Project Title	Virtual Presence in Moving Objects through 5G
Project Acronym	PriMO-5G-5G
Grant Agreement No	815191
Instrument	Research and Innovation Action
Topic	The PriMO-5G-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/

D2.1 - INITIAL DESIGN OF MEC AND NETWORK SLICE MANAGER

Work Package	WP2, 5G Core Networks
Lead Author (Org)	Jose Costa-Requena (CMC)
Contributing Author(s) (Org)	Giuseppe Destino (KCL), Anders Nordlöw (EAB), András Zahemszky (EAB), Markus Ullmann (NI), Edward Mutafungwa (AALTO), Hellaoui Hamed (AALTO), HyungJoon Jeon (EUC), Dohyun Kim(CAU), Joongheon Kim (CAU), SeHoon Yang(KT)
Due Date	30.04.2019, M10
Date	14.05.2019
Version	22.1 (Corrected Submitted)

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)

The work described in this document has been conducted within the project PriMO-5G-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.

Disclaimer

PriMO-5G-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 of Information & Communications Technology Planning & Evaluation (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 Figures

Executive Summary

The main aim of the PriMO-5G-5G project is to demonstrate an end-to-end 5G system providing immersive video services for moving objects. For this, work package 2 of the project strives to define and select the architecture for network slicing and multi-access edge computing (MEC) as part of the new 5G Next Generation Core (NGC). As a first step, this deliverable presents the design of the Multi-access edge computing as defined in ETSI [EMECspec] and network slice selection function (NSSF) specified in 3GPP [NSlice3GPP]. These two technologies i.e. MEC and network slicing have been designed in different standardisation forums i.e. ETSI and 3GPP. This report presents the proposed design for integrating these two technologies. Moreover, PriMO-5G-5G partners have identified limitations in the routing and network slice management. This report includes the description of the proposed solutions i.e. Optimal Routing and Mobile Backhaul Orchestrator (MBO) to overcome those limitations.

List of Acronyms

Acronym	Definition
3GPP	Third Generation Partnership Project
5G	Fifth-Generation Mobile Network
5G-PPP	5G Public-Private Partnership
AMF	Access and Mobility Management Function
CU	Central Unit
DU	Distributed Unit
eNB	Evolved NodeB
GUAMI	Globally Unique AMF ID
gNB	Next Generation NodeB
IAP	IP Announcement Point
KPI	Key Performance Indicator
LR	Location Register
MBO	Mobile Backhaul Orchestrator
MEC	Multi-access Edge Computing
ML	Machine Learning
NEF	Network Exposure Function
NGC	Next Generation Core
NSSF	Network Slice Selection Function
MBO	Mobile Backhaul Orchestrator
PCF	Policy Control Function
QoS	Quality of Service
RAN	Radio Access Network
SBA	Service Based architecture
SDN	Software Defined Networking
SDU	Service Data Unit
SMF	Session Management Function
SST	Service Slice Type
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management

Acronym	Definition
UE	User Equipment
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications

Terminology

Network Slice: A logical network that provides specific network capabilities and network characteristics that can be dynamically created. A given User Equipment (UE) may access to multiple slices over the same Access Network (e.g. over the same radio interface). Each slice may serve a particular service type with 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).

Network Slice instance: A set of Network Function instances and the required resources (e.g. compute, storage and networking resources) which form a deployed Network Slice.

Multi-access Edge computing (MEC): It is an evolution of cloud computing that pushes applications from centralized data centers to the network edge near the end-users. MEC is indeed one of the key pillars for meeting the demanding Key Performance Indicators (KPIs) of 5G, especially as far as low latency and bandwidth efficiency are concerned. 5G system provides a set of new functionalities that serves as enablers for edge computing. These enablers are essential for integrated MEC deployments in 5G networks. A brief explanation can be found in clause 5.13 of 23.501:

- The ability of an Application Function to influence UPF (re)selection and traffic routing via the Policy Control Function (PCF) (clause 5.6.7 of 23.501).
- The Session and Service Continuity (SSC) modes for different UE and application mobility scenarios (clause 5.6.9 of 23.501).

MEC platform: collection of functionalities required to run a MEC application on a particular virtualized infrastructure (e.g., handling DNS proxy/server, discovering and advertising MEC services, etc.).

Virtualization infrastructure: provides compute, storage, and network resources, for the purpose of running MEC applications. A virtualization infrastructure and a MEC platform, together, form a MEC host. The set of hosts are managed by *the virtualization infrastructure manager* and *the MEC platform manager*.

MEC orchestrator: functional entity responsible for the MEC system level management (e.g., triggering application instantiation and termination, triggering application relocation, etc.).

Mobile Backhaul Orchestrator (MBO): It is the module that provides access to network slice management in the mobile backhaul. The MBO provides an SDN controller to interact with the network switches in the backhaul to provide network slices as requested from the 5GC.

1 Introduction

1.1 Purpose and Scope

The main aim of the PriMO-5G-5G project is to demonstrate an end-to-end 5G system providing immersive video services for moving objects. For this, we design the architecture that will provide the required network resources to support the selected scenarios and use cases. This deliverable proposes initial design of the architecture including multi-access edge computing and network slicing functionality.

We consider firefighting to be an area where immersive video services with drones can be useful, and therefore focus on the firefighting in the development of scenarios and use cases. These uses will set the requirements for the system in terms of latency, bandwidth and reliability.

1.2 Structure of the document

This deliverable is organized as follows. Section 2 presents the 5G Core architecture, as defined by 3GPP. Section 3 describes the network slicing providing an overview including terminology and standardisation. Section 4 introduces routing optimization solution to improve user plane delivery. Section 5 describes multi-access edge computing providing initial ETSI and 3GPP standardisation. Section 6 presents the initial design for the integration of both network slicing and multi-access edge computing integrated with 5G mobile architecture. Section 0 provides concluding remarks.

1.3 Relationship to other project outcomes

This deliverable builds on past work on use case specification in WP1. Specifically, deliverable *D1.1 PriMO-5G use case scenarios* included use cases that considered the utilization of network slicing and MEC for firefighting in different contexts. Furthermore, D2.1 has strong links to WP4 research on the use of Artificial Intelligence (AI) and machine learning improved networking and edge computing in the WP1 PriMO-5G use cases. The early results for WP4 are outlined in more detail in *D4.1 Intermediate report on AI-assisted networking and edge computing*. Finally, deliverable 2.1 will provide input on the planned demonstration activities in WP5, which will include some of experimental implementation on the network slicing and MEC concepts introduced in this document.

2 5GC architecture

This section describes the 3GPP mobile architecture for 5G Core network.

Recently, 3GPP, in its Release 15, released the specification on the 5G Core Network, in TS 23.501 [3GPP23.501].

The 5GC follows a number of principles that are mainly targeted for reaching higher flexibility, supporting many different use cases. This includes the introduction of service-based principles, where network functions provide services to each other. 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 control plane is based on the Service Based Architecture (SBA). In SBA, the network functions communicate with each other via a logical communication bus and network functions can provide services to each other. A network function instance is registered to a Network Repository Function (NRF). Using the NRF, a network function instance can find other network function instances providing a certain service. The goal of such architecture is to get a higher flexibility in the overall system, and to make it easier to introduce new services.

In the 5G core, the Access and Mobility Management Function (AMF) provides the interfaces towards the Radio Access Network (RAN), the Session Management Function (SMF) keeps track of the ongoing sessions for a user, and the Unified Data Management (UDM) keeps the subscriber profiles. The User Plane Functions (UPFs) implement the user plane between the RAN and the Data Network (DN) (which can be the Internet, an operator services network or a 3rd party services network). 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.

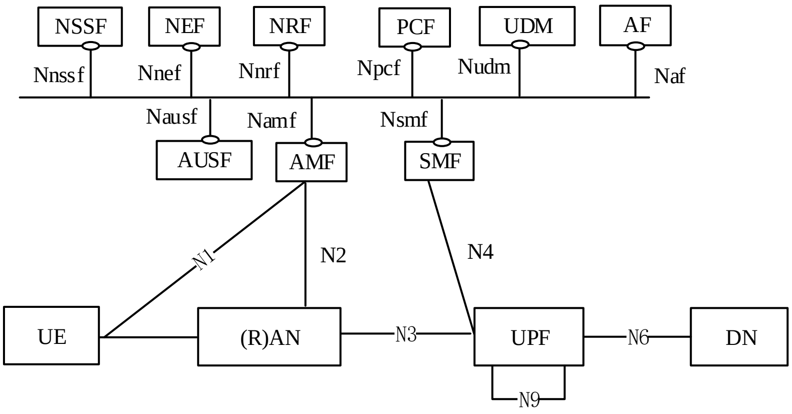


Figure 1. 5G Core Architecture, Source: 3GPP 23.501

3 Network slicing

This deliverable describes initial design of network slicing to be integrated in 5G architecture. An introduction of network slicing is defined in GSMA [NSliceGSM] which consist of the following.
From a mobile operator's point of view, a network slice is an independent end-to-end 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. The Next Generation Mobile Networks (NGMN) alliance also refers to network slice concept as follows [NSliceNGMN]. "A network slice instance may be fully or partly, logically and/or physically, isolated from another network slice instance".

The network slicing has been defined part of 5G architecture in 3GPP [NSlice3GPP]. A network slice is considered a logical end-to-end network that can be allocated to several User Equipments (UE) and it can be dynamically created and modified. A Network Slice is defined within a PLMN and shall include: the Core Network Control Plane and User Plane Network Functions and the NG Radio Access Network. A UE may access to multiple slices that are linked to a Public Land Mobile Network (PLMN) where each UE is registered. The slices are associated to given Service-level Agreement (SLA) based on bit rate, latency and packet loss.

Each slice is identified by Single Network Slice Selection Assistance Information (S-NSSAI). 3GPP has defined eight (8) S-NSSAIs in the NSSAI which is the group of S-NSSAI that is sent between the UE and the network during the registration and signalling procedure. The UE provides the network the NSSAI which then must allocate the required resources at radio, network and mobile core network functions. Figure 2 show the process for assigning AMF for the UE that requires only URLLC slice. However, if the UE requires several slices simultaneously, the process will assign a single AMF capable of handling all the requested slices.

The S-NSSAI consists of following elements:

- A Slice/Service type (SST), defined the expected requirements in terms of features and services associated to the network slice.

- A Slice Differentiator (SD), is optional and provides additional information to differentiate each slice amongst multiple Slices with the same SST to e.g. isolate traffic to different services into different slices.

In this first release the following basic slice IDs have been identified:

Slice/Service type	SST value	Characteristics
eMBB (enhanced Mobile Broadband)	1	Slice suitable for the handling of 5G enhanced Mobile broadband, useful, but not limited to the general consumer space mobile broadband applications including streaming of High Quality Video, Fast large file transfers etc.
URLLC (ultra- Reliable Low Latency Communications)	2	Slice suitable for the handling of ultra- reliable low latency communications.
MIoT (Massive IoT)	3	Slice suitable for the handling of massive IoT.

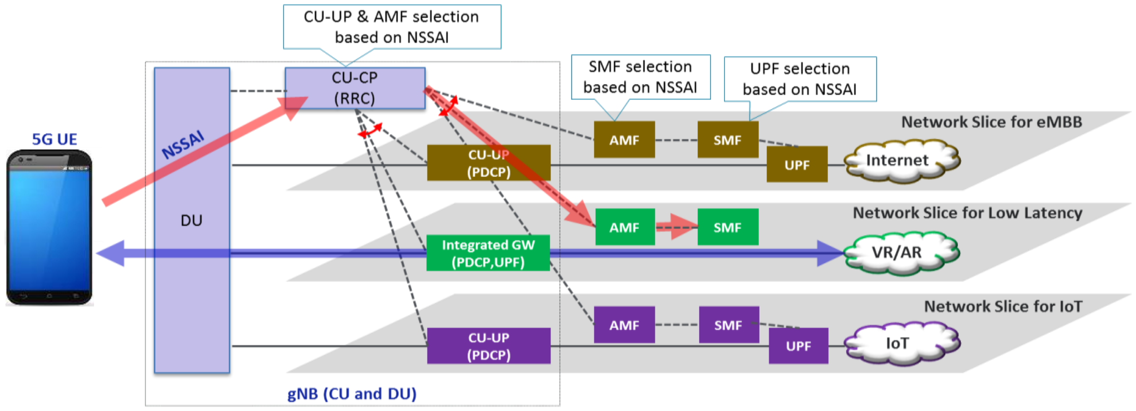


Figure 2. Example of gNB CU-CP and AMF selection scenario according to NSSAI

3.1 Network slicing design

The network slicing as specified in 3GPP documents describes the process for creating network slices based on the application or service requirements.

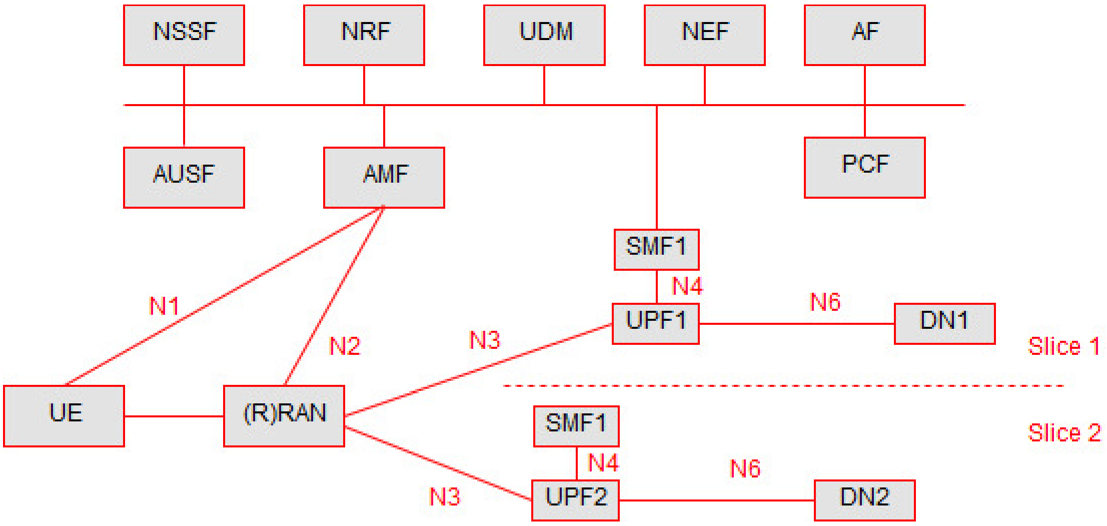


Figure 3. Network slice architecture.

As shown in previous Figure, the network slice selection function (NSSF) is the network element that handles the assignment of network slice to the service of application with selected requirements. According to TS28.801 the NSSF includes the following sub-components.

Communication Service Management Function (CSMF): Responsible for translating the communication service-related requirement to network slice related requirements.

Network Slice Management Function (NSMF): Responsible for management and orchestration of Network Slice Instance (NSI). Derive network slice subnet related requirements from network slice related requirements.

Network Slice Subnet Management Function (NSSMF): Responsible for management and orchestration of Network Slice Subnet Instance (NSSI).

3.1.1 Network slice assignment process

The network slice assignment process depends on whether the UE already has preconfigured a NSSAI from previous registrations. If the UE does not have specific or default NSSAI available during the registration the NSSF is used to discover slices assigned to the UE.

3.1.1.1 UE registration with pre-configured NSSAI

The slice selection is determined by the NSSAI which can be provided by UE during the registration process. The RAN is the first component to look into the NSSAI requested by the UE to select the appropriate AMF (TS 23.501 clause 6.3.5). The UE should use the S-NSSAI assigned to the given PLMN. The requested NSSAI provided by the UE allows the network to select the appropriate AMF, Network Slice(s) and Network Slice instance(s) for the UE (TS 23.501 clause 5.15.5).

The UE will initiate the registration and provides the assigned NSSAI from previous connections in the registration message. The AMF may query the UDM to retrieve the UE subscription which includes the Subscribed S-NSSAIs (TS 23.502, clause 4.2.2.2.2). The AMF will confirm that requested NSSAI is included in the list of allowed NSSAIs part of the UE subscription. If the UE subscription does not include the list of allowed NSSAIs the AMF will fetch it from NSSF.

After checking the requested NSSAI is allowed, the AMF must check whether it can serve the NSSAI that UE has been assigned. If the current AMF cannot handle the NSSAI requested by the UE, the AMF will query the NSSF for the allowed NSSAI.

The NSSF verifies the requested NSSAI is permitted, selects the Network Slice instance(s) (NSI) and determines the target AMF to be used for serving the UE. The NSSF may also return the NRF(s) to be used to select NFs/services (e.g. SMF, UPF) within the selected Network Slice instance(s).

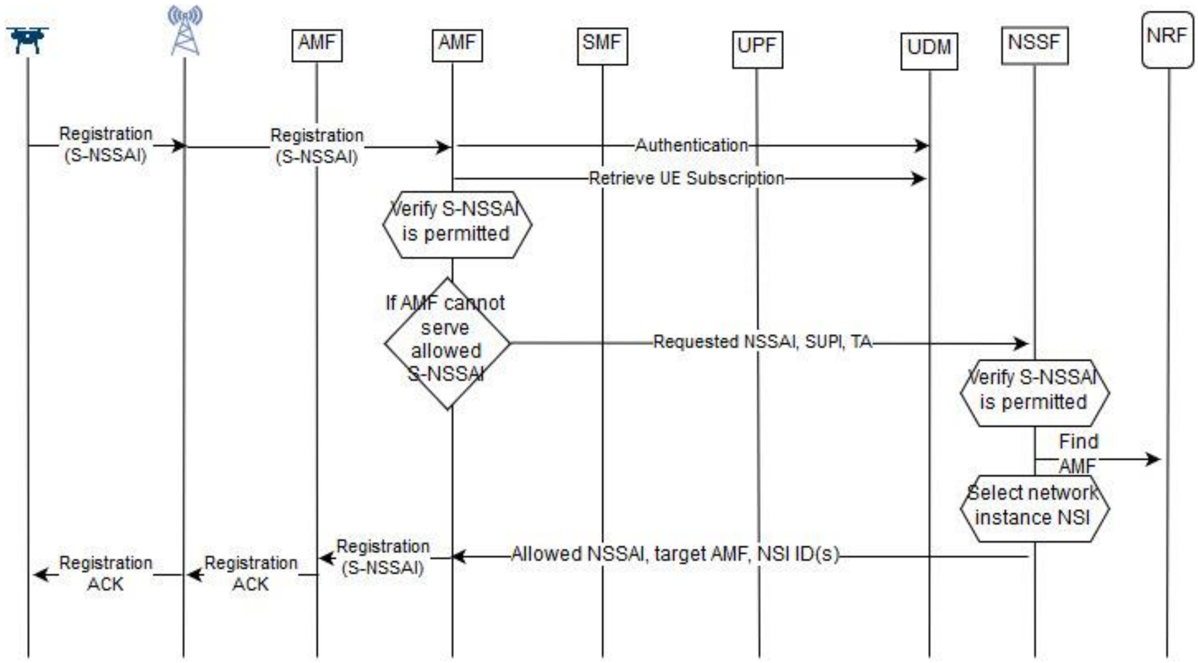

AMF
AMF
SMF
UPF
UDM
NSSF
NRF
Registration (S-NSSAI)
Registration (S-NSSAI)
Authentication
Retrieve UE Subscription
Verify S-NSSAI is permitted
If AMF cannot serve allowed S-NSSAI
Requested NSSAI, SUPI, TA
Verify S-NSSAI is permitted
Find AMF
Select network instance NSI
Registration ACK
Registration ACK
Registration (S-NSSAI)
Allowed NSSAI, target AMF, NSI ID(s)

Figure 4. Registration with slice selection with NSSAI from UE

3.1.1.2 UE registration without pre-configured NSSAI

When a UE registers with a PLMN, if for this PLMN the UE has not included a Requested NSSAI nor a Globally Unique AMF ID (GUAMI) while establishing the connection to the (R)AN, the (R)AN shall route all NAS signalling from/to this UE to/from a default AMF. The default AMF will retrieve the UE subscription from the UDM where the subscribed S-NSSAIs are included. The default AMF will contact the NSSF to find the target AMF where the registration will be forwarded and the NSI assigned to the UE.

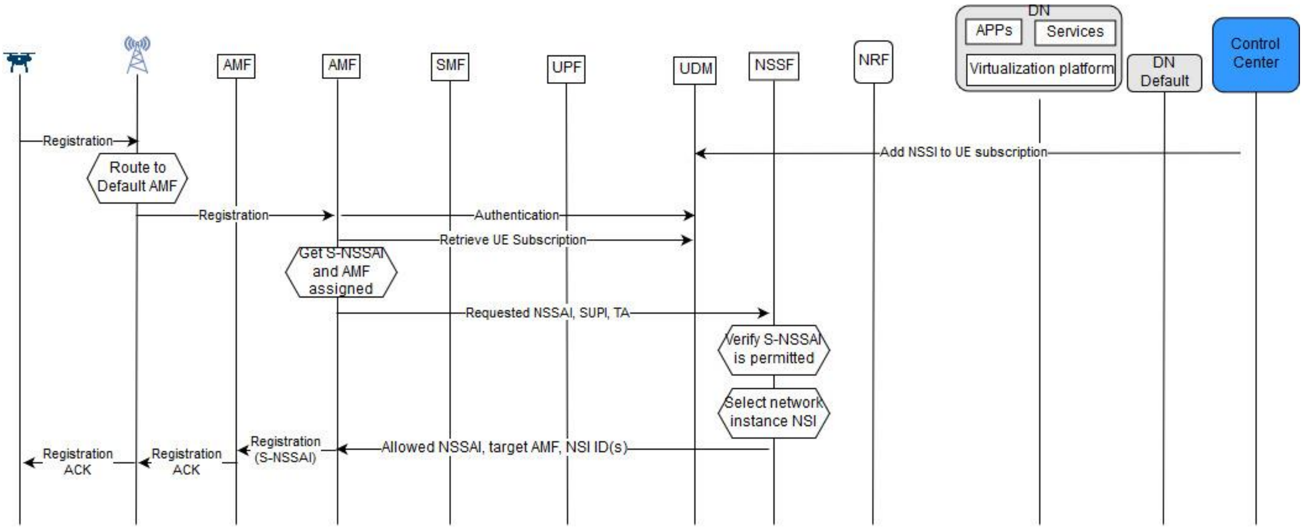


Figure 5. Registration with slice selection NSSAI from UDM

Upon successful completion of a UE's Registration procedure over an Access Type, the UE obtains from the AMF an Allowed NSSAI for this Access Type, which the UE will keep for the next registration process and at any time, the AMF may provide the UE with a new Configured NSSAI for the Serving PLMN [3GPP23.502] with UE configuration update procedure.
If the UE receives indication from the AMF that Network Slicing subscription has changed, the UE locally deletes the network slicing information it has for all PLMNs, except the Default Configured NSSAI (if present). It also updates the current PLMN network slicing configuration information with any received values from the AMF.

3.1.1.3 PDU Session establishment

SMF discovery and selection within the selected Network Slice instance is initiated by the AMF when a message to establish a PDU Session is received from the UE. The appropriate NRF is used to assist the discovery and selection tasks of the required network functions for the selected Network Slice instance.

The AMF queries the appropriate NRF to select an SMF in a Network Slice instance based on S-NSSAI, DNN, NSI-ID (if available) and other information e.g. UE subscription and local operator policies, when the UE triggers PDU Session Establishment. The selected SMF establishes a PDU Session based on S-NSSAI and DNN

After the slice assignment a PDU will be established to the UE according to the slice requirements. A PDU Session belongs to one and only one specific Network Slice instance per PLMN.

3.2 Mobile Backhaul Orchestrator

5G architecture needs to address new requirements to enable network slicing and described in previous sections, 3GPP has defined the logical elements to manage the network slices e.g. NRF, NSSF, etc. However, 3GPP does not specify how the network slice is physically allocated and managed at the level of the transport networks. The existing networking technologies used in mobile backhaul are suitable for fixed IP networks where fully distributed routing algorithms provide optimal paths based on link costs and react efficiently upon link breaks. IP networking delivers a flat net-work for best effort traffic management. However, 5G mobile networks aim at new features such as network slicing where the same network provides multiple network overlays each with different traffic requirements.

A network slice in 5G is a set of packet transport links and nodes, set of computing elements and software for the network functions to run a network using the assigned resources. A slice is set up for a particular use case of the 5G network and they can be provisioned in advance to guarantee the QoS requirements for URLLC, Massive Internet of Things (mIoT) or enhanced Mobile Broadband (eMBB) communications.

The mobile backhaul networks have fulfilled the traffic requirements based on over-dimensioning and pre-provisioning that ensure enough capacity for best-effort IP based networks. However, pre-provisioning cannot work in 5G networks since the set of assigned resources can be increased and decreased in size based on user needs and policies that change over time. The slices might be created, updated and terminated dynamically based on end-user requirements. The network slices are required to constrain unexpected high peaks of traffic e.g. M2M under pre-defined set of resources so other traffic is not affected and can keep their own allocated resources.

In order to integrate the 3GPP network slicing network functions with the physical management of network resources a Mobile Backhaul Orchestrator (MBO) is defined. The MBO utilize technologies such as SDN and Machine Learning (ML) on top of basic IP routing to efficiently manage network slices. The ML techniques are used to estimate the available resources in each link based on different network features and calculations made in the network so that it would be used as an input for the routing algorithms to decide the best route. Those features include link bandwidth usage, end to end latency, hop count, packet loss etc. The MBO uses a centralized SDN based management [SDN_MBO] of resources combined with ML to effectively allocate network resources based on new requirements for existing or new network slices.
From the different types of machine learning techniques available, we consider the supervised learning is suitable for managing the network resources. The idea behind supervised learning is that, for some inputs, we want to have certain value as an output. Thus, the supervised ML algorithms run based on the inputs received from the monitoring system until they get output values close enough to the target value which provides optimal usage of the network resources. Therefore, using this technique we evaluate if a given link is congested or not based on the network information such as jitter, bandwidth utilization, packet loss as an input. The ML will trigger some actions to change routing policies in the switches which the SDN controller integrated in the MBO will take care of.

4 Optimal Routing architecture in the mobile networks

4.1 Introduction

In this section, we present the Optimal Routing architecture. Optimal Routing is a session and mobility management scheme that is aimed to be used in 3GPP core networks.

4.1.1 Problem Statement

A Protocol Data Unit (PDU) session is defined as an association between the User Equipment (UE) and a Data Network that provides a PDU connectivity service [3GPP23.501]. A UE may have multiple PDU sessions established. In this chapter, unless otherwise noted, all the procedures are explained with a single PDU session, but the same procedures apply when the UE has multiple PDU sessions.

5G networks aim to provide seamless user experience. To achieve seamless user experience, it is important to distinguish the concepts of service continuity and session continuity. According to 3GPP 23.501 [3GPP23.501], service continuity can be defined as the uninterrupted user experience of a service, while session continuity implies that the IP address is preserved during the lifetime of a session.

3GPP has defined 3 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.

The 3 different SSC modes are illustrated on Figure 6.

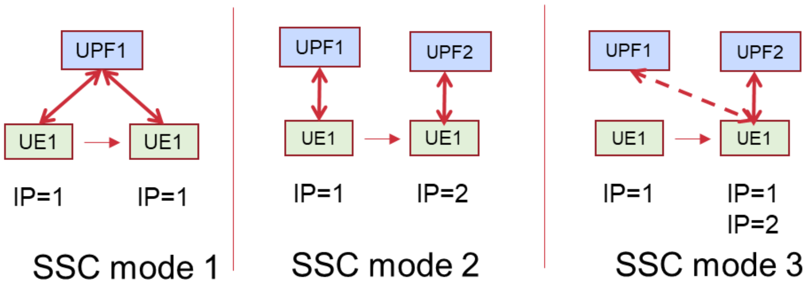


Figure 6 Session and Service Continuity (SSC) modes

In the current 3GPP Rel-15 mechanisms, using SSC modes for PDU sessions, it is not possible to achieve all of the following goals in the same time:

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

This will be explained further in the following subsections.

4.1.1.1 Challenges with SSC mode 1

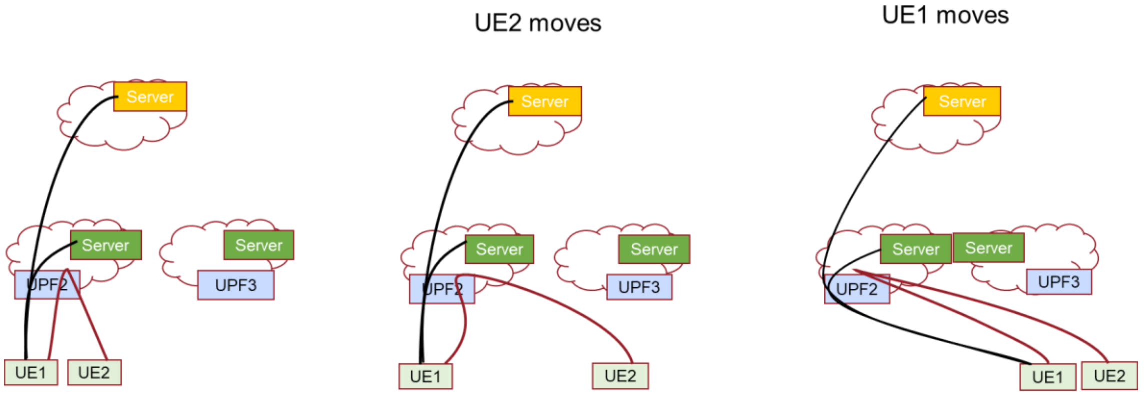


Figure 7 Single "distributed" anchor selected for UE1's and UE2's PDU Session. SSC mode 1 is used for both PDU Sessions.

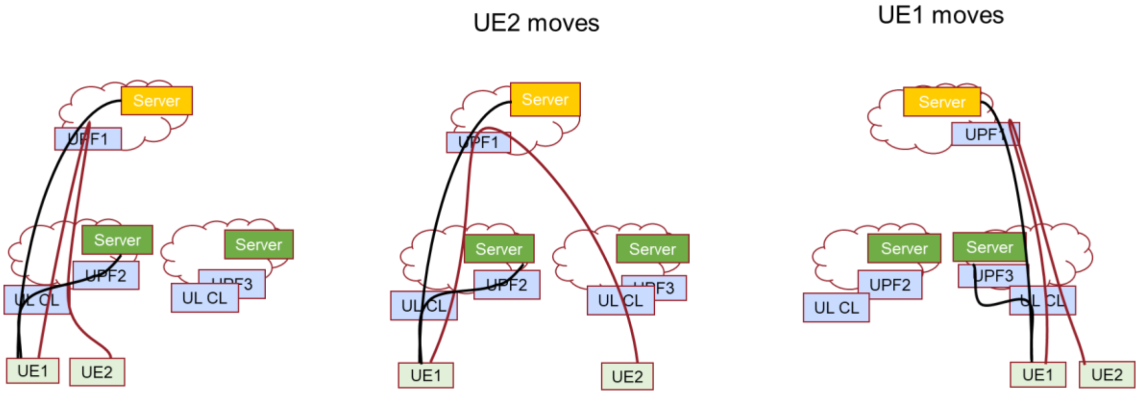


Figure 8 Single "central" anchor is selected for UE1's and UE2's PDU Session. SSC mode 1 is used for both PDU Sessions.

Figure 7 and Figure 8 explain the shortcomings of using SSC mode 1 for certain scenarios. In Figure 7, a local UPF is selected for both UE1 and UE2. While this allows communicating with servers, that are close to the user and the UE-to-UE communication is also kept local, as soon as the UEs move further away the latencies increase as the traffic must trombone back to the original UPF.

In Figure 8, a central anchor (UPF1) is selected for both UEs at the session establishment. The usage of an Uplink Classifier UPF and additional PDU Session Anchor makes sure that the UEs are able to communicate with close edge servers. However, the latency for UE-to-UE communications will be high as the traffic always has to pass the central UPF, regardless of the UE's location.

4.1.1.2 Challenges with SSC mode 2

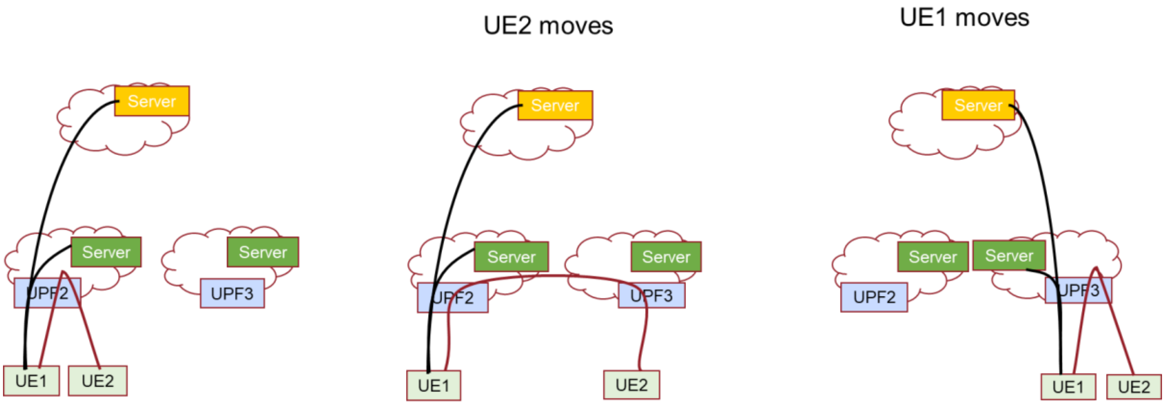


Figure 9 Single “distributed” anchor is selected for UE1’s and UE2’s PDU Session. SSC mode 2 is used for both PDU Sessions.

In Figure 9, a local UPF is selected at PDU Session Establishment both for UE1 and UE2. However, when any of the UEs move away from its UPF, the UPF may need to be re-selected to maintain low-latency communication. In SSC mode 2, this is achieved by tearing down the PDU session and establishing a new PDU session over a new UPF. In this case, there are two problems. First, there is no session continuity, and second, the UE that didn’t move needs to be informed of the new IP address of the UE that moved.

4.1.1.3 Challenges with SSC mode 3

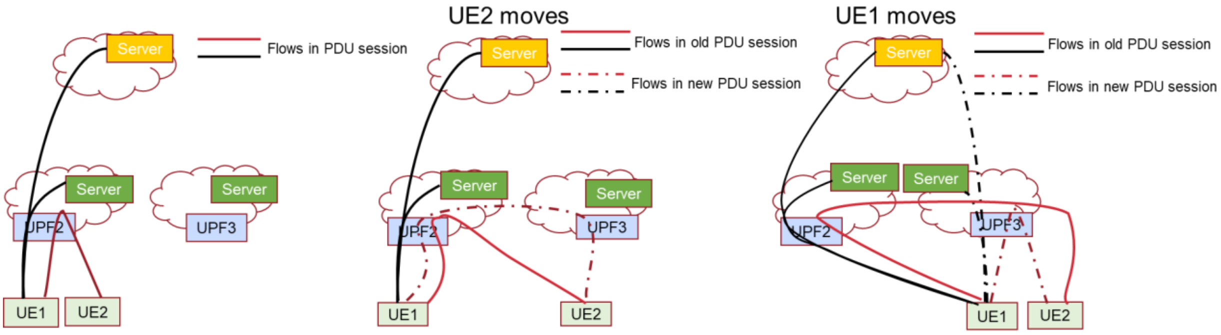


Figure 10 Single “distributed” anchor is selected for UE1’s and UE2’s PDU Session. SSC mode 3 is used for both PDU Sessions.

In Figure 10, a local UPF is selected both for UE1 and UE2’s PDU Session, and SSC mode 3 is used for those PDU Sessions. For UE-to-UE communication, the challenge is again that the UE that didn’t move needs to be informed of the new IP address of the UE that moved. Session continuity is not achieved in this scenario, and the service continuity depends on the application.

4.1.2 Optimal Routing overview

The goals of Optimal Routing are the following:

- Achieve IP address preservation within the lifetime of the session

- Maintain optimally routed paths in the mobile network even after mobility events, including
  - Ability for the UEs to communicate with close edge servers
  - Low latency communication between two UEs

As introduced in the problem statement in section 4.1.1, the above goals cannot be fulfilled simultaneously with the current mechanisms in 3GPP. 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.

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 Optimal Routing architecture is shown in a schematic illustration in Figure 11. In the figure, two remote servers communicate with the UE, and the flows go through different IAPs. The details of the picture are explained in the following subsections.

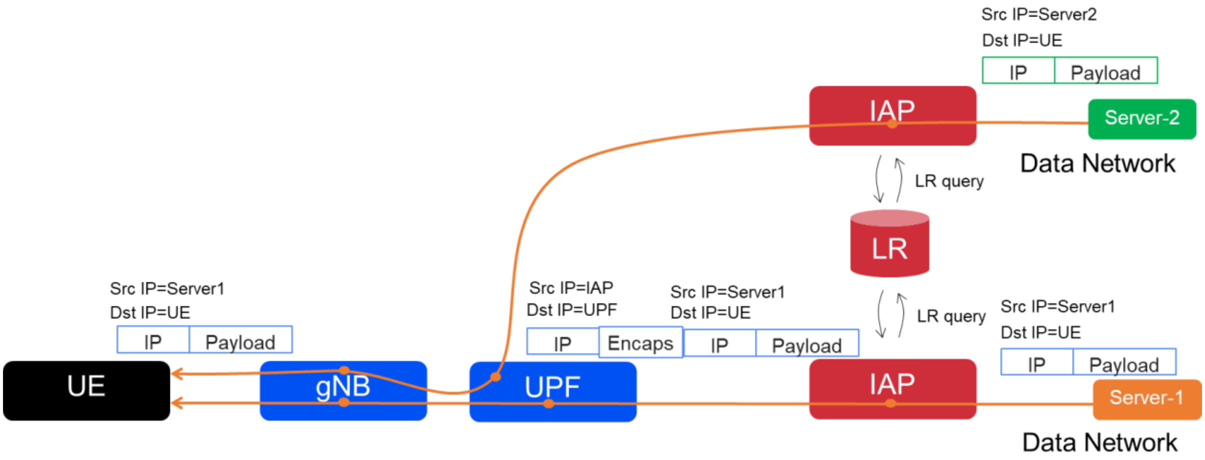


Figure 11 The Optimal Routing Architecture

4.2 IP Announcement Point (IAP)

In a typical network, there are multiple IAPs deployed. Towards the Data Network, the IAPs advertise UE IP address ranges (in the case of IPv4) or UE IP prefix ranges (in the case of IPv6). If an IAP advertises a UE's IP address, it means that it is prepared to receive packets destined to the given UE. The address space used by the operator for UE addresses may be divided into multiple ranges. Different IAPs may serve different ranges. This way, IAPs easily scale with the address space of the operator.

Multiple IAPs may advertise the same address range. In other words, multiple IAPs may receive packets for the same UE. A notable scenario is when every site in the operator's network can receive traffic for all UEs. This can be achieved by having IAPs deployed in the site in such a way that their address ranges cover the entire UE address range. It is important to note that even if multiple IAPs can treat packets to the same UE, this does not mean that all the IAPs will receive a copy of each packet for this UE. On the contrary, a single unicast packet is only received by a single IAP – the packet will be always routed to the closest IAP, which is the way routing protocols work today. There is no change in the routing protocols deployed in the network, i.e. in the Data Network plain IP routing is used.

IAP is an element that provides an entry point to the mobile network for user plane packets. When an IAP receives a downlink packet destined to the UE, it forwards the packet to the UPF serving the UE's session. For this to happen, the IAP needs to know the UE's location. As introduced above, it is the Location Register (LR) that stores the UE IP address -> UE location mappings, therefore the IAP queries the UE's location with a protocol that logically works as request/response. The received answer is stored in the IAP's local cache, so for subsequent packets there is no need to query the LR (until a timeout may happen or the entry is explicitly removed). After it received the answer, it forwards the packet towards the UPF. The IAP at this point encapsulates the packet. The exact encapsulation protocol can be different for different deployments. One example is to use GRE encapsulation from the IAP to the UPF.

As a summary, the IAP performs the following steps after receiving a downlink packet:

- Step 1: Check if UE's location is in the local cache, if yes: goto Step 3, if no: goto Step 2
- Step 2: Query the Location Register and store the answer in the local cache
- Step 3: Encapsulate the packet and send it to the UPF serving the UE's session

It is important that the local cache is always up-to-date, i.e. changes of the UE's location are propagated in a timely manner to the local cache. A solution for this is that the Location Register remembers all the IAPs that have asked for the UE's location. This way, the LR knows which IAPs currently hold state for the given UE. Whenever the UE location changes at the LR, the LR will push this change to the IAPs that need to be informed about this change.

4.3 Location Register (LR)

The Location Register is conceptually a centralized database that is storing UE IP address->UE's session location mappings, where the UE location is e.g. the IP address of a UPF where the UE's session is handled.

For designing the LR, we need to understand the connection of the LR to other control plane elements in 5GC. When a PDU session is set up, the SMF has the task: a) to allocate an IP address to the PDU session (unless it is a static address present in UDM); b) to perform UPF selection; and c) to configure the user plane so that it can handle traffic for the PDU session. Because of these tasks are already performed at the SMF, it is proposed that the SMF informs the LR of the initial UE IP address->UPF IP address mapping.

The LR needs to be informed by the control plane of the 5GC whenever the UPF changes during the lifetime of the PDU session. When the LR learns the UE's new location, it also signals this to the IAPs that currently hold state in their local caches for the given UE IP address.

When the PDU session is released, the SMF informs the LR about it, and the LR deletes the entry for the UE's session. The LR also informs the subscribed IAPs to remove the entries for the UE's session.

4.4 Procedures

4.4.1 *UE registration & PDU Session Establishment :*

During the PDU session setup, the SMF performs UPF selection, an IP address is selected to the UE, and the session is set up in the UPF and in the gNB. As an addition, the LR is informed of the UE IP address->UPF IP address mapping by SMF. The call flow is shown on Figure 12.

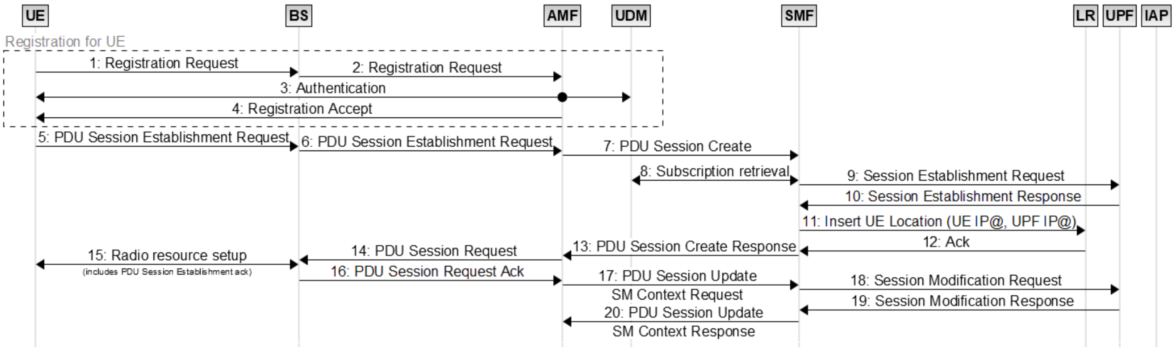

UE
BS
AMF
UDM
SMF
LR
UPF
IAP
Registration for UE
1: Registration Request
2: Registration Request
3: Authentication
4: Registration Accept
5: PDU Session Establishment Request
6: PDU Session Establishment Request
7: PDU Session Create
8: Subscription retrieval
9: Session Establishment Request
10: Session Establishment Response
11: Insert UE Location (UE IP@, UPF IP@)
12: Ack
13: PDU Session Create Response
14: PDU Session Request
15: Radio resource setup
(includes PDU Session Establishment ack)
16: PDU Session Request Ack
17: PDU Session Update
SM Context Request
18: Session Modification Request
19: Session Modification Response
20: PDU Session Update
SM Context Response

Figure 12. UE Registration and PDU Session Establishment

The call flow follows the Registration and PDU Session Establishment procedures as defined in 23.502 [3GPP23.502]. The new messages are shown as steps 11 and 12 in the call flow.

After the PDU session has been established towards a Data Network, the UE is ready to send and receive IP packets towards and from the Data Network. As described above, all downlink packets are sent via an IAP. However, uplink packets do not pass IAPs: after the UPF finished processing the packet, it can be routed directly to the Data Network. This is depicted on Figure 13.

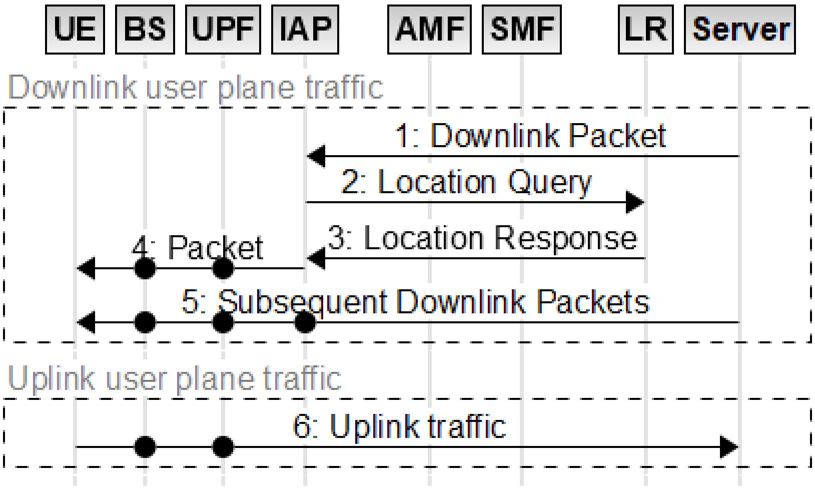


Figure 13 Downlink and uplink user plane traffic

4.4.2 *Xn Handover*

Xn handover is a procedure executed when a UE changes gNB, and the Xn reference point exists between the source and target gNBs. In this case, the UPF terminating the N3 interface does not change. Therefore, there is no interaction needed to update the LR and the IAPs.

4.4.3 Xn Handover with UPF change

In this procedure, the UE performs a handover that results in changing the UPF terminating the N3 interface. Assuming that a single UPF is handling the PDU session, this also means that the UPF serving the UE's session is changing. It the current 3GPP specifications, it is not possible to change this UPF. The procedure will contain steps to update the LR, and in turn, the LR is required to inform the relevant IAPs to change the information in their local caches. A call flow showing the procedure is presented on Figure 14.

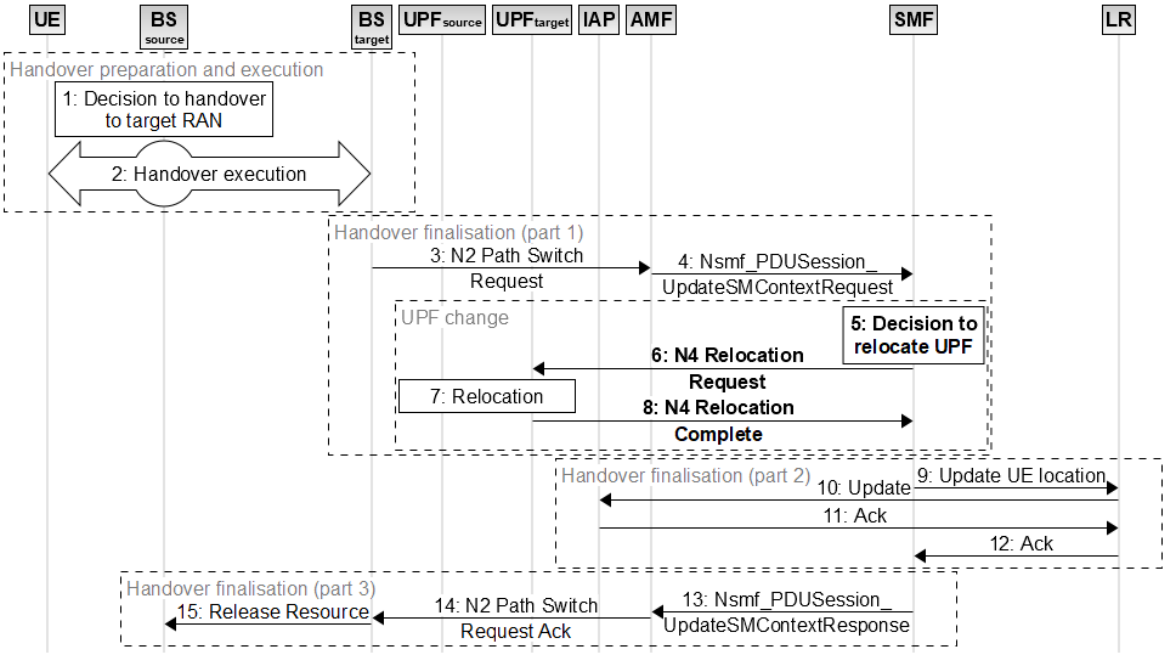

UE
BS source
BS target
UPF source
UPF target
IAP
AMF
SMF
LR
Handover preparation and execution
1: Decision to handover to target RAN
2: Handover execution
Handover finalisation (part 1)
3: N2 Path Switch Request
4: Nsmf_PDUSession_ UpdateSMContextRequest
UPF change
5: Decision to relocate UPF
6: N4 Relocation Request
7: Relocation
8: N4 Relocation Complete
Handover finalisation (part 2)
9: Update UE location
10: Update
11: Ack
12: Ack
Handover finalisation (part 3)
15: Release Resource
14: N2 Path Switch Request Ack
13: Nsmf_PDUSession_ UpdateSMContextResponse

Figure 14 Xn handover with UPF change

The procedure starts with RAN handover, where the Base Station is changed from source to target. At the end of the handover execution phase, all packets are forwarded from the source Base Station to the target Base Station. At step 3, the target Base Station sends an N2 Path Switch Request message to inform the AMF that the UE changed its Base Station, and the list of PDU sessions to be switched (here, there is a single PDU session for the UE). At Step 4, the AMF sends to the SMF a request indicating that the PDU session should be switched. At Step 5, the SMF decides that the UPF should be changed for this session, and in step 6, it sends a request on N4 for relocation, including the tunneling information to the target Base Station. In block 7, the actual relocation happens, and in step 8 the SMF is informed that the N4 relocation is complete. This message includes the CN (Core Network) Tunnel info to be sent to the Base Station. At this point, the target UPF processes the user plane packets for the PDU session and packets arriving to the source UPF are forwarded to target UPF. In step 9, the SMF informs the LR that the UE's location has changed, i.e. it sends the target UPF IP address. The LR, in turn, updates all the subscribed IAPs with the new IP address. In step 13, the SMF informs the AMF about the successful switching, including the CN Tunnel Info, which is sent by the SMF to target Base Station. This procedure builds on the handover procedure that is already part of the current 3GPP standard in TS 23.502 [3GPP23.502]. Optimal Routing adds the LR update, the IAP local cache updates and enables the UPF relocation.

4.5 Scenarios

4.5.1 UE simultaneously accessing local and central servers in the same Data Network

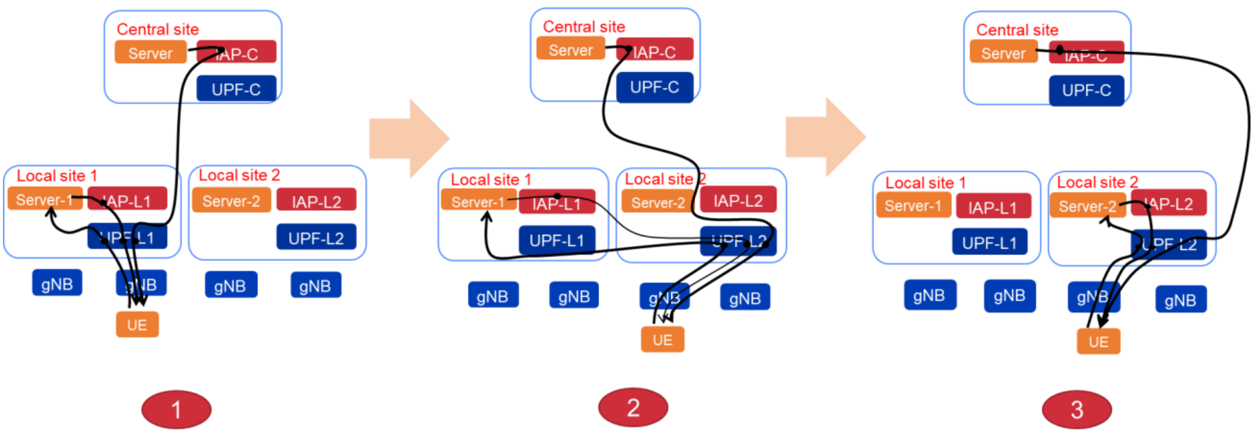


Figure 15 Optimal Routing example when a UE is communicating with local and central servers in the same Data Network.

Figure 15 shows the scenario where the UE communicates via both a server in a central site and with a server in the local site. For this specific PDU session, it is assumed that a UPF on a local site is selected by the SMF to provide the user plane. The scenario assumes a single UPF per session. In step 1, the UE is on UPF-L1 (the UPF on Local Site 1), i.e. it is UPF-L1 that terminates the N3 interface. Packets coming from the server at the central site pass the central IAP, while packets coming from the local server pass the IAP on Local Site 1. In step 2 of the Figure, the UE changes base station and UPF (based on SMF decision), so the UPF terminating the N3 interface is now UPF-L2. The signalling in the control plane for this scenario was shown in Figure 14. The IAPs now encapsulate the downlink packets and send them to UPF-L2. Note that the servers are not changed, so the UE is still communicating with a server in the Local Site 1. In case application server relocation is implemented from Server-1 to Server-2 (see step 3), and Server-2 has the same (anycast) address as Server-1, the UE will be able to continue the application session from a topologically closer server.

As an extension to the scenario, the UE may also communicate with servers on the Internet via e.g. peering points in the central sites. In this case, it is still possible to use a single PDU session for both the Internet traffic and traffic from the central and local site clouds. As for NAT functionality, if IPv6 is used, NAT is not needed in the network. If IPv4 is used, and NAT is needed between the public Internet and the operator’s network, a NAT can be placed in a central location, and it will be between the DN and the IAP. This also means that there is only a single place where Internet traffic to a given UE can enter the network, but it is still possible to use the private IP address for communicating with the servers on the central and local sites, and therefore the UE can still communicate via central and multiple local servers at the same time, with a single session and a single IP address. As an alternative, the NAT can be placed in the UPF, so traffic from the Internet can enter at multiple IAPs.

4.5.2 UE-to-UE communication

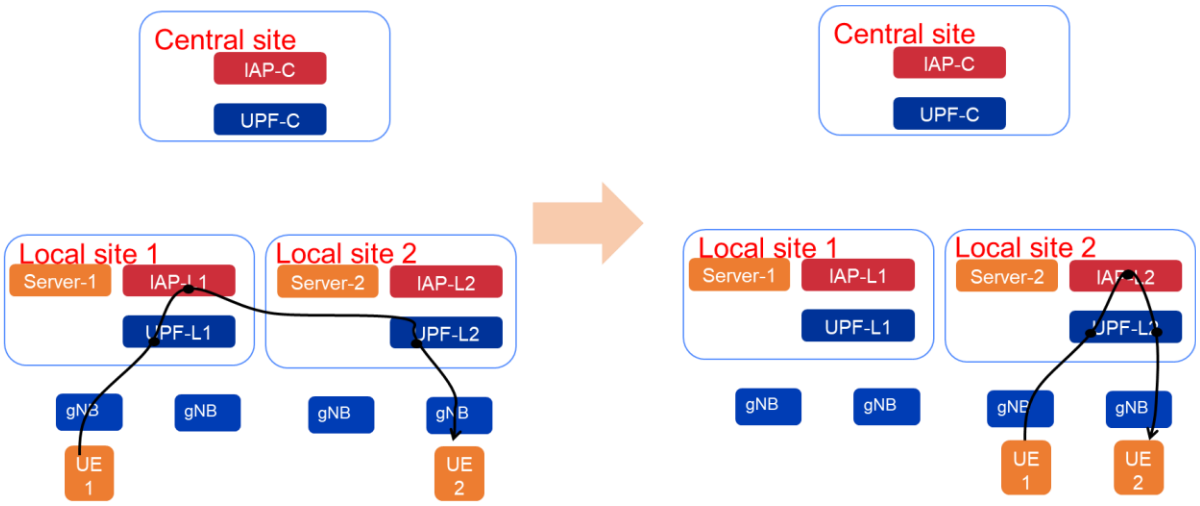


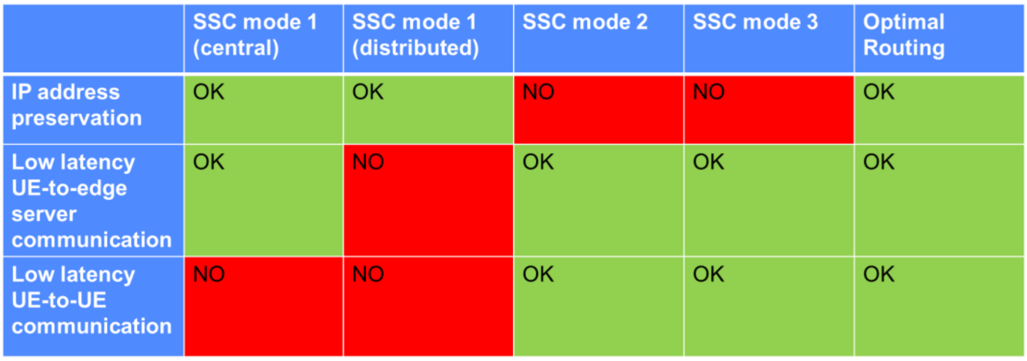
Figure 16 Two UEs communicating via each other in the Optimal Routing architecture

In Figure 16, we show a case where UE1 has a PDU session with UPF-L1 (UPF on the local site 1), while UE2 has a PDU session with UPF-L2. In this case, the two UEs can directly communicate, and packets sent from UE1 to UE2 follow the path: UE1->gNB->UPF-L1 (processing uplink packet from UE1) ->IAP-L1 ->UPF-L2 (processing downlink packets for UE2) ->gNB->UE2. Packets from UE2 follow the path: UE2->gNB->UPF-L2 (uplink packets from UE2)->IAP-L2->UPF-L1 (downlink packets for UE1)->gNB->UE1. In other words, the usage of IAPs and the UPFs at local sites makes it possible that the traffic does not need to travel via a central site.

If UE1 moves to Local Site 2, it changes base station and UPF. After the move, the communication between the two UEs is kept within Local Site 2. In this case, no central site or old local site needs to be visited. Also, the communication between the two UEs is uninterrupted, as the IP address of UE1 is not changed when its UPF got reallocated.

4.6 Summary

As presented above, the Optimal Routing architecture aims to combine the benefits of SSC modes 1, 2 and 3, while addressing their shortcomings. This makes it attractive for not only edge server to UE, but also for UE-to-UE communications. Optimal Routing can be used to address local edge server to UE and UE-to-UE communication patterns within the same PDU session with IP address preservation. The following analysis is based on the problem statement in section 4.1.1 and includes the characteristics of Optimal Routing.



	SSC mode 1 (central)	SSC mode 1 (distributed)	SSC mode 2	SSC mode 3	Optimal Routing
IP address preservation	OK	OK	NO	NO	OK
Low latency UE-to-edge server communication	OK	NO	OK	OK	OK
Low latency UE-to-UE communication	NO	NO	OK	OK	OK

Figure 17 Characteristics of different SSC modes and Optimal Routing

5 Multi-access edge computing

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 regards 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.

5.1 Standardisation

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 [3GPP23.501]. 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 [ETSIspecs], [ETSIreq], [ETSIwp][ETSIdep].

5.2 Multi-access edge computing design

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

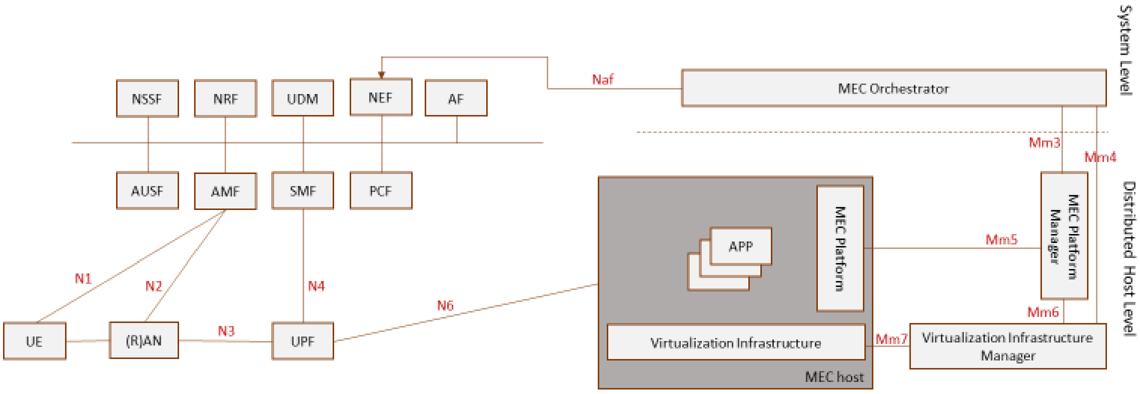


Figure 18. 5GC system and MEC integration.

The MEC system reference architecture, defined in [ETSI_GS_MEC_003], consists of MEC hosts and functional entities required to tun 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 particular 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 of 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 [ETSI_WP28] and [3GPP23.501] 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.

5.2.1 MEC application instantiation

The instantiation of an application in the MEC starts with a request to the MEC orchestrator. The latter checks the application instance configuration data and authorizes the request. The orchestrator also selects the MEC host and sends an instantiate application request to the MEC Platform Manager. The MEC Platform Manager initiates a resource allocation request to the virtualization infrastructure manager specifying the requested resource. The virtualization infrastructure manager allocates the resources according to the request of the MEC Platform Manager and sends the response back to the MEC Platform Manager. The latter provides the MEC platform with the configuration including the traffic rules to be configured, the required and optional services. The MEC platform configures the traffic rules for the application instance. Once the instance starts running normally and the traffic rules are activated, the MEC platform sends a configuration response to the MEC platform manager. A response will be communicated to the MEC orchestrator then to the initial requestor. Figure 19 summarizes the instantiation of MEC application.

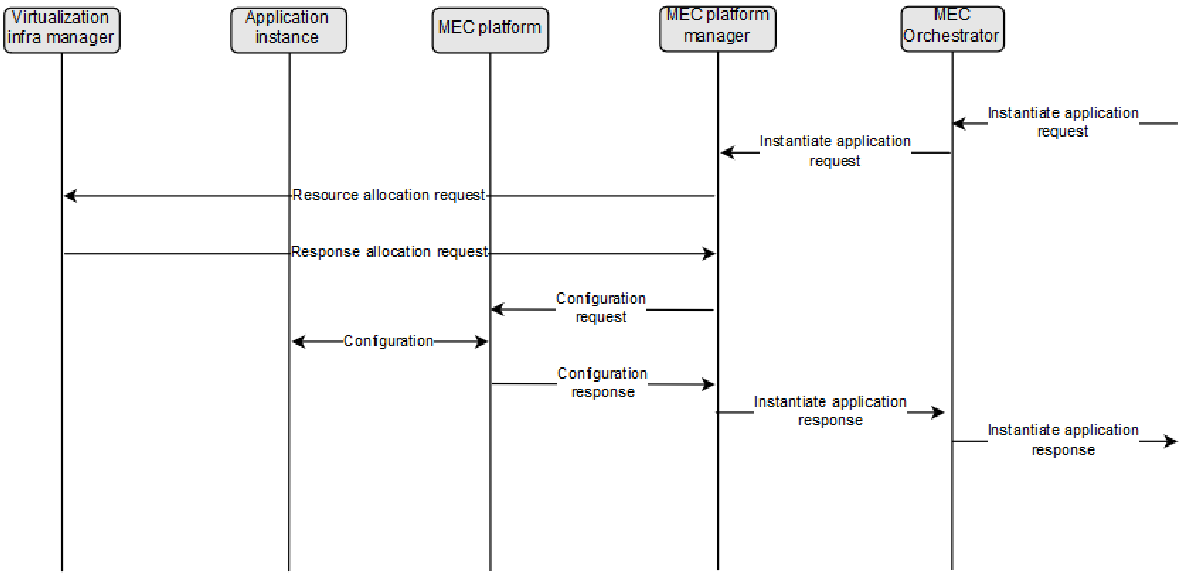

Virtualization infra manager
Application instance
MEC platform
MEC platform manager
MEC Orchestrator
Instantiate application request
Instantiate application request
Resource allocation request
Response allocation request
Configuration request
Configuration
Configuration response
Instantiate application response
Instantiate application response

Figure 19. Instantiation of MEC application

5.2.2 MEC application start-up & traffic routing

In the 5G integrated MEC architecture, traffic needs to be routed to the targeted applications. The User Plane Function (UPF) has a key role in this architecture. From a MEC perspective, the UPF is the data plane for routing the traffic to the desired applications. It is influenced by MEC through control plane interactions with 5GC NFs.

When a MEC application is instantiated, no traffic is routed to the application until the application is running and the UPF is configured to route the traffic towards it. The start-up follows the application instantiation procedure. The MEC application instance informs the MEC platform that it is up-running and performs the additional authentication/configuration when required. The data plane should thereafter be configured to route the traffic towards the MEC application. This configuration is done by the MEC platform. It interacts with the Policy Control Function (PCF) to request traffic routing by sending information that identifies the traffic to be routed. The request intends to target the existing or future PDU sessions of UEs and is sent via NEF when MEC functional entities are not allowed to interact directly with 5GC NFs. The PCF transforms the request into policies and provides the routing rules to the appropriate SMF. Based on the received information, the SMF identifies the target UPF and initiates configuration of the traffic rules. Traffic steering is provided in clause 5.6.7 of [3GPP23.501]. The flow diagram is provided in Figure 20.

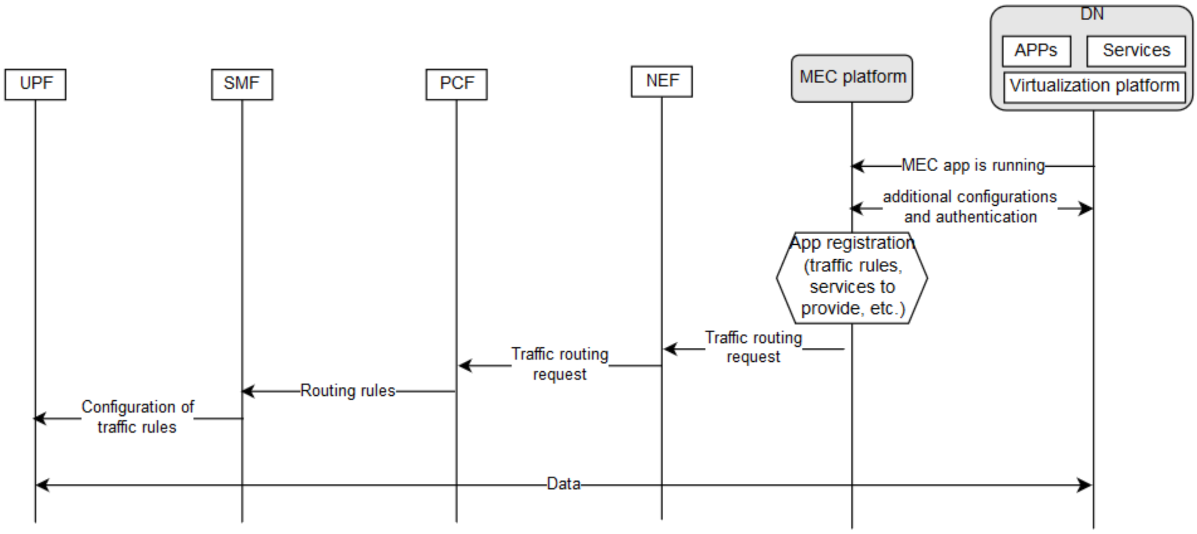


Figure 20: MEC application start-up and traffic steering

5.2.3 MEC application mobility

The users are expected to be mobile and their movements could make the current location of the serving applications at the edge non-optimal. To maintain the application requirements in a mobile environment, application mobility is required. This is translated into changing the location of the application instance serving the users. For application mobility, the MEC should ensure:

- continuity of the service;
- mobility of the application;
- mobility of user context in case of stateful applications.

MEC application mobility is a work in progress in ETSI ISG MEC. The work program MEC-021 is established to target the related issues (publication expected in August 2019). The principle of MEC application mobility is illustrated in Figure 21.

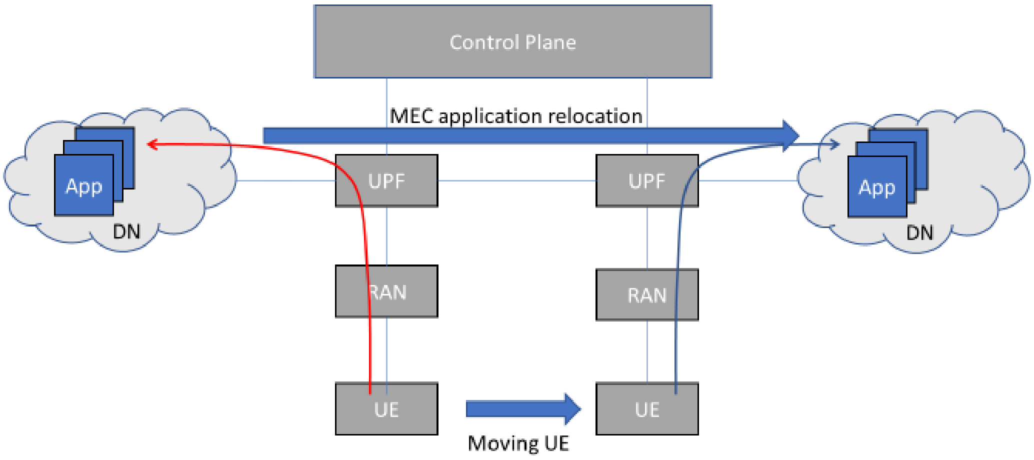


Figure 21: Principle of MEC application mobility

MEC service relocation can be triggered by, among others, relocation of the UPF. Indeed, in handover scenarios without UPF change, there is no need to relocate MEC services. 5G system provides tools for MEC functional entities to monitor the mobility events related to MEC application users. A MEC orchestrator can subscribe to user plane path management event notifications from the SMF, so it can receive notifications about path change. The subscription can be for early and/or late notifications. In the case of a subscription for early notifications, the SMF sends the notifications before the new uplink path is configured. In the case of a subscription for late notifications, the SMF sends the notification after the new UP path has been configured. The notifications will be used for traffic routing or application relocation.

ETSI gr_MEC018 report defines 5 phases for MEC application relocation (as shown in Figure 22):

- Relocation initiation: all the relocation triggers that may cause the relocation procedure are considered (considering both the source and the target MEC environments).
- Relocation validation: decision for relocation is taken, relocation method is validated.
- Relocation preparation: synchronize the application related context between the two environments.
- Relocation execution: execution of the relocation procedure.
- Relocation completion: release the old and temporary resources, traffic rules updates.

In some scenarios, the MEC application serving the moving user has already an instance in the target MEC. This could occur for applications designed to serve multiple users instead of a single user. In this case, MEC application may not need to be instantiated at the target host. Only user context needs to be transferred for stateful applications. In the other hand, if the application instance at the source MEC still is needed to serve other users, it should not be torn down after the application mobility procedure.

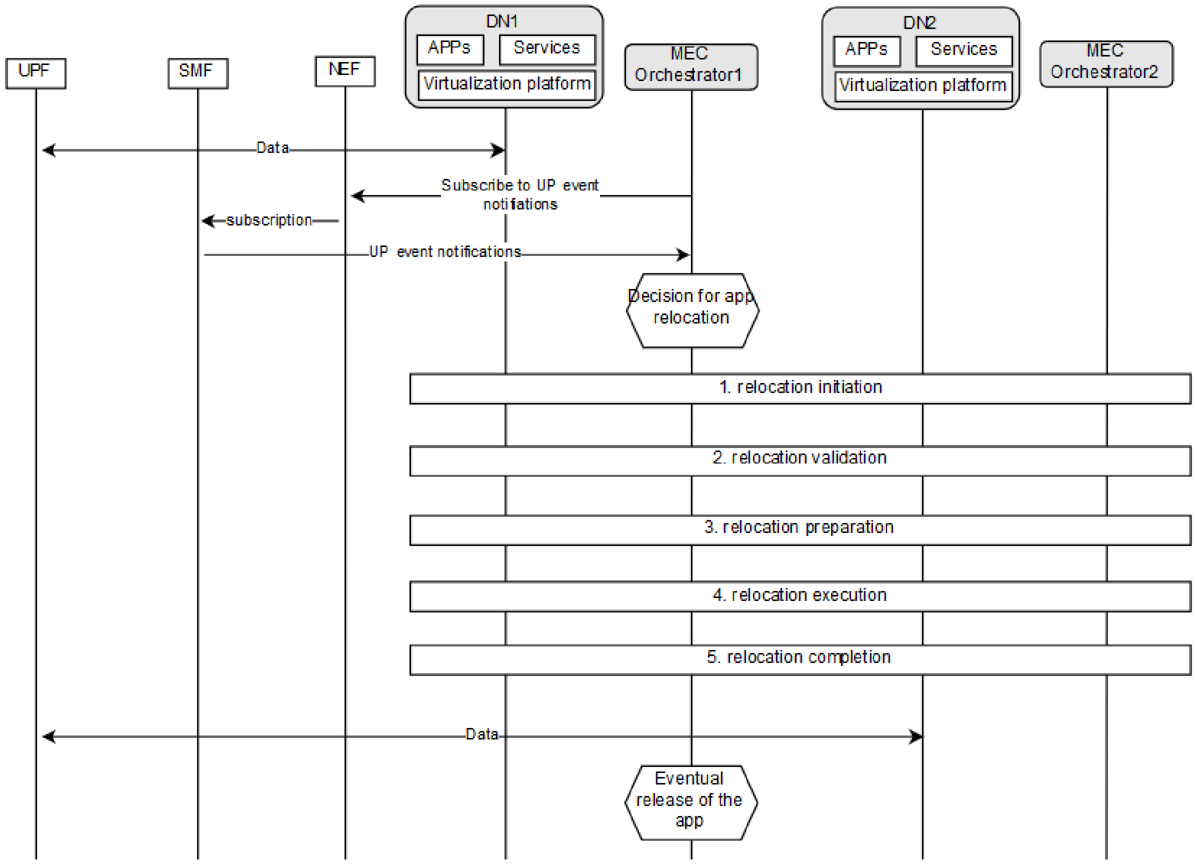


Figure 22: MEC application relocation (from MEC1 to MEC2)

6 Design Multi-access edge computing and network slicing interworking

This section provides an overview of the PriMO-5G use cases and technical requirements to be supported by the 5G system. An initial description of the proposed architecture where the network slicing and MEC are integrated together to support the PriMO-5G use cases is introduced. Besides MEC and network slicing the Optimal Routing and MBO are integrated in the 5G system to overcome some limitations in routing in 5GC and network slicing management. As a result, this section presents the proposed 5G system that integrates different technologies such as MEC, network slicing, Optimal Routing and MBO to fulfil the requirements from PriMO-5G use cases. A set of message flows are included to show the interworking between the different technologies following 3GPP procedures for mobile registration and connection setup.

6.1 Use Case Analysis

The PriMO-5G use cases consists of drones, robots and other components that will be used for firefighting in different context such as urban or forest areas. The 5G system has to support these use cases and integrate technologies such as MEC, network slicing, MBO, optimal routing to fulfil the needed requirements in terms of reliability, latency and bandwidth. In the PriMO-5G use cases, drones are equipped with fast and reliable wireless communications for information acquisition and communication. The drones can provide visual and location information for the robots, humans, and the incident commander. The drones can serve as relay for the communications between the ones at the scene (i.e. between robots and human firefighters on the field or in the command center) and the incident commander behind the scene. Aerial relay is an effective tool ensuring the reliable communications for firefighting.

The fleet of drones provide the information of the location and the surroundings. The robot sends local information such as video and sensory measurements to the drones, which can be delivered to the command center. The drones also send the command center high definition video taken from their cameras.

A mobile command center is assumed to be located in the fire truck It has multi-access edge computing (MEC) capability to process real-time immersive video to the incident commander sitting in the fire engine. If the fire break-out is in a large scale, the mobile command center may need wireless backhaul connectivity for communication with higher-level decision maker.

Operational requirements

- The firefighting robot is aware of its location and its surroundings.
- The firefighting robot receives timely commands from the incident commander.
- The incident commander and fire fighters have real-time immersive video of the fire scene.

Functional requirements

- Visual crowdsensing of the videos taken from the drones. Processing of the visual crowdsensing can be performed either at the drones or at the MEC located in the mobile command center depending on the computing power of the drones and the communication quality between the drone and the command center.
- Virtual reality (VR) video processing from the information obtained by the drones and the robot. The processing can be performed at the MEC in the command center.

Technical requirements (from the communication perspective)

- Communication between the drones for fleet control and visual crowdsensing,
- Communication between the drone and the firefighting robot for the awareness of fire scene and command messages (forward link), and for local information obtained from the robot

(reverse link),

- Communication between the drone and the mobile command center for the video taken from the drones and the relaying of local information (reverse link), and the relaying of command messages for the robot (forward link), and
- Wireless backhaul between the mobile command center and communication infrastructure.

6.2 Architecture design

The initial architecture design is described in this section includes 3GPP components such as 5G Core with all the network functions required for implementing network slicing together with ETSI defined MEC. In order to support PriMO-5G use cases, new solutions are proposed to deploy network slicing and MEC such as the optimal routing and MBO that includes SDN functionality enhanced with Machine Learning (ML). As described in previous section the Optimal Routing solution is proposed to improve the service continuity during mobility events while the MBO provides the interaction between the 3GPP Network Functions for network slicing and the physical transport network. Figure 23 shows not only the 3GPP and ETSI components required for implementing network slicing and MEC but also integration of Optimal Routing and MBO to deliver a complete 5G system that can meet the requirements and KPIs of PriMO-5G use cases. In this section we will describe how all the components in the proposed architecture will interact to provide the required functionality.

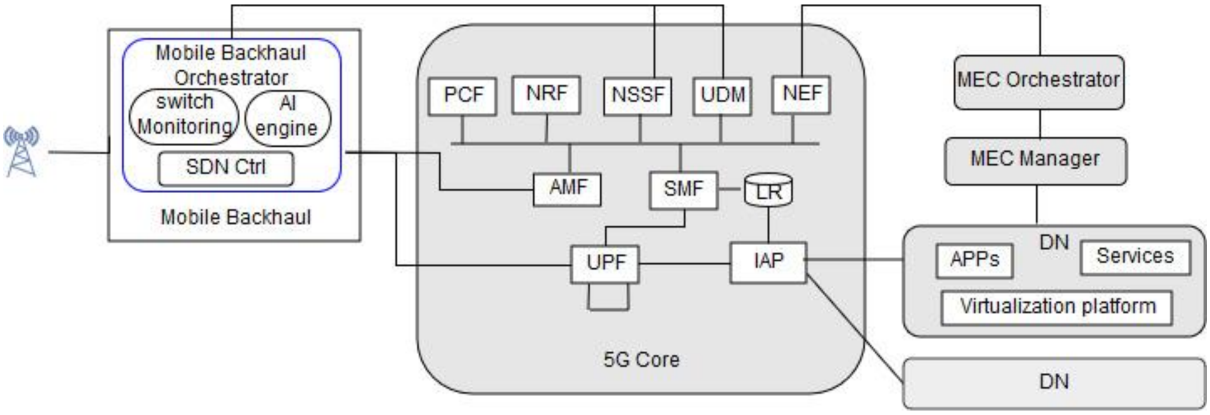


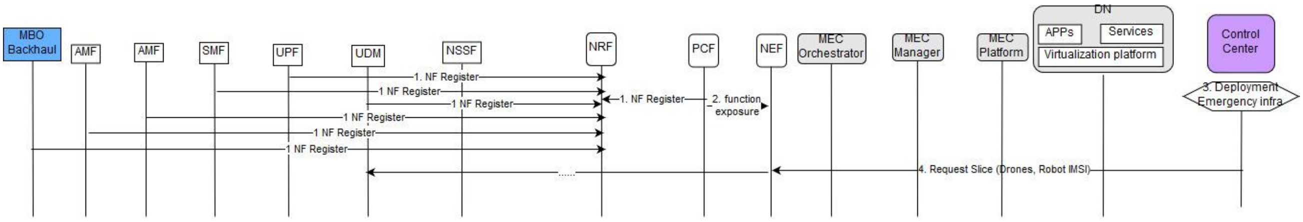
Figure 23. Network slicing and MEC integration in 5G architecture.

6.2.1 5G system message flows

This section provides the message flow to present the interaction between the different components that are part of the 5G system proposed to support PriMO-5G use cases. The system integrates MEC, network slicing, optimal routing and MBO to deliver the requirements needed but the firefighting use cases.

6.2.1.1 5G system initialization

This first step includes the flow during initialization where the system is deployed in the location of the use case. In this process all the network functions will be initialized and will proceed with the registration according to the 3GPP Service Based Architecture (SBA). In case there is no gNB available in the location of the fire, a set of gNB(s) might be deployed to ensure sufficient coverage. The Control center will interact with the 5G system to request the necessary network slice for the devices to operate with the necessary resources.



1. The network nodes e.g. SMF, UPF, PCF after connecting to the mobile network will register themselves to the NRF indicating available resources. Example of the JSON message used by the different network elements for the registration is shown in Appendix 1.
2. The PCF and UDM have to register to the Network Exposure Function to be accessible from external applications.
3. During the event of fire, the rescue team will deploy the basic mobile infrastructure including the control center in the location of the event. This setup is applicable in rural scenarios as shown in [PriMOD11].
    a. Other scenario to be supported is that rescue team uses existing mobile infrastructure from the MNO with good coverage in the location of the event. This setup is applicable in urban scenarios as shown in [PriMOD11].
    b. The Control Center has assigned URLLC SST for the devices to be used in the event
4. The control center requests a network slice to provide the low latency service for managing the drones.

6.2.1.2 UE registration and slice selection

This section describes the message flow when the UE (e.g. drone or firefighting robots register to the 5G system and get assigned the network functions (NF) i.e. AMF, SMF and UPF based on the network slice information. The UE might be registering for the first time in the system, so the slice selection has to be preloaded by the control center into the UE profile stored in the User Data Management (UDM).

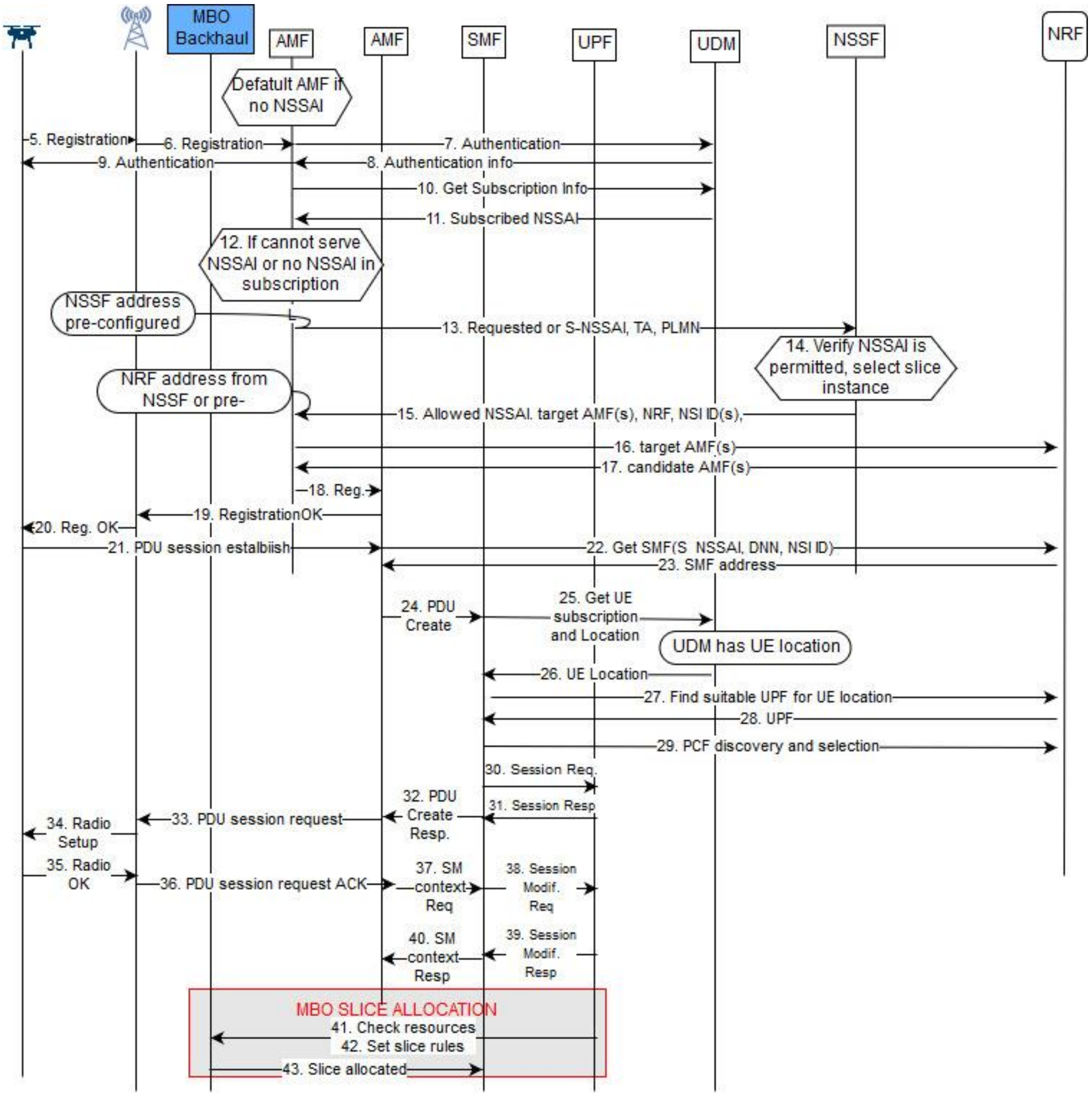
The control center can be fixed in a big truck but the infrastructure i.e. gNB can be movable and deployed on different locations during the event. The gNB can be moved with the rescue team or can be installed temporarily in the area of the fire to improve coverage. The drone could also be static to provide information and the gNB is moving.

The Network operator (e.g. through the Control Center) interacts with the NEF to register in the User Data Management (UDM) repository the Slice ID, the associated IMSI and requirements in terms of latency (i.e. ms), bandwidth (Mbs), etc.

The flow shows the process specified in 3GPP to discover the NF that will be allocated to the UE based on the network slice the Control centre assigned to the UE. This message flow does not include the usage of MEC because the requested network slice provides the required latency. The usage of MEC can be requested later if the firefighting application requires lower latency than the originally requested and is provided by the assigned network slice.
Thus, MEC will be added in the message flow and user data plane when required by the services or applications used to manage the UE during firefighting use cases.

However, if the slice assigned fulfils the needed requirements MEC might not be needed and normal network slice allocation flow will take place.

In this scenario the MBO will be required to deploy the network slice requirements into physical resources allocated by the transport network. NOTE: The numbering of the messages on the following figures in this section continue the numbering of the messages from the previous sections.



5. The UE (i.e. drone) initiates the registration towards gNB without NSSAI information
6. The gNB uses the default AMF for handling the registration process. This is the generic approach, but optimized solution could be based on dedicated gNB with pre-configured AMF for fire-fighting scenarios.
7. The AMF connects to the UDM for the UE authentication
8. The UDM returns the security tokens to the AMF to authenticate the UE

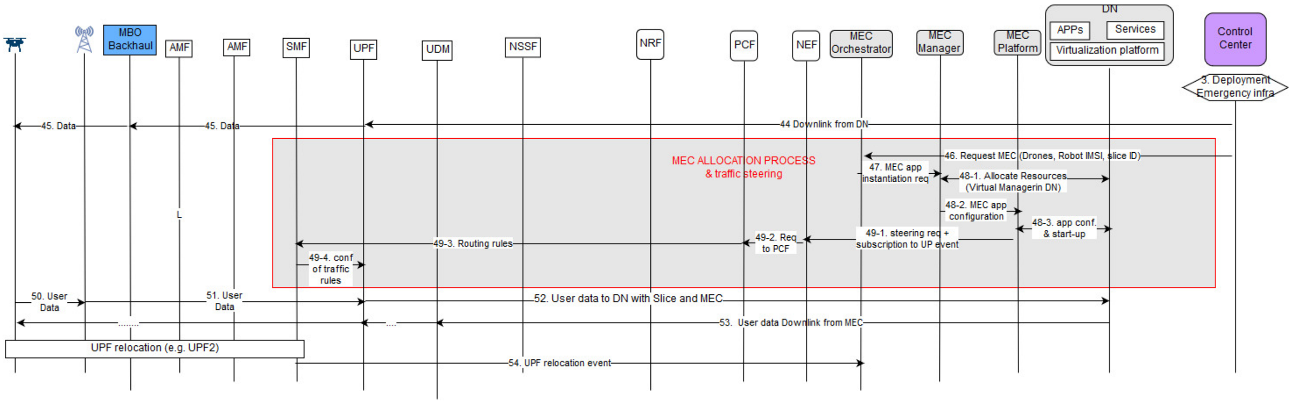
9. The AMF completes the authentication of the UE
10. The AMF gets the UE subscription and other service information and UE profile from the UDM.
11. The UDM returns the NSSAI assigned for the UE, which was added to the UDM by the emergency control center. Since there is no NSSAI provided by the UE during registration, check the assigned NSSAI to the UE in the UDM.
12. The AMF checks the NSSAI returned from UDM and checks that it cannot serve the slice associated to the NSSAI or there was no NSSAI assigned to the UE in the UDM.
13. The AMF has pre-configured the NSSF IP address and sends a request to get the slice information.
14. The NSSF verifies whether the requested slice is permitted based on the slice ID provided by the AMF in the request.
15. The NSSF returns to the AMF the confirmation that requested slice is permitted to the UE. The NSSF includes the information about the target AMF and the NRF IP so the AMF can request the information about the target AMF using example of JSON file in Appendix 2.
16. The AMF will send request to NRF to find info about the target AMF
17. The NRF will return the AMF the IP address of the target AMF
18. The default AMF will forward the registration to the target AMF to complete the registration.
19. The AMF responds with the registration complete to the gNB
20. The gNB will send the registration complete message to the UE

21. The UE sends a PDU session request to the AMF including slice and data network information.
22. The AMF sends a request to the NRF to find suitable SMF
23. The NRF responds with the address of suitable SMF
24. The AMF connects to the SMF assigned to the slice (Nsmf_PDUSession_CreateSMContext_Request procedure 4.3.2.2.1 in 23.502)
25. The SMF retrieves the subscription from the UDM.
26. The SMF responds back to the AMF (either it has accepted or rejected the request).
27. The SMF request information from NRF about available UPF that can meet the slice requirements (The SMF can use the UE location to find suitable UPF).
28. The NRF returns the UPF information that can meet the requirements.
29. The SMF connects to the NRF to discover the PCF that will subscribe for events.
30. The SMF starts the PDU session with the UPF.
31. The UPF responds with successful PDU session creation.
32. The SMF forwards the PDU session creation to the AMF
33. The AMF sends the PDU session creation to the gNB
34. The gNB sends the PDU session creation to the UE
35. The radio resources have been allocated between UE and gNB
36. The gNB sends the PDU session request acknowledge to the AMF.
37. The AMF sends the session context request to the SMF
38. The SMF sends the Session modification request to the UPF
39. The UPF sends the Session modification response to the SMF
40. The SMF sends the session context response to the AMF
41. The UPF sends a request to the Mobile Backhaul Orchestrator to check available resources for the required slice. The Monitoring system interacts with the backhaul to check available resources (i.e. the Monitoring system uses Link Layer Discovery Protocol (LLDP) to collect the information from backhaul switches).

42. The MBO will interact with the ML whether the available resources can be optimized and with the SDN controller whether new rules can be assigned in the backhaul to provide the UE with the network slice.
43. The MBO informs the UPF the network resources have been allocated and user data can be exchanged through the slice.

6.2.1.3 MEC setup

After the device is registered and got a network slice allocated the PriMO-5G application still requires lower latency and MEC platform has to be allocated to the UE. Following the standard registration process, the UE has been allocated AMF, SMF, UPF and network slice. No MEC platform has been allocated yet and mobile has all the NF allocated based on the network slice information received from the UE or stored in the UDM by the Control center. The Control center requires certain latency which is initially requested as part of the slice selection (i.e. SST =2, URLLC). However, during the initial transactions between the Control Center and the UAV the latency can be measured and if does not meet the expected requirements MEC can be requested as shown in next flow. NOTE: The numbering of the messages on the following figures in this section continue the numbering of the messages from the previous sections.


MBO Backhaul
AMF
AMF
SMF
UPF
UDM
NSSF
NRF
PCF
NEF
MEC Orchestrator
MEC Manager
MEC Platform
DN
APPs
Services
Virtualization platform
Control Center
3. Deployment Emergency infra
44 Downlink from DN
45. Data
45. Data
MEC ALLOCATION PROCESS & traffic steering
46. Request MEC (Drones, Robot IMSI, slice ID)
47. MEC app instantiation req
48-1. Allocate Resources (Virtual Managerin DN)
48-2. MEC app configuration
48-3. app conf. & start-up
49-1. steering req + subscription to UP event
49-2. Req to PCF
49-3. Routing rules
49-4. conf of traffic rules
L
50. User Data
51. User Data
52. User data to DN with Slice and MEC
....
....
53. User data Downlink from MEC
UPF relocation (e.g. UPF2)
54. UPF relocation event

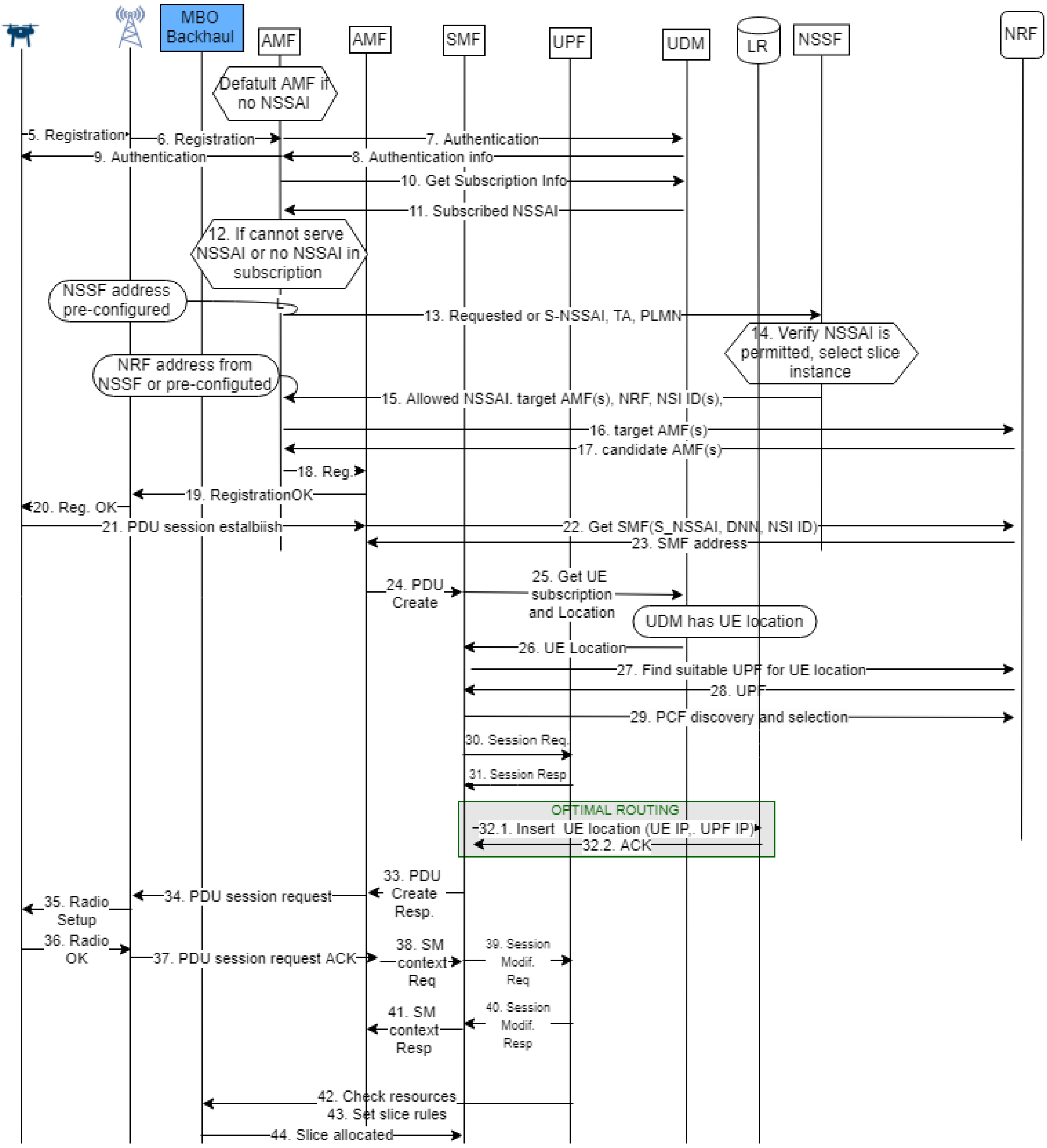
44. Downlink data is received from the DN where the Control Center is running the UPF. The downlink data might come from Control Center or application in the DN to operate the UE.
45. The UE receives the downlink packet from the UPF through the gNB
46. The control center realize the latency has to be improved and request from MEC orchestrator to allocate some processing close to the UE. The MEC orchestrator has to set up MEC support for a selected IMSI in a given slice ID
47. The MEC orchestrator interact with the MEC manager to instantiate the MEC application with the required resources.

    The MEC manager interacts with the Data Network platform to allocate the resources for hosting the application (step 48-1) and with the MEC platform for configuring the application (step 48-2). When the MEC application is instantiated, it can start-up (step 48-3).
48. The MEC platform request binding the instantiated MEC application with the target UPF(s) and subscribing to UP events. This request is addressed to the PCF through the NEF (steps 49-1 and 49-2). The PCF transforms the request into policies and provides the necessary routing rules to the appropriate SMF (step 49-3). The SMF identifies the target UPF(s) and initiate configuration of the traffic rules.
49. The UE sends uplink data to the gNB

50. The gNB forwards the uplink data to the UPF
51. The UPF forwards the uplink data to the Data Network where the application is running.
52. The UPF sends the downlink data to the UE via the gNB
53. In case of handover where the UPF has to be reallocated the MEC will receive a notification.

6.2.1.4 Optimal Routing

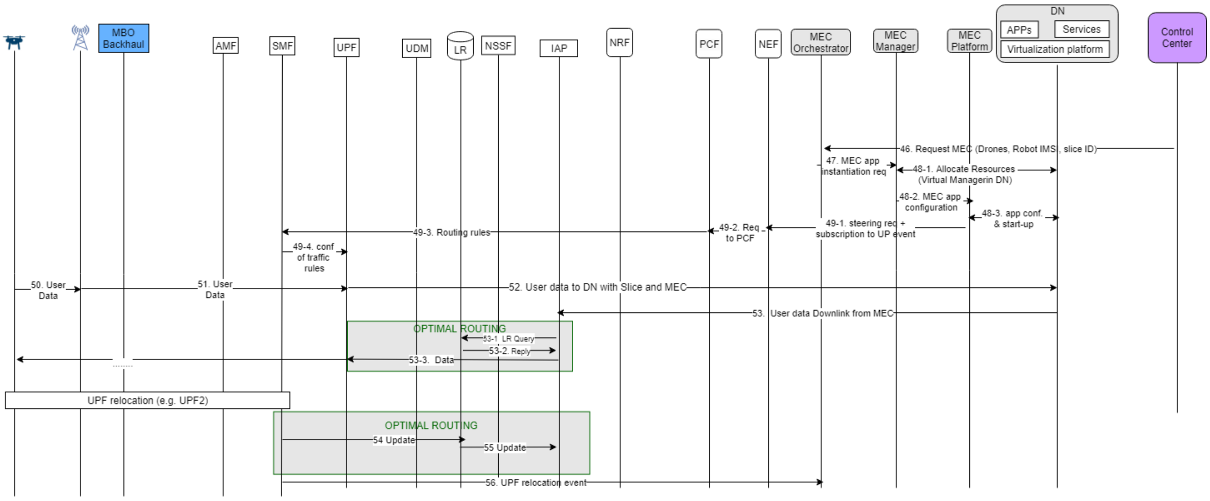
The section presents the call flow that is based on the normal setup which has been extended with a new proposal (Optimal Routing) which improves the session and service continuity and enables optimal paths after mobility. NOTE: The numbering of the messages on the following figures in this section continue the numbering of the messages from the previous sections.



The message flow diagram is similar to the previous ones but new messages are added for integrating Optimal Routing. The first change happens when the SMF starts the session with the UPF:

32.1 The SMF will insert the UE location (UPF IP address) into the LR

32.2 The LR acknowledges the registry of the UE in the LR


MBO Backhaul
AMF
SMF
UPF
UDM
LR
NSSF
IAP
NRF
PCF
NEF
MEC Orchestrator
MEC Manager
MEC Platform
DN
APPs
Services
Virtualization platform
Control Center
46. Request MEC (Drones, Robot IMSI, slice ID)
47. MEC app instantiation req
48-1. Allocate Resources (Virtual Managerin DN)
48-2. MEC app configuration
48-3. app conf. & start-up
49-1. steering req + subscription to UP event
49-2. Req to PCF
49-3. Routing rules
49-4. conf of traffic rules
50. User Data
51. User Data
52. User data to DN with Slice and MEC
53. User data Downlink from MEC
OPTIMAL ROUTING
53-1. LR Query
53-2. Reply
53-3. Data
....
UPF relocation (e.g. UPF2)
OPTIMAL ROUTING
54 Update
55 Update
56. UPF relocation event

An additional change is applied when the downlink data comes from the MEC application:

53. The application in the MEC sends the data downlink to the closest IAP

    53.1 The IAP queries the LR for the UPF IP address, where the UE has an established user plane session (this step is optional – only happens when the specific IAP does not have the information stored in its local cache)

    53.2 The LR, if the previous step was executed, responds the IAP the IP address of the UPF where the UE has an established user plane session

    53.3 The IAP forwards the downlink data to the UPF

    Note: Querying the Location Register only happens when the IAP does not have an entry stored. This means that subsequent packets typically will not trigger LR query.

As an additional change, the Location Register and the IAP(s) are updated that a new UPF is serving the UE:

54. The SMF updates the Location Register

55. The Location Register updates the IAP(s) that have active entries in their local caches

7 Conclusions

The main objective of this deliverable is to describe the architecture proposed to support the deployment of PriMO-5G use cases. In order to ensure the requirements from PriMO-5G use cases are met, the proposed architecture includes the integration of network slicing with MEC. In addition to the standard solutions defined in 3GPP and MEC, the PriMO-5G partners are contributing with new technologies such as Optimal Routing and MBO to fulfil the KPIs identified from the use cases. As a result, the PriMO-5G architecture will deliver the required enablers to deploy PriMO-5G use cases. This deliverable includes first proposal that combines network slicing with MEC, which has been specified separately in different standardisation forums. The PriMO-5G project intends to use both in addition to new contributions from the project partners to enhance the system and meet the KPIs set in PriMO-5G use case [PriMOD11]. The deliverable includes a technical overview and goes beyond to explore different possibilities to deliver different technologies together. There are still open questions regarding MEC and network slicing integration which will be analyse in depth during the project through prototypes and other validations methods.

8 References

[EMECdep] ETSI MEC Deployment towards 5G(https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp24_MEC_deployment_in_4G_5G_FINAL.pdf)

[EMECspec] ETSI MEC Specifications (https://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing)

[EMECreq] ETSI MEC Requirements (https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/01.01.01_60/gs_MEC002v010101p.pdf)

[EMECwp] ETSI MEC in 5G white paper (https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf)

[NSliceGSM] An Introduction to Network Slicing by GSMA, https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf.

[NSliceNGMN] Description of Network Slicing Concept by NGMN Alliance, https://www.ngmn.org/fileadmin/user_upload/161010_NGMN_Network_Slicing_framework_v1.0.8.pdf

[NSlice3GPP] 3GPP TS 23.501 (https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144)

[ETSIgrMEC018] Mobile Edge Computing (MEC); End to End Mobility Aspects (https://www.etsi.org/deliver/etsi_gr/MEC/001_099/018/01.01.01_60/gr_MEC018v010101p.pdf)

[ETSIspecs] ETSI MEC Specifications (https://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing)

[ETSIreq] ETSI MEC Requirements (https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/01.01.01_60/gs_MEC002v010101p.pdf)

[ETISwp] ETSI MEC in 5G white paper (https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf)

[ETSIdep] ETSI MEC Deployment towards 5G (https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp24_MEC_deployment_in_4G_5G_FINAL.pdf)

[ETSI_WP28] MEC in 5G networks, ETSI White Paper No. 28 (https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf)

[3GPP23.501] The 3GPP TS 23.502 defines Stage-2 System Architecture for the 5G System

[3GPP23.502] TS 23.502 defines procedures for the 5G System.

[3GPP23.503] TS 23.503 defines Policy and Charging Control Framework for the 5G System.

[3GPP28.532] TS 28.532 Management and orchestration of networks and network slicing

[SDN_MBO]	SDN-based UPF for Mobile Backhaul Network Slicing, EuCNC 18 18-21 June, Ljubljana, Slovenia (https://ieeexplore.ieee.org/document/8442795)
[PriMOD11]	D1.1 - PriMO-5G-5G use case scenarios

9 Appendix

9.1 Appendix 1

The body of the registration is a JSON document similar to the one below which in this case is used to register the SMF.

```
{"nfInstanceID": 3fa85f64-571-ddd-",
                  "nfType": ["SMF"],
                    "nfStatus": ["REG"],
                    "sNssais": [{"sst": 1, "sd": "sd1"}],
                 "nsiList": ["NSI ID1"],
                        "ipv4Addresses": [
                           "198.52.100.1"],
                         "allowedNssais": [{ "sst": 0, "sd": "string"}],
                 "priority": 0,
              "load": 0,
              "smfInfo": {
                           "sNssaiSmfInfoList": [{
                         "sNssai": { "sst": 0, "sd": "string"}  }],
                      "taiList": [{
                            "plmnId": {
                                     "mcc": "string",
                                     "mnc": "string"},
                               "tac": "string" }]  } } }
```

9.2 Appendix 2

Example of the JSON file returned by the NSSF is shown below.

```
{"allowedNssaiList": [{
         "allowedSnssaiList": [{
           "allowedSnssai": { "sst": 1, "sd": "sd1" }
            "nsiInformationList": [{
               "nrfId": "http://localhost:5050/nnrf-disc/v1/nf-instances" ,
               "nsiId": "NSI ID1" }]  }] }],
      "targetAmfSet": "string",
      "candidateAmfList": [
         "3fa85f64-5717-4562-b3fc-2c963f66afa6",
         "5fa85f64-5717-4562-b3fc-2c963f66afa5"] },
{"allowedNssaiList": [{
         "allowedSnssaiList": [{
           "allowedSnssai": {"sst": 2, "sd": "sd1"},
            "nsiInformationList": [{
               "nrfId": "http://localhost:5050/nnrf-disc/v1/nf-instances" ,
               "nsiId": "NSI ID2" }]  }] }],
      "targetAmfSet": "string",
      "candidateAmfList": [
         "3fa85f64-5717-4562-b3fc-2c963f66afa6",
         "4fa85f64-5717-4562-b3fc-2c963f66afa7"] }
```