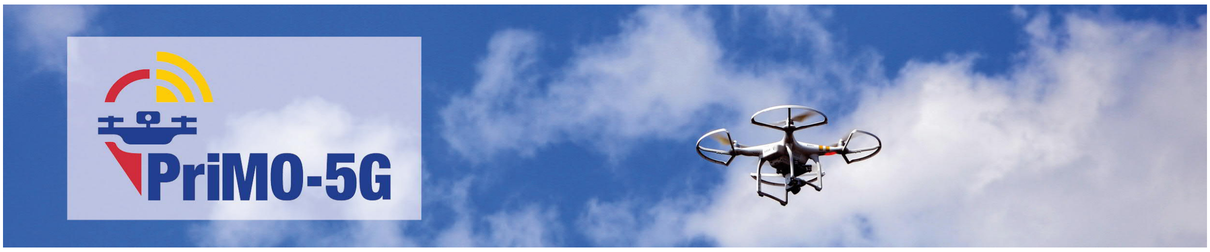


Project Title	Virtual Presence in Moving Objects through 5G
Project Acronym	PriMO-5G
Grant Agreement No	815191
Instrument	Research and Innovation Action
Topic	The PriMO-5G project addresses the area of “a) Focus on mmWave and super broadband services” in the call “EUK-02-2018: 5G” of the Horizon 2020 Work Program 2018-2020.
Start Date of Project	01.07.2018
Duration of Project	36 Months
Project Website	https://primo-5g.eu/

D1.2
END-TO-END PRIMO-5G NETWORK ARCHITECTURE

Work Package	WP1, Scenarios, architecture, economic and regulatory analysis
Lead Author (Org)	Ki Won Sung (KTH)
Contributing Author(s) (Org)	Riku Jäntti (AALTO), Nicolas Malm (AALTO), Akgul Ozgur (AALTO), Edward Mutafungwa (AALTO), Byungjin Cho (AALTO), Jose Costa-Requena (CMC), Saimanoj Katta (CMC), Abdulkadir Mohammedadem (CMC), Abraham Afriyie (CMC), András Zahemszky (EAB), Anders Nordlöw (EAB), Zere Ghebretensaé (EAB), Enric Pardo Grino (KCL), Ghizlane Mountaser (KCL), Konstantinos Antonakoglou (KCL), Toktam Mahmoodi (KCL), Ki Won Sung (KTH), HyungJoon Jeon (EUC), Roben Delos Reyes (KAIST), Soohyun Kim (KU), Seunghwan Kim (YU), Yeosun Kyung (YU), Sejin Seo (YU), Doseon Kim (YU), Seong-Lyun Kim (YU), Dong Ku Kim (YU), Sang-Hyun Park (YU), Tae Hun Jung (YU), Chan-Byoung Chae (YU)
Due Date	30.06.2020, M24
Date	26.06.2020
Version	1.0

The work described in this document has been conducted within the project PriMO-5G. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 815191. The project is also supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00170, Virtual Presence in Moving Objects through 5G). The dissemination of results herein reflects only the author’s view, and the European Commission, IITP and MSIT are not responsible for any use that may be made of the information it contains.

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Disclaimer

PriMO-5G has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 815191. The project is also supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00170, Virtual Presence in Moving Objects through 5G). The dissemination of results herein reflects only the author’s view, and the European Commission, IITP and MSIT are not responsible for any use that may be made of the information it contains.

Table of Contents

List of Tables

List of Figures

Executive Summary

In this deliverable, we present an overall system architecture which fulfils the requirements of the PriMO-5G project. The architecture consists of user equipment (UE), radio access network (RAN), core network (CN), multi-access edge computing (MEC), network management and orchestration, and public safety applications. The focus of this deliverable is to highlight the component technologies investigated in the project rather than to provide a comprehensive architecture.

The main aim of the PriMO-5G project is to demonstrate an end-to-end (E2E) 5G system providing immersive video services for moving objects. The PriMO-5G project chose the public safety, particularly firefighting, as the main use case because it is an area where immersive video services with moving objects can make a substantial improvement in the safety and efficiency of the operations. The usage of drones for firefighting can be divided into three categories. The first category is preparatory actions for the fast collection of information. Second, the drones can provide visual information to the first responder and the incident commander. Third, the drones can gather sensory and measurement data of the fire scene and surroundings.

Public safety communications are moving from a voice-centric paradigm to a data-centric paradigm enabled by LTE, 5G, and the ongoing the 3GPP evolution. In the PriMO-5G project, public safety communication requirements have further been refined by looking into advanced video services as well as aerial components, integrated AI, and edge computing. The PriMO-5G project proposes a multi-stacked architecture to cater for the next generation of data-centric public safety communication architecture. At the bottom of the architecture we have all the computing, storage, and networking resources. These sustain the whole architecture and enable the virtualization and integration of cloud-based infrastructure supporting the network functions and applications. The next layer has infrastructure components that provide the physical transmission of data. The components in this layer go through drastic evolutions. UEs are evolving dramatically with aerials, IoT devices, and with more data-centric capabilities. RAN has a variety of architectural approaches, and split options are introduced. Transport is evolved with new transport concepts and with integration of E2E capabilities, e.g. network slicing. Core networks are moving to a service-based architecture. Advanced solutions to enable edge computing, including breaking out traffic to local edge servers are introduced. To handle the physical infrastructure an orchestration layer is needed. Orchestration is happening on several layers for E2E orchestration down to domain specific orchestration. AI can be integrated to enhance placement decisions. With the 5G architecture supporting a variety of how to expose data, analytics functions can be evolved to handle operational analytics needs as well as external exposure of analytics. Service and network management is moving into being more integrated into real time communication support and is further enabled by integrated AI and other management services.

The PriMO-5G architecture consists of a set of layers to provide E2E public safety communications. The different layers are touched upon on the subsequent sections of the document. UE, as described in Section 3, deals with important types of user equipment that can be connected to a public safety communication network. These include drone for 5G live streaming, head-mounted devices for end users, and fire truck with mobile computing server. Section 4 covers RAN elements that can be operated and supportive to a wireless public safety communications system, e.g. RAN split options, integrated access and backhaul, cell-free architecture, drone base stations, lens-based mmWave communications, and moving terrestrial base stations. CN architecture in Section 5 covers different CN functions needed in a public safety system. In this section, highlighted areas are core network slicing, 5G QoS framework, enhancements for URLLC, non-public networks, isolated operation for public safety (IOPS), optimal routing, network data analytics function (NWDAF), and 5G-LAN. Edge architecture, as described in Section 6, covers functions to provide edge computing. In this section two architecture are highlighted: ETSI MEC architecture and the 3GPP edge architecture. Section 7 describes AI-enabled network management and orchestration, which covers the management data analytics function and network slice orchestration. Services and applications for public safety, as described in Section 8, covers currently defined mission critical services in 3GPP as well as opportunities for immersive applications using extended reality.

The PriMO-5G architecture can be highlighted by the three aspects. The first aspect is cellular networks for public safety with aerials. In the PriMO-5G context, public safety UE may not only be handheld devices but also aerials, e.g. drones. The cellular network may also be extended with new type of base stations or network elements providing network functionality, e.g. drones or fire trucks serving as base stations, or fire trucks hosting compute facilities. Second, edge computing provides computing and storage resources with adequate connectivity close to the devices generating traffic. It is about bringing the services closer to the location where they are to be delivered. The motivation is to reduce latency and reduce transmission costs. Example use cases include AR/VR, real-time facial recognition, video surveillance, etc. Third, end-to-end network slicing plays a crucial role in the envisaged public safety operations. Multiple network slices can be built on the common infrastructure, while each of them is realizing a wanted network characteristic supporting a customer need. Customer, in this definition, may be an enterprise, another service provider, or even the network operator itself. In the PriMO-5G context, the customer may be a public safety organization (e.g. police, firefighting organization etc.), too.

Since the PriMO-5G project aims to demonstrate an E2E 5G system providing immersive video services for moving objects, an important ingredient of the project is demonstrations and testbeds that integrate radio access and core networks developed by different project partners to showcase E2E operations of envisaged use cases. Section 9 of this deliverable describes the links between the PriMO-5G architecture and the demonstration activities of the project in the individual partner, local, intra-continental, and cross-continental levels.

This deliverable has identified a set of potential innovation and enhancements to make the overall system architecture fulfil the requirements for reliable and innovative public safety communications. The areas for the innovation and standardization include the integration of network slicing into mobile networks, enhancing artificial intelligence support for 3GPP, standardized functions for mobile integrated access and backhaul (IAB), optimal routing to allow low latency communication, and separate treatment of uplink and downlink for UAVs.

List of Acronyms

Acronym	Definition
3GPP	Third Generation Partnership Project
5G	Fifth-Generation Mobile Network
5GC	5G Core Network
5QI	5G QoS Indicator
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AN	Access Network
AR	Augmented Reality
ARP	Allocation and Retention Priority
BBU	Baseband Unit
BE	Best Effort
BLER	Block Error Rate
BS	Base Station
CN	Core Network
CoMP	Coordinated MultiPoint
CPRI	Common Protocol Radio Interface
DRB	Data Radio Bearer
DRX	Discontinuous Reception
DU	Distributed Unit
E2E	End to End
eMBB	Enhanced Mobile Broadband
eNB	Evolved Node B
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
FCAPS	Fault, Configuration, Accounting, Performance, Security
gNB	Next Generation Node B
GBR	Guaranteed Bit Rate
GCS	Ground Control Station

Acronym	Definition
GPS	Global Positioning System
HMD	Head-mounted Device, Head-mounted Display
IAB	Integrated Access and Backhaul
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IITP	Institute for Information & communications Technology Promotion
IOPS	Isolated Operation for Public Safety
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
ISO	International Organization for Standardization
ITU	International Telecommunication Union
KPI	Key Performance Indicator
LOS	Line of Sight
M2M	Machine-to-Machine
MEC	Multi-access Edge Computing
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
mMTC	Massive Machine-type Communications
MNO	Mobile Network Operator
NAS	Non-access Stratum
NB	Node B (base station)
NG-RAN	Next Generation Radio Access Network
NG-U	User plane interface between NG-RAN and 5GC
NPN	Non-Public Network
NR	New Radio
NWDAF	Network Data Analytics Function
PDU	Protocol Data Unit
PHB	Per-Hop Behaviour
PLMN	Public Land Mobile Network

Acronym	Definition
PMR	Professional Mobile Radio
PPDR	Public Protection and Disaster Relief
PSAP	Public Safety Answering Points
PSC	Public Safety Communications
QFI	QoS Flow ID
QoS	Quality of Service
RAN	Radio Access Network
RRC	Radio Resource Control
RRU	Remote Radio Unite
RU	Radio Unit
SBA	Service Based Architecture
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Networking
SDU	Service Data Unit
SLA	Service Level Agreement
SMF	Session Management Function
S-NSSAI	Single Network Slice Selection Assistance Information
TCP	Transmission Control Protocol
ToS	Type of Service
TRxP	Transmission Reception Point
UAV	Unmanned Aerial Vehicle
UE	User Equipment
UPF	User Plain Function
URLLC	Ultra-Reliable Low Latency Communications
VLAN	Virtual Local Area Network
V2X	Vehicle-to-Everything Communications
VR	Virtual Reality
XR	Extended Reality
WP	Work Package

1 Introduction

1.1 Purpose and Scope

The main aim of the PriMO-5G project is to demonstrate an end-to-end 5G system providing immersive video services for moving objects. In this deliverable, we present an overall system architecture which fulfils the requirements of the project. The architecture consists of UE, RAN, core network, MEC, network management and orchestration, and public safety applications. The focus of this deliverable is to highlight the component technologies investigated in the project rather than to provide a comprehensive architecture. We also discuss how the architecture is related to the demonstrations that will be conducted in the WP5 of the project. Finally, we present potential enhancement of the 3GPP system architecture.

1.2 PriMO-5G scenarios and use cases

The PriMO-5G project chose the public safety, particularly firefighting, as the main use case because it is an area where immersive video services with moving objects can make a substantial improvement in the safety and efficiency of the operations.

Fires are a growing challenge for modern society. The dynamic nature of fires makes firefighting operations complex, high risk, and demanding in terms of the required firefighting resources and technologies. The use of public safety communications systems and aerial support systems is now critical for enhancing the safety and efficiency of firefighting operations. Communications technologies provide significant enhancements in situational awareness for the emergency first responders and their effectiveness in managing hazards by enabling immersive services and reducing constraints on operational data sharing. As a concrete example, we envisage that the use of UAVs, particularly drones, will make the firefighting much safer and more efficient if it is combined with 5G communications.

The usage of drones for firefighting can be divided into three categories. The first category is preparatory actions. Drones can be dispatched to the fire scene faster than fire trucks to gather the overall situational information. Furthermore, the drones can interact with the people around the scene to evacuate them to the safe area. Second, the drones can provide visual information to the first responder and the incident commander. The use of immersive video services, i.e. VR and AR, is a key to this enhancement. Third, the drones can gather sensory and measurement information of the fire scene and surroundings and report it to the firefighters and the incident commander. This will help the firefighters locate the people to rescue, identify toxic substances, and detect the possible explosion or collapse.

We divide the scenarios of the drone-assisted firefighting into the urban and rural cases as illustrated in Figure 1.2-2 and Figure 1.2-1, respectively. The detailed description of the PriMO-5G use cases can be found in [PRI19-D11] and [SMJ+19].

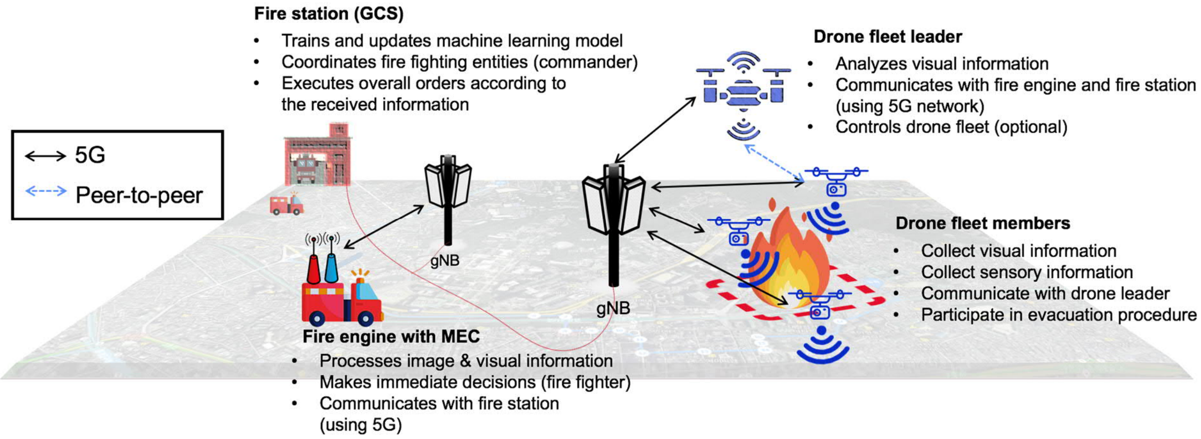

Fire station (GCS)
• Trains and updates machine learning model
• Coordinates fire fighting entities (commander)
• Executes overall orders according to the received information
Drone fleet leader
• Analyzes visual information
• Communicates with fire engine and fire station (using 5G network)
• Controls drone fleet (optional)
5G
Peer-to-peer
gNB
gNB
Drone fleet members
• Collect visual information
• Collect sensory information
• Communicate with drone leader
• Participate in evacuation procedure
Fire engine with MEC
• Processes image & visual information
• Makes immediate decisions (fire fighter)
• Communicates with fire station (using 5G)

Figure 1.2-1: An illustration of the urban firefighting. [PRI19-D11]

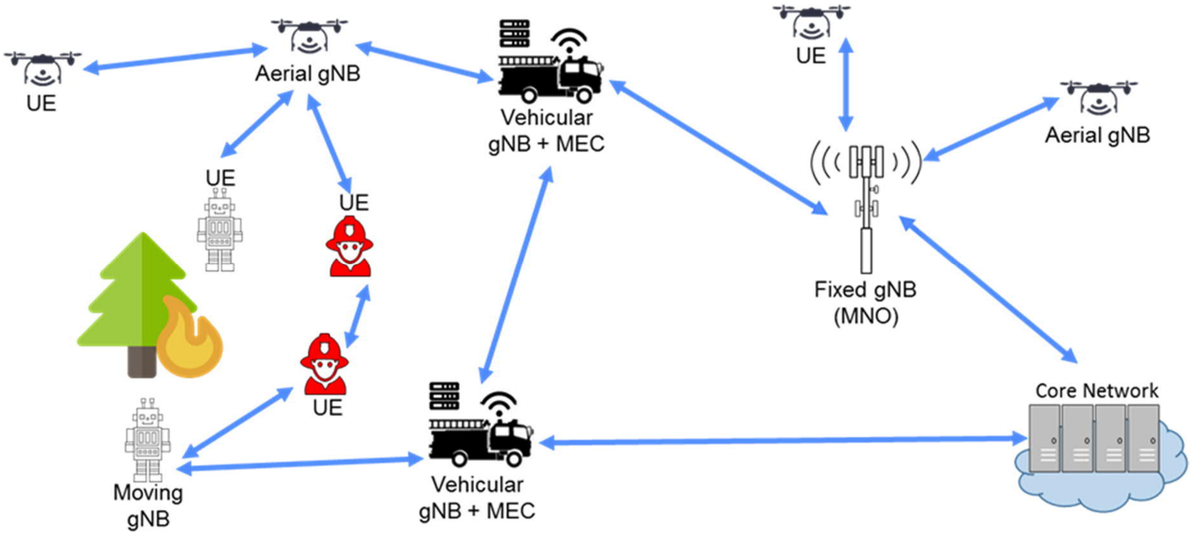


Figure 1.2-2: An illustration of the rural firefighting. [PRI19-D11]

The commonalities and the distinct characteristics of the two cases are described as below. For the urban firefighting, a crucial aspect is its density, i.e. population, network nodes, and traffic. Concentrated population around the site, both victims and passers-by, intensifies the threat of a fire accident and increases the complexity of countermeasures that need to be taken. Furthermore, diverse and abundant network nodes, which aid us in our normal lives, could hinder the operation at its critical moment. Nevertheless, we consider this setting advantageous for us because we take the urban environment as an adequate setting for a more sophisticated and versatile smart firefighting operation by fully utilizing two key enablers for our scenario: drones and 5G networks. Contrary to the urban case, the lack of existing infrastructure is a challenge to overcome in the rural firefighting. It is possible that the fire area is out of the reach of existing mobile network except for traditional voice and low data rate services. In this case, fast deployment of communications between the firefighters, incident commander, drones, and the control center is a key requirement for the smart firefighting operations.

1.3 Structure of the document

This deliverable is organized as follows. Section 2 presents the overview of the E2E architecture. Then, each component of the overall system is introduced between Section 3 and Section 8. In Section 3, UEs specific for the scope of the project are discussed. The RAN and the core network architecture are presented in Section 4 and Section 5, respectively. Section 6 describes the edge architecture. In Section 7, the network management and orchestration are discussed. Then, Section 8 presents the services and applications for public safety. This is followed by Section 9 which explains the relation with the PriMO-5G demonstrations. Finally, Section 10 provides concluding remarks with suggestions for potential enhancement of the 3GPP architecture.

1.4 Relationship to other project outcomes

This deliverable is a result of the interaction with WP2-WP4 to collect component technologies investigated in the project. Together with D1.1 [PRI19-D11] produced in Task1.1 of WP1, this deliverable will be used by WP5 for defining reference demonstration scenarios. At the same time, an input from WP5 was employed by this deliverable for identifying key technological components needed for the E2E architecture.

2 Overall Architecture

The Primo-5G public safety architecture builds on the architecture standardized in 3GPP. Therefore, this section starts with a brief overview on the 3GPP Release-16 5G architecture in Section 2.1. It is followed by the presentation of the E2E PriMO-5G architecture in Section 2.2. Finally, Section 2.3 gives an overall picture of the architecture highlights, including E2E network slicing and edge computing.

2.1 3GPP Release 16 5G architecture

The 3GPP Rel-16 5G architecture is shown in Figure 2.1-1 below:

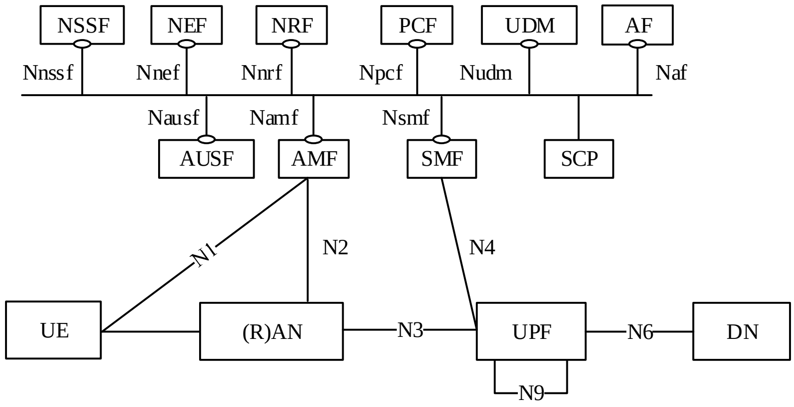


Figure 2.1-1: 3GPP Rel-16 System Architecture [3GPP-23501].

Compared to earlier releases of 3GPP, the system architecture has undergone changes both in the control plane as well as in the user plane. A service-based architecture has been introduced with service-based interfaces in the mobile packet core. Slicing concepts have been strengthened and are supported both in the service-based architecture as well as in the QoS framework for the E2E system architecture.

In core network, separate session management and mobility management functions have been reengineered and been given new entities: SMF and AMF. The concept of the user plane has been reengineered and a new entity has been introduced: UPF. The RAN has been further evolved with split options for control and user.

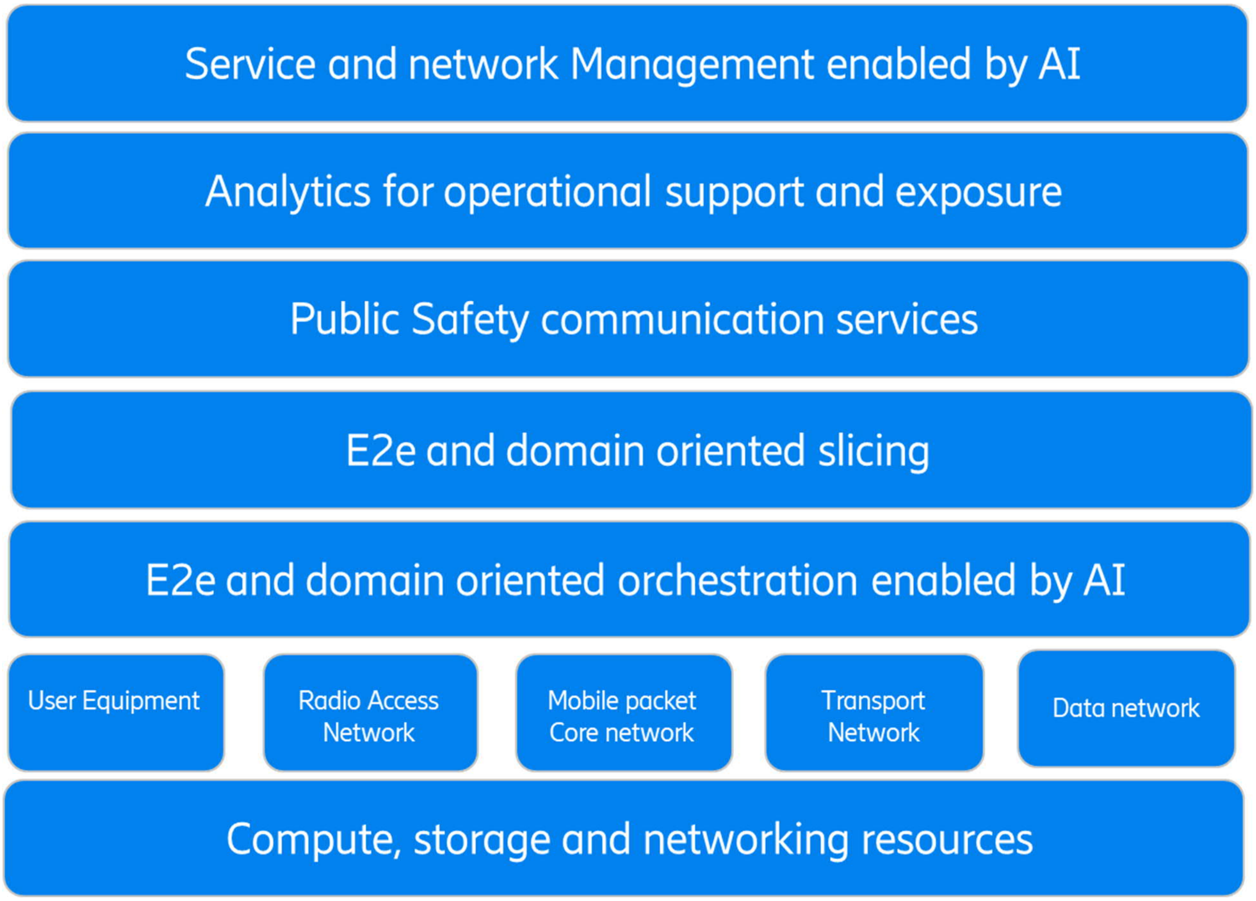
In the 3GPP Release-16 architecture, additional components have been introduced to support artificial intelligence with data management integrated into the architecture. Two new entities have been introduced: NWDAF and MDAF. Support for edge computing is evolved to support an evolution of different types of edge architectures.

Throughout this document descriptions of the system architecture are provided that will detail the overall changes. Notable sections are:

- 4. Radio Access Network Architecture
- 5. Core Network Architecture
- 6. Edge Architecture
- 7. AI-enabled Network Management and Orchestration

2.2 PriMO-5G architecture for public safety communications

Public safety communications are moving from a voice-centric paradigm to a data-centric paradigm enabled by LTE, 5G and from ongoing the 3GPP evolution. With 5G, additional capabilities are added, and other capabilities have been evolved, such as network slicing. In the PriMO-5G project, public safety communication requirements have further been refined by looking into advanced video services as well as aerial components, integrated AI, and edge computing. The PriMO-5G project proposes a multi-stacked architecture as the architectural response to cater for the next generation of public safety data centric communication architecture, see Figure 2.2-1 below:



Service and network Management enabled by AI

Analytics for operational support and exposure

Public Safety communication services

E2e and domain oriented slicing

E2e and domain oriented orchestration enabled by AI

User Equipment | Radio Access Network | Mobile packet Core network | Transport Network | Data network

Compute, storage and networking resources

Figure 2.2-1: Public safety communication architecture, moving from a voice-centric to a data-centric paradigm.

At the bottom of the architecture we have all the compute, storage, and networking resources. These are supporting the whole architecture and enables virtualization and integration of cloud-based infrastructure supporting network functions and applications.

The next layer includes infrastructure components that provide the physical transmission of data. For all of these components smaller or bigger revolutions are happening. UEs is evolving dramatically with aerials, IoT devices, and with more datacentric capabilities. RAN has a variety of architectural approaches and split options are introduced. Transport is evolved with new transport concepts and with integration of E2E capabilities for e.g. network slicing, Core networks are moving to a service-based architecture. Advanced solutions to enable edge computing, including breaking out traffic to local edge servers are introduced. This gives new options in interfacing data networks enabling both centralized and distributed computing.

To handle the physical infrastructure an orchestration layer is needed. Orchestration is happening on several layers for E2E orchestration down to domain specific orchestration. AI can be integrated to enhance placement decisions.

Public safety communication services are evolving from voice with low data requirements towards a variety of services where the public safety communication system is interacting with a variety of other functions, e.g. air traffic control.

With the 5G architecture supporting a variety of how to expose data, analytics functions can be evolved to handle operational analytics needs as well as external exposure of analytics. Service and network management is moving into being more integrated into real time communication support and is further enabled by integrated AI supporting fault, configuration, accounting, performance, security (FCAPS) and other management services.

2.2.1 Introduction to sub-areas

The PriMO-5G architecture consists of a set of layers to provide E2E public safety communications. The different layers are touched upon on the subsequent sections of the document. Each section defines an overall subarea where certain aspects have been highlighted.

User equipment, as described in Section 3, deals with important types of user equipment that can be connected to a public safety communication network. In this section highlighted areas are: Drone for 5G Live Streaming, Head-mounted devices for end users, and Fire truck with mobile computing server.

Radio access architecture, as described in Section 4, covers all types of RAN that can be operated and supportive to a wireless public safety communications system. In this section highlighted areas are: RAN split options, integrated access and backhaul, cell-free architecture, drone base stations, moving terrestrial base stations, QoS framework in RAN, RAN slicing, and RAN enhancements for URLLC.

Core network architecture, as described in Section 5, covers different core network functions and functions needed in a public safety system. In this section highlighted areas are overview of 5G core network, core network slicing, 5G QoS framework, core network enhancements for URLLC, non-public networks, Isolated Operation for Public Safety (IOPS), Optimal Routing, NWDAF, and 5G-LAN.

Edge architecture, as described in Section 6, covers functions to provide edge computing. In this section two architecture are highlighted: ETSI MEC architecture and the 3GPP edge architecture. Edge-assisted AI applications are also covered in the section.

AI-enabled network management and orchestration, as described in Section 7, covers how AI functions can enhance network management and orchestration. In section highlighted areas are the management data analytics function and network slice orchestration.

Services and applications for public safety, as described in Section 8, covers current defined services in 3GPP as well as opportunities to create applications on demand using cloud computing. In this section highlighted areas are mission critical services, and immersive applications using extended reality

Relation towards PriMO-5G demonstrators, as described in Section 9, covers component demonstrators, local and intra-continental demos and PriMO-5G intercontinental demos.

2.3 End-to-end architecture highlights

2.3.1 Cellular Network for Public Safety with aerials

Cellular networks such as the 5G network have many inherent capabilities that make them suitable for satisfying requirements the public safety communication services need.

In the PriMO-5G context, public safety UE may be handheld devices, but also aerials, e.g. drones. The cellular network may also be extended with new type of base stations or network elements providing network functionality, e.g. drones or firetrucks serving as base stations, or firetrucks hosting compute facilities.

The key capabilities of 5G cellular networks benefiting public safety (including aerials) are the following:

- Beyond-line-of-sight: Cellular networks allow the drones to be controlled from larger distances than line-of-sight would allow.

- Extended reach and capacity on-demand: The system can be on-demand extended, e.g. new base stations and sites can be deployed. With moving base stations, the deployment can dynamically change. Drone base stations or vehicular (e.g. firetruck) base stations can be directed to a given area for increased coverage and capacity.
- Mobility: This is one of the fundamental properties of cellular networks. Public safety UE utilizing the system are inherently mobile within a large area thus they would benefit from mobility capabilities.
- Connectivity for multiple services: The same infrastructure can be used for multiple services, thus eliminating the need of costly integration of multiple communication systems. Video, AR/VR, drone control (remote and local), voice etc. services can be served with the same infrastructure.

2.3.2 Edge computing

Edge computing provides compute and storage resources with adequate connectivity (networking) close to the devices generating traffic. It is about bringing the services closer to the location where they are to be delivered. The motivation is to reduce latency and reduce transmission costs. Example use cases include AR/VR, real-time facial recognition, video surveillance, etc.

Cloud applications thus can be moved closer to the end devices to edge computing locations, for example from centralized data centers to distributed data centers. This, at the same time, results in opportunities to offload processing from the end devices and perform the tasks in the edge computing locations.

Edge computing capabilities in 5G is advantageous due to the following factors:

- Application latency: The application gets closer to the user and the 5G Radio, and a few milliseconds of latency is possible to achieve. This can enable new applications and use cases in the network.
- Transport offload: Backhaul traffic is reduced, as there is no need to transfer all the traffic from the devices to central data centers.
- Processing offload: Devices can be offloaded, thus saving battery. At the same time, the user experience can be preserved.
- Local survivability: Localized application deployments may continue to be reachable if the communication links towards central components of the network fail.

The beneficial location for different applications may vary in the network, so different applications may need different edge computing locations. Applications may also be distributed requiring distributed cloud, where the application is run in multiple locations in the network. This placement may be determined by latency needs, bandwidth saving opportunities, or availability and cost of compute facilities. Placement of application resources, compute resources and network resources are enabled by the interaction between E2E orchestration and various domain orchestrations.

2.3.3 End-to-end network slicing

2.3.3.1 Network slicing

The “one-size-fits-all” paradigm for the legacy communication systems is not suitable for the new requirements mobile networks face. Flexibility, among others, is a very important requirement. The network should support a wide variety of different use cases with very diverse requirements, and a large number of deployment options. With network slicing, different E2E logical networks can be built on the top of a common and shared infrastructure layer.

The definition of “Network Slice Instance”, from NGMN is the following [NGMN16]:

“Network Slice Instance: a set of network functions, and resources to run these network functions, forming a complete instantiated logical network to meet certain network characteristics required by the Service Instance(s).

- A network slice instance may be fully or partly, logically, and/or physically, isolated from another network slice instance.
- The resources comprise of physical and logical resources.
- A Network Slice Instance may be composed of Sub-network Instances, which as a special case may be shared by multiple network slice instances. The Network Slice Instance is defined by a Network Slice Blueprint.
- Instance-specific policies and configurations are required when creating a Network Slice Instance.
- Network characteristics examples are ultra-low-latency, ultra-reliability etc.”

As a general definition, a network slice is a logical network serving a defined business purpose or customer, consisting of all required network resources configured together. The network slice is a complete network, it is for any type of access, and enabler of services. The physical or virtual resources may be dedicated to a network slice or shared between network slices. The concept is illustrated on the following schematic figure (where the gray horizontal rectangles are different network slice instances, while the blue boxes within the network slice instances are resources that may be shared between network slices or dedicated to a network slice).

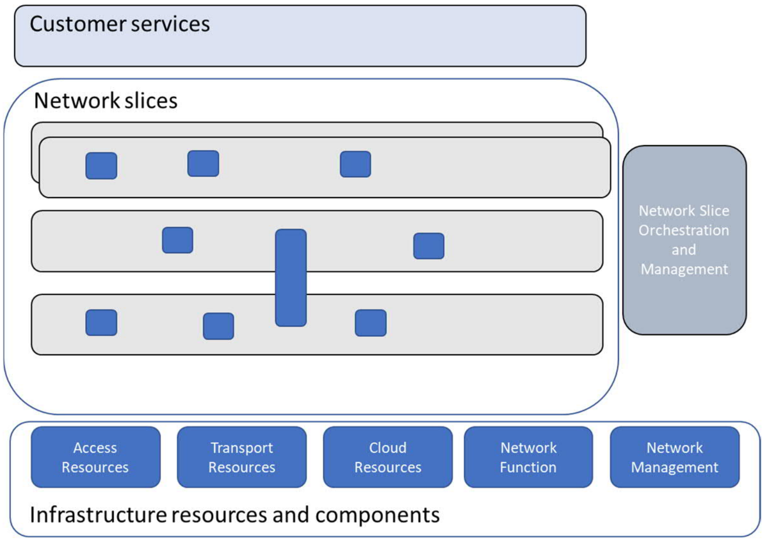


Figure 2.3-1: High-level view on network slicing.

In PriMO-5G context, network slicing allows public safety organizations to structure their network based on their needs:

- Per-public safety organization (police, firefighting)
- Per-group of services
- Following regulatory and technical requirements

Multiple network slices can be built on the common infrastructure, while each of them is realizing a wanted network characteristic supporting a customer need. Customer, in this definition, may be an enterprise, another service provider, or even the network operator itself. In the PriMO-5G context, the customer may be a public safety organization (e.g. police, firefighting organization etc.), too.

While, to a certain degree, separate logical networks can be achieved in the 4G of mobile networks, network slicing became a central component in 5G, allowed by key advancements in virtualization, automation, and SDN to achieve the desired flexibility in a cost-efficient manner.

The benefits of network slicing, used for public safety, are the following, combined with virtualization and automated orchestration and management are the following:

- Per public safety customer adaptation and optimizations lead to better customer experience
- Shorter time-to-market for making the public safety system operational
- Flexibility
- Separation of concerns

For 5G, three key use cases have been identified: eMBB, URLLC, and mMTC. These different use cases come with different requirements that a network slice should fulfill. URLLC use cases require very high availability, very high reliability, and very low latency. eMBB typically requires access to Internet and operator services, and wide area coverage. mMTC requires to support a very high amount of devices with low cost, in an energy-efficient way.

These requirements lead to the fact that the different functions of the mobile network may be placed in different parts of the mobile network. For instance, low latency often needs distribution of the user plane closer to the radio access; while on the contrary, the cost-efficiency of mobile-broadband type of deployments may lead to more centralized placements, due to economies of scale.

Public safety customer expectations may vary with regards to degree of isolation, coverage area, security requirements, service availability and service reliability requirements, priority requirements etc. may differ between customers, even for the same type of use case. These expectations may be also based on regulatory requirements, apart from traditional technical requirements, such as E2E latency.

In the following Figure 2.3-2, a network with multiple network slices is illustrated, spanning the radio access and the core network domains. Core Network User Plane (CN UP) and Core Network Control Plane (CN CP) Functions may be placed in different locations for different slices. Some CN CP Functions may be shared between the slices. In RAN, resource allocation mechanisms consider the public safety requirements to fulfill the agreed service level. Edge applications are possible to run in different locations (edge sites) in different slices, and the location can be determined based on application-specific requirements, but also considering the availability and cost of connectivity and computing capabilities, during the interaction of dynamic orchestration of the various domains.

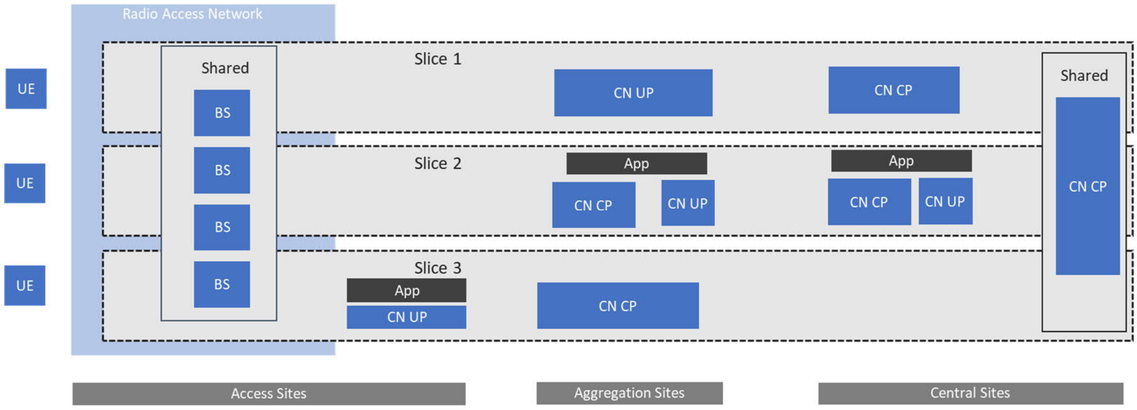


Figure 2.3-2: Edge computing and network slicing combined.

2.3.3.2 Life-cycle management and orchestration of network slices

As large number of networks slices may be provided in an operator's network, and flexibility is a key requirement, an automated network slice orchestration and management is needed to efficiently operate the network slices. This includes the process of preparation and the life-cycle management of network slice instances.

Templates are used to describe network slices. If a template already exists that can meet the customer requirement, the preparation phase can be skipped, and the creation and activation can begin. In this case, either an existing network slice instance is scaled to support the requirements of the new customer, or a new one is created utilizing the existing template. Otherwise, a new template is designed to meet the requirements, and it is verified and on-boarded, and added to the catalogue of network slice instances.

In the commissioning phase, the network slice instance is created, including the allocation and configuration of all the needed resources. The operation phase includes the activation, supervision, performance reporting (e.g. for KPI monitoring), resource capacity planning, modification, and de-activation of a network slice instance. At activation, the network slice instance is made ready to support communication services. Supervision and reporting include, e.g. monitoring, performance assurance and reporting of KPIs as agreed in the SLA. Modification may include capacity or topology changes and may be triggered by changing requirements or the supervision and reporting. The network slice instance is stopped at the de-activation step. Finally, at the decommissioning phase, all the network slice specific configurations are removed, and the network slice instance is terminated.

The phases described above can be followed in Figure 2.3-3.

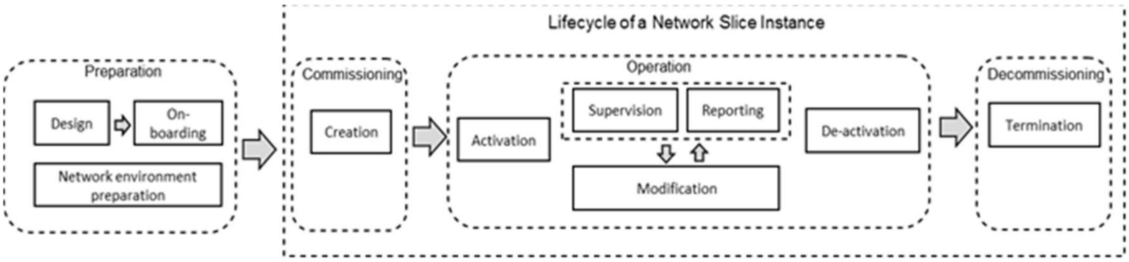


Figure 2.3-3 Preparation and lifecycle of a network slice instance. [3GPP-28530]

3 User Equipment

In this Section, the scope is extended to consider 5G UE to its fully dynamic nature. The rapid growth of UAVs can be new types of UE, and the UAV market is creating significant interest of academia as arising issues from moving nodes in cellular networks. 3GPP is considering connecting UAVs to cellular networks. However, academia is accelerating to research with the full potential of UAV communications. The E2E network of this task shows the diversity of UE, which is one of the representative characteristics of 5G. In this section, we consider not only the existing devices of users, but also UAV units that transmit and receive 5G live streaming, servers in the ground control center that support UAV missions and operations, and users of VR and AR at the end of the network. HMDs required to provide XR tangible media services such as mixed reality (MR), and mobile computing servers installed on fire trucks to deliver tangible media to firefighters, process images, and support computation are listed and described as UEs. Haptic equipment helps a better control of robots in the incident site.

3.1 Drone for 5G live streaming

In the proposed E2E system, drones could act as either end. As the gatherer of situational information granting situation awareness to other entities, drones act as primary servers responsible for transmitting situational information. On the other hand, when another entity controls their mission executions or movements, drones act as end receivers for orders. Although these aspects are on the opposite ends of a communication pipeline, their operation is never separated. To explain, controlling a drone to move to a new location changes the wireless channel conditions and the video coverage of the situation. Also, increased situation awareness granted by the drone changes how someone controls the drone by making it execute a different mission or explicitly moving it to a different location.

To support such flexibility and stability, rotor-based drones, thanks to their ability to hover, are preferable to fixed wing drones, despite their longer flight times. The choice of rotor number, from one to eight, depends on the safety and efficiency constraints. To elaborate, lower number of rotors are known to increase flight efficiency, thus increasing flight time, but also poses increased risk of fatal accidents due to their larger rotors, instability, and higher speed.

As mentioned above, drones simultaneously act as a server and a client of the 5G network in PrIMO-5G E2E architecture.

When the drone acts as the server of the 5G network, the camera mounted on the drone provides the source data for the drone to transmit. The video from the camera capturing the surrounding environment of the drone is live streamed through the 5G network to the GCS or an MEC. This video, after procedures including preprocessing, postprocessing, etc., is reproduced and consumed along the pipeline, depending on the video, system, and application specifications. Consequently, the camera being mounted on the drone should be chosen accordingly, whether it supports 4K resolution, over 60 Hz frame rate, 360 VR, 3D, infrared, thermal, and so on.

When producing a video, a drone’s camera must possesses the capabilities to mitigate the drone’s movement, ranging from subtle vibrations to quick movements. This is because drones in a mission are naturally distant from the subject of interest and seldom at complete rest. Specifically, higher resolution and framerate are required to battle distance and movement, respectively. However, the drone’s flight time is heavily dependent on how much weight it is carrying, and cameras tend to become heavier as their quality is increased, thus inevitably cutting the flight time short by lowering the flight efficiency. In consequence, the project’s E2E architecture require further advancements that the 5G network, mobile edge computing, and task offloading schemes bring, which will be documented later. To briefly explain, faster transportation to an end with an alternate display device with higher computation capability at the site is a promising alternative for a high-end camera. Therefore, if a rawer video with less processing can be reliably transmitted through 5G, then even a more compact camera with lesser capabilities could accomplish the mission equally well, acting as a heavy 5G uploader.

At the other end of the system, a drone acts as a client of the 5G communication, receiving remote control message from a server, to execute a specified mission. To execute missions and orders that depend on the video provided from the camera, the drone must receive the control message with very low latency, aided by the RAN enhancements for URLLC, which is to be introduced in Section 4.

3.2 Head-mounted devices for end users

As 5G networks provide the increased bandwidth capacity and decreased latency, the popularity of complex immersive virtual experiences is increased. Such experiences can be consumed on HMDs, VR glasses, AR glasses, or any form of emerging device.

An HMD is an image presentation gadget, worn on the head, that has a small display panel in front of one eye (monocular HMD) or each eye (binocular HMD). Usually, an HMD has one or two small displays. Eyepieces have embedded lenses and semi-transparent mirrors. The mini- display unit may comprise cathode ray tubes (CRT), liquid-crystal displays (LCDs), liquid-crystal on silicon (LCos), or organic light-emitting diodes (OLED). HMDs differ in whether they can only display computer-generated imagery (CGI), or live physical world imagery alone, or a combination. The performance parameters are listed in Table 3.2-1.

Table 3.2-1: Performance parameters of HMD.

Performance parameters
Ability to view stereoscopic images.
Resolution
Field of view (FoV)
Interpupillary distance (IPD)
Collimation
Binocular disparity
Stationary
Synchronous
Degree of freedom (DoF)
Latency

AR and VR applications can be very sensitive to network performance, with tiny disruption causing a heavy unfavourable experience on users. Exposure to HMD systems often causes unfavourable physical conditions called VR-sickness. Some parameters are highly related to multi-modal sensory cues. The conflict between or inside those factors can cause VR-sickness. For example, a video frame moves slower than eye motion; user feels dizziness. So, the latency is a critical factor of this example.

An HMD has many applications, including sports, education, military, medicine, and engineering. A scientific breakthrough, like 5G network, HMD technology also will be soaring to new heights. The current strict standards of 3GPP need to be extended to support the dynamic nature of VR, AR devices.

3.3 Fire truck with mobile computing server

In the event of a fire or emergency, various types of drones can be used over the area to identify the site of the accident and to explore ways to deal with the accident. Due to the limited size of the hardware, drones have problems with their own handling of real-time information collected by drones, with limitations in their battery, computing performance. Fire trucks can act as a central server to manage and process information collected from drones over accident sites, such as virtual edge servers. And this allows for problem solving due to drone limitations.

Flight time after the accident is also an important issue for drones to stay in the air. Sufficient batteries must always be maintained to ensure that multiple drones can fly for long periods of time while communicating smoothly between drones, or with fire trucks.

To address these problems, studies may be conducted such as exploring the optimal path of travel for charging drones or determining the order of charging for multi-drone. In the scenario in which fire trucks become one small mobile base station and drones in the area of accident inside the cell communicate with the fire trucks, fire trucks become central controllers to determine the optimal path of travel for charging drones and the sequence of charging movements for multi-drone.

3.4 Haptic Equipment

In firefighting scenario, remote-controlled robots are used to help put out fires and rescue people. The haptic communications enable the control center to control UEs' haptic equipment remotely with low latency and high reliability. An example of the use of haptic equipment is depicted in Figure 3.4-1 where a multicast system is set up to test a multi UE scenario. When the operator manipulates the master robot, a force vector data is transmitted to UEs and the slave robots at UEs regenerate the movements by the force vector data. If there is an unexpected situation such as collision or resistance, the operator can feel haptic feedback helping adjust an input control. The haptic data is compressed and packetized to connected with the wireless link over UDP in real-time.

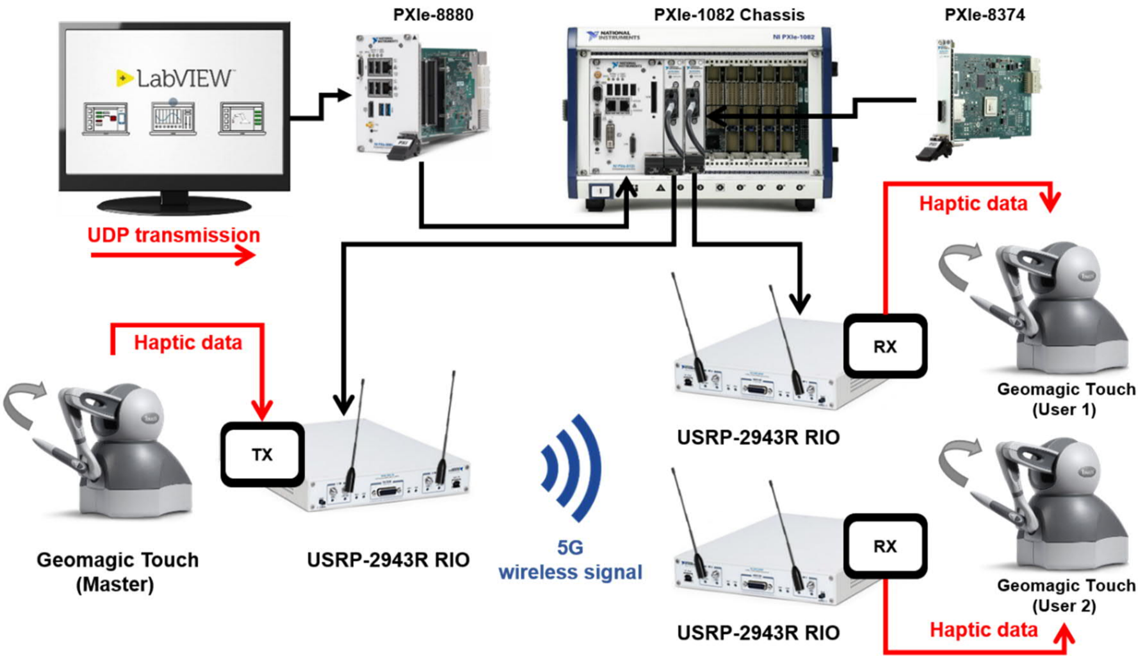


Figure 3.4-1: The system description of haptic communication.

4 Radio Access Network Architecture

The main 5G service types typically considered are eMBB with several gigabits data rate per second and reliable broadband access over large coverage areas, mMTC requiring wireless connectivity. For example, URLLC requires E2E latencies of less than 5 ms and over 99 percent reliability for V2X communication [MDB+16].

To fulfil these requirements, investigations toward an overall 5G RAN architecture that can efficiently support are still in process. To satisfy the QoS and KPIs required by the use cases described in [PRI19-D11], it is essential to improve the technology at the RAN. In this regard, this section introduces advanced technologies such as RAN split, cloud RAN (C-RAN), radio access and backhaul network convergence (refer to [3GPP-38840] and [3GPP-38874]), cell-free network structure, drone base station, moving terrestrial base stations, RAN slicing and RAN technology to support URLLC. While the document introduces various technology elements, it presents how each technology is included in the E2E network and the expected effect of them.

4.1 Radio access network split options

4.1.1 Introduction

The origins of RAN split can be traced to the early deployments of the mobile network. To reduce the loss of the RF cable connecting the radio front end to the antenna mounted at the top of a tower, the radio unit (RU) must be located close to the antenna. Therefore, most vendors provided disaggregated base stations, exposing a base station internal interface between the RU and baseband unit (BBU). The RU could then be placed close to the antenna, while the BBU, which hosts all the processing power, is hosted in a sheltered facility in a nearby location. The high bandwidth, low latency link connecting the RU to the BBU is called fronthaul and the protocol that runs over fronthaul is an industry specification known as common protocol radio interface (CPRI) [CPRI13].

4.1.2 Cloud RAN (C-RAN)

With the introduction of smart phones and the dramatic increase of mobile broadband traffic operators started facing new challenges as the revenue growth was not in pace with the growth in data volumes. To cope with the increased investment and operational cost China Mobile introduced a new RAN split architecture called C-RAN [CRAN11] which leverages the base station split by exploiting the technical advances and virtualization technology on general purpose processors in the data centers to fundamentally change the cost structure of mobile operators. In C-RAN, the radio parts called remote radio units (RRUs) are deployed close to the antenna sites and they are connected to a pool of virtualized BBUs located in a centralized location over a long distance fronthaul links. Centralizing the base band processing incurs a high fronthaul transport costs, however, it also brings a number of advantages including reduction of OPEX due to pooling gain, ability to implement advanced features such as Coordinated MultiPoint (CoMP) and interference coordination.

4.1.3 5G RAN Transport Architecture

As 5G has to support a large number of use cases with diverse requirements, 3GPP and other organizations have been exploring new functional splits that support alternative RAN deployments to address the additional requirements from the use cases. For example, in [NGMN18] and [3GPP-38801], 3GPP introduces two levels of functional splits, higher layer split (HLS) and a lower layer split (LLS), with 8 different split options using LTE protocol model as shown in Figure 4.1-1.

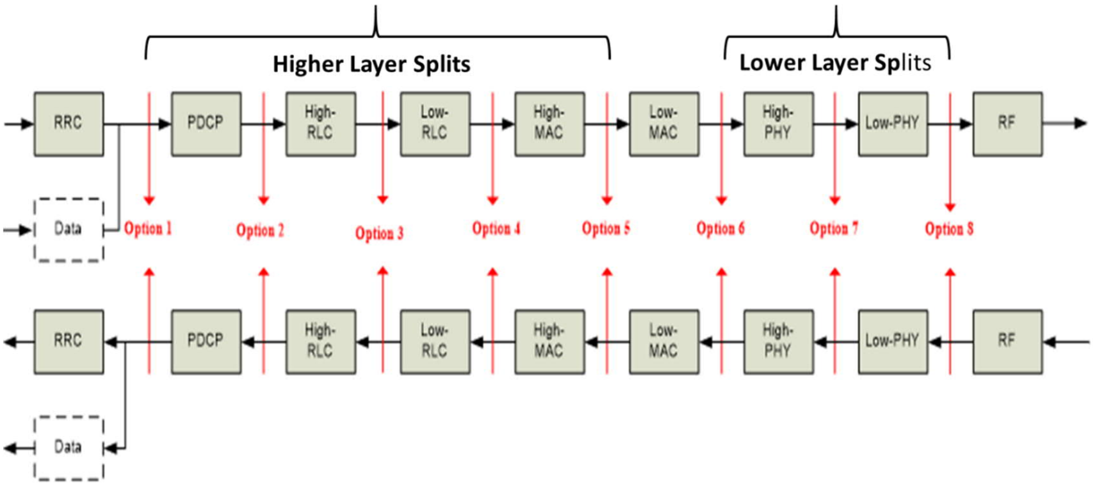


Figure 4.1-1: Function split between Central Unit (CU) and Distributed Unit (DU).

RAN functional split has a number of implications and is tradeoff between features such as latency, transport bandwidth, complexity and flexibility. In general, however, the higher layers' split options provide more relaxed latency and bandwidth requirements while the lower layers' splits would enable more centralization operational gains, CoMP, and interference suppression. And after studying the different split options 3GPP selected option 2 of Figure 4.1-1 as the HLS, and the base station was disaggregated into a central unit (CU) which hosts the RRC, PDCP and service data adaption protocol (SDAP) functions, while the distributed unit (DU) hosts the RLC, MAC and the PHY functions. SDAP is a new NR user plane protocol not shown in the Figure 4.1-1, to handle flow based QoS, such as mapping of QoS flow and radio bearer and QoS flow ID marking. The link connecting the DU and CU is called midhaul and this interface is expected to be interoperable between vendors. The CU itself was further split into user plane (CU-UP) and control plane (CU-CP) interfacing the DU over F1-U and F1-C protocol interfaces. The choice of LLS, which is mainly between option 6 and option 7 is still being studied.

4.1.4 Implication of RAN split options in PriMO-5G use cases

PriMO-5G has developed two use cases with firefighting scenario in a forest and in urban areas where UAVs are expected to assume key roles in information acquisition and communication [PRI19-D11]**.** In both cases, UAVs provide a high-altitude high definition video and sensory information of the fire scene to the Incident Commander deployed in one of the firefighting tracks.

Use cases communication requirements

The communication requirements in the UAV supported firefighting Use-Cases include:

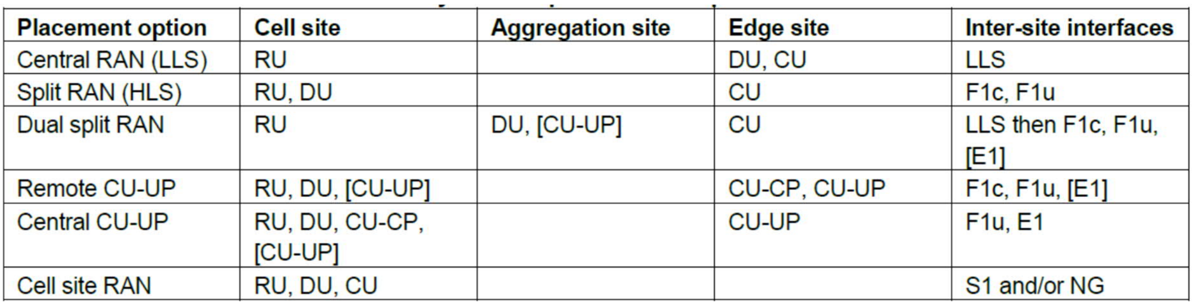
1. Command and Control (C2) communication between UAVs and the UAV controller, which requires low latency (50 ms one-way latency from eNB to UAV) and low bit rate (60-100 kbps for UL/DL) communication link [3GPP-36777].
2. Aerial, visual, and other sensory data of the fire scene taken from the air by UAVs sent to the incident commander. Depending on the video quality and the type of sensory data, the communication capacity required is at least a couple of Mbps or higher.
3. Sensor information, e.g., levels of measured toxic or hazardous substances, from firefighters and robots at the fire scene to the incident commander. This is a low bit rate, in the range of tens of kbps, and low delay link.
4. 360 degrees, VR or AR, video traffic from firefighters and robots in the fire scene to the incident commander, which requires very high data rates and very low delays. Again, depending on the

quality, a low resolution of 360 VR requires at least 25 Mbps for streaming, while high resolution comparable to HD TV requires 80 -100 Mbps.

4.1.5 5G-RAN split deployment for the PriMO-5G Use Cases:

The functional entities RU, DU, CU-UP and CU-CP can be placed at different physical locations depending on the site constraints, transport network topology and latency and capacity requirements of the application running on the cell. For the deployment of PriMO-5G use cases, the most critical communication parameters are the delay requirement for the command and control of the UAVs communication, the radio interface capacity to support the different video applications and the availability of transport capacity for fronthaul and backhaul in the location of fire scene. Table 4.1-1, from [NGMN18], summarizes the functional placement of the CU, DU and RU for different scenarios.

Table 4.1-1: Summary of RAN functional placement of the CU, DU and RU for different scenarios. [NGMN18]



Placement option	Cell site	Aggregation site	Edge site	Inter-site interfaces
Central RAN (LLS)	RU		DU, CU	LLS
Split RAN (HLS)	RU, DU		CU	F1c, F1u
Dual split RAN	RU	DU, [CU-UP]	CU	LLS then F1c, F1u, [E1]
Remote CU-UP	RU, DU, [CU-UP]		CU-CP, CU-UP	F1c, F1u, [E1]
Central CU-UP	RU, DU, CU-CP, [CU-UP]		CU-UP	F1u, E1
Cell site RAN	RU, DU, CU			S1 and/or NG

For the forest firefighting use case, it is highly unlikely to find a high capacity transport facility in location, therefore a solution based on monolithic or cell site RAN can be used. An example of such a solution is shown in Figure 4.1-2 (a), in which case all the functions of the base station are deployed in the cell site. In this case, the UPF and the main compute resource can also be co-located with the command center in the fire track or close to it in the cell site.

Similarly, the monolithic RAN solution can also be used for the urban firefighting scenario. Alternatively, when the required transport capacity is available in the fire area, an HLS solution based on remote CU-UP, where the CU is split into user and control planes with the CU-UP deployed in the cell site, while the CU-CP can be moved to the edge of the network can be used. Besides the compute resource at the cell site, this would also enable to make use of additional compute resource at the edge or central location as shown in Figure 4.1-2 (b).

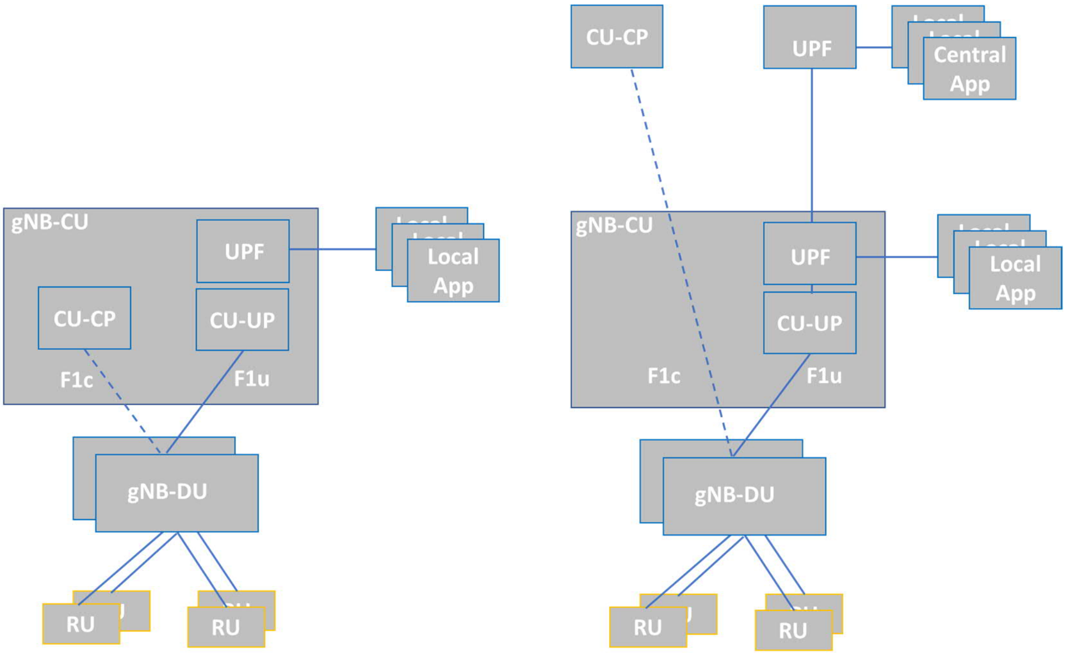


Figure 4.1-2: (a) monolithic or cell site RAN (b) Remote CU-UP HLS RAN.

4.2 Integrated Access and Backhaul

In 3GPP Release-16 the integrated access and backhaul (IAB) concept is being studied. Related specifications are [3GPP-38874] and [3GPP-38840].

A key benefit of IAB is enabling flexible and very dense deployment of NR cells without densifying the transport network proportionately. Public safety firefighting often requires the quick deployment of a public safety network fast and to provide coverage where the existing network may be out of coverage or damaged. IAB provides a possibility to fast set up a network and being able to connect devices to the IAB node. For public safety IAB nodes could be either terrestrial based or non-terrestrial based. The overall concept is shown below:

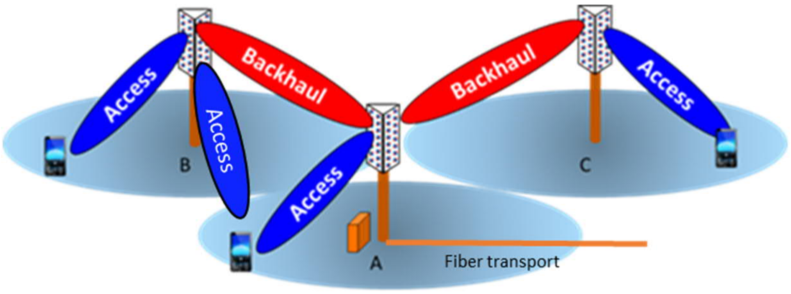


Figure 4.2-1: Integrated Access and Backhaul concept. [3GPP-172290]

The IAB concept aggregates traffic from several base stations (relay nodes) onto one base station (donor node). In Figure 4.2-1, traffic from the relay nodes to the donor node is shown as backhaul. From donor node an aggregated stream can be fed into the transport network which could be of any kind e.g. fibre or microwave. An IAB node has a north bound side that looks like a UE from the relay node to the donor node. On the south bound side, the IAB node looks like a radio base station.

Multi-hop backhauling provides more range extensions than single hop. This is especially beneficial for above-6GHz frequencies due to their limited range. This means that IAB allows a rapid deployment of smaller cells. Multi-hop backhauling further enables backhauling around obstacles, e.g. buildings in urban environment for in-clutter deployments.

The maximum number of hops in a deployment is expected to depend on many factors such as frequency, cell density, propagation environment, and traffic load. It also means that the uplink traffic characteristics will change. As aggregation occurs the uplink will be more heavily loaded.

Below in Figure 4.2-2 follows a reference diagram for the IAB architecture:

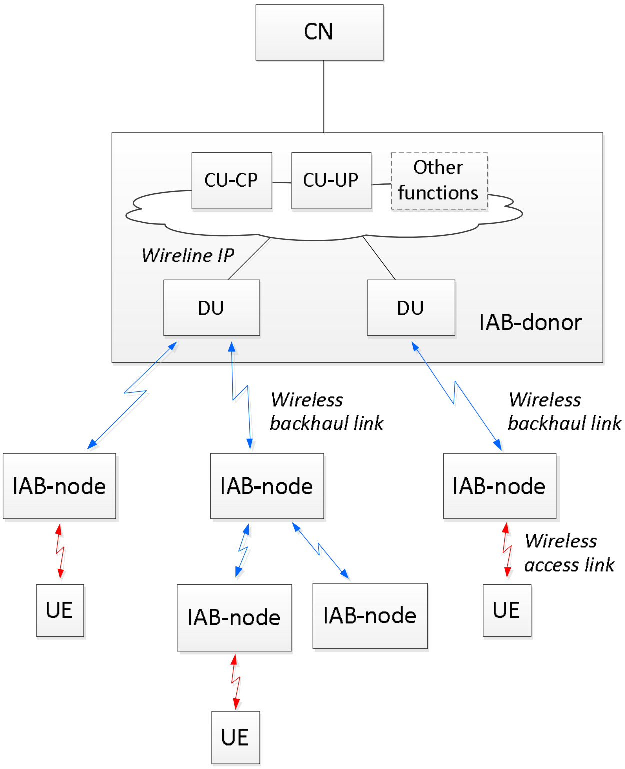


Figure 4.2-2: Reference diagram for IAB-architectures (SA mode). [3GPP-38874]

The focus of [3GPP-38874] has been on backhauling of NR-access traffic over NR backhaul links. IABs can be deployed with several topologies, see Figure 4.2-3 below:

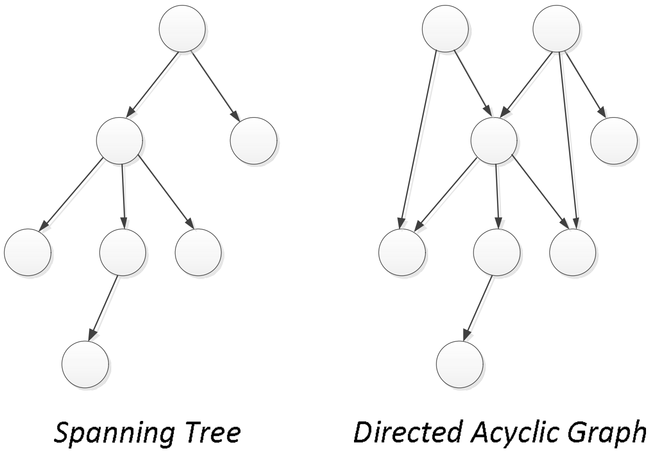


Figure 4.2-3: IAB network topologies. [3GPP-38874]

For terrestrial network deployment spanning tree will most likely be frequently used. But for swarms of drones the directed acyclic graph could also be an option where a master node for a drone could be an IAB connected to sub IABs in the swarm.

For NAS protocols an adaptation layer is implemented to support IABs; the backhaul adaptation protocol (BAP). The BAP would be above the RLC level, see Figure 4.2-4 below:

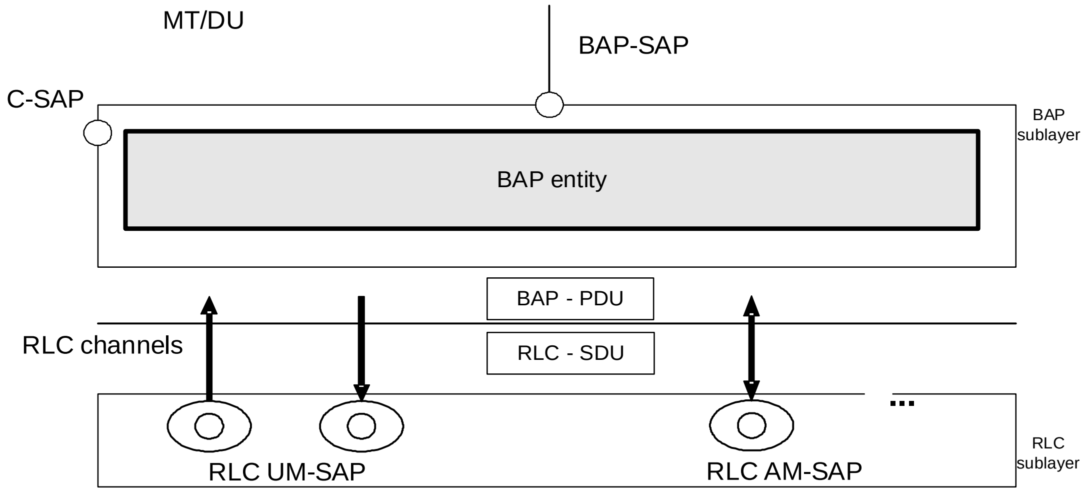


Figure 4.2-4: BAP layer, structure view. [3GPP-38340]

4.3 Cell-free architecture

The demands being placed on cellular networks continue to increase rapidly. Forecasts project up to a 1000-fold increase in the needed system capacity while concomitantly reducing latency by an order of magnitude. Meeting this increased demand is typically accomplished through network densification. While dense networks help in providing greater capacity, they increase the number of handovers that occur as UEs traverse cells more quickly. 5G networks are also aiming to support new application types. Notably, many of the envisioned use-cases involve vehicular UEs. These, by their nature, move faster than pedestrian users, further compounding the issue of the high rate of handovers. PriMO-5G scenarios call for drone swarms and fire trucks to be deployed rapidly into possibly difficult terrain with little time for planning the network's deployment. Consequently, UEs will likely often cross cell borders. This is particularly true of aerial UEs.

Cell-free architectures aim to alleviate the signalling overhead caused by frequent handovers. Instead of the traditional approach of UEs performing cell selection measurements and reporting them back to the network, positioning provides the input for the handover decision making. UEs transmit beacon signals in uplink to allow the network to compute their position. DUs utilise these beacons to perform angle-of-arrival calculations. This information is then relayed to the CU. UEs are not explicitly aware of which DU is serving them. All measurements and state information resides in the CU. The CU monitors the location of each UE to detect when it has moved to the service area of a DU other than its currently allocated one. At this point, the CU will send instructions to both the source and target DUs to execute the transfer. Since UEs are not aware of their serving DU, the transfer should in principle be seamless.

Figure 4.3-1 depicts an example scenario where a small network is deployed to assist firefighting. The CU controls three DUs placed around the building on fire as the terrain and conditions permit. UEs represent UAVs, firefighters, and other equipment. Upon connection to the network, they establish state with the CU. After this, they can move around and be served by the DUs deemed best by the CU. For instance, as UAVs circle the building, they can provide a constant video stream without interruption that would have been required by the legacy measurement-based approach. Additionally, re-deployment of DUs or arrival of additional equipment can be performed transparently from the UEs' point-of-view.

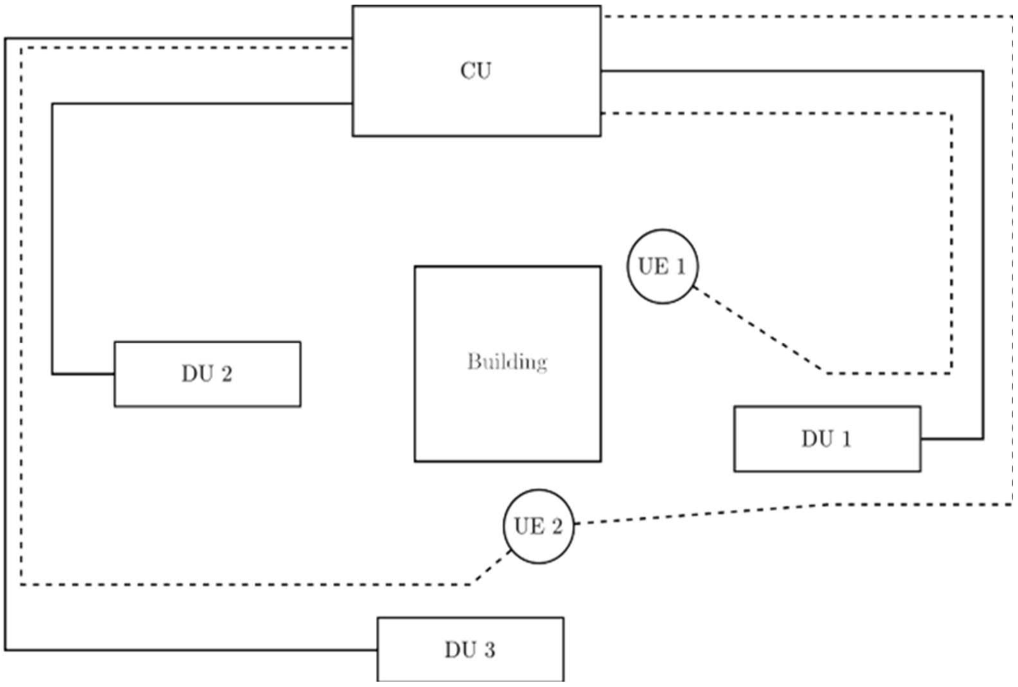


Figure 4.3-1 Cell-free network with one CU controlling three DUs (solid lines). UEs maintain state (dashed lines) with the CU regardless of which DU handles their RF transmission and reception. The selection of DU depends on the UEs' positions as they move around and into the building.

Cell-free networks provide advantages beyond signalling overhead reduction. Multiple DUs can be used to transmit the same data to UEs in an effort to improve reliability. By utilising positioning data, the CU can attempt to provide channel condition diversity through spatially separated beams. Another benefit of cell-free architectures lies in the increased flexibility of load-balancing. Instead of always using SNR as the metric to assign UEs to cells, the CU can consider more factors. For example, if a UE uses edge-computing resources present at only one DU, it may be more efficient to utilise a link with lower – but sufficient – SNR directly to the DU with the resource rather than having to relay the data to its destination.

4.4 Drone base stations

In order to provide coverage in emergency situations, spiral algorithm seems to be a good approach for deploying a swarm of drones to cover certain areas of necessity [LZZ+17]. Placement of the UAV-BS in the three-dimension space plays a significant role in their performance and quality of the service offered to the user equipment on the ground, i.e. GUEs. In contrast to the terrestrial links, where the location of the ground BS is fixed and the path losses are depending on the location of the ground user, properties of the air-to-ground channel is a function of both the G-UE and UAV-BS locations [YFZ+19].

An optimal placement may provide higher coverage with a smaller number of UAV-BSs, lower interference level, extended battery life for the UAVs, or a combination of these [LCW19]. The fundamental of the placement algorithm is to choose the optimal latitude for the different drones, which may differ according to the users they need to give coverage [AKL14].

This placement should be performed by considering the environment that affects the LOS probability [AKL14] and consider certain key performance indicators such as to guarantee coverage to all the terrestrial UE and minimizing deployment cost. Similarly, since these drones can cover a situation like a post-disaster scenario, the latency and the speed of deployment is another feature that needs to be optimized [ZLS+19].

An example of the algorithm can be shown in Figure 4.4-1. Following this, in order to improve the robustness, these algorithms need to work in case of failure such as, for any reason, any (or some) of the drones stop functioning properly. In addition, to include some resources limitations. For instance, the number of drones available to be deployed.

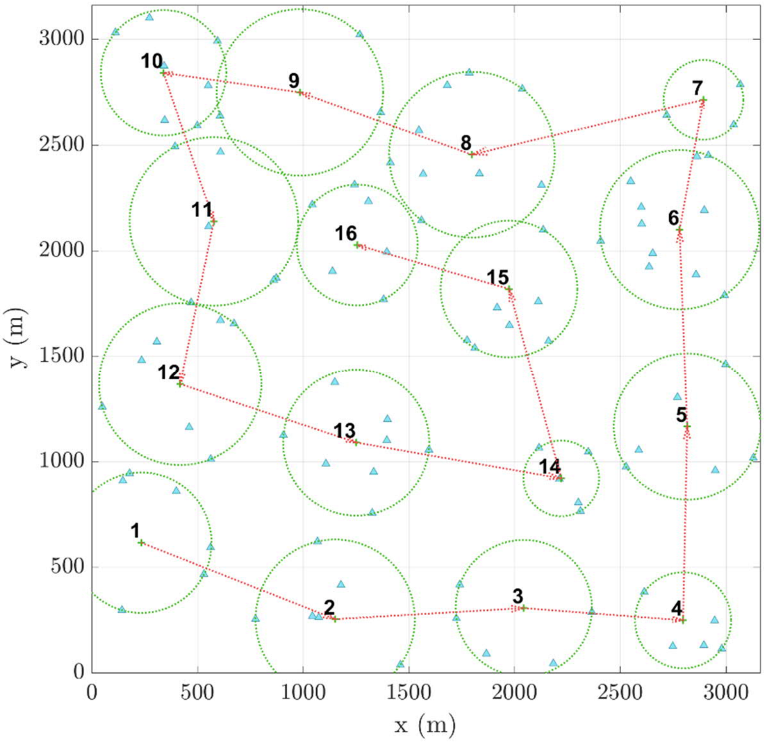


Figure 4.4-1: An example of a statoc spiral algorithm deployment.

4.5 Moving terrestrial base stations

In 3GPP Release 13, [3GPP-22346] is introduced to define IOPS. In overview, it says that “Ensuring the continued ability of Public Safety users to communicate within mission critical situations is of the utmost importance. The Isolated E-UTRAN mode of operation provides the ability to maintain a level of communications for Public Safety users, via an eNB (or set of connected eNBs), following the loss of backhaul communications. The Isolated E-UTRAN mode of operation also provides the ability to create a serving radio access network without backhaul communications, from a deployment of one or more standalone Nomadic eNBs (NeNBs)."

Moving terrestrial base station is a kind of Nomadic eNB which serves both terrestrial and drone mounted user equipment. Moving terrestrial base station can include application servers as well, e.g. call session control function (CSCF) and mission critical push-to-talk (MCPTT) server for VoLTE/VoNR and MCPTT service, respectively. AI-based vision analysis servers can be co-located with moving terrestrial base station to minimize the transmission delay and to provide the service without backhaul.

An Isolated E-UTRAN is characterized by having no, or a limited, backhaul connection as depicted in Table 4.5-1.

Table 4.5-1: Backhaul options for isolated E-UTRAN.

IOPS Scenario	Signalling backhaul status	User Data backhaul status	Comment
No backhaul	Absent	Absent	Fully Isolated E-UTRAN operation using local routing of UE-UE data traffic and possible support for access to the public internet via a local gateway
Signalling only backhaul	Limited	Absent	User data traffic offload at the E-UTRAN using local routing of UE-UE data traffic and possible support for access to the public internet via a local gateway
Limited backhaul	Limited	Limited	Selective user data traffic offload at the E-UTRAN using local routing of UE-UE data traffic and possible support for access to the public internet via a local gateway
Normal backhaul	Normal	Normal	Normal EPC connected operation

No backhaul and normal backhaul are the most commonly used scenarios and moving terrestrial base station supports both. When it detects the loss of backhaul it will initiate Isolated E-UTRAN operation, and then when it detects the recovery of backhaul, it will switch to normal EPC connected operation.

In PriMO-5G demonstrations, moving terrestrial base stations are based on 4G (LTE) technologies partially because 5G gNB is yet working in NSA mode, and partially to provide the service for legacy LTE UEs. Instead 5G NR is used to provide wireless backhaul connection for moving terrestrial base stations.

4.6 Lens-based mmWave Communication

To keep the connection between the fast-moving objects and gNB, stable and accurate beam-tracking must be supported. Also, accurate beamforming and beam tracking are essential as narrow beams are formed due to characteristics of the mmWave band. This section introduces the hybrid beamforming system with a lens antenna so that the fast beam switching method show lower complexity compared to the existing method using phased array. Also, curved lens antenna is considered, and hybrid beamforming is performed. When we use a wide frequency band, conventional analog beamforming is significantly affected by the beam-squint phenomenon. In Figure 4.6-1, a phased array antenna, the beam-squint means that the transmission angle and direction of the analog beamforming of the antenna are changed depending on the operating frequency [I95]. A hybrid beamforming system combining a lens antenna has been proposed to reduce the system cost, but the beam-squint problem in ultra-wideband systems has not been considered extensively. Therefore, in the case of wideband mmWave communication, it is essential to analyze the beam-squint of the RF lens and compare it with the phased array antenna system in the wideband. Then, the actual lens antenna should be fabricated, and a link-level experiment should be demonstrated to show the results and verify the analysis.

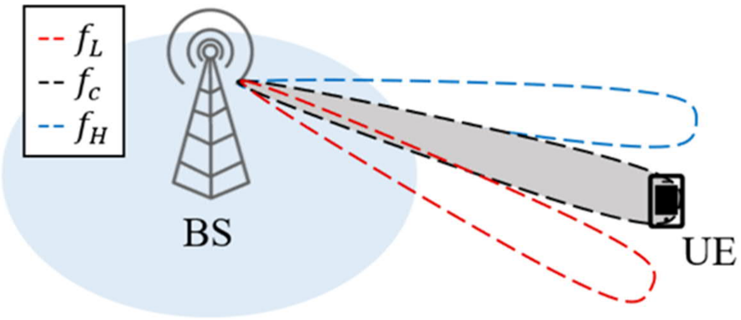


Figure 4.6-1: Distortion of beam direction with wideband frequency at BS with phased array.

4.7 QoS framework in RAN

The overall QoS framework for 5G networks is shown below. It consists of two parts, a 5GC part and NG RAN part, see Figure 4.7-1 below.

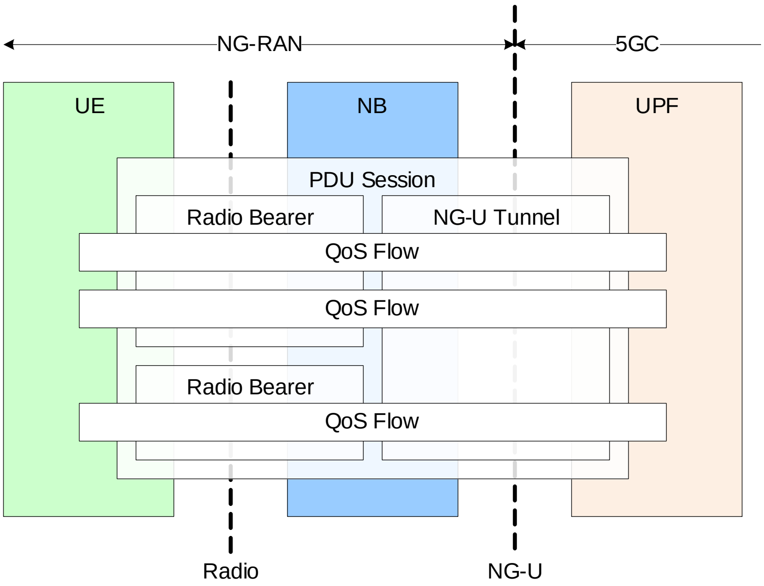


Figure 4.7-1: QoS Framework. [3GPP-38300]

Key elements of the QoS model is that PDU sessions are created in the mobile packet core where a PDU session maps to only one slice: single network slice selection assistance information (S-NSSAI). This means that a PDU session provides data transmission within a slice. The mobile packet core also defines QoS flows which contain packets with the same quality indicators (5QI). The QoS profile of a QoS Flow is sent to the (R)AN from the SMF via AMF.

NG-RAN and 5GC ensure QoS by mapping packets to appropriate QoS flows and data radio bearers (DRBs). Hence there is a 2-step mapping of IP-flows to QoS flows (NAS) and from QoS flows to DRBs (Access Stratum).

At NAS level, a QoS flow is characterised by a QoS profile provided by 5GC to NG-RAN and QoS rule(s) provided by 5GC to the UE. The QoS profile is used by NG-RAN to determine the treatment on the radio interface while the QoS rules dictates the mapping between uplink User Plane traffic and QoS flows to the UE. The QoS profile of a QoS flow contains QoS parameters which at a minimum contains A 5G QoS identifier (5QI) and an allocation and retention priority (ARP). The ARP priority level defines the relative importance of a resource request. This allows deciding whether a new QoS Flow may be accepted or needs to be rejected in the case of resource limitations.

At Access Stratum level, the DRB defines the packet treatment on the radio interface (Uu). A DRB serves packets with the same packet forwarding treatment. The QoS flow to DRB mapping by NG-RAN is based on QFI and the associated QoS profiles (i.e. QoS parameters and QoS characteristics). Separate DRBs may be established for QoS flows requiring different packet forwarding treatment, or several QoS Flows belonging to the same PDU session can be multiplexed in the same DRB. The RAN decides on which DRBs to be established.

At user plane the following access stratum stack is defined as in Figure 4.7-2:

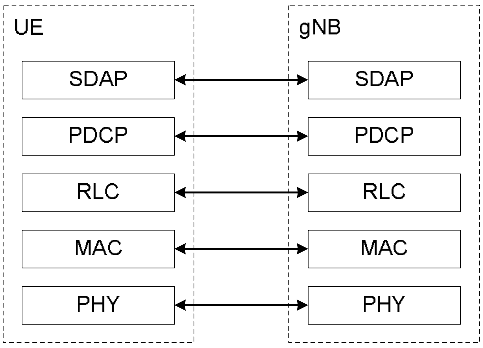


Figure 4.7-2: User Plane Protocol Stack. [3GPP-38300]

In 5G the service data adaptation protocol (SDAP) layer is introduced. The SDAP sublayer maps QoS flows to DRBs. One or more QoS flows may be mapped onto one DRB. Below follows Figure 4.7-3 on functional view for SDAP in down link and in uplink traffic:

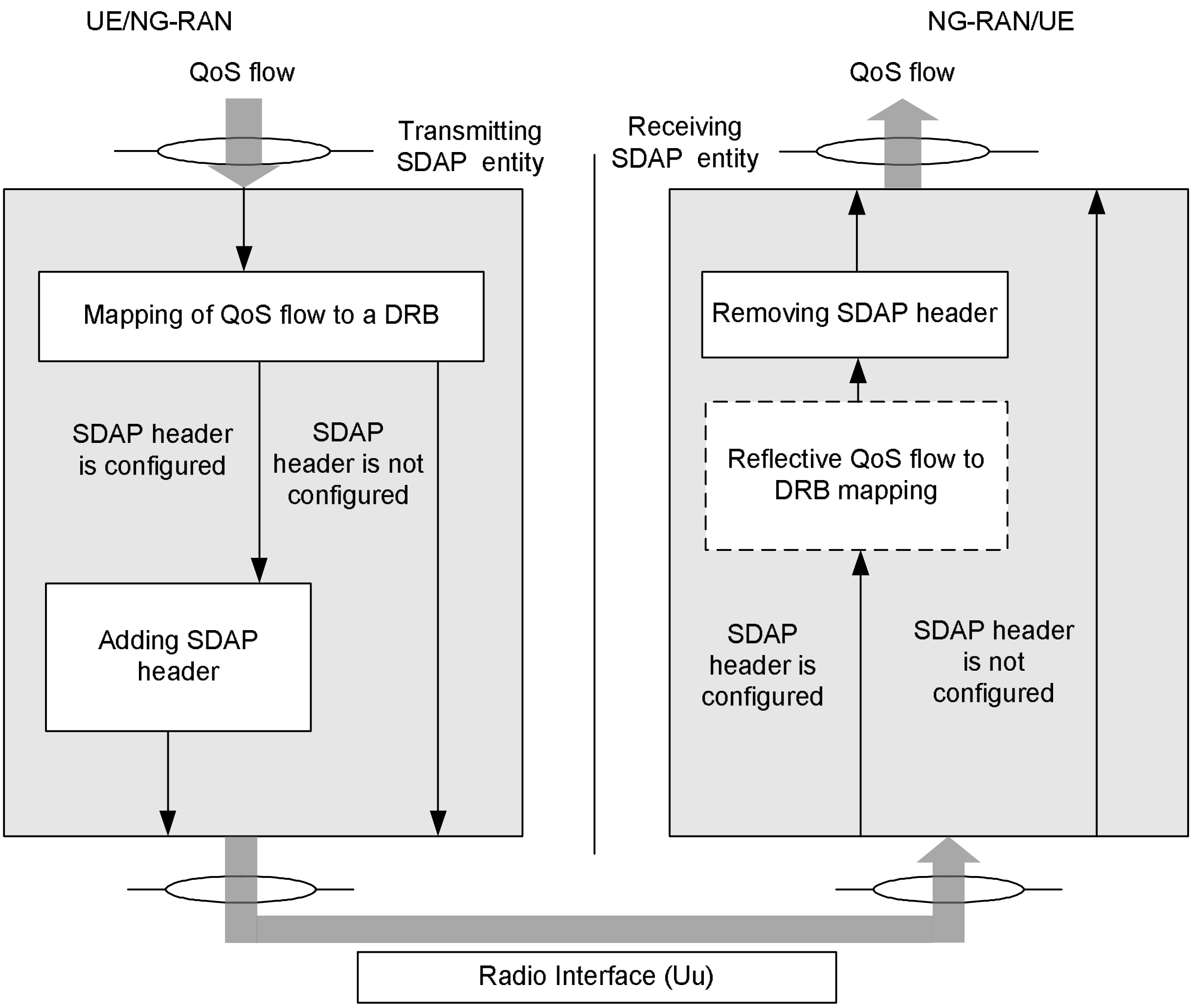

UE/NG-RAN
NG-RAN/UE
QoS flow
QoS flow
Transmitting SDAP entity
Receiving SDAP entity
Mapping of QoS flow to a DRB
SDAP header is configured
SDAP header is not configured
Adding SDAP header
Removing SDAP header
Reflective QoS flow to DRB mapping
SDAP header is configured
SDAP header is not configured
Radio Interface (Uu)

Figure 4.7-3: SDAP layer, functional view. [3GPP-37324]

In the downlink to the UE from NG-RAN a mapping of a QoS flow to a DRB is done, either with a SDAP header included or not.

In the uplink to the NG RAN from the UE there are two alternatives available depending on if reflective or explicit signalling is used. In reflective signalling the UE monitors the QFI(s) of the downlink packets and applies the same mapping in the uplink (QFI can be read from SDAP header); that is, for a DRB. In explicit signalling, the QoS flow to DRB mapping rules are explicitly signalled by RRC. The RAN decides if reflective or explicit signalling should be used.

4.8 RAN slicing

3GPP is very flexible in terms of the resource allocation strategies for intra-slice and inter-slice scenarios [3GPP-38801]. The earlier research focuses on static or nearly static resource allocations for different slices based on the well-defined SLAs and the statistical analysis. Although static RAN slicing is the simplest and the most convenient model to achieve QoS expectations, it cannot provide the flexibility or efficiency to support evolution to the next generation of wireless communications. On the other hand, as stated by 3GPP (RP-193254), the RAN slicing and resource allocation mechanisms should be considered as a way to manage the dynamic requirements of the over-the-top service providers. As such, the dynamic network conditions, as well as the evolving business needs of over-the-top service providers, have to be considered in the resource management procedure.

In cloud RAN, each slice might be configured with different functional split offering diverse solutions with distinctive QOS requirements such as latency, jitter, and data rate, which increases the deployment flexibility of RAN slicing. In flexible functional split, functions are flexibly placed between CU and DU based on the needs of a service providing different level of centralisation. For example, a slice with a lower layer split would typically be suitable for services requiring high computation or high mobility benefiting from pooling and high processing capabilities provided by CU. While slices with higher layer split, where real-time functions are located in DU, would satisfy the requirements of low latency services. Resources of CU/DU might be dedicated to a slice or common to several slices. In the formal whereby the resources are static, the QoS of the slice is guaranteed but might use resources inefficiently. In the latter, resources are allocated flexibly improving multiplexing gain at the cost of complex implementation.

Dynamic RAN slicing, i.e. dynamically determining the resource allocations to different slices, is one of the key aspects in 5G network management. As outlined in [3GPP-38801], the dynamic RAN slicing strategies have to guarantee the management aspects such as resource isolation, fairness, and SLA guarantees. The multiplexing gain that can be obtained through dynamic resource allocations causes a high level of interrelatedness among different slices. This interrelatedness complicates the resource isolation, i.e. minimizing the impact of one slice on the other one. The lack of resource isolation can result in unfair resource allocations, e.g. domination of the network resources by a slice or a service provider, and result in SLA violations.

In the focus of PriMO-5G, the considered emergency communications have a dynamic nature. Specifically, it is hard to predict the beginning or the duration of the communications. The conventional static RAN slicing approach requires the network resources to be continuously reserved for the possibility of an emergency. However, these reserved resources are unused for most of the time due to the occurrence frequency of such emergency communications. Dynamic allocation of the resources to slices can provide the required flexibility, efficiency, and scalability. In the detection of emergency communication, the resources can be dynamically allocated to the respective slices.

4.9 RAN Enhancements for URLLC

URLLC is one of the critical features of the 5G technology. Power distribution, smart factory and transport industry are expected to have a tremendous impact on the 4th Industrial Revolution as a representative URLLC application field. According to [3GPP-38824] in Release 16, in order to activate URLLC technology, E2E latency of less than 5 ms and reliability performance of less than block error rate (BLER) = 10^{-6} must be achieved.

Extending new use cases beyond initial mobile broadband use cases is a main component of not only the wireless access evolution, but also the evolution of LTE and new NR radio-access technology. This includes mMTC use cases characterized by requirements on very low device cost and very long device battery life, often also associated with a requirement on very wide-area coverage.

It also includes critical MTC applications, such as industrial process automation and manufacturing, energy distribution and intelligent transport systems. These applications are typically associated with requirements for very high communication reliability and the possibility for very low latency. In the standardization community, both within 3GPP and ITU, critical MTC applications are often referred to as URLLC.

Release-15 provided basic URLLC functionality:

1. Lower latency by supporting:
   A. Higher subcarrier spacing, with shorter transmission durations.
   B. Mini-slots with fewer number of symbols.
   C. Frequent PDCCH monitoring reducing the latency of the layer-1 control information.
   D. Configured-grant, which allows the UE to autonomously transmit uplink data without having

to send a scheduling request and wait for the uplink grant.
   E. Downlink pre-emption.

2. Higher reliability by supporting:
   A. Multi-slot repetition.
   B. Low spectral efficiency MCS/CQI tables.
   C. PDCP duplication.

Release-16 further enhances the NR support for URLLC services by enabling latency in the range of 0.5 to 1 ms and improved reliability with a target error rate of 10^{-6}. This allows the support of new use cases, such as factory automation and transport industry as well as improving the performance of Release-15 use cases such as AR/VR and gaming.

To achieve this, the Release-16 work item Physical Layer Enhancements for NR URLLC focuses on the following areas of improvements [3GPP-191584]:

1. PDCCH enhancements focusing on:
   A. Configurable field sizes for downlink control information for improved reliability.
   B. Increased PDCCH monitoring capability to minimize scheduling block/delay.

2. UCI enhancements focusing on:
   A. Support of multiple HARQ-ACK feedback occasions per slot to reduce latency.
   B. Construction of multiple HARQ-ACK codebooks with different treatment intended for different services.

3. PUSCH enhancements by supporting cross-slot-boundary scheduling for both dynamic PUSCH grant and configured PUSCH grant.

4. Scheduling and HARQ enhancements including the support of:
   A. Out-of-order HARQ-ACK allowing the HARQ-ACK of a more recently scheduled low-latency transmission to be transmitted before the HARQ-ACK of regular transmission that was scheduled earlier.
   B. Out-of-order PUSCH scheduling allowing a low-latency PUSCH that has been scheduled after a normal PUSCH to be transmitted before the normal PUSCH.
   C. Overlapping dynamic PDSCHs.

5. Inter-UE prioritization and multiplexing focusing on:
   A. UL preemption by allowing the gNB to interrupt data transmission from one user to accommodate higher-priority data from another user.
   B. Enhanced UL power control to enable power boosting for URLLC UL transmissions overlapping with some eMBB transmission.

6. Configured-grant enhancements by supporting multiple active configurations, to accommodate different service flows and to reduce the alignment time for URLLC UL transmissions.

Physical layer feedback enhancements for meeting URLLC requirements are as follows [3GPP-193233]:

1. UE feedback enhancement for HARQ-ACK (RAN1)
2. CSI feedback enhancements to allow for more accurate MCS selection (RAN2)

Potential enhancements to ensure Release-16 feature compatibility with unlicensed band URLLC/IIoT operation in controlled environment (RAN1, RAN2)

1. Detailed objectives to be clarified at RAN#87 based on essential issues to be identified in RAN#87 (if any)

Intra-UE multiplexing/prioritization of traffic with different priority based on work done in Release-16

(RAN1)

1. Specify multiplexing behaviour among HARQ-ACK/SR/CSI and PUSCH for traffic with different priorities

Enhancements for support of time synchronization (RAN2)

1. RAN impacts of SA2 work on uplink time synchronization for TSN, if any (RAN2)
2. Propagation delay compensation enhancements (including mobility issues) (RAN1, RAN2, RAN3, RAN4)

RAN enhancements based on new QoS related parameters if any, e.g. survival time, decided from SA2 (RAN2, RAN3).

Table 4.9-1: Representative use cases for Release-16 NR URLLC evaluation. [3GPP-38824]

Use case	Reliability (%)	Latency	Data packet size and traffic model	Description
Power distribution	99.9999	5 ms (end to end latency) Note: 2-3 ms air interface latency	DL & UL: 100 bytes ftp model 3 with arrival interval 100 ms	Power distribution grid fault and outage management (TR 22.804:5.6.4)
	99.999	15 ms (end to end latency) Note: 6-7 ms air interface latency	DL & UL: 250 bytes Periodic and deterministic with arrival interval 0.833 ms Random offset between UEs	Differential protection (TR 22.804:5.6.6)
Factory automation	99.9999	2 ms (end to end latency) Note: 1 ms air interface latency	DL & UL: 32 bytes Periodic deterministic traffic model with data arrival interval 2 ms	Motion control
Rel-15 enabled use case (e.g. AR/VR)	99.999	1 ms (air interface delay) for 32 bytes 1 ms and 4 ms (air interface delay) for 200 bytes	DL & UL: 32 and 200 bytes FTP model 3 or periodic with different arrival rates	
	99.9	7 ms (air interface delay)	DL & UL: 4096 and 10 K bytes FTP model 3 or periodic with different arrival rates	
Transport Industry	99.999	5 ms (end to end latency) Note: 3 ms air interface latency	UL: 2.5 Mpbs; Packet size 5220 bytes DL: 1Mbps; Packet size 2083 bytes Note: Data arrival rate 60 packets per second for periodic traffic model	Remote driving (TS 22.186: 5.5)
	99.999	10 ms (end to end latency) Note: 7ms air interface latency	UL&DL: 1.1 Mbps; Packet size 1370 bytes Note: Data arrival rate 100 packets per second for periodic traffic model	Intelligent transport system (ITS) (TS 23.501, TS 22.261)

According to [3GPP-38824], the identified use cases for Release-16 URLLC include factory automation, transport industry, electrical power distribution and Release-15 enabled use case. Evaluations are performed for the representative use cases for the identified use cases. Table 4.9-1 shows the representative use cases for Release-16 NR URLLC evaluation.

In order to improve the performance of RAN technology, URLLC is an important factor for improving RAN technology performance. Two main requirements, latency and reliability must be satisfied with values beyond the performance requirements of Table 4.9-1 in the 3GPP standard document.

First, in order to achieve high reliability, there is a method of lowering or re-transmitting a code rate of a channel code.

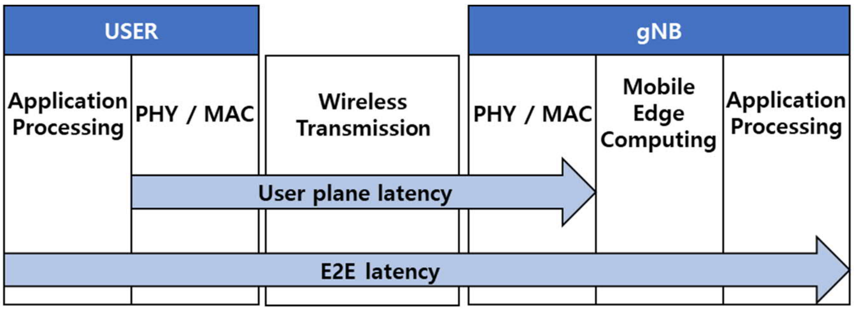


Figure 4.9-1: Indicator of latency. [PKP+19]

Second, latency can be divided into two types, User plane latency and E2E latency as shown in Figure 4.9-1.

1. User plane latency refers to the packet transmission time between the UE and the base station in the ACTIVE state. In more detail, it means the time taken to transmit from the TX wireless protocol PHY/MAC layer to the RX wireless protocol PHY/MAC layer through the wireless interface. In order to reduce the latency, using short length of block channel code. In addition, increasing the subcarrier spacing to 30 kHz, 60 kHz and 120 kHz as well as 15 kHz can reduce latency by reducing the transmission time interval (TTI).
2. E2E latency refers to the time from the TX sends application data to the Rx successfully receives the data at the RX. E2E latency consists of wireless transmission delay, queuing delay in MEC or base station, processing delay and re-transmission time. Among them, the MEC method can reduce latency and improve RAN performance. In general, the service time for transmitting and receiving communication data is composed of the sum of the communication time and the calculation time. By using the edge of the calculation process that had to be processed in the existing cloud server, the calculation time can be reduced. As a result, the computation time can be reduced by the edge system, so it reduces the latency.

URLLC L1 improvements (RAN1) for further improved reliability/latency and for other requirements related to the use cases identified:

1. PDCCH enhancements. Study focus on Compact DCI, PDCCH repetition, increased PDCCH monitoring capability
2. UCI enhancements. Study focus on Enhanced HARQ feedback methods (increased number of HARQ transmission possibilities within a slot), CSI feedback enhancements

3. PUSCH enhancements. Study focus on mini-slot level hopping & retransmission/repetition enhancements.
4. Enhancements to scheduling/HARQ/CSI processing timeline (UE and gNB), (for existing TTI durations)

Enhanced multiplexing considering different latency and reliability requirements (RAN1):

1. UL inter UE Tx prioritization/multiplexing

Enhanced UL configured grant (grant free) transmissions, with study focusing on improved configured grant operation, example methods such as explicit HARQ-ACK, ensuring K repetitions and mini-slot repetitions within a slot. (RAN1/RAN2)

5 Core Network Architecture

The main functionality of the core network is to connect UE served by the RAN to data networks. The 5G Core network (5GC) was initially defined by 3GPP in Release-15 [3GPP-23501], which was finalized in early 2019. However, in future releases, development of the 5GC continues.

5.1 Overview of 5G Core Network

The 5GC follows a number of principles that are mainly targeted for reaching higher flexibility and supporting many different use cases, including for example mobile broadband, industrial automation, public safety, etc. The requirement of flexibility, with advancement in virtualization and containerization techniques, and the introduction of microservices, led to the introduction of service-based principles, where network functions provide services to each other. The service-based architecture also provides easy extendibility, so new network functions can be introduced easier into the architecture. A clean control plane/user plane split allows independent scaling of control plane and user plane functions, and also supports flexible deployments in terms of where the user plane can run (this principle was, in fact, already introduced in EPC in Release 14). The architecture allows for different network configurations in different network slices.

The 5GC consists of functionality that can divided into control plane and user plane. The architecture is depicted at the Figure 5.1-1 below:

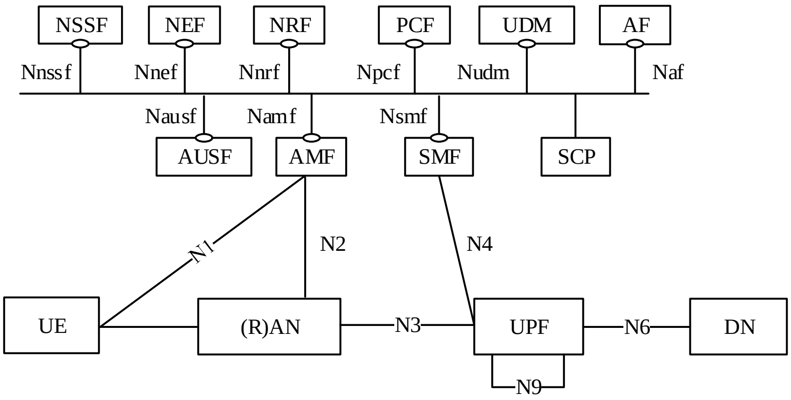


Figure 5.1-1 5G System architecture [3GPP-23501].

The following are the minimal set of mandatory functions that has to be there in any 5GC deployment:

- Access and Mobility Management Function (AMF, control plane): The AMF has signaling connections to the UEs (N1) and to the RAN (N2). Nearly all signaling call flows involves the AMF; it allows the devices to register and authenticate themselves to the network, manages mobility so the devices can move between radio cells, and reachability of the devices that are in idle mode.
- Session Management Function (SMF, control plane): The SMF is responsible for creating, modifying, and releasing user sessions, and allocating IP addresses to the sessions. It is also responsible for selecting UPF(s) for the given sessions. It communicates with UPFs on the (point-to-point) N4 interface, while it uses the service-based architecture to interact the other NFs. All session management related information from the devices reaches the SMF via the AMF.
- UPF (User Plane Function, user plane): The UPF performs forwarding user plane packets to and from the devices. Its functionality is controlled by the SMF. Besides forwarding user plane packets, it may also perform generating traffic usage reports, traffic inspection, executing

policies, QoS marking the packets towards the RAN or towards external networks. It may also buffer user plane packets in case the UE is in idle mode, and triggers paging. The interface between the NG-RAN and the UPF is N3. In a single user session, there could be multiple UPFs chained, and they are connected with the N9 interface. Finally, the N6 interface is towards the Data Network.

- AUSF (Authentication Server Function, control plane): the AUSF's main task is to authenticate the devices.
- UDM (Unified Data Management Function, control plane): the UDM is a front-end for user subscription data stored in UDR.
- UDR (Unified Data Repository, control plane): The UDR is a database storing subscription information, policies, and other data.
- NRF (Network Repository Function, control plane): The NRF keeps track of all services available of all Network Functions. At service registration, the NFs register their services to the NRF and later, at service discovery, the NRF supports the NFs to find the required services. This way, the NFs need to be configured only one or more NRF IP addresses, instead of having configured all the other NF's IP addresses. The NRF keeps an NF profile of the NF instances, that includes information including NF type, address, supported services etc.

There are other important NFs in the 5GC, for example:

- The Network Slice Selection Function (NSSF) is used to assist slice selection.
- The Network Exposure Function (NEF) is mainly responsible for exposure of capabilities and events.
- The Policy Control Function (PCF) governs the network behavior via policy decisions.
- The AF (Application Function) provide a way for applications to interact with the 5GC
- The NWDAF (Network Data Analytics Function) to provide statistics and predictions to other Network Functions

The connectivity model in 5GC is based on PDU Sessions. A PDU Session is an association between a UE and a data network. A UE, after it performed registration to the 5G System, may establish PDU sessions towards data networks, e.g. to the Internet, IMS, or to other specific purpose networks, etc. Different type of PDU sessions can carry IP (IP based PDU Session types – Ipv4), Ethernet (Ethernet PDU Session Type), or any other protocol (Unstructured PDU Session Type). In the case of IP based PDU session types the 5GC (and more specifically, the SMF) is responsible for IP address allocation to the UE (either IPv4 address or IPv6 prefix). The following are some of the main properties that characterize a PDU Session: PDU Session ID, slice Identifier (S-NSSAI), data network name (DNN). These parameters are determined at the establishment of the PDU session and do not change during the lifetime of the session. The PDU Session Establishment is handled by the SMF.

A key concept in 5GC is slicing. While definitions in different parts of the industry may vary, in general, a slice is an E2E logical network to serve a specific purpose or a customer. In 3GPP, a slice consists of a radio network and a core network, and some parts of the network resources can be shared between slices, while some parts are unique to a single slice.

The following subsections will introduce functionalities and advancements of the 5GC that are relevant to the PriMO-5G public safety use cases:

- Slicing to allow different logical networks on the top of a common infrastructure
- QoS framework for supporting differentiation between many different services within a slice

- Core Network enhancements targeting URLLC use cases, mainly to improve reliability by redundancy
- Non-public Networks as a means of deploying networks for special private purpose
- Isolated Operations for Public Safety (IOPS) so communication can continue in case the connectivity to the central network components are lost
- Optimal routing to provide low latency paths for edge scenarios and UE-to-UE communication, while keeping the IP session continuity
- NWDAF to provide analytics (statistics and predictions) to other NFs
- 5G-LAN as a means to provide private group communication

5.2 Core Network Slicing

An introduction of network slicing is defined in GSMA which consists of the following: "From a mobile operator's point of view, a network slice is an independent E2E logical network that runs on a shared physical infrastructure, capable of providing a negotiated service quality. The technology enabling network slicing is transparent to business customers. A network slice could span across multiple parts of the network (e.g. terminal, access network, core network and transport network) and could also be deployed across multiple operators. A network slice comprises dedicated and/or shared resources, e.g. in terms of processing power, storage, and bandwidth and has isolation from the other network slices." [GSMA20]

The Next Generation Mobile Networks (NGMN) alliance also refers to network slice concept as follows [NGMN16]. "A network slice instance may be fully or partly, logically and/or physically, isolated from another network slice instance".

From 3GPP networks point of view, network slicing requires having a logical network that provides specific network capabilities and network characteristics that can be dynamically created. The mobile core will assign a network slice instance to the UE which consists of a set of Network Function instances and the required resources (e.g. compute, storage, and networking resources) from RAN, network, and core functions.

Each slice may serve a service type agreed upon service-level agreement (SLA). A network slice is defined within a Public Land Mobile Network (PLMN) and includes the Core Network Control Plane and User Plane Network Functions as well as the 5G Access Network (AN).

This requires that not only the RAN and mobile backhaul infrastructure supports isolation of resources but also additional instances of network functions such as AMF, SMF, UPF might have to be created and allocated for the different slices.

A network slice is identified by S-NSSAI (Single Network Slice Selection Assistance Information) and has two parameters: SST (Slice/Service Type) and optionally, an SD (Slice Differentiator). The network will assist the UE to connect to the required slice: first the RAN will select an AMF based on the requested S-NSSAI values by the UE. This AMF will either decide to serve the UE or perform a new slice selection. For this, it may use the services of NSSF (Network Slice Selection Function).

In Figure 5.2-1, an example of three slices are shown. Slice 1 has a dedicated AMF, SMF, and UPF, while Slice 2 and Slice 3 share an AMF, while having dedicated SMFs and UPFs. A single UE may connect to multiple slices, e.g. UE2 may be connected to Slice2 and Slice3, and served by AMF2.

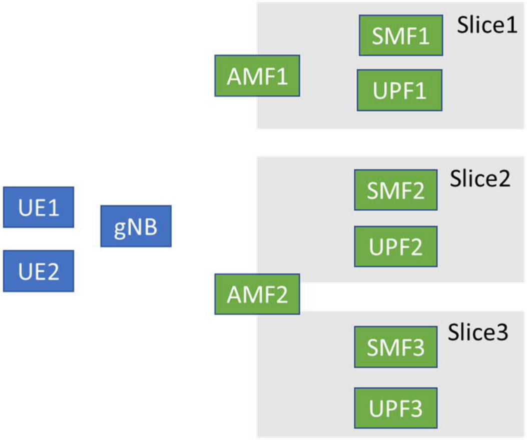


Figure 5.2-1: Simplified view on slicing.

The allocation of core network functions for different slices requires an overlay orchestrator as the NFV MANO (network functions virtualization management and network orchestration) defined in ETSI [ETSI ISG NFV]. The ETSI MANO provides standard mechanism to manage resources such as computing, network, and storage in a data center where multiple virtual machine (VM) can be allocated to run different core instances allocated to the network slices. Within each slice the core network will manage the RAN and backhaul resources to fulfil the SLA assigned to each slice.

5.3 5G QoS Framework

The 5G QoS Framework is involved in the operation of both the control plane and the user plane. The control plane involves key components in the service-based architecture (e.g. AMF, SMF, PCF, NSSF) with control traffic both down to the user plane, handling of QoS between UPF and AN both also QoS control traffic between AMF and UE.

The overall QoS framework is connected to the slices in such a way that a PDU session belongs to only one Single Network Slice Selection Assistance Information (S-NSSAI) instance per PLMN.

5.3.1 QoS framework in the User Plane between UPF and AN

Below follows Figure 5.3-1 that shows the principles for classification and user plane marking of QoS flows.

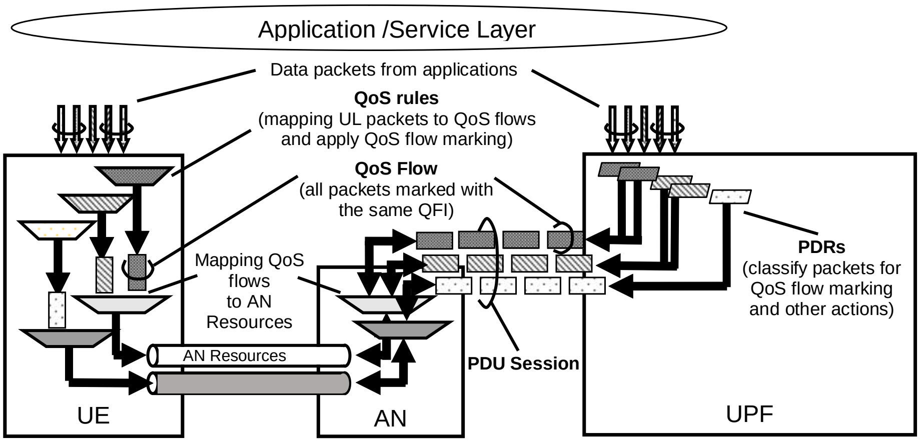


Figure 5.3-1: The principle for classification and User Plane marking for QoS Flows and mapping to AN Resources [3GPP-23501].

It starts with that a PDU session is established with corresponding slice (S-NSSAI) if a slice is used, towards a data network identified with Data Network Name (DNN). This means that the PDU session enables data transmission in a network slice instance, towards the given Data Network. A single PDU session may contain one or more QoS Flows. Arriving packets from the data network will first meet a packet detection rule (PDR) that contains information required to classify a packet arriving at the UP function. The SMF is responsible for instructing the UPF about how to detect user data traffic belonging to a PDR. The other parameters provided within a PDR describe how the UP function shall treat a packet that matches the detection information.

Based on the PDRs, QoS flows can be differentiated which contains packets with the same Quality Flow Indicators (QFI). The 5G QoS model supports both QoS Flows that require guaranteed flow bit rate (GBR QoS Flows) and QoS Flows that do not require guaranteed flow bit rate (Non-GBR QoS Flows). The 5G QoS model also supports Reflective QoS.

The QoS Flow is the finest granularity of QoS differentiation in the PDU Session. A QoS Flow ID (QFI) is used to identify a QoS Flow in the 5G System. User Plane traffic with the same QFI within a PDU Session receives the same traffic forwarding treatment

Depending on type of QoS flow different parameters can be sent, including 5G QoS Identifier (5QI) mapped to 5G QoS characteristics (i.e. access node-specific parameters that control QoS forwarding treatment for the QoS Flow). The parameters set for 5G QoS characteristics are:

- Resource Type (guaranteed bit rate (GBR), Delay critical GBR or Non-GBR).
- Priority level.
- Packet delay budget.
- Packet error rate.
- Averaging window (for GBR and Delay-critical GBR resource type only).
- Maximum Data Burst Volume (for Delay-critical GBR resource type only).

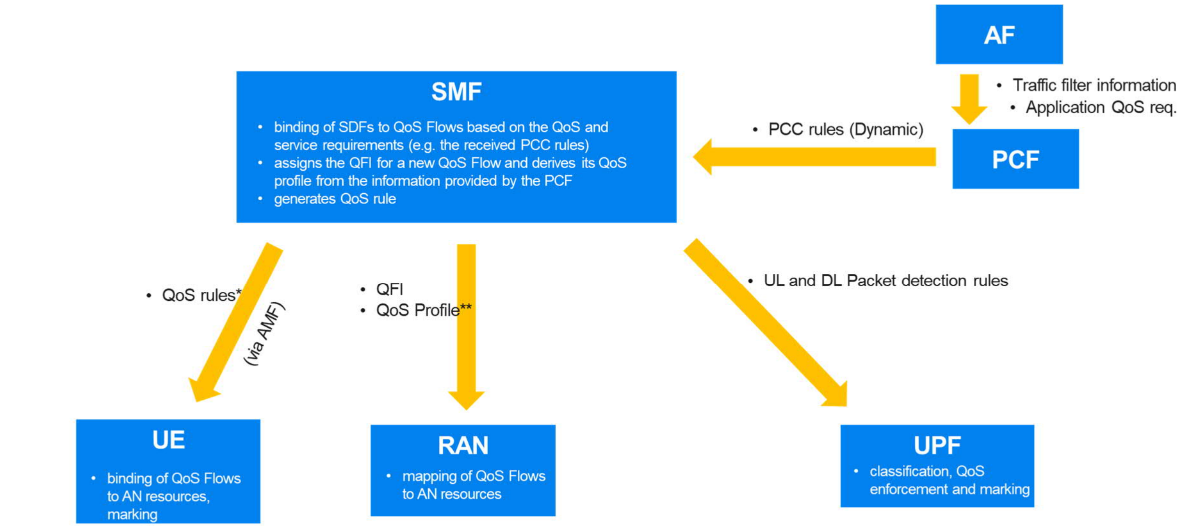
The 5QI value is just a scalar mapped to a certain 5G QoS characteristics case. There are also 5QI values for non-standardized 5G QoS areas. There is also an opportunity to tweak parameters on a standardized 5QI QoS area towards the 3GPP system.

5.3.2 QoS framework operations in the control plane

PDU Sessions are established (upon UE request), modified (upon UE and 5GC request) and released (upon UE and 5GC request) using NAS session management signalling exchanged over N1 between the UE, (AMF: relaying SM info) and the SMF. By definition, a single PDU session always belongs to a single slice and the CN interacts with the RAN and the UE to implement E2E quality-of-service for the PDU session.

A QoS Flow is controlled by the SMF and may be preconfigured, or established via the PDU Session Establishment procedure, or the PDU Session Modification procedure. The QoS profile of a QoS Flow is sent to the (R)AN from the SMF via AMF.

The entities participating and the control signalling are shown in Figure 5.3-2 below:


AF
• Traffic filter information
• Application QoS req.
PCF
• PCC rules (Dynamic)
SMF
• binding of SDFs to QoS Flows based on the QoS and service requirements (e.g. the received PCC rules)
• assigns the QFI for a new QoS Flow and derives its QoS profile from the information provided by the PCF
• generates QoS rule
• QoS rules*
(via AMF)
• QFI
• QoS Profile**
• UL and DL Packet detection rules
UE
• binding of QoS Flows to AN resources, marking
RAN
• mapping of QoS Flows to AN resources
UPF
• classification, QoS enforcement and marking

• *May be also pre-configured or derived by the UE using reflective QoS • **May be also pre-configured in the RAN

Figure 5.3-2: Entities and signalling in QoS Framework (simplified).

Signalling of QoS information to the UPF: The SMF performs the binding of service data flows (SDFs) to QoS Flows based on the QoS and service requirements. The SMF assigns the QFI for a new QoS Flow and derives its QoS profile, corresponding UPF instructions and QoS Rule(s) from the policy and charging control (PCC) rules and other information provided by the PCF.

Between SMF and UPF there is N4 interface. The N4 Session Establishment procedure as well as the N4 Session Modification procedure provide the control parameters to the UPF, the N4 Session Release procedure removes all control parameters related to an N4 session, and the N4 Session Level Reporting procedure informs the SMF about events related to the PDU Session that are detected by the UPF. Notable parameters in the N4 interface are the PDR rules and the QoS Enforcement Rules (QER), that contain information related to QoS enforcement of traffic identified by PDR(s) that is sent from the SMF to the UPF(bit rate limitation, packet marking, QoS Flow ID).

The SMF provides the following information to the UPF enabling classification, bandwidth enforcement and marking of User Plane traffic: a DL packet detection rule (PDR) containing the DL part of the SDF template and an UL PDR containing the UL part of the SDF template, the PDR precedence value, QoS related information (e.g. MBR for an SDF, Guaranteed Flow Bit Rate (GFBR) and Maximum Flow Bit Rate (MFBR) for a GBR QoS Flow) and the corresponding packet marking information (e.g. the QFI, the transport level packet marking value). The QoS rule precedence value and the PDR precedence value determine the order in which a QoS rule or a PDR, respectively, shall be evaluated.

Signalling of QoS information from AF: The AF may request QoS towards the PCF to ensure better service experience. This is usually preceded by application layer signalling from the UE to the AF.

Signalling of QoS information to the NG-RAN: If explicitly signalled, the QoS profile of a QoS Flow is sent to the (R)AN from the SMF via AMF. In other cases, the QoS profile can also be pre-configured in the (R)AN. The QoS profile contains per-flow QoS information, including but not limited to, 5QI, Allocation and Retention Priority (ARP), Reflective QoS Attribute (RQA), and Flow Bit Rates (Guaranteed Flow Bitrate and Maximum Flow Bitrate). If the 5QI value is not from the standardized values, then the 5G QoS characteristics are signalled, too (Resource Type, Priority Level, Packet Delay Budget, Packet Error Rate, Averaging Window, Maximum Data Burst Volume).

Signalling of QoS information to the UE: The UE performs the classification and marking of UL user plane traffic, i.e. the association of UL traffic to QoS Flows, based on QoS rules. These QoS rules may be explicitly provided to the UE (i.e. explicitly signalled QoS rules using the PDU Session Establishment/Modification procedure), pre-configured in the UE or implicitly derived by the UE by applying Reflective QoS (Replicate downlink QoS setting in uplink QoS settings). A QoS rule contains the QFI of the associated QoS Flow, a Packet Filter Set and a precedence value. An explicitly signalled QoS rule contains a QoS rule identifier which is unique within the PDU Session and is generated by SMF. When the UE informs the network about the number of supported Packet Filters for signalled QoS rules for the PDU Session (during PDU Session Establishment procedure/PDU Session Modification procedure), the SMF shall ensure that the sum of the Packet Filters used by all signalled QoS rules for a PDU Session does not exceed the number indicated by the UE. A default QoS rule is required to be sent to the UE for every PDU Session establishment and it is associated with a QoS Flow. For IP type PDU Session or Ethernet type PDU Session, the default QoS rule is the only QoS rule of a PDU Session which may contain a Packet Filter Set that allows all UL packets, and in this case, the highest precedence value shall be used for the QoS rule.

Influencing E2E QoS with advanced session management functionality: In general, there are ways specified to establish efficient user plane paths. This means, e.g. to support local breakout, and to ensure that the communication follows low-latency paths. This is very important when there is a need to communicate to local (edge) servers close to the UE. The SMF can provide session breakout by introducing extra User Plane Functions on the paths. The implication is that lower delays can be provided E2E, provided that the application server is also placed closer to the UE. These mechanisms (local breakout and AF influence on traffic routing) will be introduced in Section 6.3.2 (Release-17 3GPP SA2 work on enhancement of support for Edge computing).

5.4 Core Network Enhancements for URLLC

In Release-16, 3GPP studied architectural enhancements for supporting URLLC services in 5G System. The objectives were meeting the URLLC requirements on latency, jitter and reliability, as defined in [3GPP-22261]:

- Low latency, e.g. less than 5 ms between UE and DN
- High reliability, e.g. equal or higher than 99.999% between UE and DN
- Low jitter, e.g. less than 100 µs between UE and DN

3GPP SA2 has defined multiple enhancements/supporting technologies for URLLC. These include the following (focusing on Core Network):

- Definition of standardized 5QIs for URLLC traffic: for providing low latency
- Standardized SST (slice type) for URLLC, to facilitate provisioning of a network E2E with very specific requirements
- The possibility to divide the Packet delay budget (PDB) to Core Network and RAN part, and the SMF can provide specific CN PDB based on the 5QI
- To address high reliability, 3GPP worked on solutions supporting E2E redundancy, or providing redundancy on specific transport segments
    - Multiple user plane paths with dual connectivity
    - Multiple user plane paths with multiple UEs per device
    - Redundant transmission on N3 and N9 interfaces

The redundancy solutions are described further in this subsection.

5.4.1 Multiple user plane paths with dual connectivity

In this solution, the terminal sets up two Protocol Data Unit (PDU) sessions towards the same DN, and the network attempts to make the paths of the two PDU sessions independent whenever it is possible. There are two PDU sessions involved in the solution: the first spans from the UE via gNB1 to UPF1, acting as the PDU Session Anchor, while the second spans from the UE via gNB2 to UPF2, acting as the PDU Session Anchor. The independent paths may continue beyond the 3GPP network. Redundancy Handling Functions (RHFs) deployed in Host A (the device) and Host B (in the network), or optionally in another point in the network side. These functions are required to make use of the independent paths. One example for the RHF is the IEEE TSN Frame Replication and Elimination for Reliability (FRER).

The solution architecture is depicted in Figure 5.4-1. It is based on Dual Connectivity. In this case, a single UE has both connectivity to a Master gNB (MgNB) and a Secondary gNB (SgNB) (e.g. gNB1 is assuming a role of a Master gNB and gNB2 is assuming a role of a Secondary gNB). The two PDU sessions are spanning via MgNB and UPF1, and SgNB and UPF2, respectively. UPF1 and UPF2 are controlled by SMF1 and SMF2, but they can coincide, and both UPFs provide connectivity to the same DN.

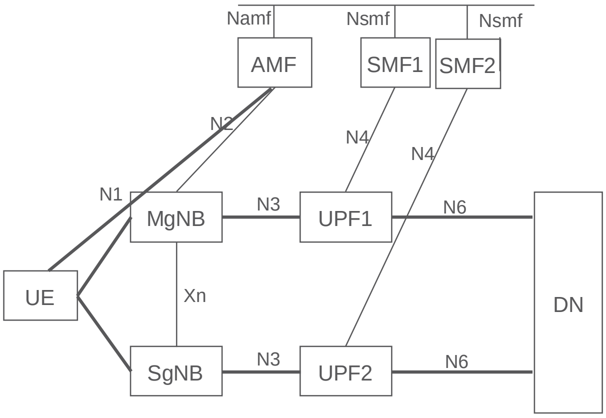


Figure 5.4-1 Solution architecture for multiple user plane paths with dual connectivity [3GPP-23501].

5.4.2 Multiple user plane paths with multiple UEs per device

This solution relies on multiple UEs in the same device for achieving redundancy. The terminal device will set up multiple PDU sessions over the 5G System (5GS), and the network will attempt to make the path of the redundant multiple PDU sessions independent, whenever possible. The redundancy mechanism is defined outside 3GPP, and one example is IEEE TSN's FRER mechanism deployed in the host within the device and on the network side.

In this solution, the device is equipped with two UEs, UE1 and UE2. The first PDU session spans from UE1 via gNB1 to UPF1, while the second PDU session spans from UE2 via gNB2 to UPF2. The independent paths may span beyond the 3GPP network (i.e. the merging of the two paths will happen somewhere in the Data Network, possibly even in the end-host). The architecture view can be seen in Figure 5.4-2 below. The two UPFs connect to the same Data Network.

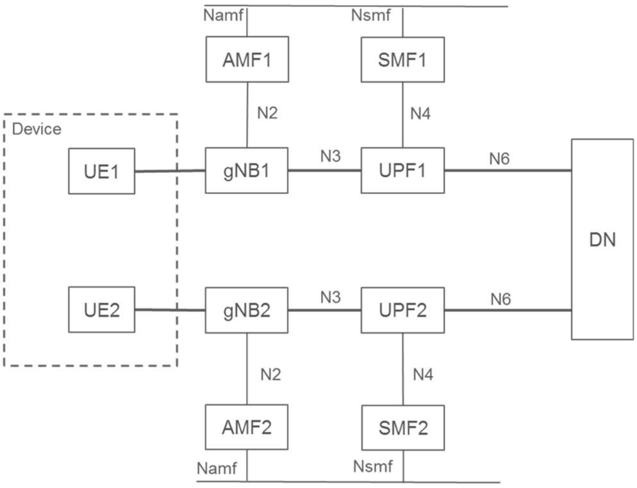


Figure 5.4-2. Solution architecture for multiple user plane paths with multiple UEs per device [3GPP-23501].

For selecting different gNBs for the different UEs, a concept called *Reliability groups* is used. UEs of the devices and the cells of gNBs are grouped into more than one reliability groups. If we select a cell being in the same reliability group as the UE, it can be ensured that UEs in the same device can be assigned to different gNBs. (In the normal case, all cells of a gNB are in the same reliability group, but distributed implementations may allow multiple reliability groups in a gNB.) This is illustrated with an example in Figure 5.4-3 below where UE1 and the cells of gNB1 belong to Reliability Group A, while UE2 and the cells of gNB2 belong to Reliability Group B.

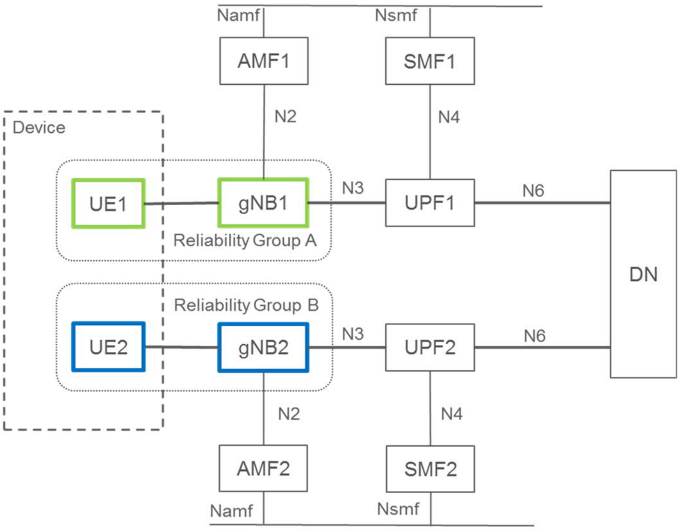


Figure 5.4-3: Example for Reliability groups [3GPP-23501]. .

5.4.3 Redundant transmission on N3 and N9 interfaces

This solution is based on creating redundant tunnels for the same traffic. If the traffic is to be duplicated on N3, then there are two tunnels set up between the gNB and the UPF. Similarly, if redundant traffic is needed on the N9 interface, then multiple tunnels are established between the UPFs.

When the multiple tunnels are set up between the gNB and the UPF, in the uplink direction it is the gNB's responsibility to duplicate the packets and send them in the different tunnels, while the UPF is responsible for identifying and discarding duplicates. In the downlink direction, the UPF has to create duplicates and send them via the different tunnels, and the gNB should discard the duplicates if more than one arrived.

5.4.4 Relation to PriMO-5G

Certain public safety operations and future applications in public safety require URLLC services. For example, drone control requires low latency and high reliability, as packet losses between the drone and the controller are not tolerated. In this case, the above described redundancy mechanisms could increase the chance that a single packet is delivered successfully between the controller and the drone, even if one of the duplicates gets lost.

5.5 Non-public networks

Mobile network services are commonly realized over a network infrastructure offered to the general public, however, the 3GPP 5G standards also allow ways to deploy a network to be used as Non-Public Network (NPN). In contrast to a Public Network (PN), an NPN is a network deployed to provide communication services for a clearly defined group of users inside an organization or group of organizations [ACIA19]. A 5G NPN can be deployed within organization premises e.g. campus or factory, or even as a wide-area network for a specific use.

Generally, an NPN can be deployed in conjunction with PLMNs, where the PLMN itself is not limited to serve the NPN(s), but they also operate as a regular public network to its subscribers. This form of NPN deployment is also known as integrated NPN, whereas standalone NPNs without interaction with the PLMN are also known as an isolated NPN. These deployment options, on a very high-level, permit NPNs to be completely isolated networks or partially integrated with PLMNs. While the isolated deployment is straightforward, i.e. a NPN has all necessary components such as a 5G RAN and Core Network running dedicated and isolated for the specific NPN's use, the integrated deployment has many aspects to be considered such as frequency allocation, e.g. in cases the RAN is to be shared, operation and management, security, trust and isolation, device connectivity when it comes to service coverage, reachability and service continuity (e.g. roaming), QoS, economic feasibility, etc.

According to the 3GPP definition in [3GPP-22261], "Non-public networks are intended for the sole use of a private entity such as an enterprise, and may be deployed in a variety of configurations, utilizing both virtual and physical elements". 3GPP in Release-16 introduced new functionality to support NPNs, in order to fulfil requirements listed in [3GPP-22261]. Architectural aspects are documented in [3GPP-23501].

5.5.1 Stand-alone Non-Public networks (SNPNs)

Public Landline Mobile networks (PLMNs) are identified by PLMN ids, where the PLMN id consists of a Mobile Country Code (MCC) and a Mobile Network Code. The MCC, which is 3-digit long, identifies a country, except for the 90x range, which is country-agnostic, and therefore could be used for stand-alone NPNs. The MNC, which is 2 or 3-digit long, identifies a mobile network inside the country, or a mobile network, if the MCC is country-agnostic. To be able to allow more standalone non-public networks, as the MCC+MNC pair would allow in the country-agnostic range, a new identifier has been added by 3GPP.

An SNPN is identified by the combination of a PLMN ID and a Network identifier (NID). The NG-RAN node as part of SNPN provide access to UE by broadcasting the following information: one or more PLMN ids, list of NIDs (per PLMN ID), this information identifies the SNPNs the NG-RAN provides access to, and another optional information (human-readable network name, and information to prevent UEs not supporting SNPNs accessing the cell).

An example Standalone Non-public network is show conceptually in the following Figure 5.5-1:

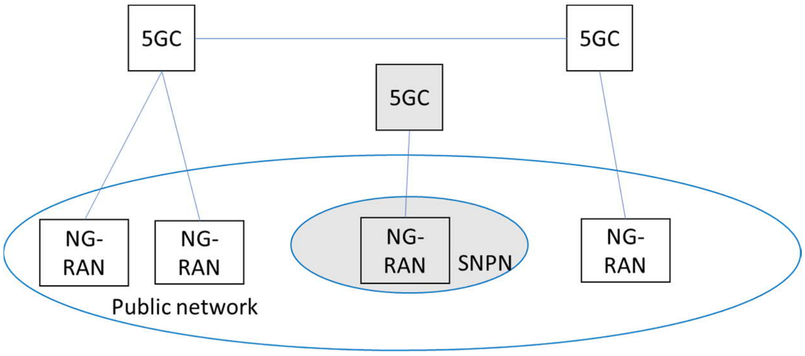


Figure 5.5-1: Standalone Non-public network example.

It is possible for a UE that has registered to an SNPN to perform another registration with a PLMN, via the SNPN user plane, to be able to access PLMN services via stand-alone non-public networks. The other way around, a UE, that is already registered to the PLMN can also perform another registration via the SNPN to the PLMN.

Service continuity can be also achieved: e.g. when a UE moves from an access network belonging to a SNPN to an access network belonging to the NPN, it is possible to handover the PDU session (set up to the PLMN via the SNPN) and preserve the IP address. Similarly, it is also possible to achieve service continuity for the PDU sessions in the other direction.

5.5.2 Public network integrated NPNs

Public network integrated NPNs can be enabled by means of dedicated Data Network Names (DNNs) or by network slicing. The Data Network Name (DNN) is a name that identifies a Data Network (similar to APN used in the Evolved Packet System). A network slice is a logical network serving a defined purpose or customer, consisting of all the required network resources configured together. By accessing a specific DNN or a specific slice, connectivity to Public network integrated NPNs can be achieved. Different level of sharing of the resources between the PLMN and the NPN is possible (see the next section on the deployment options).

An example deployment for Public network integrated NPNs (PNI NPNs) is shown conceptually in the following Figure 5.5-2:

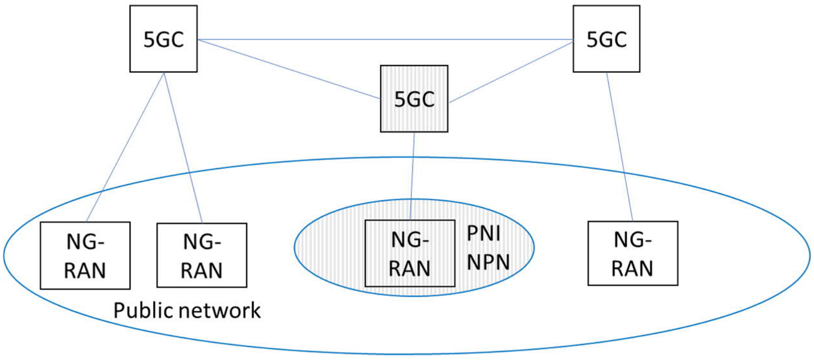


Figure 5.5-2: Example for Public network integrated NPN (PNI NPN).

For PNI NPNs, Closed Access Groups (CAGs) may be used to apply access control. It is used to prevent UEs who are not permitted to access the NPN to even attempting to access the cell (normally, a rejection would occur that means utilizing network resources unnecessarily. A CAG identifies a group of subscribers that are permitted to access one or more CAG cells associated to the CAG. CAG is used to prevent UEs which are not allowed to access the NPN via the associated cells, from automatically selecting and accessing the associated cells.

A CAG is identified by a CAG Id, which is unique inside the given PLMN id. The CAG cells broadcast one or more CAG Ids per PLMN, and a UE is configured with an Allowed CAG list, and optionally, with a flag specifying that the UE is only allowed to access CAG cells.

5.5.3 Non-public network deployment options

Following 5G-ACIA [ACIA19], and the NGMN White Paper [NGMN19] dealing with NPNs for URLLC and factories of future, multiple deployment options can be considered. However, they are also valid for other use cases as the originally described industry cases, for example even for public safety domain. Figure 5.5-3 below describes four different deployment options.

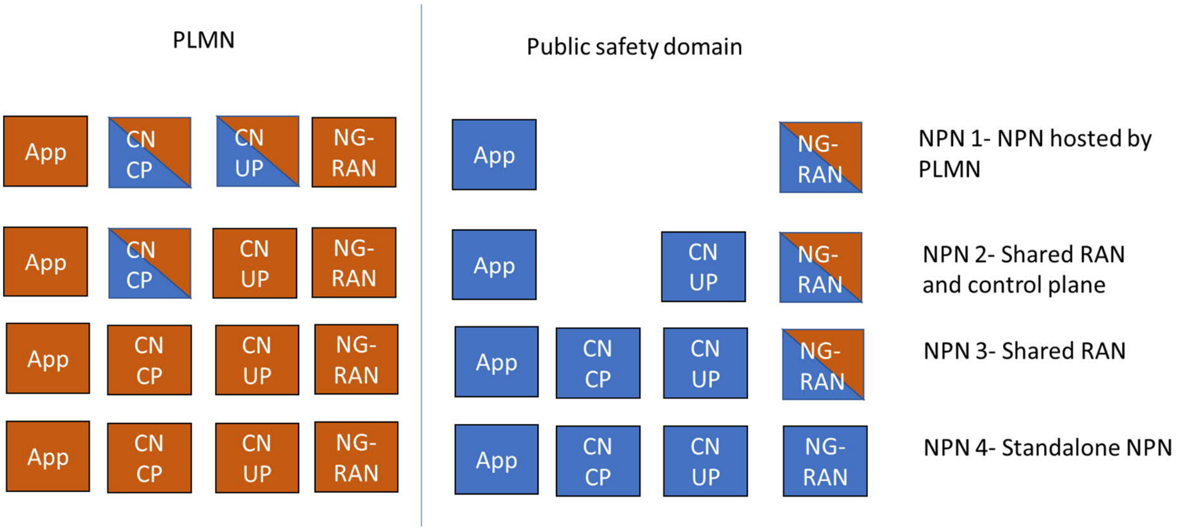


Figure 5.5-3 NPN deployment options

These are evaluated in further details in the following Table 5.5-1, which is based on NGMN work [NGMN19].

Table 5.5-1 Non-public network deployment option characteristics

Deployment option No.	NPN deployment options	Characteristic/Details	Key observations
NPN 1	NPN hosted by the PLMN	NPN and PLMN traffic are isolated and treated differently through virtualization of network functions, DNNs, slicing, etc.	• NPN data traffic and control traffic are leaving the public safety organizations • NPN devices are PLMN subscribers, i.e. SIM-based devices. • Global connectivity enabled by operator roaming agreements.
NPN 2	Shared RAN and control	The NPN RAN and control plane are shared and hosted by the PLMN. Segregation can be	• NPN data traffic remains within the public safety

	plane	realized via usage of different APNs, or slicing IDs, etc.	organization and NPN control traffic is leaving the public safety organization • NPN devices are PLMN subscribers, i.e. SIM-based devices. • Seamless roaming is possible for NPN devices aiming to connect to an NPN service.
NPN 3	Shared RAN	The NPN has its own ID. Only the RAN is shared with the PLMN, all other network functions remain segregated, also data flows remains local. It can be realized by: • Multi-Operator Core Network (MOCN), where two or more entities, e.g. PLMN and NPNs, are sharing eNodeB/ gNodeB and spectrum • Multi-Operator RAN (MORAN), where two or more entities are sharing eNodeB/gNodeB with non-shared spectrum	• NPN data traffic and NPN control traffic remain within the public safety domain. • Dedicated spectrum is required for deployment considering MORAN. • Spectrum can be shared when MOCN is considered for deployment.
NPN 4	Standalone	• All NPN functionalities are within the public safety organization. • NPN is a fully separate physical network from the PLMN with dedicated NPN ID. • However, dual subscription with NPN and PLMN is possible. • Access to PLMN services can be realized via an optional firewall connection and roaming agreement.	• Requires deployment of all 3GPP functions as part of the NPN • NPN data and control traffic will remain within the public safety organization. • Dedicated or leased spectrum for deployment is required.

5.5.4 Relation to PriMO-5G

Following the definitions, a non-public network is a network not intended for public use and intended to be used by organizations or a group of organizations. In this context, public safety organizations can use an NPN to be able to handle their communication needs. Non-public networks can come in many flavours, from isolated (standalone) NPNs to Public Network integrated NPNs, where some of the components from the PLMNs are shared between the public network and an NPN: the choice of the exact flavour depend on multiple factors, including device connectivity (need for service continuity, extended coverage, etc.), service and network isolation, QoS requirements, Operation and Management aspects, security and trust, economic considerations, spectrum related questions etc.

5.6 Isolated Operation for Public Safety

The Isolated Operation for Public Safety (IOPS) provides the ability to maintain a level of communications for public safety users, via an IOPS-capable eNB (or set of connected IOPS-capable eNBs), following the loss of backhaul communications ([3GPP-23401], [3GPP-23180]).

The following Mission Critical Push To Talk (MCPTT) service features are supported in the IOPS mode of operation: MCPTT group call, MCPTT emergency group call, MCPTT private call, MCPTT emergency private call, and the MCData short data service.

To support this there are a set of functional models defined for signalling control plane, application plane and data plane. Below will the functional model be shown for the application plane in Figure 5.6-1:

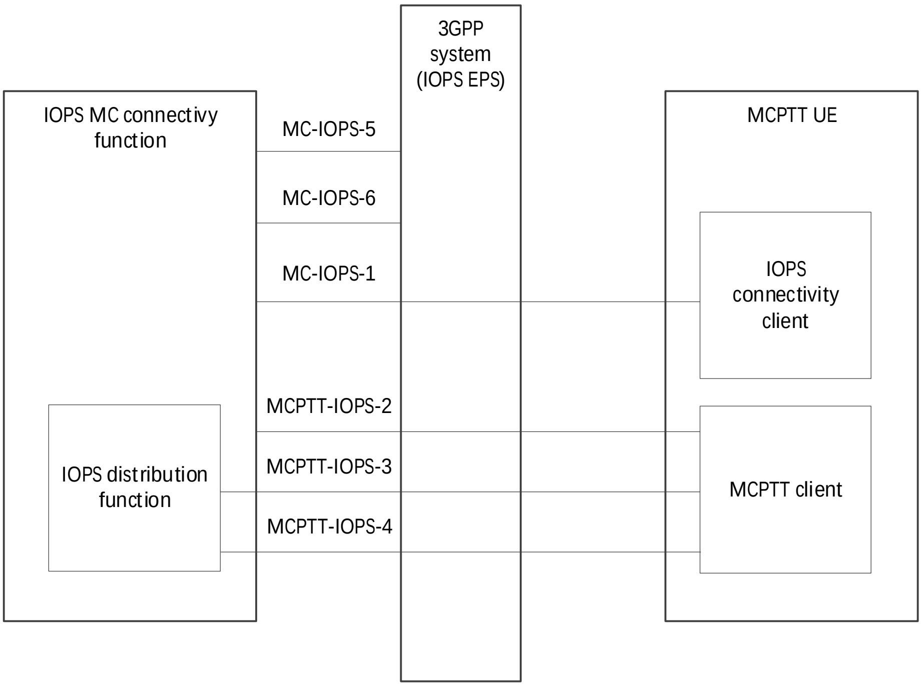


Figure 5.6-1: MCPTT functional model for the application plane in the IOPS mode of operation including a local IOPS EPS [3GPP-23180].

- The IOPS MC connectivity function entity provides supporting MC services in the IOPS mode of operation. It includes registering and discovering users within the IOPS MC system.
- The IOPS distribution function entity provides support for distributing the IP packets containing the MC service application data received from a MC service UE in the IOPS mode of operation.
- The IOPS connectivity client functional entity provides support for enabling that a user at the MC service client is registered and discovered by the IOPS MC system in the IOPS mode of operation.
- The MC service client functional entity, e.g. a MCPTT client and a MCData client, supporting

the IOPS mode of operation acts as the user agent for the corresponding MC service transactions as well as all IOPS related application transactions.

Both the MCPTT and the MCData functional model have a similar architectural set up.

A PLMN identity is dedicated to IOPS mode of operation and is broadcast in System Information by the eNB when IOPS mode is in operation. Only authorized IOPS-enabled UEs can access a PLMN indicated as an IOPS PLMN.

The decision by an IOPS-capable eNB to enter IOPS mode of operation is made in accordance with the local policies of the RAN operator. Such policies can be affected by any RAN sharing agreements that are in place. The local EPC system will be activated and the eNB will announce IOPS mode. The Local EPC instance includes at least MME, SGW/PDN GW and HSS functionality.

When the eNB detects that the normal backhaul link is up and running again the eNB will detach the UE from IOPS operation, deactivate the local EPC and announce normal operation. The eNB will again attach to the macro EPC.

5.7 Optimal routing

5.7.1 Introduction

3GPP has defined three Session and Service Continuity (SSC) modes for PDU sessions. With SSC mode 1, the UPF established at the PDU Session Establishment is maintained. The UE’s PDU Session IP address does not change, even after mobility. SSC mode 2 means that the network may trigger the release of the PDU Session and instruct the UE to establish a new PDU session. In this scenario, the IP address changes and a new PDU Session Anchor UPF may be selected. Finally, SSC mode 3 introduces make-before-break, where the network ensures that there is no loss for connectivity. The network allows the establishment of UE connectivity via a new PDU Session Anchor UPF to the same Data Network before connectivity between the UE and the previous PDU Session Anchor is released. SSC mode 3 involves changing the IP address.

Unfortunately, with the defined SSC modes it is not possible to achieve the following goals simultaneously:

- Ability to communicate with close edge servers
- Low latency communication between two UEs
- Disruption in communication is to be minimized/eliminated

However, in certain scenarios, it is desired to maintain the ability to communicate with close edge servers, offer the possibility of low latency communication between two UEs, and keep these properties even after mobility in such a way that there is no/little disruption in communication. Therefore, a new user plane handling is proposed: Optimal Routing.

5.7.2 Optimal Routing Architecture

The Optimal Routing Architecture introduces conceptually two new elements in the 5G Core architecture. First, the IAP (IP Announcement Point), which is a user plane element, and second, the LR (Location Register), which is a control plane element.

All downlink packets (sent from a host on a Data Network, e.g. from the Internet or from an operator services network) will pass an IAP on its way to the UE. Uplink packets do not pass the IAPs. Downlink packets reach the IAPs via regular IP routing and forwarding. In a typical case, when multiple IAPs are deployed in the network, different packets to the same UE (arriving from different sources) may pass different IAPs. The IAP is responsible to encapsulate and forward the packet to the UPF, where the UE’s user plane session is handled. The IAP receives the information where to send the packets by querying the Location Register.

The LR is conceptually a centralized database that stores the UE's location information. We define location as the IP address of the UPF, where the UE's session is currently established. Therefore, the LR is a database that stores UE (PDU Session) IP address -> UPF IP address mappings. In case the UPFs are chained, i.e. there are multiple UPFs that need to be passed between the DN and the RAN, the UE's location is the IP address of the "top-most" UPF. In this document, we assume that a single UPF is handling the UE's session. Also, a UE can have multiple PDU sessions, and it is possible that the different sessions will be handled in different UPFs. If the UE has multiple sessions, there will be multiple entries in the LR corresponding for the UE, one UPF IP address for each PDU session (UE) IP address.

The LR entries may be signaled from the SMF at different procedures, e.g. entries are inserted during PDU Session Establishment or modified during Xn handovers involving UPF change.

The IAPs can store the information received in their local cache. It means that there is no need to do a query to a remote location every time a new user plane packet is received. The control plane of the 5GC makes sure that the entries are up to date: the LR keeps track of the IAPs that asked for a certain UE IP address and e.g. during Xn handover involving UPF change, these IAPs are informed. These procedures are detailed in the PriMO-5G deliverable D2.1 [PRI19-D21].

The Optimal Routing architecture is shown in a schematic illustration in Figure 5.7-1. In the figure, two remote servers communicate with the UE, and the flows go through different IAPs. Again, for more information on how the LR and IAP works, please see D2.1 [PRI19-D21].

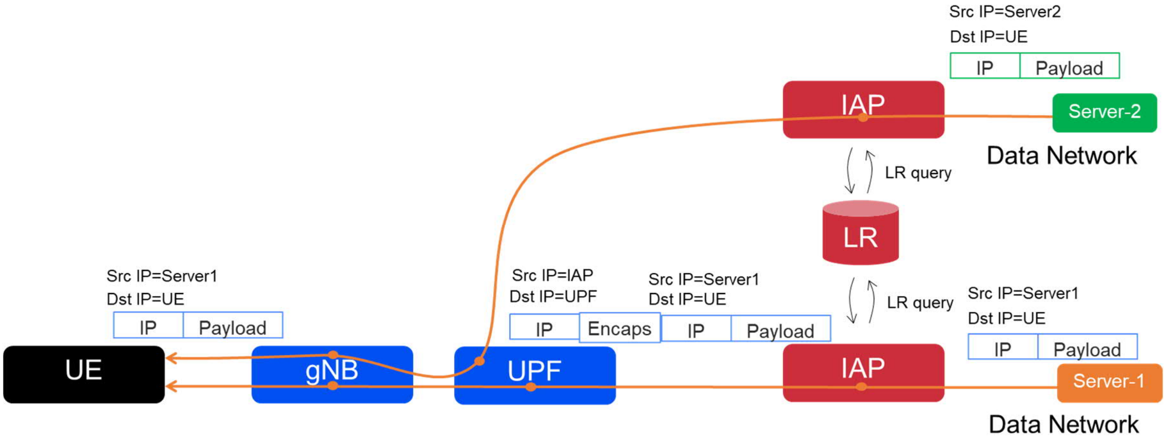


Figure 5.7-1: The Optimal Routing Architecture.

5.7.3 *Scenarios*

There are two main scenarios for Optimal Routing:

- *UE simultaneously communicating with Central (or Internet) servers and Local (edge) servers*. This is depicted in Figure 5.7-2:. In this case, at Step 1, the UE communicates with a central server and also with Server-1 at Local Site 1. At this point UPF-L1 is handling the user plane session for this PDU session. After changing gNB and UPF simultaneously (in Step 2), the related IAPs will encapsulate the packets to UPF-L2. At this point the UE still communicates with the old Server-1 at Local Site 1. The change to UPF-L2 and keeping the IP address enables to continue the communication with Server-2 at Local Site 2 with low latency. However, to fully achieve the situation depicted at Step 3, the application context has to be transferred from Server-1 to Server-2.

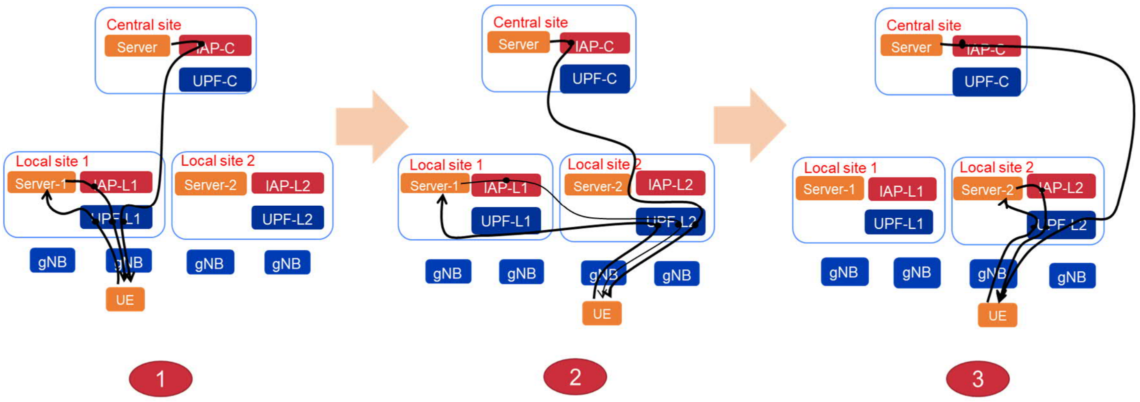


Figure 5.7-2: UE simultaneously communicating with central and local servers.

- *UE-to-UE communication.* This is depicted in Figure 5.7-3:. First, UE1 is handled in UPF-L1 and U2 is handled in UPF-L2. When UE1 moves closer and after a handover, its user plane session will be handled at UPF-L2, the two UEs can communicate with each other without the packets visiting the old local site or any central site.

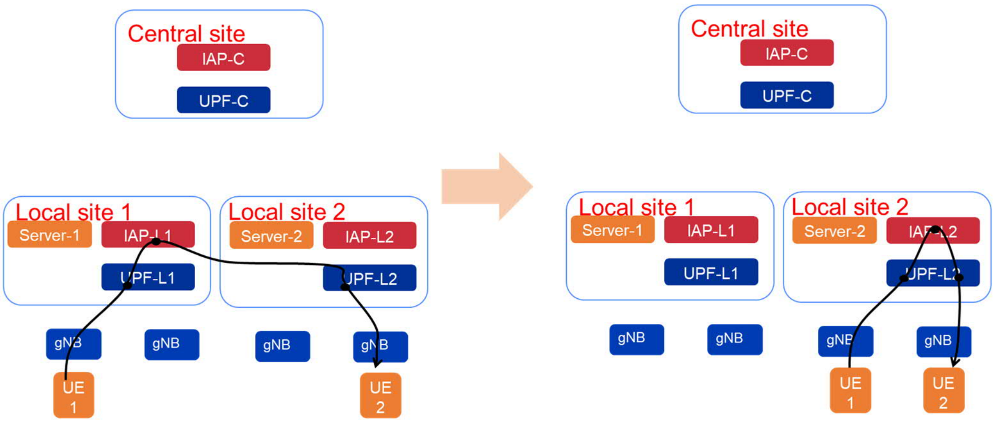


Figure 5.7-3: UE-to-UE communication.

The scenarios can be combined: it is possible that a UE communicates via central (Internet) or local (edge) server, and also with other UEs simultaneously on low latency paths, even after mobility events and without IP address changes. All the communication above can be reached with a single PDU session per UE, which means that there may be less state in the network to maintain and puts less requirements on UEs.

5.7.4 Relation to PriMO-5G

In PriMO-5G use cases, UEs can be e.g. equipment held by firefighters, or drones. In certain scenarios drones communicate with firefighters, e.g. sending visual information to them via interactive video services. In this UE-to-UE scenario, low latency is desired, even after mobility events.

Drones may be controlled by central operation centers (central servers), or also by firefighters with UEs. This latter is another UE-to-UE scenario important for PriMO-5G. In this case, it is important that the control communication is not disrupted at mobility events, and the low latency is always kept. By keeping the IP address, the application logic may be simplified, as there is no need to re-discover the IP address.

In many cases, edge computing is desired to process e.g. video information coming from the drones, or to support AR/VR communication. At the same time, the drones (or other UEs) may communicate with other UEs and central servers (e.g. with central operation center).

In all the above cases, Optimal Routing provides a framework for achieving the goals of low latency and keeping the IP address. Furthermore, the different use cases can be combined (edge computing and UE-to-UE scenarios).

5.8 Network data analytics function

NWDAF is part of the service-based architecture in 5G Core, see Figure 5.8-1 below:

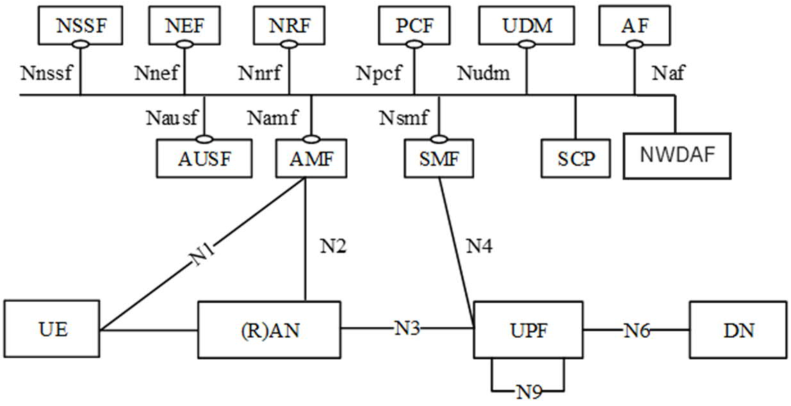


Figure 5.8-1: Service based architecture with NWDAF.

The main purpose of NWDAF is to collect data and perform data analytics to support the internal needs of the 3GPP systems from network functions. There is also an opportunity for an application function to subscribe to NWDAF data and to access analytics. NWDAF has been introduced as a placeholder in Release-15, and in Release-16 there was intensive work on NWDAF services. NWDAF is defined in 3GPP in [3GPP-23228]. See Figure 5.8-2.

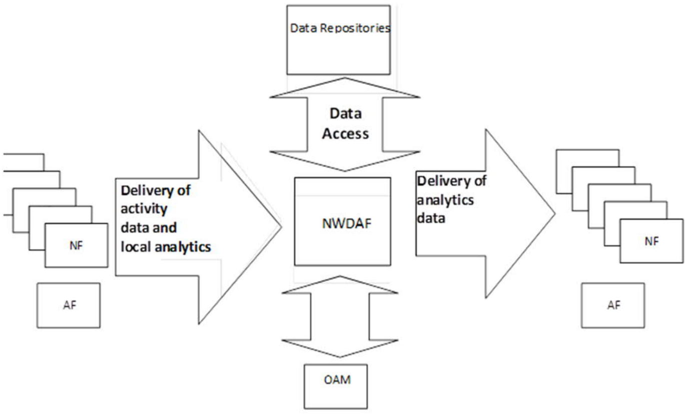

Data Repositories
Data Access
Delivery of activity data and local analytics
NWDAF
Delivery of analytics data
NF
AF
NF
AF
OAM

Figure 5.8-2: NWDAF in the core for data collection and analytics to support network functions and application functions [3GPP-23791].

The NWDAF utilizes the existing service-based interfaces to communicate with other 5GC network functions and OAM. NWDAF can collect data from network functions and application functions, retrieve additional data from OAM or other data repositories and deliver data to network functions and application functions. The analytics provided by NWDAF is either statistical information of past events or predictive information. In Release-16, NWDAF only supports non-roaming architectures. Multiple instances of NWDAF can exist in the network.

To discover NWDAFs, network function consumers to NWDAF can utilise the NRF to access data that can support the selection of NWDAF, e.g. NWDAF Serving Area information that contains a list of TAIs for which the NWDAF can provide analytics.

Data collection can be made from various sources:

- Network functions (e.g. SMF, AMF, etc.), via the event exposure service offered by each network function or NRF
- OAM data from the OAM systems: mainly generic data, performance data and fault supervision data and 5G E2E KPIs,
- Application functions based on instance needs (e.g. to request service experience data from an application)

There is a growing list of services that can be provided from the NDWAF analytics services. Below follows some of the services that may be provided:

- Slice oriented services; like load level information
- Application oriented services; like observed service experience (e.g. MOS values),
- Network function-oriented services; like NF load levels, user data congestion analytics which can relate to the congestion experienced when transferring data over the control plane/user plane
- Network related services; like QoS sustainability statistics for a list of TA or cells
- UE related services; like network performance analytics on UEs for an area of interest towards a NF or a group of NFs both for the past as well with predictions, could be UE mobility analysis which describes UE cell or TA coverage including predictions, UE communication analytics which can give data about traffic characterizations and traffic volumes both for the past as well as predictions, identification of UEs with abnormal behaviour

As an example, the Nnwdaf enables the PCF to subscribe to and be notified on slice load level analytics. The following information are notified by the NWDAF: Identifier of network slice instance, Load level information of network slice instance. This may enable the PCF to set up rules regarding PDU session establishment.

For public safety many of these data collection and analytics services can be of great value:

- Slice level: review of load levels and indications on if a certain slice may be at risk
- Application level: MOS levels for internal communication and maybe divided per public safety function (police, medical etc)
- Network function level: secure deployment and operation of AFs for the public safety network
- Network related services: QoS sustainability statistics could be used to determine if more base stations should be provided and/or if new TAs should be configured
- UE related services: UE analytics to tune network and services

5.9 5G-LAN

The next generation of industrialization integrates fixed LAN and cellular devices to provide private virtual network (VN).

The industrial automation infrastructure demands a reliable transport as well as seamless connectivity of sensors, actuators and controlling devices distributed in both cellular and fixed networks. To address the needs of next generation industrial networks 5G system should provide virtual network functionality.

Industrial applications require robust and reliable data transport. Over time several technologies have been applied, but Ethernet-based networks are the most widely adopted infrastructure in industrial practice today. Such a proliferation of Ethernet stems from the facts that it is standardized and is a de-facto transport for TCP/IP-based industrial wireline communication, which also facilitates integration of industrial applications with Information Technology (IT) services. The use of Ethernet in industrial automation has been realized via protocols like EtherCAT, Ethernet Powerlink and PROFINET thus VN should be seamless integrated with 5G system. In the event of firefighting emergency, a rescue team can deploy a 5G VN that includes all the mobile devices part of the rescue team. The 5G-LAN integration allows to create secure network integration also fixed devices using either IP or Ethernet transport. The rescue team could utilize multiple 5G VNs each having own network slice to private communication with each session associated to different 5G VN.

3GPP Release-16 defined the VN management as part of Network Exposure Function (NEF) that will allow the emergency rescue team to setup the group members of the 5G VN.

6 Edge Architecture

Cloud computing offers storage, computational, and networking facilities within a single or multiple virtualization platform for enabling different services for mobile networks. Such infrastructure services can be offered by separate service providers. However, cloud computing has shortcomings with regard to emerging applications that require ultra-short latency. These limitations are principally due to the centralized cloud computing architecture. Multi-access Edge computing (MEC) represents a vital solution to these limitations. By pushing the computing resources to the edge servers that are near to users, it allows reducing the delay and enables applications requiring response time in the range of milliseconds. This section provides an overview on MEC integration with 5GC.

6.1 Overview

The MEC has been specified mainly in ETSI. The 3GPP provides the enablers for integrating MEC into 5G networks but does not specify the functionality [3GPP-23501]. On the other hand, ETSI ISG MEC (Industry Specification Group for Multi-access Edge Computing) is the home of technical standards for edge computing. The specifications in ETSI define the requirements, deployment scenarios and interfaces [ETSI20], [ETSI16], [ETSI18a], [ETSI18b].

The design of MEC requires integration of 5G architecture defined in 3GPP and MEC defined in ETSI as shown in the following diagram in Figure 6.1-1.

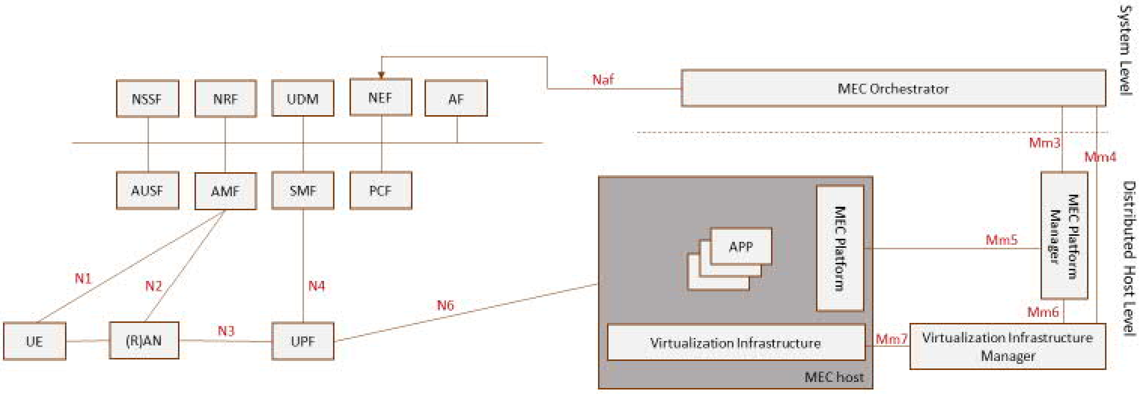


Figure 6.1-1. 5GC system and MEC integration.

6.2 ETSI MEC architecture

The MEC system reference architecture, defined in [ETSI18b], consists of MEC hosts and functional entities required to run the MEC applications within an operator network. Two levels are distinguished: distributed host level and system level. The distributed host level includes MEC hosts and management entities. Each MEC host contains a virtualization infrastructure, which provides compute, storage, and network resources, and a MEC platform. The latter ensures the required functionalities to run a MEC application on a virtualized infrastructure (this include discovering and advertising MEC services, instructing data plane according to traffic rules, configuring DNS proxy/server, etc.). The management of the host level consists of the MEC platform manager and the virtualization infrastructure manager. They are responsible for handling the functionalities of a particular MEC host and the application running on it. The system level includes the MEC orchestrator as a core functional entity. The MEC orchestrator has an overview of the complete MEC system. Its functionalities include triggering application instantiation and termination, triggering application relocation, etc.

The figure above shows the 3GPP 5G system architecture on the left while MEC system architecture is on the right. Initial efforts to integrate these two architectures are provided in [ETSI18b] and [3GPP-

23501] Clause 5.13. The design approach of 5G architecture allows mapping MEC onto Application Functions (AF). At the MEC system level, the MEC orchestrator interacts with the Network Exposure Function (NEF) of the 5G system through the Naf interface (interface exhibited by AF). At the MEC host level, the MEC platform is the entity that interacts with 5G NFs and takes care of controlling the traffic steering to the MEC applications. The host level functional entities are deployed in a Data Network (DN) that could be external to the 5G system.

6.3 3GPP edge architecture

Edge computing is about bringing the services closer to the location where they are to be delivered. Services include computing power and memory running the requested application. Applications, data, and computing power are pushed away from centralized points (central data centers) to locations closer to the user (e.g. distributed data centers). The motivation is to reduce latency and reduce transmission costs. Example use cases include AR/VR, real-time facial recognition, video surveillance, etc.

Currently, edge computing is specified across various standardization bodies and open source fora, including 3GPP, ETSI, CNCF (Cloud Native Computing Foundation), ONAP (Open Network Automation Platform) and LF Edge. In this subsection, we focus on the work being done in 3GPP.

The subsection is built up the following way. First, 3GPP general tools for connectivity supporting edge computing is introduced in Section 6.3.1, then potential enhancements are described in Section 6.3.2. Finally, the 3GPP defined edge application architecture is described in Section 6.3.3.

6.3.1 General tools to enable edge computing

So far, 3GPP SA2 has mainly focused on access and connectivity aspects that act as enablers for edge computing. These general tools can provide efficient user plane paths between UEs and application servers. These generic connectivity tools allow different edge architectures, including but not limited to ETSI MEC and 3GPP EDGEAPP.

3GPP specified general tools that can provide efficient user plane path. They can be used as enablers for edge computing:

- **UPF selection**: the SMF is responsible for UPF selection. As a prerequisite, the SMF has to be aware of which UPFs are available and what properties they have (e.g. capabilities, load, location). This may be achieved different ways. First, the SMF can be configured with the available UPFs via O&M, and the configuration may include topology related information. Second, available UPFs may be discovered via the Network Repository Function (NRF). Third, capabilities, and e.g. UPF load information can be received from the UPF when the basic N4 connection is set up. UPF selection is performed at the SMF at certain events, e.g. at PDU Session Establishment or mobility events. At this point, the SMF can take into account various information to select UPF(s), for example:
    - UPF dynamic load
    - UPF relative capacity
    - UPF location
    - UE location information
    - UPF capabilities
    - Data Network Name
    - PDU Session Type
    - SSC mode for the PDU session
    - UE subscription profile
    - Data Network Access Identifier (DNAI)
    - S-NSSAI
    - Access technology
    - User plane topology

- **Selective traffic routing to Data Network (Uplink Classifier and IPv6 multi-homing)**: In the simplest case, for each PDU session, there is a single PDU Session Anchor UPF (PSA UPF), and thus a single N6 interface towards a Data Network. However, it is possible to select multiple UPFs for a single PDU session, and there may be multiple PSA UPFs associated with the PDU session, each providing access to different parts of the same Data Network, and traffic can be selectively routed to different N6 interfaces. For edge computing, one PSA UPF can correspond to a UPF with N6 interface towards a local edge site, where edge applications can run close to the users, and another PSA UPF can correspond to a UPF interfacing a central data center and the Internet peering point. This can be achieved by either inserting Uplink Classifier (UL CL) UPF, or Branching Point (BP) UPF in the path.
  - UL CL is a functionality to direct traffic to different PSA UPFs in the uplink direction, while in downlink direction the UL CL UPF merges traffic coming from different PSA UPFs to the UE. The classifier works on traffic filters, that is inserted by the SMF. By checking the filtering rules and examining the destination IP Address (or IPv6 prefix) on the uplink packets, the UL CL decides which PSA UPF to send the packet to. Adding and removing UL CL and additional PSA UPFs can be performed by the SMF any time during the lifetime of a PDU Session.
  - The Branching Point functionality works with IPv6 multi-homing. IPv6 multi-homing means that multiple IPv6 prefixes are assigned to the UE within a single PDU session. Each IPv6 prefix will be served by a different PSA UPF. The common, Branching Point UPF examines the source IPv6 address of the uplink packets, and based on that, it will forward them to the corresponding PSA UPF. Similar to the UL CL case, the SMF at any point of time insert and remove a Branching Point during the lifetime of the PDU session. The key difference between the UL CL solution and the Branching Point solution is that the first one is purely network-based, while the second one needs UE involvement and support.

- **SSC modes**: At the start of a PDU Session, a PSA UPF is allocated that will be also the IP anchor for the PDU session. However, after mobility, a new PSA UPF may be closer to the UE and it would be beneficial to change the PSA UPF. However, this change would mean IP address change and different applications/services handle IP address changes differently. Therefore, 3 Session and Service Continuity (SSC) modes are defined: SSC mode 1, 2, and 3. The SSC mode is assigned to the PDU session when it is established. SSC mode 1 preserves the PSA UPF and the IP address, while SSC mode 2 and 3 allows the PSA UPF to change, but this involves assigning a new IP address. SSC mode 2 works with “break-before-make”, while SSC mode 3 works with “make-before-break” fashion. Considering edge scenarios, the SSC modes 2 and 3 help to set up and maintain efficient user plane paths, however, with the price of the IP address change that may not be tolerated/supported by some services.

- **Application Function influence on traffic routing**: This is a control plane solution where an AF can provide input to the 5GC on how certain traffic should be routed. It is up to the 5GC (and in particular the SMF) to honour these by the available tools, e.g. UPF selection, SSC modes, UL CL/BP insertion. The AF either sends the request to the PCF, when it is allowed, or the request goes via the Network Exposure Function (NEF). The AF in this request may provide the following information:
  - Traffic descriptor
  - Potential location of applications, represented by Data Network Access Identifier
  - UE identifier
  - N6 routing information

- **Network capability exposure**: The Network Exposure Function’s (NEF) main functionality is to support interaction with external applications. There are multiple use cases for NEF, while the AF influencing the traffic routing is the one that is most relevant to edge scenarios.

- **Local Area Data Networks (LADN):** This feature allows enabling access to Data Networks from specific areas. The LADN is available in the so-called LADN Service Area, which is a set of Tracking Areas. Outside the area, the UE is not able to reach the DNN.

The mechanisms described above support connectivity to edge services. Based on the above, there are at least 3 different connectivity models for edge applications that are supported by current 3GPP mechanisms:

- **Distributed Anchor Point**: The single PSA UPF for the PDU session is distributed in the network, e.g. to local sites. All application traffic is using the same PDU session. The PSA UPF may be changed after UE mobility, with UE IP address change (with SSC modes 2 and 3).
- **Session Breakout**: A PDU session is served by two PSA UPFs: one PSA UPF is located in a central site, while another one in the local site. An UL CL or a BP diverts traffic to the local PSA UPF interfacing e.g. local data center hosting the application server. The local PSA UPF may be changed as the user moves.
- **Multiple PDU Sessions**: edge application traffic and other traffic uses different PDU sessions. The edge application traffic has a PDU session with a PSA UPF in a local site, while the other traffic has a PDU session with a central PSA UPF. After mobility, the local PSA UPF may be changed (with SSC modes 2 and 3).

These connectivity models are shown in Figure 6.3-1:.

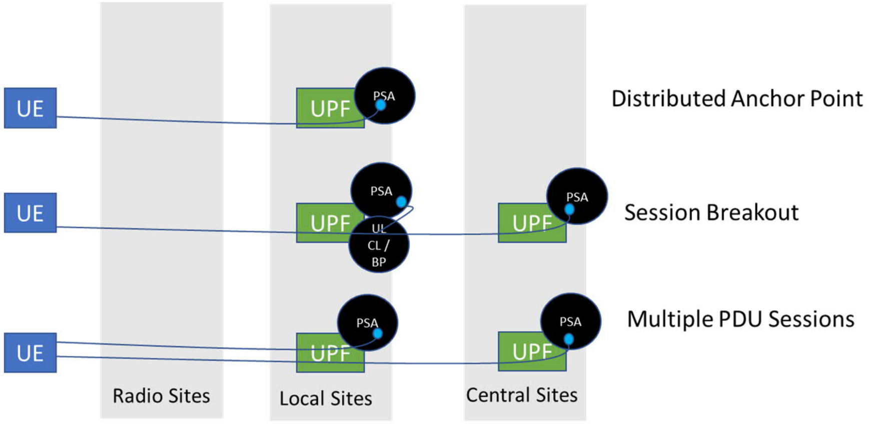


Figure 6.3-1: Connectivity models for edge computing.

Furthermore, the Optimal Routing solution described in Section 5.7 allows a new connectivity model, where all traffic is utilizing a single PDU session, and the UE IP address is not changing. This supports both edge scenarios and UE-to-UE traffic without the need to visit central sites.

6.3.2 Release-17 3GPP SA2 work on enhancement of support for Edge computing

At the time of writing this document (1st Quarter 2020), 3GPP SA2 is working in Release-17 on enhancement of support for edge computing. The on-going work is in the study phase, and the key issues and solution proposals are documented in [3GPP-23748].

Among others, key issues most relevant to PriMO-5G include edge application server discovery and edge relocation. In the first, the question is that how the application/UE discover the IP address of a suitable edge application server, as typically multiple edge application servers will be available to provide the same service, e.g. on different local sites. This also involves the question of how to support re-discovery of edge application server if the previous one becomes non-optimal or unavailable for the UE (e.g. due to mobility). The second key issue on edge relocation focuses on the issue on how to handle application server relocation.

As the time of the writing, only solutions for the first key issue is available. In one of these solutions in [3GPP-23748], the Mobile Network Operator deploys a new DNS component, referred to as “DNS AF”. The DNS AF receives a DNS request from a UE, gets UE location information and determines at least one suitable local PDU session anchor (PSA) point for that UE location and application. The DNS AF adds the corresponding N6 access location(s) as ECS (EDNS Client Subnet) option(s) to the DNS request. Now, it is up to the Service Provider to select a suitable edge application server that matches the given location(s). The 5GC, after receiving the DNS reply, inserts the uplink classifier onto the user plane path and sets up the rules to steer the traffic towards the selected edge application server. In this solution, the UE is agnostic to the edge computing.

6.3.3 3GPP Application Architecture for the edge

The 3GPP SA6 WG is working on an application architecture for the edge, in the activity called EDGEAPP. The architecture aims to enable applications to be hosted in the edge of the 3GPP network. 3GPP designed the architecture following the key architecture principles, listed in [3GPP-23758], such as:

- Application Client portability
- Edge Application Server's portability
- Service differentiation
- Flexible deployment
- Interworking with 3GPP network

The normative work is on-going at the time of the writing of this document, and it is documented in [3GPP-23558]. The application architecture is depicted in Figure 6.3-2::

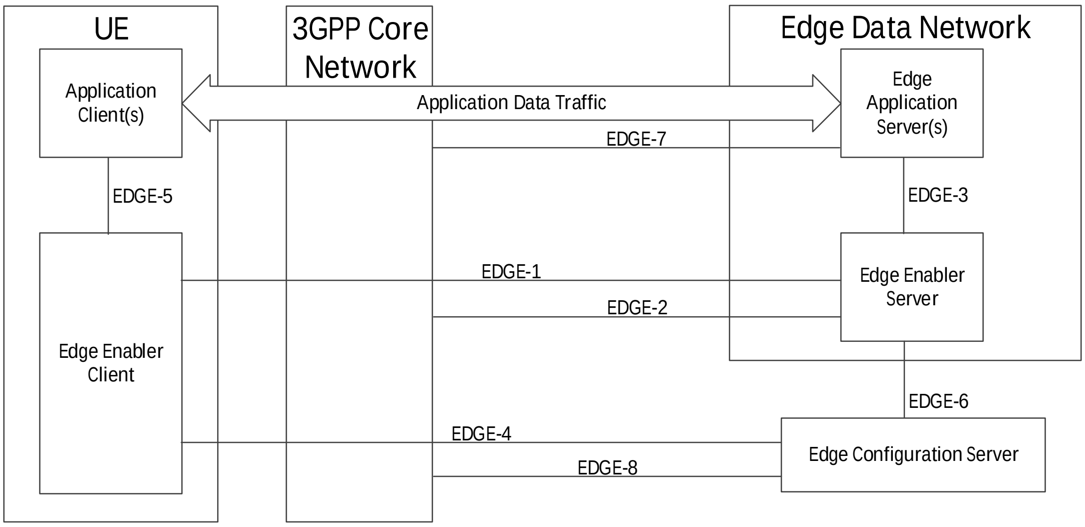


Figure 6.3-2: 3GPP EDGEAPP - Edge application architecture.

In the architecture depicted above, the Edge Enabler Server provides supporting functions needed for Edge Application Servers to run in an Edge Data Network. The Edge Enabler Client provides supporting functions needed for Application Client(s). Finally, the Edge Configuration Server provides supporting functions needed for the UE to connect with an Edge Enabler Server. It also supports Edge Data Network configurations provisioning to the 3GPP network.

6.4 Edge-assisted AI applications

The MEC system architecture is designed to enable applications or parts of applications with a variety of computation, network, and storage requirements to run in virtualized environments as part of an edge cloud that end devices interact with. Machine learning applications that incorporate for instance neural networks can benefit from MEC, both in terms of improving performance of training and inference. This is not only due to the computation offloading and therefore reduced energy consumption on the user equipment, but also because using MEC, it is possible to increase the performance of AI-based applications in terms of algorithm convergence. Such algorithms can be used for data classification, data regression and action-selection applications [WHL+20].

Edge-assisted AI applications can be divided in two types of architectures with respect to which parts at the edge of the network, also taking into account the end devices, can be used for deploying AI applications including their computation-related parts. These two types are the edge cloud architecture, where only the edge servers process data directly delivered from the end devices, and the end-edge cloud, where the edge servers and end devices collaboratively perform AI algorithm tasks [RZH+19].

Regarding the edge cloud architecture, edge-assisted AI applications can make use of machine learning algorithms such as supervised, unsupervised or reinforcement learning. In all cases, the server interacts with the end devices by retrieving input data to process them. As an example, in the case of supervised learning for object recognition, the end device may transmit a video stream to the edge cloud server where an already trained model is deployed for inference. When needed, results can be transmitted back to the end device. An example of mobile edge computing is the use of UAVs as reinforcement learning agents used to improve decision making for e.g., path planning or resource allocation when acting as base stations [LCG+20].

Considering the end-edge cloud architecture, in the past few years, due to the constant increase in computational power of end devices and the decrease in their power consumption to achieve improved performance, end-edge-assisted AI systems have been proposed and are currently under research and development. Two major AI paradigms are the Distributed Deep Neural Networks (D-DNN) and Federated Learning systems. The first one essentially requires the split of the neural network layers so that initial layers of the DNN are executed by the end user device, while the rest are executed by the edge cloud [TMK17]. On the other hand, Federated Learning is making use of a collaboration scheme between multiple end user devices and the edge cloud [YLC+19]. Each of the end-user devices is responsible for training and executing a local DNN. The edge cloud is responsible for retrieving and aggregating the DNN parameters of local models, produced by the end user devices in order to update a global model that can be downloaded by the end user equipment. In this case. communication between the end user equipment and the edge cloud can happen in a synchronous or asynchronous manner depending on the aggregation algorithm running on the edge cloud.

7 AI-enabled Network Management and Orchestration

Administering computer networks requires careful decision-making to ensure that the network can continuously support existing network services and accommodate new ones. These decisions rely heavily on a good assessment of previous and current network management data, which is complex due to the diversity of services that the network provides.

Recent successes in AI show the power of AI-enabled technologies in making better decisions. AI can uncover the underlying patterns in the data provided by the network to understand the network dynamics better or to detect anomalies more accurately. This section discuss how AI can guide network management and orchestration in making better informed decisions to achieve optimal network performance and to reduce network errors.

7.1 Management data analytics function

5G networks are designed to support network slicing and a plethora of network services with diverse requirements in terms of latency, capacity, and reliability, which present significant management and operational challenges. To address these performance challenges and securely maintain the SLAs of the services the management system has to proactively detect and manage any performance degradation or faults preventing SLA violations and the resulting business and financial repercussions. Such a management system benefits from management data analytics services, which can collect real-time performance data that can be used by analytic applications to detect the potential issues in advance and take appropriate actions to prevent or mitigate the issues. Leveraging on the recent progress of AI and machine learning, 3GPP is working on a number of specifications, which can be used to bring the necessary intelligence and automation to the network service management & orchestration

7.1.1 Management Data Analytics Service (MDAS)

3GPP Release-16 specifies a framework of management service, which offers management capabilities that are accessed by management service consumers via standardized service interface composed of individually specified management service components. Leveraging the Service Based Architecture (SBA), management services, such as performance management, configuration management and fault supervision services are exposed to authorized service consumers, which in turn may expose these services to other service consumers. In this context, a management data analytics Service (MDAS) can be used to provide data analytics of different levels of the network. For example, the MDAS at a network function (NF) level can collect the NF's load related performance data, e.g. resource usage status of the NF. The analysis of the collected data may provide forecast of resource usage information in a predefined future time. The performance data of several NFs can, together with other management data, then be analyzed and transformed into one or more management data analytics that can be applied at a domain level (RAN, CN, NSSI) or in a centralized manner, e.g. at PLMN level [3GPP-28533]. A domain-level MDAS provides domain specific analytics, e.g. resource usage prediction in a CN or failure prediction in a NSSI, etc. A centralized MDAS can provide E2E or cross-domain analytics service, e.g. resource usage or failure prediction in an NSI, optimal CN node placement for ensuring lowest latency in the connected RAN, etc.

7.1.2 MDAS for Communication Service Instance (CSI)

5G system is expected to provide a variety of Communication Services Instances (CSIs), each one supported by or more Network Slices Instances (NSIs). The MDAS can help to perform management tasks in the preparation, commissioning, operation as well as in the termination of the Life Cycle Management phases of the Network Slices Instances supporting the CSI [3GPP-28535]. For example, MDAS can support service provisioning by preparing service catalogues, evaluating network requirements for a new service, and carrying out feasibility check. During the operation phase of NSI, the MDAS assumes the role of analytics in the management loop supporting communication service assurance service. In this case, the MDAS takes the network and service management related data collected by the observation stage of the management loop, processes and analyzes the data, possibly using AI and ML models, and provides the analytics reports for root cause analysis of ongoing issues and prediction of network or service demands [3GPP-28809]. The MDAS also classifies and correlates the input data (current and historical data), learn and recognize the data patterns, and makes analysis to derive inference, insight and predictions and feeds this information to the decision stage. The decision phase of the management loop then decides of when and what management actions to be taken based on MDAS reports and other management data (e.g., historical decisions made). The decision may be made by the consumer of MDAS (closed loop), or a human operator (open loop) and feeds the information to the execution stage, for the execution of the management actions according to the decisions. During the execution step, the actions are carried out to the managed networks and services, and the reports (e.g., notifications, logs) of the executed actions are provided.

7.1.3 Management interaction with NWDAF

The MDAS producer provides the analytics data for management based on the data from different types of RAN and CN network functions. However, the MDAS producer may also consume the analytics result of NWDAF or feed its results to it as shown in Figure 7.1-1 [3GPP-28809].

- The NWDAF may consume the MDAS for identified scenarios and provide analytics service for 5GC NF for control purpose.
- The CN Domain MDAS producer may consume the service provided by NWDAF and other 5GC NFs and provide analytics data for management purpose.
- The RAN Domain MDAS producer may consume the service provided by gNB and provide analytics data for management purpose.

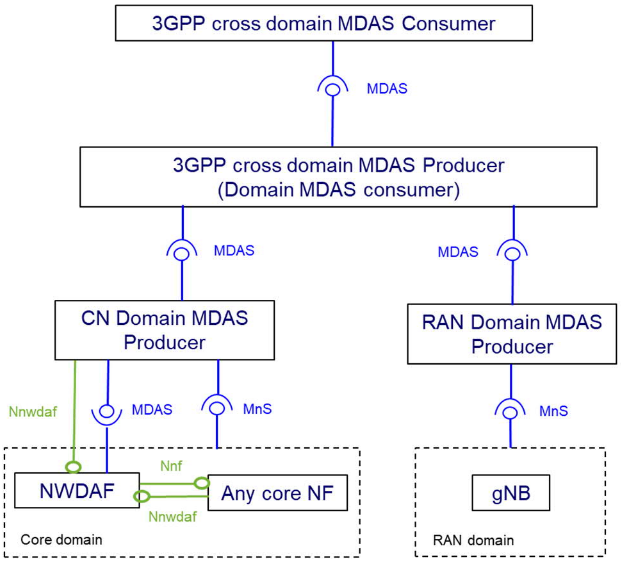


Figure 7.1-1: Coordination between NWDAF and MDAS producer for data analytics.

7.1.4 Implication of MDAS to PriMO-5G uses cases

The firefighting scenario use cases developed in PriMO-5G assumes deployment of communication service insurances that can support the command and Control (C2) communication between the UAV and UAV controller, aerial visual and other sensory data between UAVs and the Incident Commander, and VR/AR video traffic between firefighters at the fire scene and the Incident Commander.

The MDAS will play a critical role in supporting the management of the network slices instances that support the communication services, e.g. in the closed loop management of fault supervision and performance assurance of the communication service.

7.2 Network slice orchestration

As a part of the 5G service-based architecture (SBA), the orchestrator of network slicing isolates traffic and delivers expected requirements in terms of latency and bandwidth. The 5G SBA facilitates the deployment and configuration of network functions (NFs). It enables the self-discovery of NFs available in the network and selects optimal functions based on the service requirements. When integrating 5G Network Slicing into an existing fixed infrastructure, we assume there are already existing network configurations in place such as predefined VLANs. However, the orchestrator of network slices in a 5G system requires the mapping of pre-configured settings in the mobile backhaul, e.g. VLANs with equivalent network resources in the 5G network such as slices and MEC functions.

Deployment of network slicing requires a network orchestrator to facilitate the mapping of slices with available transport network. A new network function is required to hide the complexity of 5G networks and to translate the existing transport network settings into slices and QoS parameters for the RAN. SDN is proposed to be the underlying technology for this new NF for interacting with the network switches to separate physical network resources for each slice and then request specific QoS settings from the RAN. Thus, the new NF named Mobile Backhaul Orchestrator (MBO) will manage the network resources to support network slicing and will request specific 5CI from the RAN. The overall description of the 5G and fixed LAN integration is described in the following Figure 7.2-1.

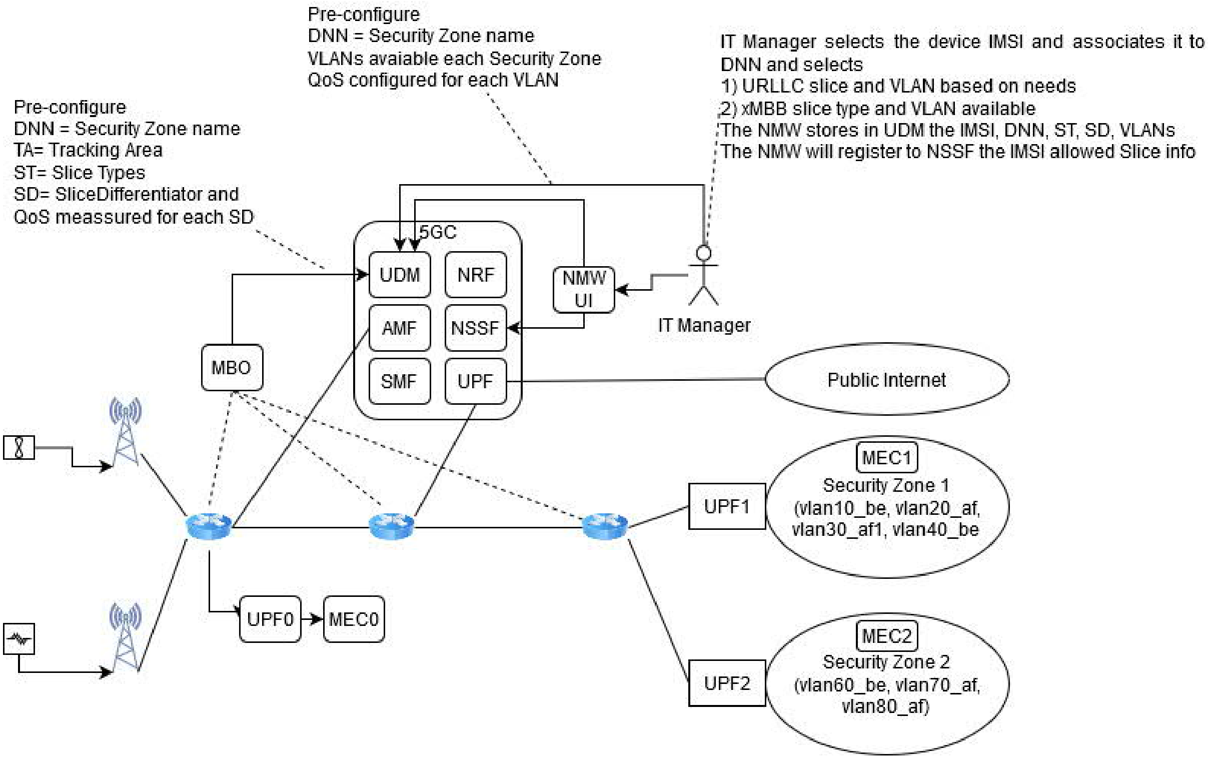

Pre-configure
DNN = Security Zone name
VLANs avaiable each Security Zone
QoS configured for each VLAN
IT Manager selects the device IMSI and associates it to DNN and selects
1) URLLC slice and VLAN based on needs
2) xMBB slice type and VLAN available
The NMW stores in UDM the IMSI, DNN, ST, SD, VLANs
The NMW will register to NSSF the IMSI allowed Slice info
Pre-configure
DNN = Security Zone name
TA= Tracking Area
ST= Slice Types
SD= SliceDifferentiator and
QoS meassured for each SD
5GC
UDM
NRF
AMF
NSSF
SMF
UPF
NMW UI
IT Manager
MBO
Public Internet
MEC1
Security Zone 1
(vlan10_be, vlan20_af, vlan30_af1, vlan40_be
UPF1
UPF0
MEC0
MEC2
Security Zone 2
(vlan60_be, vlan70_af, vlan80_af)
UPF2

Figure 7.2-1: Overall description of the 5G and fixed LAN integration.

8 Services and Applications for Public Safety

8.1 Mission Critical services

Mission critical (MC) services consists of Mission Critical Push To Talk (MCPTT) and Proximity Services (ProSe) services to MCPTT, Mission Critical Broadcast service, and MC Critical Data service. These services share common requirements on MC networks as defined in [3GPP-22280]. The MC services can either be supported in on-network cases or off-network cases. The MC services support both private as well as group communications.

8.1.1 Mission critical broadcast services

MC services support broadcast communication from authorized MCX Service Group Members. Different user broadcast groups can be created. This enables a user to be a part of several broadcast groups and get multiple receiving, called monitoring by some organizations. The broadcast information provides MCX users with the current information about police, fire or critical medical events that are occurring within their jurisdictions. This is useful for dispatchers or those that might not be the primary support for that event at that moment. The information gained by monitoring might be useful for the dispatcher to determine any actions to take or be useful later if the MCX User is deployed to provide additional support for that event. Late communication entry is supported.

8.1.2 Mission critical push-to-talk (MCPTT)

The MCPTT is intended to support communication between several users (in a group call), where each user can gain access to talk resembling the feeling of a walkie talkie, i.e. within a small communication group talk to everyone given the channel is free. MCPTT is described in 3GPP in several technical specifications, most notably: [3GPP-22179] and [3GPP-23379].

Public safety workers often operate in groups. For their communications public safety workers are organized in groups. People that are working together communicate in the same MCPTT Group, the group communication helping them to coordinate quickly. People with different tasks often communicate in separate MCPTT Groups.

To help the public safety worker to quickly find and select the correct MCPTT Group for the task, the MCPTT Groups in the radio are often structured in folders and/or accessible via key-shortcuts. An MCPTT Service provides Group Call and Private Call capabilities, which have various process flows, states and permissions associated with them. Furthermore, MCPTT Priority and QoS is situational. The MCPTT Service is intended to provide a real-time priority and QoS experience for MCPTT calls, as public safety users have significant dynamic operational conditions that determine their priority.

Private Calls allow two MCPTT Users to communicate directly with each other without the use of MCPTT Groups. Two commencement modes of Private Calls are supported: Manual Commencement Private Call and Automatic Commencement Private Call. Manual Commencement Private Calls mimic a telephone conversation where the called party receives a notification that they are being requested to join a Private Call, and the called party may accept, reject, or ignore the call request. Automatic Commencement Private Calls mimic the immediate setup and voice propagation of Group Call operation between two users where the calling party initiates an Automatic Commencement Private Call to another user and sends audio without any additional call setup delay beyond Group Calls.

The MCPTT Service while operating in off-network mode comprises a set or collection of functions necessary to provide MCPTT using a ProSe direct (UE-to-UE) communication path (ProSe direct communication path) for transport. The ProSe direct communication path does not traverse the network infrastructure. Figure 8.1-1: below follows the functional model for the application plane of the MCPTT service.

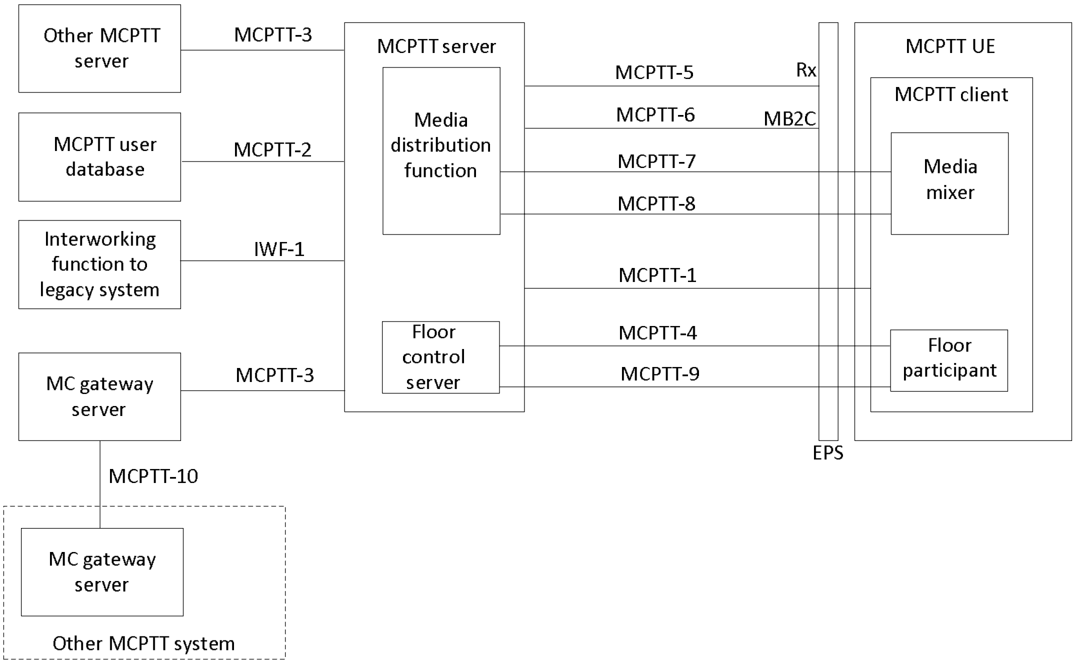


Figure 8.1-1: Functional model for the application plane of the MCPTT service [3GPP-23379].

The MCPTT service can also be provided in off-network operation conditions. Below Figure 8.1-2 shows the functional model for such a situation:

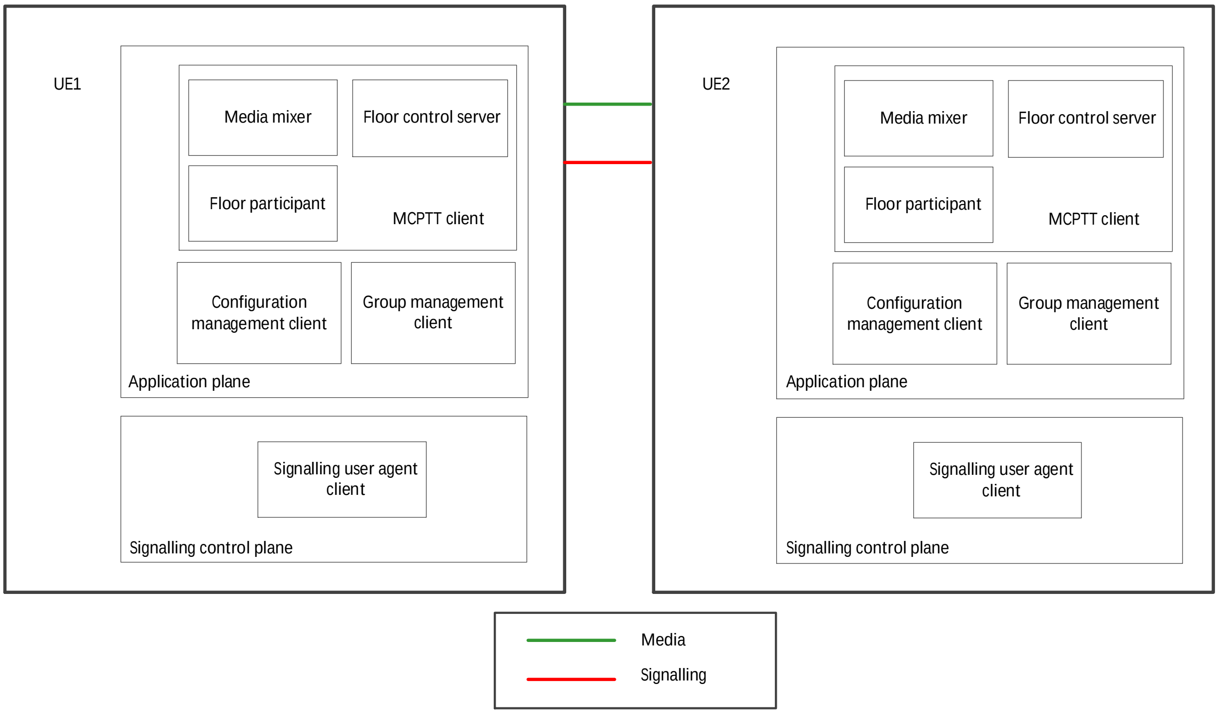


Figure 8.1-2: Functional model for off-network operation of MCPTT service [3GPP-23379].

8.1.3 *Proximity Services*

Proximity Services (ProSe) are services that can be provided by the 3GPP system based on UEs being in proximity to each other [3GPP-23303]. As mentioned on Section 8.1.1, the MCPTT Service while operating in off-network mode uses a ProSe direct (UE-to-UE) Communication path (ProSe direct communication path) for transport.

Figure 8.1-3: below follows a figure showing the non-roaming architectural model for ProSe services.

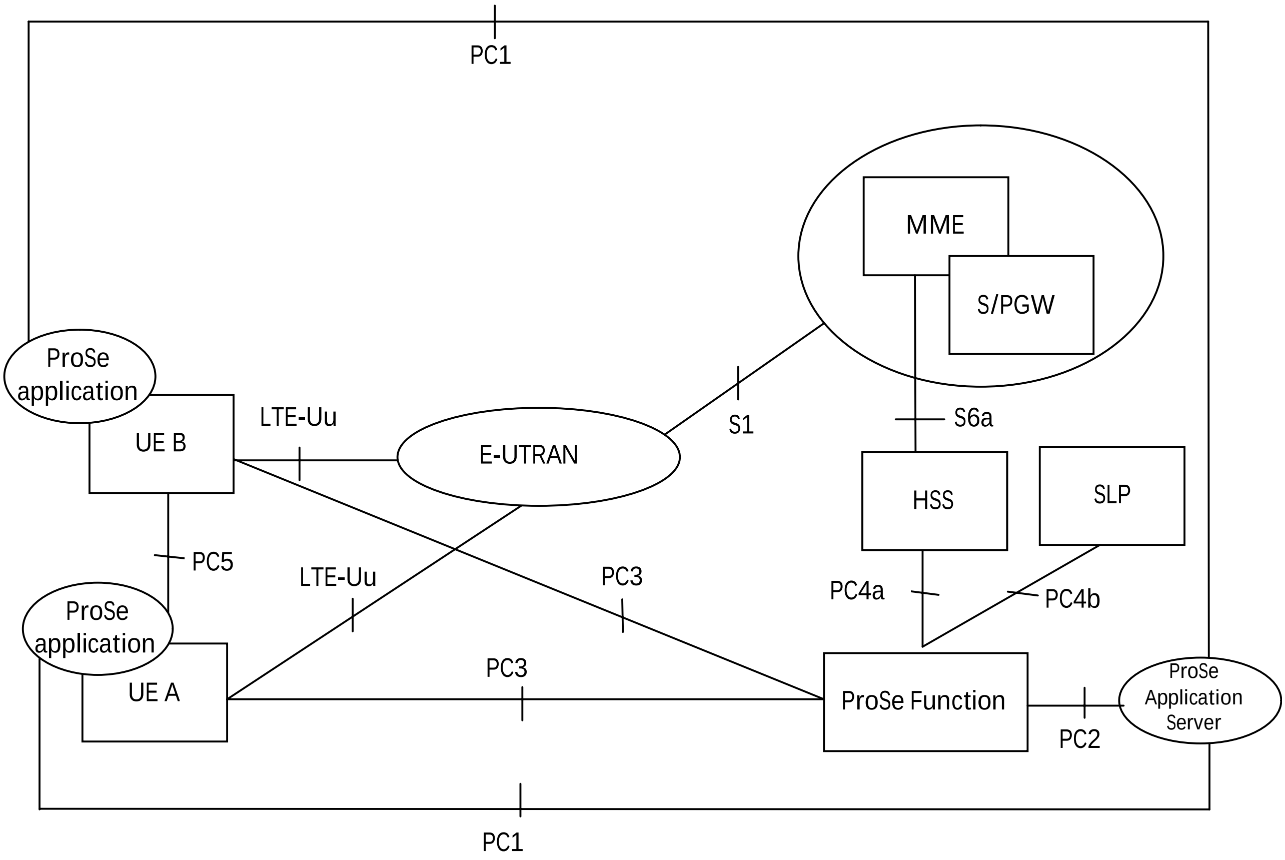


Figure 8.1-3: The non-roaming architectural model for ProSe services.

The ProSe Function is the logical function that is used for network related actions required for ProSe. Furthermore, the ProSe Function plays different roles for each of the features of ProSe. On each UE a ProSE application exist which enables communication between the UEs through PC5.

The ProSe Function consists of three main sub-functions that perform different roles depending on the ProSe feature, namely:

1. Direct Provisioning Function (DPF) is used to provision the UE with necessary parameters in order use ProSe Direct Discovery and Prose Direct Communication.
2. Direct Discovery Name Management Function is used for open Prose Direct Discovery to allocate and process the mapping of ProSe Applications IDs and ProSe Application Codes used in ProSe Direct Discovery. It uses ProSe related subscriber data stored in HSS for authorisation for each discovery request. It also provides the UE with the necessary security material in order to protect discovery messages transmitted over the air. In restricted ProSe Direct Discovery, it also interacts with the Application Server via PC2 reference points for the authorization of the discovery requests.

3. EPC-level Discovery ProSe Function has a reference point towards the Application Server (PC2), towards other ProSe Functions (PC6), towards the HSS (PC4a) and the UE (PC3).

The user's profile in the HSS contains the subscription information to give the user permission to use ProSe.

Any ProSe-enabled UE may support the following functions: Exchange of ProSe control information between ProSe-enabled UE and the ProSe Function over PC3 reference point and procedures for open and restricted ProSe Direct Discovery of other ProSe-enabled UEs over PC5 reference point.

A Proximity request is initiated by UE A and activates procedures towards ProSe function, HSS and towards UE B to set up the proximity service. The ProSe UE-to-Network Relay entity provides the functionality to support unicast traffic (UL and DL) connectivity to the network for Remote UEs for Public Safety (see Figure 8.1-4:).

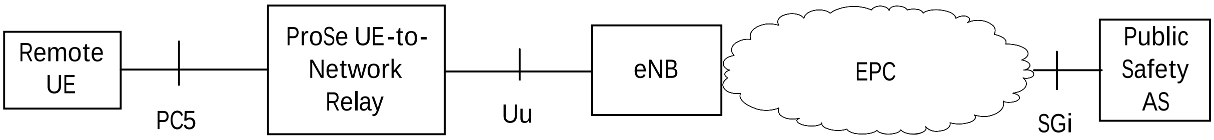


Figure 8.1-4: Architecture model using a ProSe UE-to-Network Relay for Public Safety.

8.1.4 Mission Critical Data

MCData services suite consists of the following sub-services: short data service (SDS) and file distribution (FD). MCData is described in [3GPP-22281] and [3GPP-23282].

The SDS feature of the MCData Service could be considered as a basic protocol carrying a limited size, but variable content, payload message. This message could be text or could be marked for extensible purposes including short binary messages for application communication. Messaging could be one-to-one messaging or could be group messaging.

File distribution and data streaming are fundamental capabilities of the MCData Service. File distribution can be used to provide a standalone file transfer capability or can be invoked by a controlling application to support the purpose of the application. Whereas data streaming can be used to provide a standalone data streaming capability or can be invoked by a controlling application to support the purpose of the application. IP connectivity can be used for MCData applications that are based on the IP client-server paradigm. The UE can contain a client using a service in the network.

Robots as defined in [3GPP-22281] will be used more and more to provide unique services to mission critical organizations. Consequently, critical communications users need a common communication framework for robots which can take advantage of different transport technologies such as a 3GPP system. The MCData Service, working in conjunction with existing robot control capabilities, will provide mechanisms to do that.

Figure 8.1-5: below shows the Generic application plane functional model for MCData.

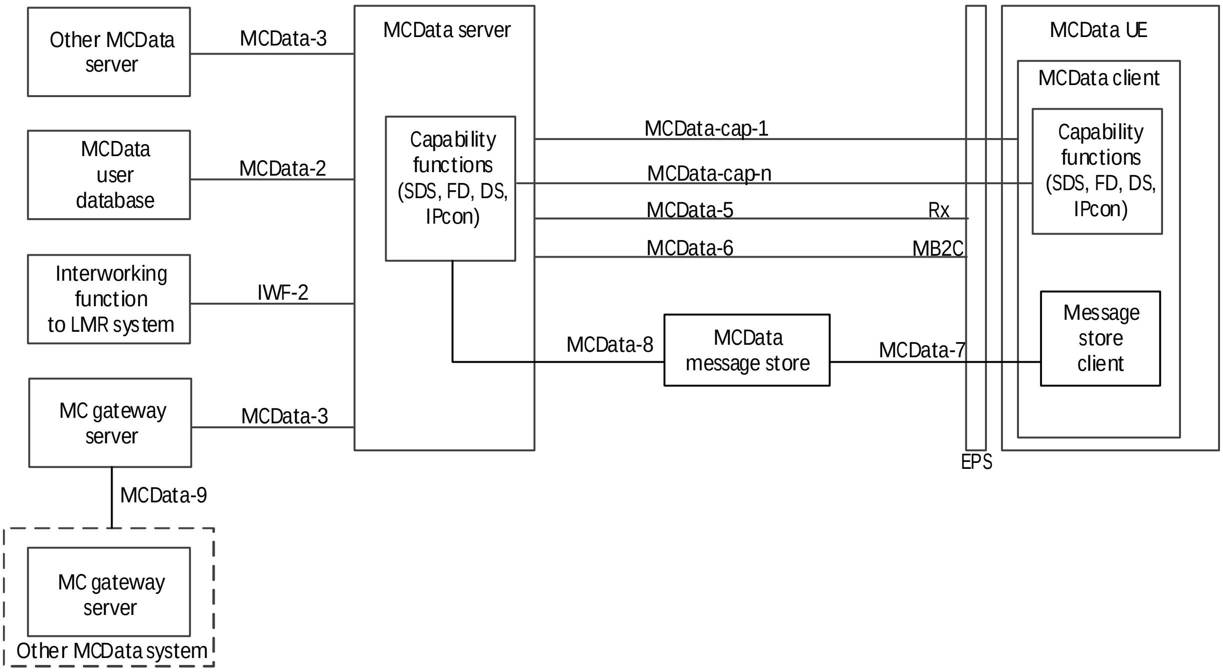


Figure 8.1-5: Generic application plane functional model for MCData, [3GPP-23282].

8.1.5 Mission Critical Video

The MCVideo service supports video media communication between several users (i.e. group call), where each user has the ability to gain access to the permission to stream video in an arbitrated manner. The MCVideo service also supports private calls between two users. For further 3GPP reference, see [3GPP-22281]. Figure 8.1-6: below depicts the functional model for application plane of MCVideo service.

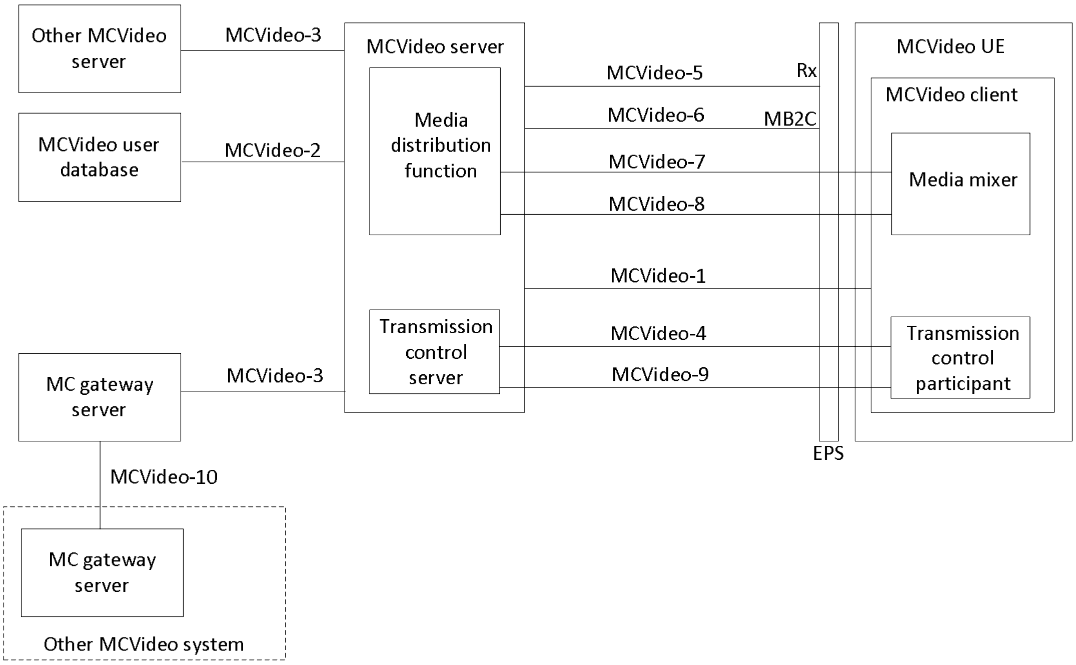


Figure 8.1-6: Functional model for application plane of MCVideo service [3GPP-22281].

8.2 Immersive applications using extended reality

Immersive technologies are a core enabler - assimilating massive amount of multi-dimensional data created by the growth of connected products and the IoT, which extends the reality users experience by blending the virtual and real worlds. Extended reality (XR) is an emerging umbrella term for all the immersive technologies as shown in Figure 8.2-1, that encompasses all virtual or combined real-virtual environment and human/machine interaction compounds including augmented reality (AR)[1], virtual reality (VR)[2] and mixed reality (MR)[3], embraced through a wide scope of applications including critical service for society.

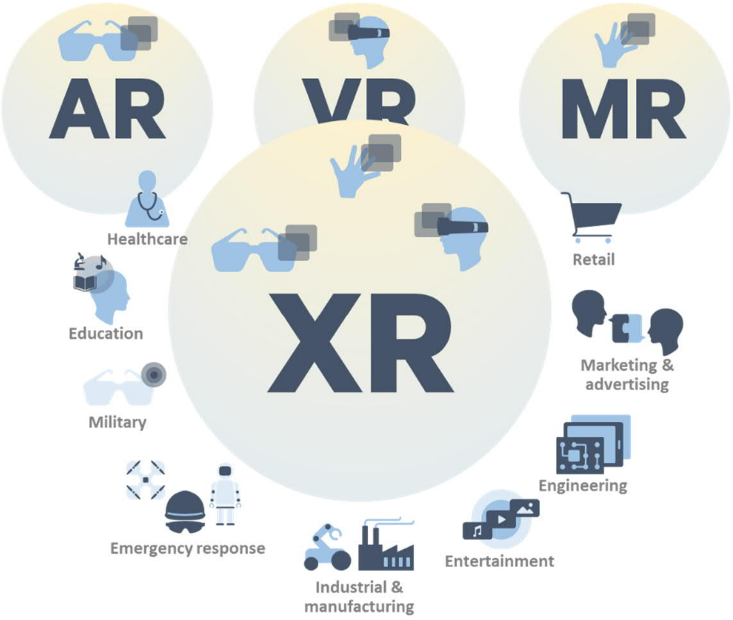


Figure 8.2-1: Different Types of Realities and Some Applications [3GPP-26928].

While there are many potential XR application scenarios in various areas with commercial value, such as manufacturing, education, remote work, retail, marketing, real estate, and so on, public safety is one important area of focus in today's society, where XR will bring significant values by providing immersive user experience for emergency preparedness and response services.

- Preparedness with first responder training: XR technology allows first responders to gain access to a set of risky training scenarios [EON20], helping them better prepare for accurate rescue decisions under pressure and serving as a great psychological and emotional aid without having to be in danger. In training process, safety professionals immerse themselves with various threats in virtual forms. Such an immersive training service allows firefighters to execute virtual evacuation drills with flames and smokes during fires [FLA20], border officers to

[1] In AR, users are not isolated from the real world and can still interact and see what is going on in front of them.

[2] In VR, users are fully immersed in a simulated digital environment.

[3] In MR, digital and real-world objects co-exist and can interact with one another in real-time.

discipline how to respond to the terrorist threat in a cruise ship terminal [AUG20], and police officers to visit hostage situations such as street gun battles [GOV20].

- Response with situational awareness: XR enables to offer clear views or insights of catastrophic situations by overlaying vital information of inaccessible and/or low-visibility areas on top of geographic information system (GIS) map, to on-the-ground responders for mitigating potentially tragic consequences [MCM+11]. For example, integrating UAV sourced information to 3D GIS map enables to keep track of to where rapid-fire evolution goes with displaying data in augmented reality, including street names, individual responder locations and key points of interest [EDG20]. And safety professionals equipped with the mission gear such as AR glasses, binaural microphones/headphones, a 360-degree helmet/drone-mounted infrared/thermal camera and other sensors, enable to navigate burning space to rescue victims by looking through smoke, toxic gases and darkness [SBC+19], [QWA20].

XR-powered immersive experience can be fulfilled by satisfying XR device extreme requirements [QAL16] across visual quality, sound quality, and intuitive interactions. Making the XR experience truly immersive is enabled by meeting E2E traffic connectivity requirements such as bitrates, latency, and reliability, depending on the XR architecture [3GPP-26928]. For example, XR distributed computing architecture utilizes resource of XR edge server for computation workload and takes advantage of the low latency 5G connection and flexible requirements of distributed environment, while other architectures such as device-based one impose less stringent requirements. In the sense, edge computing is the inevitably key component to enable the immersive service for public safety.

9 Relation towards PriMO-5G demonstrators

The PriMO-5G project aims to demonstrate an E2E 5G system providing immersive video services for moving objects. This is achieved by both local and cross-continental testbeds that integrate radio access and core networks developed by different project partners to showcase E2E operations of envisaged use cases, specifically those related to firefighting.

To that end, the experimentation activities in PriMO-5G are conducted in multiple phases. In the initial phase, the focus is on testing and demonstrating the key enabling radio, edge and core network components and applications. The primary goal of the component testing and demonstration activities is to provide insights on the capabilities of these components from the perspective of the E2E operations of the PriMO-5G use cases. Moreover, these activities enable the consortium members to identify and/or enhance the components needed for the subsequent system integration phase. The system integration phase is envisioned to occur over selected partner sites on European and Korean sides. Finally, in the third phase the testbeds on the European and Korean sides will be interconnected to demonstrate global applicability and feasibility of E2E operations of PriMO-5G use cases.

This section describes the links between the different demonstrators and the PriMO-5G architecture outlined in this report. Detailed information about the PriMO-5G demos can be found in [PRI19-D52].

9.1 Component demonstrators

9.1.1 Demos from European partners

This section describes the component-level demonstration activities from the European partners. Altogether, there are 8 demonstrations as summarized in Table 9.1-1.

Table 9.1-1: Component demos from European partners.

Component demo	Brief description (including ownership)	Relation to PriMO-5G architecture
Cell-free UDN **(AALTO)**	The demonstrator studies how aerial and ground UEs are served by an ultra-dense user-centric network. Contrary to traditional architectures, user-centric networks do not require UEs to periodically interrupt data transmissions in order to perform neighbouring BS measurements, which are then reported to the serving BS. Instead, the network estimates the physical position of each UE. Positioning relies on the UEs transmitting beacons for AoA estimation.	Directly links to the cell free architecture described in Section 4.3
Real-time video broadcast **(CMC)**	Demonstrates the streaming of live video from UAVs to mobile devices located in the same service area. The demo consists of UAV that streams live video uplink to the mobile packet core from where the video can be broadcasted downlink through the eMBMS system to all the mobile devices to receive live video and monitor the progress of the emergency situation. Current demo uses 4G LTE radio for the uplink video streaming and 4G LTE radio for the downlink broadcast since eMBMS service specified in 5GC is still based on the 3GPP specifications Release 13 that uses 4G LTE radio access.	The eMBMS demonstrated here (and future 5G broadcast) provides the foundations to the Mission Critical Broadcast services described in Section 8.1.1.
UAV-UE video broadcast **(CMC)**	This demo uses similar equipment as the real-time video broadcast (described above), but streaming is not real time. Instead this demo uses the new xMB interface defined in 3GPP that allow broadcasters to schedule the session, select the service area and the content for broadcasting.	(same as above)
5G network slicing	The network slicing demo was performed for separating traffic from NB-IoT devices. The motivation for using network slicing	The new slicing supports the

Component demo	Brief description (including ownership)	Relation to PriMO-5G architecture
(CMC)	in specific cases like IoT sensors is because traffic would be unpredictable due to burst communication from large number of sensors. Currently there is no comprehensive solution that allows dynamically manage network slicing in mobile networks. In this demo we deployed network slicing based on the integration of SDN in the mobile backhaul. SDN will be integrated in the network edge to deliver network slicing and isolate traffic from various applications and allows sharing existing mobile infrastructure with different service providers.	deployment of XR/VR in different locations to improve response time.
MEC orchestrator **(CMC)**	Demonstrate the usage of Service Based Architecture (SBA) to deploy different network functions in separate locations if they are connected through IP network. In this demo we deploy different instances of the UPF network function in different location.	(same as above)
5G-NR transceiver **(NI)**	Demonstrate the vertical stack integration of NI's FPGA-based 5G NR PHY layer with the 3rd party 5G NR protocol layer. Testing the vertically integrated NI 5G NR UE stack against a gNB emulator is the main idea of this component demo. The gNB emulator consists of a real-time playback of 5G NR DL signals with pre-generated content. Those 5G NR DL signals are received, demodulated and decoded in real-time by the NI 5G NR UE stack, which has been configured via 3GPP-compliant RRC test vector injection.	The component demo is focusing on the integration of components for the radio access network and therefore relates to the RAN part of the overall PriMO-5G architecture in Section 2. As it is drawn towards high-data rate communication for video streaming, the application is related to the AR/VR usage described in Section 8.
Cross-Domain **(KCL)**	The demo consists of CMC 5G core located in Finland and gNB located in KCL in the UK. The gNB is based on Open Air Interface (OAI) software. running on a Linux PC at KCL, supports NSA functionality with a full software implementation of LTE standards. The gNB is configured for FDD band 7.	Demonstrate a scenario where mobile operator deploys a remote network to control a connection. For example, smart firefighting with UAVs in urban area where the mobile operator has coverage and signalling is handled in remote server. This is related to the RAN part of the overall PriMO-5G architecture in Section 2.
Optimal routing **(EAB)**	Demo aims to visualize and demonstrate the feasibility of Optimal Routing. It is achieved by implementing the main functional entities in the concept, that is: IAP (IP Announcement Point) and the LR (Location Register. The IAP in the user plane receives plain IP packets from the Data Networks. and determines which UPF the packet has to be	Directly links to the optimal routing architecture described in Section 5.7.

Component demo	Brief description (including ownership)	Relation to PriMO-5G architecture
	sent to, either by sending a query to the LR, or looking up from its local cache. The LR stores the UE IP address -> UPF mappings. These entries are inserted, updated and deleted by the Session Management Function (SMF).	

9.1.2 Demos from Korean partners

The component-level demonstration activities from the Korean partners are depicted in this section. Altogether, there are 4 demonstrations as summarized in Table 9.1-2.

Table 9.1-2: Component demos from Korean partners.

Component demo	Brief description (including ownership)	Relation to PriMO-5G architecture
Aerial video streaming system with real-time object detection and super-resolution **(YU, KU, KT)**	This demo is a representative implementation of urban firefighting. It constitutes components including YU's UAVs and ground control station, KT's 5G gNB and core, and KU's vehicle capable of edge computing. In this demo, to advance the research and implementation of 5G system using UAVs, the PriMO-5G team implements an E2E system that showcases seamless real-time streaming of immersive media through 5G and post-processing using AI techniques for object detection and super resolution.	The demo has multiple links including edge-assisted AI applications of Section 6.4 and immersive applications of Section 8.2.
Streaming aerial video system through LTE NIB **(EUC)**	This demo is similar to the aerial video streaming demo described above, that is it also demonstrates the streaming of live video from UAVs to mobile devices. But in this demo the video streaming is served by MC (Mission Critical) services defined in 3GPP Rel.13. For this, NIB includes LTE eNB, EPC, and MC server.	This demo is linked to the moving base stations Section 4.5 and mission critical broadcast services of Section 8.1.1
Lens based mmWave communications **(YU)**	The demo introduces the hybrid beamforming system with a lens antenna so that the fast beam switching method show lower complexity compared to the existing method using phased array. Also, both curved and planar lenses are considered, and hybrid beamforming is performed. All of these are implemented in the NI hardware and have high throughput, and high received power in mind. The system operating frequency will support wide bandwidth of up to 800MHz at 28.5 GHz.	Directly linked to the Lens-based mmWave Communiations in Section 4.6.
Haptic communications **(YU)**	Demonstration for testing a proof of concept of haptic communications. In firefighting scenario, remote-controlled robots are used to help put out fires and rescue people. The haptic communications enable the control centre to control UEs' haptic equipment remotely with low latency and high reliability.	This demo is directly linked to Haptic Equipment in Section 3.4.

9.2 Local and Intra-continental demos

There are several activities of system integration in the local and intra-continental level. Table 9.2-1 in this section summarizes the ongoing and planned demos.

Table 9.2-1: Local and intra-continental system integration demos.

Component demo (ownership)	Brief description (including ownership)	Relation to PriMO-5G architecture
CMC-NI system demos **(CMC, NI)**	The objective of the overall integration and demo efforts of CMC and NI is to showcase the integration of the CMC 5G core towards the mmWave capable 5G gNB from NI. A first demo/test showcases the integration of the CMC 5G core with the NI 5G gNB. In this test CMC 5G core is running in NI premises with NI gNB running locally in Germany. The objective is to test the interoperability between a 5G core and NI mmWave gNB with PAL. A second separate demo showcases the integration of the NI 26 GHz mmWave antenna arrays on NIs 5G gNB and UE.	This supports the usage of mmWave links for local deployment with demand of high bandwidth for XR/VR and video streaming.
KCL-CMC system demos **(KCL, CMC)**	In this demo CMC 5G core is running in KCL in the UK and AALTO gNB running in Finland. The objective is to test an architecture where network functions are deployed in different locations. KCL and Aalto are connected using a remote connectivity through VPN. The setup consists of CMC 5G with NSA and SA functionalities running on a Linux PC at KCL, which is connected to the Aalto gNB manufactured by Nokia and 5G UE.	This case supports the network slicing with intercontinental deployment of applications in local and remote sites.
YU-KT-KU system demos **(YU, KT, KU)**	This system demo is a continuation work of the aerial video streaming demo from Table 9.1-2. The developments are targeted in the demonstrations: • Immersive video streaming using video captured from the UAV's camera that is then stitched in real-time to produce and encode the raw videos into a single 4K video. • Real-time object detection which investigates two well-known tradeoffs: computation vs. communication and exploration vs. exploitation. • AI-based image post-processing (to be completed)	The demo has multiple links including edge-assisted AI applications of Section 6.4 and immersive applications of Section 8.2.
EUC-KT-YU system demos **(EUC, KT, YU)**	In the demo 5G NR is used as a wireless backhaul of 4G LTE so that the portable base station provides the public safety service even without wired backhaul. In this test, we will try two different scenarios. First scenario is to locate eNB in EUC portable base station, and then eNB communicates with the KT core network components at remote site through 5G backhaul. Second scenario is to locate eNB,core network and MCPTT server components in EUC portable base station.	The first scenario of this demo is linked to the moving base stations Section 4.5 and mission critical broadcast services of Section 8.1.1, whereas, the second scenario without backhauling provides a precursor for standalone NPN and IOPS in Sections 5.5.1 and 5.6, respectively.

9.3 PriMO-5G intercontinental demos

9.3.1 Description of Finland-KR end-to-end system demo

The intercontinental system demos aim at demonstrating the flexibility of 5G Service Based Architecture (SBA) that allow having different network slices where network functions can be running in different interconnected 5G networks (in this case the 5G testbeds in Finland and Korea). As the 5G SBA suggests, a public safety mission would require the architecture to be constructed in a way that is most fitting to achieve a success in its own criteria, such as, maximum public protection, fast recovery, and so on. Accordingly, an E2E network slice should be constructed and allocated to meet the requirements of a specific mission. To that end, in PriMO-5G intercontinental firefighting scenarios, a slice could be allocated to route traffic from UAV to a local MEC platform for processing video collected by the UAVs during firefighting event in an affected country. Furthermore, an intercontinental slice between 5G networks in assisting and affected countries, would allow for support staff at Emergency Operations Centers (EOC) of assisting country to have richer and current situation picture through immersive video transported over the slice.

9.3.2 Relation to PriMO-5G architecture

The intercontinental demos leverage the E2E network slicing approaches described in Section 2.3.3 and Section 5.2 to demonstrate of flexibility and shared use of network functions by their instantiation across two 5GC deployments (in Europe and Korea). This allows for creation of network slices with convenient discovery and placement of network functions (e.g. UPF, MEC) closer to scene of disaster in affected country even as the other network functions are placed in an assisting country in another continent.

10 Conclusions

10.1 Summary of the document

The main aim of the PriMO-5G project is to demonstrate an E2E 5G system providing immersive video services for moving objects. This deliverable has described an overall system architecture to fulfil the aim of the PriMO-5G project. The deliverable started with the summary of PriMO-5G scenarios and use cases in Section 1. As previously presented [PRI19-D11] and [SMJ+19], PriMO-5G chose the public safety, particularly firefighting, as the main use case because it is an area where immersive video services with moving objects can make a substantial improvement in the safety and efficiency of the operations. In Section 2, the overview of E2E architecture has been provided with the highlights of cellular network with aerials, edge computing, and network slicing. Then, main body of the deliverable, i.e. between Section 3 and Section 8, introduced each component of the overall architecture. Section 3 discussed important types of user equipment that can be connected to a public safety communication network. Section 4 covered RAN elements that can be operated and supportive to a wireless public safety communications system, while Section 5 covered different CN functions needed in a public safety system. Edge architecture and the functions to provide edge computing appeared in Section 6. In Section 7, AI-enabled network management and orchestration were touched upon. Section 8 detailed currently defined mission critical services in 3GPP as well as opportunities for immersive applications using extended reality. Finally, Section 9 explained the relations of the overall system architecture with the PriMO-5G demonstration activities.

10.2 Potential Enhancements of 3GPP Architecture

This deliverable has presented the overall system architecture to fulfil the requirements for delivering a reliable and innovative public safety application. In this section, we collect a set of potential innovation and enhancements that would require standardisation to ensure multi-vendor interoperability. Also, we list the identified areas that cover the whole architecture, which consists of UE, RAN, core network, MEC, network management and orchestration.

One area for innovation and standardisation is the integration of network slicing into mobile networks. Network slicing and MEC are to deliver an integrated solution that benefits from both technologies. 3GPP has defined network slicing to allocate network resources to independent instance of the network allocated for selected UEs or applications. On the other hand, MEC is defined to run application processing closer to the UE. Both network slicing and MEC can be deployed separately or they can be integrated together to support more demanding use cases. The integration of network slicing together with MEC requires the following enhancements:

- Association of MEC infrastructure with UPF that will handle the user plane.
- Discovery of the location for MEC and UPF that will deliver low latency.
- Integration of network selection with MEC to deliver both network resources and MEC for computational resources.

There are many opportunities to enhance artificial intelligence support for 3GPP. Standardization is ongoing to advance the NWDAF and MDAF functions. One potential area to standardize would be to look at information needs for standard procedures and requirements for AI processing: potentially starting with FCAPS. For these procedures, AI frameworks including 3GPP support for measurement objects and data processing could be advanced in standardization. Information objects could be defined connected to the procedures. 3GPP functions could then be identified that supports the information objects. Necessary measurement objects connected to the 3GPP functions could then in the next step be defined. The measurement objects could be hierarchical and some intermediate processing could also be suggested. This could support real-time management of the 3GPP system. So, in summary one can follow the following logic: procedure information needs-> information objects -> 3GPP functions -> measurement objects. For major external services, a similar approach could be done to specify exposure support. There is also an opportunity to define AI support for direct 3GPP functions. E.g. scheduling or beamforming control.

IAB offers an opportunity to build coverage in a hierarchical manner. There is an opportunity to advance and standardize functions for mobile IABs. This is applicable for both aerial IABs as well as IABs based on cell on wheels.

The Optimal Routing solution presented in Section 5.7 of this document improves the 3GPP architecture by allowing low latency UE-to-UE and UE-to-server communication, even after mobility events, without breaking session continuity. The solution is based on decoupling the IP anchor from the UPF and on allowing change of UPF for the PDU session without IP address change of the UE's PDU session. Control plane mechanisms are introduced to keep track which UPF serves the session, and to update user plane forwarding. By avoiding changing the IP address, even at UPF changes, session continuity is kept, and service continuity is maintained. Changing the UPF without (UE) IP address change potentially opens up easier relocation of application servers located in edge sites.

In the proposed E2E system, UAVs could act as either end. As the gatherer of situational information granting situation awareness to other entities, UAVs act as primary servers responsible for transmitting situational information. On the other hand, when another entity controls their mission executions or movements, UAVs act as end receivers for orders. Thus, UAVs may require different 5QI depending on the flow and bandwidth latency. A single UAV may have one flow uplink that requires high bandwidth and second flow downlink for control commands that requires a different 5QI. Also, a UAV may request multiple PDU where 5QI is for uplink or downlink only.

References

[3GPP-172290] 3GPP Work Item Description RP-172290, Study on Integrated Access and Backhaul for NR, 2017.

[3GPP-191584] 3GPP Work Item Description RP-191584, Enhanced Industrial Internet of Things (IoT) and URLLC, 2019.

[3GPP-193233] 3GPP Work Item Description RP-193233, Physical layer enhancements for NR ultra-reliable and low latency communication (URLLC), 2019.

[3GPP-22179] 3GPP Technical Specification 22.179, Mission Critical Push to Talk (MCPTT); Stage 1, Release 13.

[3GPP-22261] 3GPP Technical Specification 22.261, Service requirements for the 5G system, Release 15, 2016.

[3GPP-22280] 3GPP Technical Specification 22.280, Mission Critical (MC) services common requirements, Release 14, 2016.

[3GPP-22281] 3GPP Technical Specification 22.281, Mission Critical (MC) video, Release 14, 2016.

[3GPP-22282] 3GPP Technical Specification 22.282, Mission Critical (MC) data, Release 14, 2016.

[3GPP-22346] 3GPP Technical Specification 22.346, Isolated Evolved Universal Terrestrial Radio Access Network (E-UTRAN) operation for public safety; Stage 1, Release 13, 2015.

[3GPP-23180] 3GPP Technical Specification 23.180, Mission Critical (MC) services support in the Isolated Operation for Public Safety (IOPS) mode of operation, Release 17, 2019.

[3GPP-23228] 3GPP Technical Specification 23.228, Architecture enhancements for 5G System (5GS) to support network data analytics services, Release 16, 2018.

[3GPP-23282] 3GPP Technical Specification 23.282, Functional architecture and information flows to support Mission Critical Data (MCData); Stage 2, Release 14, 2016.

[3GPP-23303] 3GPP Technical Specification 23.303, Proximity-based services (ProSe); Stage 2, Release 12, 2016.

[3GPP-23379] 3GPP Technical Specification 23.379, Functional architecture and information flows to support Mission Critical Push To Talk (MCPTT); Stage 2, Release 14, 2016.

[3GPP-23401] 3GPP Technical Specification 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, Release 8.

[3GPP-23501] 3GPP Technical Specification 23.501, System architecture for the 5G System (5GS), Release 15, 2016.

[3GPP-23558] 3GPP Technical Specification 23.558, Architecture for enabling Edge Applications (EA), Release 17, 2019.

[3GPP-23748] 3GPP Technical Report 23.748, Study on enhancement of support for Edge Computing in 5G Core network (5GC), Release 17, 2019.

[3GPP-23758] 3GPP Technical Report 23.758, Study on application architecture for enabling Edge Applications, Release 17, 2019.

[3GPP-23791] 3GPP Technical Report 23.791, Study of enablers for Network Automation for 5G, Release 16, 2017.

[3GPP-26928] 3GPP Technical Report 26.928, Extended Reality (XR) in 5G, Release 16, 2018.

[3GPP-28530] 3GPP Technical Specification 28.530, Management and orchestration; Concepts, use cases and requirements, Release 15, 2017.

[3GPP-28533] 3GPP Technical Specification 28.533, Management and orchestration; Architecture framework, Release 15, 2018.

[3GPP-28535] 3GPP Technical Specification 28.535, Management services for communication service assurance; Requirements, Release 16, 2019.

[3GPP-28809] 3GPP Technical Report 28.809, Study on enhancement of management data analytics, Release 16, 2019.

[3GPP-36777] 3GPP Technical Report 36,777, Enhanced LTE support for aerial vehicles, Release 15, 2017.

[3GPP-37324] 3GPP Technical Specification 37.324, Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Service Data Adaptation Protocol (SDAP) specification, 2017.

[3GPP-38300] 3GPP Technical Specification 38.300, NR; Overall description; Stage-2, Release 15, 2017.

[3GPP-38340] 3GPP Technical Specification 38.340, NR; Backhaul Adaptation Protocol (BAP) specification, Release 16, 2019.

[3GPP-38801] 3GPP Technical Report 38.801, Study on new radio access technology: Radio access architecture and interfaces, Release 14, 2016.

[3GPP-38824] 3GPP Technical Report 38.824, Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC), Release 16, 2018.

[3GPP-38840] 3GPP Technical Specification 38.840, NR; Backhaul Adaptation Protocol (BAP) specification, Release 16, 2018.

[3GPP-38874] 3GPP Technical Report 38.874, NR; Study on integrated access and backhaul, Release 15, 2017.

[ACIA19] 5G-ACIA, 5G Non-Public Networks for Industrial Scenarios, July 2019, https://www.5g-acia.org/publications/5g-non-public-networks-for-industrial-scenarios-white-paper/

[AKL14] A. Al-Hourani, S. Kandeepan, and S. Lardner, “Optimal lap altitude for maximum coverage,” IEEE Wireless Communications Letters, vol. 3, no. 6, pp. 569–572, Dec 2014.

[AUG20] http://www.auggmed-project.eu/

[CPRI13] Common Public Radio Interface (CPRI); Interface Specification v6.0, available at http://www.cpri.info/downloads/CPRI_v_6_0_2013-08-30.pdf

[CRAN11] C-RAN The Road Towards Green RAN, Ver 2.5 White Paper, Oct 2011, https://pdfs.semanticscholar.org/eaa3/ca62c9d5653e4f2318aed9ddb8992a505d3c.pdf

[EDG20] https://edgybees.com/

[EON20] https://eonreality.com/

[ETSI16] ETSI GS MEC 002 V1.1.1, Mobile Edge Computing (MEC); Technical Requirements, 2016, https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/01.01.01_60/gs_MEC002v010101p.pdf

[ETSI18a] ETSI White Paper No. 24, MEC Deployments in 4G and Evolution Towards 5G, February 2018, https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp24_MEC_deployment_in_4G_5G_FINAL.pdf

[ETSI18b] ETSI White Paper No. 28, MEC in 5G networks, June 2018, https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf

[ETSI20] ETSI MEC Spécifications, available at https://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing

[FLA20] https://www.flaimsystems.com/

[GOV20] https://www.govred.com/

[GSMA20] GSMA White Paper, An Introduction to Network Slicing, Jan. 2020, https://www.gsma.com/futurenetworks/resources/an-introduction-to-network-slicing-2/

[I95] T. K. Ishii, ed.,‘Handbook of microwave technology,’ Elsevier, 1995

[LCG+20] M. Li, N. Cheng, J. Gao, Y. Wang, L. Zhao and X. Shen, "Energy-Efficient UAV-Assisted Mobile Edge Computing: Resource Allocation and Trajectory Optimization," in IEEE Transactions on Vehicular Technology, vol. 69, no. 3, pp. 3424-3438, March 2020. https://ieeexplore.ieee.org/document/8964328

[LCW19] C. Lai, C. Chen, and L. Wang, “On-demand density-aware UAV base station 3d placement for arbitrarily distributed users with guaranteed data rates,” IEEE Wireless Communications Letters, vol. 8, no. 3, pp. 913–916, June 2019.

[LZZ+17] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, “Placement optimization of UAV-mounted mobile base stations,” IEEE Communications Letters, vol. 21, no. 3, pp. 604–607, March 2017.

[MCM+11] L. Merino, et al., “Automatic forest fire monitoring and measurement using unmanned aerial vehicles." VI International Conference on Forest Fire Research. 2010.

[MDB+16] P. Marsch, et al., 5G Radio Access Network Architecture: Design Guidelines and Key Considerations, IEEE Communications Magazine Vol. 54,No. 11, 2016.

[NGMN16] NGMN White paper, Description of Network Slicing Concept, Jan. 2016, https://www.ngmn.org/wp-content/uploads/160113_NGMN_Network_Slicing_v1_0.pdf

[NGMN18] NGMN White paper, NGMN Overview on 5G RAN Functional Decomposition, Feb. 2018, https://www.ngmn.org/publications/ngmn-overview-on-5g-ran-functional-decomposition.html

[NGMN19] NGMN White paper, 5G E2E Technology to Support Verticals URLLC Requirements, November 2019, https://www.ngmn.org/wp-content/uploads/Publications/2019/200210-NGMN_Verticals_URLLC_Requirements_v16.pdf

[PKP+19] O. Park, et al., “Technical Trends of Ultra-Reliable Low-Latency Communication for 5G”, ETRI Electronics and Telecommunications Trends, Vol. 34, No. 6, 2019, DOI: 10.22648/ETRI.2019.J.340604

[PRI19-D11] PriMO-5G Deliverable D1.1, PriMO-5G use case scenarios, February 2019, https://primo-5g.eu/download/357/

[PRI19-D21] PriMO-5G Deliverable D2.1, Initial design of MEC and Network Slice Manager, April 2019, https://primo-5g.eu/download/389/

[PRI19-D52] PriMO-5G Deliverable D2.1, Intermediate report – component demonstrations and system integration plans, May 2019.

[QAL16] Qualcomm, Making immersive virtual reality possible in mobile, March 2016, https://www.qualcomm.com/media/documents/files/whitepaper-making-immersive-virtual-reality-possible-in-mobile.pdf

[QWA20] https://www.qwake.tech/

[RZH+19] Ren, J., Zhang, D., He, S., Zhang, Y. and Li, T., 2019. A Survey on End-Edge-Cloud Orchestrated Network Computing Paradigms: Transparent Computing, Mobile Edge Computing, Fog Computing, and Cloudlet. ACM Computing Surveys (CSUR), 52(6), pp.1-36, 2019. https://dl.acm.org/doi/abs/10.1145/3362031

[SBC+19] S. Sharma, et al., "Emergency Response Using HoloLens for Building Evacuation”, vol 11574. HCII 2019: Virtual, Augmented and Mixed Reality. Multimodal Interaction, Springer, pp. 299-311.

[SMJ+19] K. W. Sung, et al., PriMO-5G: making firefighting smarter with immersive videos through 5G, in proc. IEEE 2nd 5G World Forum (5GWF), 2019.

[TMK17] S. Teerapittayanon, B. McDanel and H. T. Kung, "Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices," 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, 2017, pp. 328-339. https://ieeexplore.ieee.org/document/7979979

[WHL+20] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan and X. Chen, "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, 2020. https://ieeexplore.ieee.org/abstract/document/8976180

[YFZ+19] C. Yan, L. Fu, J. Zhang and J. Wang, "A Comprehensive Survey on UAV Communication Channel Modeling," in IEEE Access, vol. 7, pp. 107769-107792, 2019. doi: 10.1109/ACCESS.2019.2933173.

[YLC+19] Yang, Q., Liu, Y., Chen, T. and Tong, Y., 2019. “Federated machine learning: Concept and applications,” ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 2019, pp.1-19. https://dl.acm.org/doi/10.1145/3298981

[ZLS+19] N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F. R. Yu, and K. Wong, “UAV-assisted emergency networks in disasters,” IEEE Wireless Communications, vol. 26, no. 1, pp. 45–51, February 2019.