Good architecture forms the foundation of communication networks. The future network will be divided horizontally. Increasing needs from enterprises and others will require a high level of programmability/configurability in the future network. This also calls for automation, machine learning and simplification to support the important use of abstraction to hide potential architecture complexity as it evolves into a cognitive network. This is a description of the network architecture and how it evolves over time.
1 Executive summary ................................................................. 4
2 Network trends ........................................................................ 6
3 Network Capabilities .................................................................. 8
   3.1 Service and Management Automation .................................. 8
   3.2 Artificial Intelligence ....................................................... 9
   3.3 Network slicing ............................................................. 10
   3.4 Network Reliability, Availability and Resilience ................. 11
   3.5 Security ......................................................................... 12
   3.6 Service exposure ......................................................... 13
4 Network architecture domains ................................................... 15
   4.1 Introduction .................................................................... 15
   4.2 Management, Orchestration, Monetization .......................... 17
   4.3 Access, Mobility & Network Applications ....................... 19
   4.4 Cloud Infrastructure ..................................................... 23
   4.5 Data pipeline .................................................................. 24
   4.6 Transport ...................................................................... 25
5 Network architecture examples ............................................... 26
   5.1 Wide area public network/Evolved MBB ............................ 26
   5.2 Wide area Private Networks ........................................... 28
   5.3 Private and Local Dedicated Networks .............................. 31
6 CTO network trends ................................................................ 33
1 Executive summary

The unprecedented events surrounding the COVID 19 pandemic showed a rapid adoption rate of digitalization made possible thanks to the existing capabilities in the communications infrastructure, both the mobile and the fixed. It has increased awareness in both business and society that availability, reliability, affordability and sustainability are all essential aspects of the digital infrastructure.

Future communication technologies will enable a fully digitalized, automated and programmable world of connected humans, machines, things and places. Networks will be enhanced to support time-critical communication e.g. for AR/VR/XR and drones.

The desire for more advanced use cases in which the physical and digital worlds converge can already be seen in 5G and is even more evident in the emerging development of 6G.

The rapidly evolving digital transformation has given rise to a new paradigm in enterprise and across a wide range of industries displaying a multitude of specific requirements to serve distinct use cases in a number of different areas.

Society has grown to depend on the mobile broadband network being available 24/7 delivering a complete range of data and communication services turning the MBB (Mobile Broad Band) network into a society critical resource. Despite this, regulatory activity is rather low.

The need and ability for governments to handle various kinds of crises and the digital transformation for Public Safety agencies and Critical Infrastructure requires technology with mission-critical grade. One way of addressing this is to adopt standardized technology and ecosystem, like cellular 3GPP technology creating a symbiotic approach between Public Safety and Commercial networks.

This development of trends of the mobile broadband network becoming a society-critical service, enterprises and mission critical networks adopting 5G technology creates a universal need for Network Reliability, Availability and Resilience (NRAR) to identify and mitigate the weakest link in an E2E chain to protect network service characteristics.

Increasing security threats and attacks means all networks of today must be considered as untrusted. This is the assumption of the Zero Trust (ZT) security model adopted by communication networks.

Future deployments will be less node-centric for both RAN and Core network; cloud technologies will be essential. The cloud paradigm transforms how networks, network functions and applications are built, deployed and managed.

The external interfaces and the definition of the functional behavior of the Radio Access Network (RAN) are standardized by 3GPP and in O-RAN specifications. The high-level specifications, leave room for innovation to enhance the network with RAN-internal value-add features which, over the years, have resulted in continuous improvements. Determining an optimal RAN architectural split
requires a finer level of granularity than that offered by 3GPP using also e.g. preferred execution environment, spectrum efficiency, etc.

With the introduction of 5G Stand Alone deployments of 5G Core (5GC) deployments increase. Evolved Packet Core (EPC) will remain alongside 5GC due to the co-existence of 5G and LTE for some time. The use of eMBB for society critical use, new business verticals and wide-area private networks results in expectations being high from CSPs on 5G System (5GS) capabilities. Mission Critical Push-To-Talk (MC-PTT) is delivering the original service -communication services to the focus area of public safety.

For CSPs that seek to extend their offerings to address specific needs of enterprises, it is key to expand outside telecom to explore the exposure of network capabilities that are easy to consume and shaped to fit the needs and desired use cases of enterprises and their partners. To be successful, they should expand their service portfolio and turn their network into a programmable platform with the capability to onboard new applications to create a business-driven ecosystem of developers, applications, enterprises and their partners.

All above future evolution of the communication network, including the use of Cloud environments, will depend on a strong management and orchestration function. Much of the network intelligence will grow towards automation and the zero-touch vision developing into a data- and intent-driven cognitive network. AI is thus rapidly becoming important to aid with the increasing complexity in all layers of the network.

Going forward, 5G will be the main digital infrastructure for consumers with mobile and fixed wireless residential access supporting augmented/virtual reality and artificial intelligence (AI) based services.

6G is emerging though still in its conceptional phases. Some initial trends can be unveiled like limitless connectivity. A set of multi-vendor interfaces needs to be selected carefully to ensure openness in networks and the ecosystem while minimizing system complexity, ensuring development agility and a robust and resilient network.

The demands on the future networks drive for development into cognitive networks leveraging AI to satisfy our expected need for a resilient and available network not only in the traditional MBB segment but also in Industrial and Mission Critical segments. Finally we can start to envisage the future network architecture evolving as the initial phases of 6G are emerging.
Network trends

Future technologies will enable a fully digitalized, automated and programmable world of connected humans, machines, things and places. All experiences and sensations will be transparent across the boundaries of physical and virtual realities. Traffic in future networks will be generated not only by human communication but also by connected, intelligent machines and bots that are embedded with artificial intelligence (AI). Networks will be enhanced to support time-critical communication e.g. for AR/VR/XR and drones.

The machines and other ‘things’ that make up the Internet of Things (IoT) require even more sophisticated communication than humans do. For example, connected, intelligent machines must be able to interact dynamically with the network. In a digitalized and programmable environment, sensor data will be used to support the development of pervasive cyber-physical systems consisting of physical objects connected to collaborative digital twins. Future network capabilities will eventually include support for the transfer of sensing modalities such as sensations and smell.

The ability to support digitalization efforts of other industries is a major opportunity for CSPs going forward. Different needs apply to different industries as they digitalize their sales, services, products, supply, delivery, marketing, R&D etc. Future CSP competitiveness will rely on a much broader set of capabilities. Entirely new business models will be possible through advances in connectivity, software, mobile devices and cloud. These strategic shifts will increasingly become a matter of survival in the digital market. The term user will be seen from the widest context, it can be anything from a person to an actuator, a sensor, a data center, a content creator, etc.

The interconnect between different kinds of networks, from local to wide-area coverage, builds a global network that provides a platform for pervasive global services. The inherent mobility within and between the networks creates unprecedented coverage both indoors and outdoors. Utilizing all these network assets enables a distributed environment for access, compute and storage. These assets are virtualized, distributed across the network, and are made available where they are needed and are most efficient. Applications and processes are dynamically deployed throughout the network. Network slicing enables streamlined connections for different applications, enhancing the efficiency of the total usage of the network.

The network architecture becomes horizontalized and takes into consideration the cost of acquiring, installing and operating a future communication and connectivity network. It calls for automation, machine learning and simplification as well as supporting the important use of abstraction to hide potential architecture complexity.

Autonomous operation including deployment and assurance is an inherent capability of the network platform to enable cost-efficiency. The management and operation are becoming more specified in the industry to both support increasing automation and reduce the integration costs incurred by fragmented operations support systems (OSS). This allows increasing innovation targeting support for different vendors applications on an industry aligned automation
platform. The OSS evolves from managing the network as a resource to managing the network as a service with intent.

Built-in, automated security functions protect the network and the integrity of its users from external threats. Just as important are network reliability, availability and resilience to fulfil expectations from industry and society. These include support for new use cases with built-in criticality, mobile broadband becoming critical for society and transition to cloud and IT based architectures.

Future deployments will be less node-centric, and both RAN and Core Networks (CN) will have more common platforms and increasingly be deployed on Cloud. This removes some of the reasons to duplicate functionalities, such as having the RAN rely on the CN as a “data store” for idle devices. Consequently, it is important to revisit some architecture assumptions behind today’s functional separation between RAN and CN.

A factor that is common to all future deployment scenarios is the requirement for a superior transport network to be flexible, scalable, and reliable in order to support demanding future use cases and novel deployment options, such as a mixture of distributed RAN and centralized/cloud RAN. This is achieved by AI-powered programmability, multi-service abstraction/virtualization on heterogeneous networks, and closed-loop automation to keep transport networks flexible and manageable.

The horizontal architecture will ensure that interfaces are built with business needs as the primary driver. If well aligned in both content and time, it will catalyze and progress new business. Open business interfaces that support a flexible approach to building new services will provide for a shorter time to customer and nurture innovation across industry and society.

The network offering will be consumed through an automated digital marketplace. Network services and data are available through consistent and open business interfaces for the applications (APIs). Data, such as location, connectivity conditions and user behavior, can be made available from the network platform.
Network Capabilities

3.1 Service and Management Automation

Automation has broad applicability and covers several areas like delivery, deployment, data management, etc. This chapter describes mainly the Service and Management Automation aspect and how automation performs such task without human intervention.

The network will evolve to a data and intent driven cognitive network introducing a higher level of automation and optimization to lower OPEX, increase performance, robustness, and availability as well as to automate service provisioning and assurance.

The intent-based operations offer a higher level of abstraction to the network operator, changing the way networks are operated by describing what the network should do (the intent) instead of describing how it should be done. The changed approach to management, will shift the operator from managing networks manually to managing automation.

New technologies and concepts like cognitive frameworks for knowledge management and logical reasoning combined with data-driven knowledge creation and reasoning will enable the transition to such a cognitive network. Automation introduces a higher level of intelligence in the network, new learning architectures for multi-vendor deployments, new concepts for the RAN, Core and OSS to further optimize the performance and efficiency. The cognitive network will handle many of today's system management tasks autonomously.

A key part of managing automation is the control loops. This is achieved by designing the insights and policies required for a use case, and the decisions that need to be made.

Insights are the analytics outcome when data is processed. Policies represent the rules governing the decisions to be made by the decision function in the control loop. The decisions are actuated by the execution function, which interacts with production domains by requesting actions such as update, configure or heal. The control loops are implemented by multiple support and network functions, they can act in a hierarchical way and delegating some level of control to other domains.

By leveraging analytics, AI and policy, control loops become adaptive and are central to assuring and optimizing the deployed services and resources.

Related articles/Additional reading:
[1] Cognitive processes for adaptive intent-based networking
[3] AI-enabled RAN automation
3.2 Artificial Intelligence

Artificial Intelligence (AI) will be used at all levels of the network and increasingly important as both networks and services are becoming data- and intent-driven, introducing a higher level of automation to mitigate an increasing network (configuration) complexity, while AI-mechanisms also support performance improvements.

The cognitive network will handle many of today’s management tasks. It will introduce an Intent function to achieve a higher level of autonomy, introduce a higher level of intelligence in the network and effectively relieve the network operator of direct network management tasks. This autonomy will change the way networks are operated by setting requirements and goals that reflect the operator’s business objectives and customer needs. Intent is the mechanism for communicating such goals and requirements.

Using AI increases the demand for following Ethics guidelines for AI by cross government and industry stakeholders. Uptake of AI might also be hindered by trust issues with new AI technologies and their increasing algorithmic complexity and autonomous evolution over time. It is essential that AI systems allow for human oversight, influence, and intervention. Through Trustworthy AI the use of AI shall among other things be safe, fair, non-discriminating, sustainable and respect the privacy of individuals or groups.

Finally, to be able to use AI at scale in products and services the process of developing and maintaining AI/ML for commercial products is formalized in a similar way to life-cycle management (LCM) for SW applications. This includes using applicable parts of MLOps from Data science experiments, through development and training to deployment.

Related articles/Additional reading:
3.3 Network slicing

Network Slicing is a model to realize specific and dedicated connectivity services over CSP networks already providing an extensive number of services/use cases supporting opportunities for various new services.

The network slicing techniques and capabilities enable e.g. private 5G networks to open up for new business opportunities. Operators should leverage their own business requirements and create the appropriate fit-for-purpose slices to cover each of the use cases. The diversity of new commercial and technical requirements has significant implications on how networks are built and managed. Figure 2 below depicts an example of slices being deployed across the entire network for different purposes as indicated by the colors.

Three main approaches to offering and delivering private 5G networks where slicing can also be leveraged for e.g. customization or resource isolation, are described below.

The **standalone approach**: on-premises private 5G network deployments, independent and isolated from the public 5G infrastructure. Network slicing can be used to customize the behavior for different use cases/traffic types.

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

The **hybrid approach** is provided by combining infrastructure adopted for public services with infrastructure at a customer’s premises using network slicing to customize and isolate slices per enterprise customer and use case/traffic type. A good example of a hybrid private 5G network is an airport where dedicated on-premises hardware is integrated with and reuses the public infrastructure to improve service and cost-efficiency.
Mobile broadband has become a society-critical service in recent years, with enterprises, governments and private citizens alike relying on its availability, reliability and resilience around the clock. Living up to continuously rising expectations while simultaneously evolving networks to meet the requirements of emerging use cases beyond MBB will require the ability to deliver increasingly higher levels of network robustness.

5GS has been designed to provide the robustness required to support the growth of conventional MBB services, while also offering network support to new business segments and use cases with more advanced requirements in terms of NRAR. 5GS delivers new capabilities that enable enterprises with business-critical use cases in segments such as manufacturing, ports and automotive to take a major step forward in their digitalization journeys by replacing older means of communication with the 5GS. These new capabilities are also beneficial for mission-critical networks like national security and public safety deployments being modernized.

It is important to consider all parts of the network in the definition of robustness (as illustrated by the green part in Figure 3), as the weakest link in the E2E chain sets the limits for the network service characteristics. In addition, network-level design must include consideration of both sunny day scenarios and different disaster/failure cases in all parts of the network. The large orange section represents both new critical use cases and society-critical use cases with new and tougher requirements. The orange line between the application client and the server, highlights the significance of the E2E perspective.

Figure 3 Shifting focus from node/NF-level to network robustness for demanding E2E applications
While both 4G and 5G are able to provide the high level of robustness required to deliver such services today, new and emerging use cases require the addition of new features and mechanisms in the network robustness toolbox. 5GS has been designed to meet even the most challenging network robustness requirements. Beyond that, the creation of robust networks also requires careful network planning and deployment.

The 5GS robustness toolbox consists of both standardized and vendor-specific network features and mechanisms. Highly flexible, it gives CSPs the power to activate the most appropriate mechanisms depending on the use cases and the deployment variants. The toolbox also enables CSPs to activate different mechanisms for different user equipment within a single network.

Related articles/Additional reading:
[10] Robustness evolution: Building robust critical networks with the 5G System

3.5 Security

All networks of today must be considered as untrusted, assuming a breach. This is the assumption of the zero trust (ZT) security model; to never make assumptions about trustworthiness. The ZT security model is based on the NIST Cyber Security Framework that specifies a model to approach security including, Identify, Protect, Detect, Response and Recover.

Telecommunications standards have already evolved the telecom security model by adopting Zero Trust principles and shifted from a perimeter-based security model in 3G/4G to a Zero Trust Architecture (ZTA) and Detect & Response in 5G.

ZTA includes principles for secure and trusted horizontal communication over a network, e.g for two endpoints to communicate securely over any network or for mutual authentication between network functions. All traffic is encrypted, integrity is protected, and all access must be authorized by the Network Repository Function (NRF) acting as an authorization server.

Dynamic Access Control(DAC) is a principle in ZTA, where complimentary information about the subject/resource is considered in the access control process, like log insights, device information, location, etc.

Detect and Response cover two main areas; local and remote detect response possibilities (loop size). While a single NF could detect deviating behaviour and act, information from more NFs and other data sources may be needed to detect suspicious activity.

Detect & response functionality has the intention to act in a highly automated way. A response could affect the integrity and availability of a system and even if some anomalies have been detected, the system could still be operational.

Related articles/Additional reading:
[12] 5G security -trustworthy 5G system
[13] Security
3.6 Service exposure

As CSPs seek to extend their offerings to address specific needs of enterprises, it is key to expand outside telecom to explore the exposure of network capabilities that are 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 lastly leverage their connectivity offerings and combine with cloud and edge offerings from different players.

Exposure can be applied in different locations, both in the network and in the device as illustrated in Figure 4 below.

![Figure 4 Exposure Interfaces](image)

- **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.
- **C Interface layer** contains a 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.
- **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 from Modem vendors.

Although the Z and C layers are expressed as thin lines in Figure 4 these can contain a set of functions that are common to all exposed services, e.g. discovery, access control, identity management, throttling, monitoring etc and are realized by the framework in picture 5. This is to further drive that the different consumers of the API (developers, integrators, enterprises etc) get a consistent experience and enabling scale and the Application does not have to "proxy" through the Management, Orchestration and Monetization layer.

The functional architecture for service exposure is built around four customer scenarios, see Figure 5, **Internal**, applications for control, monitoring, optimization, **B2C**, for consumers directly using services via web or app support largely with self-service management, **B2B** consists of partners that use services.
support their business. The B2B2X scenario is made up of more complex value chains.

![Figure 5 Service exposure functional framework](image)

OSS and BSS will take a central role in facilitating exposure to enable new business models for CSPs as well as handling commercial and operational management of exposed services.

Related articles/Additional reading:

[15] Service exposure and automated life-cycle management: the key enablers for 5G services
[16] Network Exposure
4 Network architecture domains

4.1 Introduction

The high-level network architecture, Figure 6, represents a transformational trend in how networks are built, operated and opened up for innovation. Instead of dedicated, well-defined, and vertically integrated nodes connected in a static network setup, the networks are evolving towards a more dynamically adaptable architecture where Network Functions (NF) and applications are running where and when they are needed to optimize performance, cost and business agility. This network transformation is enabled through the horizontalization of the network architecture where distributed cloud resources, joint data pipelines and Open APIs open up for the programmability and flexibility needed - both inside the network and to the outside world.

With 5G we have created a powerful innovation platform that can be used to meet the communication and connectivity needs of virtually any sector of industry or society. Driving openness in 5G and beyond to foster new application development and new business is critically important.

A central part of openness in 5G networks, and beyond, is the horizontal architecture. It is important to leverage the strength of business relevant vertical/functional interfaces and additional horizontal interfaces in the network platform for flexibility in a cloud world, all coordinated across relevant standardization bodies, open-source projects, alliances and partnerships.

Separation through horizontalization between HW, cloud, transport, data pipelines, network applications, management, and monetization will consequently make interfaces supporting horizontalization more important for multi-vendor interfaces.

In this transition, correct modularization of network functionality is crucial. The open air-interface, open RAN-packet core interface and global roaming specified in 3GPP is the basis, and remain so, for the global scale with a strong ecosystem.

![Figure 6 High-Level Network Architecture](image)
The horizontal architecture functional domains will be further described in the following chapters. Vertical topological domains span from “Devices / Local networks” on the left all the way to “Global sites” on the right.

Global connectivity and services have by tradition been deployed in a federated model, where the interfaces are well standardized and offered by any service provider. The complexity with multiple networks has been hidden through interoperability and inter service providers exchange models. However, the rapid deployment of new features makes the traditional standardized federated model hard to use. New methods of enabling exposure of assets from multiple networks is needed, like network asset facilitation and exchange, or even, on service providers request, aggregation into a single offering.

Figure 7 Network architecture business interfaces

The architecture supports various types of interfaces as illustrated in Figure 7.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Between CSP and its access customers, Access interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Between CSPs, Interconnect /roaming interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Between CSP and application service customers, Service interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Between CSP and cloud platform service providers, Cloud platform interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Between CSP and transport service providers, Transport interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>Between CSP and suppliers, Data sharing, Managed service, CI/CD interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Between CSP suppliers, Network internal interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Between CSP suppliers, Multi domain management interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>Within Developer environment encapsulating/wrapping C layer as needed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Related articles/Additional reading:
[17] Openness in mobile networks
### 4.2 Management, Orchestration, Monetization

Management of the cloud environment and new services become even more important as all services need to be managed in real-time maintaining the required service performance. This dynamic and competitive environment requires management of the networks and their supporting systems to be less expensive to manage and maintain while expanding their role in the service definition. The network needs to be self-provisioned to drive down costs in an instant-access, cloud, and application-driven world. The management system needs to be scalable, with an agile IT operational model that bridges the physical and virtual environment and provide services deployable at speed with “zero touch” fulfillment and assurance.

A future reference management architecture will be based on founding principles to support a real time management environment of the ecosystem, such as an open micro service-based architecture that enables data centric applications – where the application logic should have control of how, what, and when the data can be exposed. It will also support the separation of functional management and realization management with the realization management and resource automation tends towards a gitops style paradigm.

Figure 8 below describes the major functional components, that will be used to realize automation. This view captures a design-time set of functions (universal design studio) and a run-time set of functions (all others) that will interwork to realize the automated life-cycle management of network functions and services.

Service Orchestration functionalities are at the foundation of the architecture to enable the service- and intent-based automation. Machine learning capabilities are present to enable the intent-based automation capabilities.

![Figure 8 Service and Network Automation Platform](image)

The automation platform will support a broad range of OSS (Operational Support System) use cases, from very specific automation / optimization applications to feature-rich and complex applications.
One of the primary roles of BSS (Business Support System) is to manage a CSP’s relationships with its stakeholders by keeping track of Service Level Agreements (SLA), handling orders, generating reports, sending invoices, etc. In the past, these stakeholders were generally limited to consumers, resellers, partners, and suppliers. In the 5G/IoT business context, though, more complex ecosystems are arising that BSS must evolve to support.

Examples of new stakeholder groups that need to be considered in the 5G/IoT business context include:

- Enterprises and industry verticals that require solutions beyond telecoms
- New types of suppliers such as IoT device providers and suppliers of eSIM (embedded SIM) and related technologies
- Platform providers that specialize in specific IoT or edge clusters or groups of IoT use cases.
- Integrators that specialize in specific verticals that combine capabilities from multiple stakeholders to address consumer needs.

From a high-level architectural viewpoint, BSS in the 5G/IoT ecosystem closely resemble the traditional BSS depicted in Figure 9, with similar interfaces to surrounding systems. The BSS architecture in is presented in the Open Digital Architecture format.

![Figure 9 5G reference architecture for BSS](image)

Related articles/Additional reading:

[18] Open, intelligent and model-driven: evolving OSS
[19] 5G BSS: Evolving BSS to fit the 5G economy
[20] Network Automation
[21] Telecom BSS
4.3 Access, Mobility & Network Applications

Radio Access
The external interfaces and the definition of the functional behavior of the Radio Access Network are standardized by 3GPP and adopted in the specifications of O-RAN. The high-level specifications, leave room for innovation to enhance the network with RAN-internal value-add features - a flexibility that over many years have resulted in continuous improvement in many areas.

Deployment options stretch from distributed to more centralized and the most appropriate option largely depends on the deployment area (urban, suburban, or rural) and the availability of dark fiber. In all options, outdoor site deployments can be either macro sites (rooftop/mast mounted) or street sites (pole, wall).

To determine the optimal architectural split, the RAN architecture needs to be examined with a finer level of granularity than that offered by 3GPP. The target functional composition should be analyzed and defined by, e.g. function and interface characteristics, preferred execution environment, spectrum efficiency.

Deployment options stretch from distributed to more centralized and the most appropriate option largely depends on the deployment area (urban, suburban, or rural) and the availability of dark fiber. In all options, outdoor site deployments can be either macro sites (rooftop/mast mounted) or street sites (pole, wall).

Today's RAN architecture consists of a number of functions, where these can be mapped to one or several 3GPP network functions. The cloud native RAN architecture consists of several micro services connected in a communication framework and distributed on different workloads and locations, representing one or several 3GPP/ORAN network functions. Standard interfaces are used for external communication.

The flexibility of locating RAN functionality in different locations in NR RAN architecture and the ability to support more radio sites increase the need for network automation, simplifying installation, deployment, and operation.

The RAN automation strategy and target architecture are based on an automation platform (SMO) that manages a RAN with physical and cloud-based network functions. For the Non-RT RIC, the RAN automation applications are called rApps.

Addressing Cloud RAN automation needs, O-RAN has introduced another set of requirements with significant impacts on the Management and Orchestration strategy through the introduction of new interfaces (A1, O1, O2, R1), new defined functions and functional layering (Real-Time RIC, Non-Real-Time RIC, and SMO), and the introduction of an open and multi-vendor set of requirements.

Related articles/Additional reading:
[22] Artificial intelligence in RAN – a software framework for AI-driven RAN automation
[23] 5G New Radio RAN and transport choices that minimize TCO
[24] Enhancing RAN performance with AI
[25] 5G RAN
Packet core
The EPC network architecture is in a steady state. No major network architecture changes are expected from standardization of 5GC network functions.

Although initial 5GC deployments happened in 2020/21, packet core solutions with both EPC and 5GC capabilities will be required for a long time due to the co-existence of LTE and NR.

The expectations on 5GS capabilities for both eMBB, new business verticals and wide-area private networks are very high from CSPs. Intensified society critical use of eMBB increase the demand for reliability and availability in 5GC.

Other 5G use cases require ultra-low latency and very efficient packet processing adding demands to the user plane compared to previous generations including Time synchronization service (5G TSN), API exposure capabilities and improved resilience and redundancy (NRAR) with, e.g., dual user planes, support for non-public networks (NPN).

Initial 5GC solutions have focused on large scale eMBB deployments but will need to also target smaller networks and small scale standalone deployments for, e.g., Enterprise and Industry. 5GC will support voice migration to VoNR (Voice over New Radio) on top of the already deployed VoWiFi (Voice over WiFi) solutions.

The core will exist in an environment that is cloud-based, applying cloud native design principles for scalability, dynamic orchestration of network resources, and a modular and highly resilient base architecture.

5G Core is built on a new architecture paradigm, Service Based architecture, with new reference points and services used between network functions in the control plane. While 5G Core has advantages over EPC, it is worth highlighting that it is a new architecture and network paradigm that needs to be deployed in the CSPs network.

The functional view of the dual core in Figure 10 reflects the network functions in EPC and 5GC that are foreseen in the solution for 4G-5G. Interworking between 5GC and EPC is needed for a successful launch of 5GC and while full coverage of NR SA is achieved.
The migration between EPC to 5GC will require coexistence of functions for quite some time including Subscription, Policy and Exposure Management functions according to 3GPP.

For use cases beyond eMBB, the traditional model where a CSP owns and manages individual subscriptions must be changed to allow for an NPN owner (e.g., an industrial/critical IoT for factory automation use cases) to manage its users on an individual basis.

New use cases for Adaptive policy control based on analytics and insights will introduce the use of AI/LM algorithms.

The SCEF/NEF provides a unified Exposure interface covering 4G and 5G, and hiding the complexity of the underlying network topology from e.g. application developers by providing a common API framework (CAPIF) allowing for consumption of the service APIs in an easier way.

Related articles/Additional reading:
[26] 5G migration strategy from EPS to 5G system
[27] 5G Core

Communication Services
Communication was the original service for which telecom networks were developed and the interoperable communication services of today still must be backwards compatible to a legacy of older generations of the service still in use.

The services and their IMS based architecture are reused for all 3GPP accesses including 5G and is expected to be at least as good as VoLTE. The migration from EPC and 5G Core is supported seamlessly.

The IMS architecture is built to support new innovative services and communication means like XR with emerging business models, evolved ecosystem and more complex value chains arising needing more flexibility and automation with new delivery models, etc. As an example, XR, holographic conferencing will be supported by Interactive Calling.
The introduction of Cloud will fundamentally change how IMS applications are deployed and managed but does not mean that the traffic view of IMS network architecture will change. The cloud mechanisms and functions are complementary to those of the IMS Network architecture.

The IMS architecture has evolved in 3GPP to support service-based interfaces to the 5GC while aligning to a Service Based Architecture (Figure 11 below).

![Figure 11 IMS Service based interface introduction](image)

Communication Services are demanded by Enterprise segments. There are different Private network deployment models where communication services may play a role.

- PNI-NPN where communication services could be provided with a network slicing deployment
- SNPN with independent IMS provider, IMS is deployed and provided by a third-party provider to Enterprise, i.e., an operator

Mission Critical Push-To-Talk (MC-PTT) is delivering communication services to the focus area of public safety. Even if network entities and reference points are new, there are many similarities to the IMS reference architecture as illustrated in Figure 12.
Cloud Infrastructure

Cloud native technologies have had an enormous impact on how as a service product are developed, deployed, and operated. This in combination with virtualization technologies has been seen as a key technology to use when building future telco networks.

Cloud Native Technologies offer some in telco system realization but also many challenges which may not be visible initially.

Examples of such benefits is the ability to handle separation of infrastructure and SW as well as increased scalability and flexibility.

Cloud native technologies are developed and evolved from a service centric product delivery paradigm and thus it’s necessary to ensure a DevOps approach to software development, delivery, and deployment.

In cloud native, integration towards the (Hyperscaler Cloud Provider) HCPs is a way of using services already provided by the HCPs. Though this can benefit a solution it may create lock in effects incurring future costs, e.g. due to the need to support very different variants of services, etc. Making your e2e orchestration solution interact with the HCP orchestration solutions and pipelines will be required.

Telco system are inherently distributed systems, cloud infrastructure mostly not. The orchestration functional distribution can be used but for geographical distribution, the concept of a function distributed on several container
environments can be used and still treated as one functional entity from the OSS layer.

A benefit with cloud native implementation is the possibility to leverage functionality provided by the underlaying platform. Several deployment variants of the Cloud native Network Function is when realizing the Cloud native Network Function by using 3P provided PaaS is described in Figure 13.

Deploying on a 3P PaaS has an impact on the management view of the system and consideration to how application are built and how responsibility is divided between the PaaS and application provider.

![Figure 13 PaaS deployment variants](image)

**Figure 13 PaaS deployment variants**

Related articles/Additional reading:
- [32] Creating the next-generation edge-cloud ecosystem
- [33] Edge computing and deployment strategies for communication service providers
- [34] Cloud Infrastructure
- [35] Cloud Native
- [36] Edge Computing

4.5 Data pipeline

The purpose of data pipelines is to facilitate access to high-quality data for applications ranging from network management and customer experience management (CEM) to business analytics, product serviceability, artificial intelligence (AI)/machine learning (ML) model training and a lot more. As many of these applications use the same data sets, a harmonized data ingestion architecture boosts efficiency and makes it possible to focus application development resources on use case realizations rather than on data management.

A harmonized data ingestion architecture, Figure 14, is built on the idea that data should be collected once and then shared with any application that needs it. The proposed architecture can be deployed both in customer networks and as
service in application clusters (ACs) external to customer networks.

The key functional entities in the data ingestion architecture are data sources, data collectors, the extensible data collection architecture (EDCA), the data relay gateway (DRG), the data distribution central (DDC) functionality, the global data catalog (GDC) and the globally federated data mesh.

**Figure 14 Data ingestion architecture**

Related articles/Additional reading:

[37] Data ingestion architecture for telecom applications

### 4.6 Transport

Transport network evolution is mainly driven by external factors like the introduction of 5G with the need for more bandwidth, tighter latency requirements and new E2E services driven by operators and enterprises including for example network slicing.

Evolution of network functions into CNFs/VNFs introduces a new deployment architecture implying that the traditional WAN Transport needs to interact with the Data Center networking with CloudRAN being one obvious example.

Other implications are that traditional transport technologies and architectures used in core transport networks will be pushed further out in the RAN transport network especially in dense metro areas. Additionally UnderLay/OverLay technologies allows for mixing of fronthaul/backhaul traffic in RAN transport further out in the network.

As for many other network areas, automation is a strong trend in transport including intent based configuration. The basis for automation can be found in efficiency but with the introduction of the flexible 5G NR architecture, slicing, auto-provisioning of security, etc automation is a must.

It is fundamental to have a cost-efficient deployment to handle smaller sites. Traditional DC sites normally have separate infrastructure for DC fabric and DC GW. On the smaller sites fronthaul switching will also be required. A unified
solution for fronthaul, DC Fabric, and DC GW is a fundamental part of this architecture. A direct consequence of this unified solution is a single SDN controller for the whole domain.

With 5G proliferation, massive amounts of millimeter wave (mmWave) spectrum will be available. This opens up for the introduction of Integrated Access Backhaul (IAB) further discussed in [40] below.

Related articles/Additional reading:
[38] Enabling intelligent transport in 5G networks
[39] 5G New Radio RAN and transport choices that minimize TCO
[40] Integrated access and backhaul – a new type of wireless backhaul in 5G
[41] 5G synchronization requirements and solutions
[42] 5G transport

5 Network architecture examples

5.1 Wide area public network/Evolved MBB

The evolved MBB (eMBB) network architecture use case is a wide-area public network that delivers different data and communication services for consumers, business users, mission-critical and IoT use cases. Today networks are expected to be available 24/7 being a society-critical resource, however yet not surrounded by much regulations.

The typical and dominant end-user device using eMBB is a smartphone, but other device types and gadgets are be laptops, gaming devices and AR/VR devices taking advantage of the low latency and high data rates offered.

The introduction of the eMBB network architecture introduces capabilities far beyond those of previous cellular generations; support for massive data rates, very low and bounded latency, ultra-high reliability and availability, extended QoS framework, distributed Cloud, etc.

The target architecture consists of 5G NR standalone and 5G Core, but will, however, coexist and interwork with 4G for many years to come as described also in earlier chapters.

The functional network architecture for eMBB includes the RAN, Packet Core, UDM & Policy and Communication Services functional domains. The target architecture for eMBB will be based on a combined EPS/5GS network including LTE and NR RAN and the EPC-5GC tight interworking architecture as shown below. There will be only limited 2G/3G dependency for, e.g., for NNI and outbound roaming. Figure 15 depicts this functional target architecture for eMBB with a focus on the traffic network functions in a CSP network.
Below Figure 16 shows an example of the eMBB deployment architecture using a distribution of functionality over RAN access, regional and national DC sites. All core UP related functionality deployed in the regional DC site and the core CP related functionality is in both sites. E2E OSS and BSS related functionality is in the National DC site. The RAN and Core & IMS management and orchestration are placed in Regional DC to meet regional survivability requirements.

New value-add services, like low-latency consumer applications (e.g. for gaming/AR/VR) supported by the CSP may require special attention. These low-latency applications are placed in regional data centers but may also be placed further out depending on the need of the service, e.g., to local access sites.

Related articles/Additional reading:
[43] 5G network
[44] 5G deployment considerations
5.2 Wide area Private Networks

The need and ability for governments to handle various kinds of crises and the digital transformation for Public Safety agencies and Critical Infrastructure require technology that supports and facilitates cooperation between various units involved.

Governments who plan to deploy their own dedicated networks for mission-critical communication should adopt standardized technology and ecosystem, like cellular 3GPP technology. They can also choose a symbiotic network approach, a configuration allowing for a government mission critical network to interact with commercial networks benefitting both governmental and commercial players (see also [45])

Mission-critical grade performance has to be ensured in several areas like: Network availability, Multi Network Operations, Coverage and capacity, Security and hardening, etc

As communication becomes increasingly mobile, the growing desire of governments to take advantage of advancements in technology, such as augmented reality and drones, is increasing the need for more bandwidth and lower delays in communication.

Figure 17 illustrates a few examples of Mission critical use cases all depending on a wide area network scenarios and requiring different capabilities and characteristics from the network.

Figure 17 Examples of Critical communication areas

The journey of legacy Land Mobile Radio (LMR) to 3GPP networks allow for a wider range of deployment alternatives as described in Figure 18 and further explored in [47].
Figure 18 Examples, Mission Critical 4G Networks, leveraging existing coverage

**Dedicated Network.** Critical communications network run on separate infrastructure, but both networks can share physical facilities.

**Shared RAN – dedicated spectrum usage.** RAN sharing with commercial network, complemented by dedicated apps, Core and RAN infrastructure using dedicated spectrum.

**Shared RAN – dynamic spectrum usage.** Critical communication users and consumers benefit from shared access to the commercial network’s RAN as well as dedicated RAN, though critical communication users have priority.

**Secure MVNO.** Model relies on shared use of the commercial RAN, together with roaming capabilities in the core, but complemented with a dedicated core (either full or upper part only) and applications to ensure partitioning of sensitive user and network data.

The evolution from a mission critical 4G network to a mission critical 5G network is summarized in

Figure 19.
Some of the 3GPP mission critical enablers are already standardized, some will come later, but non-stand-alone 5G network can already be used. The mission critical services are then terminated on the LTE-leg while the NR-leg is used for data off-load.

In a couple of years with improved NR coverage and mission critical devices supporting NR, the network can be leveraged for, e.g., real-time drone control, or multiple real-time body cam streaming leveraging the high throughput/low latency of NR. MCPTT will still be supported on LTE should certain critical networking capabilities, e.g. broadcast, not be available on 5G.

Approaching 2024-2025, MCN will be able to leverage the full 5G stand-alone network (with services like MCPTT being available sooner), running all MC-services on mission critical 5G devices. Networks could still cater for mission critical LTE-only devices and coverage in areas with only LTE, but it will be possible to fully move to 5G completely.

Figure 19 The Journey to mission critical 5G

Related articles/Additional reading:
[45] Whitepaper on critical communication secured networks
[46] Mission critical communication
[47] Ericsson Mission Critical Networks
[48] Infographic - The Policeman at work
[49] Explore future-Mission-Critical-Services
5.3 Private and Local Dedicated Networks

Private mobile networks are targeting enterprises with private 4G and 5G technologies, as well as, with the objectives to make mobile technology easy to deploy and easy to use for enterprise end users.

The below Figure 20 shows some examples of industry verticals where private 4G and 5G systems are being deployed.

Figure 20 Sample of industry verticals using 4G and 5G mobility

Different vertical industries have their own requirements to serve specific use cases in a specific area. The area may cover a large seaport and include several macro radios, for example, or just cover the indoor space in a single warehouse in the other extreme. On the other hand most industries require high availability, local survivability, high security, and privacy.

Mobile private networks will largely outnumber the typical mobile operator networks in most countries in the future. For example, spectrum allocation models may need to be revised to handle this growth through, e.g., industry spectrum.

Characteristics of these private networks will differ from operator mobile networks being smaller and having stricter dependencies like low latency. To be successful a highly flexible and cost-effective architecture is required to minimize needs for system integration. For example, the networks may not support all typical public network services.

Scalable solutions to deliver small footprint products for both Core and RAN, and as well as a core network based on cloud native principles are needed to satisfy needs for dynamic deployment on different sites with a focus on end customer requirements. NRAR capabilities are a must.

Automation will be required as will enabling operations capabilities for both enterprise and CSP.

3GPP specifies two Non-Public Network (NPN) deployment models, Public Network Integrated NPN (PNI-NPN) and Standalone NPN (SNPN).

Private networks require integration into the different physical domains of the enterprises, enterprise OT production domains, and enterprise IT domains. An enterprise may have several, geographically distributed domains, where some are served by the same NPN and others are not.

Enterprises that favor SNPN deployment model have common requirements, e.g., that services have local survivability when connectivity towards any cloud services fails. A new demand seen on the market is the request for small and low-cost remote sites that are attached to the central site of a dedicated network.
Since a NPN deployed at the enterprise premises and operated by a CSP comprises of all network nodes that make up a complete 5G network, special attention is needed for e.g., Life cycle management, HW footprint, SW licensing cost.

![5G NPN deployment architecture with small remote sites (OT1 & OT3)](image)

NPNs are in their complexity comparable to MBB PLMNs but typically host 1k – 20k subscribers meaning that efforts to operate and maintain such NPNs must be scaled down accordingly.

Since the NPN often is regarded as a production asset, the enterprise demands operational control and full transparency. For enterprises with limited 3GPP expertise, simple and operational IT-like APIs that expose 5G NW capabilities (e.g QoS, localization, etc) to facilitate integration with OT systems are essential.

Related articles/Additional reading:
- [50] Critical IoT connectivity: Ideal for time-critical communications
- [51] Boosting smart manufacturing with 5G wireless connectivity
- [52] Optimizing UICC modules for IoT applications
- [53] 5G-TSN integration meets networking requirements for industrial automation
- [54] Industry 4.0
- [55] 5G spectrum for local industrial networks
- [56] Critical capabilities for private 5G networks
- [57] Cellular IoT in the 5G era
CTO network trends

The COVID-19 pandemic has created a high level of awareness of the important role of the digital infrastructure to deliver society critical, economic, and governmental functions.

While this acceleration of digitalization during the pandemic was made possible by existing capabilities, 5G will be the main digital infrastructure for consumers with e.g., XR developing as well as for enterprises increasing productivity through increased automation, networks slicing, etc.

Below 5 Technology trends discuss the evolution of networking according to the Ericsson CTO.

#1: Digital representation for the networked reality
Describing the future use of low-level processing in the network to provide e.g. multisensory digital experiences (Internet of Senses) and accurately representing the surroundings are using network-generated maps (Spatial Mapping), etc.

#2: Adaptable limitless connectivity
One primary goal of 6G access is to deliver adaptable limitless connectivity. Multi-vendor interfaces will ensure openness both in networks and in ecosystem at large while minimizing system complexity ensuring end-to-end performance, resilience mechanisms, etc to different applications.

#3: Integrity of trustworthy systems
Providing security services will be essential in order to manage and verify compliance to security, safety, resiliency and privacy demands. One such example is the identity-centric approach with a Zero Trust architecture.

#4: Federated cognitive networks
The network automation path started in 5G, will continue making 6G networks cognitive and acting autonomously to optimize its performance based on intent and supported by trustworthy AI

#5: A unified network compute fabric
The convergence of internet, telecom, media and information technologies will lead to the creation of a unified, global system of interconnected components requiring a network compute fabric to facilitate unification across ecosystems, application management, execution environments, and exposure of network and compute capabilities. 6G will evolve from today’s roaming models based on smart contracts for the federated ecosystem.