11.3 5G CORE NETWORK

Core networks traditionally have been designed as a single architecture addressing a range of requirements and supporting backward compatibility. This one size fits all approach has been successful in keeping the costs down to a reasonable level and by supporting legacy circuit switched and today's packet switched functionalities.

This core network, however, is rigid in the sense that it is not flexible enough to accommodate the customized and variable connectivity needs of individual users and businesses that are expected in the future. However, with virtualization, NFV, SDN, and network slicing, it is possible to make core networks more flexible and scalable. Thus, the next generation core network is expected to exist in a cloud-based environment with a high degree of virtualization and software-based networking. Such flexibility is needed to support a variety of access networks such as 3G, LTE, 4G, WiFi, and tomorrow's 5G.

11.3.1 Components of Core Network/High Level Architecture

The current EPC will further evolve to support virtualization and network slicing to become NGC applicable for 5G networks.

Network slicing is often termed as logical instantiation of a network possibly due to virtualization technologies [23]. The concept is seen as the natural extension/evolution of the current network sharing methodologies [24]. Network slicing is one of the promising techniques that will likely exist in both radio access and core networks. It allows multiple logical networks to be created on top of a common physical infrastructure. Either DCN (Dedicated Core Network) or a combination of NFV and SDN can be used as a technology to enable network slicing along with orchestration and analytics [25,26]. DCN or Décor as defined in 3GPP TS 23.401 [27] is a feature that enables an operator to deploy multiple logical mobile core networks connected to the same RAT or multiple RATs (e.g. GERAN, UTRAN, E-UTRAN, WB-E-UTRAN and NB-IoT). A DCN consists of one or more MME/SGSN and it may be comprised of one or more SGW/PGW/PCRF. This feature enables subscribers to be allocated to and served by a DCN based on subscription information (e.g., "UE Usage Type").

With 5G, a single terminal can use multiple services with different characteristics almost simultaneously. In such cases, a network slice can be created for each service, requiring all such slices to coordinate control for that particular single terminal. These slices can be mapped to respective radio and core network slices to provide end-to-end connectivity. The methodology is currently being specified for selecting radio/core networks particularities for supporting slicing in existing as well as in future 5G systems.

Control and User Planes' Separation: The separation of control and user planes is one of the key principles of 5G core network architecture. 3GPP started a study in TR 23.714 [28] on user/control planes' separation involving core network elements such as P-GW, Traffic Detection Function (TDF), and so on. This separation allows independent scaling of each plane and migration toward cloud-based architecture. For example, the control plane can be placed in a centralized location with complex hardware and processing capabilities. On the other hand, the user plane can be distributed to a larger number of local sites making reachability from the perspective of a user easier. A good example of this will be content caching in local sites instead of securing it from the main server sitting thousands of miles away. This separation is the fundamental concept of SDN and having SDN will make core networks more flexible.

Further details on network slicing, NFV, and SDN can be found in Chapter 5 5G Concepts.

4.1 5G System Architecture

As in the 4G (LTE/EPC) and previous generations, the 3GPP 5G system defines the architecture for communication between a User Equipment (UE) and an end point, such as an Application Server (AS) in the Data Network (DN), or another UE. The interaction between the UE and the Data Network is via the Access Network and Core Network as defined by 3GPP Standards. Figure 4.1 depicts a simple

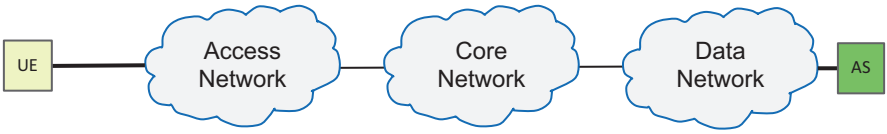


Fig. 4.1 End-to-end architecture

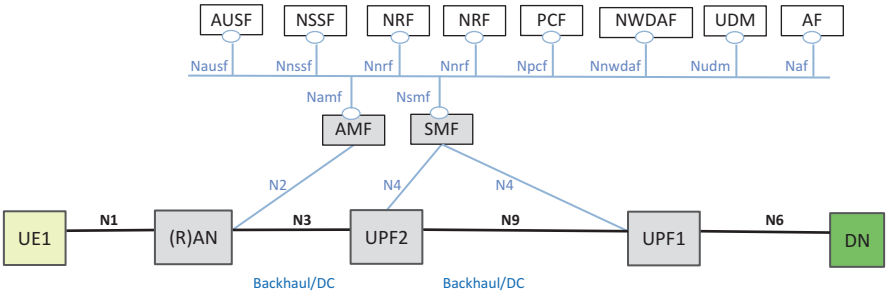


Fig. 4.2 5G System architecture

representation of an end-to-end architecture. In this chapter we focus on describing the 5G Core as defined by 3GPP 5G standards for PLMN [1–3]. The Access Network in 3GPP is referred to as Radio Access Network (RAN).

At a very high level the Core and RAN consist of several Network Functions which are associated with Control Plane and User Plane functionalities. The actual data (also refer it as user data) is normally transported via a path in the User Plane, while the Control Plane is used to establish the path in the User Plane. The Short Message Service (SMS) is an exception in which the data (short message) is communicated via the Control Plane.

The 5G System architecture (5GS) is represented in two ways in the 3GPP standards, one is a service-based representation in which the control plane network functions access each other's services, and the other is a reference point representation in which the interaction between the network functions is shown with point-to-point reference points. In this chapter we use the service-based representation since the 5G architecture is defined as service-based architecture. The 3GPP 5GS service-based non-roaming reference architecture in shown in Fig. 4.2. In Release-15 specifications the Service-based interfaces are defined within the Control Plane only. In 3GPP terminology, "a network function can be implemented either as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, or as a virtualized function instantiated on an appropriate platform, e.g. on a cloud infrastructure."

The EPC in Release-14 was enhanced with an optional feature that allowed separation of control plane and user plane. In this feature, the Serving Gateway (SGW) and Packet Gateway (PGW) are divided into distinct control plane and user plane functions (e.g., SGW-C and SGW-U). This optional feature provided more flexibility and efficiency in network deployment—See [4] for details. In 5G architecture, the separation of control plane and user plane is an inherent capability. The Session

Management Function (SMF) handles the control plane functionality for setup and management of sessions while the actual user data is routed through the User Plane Function (UPF). The UPF selection (or re-selection) is handled by SMF. The deployment options allow for centrally located UPF and/or distributed UPF located close to or at the Access Network.

In EPC, the mobility management functionality and session management functionality are handled by Mobility Management Entity (MME). In 5GC, these functionalities are handled by separate entities. The Access and Mobility Management function (AMF) handles the mobility management and procedures. AMF is termination point for control plane connection from (Radio) Access Network ((R)AN) and UE. The connection between UE and AMF (which traversed through RAN) is referred to as Non-Access Stratum (NAS). The Session Management Function (SMF) handles the session management procedures. The separation of the mobility and session management functionalities allows for one AMF to support different Access Networks (3GPP and non-3GPP), while SMF can be tailored for specific Accesses.

Figure 4.3 shows a Roaming architecture with local breakout at the Visited PLMN (VPLMN). In this scenario the Unified Data Management (UDM), which includes the subscription information, and Authentication Server Function (AUSF), which includes authentication/authorization data, are located in the Home PLMN (HPLMN). There are Security Edge Protection Proxies (SEPP) that protect the communication between the Home and Visited PLMNS. UE communicates to Data Network (DN) via the User Plane Functions (UPF) in the VPLMN. The AMF and the Session Management Function (SMF) which handle the mobility and the session management for the UE are located in the VPLMN as well.

4.2 5G Core (5GC) Service-Based Architecture

A major change in the 5G Core (5GC) architecture compared to EPC and the previous generations is the introduction of the service-based architecture. In EPC architecture the control plane functions communicate with each other via the direct interfaces (or reference points) with a standardized set of messages. In the service-based architecture, the Network Functions (NF), using a common framework,

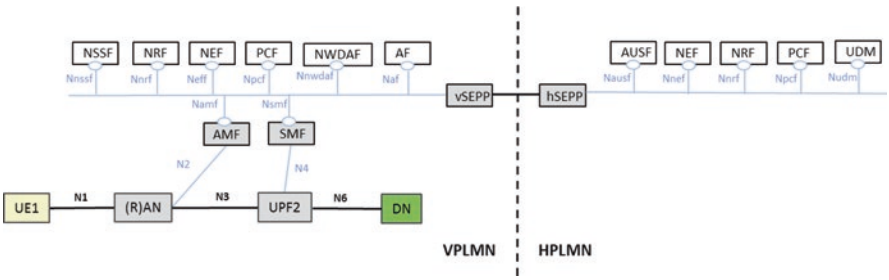


Fig. 4.3 Roaming 5G System architecture—Local breakout scenario

expose their services for use by other network functions. In the 5GC architecture model the interfaces between the networks functions are referred to as Service-Based Interfaces (SBI). The Service Framework defines the interaction between the NFs over SBI using a Producer-Consumer model. As such a service offered by a NF (Producer) could be used by another NF (Consumer) that is authorized to use the service. The services are generally referred to as "NF Service" in 3GPP specifications.

The interaction between the NFs may be a "Request-response" or a "Subscribe-Notify" mechanism. In the "Request-response" model a NF (consumer) request another NF (producer) to provide a service and/or perform a certain action. See Fig. 4.4. In "Subscribe-Notify" model a NF (consumer) subscribes to the services offered by another NF (producer) which notifies the subscriber of the result (Fig. 4.5).

As can be seen in Fig. 4.3, in the 5G System Architecture, each network function has an associated service-based interface designation. For example, "Namf" designates the services exhibited by the Access and Mobility Management function (AMF). 3GPP specifications define a set of Services that are offered/supported by each Network Function. For example, the NF services specified for AMF are shown in Table 4.1. The details for Service descriptions are described in [2].

There are three main procedures associated with the Service Framework as defined in 3GPP—see [1, 5] for details:

NF service registration and de-registration: to make the Network Repository Function (NRF) aware of the available NF instances and supported services.

NF service discovery: enables a NF (Consumer) to discover NF instance(s) (Producer) that provide the expected NF service(s). A NF typically performs a Services Discovery procedure with NRF for NF and NF service discovery.

NF service authorization: to ensure the NF Service Consumer is authorized to access the NF service provided by the NF Service Provider (Producer).

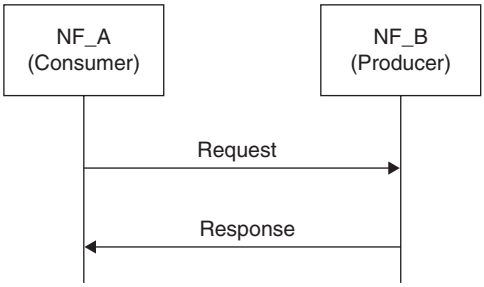


Fig. 4.4 "Request-response" NF Service illustration

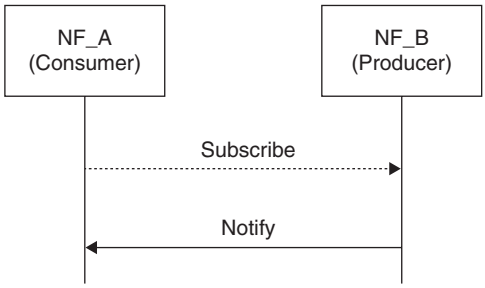


Fig. 4.5 "Subscribe-Notify" NF Service illustration 1

Table 4.1 Namf services

Service name	Description
Namf_ communication	Enables an NF consumer to communicate with the UE and/or the AN through the AMF This service enables SMF to request EBI allocation to support interworking with EPS
Namf_ EventExposure	Enables other NF consumers to subscribe or get notified of the mobility-related events and statistics
Namf_MT	Enables an NF consumer to make sure UE is reachable
Namf_Location	Enables an NF consumer to request location information for a target UE

4.2.1 Example of NF Service Registration

In this example (Fig. 4.6) AMF as the NF service consumer sends a HTTP PUT request to NRF with the resource URI representing the NF Instance. The request contains Nnrf_NFManagement_NFRegister request message (the NF profile of NF service consumer) to NRF to inform the NRF of its NF profile. The NF profile of NF service consumer includes information such as NF type, FQDN or IP address of NF, and Names of supported services. The NRF authorizes the request and upon success stores the NF profile of NF service consumer and marks the NF service consumer available. The NRF acknowledge the success of AMF Registration by returning a HTTP 201Created response containing the Nnrf_NFManagement_NF Register response (including the NF profile). See 3GPP TS 23.501 [1] and 3GPP TS 29.510 [5] for details.

4.2.2 Example of NF Service Discovery

In this example (Fig. 4.7) the AMF as NF service consumer intends to discover NF instances or services available in the network for a targeted NF type. The AMF sends HTTP GET request to NRF in the same PLMN by invoking Nnrf_ NFDiscovery_Request. This request contains Expected NF service Name, NF Type of the expected NF instance, and NF type of the NF consumer and may also include

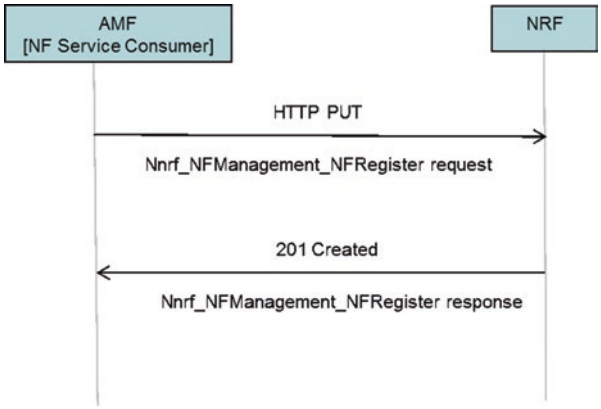


Fig. 4.6 Nnrf_NF Registration procedure

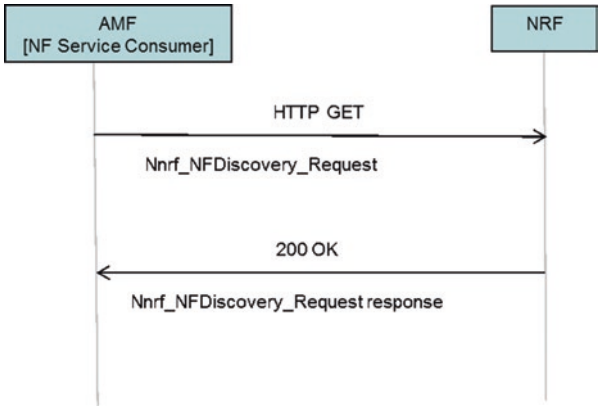


Fig. 4.7 Nnrf_NF service Discovery

other information/parameters such as Subscription Permanent Identifier (SUPI) and AMF Region ID. The NRF authorizes the request, and if allowed the NRF determines the discovered NF instance(s) or NF service instance(s) and provides the search results to the NF service consumer (e.g., AMF) in a HTTP 200 OK. See 3GPP TS 23.501 [1] and 3GPP TS 29.510 [5] for details.

4.3 Network Slicing

From the 3GPP point of view, a 5G network slice is viewed as a logical network with specific functions/elements dedicated for a particular use case, service type, traffic type, or other business arrangements with agreed-upon Service-level

Agreement (SLA). It is important to note that 3GPP only defines network slicing for 3GPP defined system architecture and does not address transport network slicing or resource slicing of components.

The most commonly discussed slice types in industry are enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and massive IoT (mIoT). However, there could be many more network slices. In 4G systems (EPS) there is an optional feature called eDecor to support Dedicated Core Networks (DCNs) to allow selection of the core networks based on UE's subscription and usage type. The network slicing in 5GS is a more complete solution that provides capabilities for composing multiple dedicated end-to-end networks as slices.

An end-to-end Network Slice includes the Core Network Control Plane and User Plane Network Functions as well as the Access Network (AN). The Access Network could be the Next Generation (NG) Radio Access Network described in 3GPP TS 38.300 [6], or the non-3GPP Access Network with the Non-3GPP InterWorking Function (N3IWF). To emphasize that there could be multiple instances of a network slice, the 3GPP 5GS specifications define the term "Network Slice instance" as set of Network Function instances and resources (e.g., compute, storage, and networking resources) which form a Network Slice.

In 5GS, the Network Slice Selection Assistance Information (NSSAI) is a collection of identifications for network slices. A network slice is identified by a term referred to as Single-NSSAI (S-NSSAI). The S-NSSAI signaled by the UE to the network assists the network in selecting a particular Network Slice instance. An S-NSSAI comprises a Slice/Service type (SST) and an optional Slice Differentiator (SD) which may be used to differentiate among multiple Network Slices of the same Slice/Service type.

An S-NSSAI can have standard values or nonstandard values. The S-NSSAI with standard value means that it comprises an SST with a standardized SST value. An S-NSSAI with a nonstandard value identifies a single Network Slice within the PLMN with which it is associated.

3GPP has defined some standardized SST values in TS 23.501 [1]. These SST values are to reflect the most commonly used Slice/Service Types and will assist with global interoperability for slicing. The support of all standardized SST values is not required in a PLMN (Table 4.2).

Figure 4.8 shows an example of three Network Slices in 5GS. For Slice 1 and Slice 2, the Access and Mobility Management Function (AMF) instance that is

Table 4.2 Standardized SST values

Slice/service type	SST value	Characteristics
eMBB	1	Slice suitable for the handling of 5G enhanced mobile broadband
URLLC	2	Slice suitable for the handling of ultra-reliable low latency communications
MIoT	3	Slice suitable for the handling of massive IoT

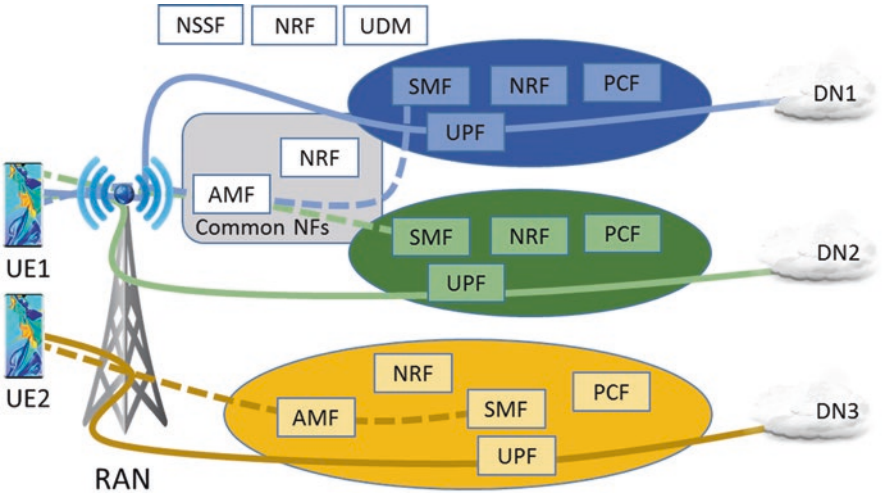


Fig. 4.8 Example of Network Slices in 5GS

serving the UE1 and UE2 is common (or logically belongs) to all the Network Slice instances that are serving them. The UE in Slice 3 is served by another AMF. Other network functions, such as the Session Management Function (SMF) or the User Plan Function (UPF) may be specific to each Network Slice.

The Network Slice instance selection for a UE is normally triggered as part of the registration procedure by the first AMF that receives the registration request from the UE. The AMF retrieves the slices that are allowed by the user subscription and may interact with the Network Slice Selection Function (NSSF) to select the appropriate Network Slice instance (e.g., based on Allowed S-NSSAIs, PLMN ID). The NSSF contains the Operators' policies for slice selection. Alternatively, the slice selection policies may be configured in the AMF.

The data connection between the UE and Data Network (DN) is referred to as PDU session in 5GS. In 3GPP Release-15 a PDU Session is associated to one S-NSSAI and one DNN (Data Network Name). The establishment of a PDU session is triggered when the AMF receives a Session Management message from UE. The AMF discovers candidate Session Management Functions (SMF) using multiple parameters (including the S-NSSAI provided in the UE request) and selects the appropriate SMF. The selection of the User Plane Function (UPF) is performed by the SMF. The Network Repository Function (NRF) is used for the discovery of the required Network Functions using the selected Network Slice instance—the detailed procedures are specified in 3GPP TS 23.502 [2]. The data transmission can take place after a PDU session to a Data Network is established in a Network Slice. The S-NSSAI associated with a PDU Session is provided to the (R)AN, and to the policy and charging entities, to apply slice specific policies.

For roaming scenarios, S-NSSAI values applicable in the Visited PLMN (VPLMN) are used to discover a SMF instance in the VPLMN and in Home–Routed deployments S-NSSAI values applicable in the Home PLMN (HPLMN) are also used to discover a SMF instance in the HPLMN.

5.7 NETWORK SLICING

Network slicing is an end-to-end logical instance of a network with at least the following attributes [108–113]:

a. Runs on a physical or virtual network
b. Optimizes use of network for each intended usage scenario
c. Uses a set of access and core network functions
d. Is controlled and managed independently
e. Created on demand, and
f. Does not interfere with other functions and services on coexisting slices

Network softwarization is an emerging trend that seeks to transform a network's designing, planning, implementation, and operations through software programming. Network softwarization provides necessary flexibility and modularity to create logical networks (i.e., network slices) through NFV (Network Functions Virtualization) and SDN (Software Defined Networking) technologies [109,111].

The global standardization efforts are in the early stage and are primarily focused on vertical slicing (defined later). Industry forums such as 5GPPP (5G Infrastructure Public Private Partnership) and NGMN (Next Generation Mobile Networks) Alliance are leading the way, while SDOs like

3GPP have just started work [114]. Network slicing is identified as a key technology for 5G, but a great amount of work is needed to turn it into a successful reality.

5.7.1 E2E Slicing

Network slicing can be considered an evolution of network sharing, which is a proven business model for operators to reduce CAPEX and OPEX. Network slicing goes beyond sharing and envisions using virtualization and softwarization to improve user experience, increase network usage, and add on to operators' revenues. The slicing manifests the resolution of many complex issues including slice design, instantiation, implementation, and operations that requires new thinking. For instance, the design of a slice for a particular case requires enablement of functions in control and data planes, instantiation demands provision mechanisms over infrastructure, implementation seeks availability of all required network elements and the respective functions, while for operations slice configuration and monitoring setups are required in addition to other procedures.

Each network slice may consist of air-interface, radio access, and core networks as shown in Figure 5.11. Fronthaul in the case of C-RANs and transport area may have to be considered for some cases. Furthermore, virtualization and softwarization through NFV and SDN are also key building blocks of network slicing. The slicing is realized through network functions which provide the tailored capabilities needed to entertain a specific demand. Network functions could be slice-specific or could be used across multiple slices. The network functions could be physical (comprised of both hardware and software) or virtual (where software is decoupled from the hardware it runs on) [109].

In today's cut throat competition environment, operators are working to maximize their return on investment and utilization of their networks. E2E (end-to-end) slicing is an instrument which can achieve this goal, is making headlines, and is an important area of 5G research and development.

5.7.1.1 Slicing in RAN

A slice in the RAN relies on Radio Access Technology (RAT) and configuration of radio resources to carry what it is intended to deliver. There could be separate RATs in 5G designed to address different services. For example, one RAT could be for IoT and another perhaps for mobile broadband. RAN related configuration, which is customized to a particular slice, may include access control, load balancing, and resource scheduling [111,112]. To activate a slice, an access point/base station may allocate radio resources for the slice and enable all radio and network functions required for the operation of the slice. The slices in RAN may require slice-specific control-plane/user-plane and slice on/off operations [115].

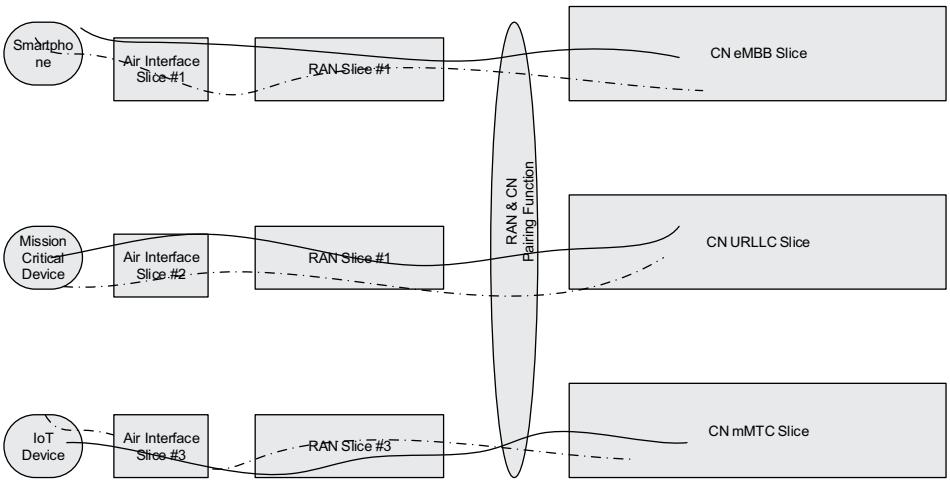


FIGURE 5.11 E2E network slice.

The slice in RAN will share radio resources such as time/frequency/space with other slices, along with the corresponding communication hardware either in a dynamic or static fashion according to the configuration rules. The radio resource sharing is performed by a central scheduler that resides in the 5G base station. If static resource sharing is used, the slice will get guaranteed resource allocation, while with dynamic resource sharing, usage is optimized. The type and amount of RAN resources required for a network slice depends on the service such as mobile broadband, IoT, autonomous driving, and their respective QoS requirements, that is, high capacity, massive connections, and ultra-low latency, respectively [111,112].

As far as C-plane (control plane) and U-plane (user plane) configurations are concerned, three alternatives are possible. Alternative 1 may allow a common C-plane across all slices and a dedicated U-plane for each slice. Alternative 2 may provide a dedicated C-plane and U-plane for each of the slices, while the third alternative would be a case with a common U-plane and a dedicated C-plane for each of the slices. Common C-plane slice functions include functions in idle mode such as paging, cell selection, and so on, while the functions in connected mode such as handover, dedicated bearer setup, and so on, can be categorized into slice-specific control plane functions [111,112,115].

5.7.1.2 Slicing in Core Network

Traditionally, the core networks have been architected as a single network serving multiple purposes and have been tailored to one or more RANs while supporting backward compatibility and interoperability. Network slicing, if implemented correctly, allows core networks to be logically separated making each core network slice operate independently while likely running on the same shared infrastructure.

A key element of 5G architecture is the separation of control and user planes' functions in the core network that allows selective choice of the U-plane functions needed for different slices and distribution of U-plane to sites closer to the devices, besides other features. The C-plane is agnostic to many U-plane functions such as physical deployments and L3 transport specifics. The control plane, as the name suggests, manages signaling messages, location information, cell selection/reselection, and so on. The C-plane can be placed in a central location, making management and operations less complex, whereas the user plane can be distributed to a number of sites. By bringing the U-plane closer to the users, the round-trip time between user and network services can be shortened. U-plane functionality can be deployed to address specific use cases. For example, an MBB (mobile broadband) service can be divided into video streaming and web browsing subservices which can be implemented by different feature sets within a network slice [112].

The slices can be defined to different support services/applications with a targeted set of radio/core network functions. Slice pairing functions are defined to pair RAN and core network (CN) slices to form end-to-end slices. Mapping among devices, RAN and CN slices, can be 1:1:1 or 1:M:N, for example, a device can have RAN slices while a RAN slice can connect to multiple CN slices [112,118].

5.7.2 SDN and NFV in Slicing

SDN and NFV are essential for the effective working of network slicing. SDN separates the C-plane and U-plane to optimize the performance of the network while NFV enables virtualization of networks.

5.7.2.1 SDN Overview

The SDN architecture defined by Open Networking Foundation (ONF) enables a common architecture to efficiently support diverse slices tailored for different services with different requirements. The SDN architecture is technology neutral, thus it can support wired, wireless, and mobile technologies. SDN consists of two key components, namely resources and controllers. Resources could be anything, but in this case, it could be a physical network function consisting of a piece of hardware and

software or a virtual network function (VNF) where the software is decoupled from the hardware for the realization of a particular slice. Resources are managed through controllers which are central entities in SDN architecture. A controller maintains isolation in the control and data planes, allowing each network slice instance to be operated as a distinct and logically separate network. It provides resource isolation as well for a multitude of slices running on a common infrastructure. In addition to resources and controllers, the following key concepts are critical for the implementation of SDN based networks [109,116]:

- *Virtualization* is a function of a controller to abstract the underlying resources it manages. Network virtualization assists in the creation of isolated virtual networks that are dissociated from the underlying physical network and run on top of it.
- *Orchestration* by definition brings disparate things into a coherent whole. Within slicing, it is defined as the responsibility of the controller to dispatch resources to address the demands of the client in an optimal manner.
- *Recursion*, which could be hierarchical or federated, allows the controller to use resources from a lower level controller and provides services to a higher level controller (hierarchical scheme) while federation works on an equal level.

SDN provides extensive control plane functions for enabling network slicing, however, to efficiently manage the lifecycle of slices and their constituent resources, NFV is needed.

5.7.2.2 NFV Overview

ETSI NFV standard GS (Group Specification) NFV 002 [117] defines the lifecycle management of network services. A network service according to ETSI GS NFV 003 [118] is a composition of network functions and is defined by its functional and behavioral specifications. Thus, the concept of lifecycle management can be reused for network slicing [112].

NFV envisions the implementation of network functions as software-only entities (virtualization) that run over the NFV Infrastructure (NFVI). Virtualization means that the Network Function (NF) and part of the network infrastructure are implemented in software, that is, a decoupling of software from hardware. This decoupling allows the independent evolution of each. Along the same lines, an end-to-end network service (voice, data, IoT, etc.) can be described by an NF Forwarding Graph* of interconnected NFs and end points. To facilitate virtualization, the NFV reference architectural framework defined by ETSI can be used. This framework enables dynamic construction and management of VNF instances. It also manages the relationships between VNFs related to data, control, and other attributes.

5.7.2.3 SDN + NFV for Slicing

To take advantage of the benefits of both SDN and NFV for network slicing, an appropriate cooperation between the two is required. However, integration of both entities into a common framework is not an easy job.

Consider a slice in LTE that may consist of a radio access component eNB, and core network components such as Mobility Management Entity (MME), Serving Gateway (SGW), Packet Gateway (PGW), and so on, to provide broadband to a population of mobile users. The eNB is a PNF (physical network function) while other entities could be either PNF or VNF. These PNFs and VNFs, which are resources as far as SDN is concerned, can be made part of the end-to-end network slice as defined as shown in Figure 5.12.

These VNFs of an LTE slice can interact with ETSI NFV MANO (Management and Orchestration) [119] framework (part of the NFV reference architectural framework defined in [117]). Figure 5.13

* A graph of logical links connecting NF nodes for the purpose of describing traffic flow between these network functions is an NF forwarding graph [118].

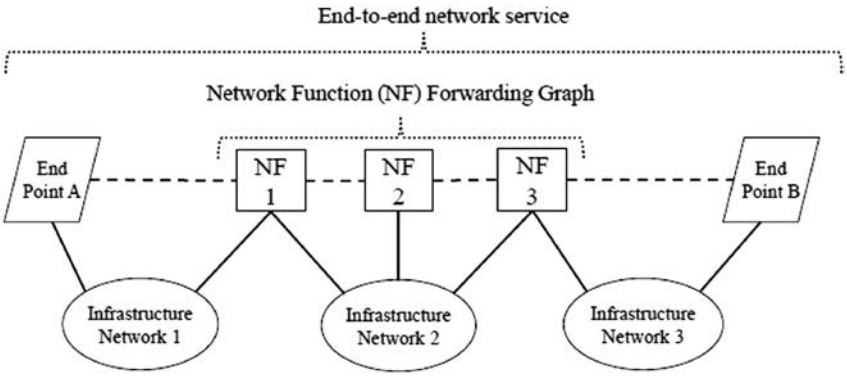


FIGURE 5.12 E2E network service. (From ETSI GS NFV 002 (V1.1.1) 2013. Network Functions Virtualisation (NFV); Architectural Framework, ETSI, October [117].)

shows a potential architecture which manages both virtualized and nonvirtualized network functions using NFV and SDN for 3GPP technologies. For the nonvirtualized components or PNFs, NFV MANO may not be used and thus a separate lifecycle management is required. PNFs are not under the control of MANO, but are also shared between numerous slices, particularly the PNF part of the radio access network. 3GPP currently has a study item on the management of such nonvirtualized

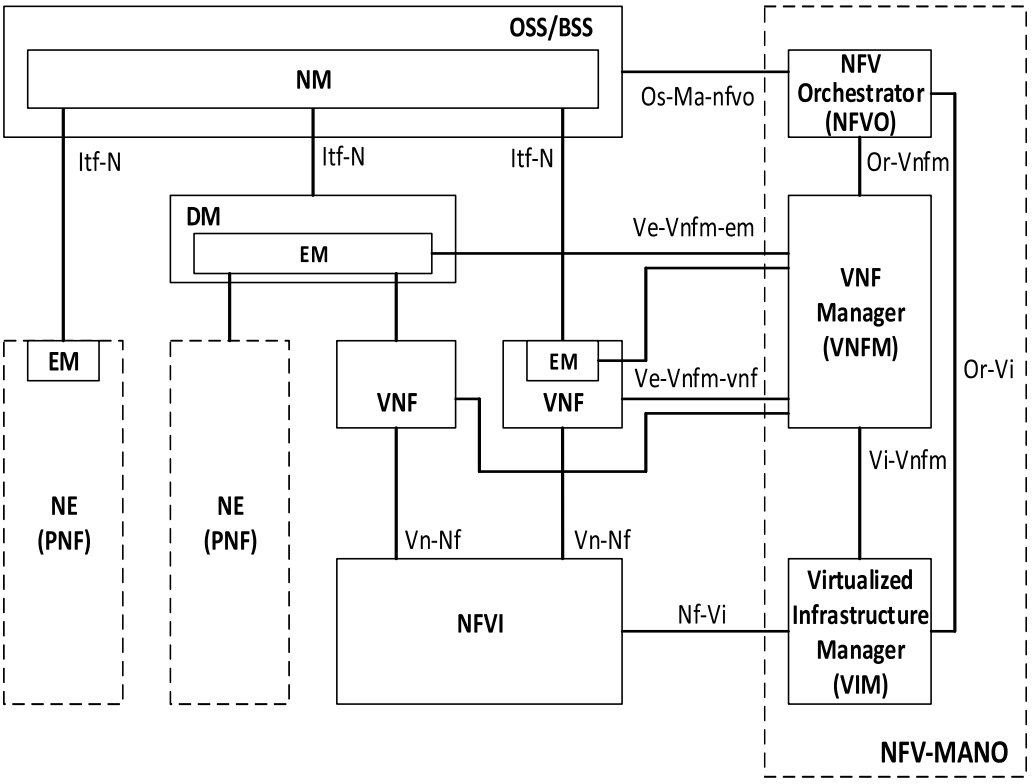


FIGURE 5.13 Mixed network management mapping relationship between 3GPP and NFV-MANO architectural framework. (From 3GPP TR 32.842 (V13.1.0) 2015. Telecommunication Management; Study on Network Management of Virtualized Networks. Technical Report (Release 13), Technical Specification Group Services and System Aspects, 3GPP, December [120].)

parts [120]. The same report, that is, TR 32.842, also clarifies the relationship between the 3GPP management framework and NFV-MANO framework.

3GPP defined VNFs will still require some underlying hardware. These hardware resources are expected to be managed independently from the virtualized entities required for the 3GPP system. The NFV objective is that the VNFs are procured independently from the hardware resources and applied to partially virtualized (e.g., via SDN) or completely virtualized systems. The term partial versus complete partial virtualization is defined in ETSI GS NFV-INF 001 [121]. According to the said specification, a large NF or an NE (network element) can be broken down into a number of constituent NFs. If all such constituents are implemented as VNFs, then the virtualization of such a large-scale NF or NE can be considered complete. If not all but only some constituent NFs are implemented as VNFs, then virtualization of such a large-scale NF is said to be partial.

5.7.3 Benefits/Challenges/Future

Virtualization and softwarization bring a number of challenges and opportunities. Some examples of network slicing include a slice serving a utility company (requesting a water tanker, etc.), a slice for a Mobile Virtual Network Operator (MVNO), a slice for streaming video service, and so on.

Today, one of the difficult challenges faced by operators is the clogging of networks by certain devices, impacting the service delivery to other users. Network slicing may assist in containing such rouge devices to particular slice(s) while keeping other slices unaffected. A misbehaving sensor, for example in a network slice, will not impact a critical public safety service running on another slice.

Network slices are deployed over a common underlying infrastructure which has a finite number of resources. This implementation has two challenges, namely isolation and resource management. Without proper isolation, slices may not be able to perform adequately. However, if slices are assigned dedicated resources, these may lead to over-provisioning. Resource management mechanisms are needed to strike a balance for the implementation of dedicated and shared resources.

For close to perfect collaboration between SDN and NFV, it would be mandatory to formalize interfaces through which either could query or invoke the services of the other. Along the same lines, interfaces between 3GPP and NFV MANO also require research and investigation for smooth operation. For the success of this collaboration, it is important that the three different disciplines understand each other's concepts and terminology and resolve the differences for a successful outcome [122].

For the success of slicing, new business models are required that necessitates innovative partnerships between several players including but not limited to traditional service providers, OTT (over-the-top) service providers, utility companies, media houses, and so on, and business friendly regulatory frameworks.

In a nutshell, present day techniques provide a number of services over one network; 5G network slicing can set up an optimized network environment for every service (*at least this is the expectation*) [123].

Virtualization: NFV and SDN are essential in 5G networks to reduce costs and bring added value to network infrastructure. NFV is the process of moving/forwarding tasks such as load balancing, firewalls, and so on away from dedicated hardware into a virtualized environment [15]. NFV enables the execution of software-based network function on general purpose hardware by leveraging virtualization techniques. The virtualization technologies allow breakup of the software of network functions from dedicated hardware [16]. Softwarization allows implementation of network functions in software, including virtualization of such functions and programmability by setting appropriate interfaces. It is an approach to use software programming to design, implement, and maintain network equipment and services. In SDN, the control plane is decoupled from the data plane and is managed by a logically centralized controller that has a holistic view of the network [11]. Softwarization in RAN may allow some functions such as PDCP and RRC to be implemented as VNFs. Softwarization can also be used to implement certain core and transport functions. The original aim of combining NFV and SDN was to decouple services from physical resources allowing flexibility and adaptability in the network. When NFV and SDN come together, they provide the additional benefit of detaching lifecycle management from physical constraints [16].

Network slice, as discussed in Chapter 5, supports the connectivity of a particular use case through a collection of 5G network functions, and specific configurations in RAN, transport, and core networks. Network functions provide connectivity, storage, and computation. Details on network functions can be found in [11]. Finally, 5G is not all about connectivity, but also demands high end computation and storage. Computation and storage requirements also vary among the different network areas and elements. For example, a BBU pool may have less stringent needs as compared to a core network packet gateway. Similarly, the transport network encompasses several aggregation nodes that need to offer computing and storage capabilities.

Edge Computing

- Edge computing is an emerging computing paradigm which refers to a range of networks and devices at or near the user.
- Edge is about processing data closer to where it's being generated, enabling processing at greater speeds and volumes, leading to greater action-led results in real time.
- It offers some unique advantages over traditional models, where computing power is centralized at an on-premise data center.
- Putting compute at the edge allows companies to improve how they manage and use physical assets and create new interactive, human experiences.
- Some examples of edge use cases include self-driving cars, autonomous robots, smart equipment data and automated retail.

https://www.accenture.com/in-en/insights/cloud/edge-computing-index#:~:text=Edge%20computing%E2%80%94or%20just%20%E2%80%9Cedge,as%20possible%20to%20data%20sources.

Edge Computing

- Possible components of edge include:

1. Edge devices:
2. Network edge:
3. On-premises infrastructure:

https://www.accenture.com/in-en/insights/cloud/edge-computing-index#:~:text=Edge%20computing%E2%80%94or%20just%20%E2%80%9Cedge,as%20possible%20to%20data%20sources.

Edge Computing

- Possible components of edge include**: 1. Edge devices**:

- We already use devices that do edge computing every day—like smart speakers, watches and phones –

devices which are locally collecting and processing data while touching the physical world.

- Internet of Things (IoT) devices, point of sales (POS) systems, robots, vehicles and sensors can all be edge devices—if they compute locally and talk to the cloud.

https://www.accenture.com/in-en/insights/cloud/edge-computing-index#:~:text=Edge%20computing%E2%80%94or%20just%20%E2%80%9Cedge,as%20possible%20to%20data%20sources.

Edge Computing

- Possible components of edge include**: 2. Network edge**:

- Edge computing doesn’t require a separate “edge network” to exist (it could be located on individual edge devices or a router, for example).
- When a separate network is involved, this is just another location in the continuum between users and the cloud and this is where 5G can come into play.
- 5G brings extremely powerful wireless connectivity to edge computing with low latency and high cellular speed, which brings exciting opportunities like autonomous drones, remote telesurgery, smart city projects and much more.
- The network edge can be particularly useful in cases where it is too costly and complicated to put compute on premises and yet high responsiveness is required (meaning the cloud is too distant).

https://www.accenture.com/in-en/insights/cloud/edge-computing-index#:~:text=Edge%20computing%E2%80%94or%20just%20%E2%80%9Cedge,as%20possible%20to%20data%20sources.

Edge Computing

- Possible components of edge include: **3.** On-premises infrastructure:

- These are for managing local systems and
  connecting to the network and
  could be servers, routers, containers, hubs or bridges.

https://www.accenture.com/in-en/insights/cloud/edge-computing-index#:~:text=Edge%20computing%E2%80%94or%20just%20%E2%80%9Cedge,as%20possible%20to%20data%20sources.

Edge Computing

- Edge computing is considered a key technology for efficient routing to application servers hosted by the operator or a third party to achieve low latency as well as efficient use of the transport network.
- In the context of 3GPP, the Edge Computing refers to the scenarios where the services need to be hosted close to the access network of the UE (e.g., at or close to the RAN).
- As described earlier, in 5GS the routing of the data (or user) traffic is done via UPF interface to a Data Network.
- The 5G core network supports the capability to select a UPF that allows routing of traffic to a local Data Network that is close to the UE's access network.
- This includes the local breakout scenarios for roaming UEs as well as non-roaming scenarios.

Edge Computing

- The decision for selection (or reselection) of UPF for local routing may be based on the information from an Edge Computing Application Function (AF) and/or to other criteria such as the subscription, location, and policies.
- Depending on the operator's policy and arrangements with the third parties, an AF may access the 5G core directly or indirectly via the Network Exposure Function (NEF).
- For example,
  an external AF at the edge data center could influence the routing of the traffic by altering the SMF routing decisions via its interaction with the Policy Control Function (PCF).

Edge Computing

- The figure shows local and central data networks (DN) that a UE can reach for edge compute services.
- In this example, each of the UEs (UE1–3) requests the SMF to establish PDU sessions with a central and local anchor.
- The SMF establishes UPF anchor (PSA-0) as the central anchor for each of the PDU sessions, but they each have different UPF PSA for session breakout since the aim is to breakout to a local N6 network that is proximate to the edge application server (EAS).

Edge Computing

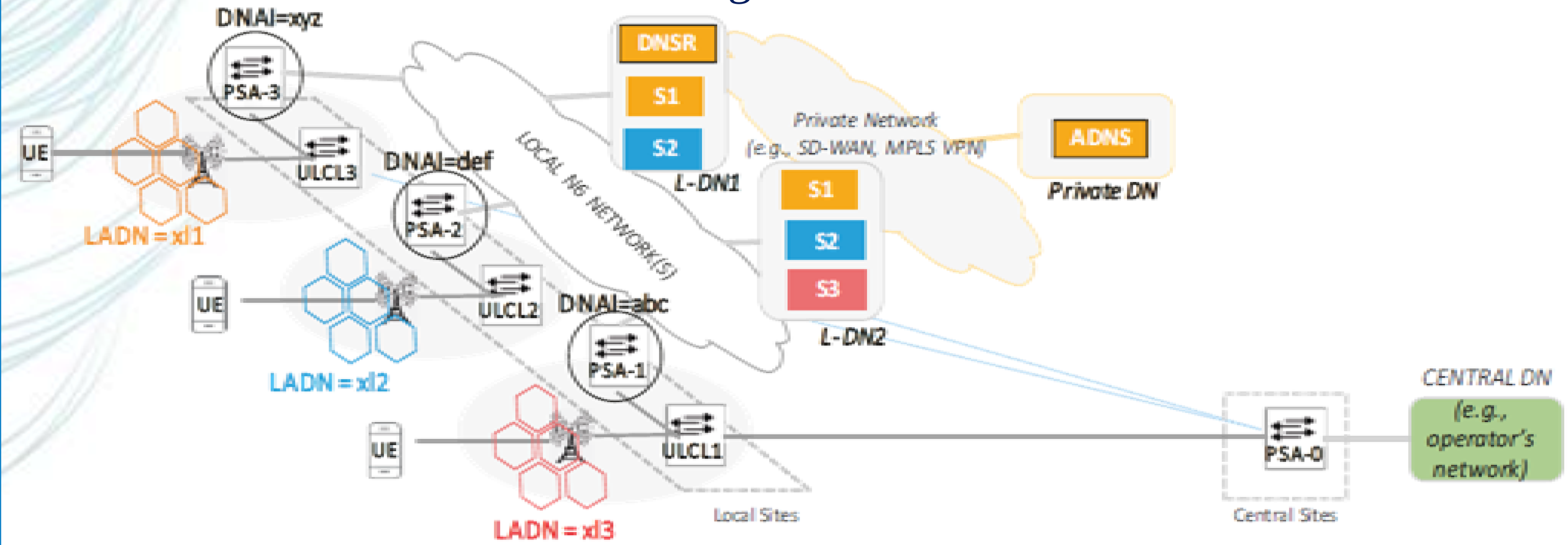
- Edge computing is considered a key technology for efficient routing to application servers hosted by the operator or a third party to achieve low latency as well as efficient use of the transport network.
- In the context of 3GPP, the Edge Computing refers to the scenarios where the services need to be hosted close to the access network of the UE (e.g., at or close to the RAN).
- As described earlier, in 5GS the routing of the data (or user) traffic is done via UPF interface to a Data Network.
- The 5G core network supports the capability to select a UPF that allows routing of traffic to a local Data Network that is close to the UE's access network.
- This includes the local breakout scenarios for roaming UEs as well as non-roaming scenarios.

Edge Computing

- For end-to-end QoS and other service guarantees (reliability, protection, etc.), the PDU session segment and the N6 segment to the application server are important to manage.
- 3GPP-specified QoS with URLLC considers the PDU session segment.
- For the N6 segment, IP routing and transport network underlays with SD-WAN, MPLS, TSN, or other means of providing reliable service guarantees is necessary.
- If the PDU session segment and the N6 segments are operated by different providers (e.g., MNO, IP transport provider, application service provider) there needs to be coordination to support end-to-end guarantees

Edge Computing

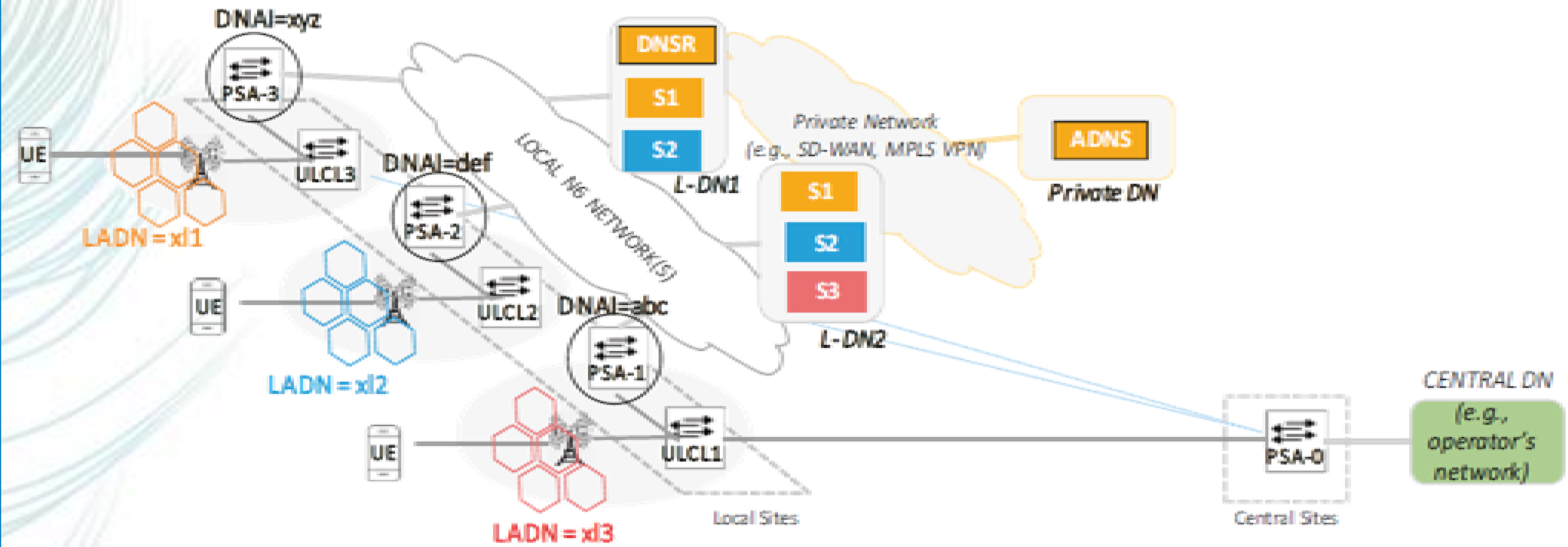
- Another concept introduced in 5G standards to support routing to application data networks (DN) based on the origin/location of the request from the UE. This is shown in Figure.



Edge Computing

- The UE is configured with LADN (Local Area Data Network) information on a per PLMN basis for 3GPP accesses.
- LADN Service Area applies to a set of tracking areas (TA) of PLMN and is configured in AMF and is only accessible by the UE at those specific locations.
- An LADN DNN (Data Network Name) corresponds to an LADN service that the UE is configured with during registration.
- Location-based selection and direction of application data packets is facilitated by LADN.
- For example, in the figure above, there are three LADN (xl1, xl2, xl3) comprising of a set of TAs (some of which may overlap).
- A UE that registers and is configured to use these sets of services (S1, S2, S3).
- Each of these sets may consist of applications that can only be accessed from a DNAI (Data Network Access Identifier).

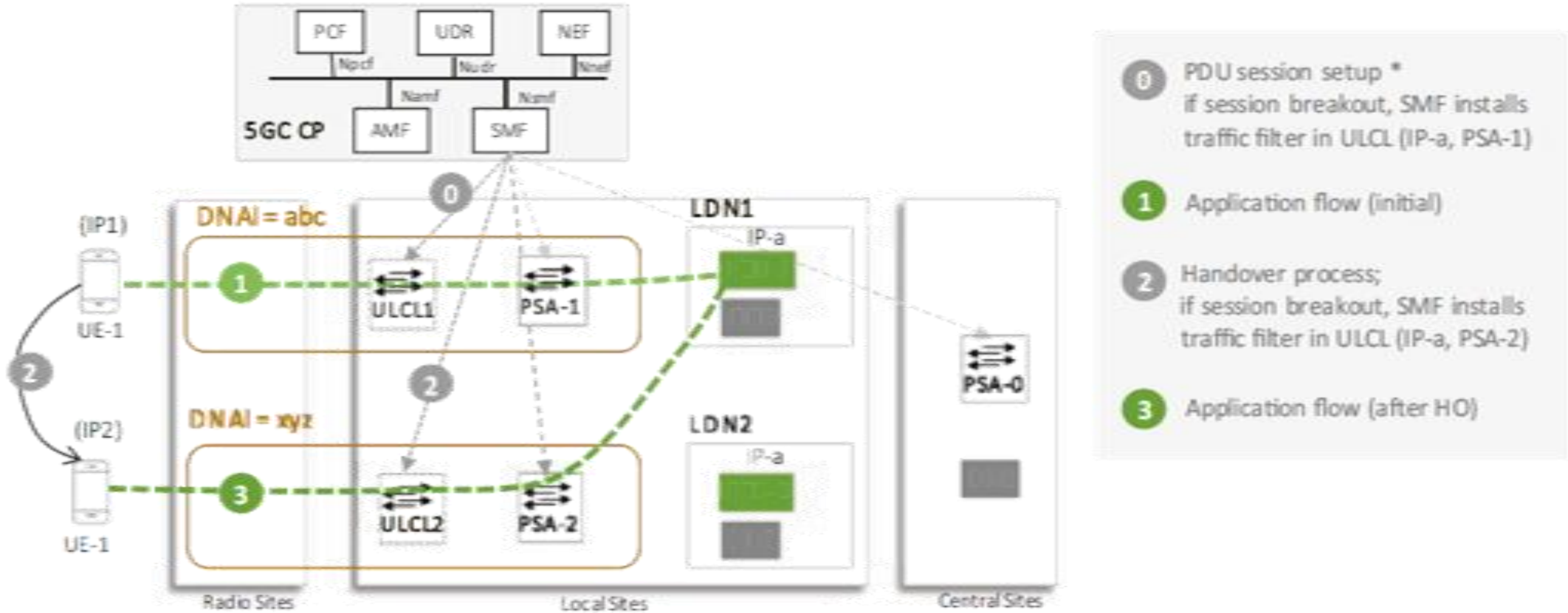
Edge Computing



Mobile Edge Computing

- A data network (DN) where applications are located is identified by a DNAI.
- For example,
  UEs with LADN xl1 is allowed to access the set of services in S1 via DNAI = xyz.
- Route filters in ULCL perform access filtering to limit access to only authorized sessions.

Mobile Edge Computing

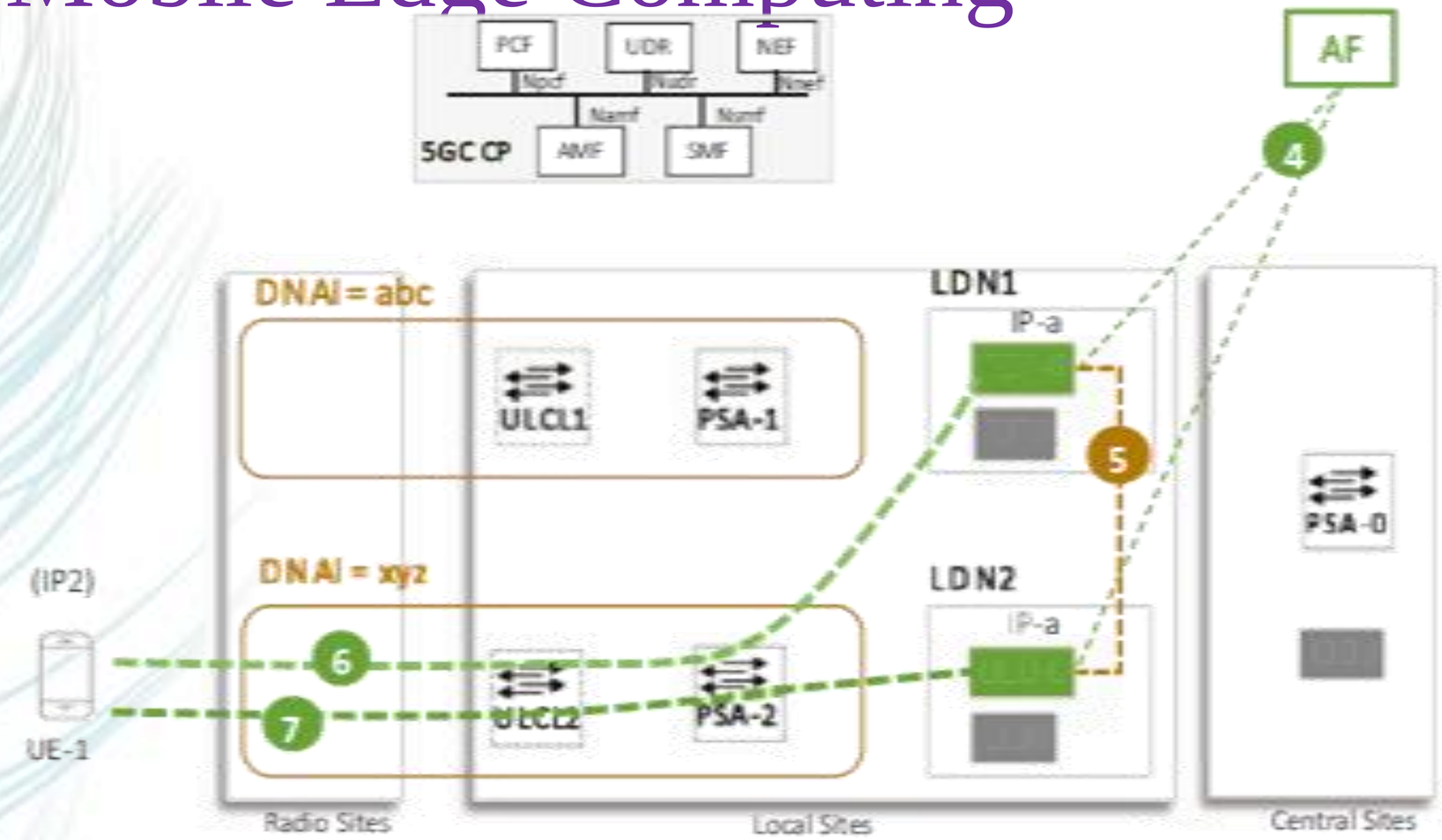


- For stateful applications and UE mobility (new IP access/DNAI), it results in selecting a new LDN (Local Data Network).
- The UE continues to be served by the old LDN location where its data is stored.
  In the figure, after the PDU session is set up (step 0),
  the UE is served by EAS-1 in the local data network (step 1).
  After handover (step 2), the UE continues to be served by EAS-1 (step 3).

Mobile Edge Computing

- The rationale is that not all UE mobility requires application server relocation.
- It may be that the current application server is close enough, or that closer application servers may be overloaded.
- The application domain orchestrator determines the appropriate change of application server based on UE mobility, but also managing congestion, site outage, mitigation of DDoS attacks, and other administrative policy.
- The application domain may also decide to move the serving application server for a variety of reasons.
- The trigger may be that following handover, the UE is better served by a closer server, but the change of application server may be initiated to balance application load, handle server failure, or comply with domain policies on transit cost.

Mobile Edge Computing



4 EAS1 reports new addr (IP2)
AS initiates relocation procedures

5 State is copied between
EAS1, EAS2

6 EAS1 issues redirect to EAS2

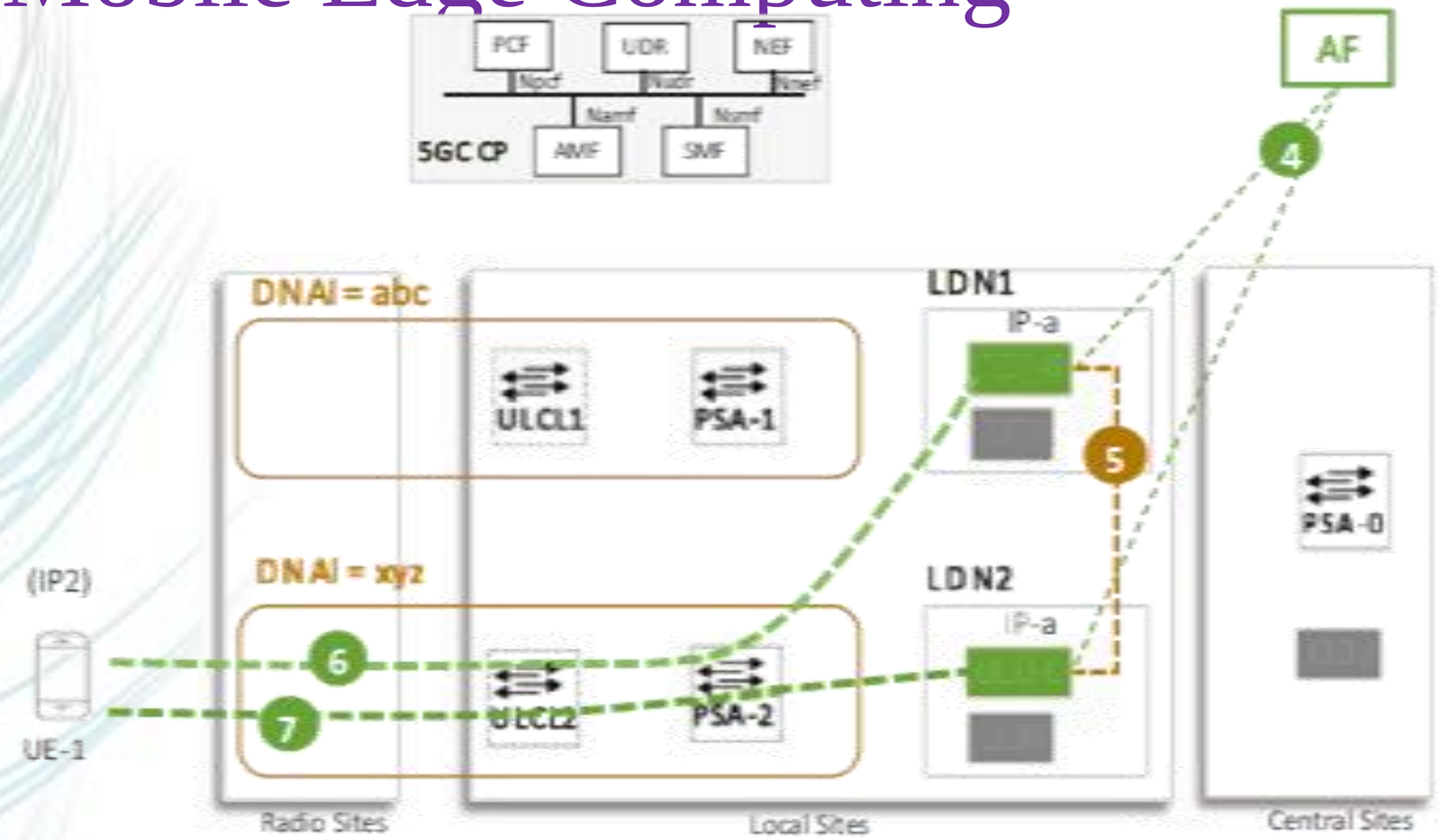
7 Application flow (after redirect)

- Figure shows an overview of possible steps to achieve the relocation of application server.
- 3GPP procedures also describe application server relocation using AF influenced routing.
- In this case, the AF coordinates both PDU session and application server relocation.

Mobile Edge Computing

- When the application domain orchestrator (AF in figure) determines that the server should be relocated based on IP address change or other criteria (step 4), it selects the new server (EAS2).
- The old server (EAS1) and new server (EAS2) replicate state (step 5) for the UE during this transition.
- The old server (EAS1) then sends an application layer redirect request, e.g., HTTPS redirect (step 6) to the new server (EAS2).
- The UE has the URL translated and initiates application message to EAS2 (step 7), and the replication of state across EAS1/EAS2 is stopped.
- Thus, PDU session relocation and application relocation can both be accomplished with minimum coordination.

Mobile Edge Computing



4 EAS1 reports new addr (IP2)
AS initiates relocation procedures

5 State is copied between
EAS1, EAS2

6 EAS1 issues redirect to EAS2

7 Application flow (after redirect)

- Engineered backhauls and IP transport along with new protocols such as QUIC enable a good end-to-end QoS during this transition by using zero RTT application session establishment capability.

Mobile Edge Computing (MEC)

- Multi-access edge computing (MEC), formerly **mobile edge computing**, is an ETSI-defined network architecture concept
- MEC enables cloud computing capabilities and an IT service environment at the edge of the cellular network and, more in general at the edge of any network.
- The basic idea behind MEC is that by running applications and performing related processing tasks closer to the cellular customer, network congestion is reduced and applications perform better.
- MEC technology is designed to be implemented at the cellular base stations or other edge nodes, and enables flexible and rapid deployment of new applications and services for customers.
- Combining elements of information technology and telecommunications networking, MEC also allows cellular operators to open their radio access network (RAN) to authorized third parties, such as application developers and content providers.

https://en.wikipedia.org/wiki/Multi-access_edge_computing#:~:text=Article%20Talk,the%20edge%20of%20any%20network.

Quality of Service Requirements

- Quality of service (QoS) requirements are technical specifications that specify the system quality of features such as performance, availability, scalability, and serviceability.

System Quality	Description
Performance	The measurement of response time and throughput with respect to user load conditions.
Availability	A measure of how often a system's resources and services are accessible to end users, often expressed as the uptime of a system.
Scalability	The ability to add capacity (and users) to a deployed system over time. Scalability typically involves adding resources to the system but should not require changes to the deployment architecture
Security	A complex combination of factors that describe the integrity of a system and its users. Security includes authentication and authorization of users, security of data, and secure access to a deployed system
Latent capacity	The ability of a system to handle unusual peak loads without additional resources. Latent capacity is a factor in availability, performance, and scalability qualities.
Serviceability	The ease by which a deployed system can be maintained, including monitoring the system, repairing problems that arise, and upgrading hardware and software components.

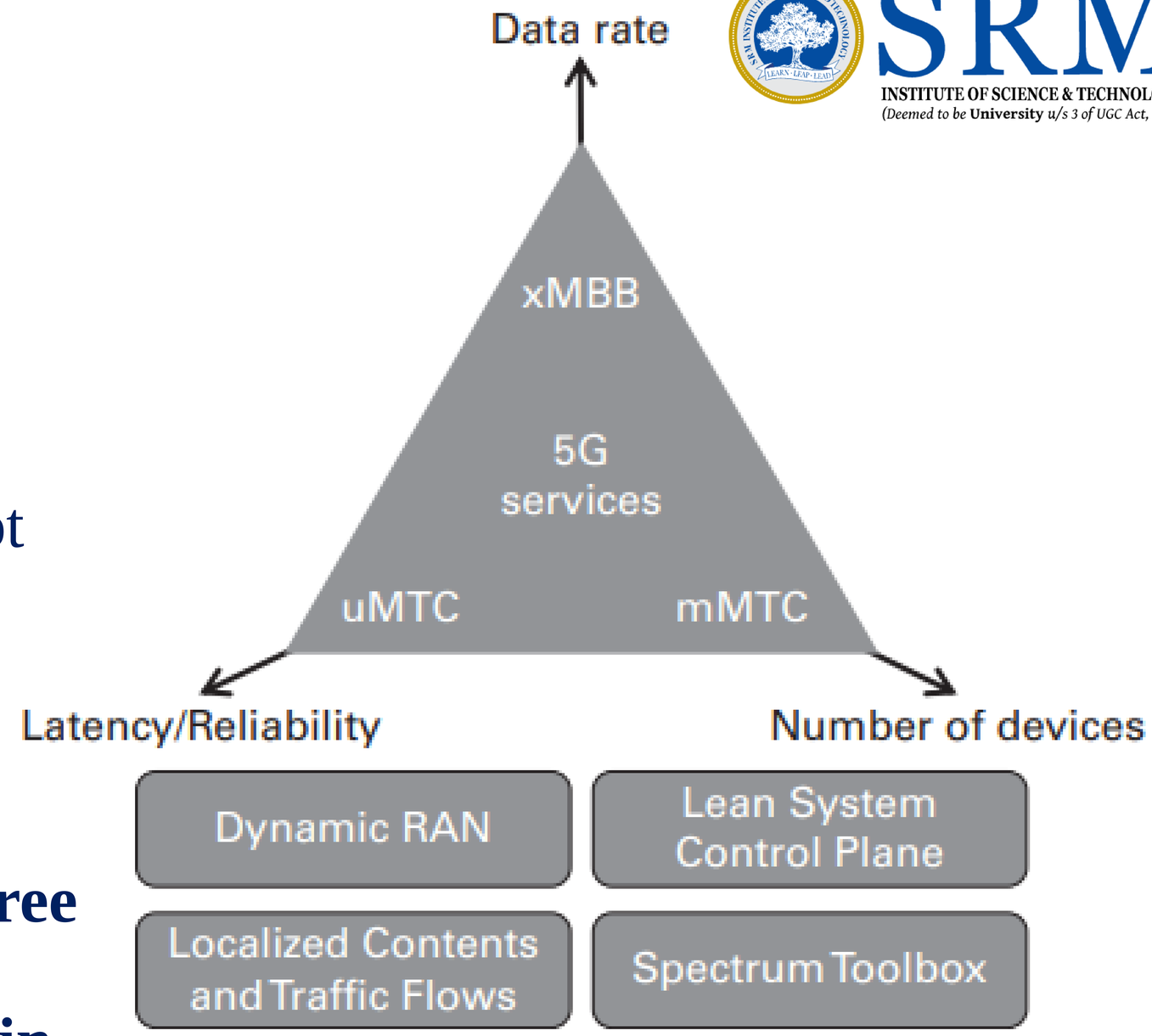
https://docs.oracle.com/cd/E19636-01/819-2326/gaxqg/index.html#:~:text=Quality%20of%20service%20(QoS)%20requirements,specified%20in%20the%20business%20requirements.

Radio Access Network

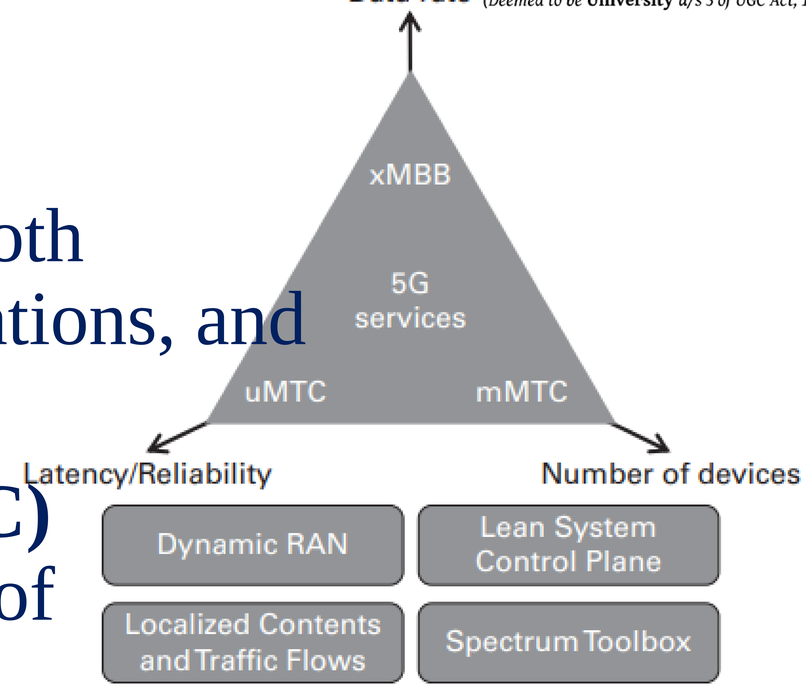
- In the 5G vision, access to information and sharing of data are possible anywhere and anytime to anyone and anything.
- To make this vision a reality, the capabilities of 5G systems must extend far beyond those of previous generations.
- 5G systems must exhibit greater flexibility than previous generations, and involve farther-reaching integration including not only the traditional radio access networks, but also core network, transport and application layers.
- **Altogether, this requires a new way of thinking in 5G wireless access, network architecture and applications**

Radio Access Network

- Because of the wide range of requirements, the earlier generations' one-size-fits-all approach will not work for 5G.
- **Therefore, the proposed 5G system concept, aligns the requirements, and combines technology components into three generic 5G communication services, supported by four main enablers, as shown in Figure**



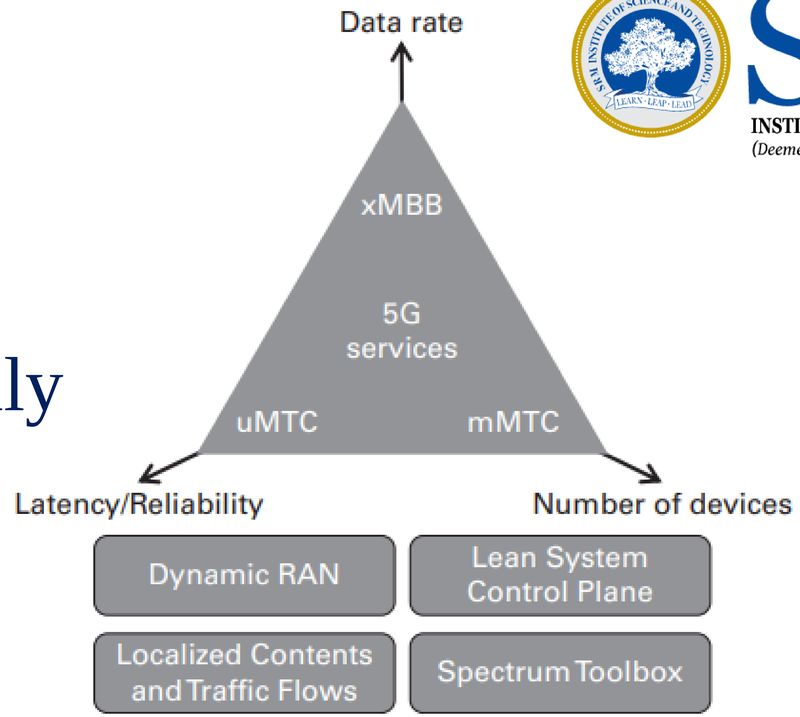
Radio Access Network



- **Extreme Mobile BroadBand (xMBB)** provides both extreme high data-rate and low latency communications, and extreme coverage.
- **Massive Machine-Type Communication (mMTC)** provides wireless connectivity for tens of billions of network-enabled devices, scalable connectivity for increasing number of devices, efficient transmission of small payloads, wide area coverage and deep penetration are prioritized over data rates.
- **Ultra-reliable Machine-Type Communication (uMTC)** provides ultra-reliable low-latency communication links for network services with extreme requirements on availability, latency and reliability

Radio Access Network

- The generic 5G services will not necessarily use the same air interface.
- The preferred waveform depends on design decisions, and how the generic 5G services are mixed.



- A flexible OFDM-based air interface is the most suitable for **xMBB**, whereas new air interfaces as FBMC and UF-OFDM may be promising for **uMTC** where fast synchronization is necessary

Radio Access Network

Dynamic Radio Access Network (DyRAN)

- The Dynamic Radio Access Network (DyRAN) provides a RAN that adapts to rapid spatio-temporal changes in user needs and the mix of the generic 5G services.
  The DyRAN incorporates elements such as

1. • Ultra-Dense Networks,
2. • Moving Networks (i.e. nomadic nodes and moving relay nodes),
3. • Antenna beams,
4. • Devices acting as temporary access nodes and
5. • D2D communication for both access and backhaul.

Radio Access Network
Dynamic Radio Access Network (DyRAN)

1. Ultra-dense networks

- Network densification is a straightforward way to increase the network capacity, and network densification will continue from macro-cellular networks through small cells to Ultra-Dense Network (UDNs).
- UDNs will be deployed both outdoors and indoors, and can have inter-site distances down to a few meters.
- UDNs target user data rates on the order of 10 Gbps, which translates to requirements on high (local) area capacity and high throughput.
- Providing this in an energy-efficient manner requires access to large, preferably contiguous, bandwidth, which is only realistic in the cmW and mmW bands.

Radio Access Network
Dynamic Radio Access Network (DyRAN)

2. Moving Networks (i.e. nomadic nodes and moving relay nodes),

Consist of nomadic nodes and/or moving relays nodes.

- **Moving relay nodes** are wireless access nodes that provide communication capabilities to in-vehicle users, especially in high-mobility scenarios.
- Typical moving relay nodes would be trains, busses and trams, but possibly also cars.
- Moving relay nodes can overcome the outdoor to indoor penetration losses due to metalized windows.

Radio Access Network
Dynamic Radio Access Network (DyRAN)

2. Moving Networks (i.e. nomadic nodes and moving relay nodes),

Consist of nomadic nodes and/or moving relays nodes.

- **Nomadic nodes** are a new kind of network node, where the on-board communication capabilities of vehicles are utilized to make the vehicles serve as a temporary access node for both in-vehicle and outside users.
- Nomadic nodes enable network densification to meet traffic demands varying over time and space.
- Nomadic nodes resemble UDN nodes, but offer their services as temporary access nodes at non-predictable locations and at non-predictable times, and any solution must handle this dynamic behaviour.

Radio Access Network
Dynamic Radio Access Network (DyRAN)

3. Antenna beams,

Beamforming
i.e. the forming of antenna beams, can be used for example to increase the Signal to Interference plus Noise Ratio (SINR) in a local area, in the context of massive Multiple Input Multiple Output (MIMO) or Coordinated Multi-Point (CoMP).

Though the antenna site itself is fixed in location, the beam-direction is dynamic in space and time, and the illuminated area can be considered as a virtual cell.

The virtual cell created by beamforming is more controllable than nomadic nodes.

Radio Access Network
Dynamic Radio Access Network (DyRAN)

4. Wireless devices as temporary network nodes

- High-end wireless devices, such as smartphones and tablets, have capabilities similar to inexpensive UDN nodes.
- A device equipped with D2D capability can act as a temporary infrastructure node for e.g. coverage extension.
- In this mode, a device may take certain network management roles, e.g. resource allocation between D2D pairs, or **Massive Machine-Type Communication** (mMTC) gateway functionality.
- However, admitting user devices into the RAN as temporary access nodes lead to trust issues that need to be resolved

Radio Access Network
Dynamic Radio Access Network (DyRAN)

5. Device-to-device communication

- Flexible D2D communication is a key element in the DyRAN where it can be used for access, offloading the U-plane to a D2D-link and backhaul.
- After device discovery, the most suitable communication mode will be selected based on various criteria, e.g. capacity needs and interference levels.
- D2D communication is also applicable in wireless self-backhauling.