexity at system level at introduction of new functionality. Preparation to manage the complexity in all steps from specification, development to operations will be important, e.g., using automation.

To further smoothen the introduction of 6G, the reuse of the industry's investments in 5G Core is important. This includes the granular QoS framework, the comprehensive support of network slicing, support for time-sensitive and reliable communication, and exposure of these network capabilities to address new end-user service opportunities. A part of the evolution of 5GC also includes possibilities for optimizations and simplification of the SBA, for example by reducing the dependencies across NFs or removing unnecessary flexibility to make the standardized architecture future proof. It could also be considered to bundle new features and functions with existing NFs, e.g., based on main consumer of the function, to reduce the pace of increase of the number of NFs and SBIs.

In addition new use cases requiring new or evolved functionality can be expected in the evolution towards a new 6G RAT.

While 5GC is expected to evolve with NFs being common to 5G and 6G, 6G-only NFs must be expected. Other aspects that need further exploration is the area of energy efficiency in the core network deployments. This can partly be addressed in the functional architecture to support new RAN efficiency methods. Still the major opportunity is to enable power savings in the implementation architecture and the underlying cloud infrastructure, rather than in the 3GPP standardized architecture.

With the approach to base the core network on an evolved 5GC as opposed to start from a clean sheet of paper, the operators will be best positioned to monetize 6G from the beginning.

Related articles/additional reading:
Indirect communication for service-based architecture in 5G core
Energy-efficient packet processing in 5G mobile systems
Toward 5G Advanced: overview of 3GPP releases 17 & 18 - Ericsson

[2] Not all NFs expose SBIs in the release where they are introduced, but often consume SBIs.

5.5 Communication services

An IMS-based communication system has been the foundation for voice services in 4G and 5G. It is assumed that IMS will be the communication engine also for future regulatory telephony communication services like voice and emergency services in 2030/6G timeframe. The main reasons include the maturity and wide support in both networks and device eco-system. It can also be expected that 6G interworking with 5G (and 4G) and Wi-Fi voice using IMS will be required.

Technology and architecture for realization of new conversational use-cases is an ongoing industry discussion for 5G already. A prime example is XR conversational services, which are studied in 3GPP. Different models are considered, ranging from IMS extension to pure OTT models. We anticipate the usage in 5G will be the baseline for 6G.

Related articles/additional reading:
Voice and communication services using 4G and 5G

5.6 Data Pipeline

Data is key to several processes and currently CSPs have created multiple independent data pipelines for specific purposes, to rigid and difficult to evolve.

The result is inefficiency by replicating the same data multiple times as well as bringing an approach which is hard to govern. We foresee an evolution where data management capabilities become a fundamental part of the CSP's architecture and are introduced in a way that enables a stepwise and multivendor approach. Data management is a cross-cutting concern covering multiple domains, data sources and data of varying characteristics.

There are a few principles to follow:

- Collect once, use many
  - Data is collected once and can be used by many consumers
- Data is managed in a federated approach
  - There can be multiple data islands that are federated.
- Data is used in transparent, compliant and ethical ways
- Data Quality is ensured during the data lifecycle
- Data access is authorized enabling legal and ethical use.

Technically this also means that:

- An application can discover data which it can consume.
- An application can request to receive data to be consumed.
- Any resulting insights can be feedback as data into the data management functionality.
- Data lineage is supported.
- This enables data governance.

The data ingest architecture (for which further details can be read in below article Ericsson Technology Review -Data ingestion architecture for telecom applications) is represented below:

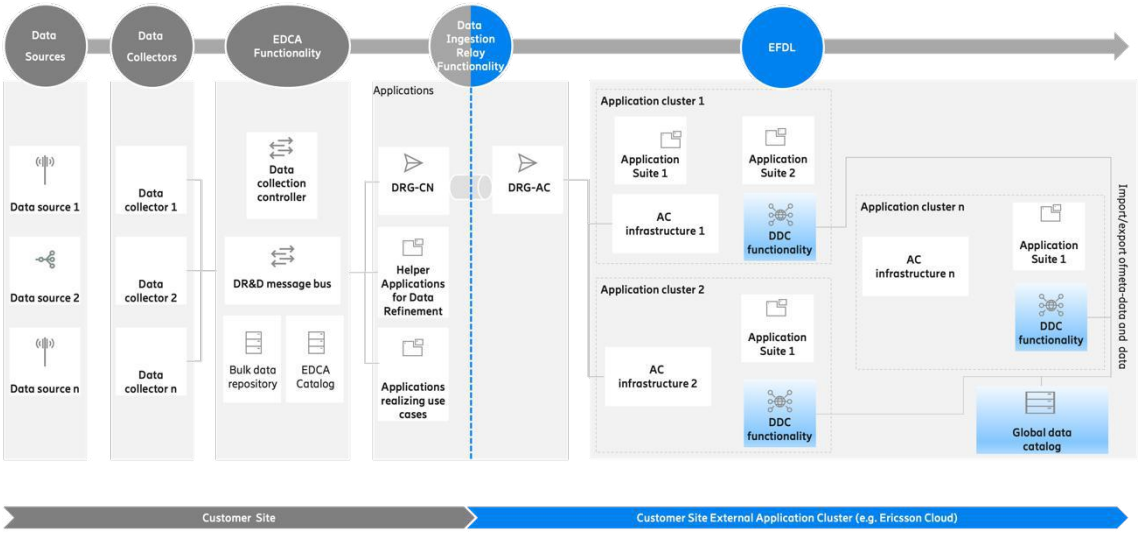


Figure 19 Data Ingestion Architecture

Several data collection instances may exist in a network and such an instance can be part of multiple functions in the network which work in a federated approach – e.g. it can be part of the SMO as one data set of data islands.

Related articles/additional reading:
[Data ingestion architecture for telecom applications]

6 Network architecture examples

6.1 Network Deployment Cases

3GPP based telecom networks are highly suitable as the base architecture for many different scenarios, enabling multiple use cases for several customer types.

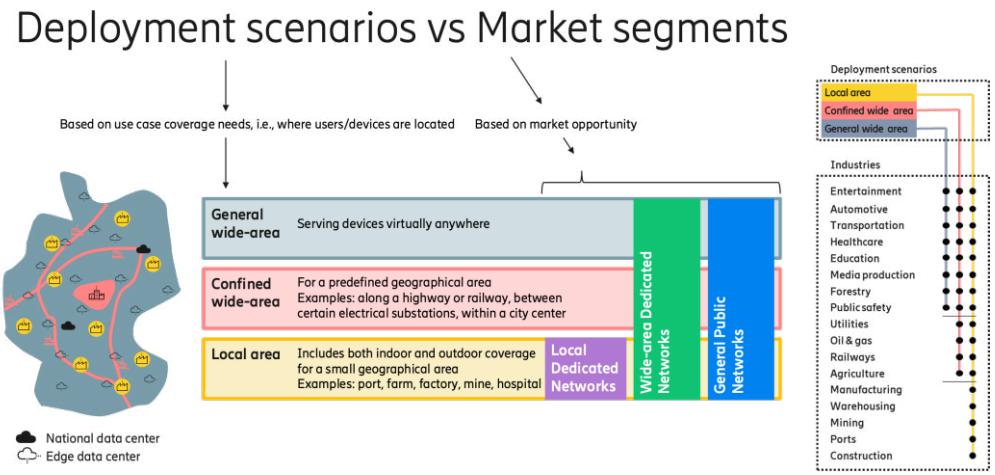


Figure 20 Deployment scenarios vs Market segments

Above Figure 20 describes three main market segments with industrial deployment scenarios stipulated. This should be viewed as a guide and variations will exist.

General public networks
The segment covers the mobile operator-provided consumer and business services (time-critical and non) and include for example Internet/data access, communication services, emergency services and XR/cloud gaming.

Wide-area dedicated networks
Wide-area dedicated networks may be deployed with general wide-area coverage or limited to a confided wide-area, meaning a limited predefined geographical area, coverage. It provides similar type of services as in the General public networks as well as enterprise services with enterprise APN and VPN, public safety, etc. and additional services like, e.g. mission critical communications services (e.g., Rail).

Local dedicated networks
Services for Local dedicated networks include connectivity and communication services for a specific local area and industry like a Mining site, a Manufacturing site, Hospitals, Airports/Ports, Stadiums, University Campus, Enterprise Campus, etc.

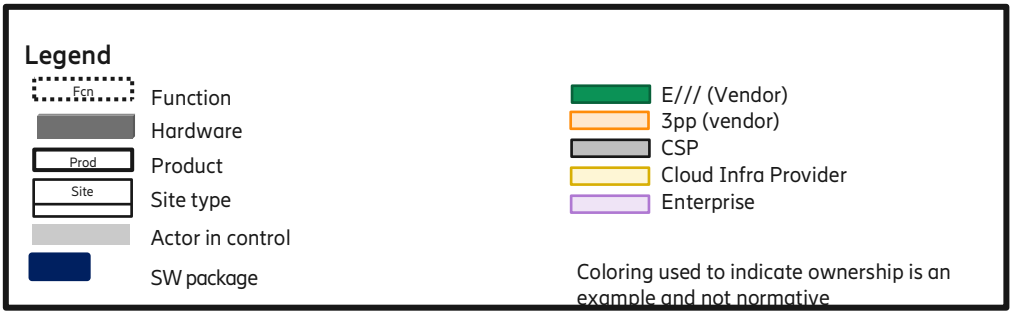
From a RAN perspective, such a local-area private network may be either realized by means of an infrastructure mapped over MNO resources or as a dedicated infrastructure with own resources. A Local dedicated network cannot support use cases requiring wide-area coverage.

A few examples of considerations to be made of functionality to be deployed across all three segments:

1. It is necessary to provide mechanisms for local and central network exposure. Deployment cases are different, but the exposed services are the same.
2. It will be central to support network slicing to allow partitioning of the network for use cases requiring differentiated characteristics.
3. NRAR (chapter 4.3) is one example of several characteristics that must be supported to ensure, in this case reliability and resilience for all above segments.
4. When developing network functions a system wide context must be understood to support a flexible usage in a large set of deployments and solutions.

Below examples describes on a high level the different segments from above. Note that these are examples with the possibility of both diversity and mutations.

Table 2 Network deployment cases Legend



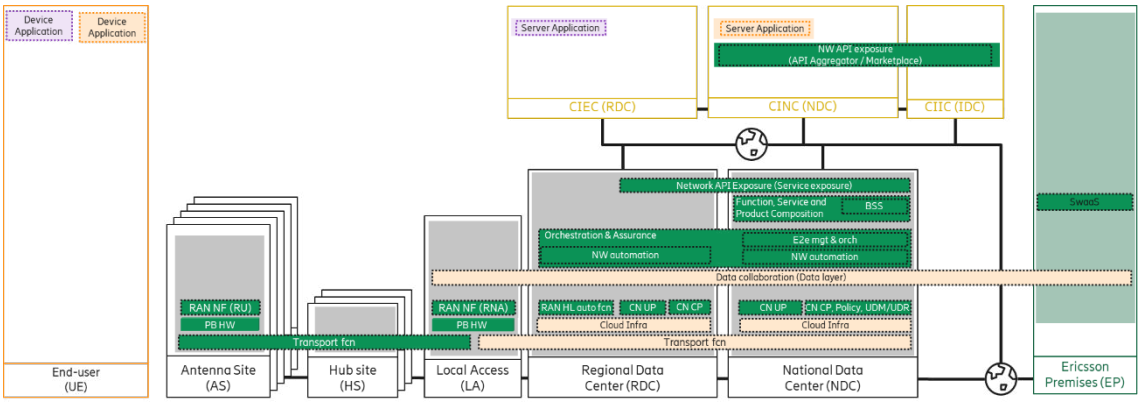


Figure 21 General public network horizontal architecture view deployed in network operator premises

In this deployment (Figure 21), the CSP owns the hardware and infrastructures of the General public network, while network API exposure function has shared ownership between CSP and Ericsson. The example shows a third party providing the cloud infrastructure in the CN sites while the RAN sites rely on purpose-built equipment.

The example depicts a third party delivering the data layer product, deployed in both CSP's sites and Ericsson premises.

Ericsson provides the applications for Network API Exposure layer. The API aggregator/ marketplace application is owned by Ericsson and deployed in a public cloud provider's data center.

The second example (Figure 22) is a Wide-area dedicated (confined) network and typically deployed by/for private enterprise/government and implies it's restricted to a limited set of users.

Examples can be railroad, public safety networks, electricity grids, etc. In some markets these types of dedicated networks can be provided as dedicated slices in a General public network, though not depicted here.

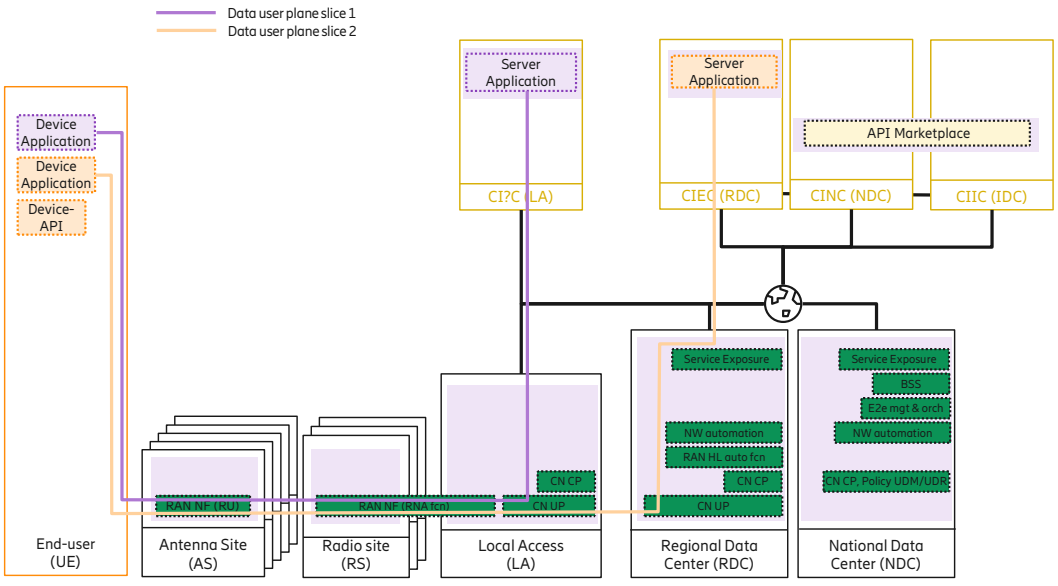


Figure 22 Wide-area dedicated (confined) network

This network requires additional characteristics such as high coverage and availability from Public Safety and latency for Rail as low as below 10ms (one-way e2e) with additional requirements of high throughput in both up- and downlink.

Slicing can be used for use-case separation and to ensure required characteristics. A distributed Core UP is another option to support, e.g. critical industries.

Several models are possible (already in use) for ownership of a wide-area confined network.

Exposure for the Wide-area dedicated networks is different compared to General Public networks with fewer users, but with high accessibility and security requirements, thus limited need for aggregation functionality.

Figure 22 shows an example of an internal (to the owner of the Wide-area dedicated network) API Marketplace provided for application developers which could be hosted either in a public or private cloud.

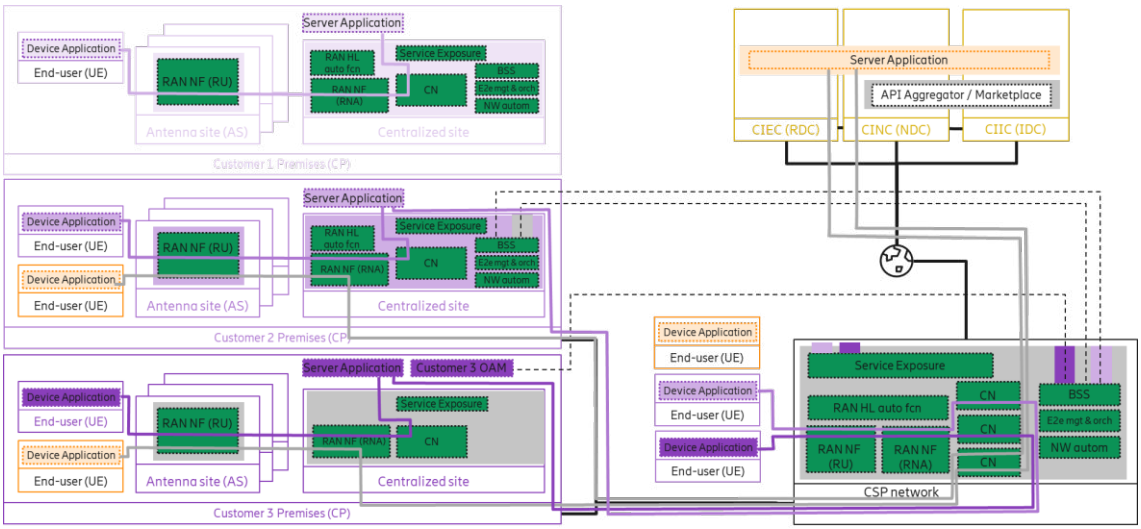


Figure 23 Local Dedicated Networks

Local dedicated networks are typically deployed for a private enterprise and three cases are described above in Figure 23:

- **Customer 1**, where the equipment is owned and managed by the enterprise and only used within the enterprise premises with private local devices.
- **Customer 2**, where the equipment is owned and managed by the enterprise. The usage is both local within the enterprise and external to a CSP network. The devices in this case can be both private local devices as well as external devices with CSP connection.
- **Customer 3**, where the equipment is local to the enterprise but owned and managed by a CSP. The usage is both local within the enterprise and external to a CSP network. The devices in this case can be both private local devices as well as external devices with CSP connection.

The two last examples offer the additional possibility of providing, e.g. slices in the General public network for the enterprise application.

Important characteristics to the use cases in this segment are privacy, throughput and latency as low as <5ms. Mobile broadband is still required as is a communication service (voice) due to the mix of uses cases in e.g. manufacturing.

Several areas exist where service exposure will be necessary. The Service Exposure functionality/API Marketplace can be owned by an enterprise or access may be provided to a marketplace provided by the CSP.

Several ownership models for dedicated networks exist with two major distinctions: 1) CSP provides local dedicated network functionality to enterprises, or 2) enterprises own the infrastructure/network functionality.

7 Abbreviations and Definitions

3GPP	3rd Generation Partnership Project
5GC	5G Core
5GS	5G System
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
ARMI	Automated RAN Management Interface
AS	Application Server
ATMI	Automated Transport Management Interface
BCSS	Business Area Cloud Software and Services
CD	Continuous Development
CI	Continuous Integration
CI/CD	Continuous Integration/Continuous Delivery
CI/CDD	Continuous integration, delivery, and deployment
CN	Core Network
CN	Customer Network
COTS	Commercial Off The Shelf
CP	Control Plane
CPaaS	Communication Platform as a Service
CSP	Communication Service Provider
DDC	Data Distribution Central
DDD	Data Driven Development
DDDE	Data Driven Development Environment
DI	Data Ingestion
DL	Downlink
EDCA	Extensible Data Collection Architecture
ETSI	European Telecommunications Standards Institute
GDPR	General Data Protection Regulation
GNP	Global Network Platform
HCP	Hyperscale Cloud Provider
HW	Hardware
IBA	Intent Based Architecture
ICO	Infrastructure Company
IMF	Intent Management Function

IMS	IP Multimedia Subsystem
ISP	In Service Performance
IT	Information Technology
KaaS	K8s-as-a-Service
L4S	Low Latency, Low Loss, and Scalable throughput
LCM	Life Cycle Management
MBB	Mobile Broadband
mIOT	Massive IOT
ML	Machine Learning
NAP	Network Automation Platform
NESAS	Network Equipment Security Assurance Scheme
NF	Network Function
NFV	Network Function Virtualization
NI-QoS	Network Initiated Quality of Service
NIST	National Institute of Standards and Technology
NN	Network Neutrality
NRAR	Network Reliability, Availability and Resilience
NSPS	National Security Public Safety
NTN	Non-Terrestrial Network
NWaaP	Network as a Platform
OCS	O-RAN Software Community
OPEX	OPeration EXpenditures
ORAN	Open Radio Access Network
O-RAN	Open RAN
OS	Operating System
OSS	Operation Support System
OTT	Over The Top
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAR	Reliability, Availability and Resilience
RAT	Radio Access Technology
SaaS	Software as a Service
SBA	Service-Based Architecture
SCAS	Security Assurance Specification product evaluation and testing
SLA	Service Level Agreement
SMO	Service Management and Orchestration
SOM	Security Orchestration and Management
SW	Software
TC	Traffic Classification
TCC	Time Critical Communication
TCO	Total Cost of Ownership
TD	Traffic Descriptor
TFT	Traffic Flow Template
TMF	TeleManagement Forum
TN	Transport Network
TTM	Time To Market
UE	User Equipment
UL	Uplink
UP	User Plane
UPF	User Plane Function
URLLC	Ultra Reliable Low Latency Communication
URSP	UE Route Selection Policy

VoLTE	Voice over LTE
ZTA	Zero Trust Architecture

Intent can be defined as a "formal specification of all expectations including requirements, goals and constraints given to a technical system". It states which goals to achieve rather than how to achieve them. Intent enables the creation of autonomous sub-systems rather than creating tightly coupled management workflows.

Cognition is a psychology term referring to an "action or process of acquiring knowledge, by reasoning or by intuition or through the senses" [Oxford]. Using cognitive technologies makes it possible to implement a technical system with cognitive capabilities using e.g. AI techniques including Machine Learning (ML) and Machine Reasoning (MR).

8 References

[1] "The ITU-R Framework for IMT-2030". ITU, July 2023

[2] NIST SP 800-207, Zero Trust Architecture, https://csrc.nist.gov/pubs/sp/800/207/final

[3] ATIS https://www.atis.org/tops-council/enhanced-zero-trust-and-5g/

[4] CISA https://www.cisa.gov/resources-tools/resources/security-guidance-5g-cloud-infrastructures

[5] "Transforming our world: the 2030 Agenda for Sustainable Development". United Nations, 2015.

[6] Jörg Niemöller, et. Al.: "Creating autonomous networks with intent-based closed loops", Ericsson Technology Review, April 2022 https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/creating-autonomous-networks-with-intent-based-closed-loops

[7] "5G Value: Turning Performance into Loyalty". ConsumerLab report. Ericsson, 2023.

[8] "Digitalization Strategy for Business Transformation". Gartner.

[9] "Why every decision on 6G must put sustainability first". Ericsson blog article, August 2023.