End to End Network Slicing

White Paper 3, OUTLOOK 21

November 2017

Executive Summary

In 2020, 5G is expected to provide a great variety of services and applications that may have different requirements on 5G system functionality and key performance. For a mobile network operator, 5G should become an end-to-end (E2E) flexible, scalable and demand-oriented system to meet the various requirements. The E2E network slicing in 5G is considered as a key driving force to achieve this challenging goal.

E2E network slicing is a logical network that provides specific network capabilities and network characteristics with logical isolation. It is a cross-domain technology that may span across access network (AN), transport, network (TN), core network (CN) and terminal domains. Each of these domains comprises of functions, platforms, protocols as well as resources. This makes network slicing require new thinking in system architecture and design.

This paper attempts to trigger the consistent understanding on the concept, objective, scenarios and requirements for E2E network slicing. Specifically, this paper will mostly focus on the system and management architecture, key enabling technologies and related procedures in each domains and security solutions.

Section 2 identifies a number of technical terms used in this paper, including network slicing instance (NSI) and network slicing template (NST). The overall objective of E2E network slicing is providing logical isolation and independent operations cross AN, CN, TN and terminal domains with an E2E management and security system. The functional requirements on slice creation, modification and deletion operation, and service and user association operation, as well as performance requirements of diverse services on slice capabilities are also introduced. To meet these requirements, an overall architecture with three fundamental layers (infrastructure layer, network slice layer and network management layer) is presented.

Section 3 proposes the network slicing design and management principles. NST design should be separated from the NSI operation. NSIs require multi-dimensional management, O&M capability exposure, capacity scalability and on-demand customization in each domain. NSIs share common infrastructure but ensure isolation, guarantee vertical industry requirements and support multi-vendor and multi-operator scenarios. The NSM architecture is based on "Layer- and Domain-based management" principle, including NSM function and network slice subnet management function. The cooperation between the two functions guarantees the E2E requirement.

Section 4 discusses the scenarios and objective for CN slicing. The system architecture for 5G to support network slicing is presented, where related functions, such as network slice selection function, access and mobility management function are introduced. In such architecture, a network slice may be implemented with combination of shared and dedicated resources. This means each CN slice may have dedicated and shared network functions. The CN slice selection, serving AMF selection and NSI selection for a UE to establish a PDU session are all based on the network slice selection assistance information.

Section 5 discusses the scenarios and requirements for TN slicing. Four-layer TN slicing model are presented, which are physical slicing, hard pipe slicing, soft pipe slicing and service layer. A potential solution for 5G TN including fronthaul transport, middlehaul transport, backhaul transport, SDN controller and network slicing is also presented. TN slicing consists of control plane slicing and forwarding plane slicing. The former provides encapsulating and scheduling capabilities for the services, and the latter needs to support L2/L3 layer high-speed forwarding, L1 layer channelization, physical isolation, and capability expansion.

Section 6 discusses the scenarios, requirements and objective for RAN slicing. Designing a RAN slicing should

consider unified logical architecture and resource logical isolation, differentiated real-time (RT) functions and non-real-time (NRT) functions, and ensure network key performance, O&M requirement and service requirement. Then, RAN slicing parameters can be divided into two parts: one is related to Q&M (platform management and network management), and the other is related to the service management (UE management, wireless function deployment, radio resource management, QoS controlling aspects, etc.). In the presented RAN slicing architecture, isolation and sharing can be supported at different RAN slices. Furthermore, a CN and TN NSI can be shared by different RAN slices. This explains the relationship between RAN slicing and PLMN. Based on such an architecture, access and network slice selection, UE context handling, PDU session handling and handover procedures are discussed in detail to describe how a RAN slicing works.

Section 7 introduces two types of terminal slicing: vertical slicing and horizontal slicing. Vertical slicing targets at supporting vertical industry and markets. It enables resource sharing among services and applications and avoids or simplifies a traditional QoS engineering problem. Horizontal slicing, as a step forward, targets at extending the capabilities of mobile devices with over-the-air resource sharing across network nodes and devices. As the industry moves towards vertical slicing as the phase-1 implementation, provision for enabling horizontal slicing as the phase-2 implementation is needed to improve user experience.

Section 8 focuses on network slicing security. There are a number of slicing security requirements on slices isolation, security mechanism differentiation, slice authentication and authorization, virtualization, slicing management, etc., Specifically, slice-specific security policies (e.g. confidentiality protection, integrity protection, cryptographic algorithms / protocols) should be configurable according to the security requirements and different security mechanisms (e.g. authentication mechanisms) can be used in different network slices. Network slicing isolation should confine potential network attacks within a single NSI. UEs can access network slices only after they have been authenticated and/or authorized to access the network slices and the capabilities to manage network slices should be under control of authorized operators. To deal with the security threats, a network slice security architecture with the authentication server function, security anchor function, unified data management function and policy control function is proposed, and a general network slice security procedure is also described to show how the security solution works.

This paper tells that implementing an E2E network slicing for 5G is a challenging goal. It needs well-developed enabling technologies, global standards and mature ecosystem. We aspire this paper serves the purpose of triggering to establish common understanding and tight cross-industry collaboration among operators, vendors, SDOs and vertical industries to promote 5G development and commercialization.

Table of Contents

1 Introduction

In 2020, 5G is expected to provide a great variety of services ranging over three fundamental scenarios: enhanced mobile broadband (eMBB), massive machine type communication (mMTC), and ultra reliability low latency communication (URLLC). These services may have different requirements on 5G system functionality (e.g. priority, charging, policy control, security, and mobility) and key performance (e.g. latency, mobility, availability, reliability and data rates). For example, eMBB focuses on high throughput, mMTC focuses on the connection terminal number, and URLLC focuses on the low latency and high reliability. For a mobile network operator (MNO), 5G should become an end-to-end (E2E) flexible, scalable and demand-oriented system to meet the various requirements.

The E2E network slicing technology has been developed to help 5G system on-demand supply [1-8]. A network slice is self-contained in terms of operation and traffic flow and can have its own network architecture, engineering mechanisms and network provision. It in general is to architect, partition and organize virtualized network resource to enable flexible support of diverse use case realizations. With network slicing, one physical network is sliced into multiple virtual networks, each architected and optimized for a specific requirement and/or specific application/service. Cloud computing, software defined network (SDN) and network function virtualization (NFV) are key technologies for network slicing.

NFV achieves the multi-tenant mode of physical hardware and splits the load balancing from the hardware. All the network functions are deployed on the virtual server of the commercial server, rather than being deployed separately on its private network device. In this way, Radio Access Network (RAN) acts as the edge of the cloud, and the core function can be regarded as the core cloud. The connection between the virtual machines located in the edge cloud and the core cloud can be configured using SDN. SDN-based 3-D network slices (i.e., resource slices, functional slices and network slices) provide users with personalized virtual network slices under the control of the SDN control platform. The virtual network functions are placed in different locations (i.e., edge clouds or core clouds) in each slice according to different service characteristics. Compared with the existing virtualization technology, E2E network slicing provides a more flexible and efficient implementation to achieve an open 5G network. It separates the network resource slices from the network function slices rather than bundles them together. Specifically, the separation of network resources and network function slices speeds up the response of the network to meet different user needs and improves the resource utilization.

The E2E network slicing is a cross-domain technology and needs cross standards developing organizations (SDOs) and cross-industry collaboration. This paper attempts to trigger the consistent understanding on enabling technologies, and system architecture for network slicing to promote 5G development and commercialization.

2 Overview

2.1 Definitions

A number of technical terms used within the present document are identified below for the purposes of consistent understanding cross different partners [5].

- Network slicing: A logical network that provides specific network capabilities and network characteristics.

- Network slice instance (NSI): A set of network function instances and the required resources (e.g. compute, storage and networking resources) which form a deployed network slice to meet certain network characteristics.
- End-to-end: A network connection enabling user access applications.
- End-to-end network slicing. A logical network may contain specific core network capabilities, access network capabilities, transport network capabilities, terminal capabilities and related characteristics.
- Network slice type: Network slice types are high-level categories for NSIs, which reflect the distinct demands for network solutions. Three fundamental network slice types have been identified for 5G: eMBB, mMTC and URLLC.
- Network slice template (NST): Network slice template is used to create NSIs, which is the output from the slice design phase.
- Tenant: The operators' customers (for example, customers from vertical industries) or the operators themselves. They utilize the NSIs to provide services to their end users. Tenants may have independent operation and maintenance (O&M) requirements, which are uniquely applicable to the NSIs.
- Physical resource isolation: physical resource allocated for one network slice cannot be used by other network slices in order to avoid negative effect between multiple network slice instances

2.2 Overall objective

The overall objective of E2E network slicing is providing logical isolation and independent operations include the terminal, RAN radio resource, core network (CN), transport network (TN) and network management for different scenarios, service and vertical industry.

2.3 Scenarios

Three fundamental scenarios have been identified for 5G: eMBB, mMTC, and URLLC. Three fundamental network slice types and a number of use cases are also identified related to these scenarios. Each use case may be served by a network slice to meet the performance requirements on 5G network. For intelligent city, smart grid, intelligent agriculture and other services, networks need to support massive equipment access sending small data. Video surveillance and mobile medical services express high requirements on the transmission rate. Automotive networking and vehicle to everything (V2X) communication and industrial control services require nearly 100% reliability and millisecond latency. IoT services need different types of networks and requirements for mobility, billing, security, policy control, latency and reliability.

As the work on 5G began in 3GPP, it will bring new type devices those are becoming increasingly popular and are expected to become essential in our everyday life. Examples include wearable devices for advanced telemedicine applications, virtual/augmented reality, UHD video and V2X applications other than smart phone. This brings emerging application scenarios for network slicing.

Vision examples of related use cases are as follows [1]:

- Self-automated car in a smart city: Bob starts his self-automated driving car that relies on V2X communication. While sitting in the car, Bob initiates a HD video streaming service through infotainment system available in the car. In this example, the V2X communication requires a low-latency but not necessarily a high throughput, whereas, the

HD video streaming requires a high throughput but is tolerate to the latency.

- Healthcare robot: A robot that is monitored by the healthcare service provider takes care of elderly people at home. The robot sends a regular report of health status and the activities interacting between the robot and the elderly people to the healthcare operator. The robot also allows the elderly people to do any Internet like services (e.g., web-surging, hearing streaming music, watching a video) or even making a call to their doctor directly in case of emergency.

2.4 Requirements

A network slice can provide the functionality of a complete network, including RAN functions and CN functions. To achieve the network slicing, 5G system shall be able to create, modify and delete a network slice, define and update the set of services and capabilities for a network slice, identify UE and its service requirements and associate it to a network slice or remove it from a network slice. Network slices may have different priority order, and the isolation between slices should be ensured, where one network slice operation shall have no or minimal impact on other ones. The slice modification operation may redefine the maximum/minimum capacity or add and remove network functions to the slice. 5G system should enable a UE access more than one network slice.

5G network is expected to cater to a plurality of devices that will connect to the network, each with its diverse set of quality of service/quality of experience (QoS/QoE) ranging over E2E latency, throughput, types of data transferred between end point devices and application servers, frequency of data transfers among other dependencies. Three categories of services and their characteristics are described in Table 2-1. For example, when the E2E latency is very tight, a part of CN and RAN function need to be deployment near the air interface. In addition, business needs multi-tenancy model to efficiently utilize network resources. 5G system should ensure network slicing to meet service level agreement (SLA) requirements for each tenancy.

Table 2-1: categories of services and their characteristics

Characteristics	mMTC	URLLC	eMBB
Availability	Regular	Very High	Regular (baseline)
E2E latency	Not highly sensitive	Extremely sensitive	Not highly sensitive
Throughput type	Low	Low/med/high	Medium
Frequency of Xfers	Low	High	High
Density	High	Medium	High
Network coverage	Full	Localized	Full

2.5 Overall architecture

Network slicing enables the operator to create networks customized to provide optimized solutions for different market scenarios which demands diverse requirements, e.g. in the areas of functionality, performance and isolation. For this, it requires native support from the overall system architecture. The overall architecture consists of three fundamental layers [10]: the infrastructure layer, network slice layer and network management layer. The infrastructure layer provides the physical or virtual resources and fundamental capability for network slicing, for instance, computing resource, storage resource, and connectivity. The network slice layer runs above the infrastructure layer and provides necessary network functions to form E2E logical networks via NSIs. The network management layer contains the conventional business support system / operation support system (BSS/OSS) and network slice management (NSM) system, which designs and manages network slicing. Moreover, it also ensures that the SLA requirements.

3 Network Slicing Management Architecture

3.1 Network slicing design and management principle

The NSM architecture has the following key principles:

- **Separate the Network slice template design and the NSI operation:** The network slice template is created based on the network capability of each technical domain and a tenant's particular requirements in the design phase. An Network Slice Instance is instantiated based on the network slice template in the operation phase, which comprises the configuration and deployment of corresponding network functions and resources in different technical domains. The separation of network slice design with the operation enables the reuse of a network slice template.
- **NSIs require multi-dimensional management:** Usually, a NSI contains multiple technical domains. Besides, an NSI may also involve multiple administrative domains which belong to different operators. To assure fast deployment of NSI, the usage of efficient multi-dimensional management is necessary through coordination and cooperation across these different domains.
- **Common infrastructure:** Network slicing takes advantages of common infrastructure among tenants from the same operator. This solution is different with the dedicated network solutions which utilize static and physically isolated networks to satisfied different tenants. It helps to reduce the service time to market (TTM) and acquire higher resources utilizatioin efficiency. Furthermore, such design is helpful for long-term technology evolution and for constructing a healthy industry ecosystem.
- **On-demand customization:** In an NSI, customization capabilities is diffenent within each technical domain. The coordination of capabilities is performed via the NSM system during the process of network slice template design, O&M and NSI deployment. An independent tailoring-process can be performed in each technical domain according to design schemes to effectively balance architectural complexity with the simplicity needed by commercial practice.
- **Isolation:** The overall architecture is in support of the NSIs' isolation which includes O&M isolation, resource isolation and security isolation. NSIs can be both physically and logically isolated in different levels.
- **Guaranteed-performance:** Different domains are integrated seamlessly to form Network Slicing which satisfies and ensures 5G performance specifications defined by industry as well as accommodates requirements from vertical industry.
- **Scalability**: As a result of virtualization, an NSI can dynamically change its occupied resources, such as scaling in/out.
- **O&M Capability Exposure:** NSIs used by tenants may be dedicated, shared or partially shared. Moreover, O&M demands from different tenants may be independent. The access is provided by NSM system to a number of O&M functions of NSIs for the tenants, for expample, allowing them to configure NSIs related parameters, such

as policy.

- **Support for multi-vendor and multi-operator scenarios:** Network slicing allows multiple technical domains composed of network elements from different vendors to be managed by a single operator. Moreover, the architecture is also in support of the scenario, in which multiple administrative domains of different operators are covered by the services from tenants.

3.2 NSM Architecture

To support network slicing, the NSM shall be based on the state of art cloud management technologies with advanced features. It makes use of a streamline of related services to provide O&M capability, which handles disadvantages of the traditional network management system, for example, lack of automatic O&M methods or long TTM. The NSM system could also help operators to enable new business opportunities by establishing an open ecosystem.

Figure 3-1 describes the overall NSM system architecture.

- CSMF is responsible for translating the communication service related requirement to network slice related requirements and communication with NSMF.
- NSMF is responsible for management and orchestration of NSI and the derivation of network slice subnet related requirements from network slice related requirements. Futhermore, the NSMF includes the function of Cross Domain Slice Manager which is responsible for the NSI lifecycle management (i.e. pre-provision, instantiation, configuration and activation , and decommissioning). It meets the E2E requirement through multi-dimensional coordination within different domains. The NSMF decomposes into sets of requirements based on each technical domain’s capabilitiy then maps each segment of requirement to the related technical domain. To guarantee the overall E2E requirement, the NSMF aggregates the network service performance of each individual technical domain. Then corresponding adjustments and configurations was performed to ensure closed-loop control. It is necessary to coordinate the interworking between diffenent NSMFs in order to support management functions across different administrative domains for multiple operators.
- NSSMF is responsible for management and orchestration of NSSI. The NSSMF includes the domain slice managers for different technical domains, e.g., access network NSSMF, core network NSSMF, and transport network NSSMF or a combined domain NSSMF. As a logical entity, the NSSMF is responsible for pre-provision, instantiation, configuration and activation, and decommissioning of subnets in a single technical or a combined domain. The NSSMF ensures the real-time assurence for decomposed E2E requirement capabilities in each single or combined domain, for example, through monitoring and fault localization. Each single or combined domain has independent requirement, particular closed-loop control of functions and resources to fulfill fast service scheduling and resource optimization.

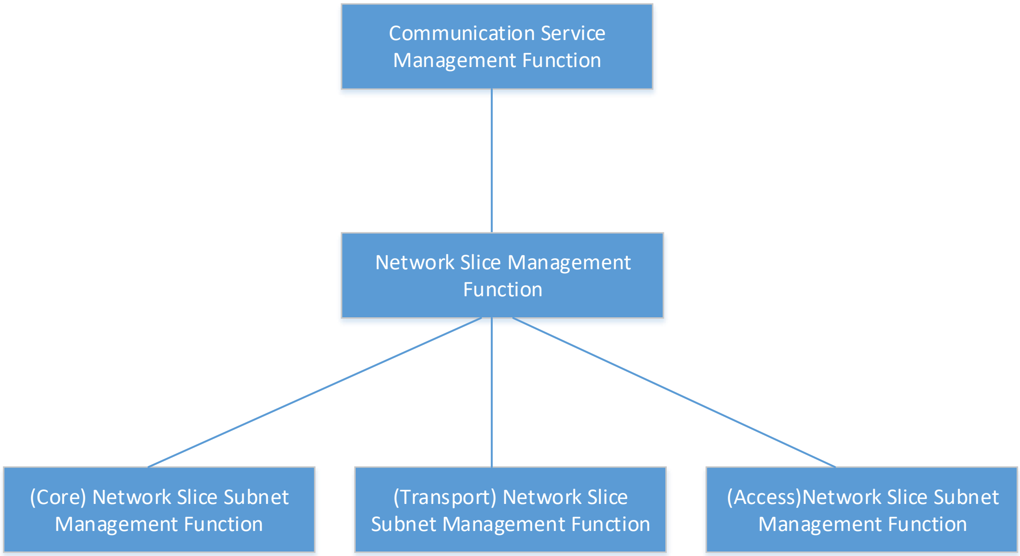

Communication Service Management Function
Network Slice Management Function
(Core) Network Slice Subnet Management Function
(Transport) Network Slice Subnet Management Function
(Access)Network Slice Subnet Management Function

Figure 3-1: Network slice management architecture

The NSM system is of great importance in the entire system architecture. As shown in Figure 3-2, NSM provides functions in the following phases:

- Preparation phase
- Instantiation, Configuration and Activation phase
- Run-time phase
- Decommissioning phase

In the preparation phase the NSI does not exist. The preparation phase includes the creation and verification of network slice template(s), the on boarding of these, preparing the necessary network environment which are used to support the lifecycle of NSIs and any other preparations that are needed in the network.

During instantiation / configuration, all resources shared/dedicated to the NSI have been created and are configured, i.e. to a state where the NSI is ready for operation. The activation step includes any actions that makes the NSI active, e.g. diverting traffic to it, provisioning databases (if dedicated to the network slice, otherwise this takes place in the preparation phase) etc. Network slice instantiation, configuration and activation can include instantiation, configuration and activation of other shared and/or non-shared network function(s).

In the run-time phase, the NSI is capable of traffic handling to support communication services of certain type(s). The run-time phase includes supervision/reporting (e.g. for KPI monitoring), as well as activities related to modification. Modification could map to several workflows related to runtime tasks, e.g. upgrade, reconfiguration, NSI scaling, changes of NSI capacity, changes of NSI topology, association and disassociation of network functions with NSI.

The decommissioning phase includes deactivation (taking the NSI out of active duty) as well as the reclamation of dedicated resources (e.g. termination or re-use of network functions) and configuration of shared/dependent resources. After decommissioning the NSI does not exist anymore.

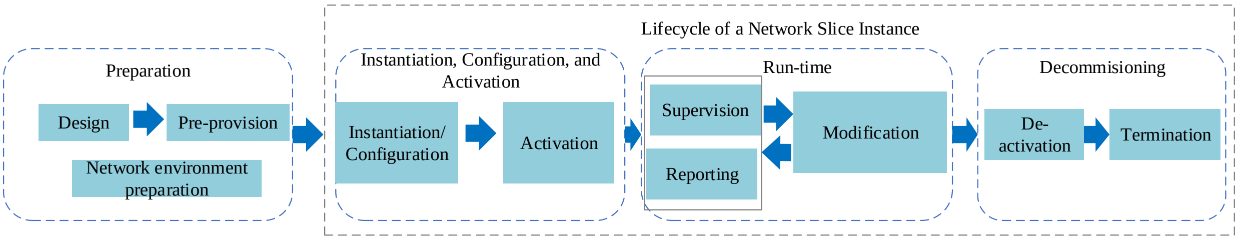

Lifecycle of a Network Slice Instance
Preparation
Design
Pre-provision
Network environment preparation
Instantiation, Configuration, and Activation
Instantiation/ Configuration
Activation
Run-time
Supervision
Reporting
Modification
Decommisioning
De-activation
Termination

Figure 3-2: Lifecycle phases of an NSI

Besides the seamlessly managing and assuring the requirement, the NSM could also predict the network status changes with the help of enhanced air interface algorithms to provide certain control and management actions for precaution. In addition, there two types of NSM system, standalone (a new management entity) and non-standalone (integrated with OSS).

4 Core Network Slicing

4.1 Scenarios and objective

4.1.1 Necessity and scenarios

Different use cases may have different functionality requirements and performance requirements. But the current network is a ‘one size fits all’ system. When a new use case is deployed, in order to fulfill its requirements, multiple network functions have to be updated and new network functions have to be deployed. So the network becomes larger and larger and hard to manage and update.

Network slicing allows the operator to provide dedicated logical networks with customer specific functionality, so it can improve the flexibility of the network. The Figure 4-1 provides a high level illustration of the network slicing.

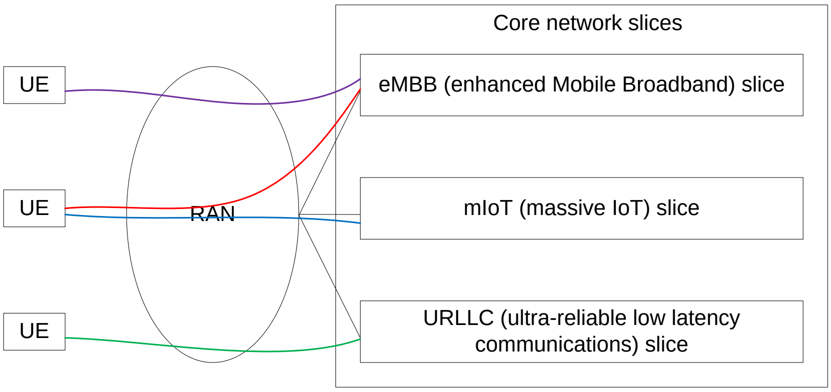

Core network slices
eMBB (enhanced Mobile Broadband) slice
mIoT (massive IoT) slice
URLLC (ultra-reliable low latency communications) slice
UE
UE
UE
RAN

Figure 4-1: network slicing high level illustration

4.1.2 Objective

Core network slicing targets are as follows:

- Dynamic management: the operator can dynamically create and manage (e.g. scale in/out, delete, modify) the network slices customized for different market scenarios.
- Support selection of network slices: the network can select a suitable network slice for specific UE, devices, services and subscribers. The network slice connected by the UE can also be changed.
- Simultaneously access: a UE can access one or more network slices simultaneously.
- Isolation: different network slices are isolated. A service in one slice will not impact services offered by other slices.

4.2 Key technologies

4.2.1 Core network slicing architecture

Overall system architecture for 5G with network slicing function is shown in Figure 4-2. A brief description of each of the network function (NF) are listed below：

- Authentication Server Function (AUSF): supports authentication server function.
- Access and Mobility Management Function (AMF): access control, mobility control, transparent proxy for routing SM message.
- Unified Data Repository (UDR): storage and retrieval of data by the UDM, PCF or NEF.
- Unstructured Data Storage Network Function (UDSF): storage and retrieval of information as unstructured data by any NF.
- Network Exposure Function (NEF): expose the services and capabilities provided by 3GPP network functions.
- NF Repository Function (NRF): maintains NF profile, supports service discovery.
- Policy Control function (PCF): decides the policy and provides them to the control plane function.
- SMF (Session Management Function): manages the PDU session e.g. PDU session establishment, modify and release.
- Unified Data Management (UDM): authentication credential processing, access authorization, registration/ mobility management and subscription management.
- User plane Function (UPF): handles the user plane traffic, e.g. traffic routing & forwarding, traffic inspection and usage reporting, handling.
- Application Function (AF): interacts with the 3GPP Core Network (CN) to provide services.
- Network Slice Selection Function (NSSF): selects the NSI, determines the allowed network slice selection assistance information (NSSAI) and AMF set to serve the UE.

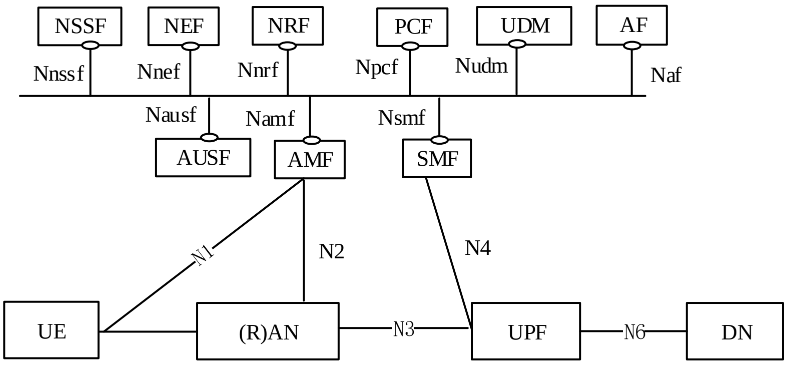


Figure 4-2: The service-based architecture for network slicing

NOTE 2: For the sake of clarity of the diagrams, the UDSF,UDR, NEF and NRF have not been depicted. However, all depicted Network Functions can interact with the UDSF, UDR, NEF and NRF as necessary.

Network slicing enables logical partitioning of CN, with each logical slice providing a complete CN capability with all network nodes that leverages shared infrastructure or separate infrastructure as required by operator. One of key capability in 5G CN is the separation of user plane and control plane capabilities. Control plane capabilities such as session management, access authentication, policy negotiation/management, user data storage is functionally independent of user plane function that handles packet forwarding, encapsulation/de-capsulation, and associated transport level specifics. This separation enables

- Slice based independent user plane.
- Selective choice of control plane functions needed for different network slices
- Distribution of user plane closer to the edge of the network Slice common or independent control plane

The network slicing architecture is shown in Figure 4-3. It includes two groups. One is the dedicated network slice and the other is the network slices sharing common CP NFs. Global network functions across multiple slices. Such functions are those related with multiple slices in the network, e.g., UE subscription repository function. The NSSF is such a common function provided to all network slices in a public land mobile network (PLMN). Common CP network functions for multiple slices with UE simultaneously connected. A UE can access multiple network slices at the same time. In this case, there should have a minimum set of NFs which can be flexibly expanded with additional NFs per slice requirement. The minimum set of common CP NFs should include AMF. The signaling (e.g., between UE and AMF function, and between new/target AMF function and old AMF function) can be reduced if mobility management (e.g. UE location update related management) is shared among different network slices when a UE simultaneously obtains services from different slices.

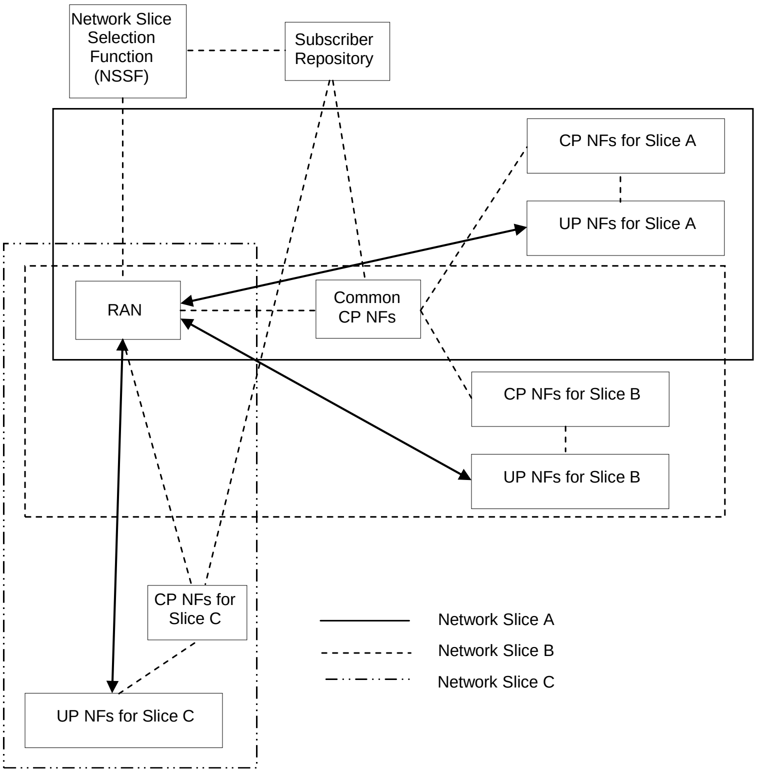


Figure 4-3: Network slicing architecture for 5G CN

An example of network slice with combination of shared and dedicated resources are illustrated in Figure 4-4, where each of the network slices have dedicated Network Functions (NFs) e.g. SMF, UPF, NEF, NRF & AF nodes. It may be possible to implement shared NF for AMF, AUSF, PCF & UDM (e.g. same operator implementing different network slice for MBB & mMTC use cases; where as an entirely dedicated set of all NFs are implemented for Slice C. While this is for illustrative purposes, it provides a perspective on the range of combinations that need to be possible to be configured for different network slices based on business-driven tops down requirements and capabilities of the resources that may need to be brought together as part of a network slice. Also evident from the diagram is the set of network functions instantiated for each of the network slices – slice B does not instantiate an AF.

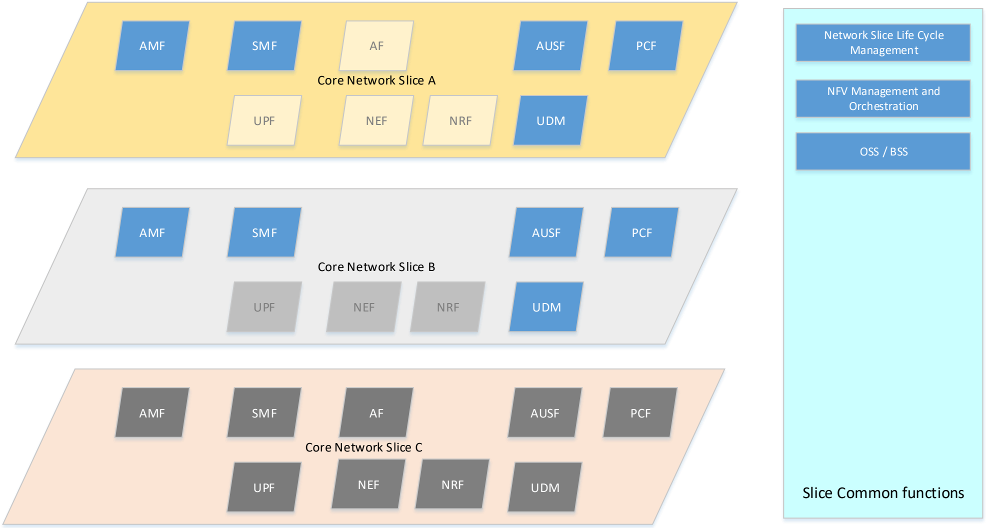


Figure 4-4: An example of network slices with shared and dedicated resources

4.2.2 Core network slicing selection

The CN part of a NSI(s) serving a given mobile device will be selected by the CN itself independent on the mobile device/UE. A UE may be served by one or more CN slices at the same time, for services that need different capabilities.

4.2.2.1 S-NSSAI and NSSAI for CN slicing selection

A collection of S-NSSAIs is called as NSSAI, and each S-NSSAI assists the network in selecting a particular NSI at run time. An S-NSSAI is comprised of:

- A slice/service type: the expected network slice behaviour in terms of features and services;
- A slice differentiator: the optional information that complements the slice/service type(s) to allow further differentiation for selecting an NSI from the potentially multiple NSIs that all comply with the indicated slice/service type.

The UE may have the following different types of NSSAIs:

- Configured NSSAI: The network(i.e. the HPLMN) may provide a Configured NSSAI per PLMN in the UE. It can be standardized value or PLMN specific value.
- Allowed NSSAI: Upon successful completion of a UE's Registration procedure, the UE may obtain from the AMF an Allowed NSSAI for this registration area, which may include information of one or more S-NSSAIs . The UE shall use only the S-NSSAIs in the Allowed NSSAI corresponding to a Network Slice for the subsequent Network Slice selection related procedures in the registration area.
- AS(layer) NSSAI: Upon successful Registration, the UE is provided with an AS NSSAI, and a Temporary ID by the serving AMF. The UE shall include this AS NSSAI in AS layer e.g. RRC Connection Establishment resulting from initial and mobility Registration messages but not Service Request to enable the (R)AN to route the NAS signalling between the UE and the appropriate AMF in case the Temporary ID is not valid.

4.2.2.2 Serving AMF selection& Registration to a set of NSI

When the UE initially accesses to the network , a set of network slices to serve the UE is determined first based on requested NSSAI provided by UE, i.e., the common network functions of the set of network slices, which at least includes an AMF, are determined.

If the (R)AN can not select an appropriate AMF based on the information included in the AS message, the (R)AN will forward the registration request message to a default AMF. The default AMF retrives the UE subscription data and interacts with the NSSF (Network Slice Selection Function) to get information for selection of a new AMF to serve the UE. The default AMF or the (R)AN determines a target AMF as the UE's serving AMF based on the information.

After successful initial registration the UE is provided with a temporary identity that is provided by the UE in AS message during subsequent accesses to enable the (R)AN to route the NAS message to the appropriate AMF, as long as the temporary identity is valid. Otherwise the (R)AN uses the AS NSSAI included in AS message to select an appropriate AMF and forwards the NAS message to the selected AMF. If the (R)AN is not able to select an AMF based on the temporary identity or the AS NSSAI, the NAS message is forwarded to a default AMF and the default AMF interacts with the NSSF to determine the information for selection of a new AMF to serve the UE. AMF forwards the requested NSSAI with UE's subscribed NSSAI retrieved from UDM to NSSF. Based on UE's subscribed NSSAI, requested NSSAI (if have), registration area and local policy, NSSF determines allowed NSSAI. Then NSSF determines the serving NSI based on allowed NSSAI, and responds the allowed NSSAI and NSI-ID of serving NSI in UE's registration area to AMF. The mapping between S-NASSI and NRF can be changed due to the operational policy of operators. When the mapping changes, NSSF can notify AMFs on the change. Notified AMFs can de-register UE or be relocated to proper AMFs, which can be retrieved from NRF.

4.2.2.3 Selection of a particular NSI for the UE for establishing a PDU session

A slice-specific network function(s) (e.g. SMF) within the NSI are selected by the AMF via the NRF during a PDU session establishment procedure based on the S-NSSAI and DNN included in the PDU session establishment request and other information e.g. UE subscription and local operator policies, when the UE triggers the establishment of a PDU session for an application.

The network operator may provision the UE with network slice selection policy (NSSP). The NSSP includes one or more NSSP rules each one associating an application with a certain S-NSSAI. A default rule that matches all applications to a default S-NSSAI may also be included.

During Registration procedure, the serving AMF will get the the Allowed NSSAI and the corresponding NRF. The AMF provides the S-NSSAI included in the PDU session establishment request to the NRF and retrieves the SMF address for the PDU session and forwards the PDU session establishment request to the SMF.

4.2.3 Core network slicing registration

Registration procedure is used to authenticate UE and authorize UE to receive services, as shown in Figure 4-5. The slice is also selected for UE.

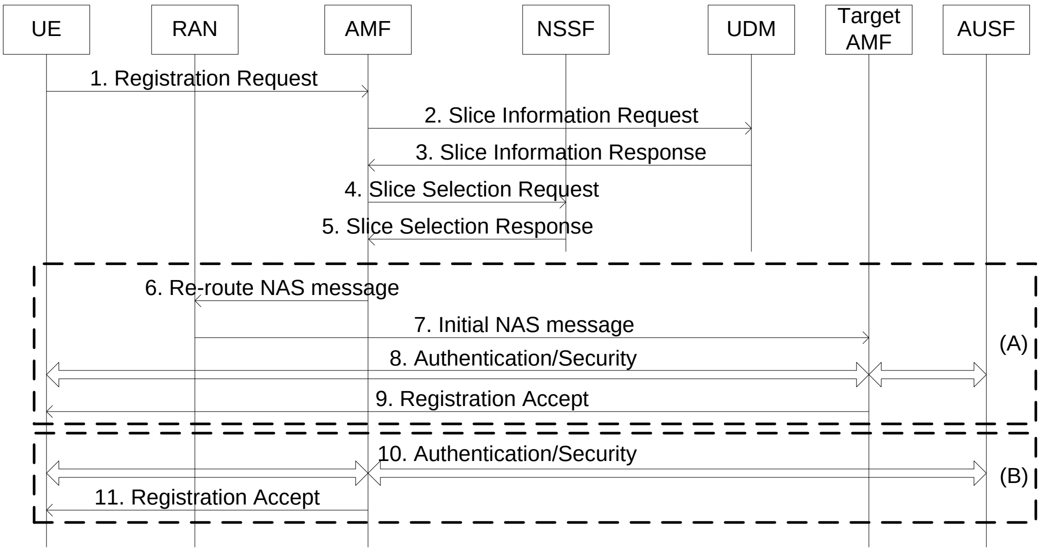

UE
RAN
AMF
NSSF
UDM
Target AMF
AUSF
1. Registration Request
2. Slice Information Request
3. Slice Information Response
4. Slice Selection Request
5. Slice Selection Response
6. Re-route NAS message
7. Initial NAS message
8. Authentication/Security
9. Registration Accept
(A)
10. Authentication/Security
11. Registration Accept
(B)

Figure 4-5: Registration procedure

1. UE sends Registration Request to AMF. The message includes requested NSSAI.
2. AMF sget subscription data from UDM.
3. UDM responds subscription data to AMF. The subscribed NSSAI includes subscribed S-NSSAI.
4. AMF sends Slice Selection Request to NSSF to select slice for UE. The message includes requested NSSAI, subscribed NSSAI, UE location.
5. NSSF determines allowed NSSAI and selects target AMF based on information provided by AMF. NSSF sends Slice Selection Response to AMF. The message includes allowed NSSAI，mapping between the S-NSSAI in the Allowed NSSAI and the NRF and AMF list. The AMF list may include AMF IP address list or FQDN list.

   AMF determines whether it is the target AMF based on whether it is included in the AMF list. If it is the target AMF, then the steps in box (B) are performed, the steps in box (A) are skipped. Otherwise the steps in box (A) are performed, the steps in box (B) are skipped.
6. AMF determines that it is not the target AMF, it sends Re-route NAS message to RAN. The message includes the Registration Request and the identifier of the target AMF.
7. RAN sends the Registration Request to the target AMF.
8. Authentication and security are performed.
9. Target AMF sends Registration Accept message to UE.
10. AMF determines that it is the target AMF. Authentication and security are performed.
11. AMF sends Registration Accept message to UE.

4.2.4 Core network slicing PDU session establishment

PDU session establishment procedure is used by UE to establish a new PDU session. During the procedure, the AMF has to select the SMF to serve the PDU session. The procedure for non-roaming and roaming with local breakout scenarios is shown in Figure 4-6.

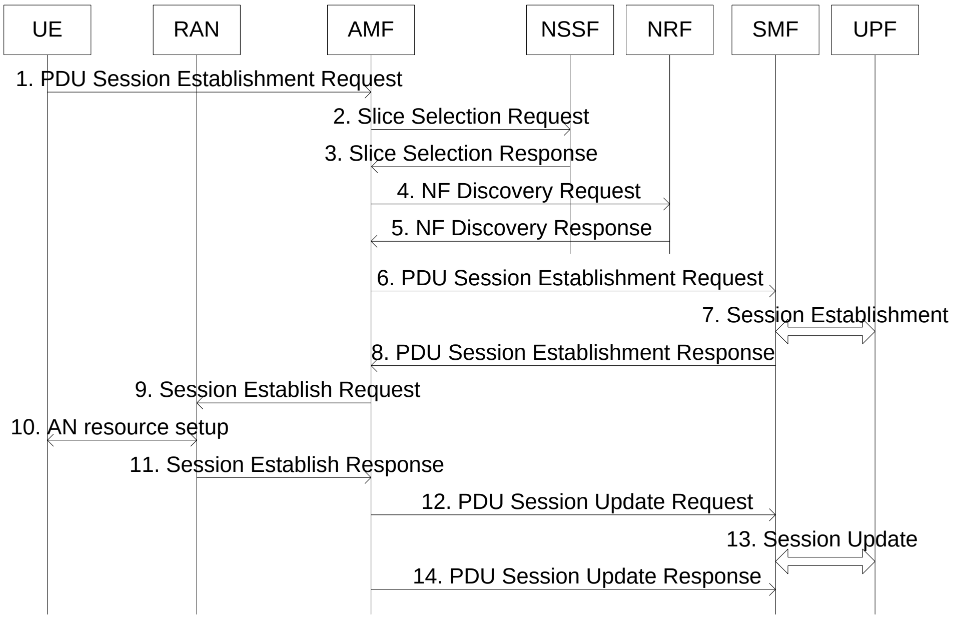

UE
RAN
AMF
NSSF
NRF
SMF
UPF
1. PDU Session Establishment Request
2. Slice Selection Request
3. Slice Selection Response
4. NF Discovery Request
5. NF Discovery Response
6. PDU Session Establishment Request
7. Session Establishment
8. PDU Session Establishment Response
9. Session Establish Request
10. AN resource setup
11. Session Establish Response
12. PDU Session Update Request
13. Session Update
14. PDU Session Update Response

Figure 4-6: PDU session establishment procedure (non-roaming or roaming with local breakout)

1. UE sends PDU Session Establishment Request to AMF. The message includes requested S-NSSAI I in the Allowed NSSAI.
2. AMF sends NF Discovery Request to NRF based on the mapping between the S-NSSAI and the NRF returned during registration procedure, the message includes requested S-NSSAI.
3. NRF selects SMF and sends the NF Discovery Response to AMF. The message includes SMF ID.
4. AMF sends the PDU Session Establishment Request to SMF.
5. SMF selects UPF and establishes session.
6. SMF sends PDU Session Establishment Response to AMF. The message includes the tunnel information of the UPF.
7. AMF sends the Session Establishment Request to RAN. The message includes tunnel information of the UPF
8. RAN establishes AN resource for the PDU session.
9. RAN sends Session Establishment Response to AMF. The message includes tunnel information of the RAN.
10. AMF sends PDU Session Update Request to SMF. The message includes tunnel information of the RAN.
11. SMF sends the tunnel information of the RAN to UPF.
12. SMF sends PDU Session Update Response to AMF.

Then the NSSF in VPLMN returns the NRF in HPLMN and NRF in VPLMN to AMF. The AMF sends the requested S-NSSAI and NRF in HPLMN to NRF in VPLMN. The NRF in VPLMN selects SMF in VPLMN and sends requested S-NSSAI to NRF in HPLMN. The NRF in HPLMN selects the SMF in HPLMN and send the ID of the SMF to the NRF in VPLMN. After receiving the ID of the SMF in HPLMN, the NRF in VPLMN sends the two SMFs to AMF.

5 Transport Network Slicing

5.1 Scenarios and requirements

With the development of 5G, it raised the following technical requirements for the transport network:

Large bandwidth: to meet requirements of eMBB, 5G base station density and single station capacity are greatly raised, and there appears demand for higher data rate and lager capacity of the transport network.

Low latency: the URLLC makes strict demands on the forwarding delay of the transport network.

Network slicing: in the age of 5G, the same transport network needs to support services of eMBB, URLLC and mMTC etc., and needs slicing network for different services.

Flexible scheduling of the services: for the separately placement of 5G Centralized Unit/Distributed Unit (CU/DU), MEC subsidence, virtualization, etc., the transport network needs to provide inter-cloud connectivity and connection-oriented carrier grade packet transport.

Transport network slicing for different services such as URLLC, eMBB and mMTC or for other reasons is a very important requirement on 5G network. Four layer networks are shown in Figure 5-1, which are Physical, Hard Pipe, Soft Pipe and Service layer network:

Physical Layer Network: It is physical resource, such as rack, port and fiber link etc.

Hard Pipe Layer Network: It is L1 hard pipe and can't statically multiplex, such as optical, OTN or FlexE layer network;

Soft Pipe Layer Network: It is L2 or L3 network, such as Ethernet or IP network;

Service Layer Network: It represents client layer network of transport network.

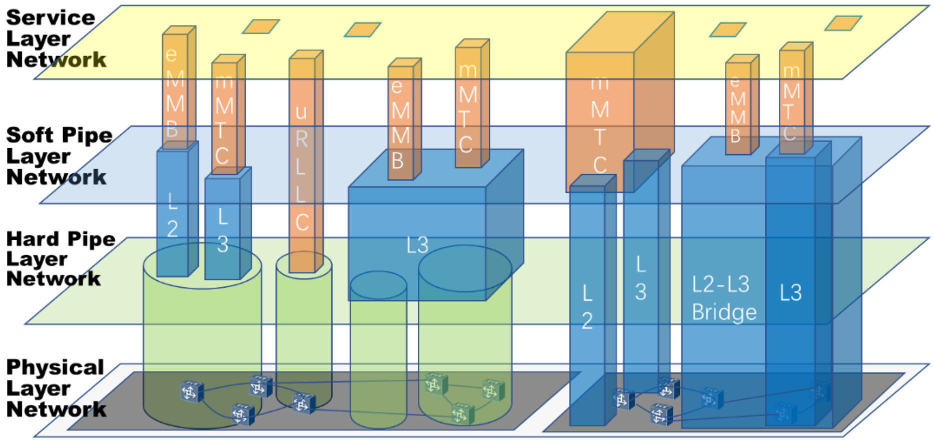


Figure 5-1: Network Slicing for different layer network

Transport network slicing happens inside specifically layer network, so it can be classified into four types:

- **Physical Slicing**: slicing of the network resource (e.g., assignments for network elements, ports, links, racks etc.)
- **Hard Pipe Slicing**: L1 channel slicing (e.g., the channels of WDM, OTN, FlexE)
- **Soft Pipe Slicing**: Ethernet / IP virtual network segment (e.g., L2 VPN, L3 VPN, L2 / L3 Bridge)
- **Service Slicing**: service type slicing (e.g., URLLC, eMBB, mMTC)

As shown in Figure 5-1, between slicing of Physical Layer Network, Hard Pipe Layer Network, Soft Pipe Layer Network and Service Layer Network, there is network slicing mapping relationships between adjacent slicing of layer network. Following are some examples that helpful to understand network slicing mapping in Figure 5-1:

(1) Mapping of Service Slicing to Hard Pipe Slicing to meet the ultra-low latency requirements for URLLC scenarios.

(2) Mapping of Service Slicing to Soft Pipe Slicing to meet the ultra-low latency requirements for URLLC scenarios.

(3) Mapping of Soft Pipe Slicing to Physical Slicing to meet the requirements of eMBB and mMTC service with L2 / L3 VPN technology.

5.2 Key technologies

Compared with previous wireless technology, 5G has proposed new requirements in terms of bandwidth, latency, network slicing, service maintenance etc. For example: Bandwidth requirements is promoted from GE / 10GE to 25GE*N / 100 GE; latency requirements of transport node is promoted from 50us to 5 us; network slicing requirements is promoted from "soft pipe" to "hard pipe"; Service maintenance is promoted from connectionless private network management to connection oriented unified control.

In the face of the above requirements, new transport network technology is needed: the forward is based on the SR over slice Ethernet / OTN over DWDM, and the control is based on SDN. New transport network should meet the needs of 5G and future transport networks through innovative technologies at the L1/L2/L3.

Figure 5-2 shows the total solution for 5G transport. It includes fronthaul transport, middlehaul transport, backhaul transport, SDN controller and network slicing.

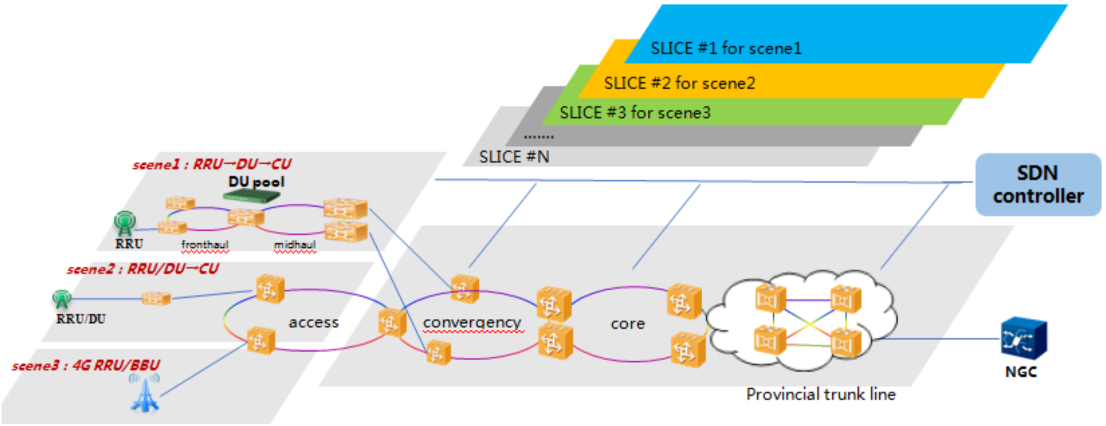


Figure 5-2: Total solution for 5G transport

Transport network slicing consists of control plane slicing and forwarding plane slicing. Control plane achieves end-to-end resource allocation, provides control interfaces for network applications, and provides resource allocation and configuration for forwarding planes. The forwarding plane provides the corresponding QoS guarantee according to the needs of different service slicing, and meets the different requirements of various services with the same physical network.

5.2.1 Control plane slicing

Control plane provides encapsulating and scheduling capabilities for the services. SR and SDN control are two key technologies for control plane slicing.

SR technology is based on application of Segment Routing to transport network. Comparing with the traditional control plane technology, SR technology realizes the decoupling between service and network. The service establishment only needs to operate at the edge node, and can be seamlessly connected with the SDN centralized control. SR provides "connection oriented" and "connectionless" pipeline to meet the 5G cloud networking flexible connectivity requirements.

SDN control provides centralized routing control capability and flexible programming ability to achieve flexible scheduling of the service. SDN centralized controlled L3VPN can achieve service flexible scheduling by providing centralized routing control, routing strategy and flexible programming ability. It can also effectively reduce the complexity of new transport equipment with suitable combination of centralized routing and distributed protocol.

5.2.2 Forwarding plane slicing

The forwarding plane needs to support L2/L3 layer high-speed forwarding, L1 layer channelization, physical isolation, and capability expansion.

Double plane switching

Different requirements of 5G service need the double plane switching of "packet switching + slot switching ". Compare with traditional L2/L3 VPN, new transport network can achieve 10 times lower latency through the slot switching. New transport network can also achieve low transport jitter based on the physical isolation properties of the slot switching, wich is better than traditional L2/L3 VPN., and it is a better scheme to meet the double requirements of L3 sink and ultra-low latency.

Hard isolation

Comparing with the soft isolation of traditional L2/L3 VPN, the transport network slicing provides a hard isolation of different services. 5G integrated mobile transport network needs to support multi services, guarantee network slice resource, and provide hard isolation between slicing and service. Comparing with the traditional QoS based (best effort) L2/L3 VPN slicing, slot switch based transport network slicing not only achieve a hard isolation between different services, but also ensure the low jitter. In contrast to the look-up table and queuing of L2/L3 VPN technology, low delay and deterministic delay can’t be guaranteed.

Capability expansion

The new transport network can combined DWDM with the traditional physical electrical layer interface. The new transport network supports large bandwidth on the physical electrical layer pipe. The new transport network can also realize the flexible extension of electric layer pipe to meet the application requirements of high bandwidth through the multiple wavelength optical layers bonding. The wavelength based bonding not only greatly increases the expansion capacity of the channel bandwidth, but also realizes the huge advantage of cost by saving the fiber. The specific transport equipment configuration can be a flexible combination of the electrical and optical layers, which can be configured as integrated equipment with both electrical and optical layers, and also can be used separately by the electrical layer. The new transport network's architecture satisfies the bandwidth requirements from fronthaul, middlehaul to backhaul.

6 Radio Access Network Slicing

6.1 Requirements, scenarios, objectives and design principles

6.1.1 Requirements

5G RAN should support E2E network slicing, making RAN network share the same infrastructure, schedule resources dynamically, and construct different logic networks to adapt to the diverse network requirements. RAN slicing includes the following requirements [9]:

- The requirements of the E2E slicing, including slice awareness, association, isolation, customization and SLA, etc.

In order to support the standardized E2E slicing in 3GPP, the access network needs to connect to the O&M, 5G new core and UE. Such requirements need to be fully defined in industry standards.

- The requirements from the RAN-particular technical characteristics, such as improving radio resource efficiency in the E2E network slice technology.
- The requirements from RAN O&M. For example, we need to provide the corresponding management methods of RAN to support the demand of virtual operators on wireless resources.

6.1.2 Scenarios and slicing granularity

RAN may support diverse network slices with eMBB, URLLC and mMTC slice/service types as well as non-standardized slice/service types. Network slices may have different slice requirements. To enable differentiated handling of traffic for network slices with different requirements, RAN has a set of different configurations for different network slices. To select the appropriate configuration for the traffic for each network slice, NG-RAN receives relevant information indicating which of the configurations applies for this specific network slice. Network can realize the different network slices by scheduling and by providing different L1/L2 configurations.

6.1.3 Objective

The objective of RAN part of NSI (RAN slicing) is to allow RAN differentiated treatment depending on each customer requirements. With slicing, it is possible for an MNO to consider customers as belonging to different tenant types with each having different service requirements that govern in terms of what slice types each tenant is eligible to use based on different subscriptions and requirements.

6.1.4 Design principle

The NSI crosses multiple network domains, including terminals, access networks, transport networks and core networks, as well as data centers. Each NSI has a specific network topology, network functions, and allocated resources. Based on this principle, RAN should have the ability to be aware of network slicing to isolate resources between different network slices and the ability to support the three important functions as follows:

- CN NSI selection: includes selecting suitable CN NSI for the UE according to the NSSAI
- RAN slicing O&M: includes RAN slicing monitor and self-control to support the management of business requirements (SLAs).
- RAN resource management: includes radio resource coordination and isolation among different slicings.RAN should support radio resource isolation that guarantees no effect on different slices especially for the access, and radio resource coordination sharing that improve the reuse benefit, such as a common node to support multiple slicing. A common scheduler can realize radio resource isolation and allocation for different slices.

Principle 1：Unified logical architecture, flexible deployment on slicing

A unified logical architecture is applied to different access technologies such as NR, EUTRAN, as well as the centralized and distributed networks. The flexible deployment can achieve in implementation based on different scenarios. For example, the RAN PDCP supports differentiated deployment in order to meet the diversified delay requirements of different slices.

Principle 2: RAN resource logical isolation

The RAN resources are diverse, and the isolation of different types of resources depends on the needs of operators and tenants. RAN resources can be separated by frequency, Time, Code, hardware equipment, software and other dimensions. Based on reuse of RAN logical resource slicing, services can be isolated logically. The resource efficiency can be improved with flexible usage.

Principle 3: Differentiated implementation

According to 5G RAN function definition, 5G RAN functions include real-time functions such as DU and non-real-time functions such as CU-C and CU-U. The non-real-time functions focus on the customization to satisfy the demands of the slicing. In addition, the real-time functions focus on the utilization efficiency of resources, and provide differentiated slicing functions through flexible configurations, such as occupying the dedicated resources or sharing a certain proportion of resources.

Principle 4: Ensuring key performance

The slicing can give the network the flexibility to introduce new business model and flexible resource usage strategy. However, the main premise of slicing is to guarantee that the network performance including spectrum efficiency, communication quality and system capacity.

6.2 Key technologies

6.2.1 RAN slicing architecture

RAN needs to meet the O&M requirements as well as assure the service requirements when NSSAI received. Therefore, in the RAN, the E2E slicing parameters can be divided into two parts: one is related to operation, and the other is related to the service of UE.

In Figure 6-1, gNB receives NSSAI and then generates the configurations and parameters for operation and control according the NSSAI.

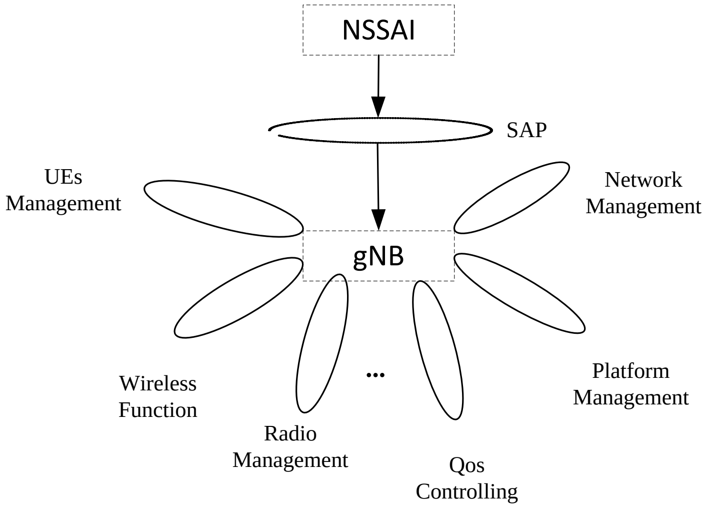


Figure 6-1: E2E slicing support in RAN

From the O&M perspective, it mainly includes the functions of platform management and network management.

Platform management includes dynamic shutdown and startup, monitoring, deployment and configuration on the platform, such as the platforms for the vertical industry and common communication may need to be isolated from each other. It also includes unified management and distribution functions on the cloud platform. For example, through virtualization technology, undifferentiated capacity and overall management on the platform is required.

Network management includes network partition for the tenant network management, network isolation and division, such as network function for the vertical industry and the common communication may be isolated and divided. It also includes radio resource division, such as using different configurations of antennas to realize the signal isolation.

From the perspective of service, service management for RAN slicing includes the UE management, wireless function deployment, radio resource management, QoS controlling aspects, and so on.

UE management includes UE context maintenance and storage for the UE, which provides accurate results for UE context establishment, release, reconfiguration and report. If the UE belongs to a different virtual operator (Tenant), UE information should be isolated and secured between slicing.

Wireless function deployment includes flexible deployment according to the requirements of E2E slicing; such as for the low-latency service the wireless function should be near the RRU to meet the latency requirement. In addition, if its latency is not so tight, a part of wireless functions placed near the RRU, and other functions are placed far away from the RRU to get the centralized benefit.

Wireless resource management includes frequency domain and time domain management with the requirement of slicing. For example, the bandwidth of 100Mhz can be divided into a number of small bandwidth for the targeted users. In addition, we can change the wireless resources in real time based on the scheduling and UE requirement. Using real time scheduling and allocation, radio resource can be soft isolated though among different virtual operators.

QoS control includes providing suitable UE service quantization definition, E2E slice guarantee, network O&M. For the ultra low latency service in virtual operator, QoS control includes guarantee the QoS requirement according to the ability of RAN. An NSI may contain different types of RAN parts of NSI, such as 3GPP NR in terms of eMBB, URLLC and mMTC etc and Non-3GPP WLAN. Consolidating fix line and mobile access in 5G is a desirable approach, which also requires further updates on the architecture design. One UE can be associated with more than one RAN parts of NSI where some L2 and L3 components in addition to L1 resource shall be shared.

RAN slicing should have the ability to meet various requirements by differentiated RAN services. RAN slicing can be implemented based on the Cloud RAN architecture. Isolation and sharing can be supported at different RAN slices, that is, there can be common sharing functions and slice-specific functions for each slice. As shown in Figure 6-2, NR and E-UTRAN has a shared RRC and a dedicated RRC respectively. At the same time, as shown in Figure 6-3, a PLMN can support multiple groups of RAN slices, and each RAN slice group is associated with a core NSI. Slices that belong to different PLMN can share or use exclusive RAN radio resources based on business requirements. On the cloud RAN platform, the transport network could also be shared between RAN slices corresponding to multiple PLMNs. Base on 5G flexible air interface design, a unified common MAC scheduler across multiple slices is applied to achieve various requirements of different types of slices. Considering different RAN slices configured, MAC scheduler allocates different time and frequency resources and takes the balance between radio resource logical isolation and usage efficiency.

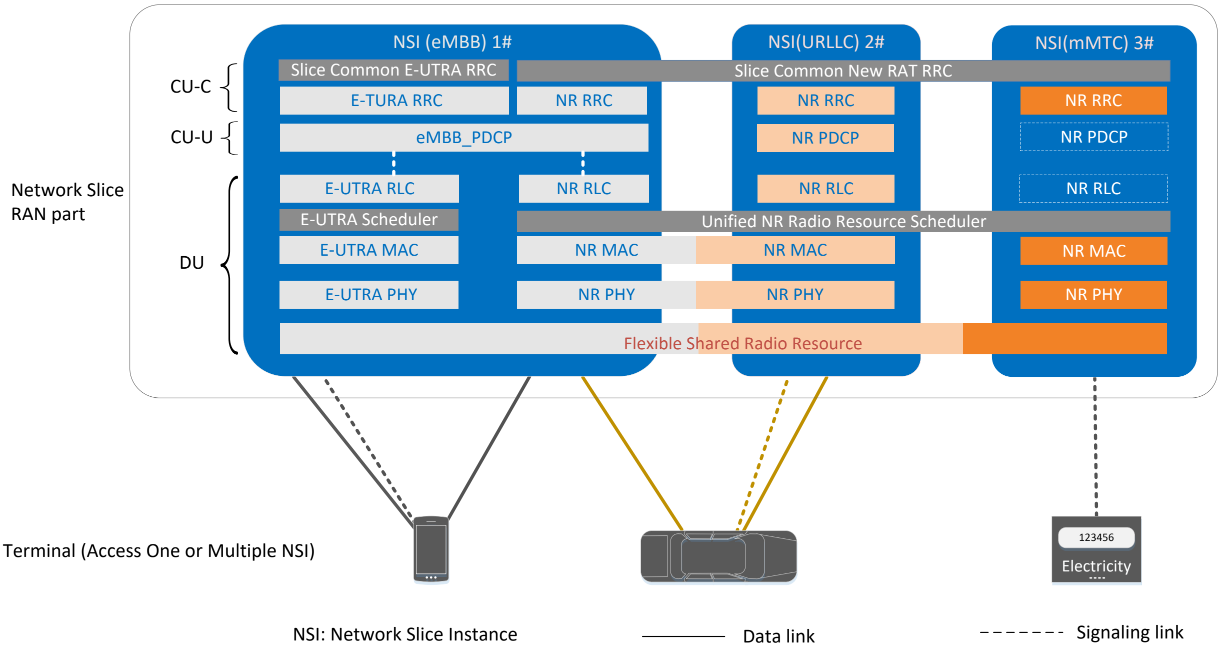


Figure 6-2 RAN architecture with network slicing support example

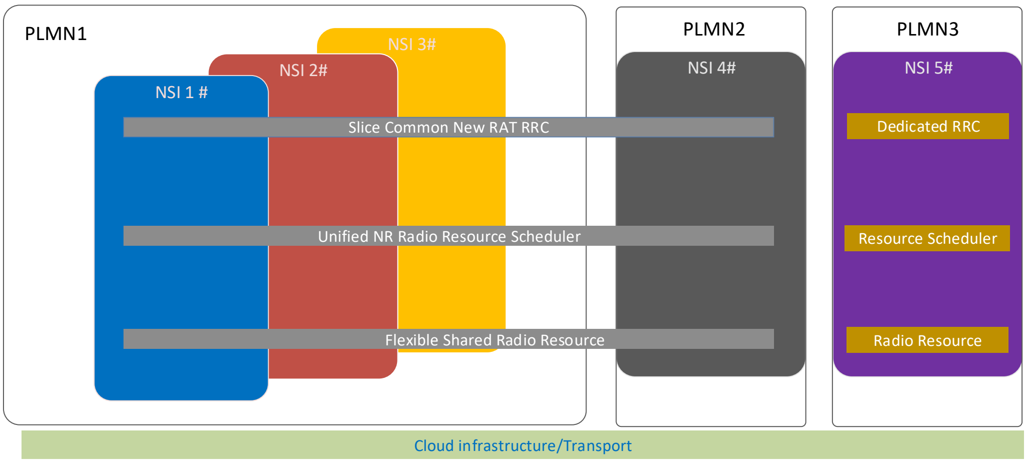


Figure 6-3 The relationship between RAN slicing and PLMN

6.2.2 RAN slicing procedures

6.2.2.1 Access and Network Slice Selection

RAN selects AMF based on a Temp ID or assistance information provided by the UE over RRC. In case a Temp ID is not available, the RAN uses the information provided by the UE during RRC connection establishment to select the appropriate AMF instance (details are FFS and subject to RAN2 standardization). If such information is also not available, the RAN routes the UE to a default AMF instance. The access and network slice selection procedure is shown in Figure 6-4[12].

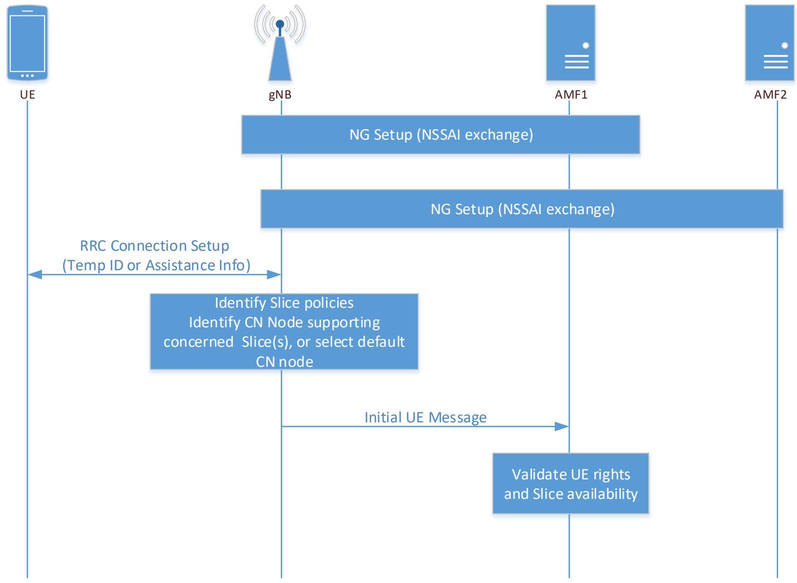

UE
gNB
AMF1
AMF2
NG Setup (NSSAI exchange)
NG Setup (NSSAI exchange)
RRC Connection Setup
(Temp ID or Assistance Info)
Identify Slice policies
Identify CN Node supporting
concerned Slice(s), or select default
CN node
Initial UE Message
Validate UE rights
and Slice availability

Figure 6-4: Access and network slice selection procedure

6.2.2.2 UE Context Handling

Following the initial access, the establishment of the RRC connection and the selection of the correct AMF, the AMF establishes the complete UE context by sending the Initial Context Setup Request message to the NG-RAN over NG-C. The message contains the S-NSSAI as part of the PDU session/s resource description. Upon successful establishment of the UE context and allocation of PDU resources to the relevant NW slice/s, the NG-RAN responds with the Initial Context Setup Response message. The UE context handling procedure is shown in Figure 6-5[12].

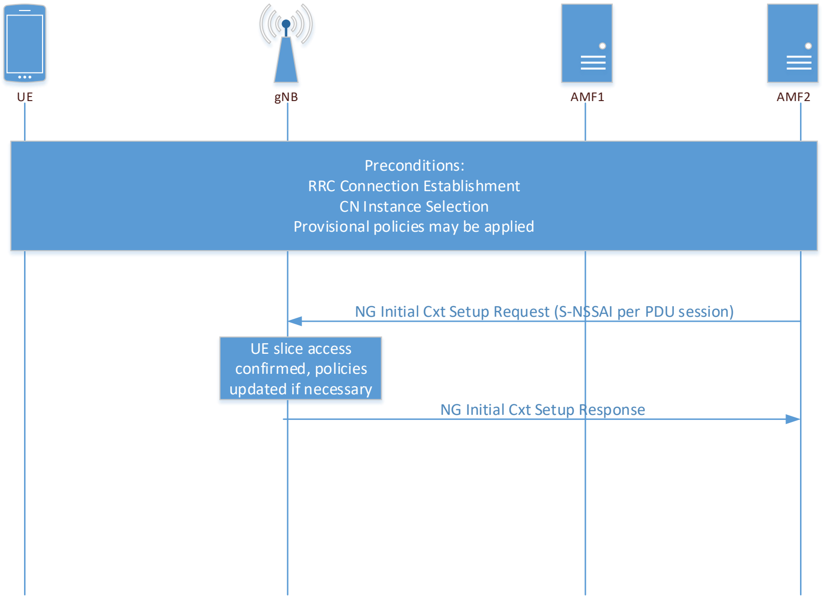

UE
gNB
AMF1
AMF2
Preconditions:
RRC Connection Establishment
CN Instance Selection
Provisional policies may be applied
NG Initial Cxt Setup Request (S-NSSAI per PDU session)
UE slice access
confirmed, policies
updated if necessary
NG Initial Cxt Setup Response

Figure 6-5: UE context handling procedure

6.2.2.3 PDU Session Handling

When new PDU sessions need to be established or existing ones modified or released, the 5GC requests the NG-RAN to allocate/release resources relative to the relevant PDU sessions by means of the PDU Session Setup/Modify/Release procedures over NG-C. In case of network slicing, S-NSSAI information is added per PDU session, so NG-RAN is enabled to apply policies at PDU session level according to the requirements of different network slice, while still being able to apply (for example) differentiated QoS within the slice. NG-RAN confirms the establishment/modification/release of a PDU session associated to a certain NW slice by responding with the PDU Session Setup/Modify/Release Response message over the NG-C interface. The PDU session handling procedure is shown in Figure 6-6[12].

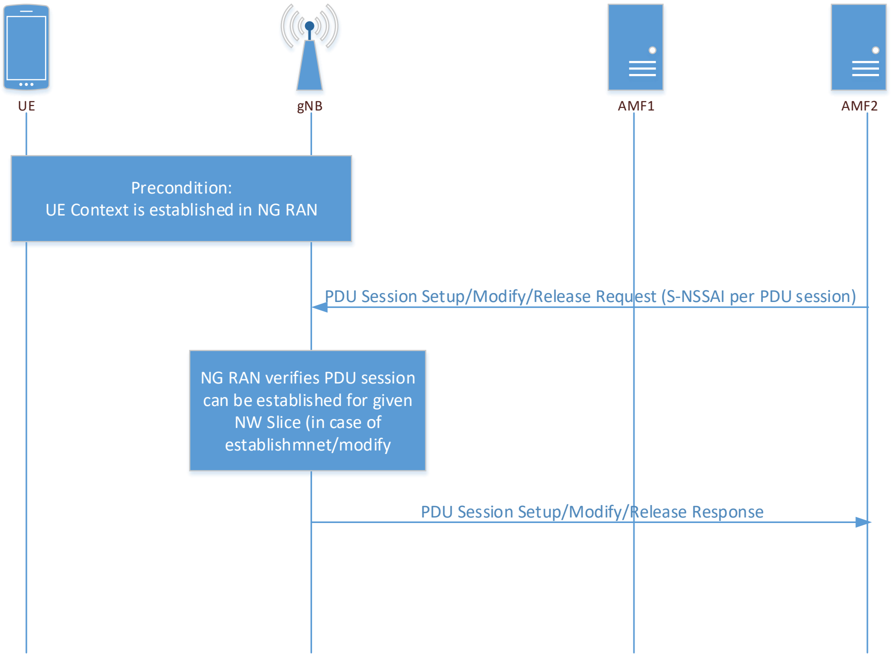

UE
gNB
AMF1
AMF2
Precondition:
UE Context is established in NG RAN
PDU Session Setup/Modify/Release Request (S-NSSAI per PDU session)
NG RAN verifies PDU session can be established for given NW Slice (in case of establishmnet/modify
PDU Session Setup/Modify/Release Response

Figure 6-6: PDU session handling procedure

6.2.2.4 Mobility

To make mobility slice-aware in case of Network Slicing, S-NSSAI is introduced as part of the PDU session information that is transferred during mobility signalling. This enables slice-aware admission and congestion control. The related handover procedure is shown in Figure 6-7[12].

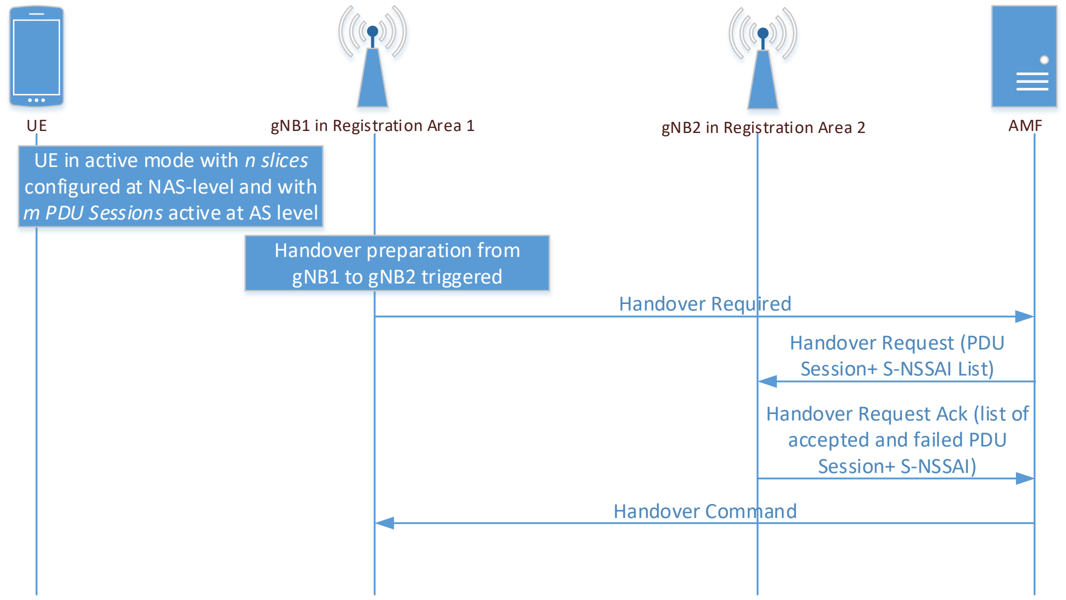

UE
gNB1 in Registration Area 1
gNB2 in Registration Area 2
AMF
UE in active mode with *n slices* configured at NAS-level and with *m PDU Sessions* active at AS level
Handover preparation from gNB1 to gNB2 triggered
Handover Required
Handover Request (PDU Session+ S-NSSAI List)
Handover Request Ack (list of accepted and failed PDU Session+ S-NSSAI)
Handover Command

Figure 6-7: Handover procedure

6.2.3 RAN real-time function slicing

According to 3GPP specifications for service interaction, new RAT functions consist of real time (RT) functions and non-real time (NRT) functions. The RT functions are preferred to be deployed in DUs close to terminals. And the RT functions should optimize the resource utilization and ensure that the security and isolation requirements of NSIs are satisfied. Specifically,

- The RT functions could either occupy the resources exclusively or share a portion of the resources according to the flexible configuration.
- Different RT functions can apply different physical layer numerologies (e.g., frame-related parameters) to provide functional differentiation for different NSIs.
- In MAC, logical channel prioritization procedure should take into account the mapping of one logical channel to one or more numerologies and/or transmission time interval durations based on the slice specific requirements.
- RLC functions (e.g. segmentation/re-segmentation and ARQ etc) are tailored to meet the slice requirements.

6.2.4 RAN non-real-time function slicing

In the realization of NSI in the AN, the NRT functions can apply cloudification and orchestration technologies, in order to satisfy the customization demands, such as tailored access functions and configuration differentiation. The NRT functions can be deployed in a CU.

7 Terminal Slicing

Network slicing is the key enabling technologies to meet two major challenges: 1) to support the vertical industry applications so that to expand the wireless industry market and 2) to further enhance device capability and user experience. We apply network slicing in two dimensions: Vertical network slicing and horizontal network slicing [11]. Vertical slicing targets at supporting vertical industry and markets. It enables resource sharing among services and applications and avoids or simplifies a traditional QoS engineering problem. Horizontal slicing, as a step forward, targets at extending the capabilities of mobile devices and enhancing user experiences. Horizontal slicing goes across and beyond platforms physical boundaries. It enables resource sharing among network nodes and devices, i.e., high capable network nodes/devices share their resources (e.g., computation, communication, storage) to enhance the capabilities of less capable network nodes/devices. Horizontal slicing requires over-the-air resource sharing across network nodes.

7.1 Vertical Slicing

In late 4G and early 5G, the wireless industry started to vertically slice big mobile broadband network into multiple virtual networks to serve vertical industries and applications in a more cost efficient manner. Each network slice can have different network architecture, and different application, control, packet and signal processing capacity to achieve optimum return on investment. A new industry or type of service can be added to an existing network instead of deploying a new network. Vertical network slicing practically segregates the traffic from a vertical application standpoint from the rest of general mobile broadband services, practically avoiding or dramatically simplifying a traditional QoS engineering problem (i.e., QoS mechanism can be imposed to each slice instead of imposed to all the traffics among all the slices). Vertical network slicing was primarily focused on core network nodes enabled by techniques such as NFV and SDN. With time we see this trend continues and expends into the radio access networks and the air interface. Examples of vertical slices developed or under developing in 4G LTE include the MTC slice, and the narrow-band internet-of-things (NB-IOT) slice. Adding new slices in LTE is to patch on to the baseline LTE framework which was primarily designed for mobile broadband communication. In 5G, a forward-compatible design is desirable to provision for adding new slices in the future.

The device can be authenticated and attached to a diverse set of network slices as needed that are each tailored to a specific purpose: small short message, streaming video, voice calls, internet browsing, chatting and so on. The independence of network slices indicated in the 3GPP requirements ensures that the high-quality video call that it will not be interrupted nor impacted by other background functions. Such services as video calling can be established with high security, while the security applied to other activities may use completely different methods. Figure 7-1 illustrates some vertical slicing example:

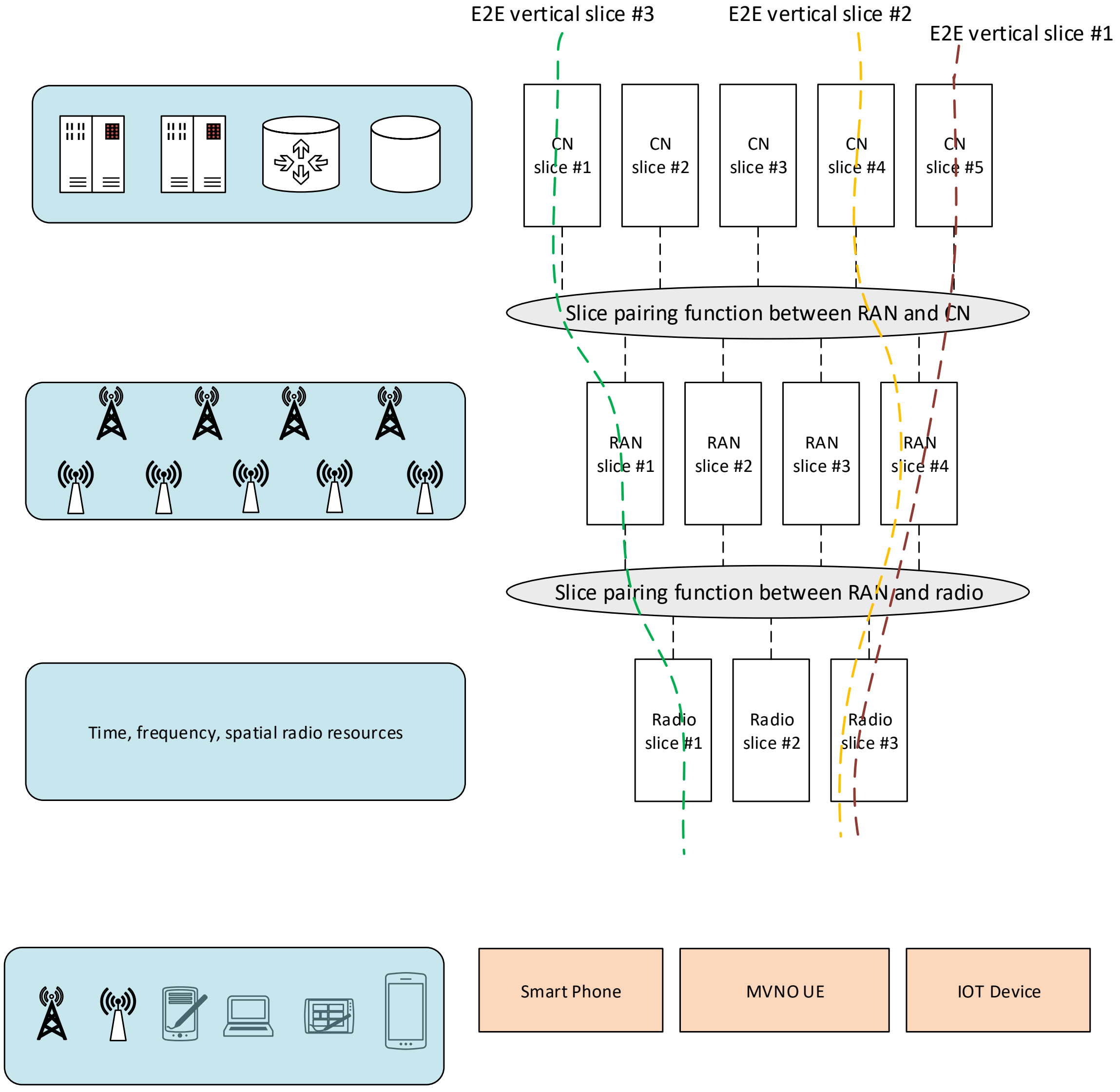


Figure 7-1: The vertical slicing examples.

7.2 Horizontal Slicing

As the work on 5G began in 3GPP, it will bring new type devices those are becoming increasingly popular and are expected to become essential in our everyday life. Examples include wearable devices for advanced telemedicine applications, virtual/augmented reality, UHD video and V2X applications other than smart phone. Despite continuous improvement of hardware, the mobile device capabilities still remain a concern due to the limitation of computing, storage and power supply resources. Horizontal slicing augments performance of mobile devices by leveraging the latest innovations in cloud computation and virtualization technologies, here virtualization technologies provide a key foundation for network slicing by enabling use of both physical and virtual resources to create the service they are designed for. moving forward, we expect the network capacity and user experience need to be further improved. However, the capacity increase does not need to be E2E uniform: the capacity scaling factor can be higher when closer to a user, and lower as we move deeper into the infrastructure network as shown in Figure 7-2 where we depicted a case with 10,000 capacity times scaling at the very edge of the network and a 10 times scaling at the network core. Such non-uniform capacity scaling is driven by the new types of user traffics and services and is founded by communication technology fundamentals: We expect a large amount of user traffic generated at the network edge due to the ever increasing number of devices and the significantly increased device sensing capability. As device sensing is local, it is desirable to keep sensed data processing and the corresponding decisions and actions to be local so that to reduce latency and improve privacy and security. As a result, the amount of data going into the deeper network will be less and the traffic scaling will be non-uniform. The non-uniform traffic scaling requires non-uniform network capacity scaling.

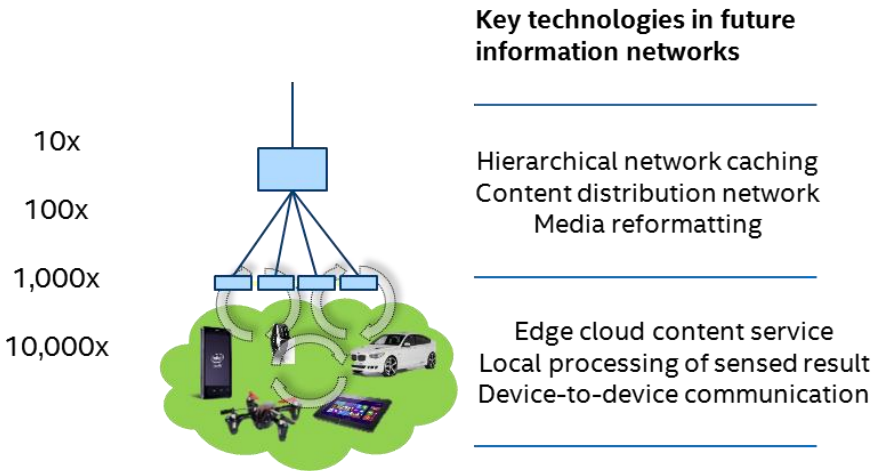
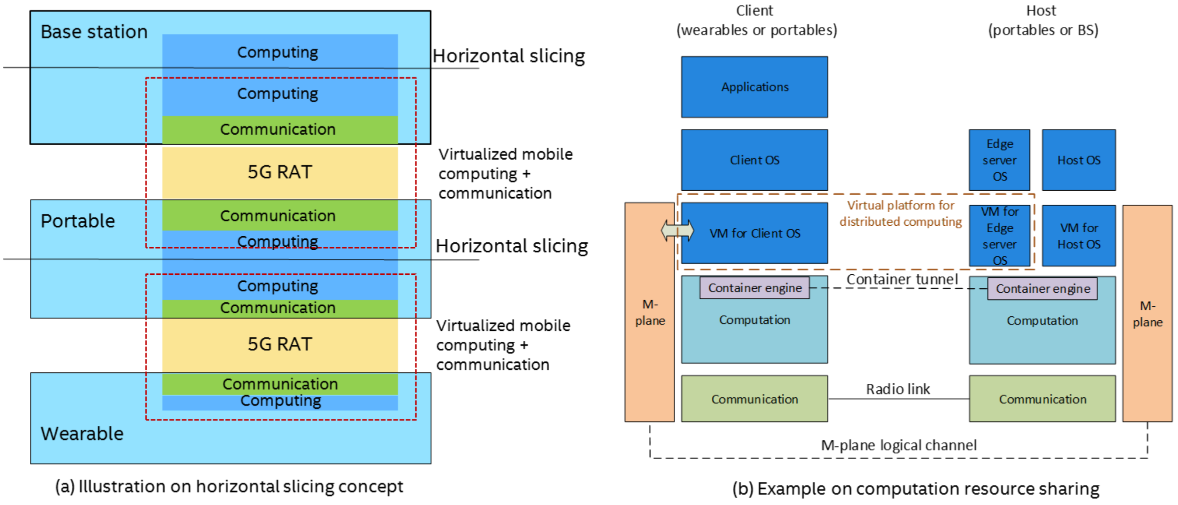

Key technologies in future information networks
10x
100x
1,000x
10,000x
Hierarchical network caching
Content distribution network
Media reformatting
Edge cloud content service
Local processing of sensed result
Device-to-device communication

Figure 7-2: Illustration on non-uniform capacity scaling

The non-uniform capacity scaling can be viewed as using computation to help communication, i.e., edge computing and processing reduce traffic towards the deeper infrastructure network. The augment of device capability and the user experience enhancement, on the other hand, largely relies on using communication to help computation. Despite that the devices further shrink in size from portable devices to wearable devices, the user expectation on computation keeps increasing. We expect computation offloading will be needed to help deliver user experience, e.g., base stations slice part of their computation resource to help computation at the portable devices, portable device slice out a part of their computation resource to help the computation of the wearable devices. Horizontal slicing helps devices to go beyond its physical limitation and is designed to accommodate the new trend of capacity scaling and enable edge cloud computing and computing offloading: The computing resources in the base station and the portable device (or high-capable device) will be horizontally sliced, and these slices, together with the computation resource slice of the wearable devices (or low-capable device) will be integrated to form a virtual computing platform through a new 5G air interface design to significantly augment the computing capability of future portable and wearable devices as well as perform local traffic processing. With horizontal slicing, the definition of E2E needs to be revisited, as the traffic flow terminates within the horizontal slice built among devices with direction communication link (and likely in close proximity).

Figure7-3 provides an example on illustrating the horizontal slicing concept. Computation resource slicing and the air interface supporting computation resource sharing need to be jointly designed for optimized performance (Figure 7.3(a) illustrates the horizontal slicing concept. Figure 7.3(b) presents an example architecture on realizing computation slicing and resource sharing). The main building blocks are a communication module, a computation module, a management-plane (M-plane) module, the virtual machines (VM) formed by virtualization technique and the operation systems (OS) running on the VMs. Computation slicing is managed by the M-plane and implemented below OS. The M-planes monitors the system resource usage and the radio link condition. When the M-plane of the client sees benefit (in terms of performance, costs, etc.) of computation offloading or when the client OS requests resource beyond the client computation can support, it asks the host M-plane for sharing computation resource. The signaling exchange between the M-planes at the client and host are carried as air interface L3 signaling message. If the client's request for resource is accepted by the host, the client M-plane informs the client VM on the available of the resource. The client VM then slices the sliceable application based on the information from the M-plane and convey the generated executable code to the container engine, where container is developed to facilitate distribution and execution of sliceable application [8]. The container engine packs the sliced executable code into container and convey the data packet carrying the container to the communication module. The communication module packs the container data into L2 PDU and generate L1 data block and transmit to the host via the selected radio link. The communication module at the host decodes the received data blocks and convey to the container engine. The host container engine unpacks the received container and hands the executable code to the edge server VM. Upon completion of the computation task, the container engine of the host packs the executed results in the container and deliveries it back to the client container engine via communication link. The client container unpacks the executed result. The client OS then applies the received executed results.



(a) Illustration on horizontal slicing concept

(b) Example on computation resource sharing

Figure 7-3: Illustration on enhancing device capability using horizontal slicing

Figure 7-4 illustrates the described computation resource sharing procedure. Note that in the described example, the computation task slicing is implemented below OS. Alternatively, the slicing can be implemented at OS or above. In this case, the container information can be treated as normal user traffic and no special design is needed in the air interface. However, the performance (such as processing delay) will be affected by the OS performance. Implementing slicing at below OS is expected to give better performance. In this way, the host computing resource is virtualized into two parts. One part is used by its own applications, running on its own OS. Another part is allocated for client, and used by the client as remote resource. The application at the client has to be sliceable to be executed at two VMs simultaneously. One at the client, the other at the host. The client VM may serve as the master VM, and the host VM may serve as a co-processing engine to serve the client master VM. The traffic carrying the container can be made visible to the communication link. A dedicated L2 logical channel can be specified to carry the container traffic.

The global effort on network slicing is still in the early stage and is mostly focused on vertical slicing. As the industry moves towards vertical slicing as the phase-1 implementation, provision for enabling horizontal slicing as the phase-2 implementation is needed to improve user experience.

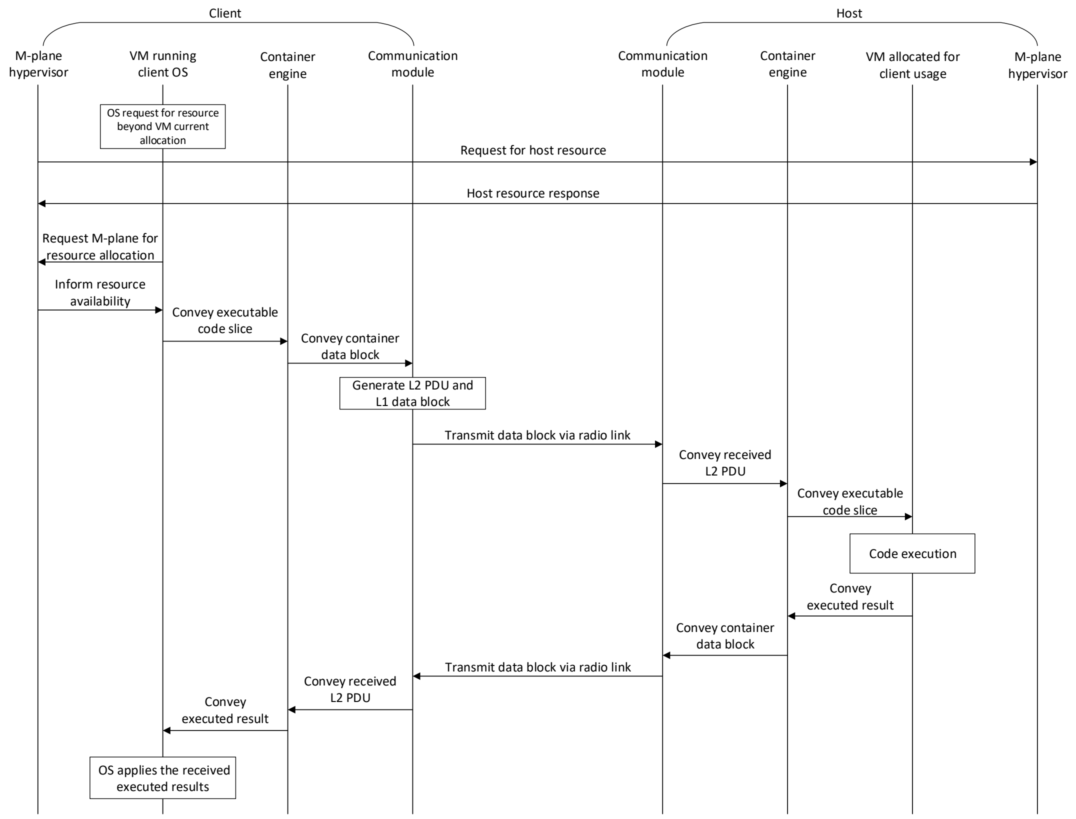

Client
Host
M-plane hypervisor
VM running client OS
Container engine
Communication module
Communication module
Container engine
VM allocated for client usage
M-plane hypervisor
OS request for resource beyond VM current allocation
Request for host resource
Host resource response
Request M-plane for resource allocation
Inform resource availability
Convey executable code slice
Convey container data block
Generate L2 PDU and L1 data block
Transmit data block via radio link
Convey received L2 PDU
Convey executable code slice
Code execution
Convey executed result
Convey container data block
Transmit data block via radio link
Convey received L2 PDU
Convey executed result
OS applies the received executed results

Figure 7-4: Computation resource sharing procedure

8 E2E Network Slicing Security

8.1 Security threats and requirements

8.1.1 Isolation of network slices

Network slices may be arranged to different applications or tenants. Without proper isolation between network slices, an attacker may launch attacks from one slice to others. For example, capacity elasticity of one slice may consume the resources of other slices. attackers may launch DoS attacks through this way.

A UE may simultaneously activate multiple network slices for different applications. Without proper cryptographic protection in slices, attackers may eavesdrop or tamper the data belonging to other slices. in a particular case where one slice is serving an UE through a untrusted non-3GPP access and another slice is serving the UE through a 3GPP access, the attacks on data confidentiality and integrity would be possible.

In order to maintain proper isolation between network slices, the following security requirements shall be considered:

- The resources allocated to different NSIs should have no impact on each other.
- Network slicing isolation can confine potential network attacks within a single NSI.
- Data confidentiality and integrity protection in network slices and data leakage avoidance between network slices should be supported.

8.1.2 Security mechanism differentiation for network slices

The security requirements of eMBB, eMTC and URLLC services will be different on 5G. For example, eMBB services may adopt enhanced traditional security mechanisms used in LTE, mMTC services may require quick authentication protocol and lightweight cryptographic algorithms, URLLC services may require strong authentication protocol, cryptographic algorithms and credential management. These services in 5G network could be arranged into different application-specific network slices. It means network slices may have their own security policies that require different security mechanisms to be applied according to the usages of slices. Moreover, the tenants of network slices may have different security requirements to the slices. Then, accesses and sessions may be compromised if there are no proper security configuration in network slices. In order to support differentiated security mechanisms, the following security requirements shall be considered:

- Slice-specific security policies (e.g. confidentiality protection, integrity protection, cryptographic algorithms and / or protocols) should be configurable according to the security requirements of network slices.
- Different security mechanisms (e.g. authentication mechanisms) can be used in different network slices.

8.1.3 Network slice authentication and authorization

A UE can simultaneously access multiple services delivered by different network slices. There may be a variety of network access, e.g. 3GPP, trusted non-3GPP and untrusted non-3GPP access. Furthermore, in the context of IoT, there will be a proliferation in the types and the number of connected devices such as sensors and smart wearable devices.

Services provided by a network slice could belong to the MNO or a 3rd party service provider. In the case where a network slice is allocated to a 3rd party, the 3rd party may require the ability of authenticating and/or authorizing the UEs that want to gain access to the slice.

The potential security threatens related to the authentication and authorization of network slices are:

- User's privacy information used in the network slice selection procedure may be intercepted or eavesdropped.
- If there is no authentication and/or authorization for using network slices, unauthorized users may connect to network slices and consume resources
- If there is no proper security mechanism for the authorization of network selection, this will open up for different types of attacks (e.g. impersonation and DoS).

In order to support authentication and authorization of network slices, the following security requirements shall be considered:

- UEs can access network slices only after they have been authenticated and/or authorized.
- Network slice authentication and authorization may be executed by the MNO or a 3rd party. UEs should have a primary authentication by MNO when they have access to a slice.
- Users' privacy shall be protected during network slice authentication and authorization.

8.1.4 Security on virtualization

Virtualization support is needed for network slicing but brings some threats:

- Lack of logical and physical isolation between distinct virtualized network function (VNF) hosted by the same hypervisor.
- DoS effects, e.g. starvation of resources allocated to VNFs or network slices.
- Integrity of hypervisor and hosted VNFs.

Then, the platform which slice run on should assure a level of virtualization security like isolation.

8.1.5 Security on management of slicing

Attackers may illegally obtain capabilities to manage slices or on-going services and launch attacks to slices (e.g. terminate a slice or compromise a critical network function). The 3GPP systems provide capabilities for authorized third parties to create, manage a network slice configuration (e.g. scale slices) via suitable application program interfaces (APIs). These interfaces can be utilized to launch attacks by unauthorized third parties. In order to support authentication and authorization of network slices, the following security requirements shall be considered:

- The capabilities to manage network slices should be under control of authorized operators
- The APIs should be accessed by authorized third parties.

8.2 Key Technologies

8.2.1 Network slice security architecture

Network slice security architecture is shown in the Figure 8-1. The core network components related to security are:

- Security Anchor Function (SEAF): An authentication function in the core network that interacts with the AUSF and the UE and receives from the AUSF the intermediate key that was established as a result of the UE authentication process. The SEAF also interacts with the AMF, e.g. during initial attach. In the roaming case, an SEAF resides in the visited network.
- Authentication Server Function (AUSF): An authentication function that interacts with the UDM and terminates requests from the SEAF. The AUSF retrieves subscriber profile information of UEs. AUSF is responsible for performing authentication functionality, e.g. EAP-AKA and EAP-AKA'.
- Unified Data Management (UDM): The UDM stores the service area restrictions of a UE as part of the UE's subscription data. This function stores the long-term security credentials used in authentication and executes any cryptographic algorithms that use the long-term security credentials as input. It also stores the (security-related part of the) subscriber profile.
- Policy Control function (PCF): Provide security policies that describe what security control shall be applied to an UE that tries to establish a UP session in a network slice.

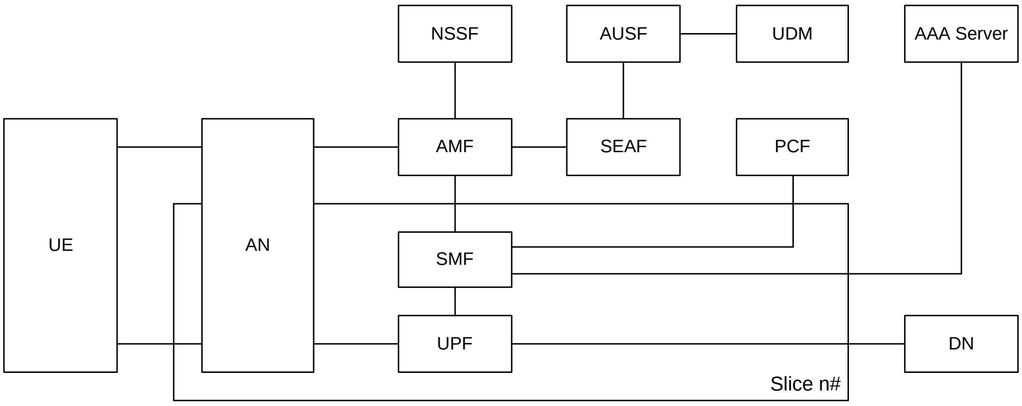


Figure 8-1: Network slice security architecture

8.2.2 Network slice security procedure

A general network slice security procedure is shown in Figure 8-2, which includes the following steps:

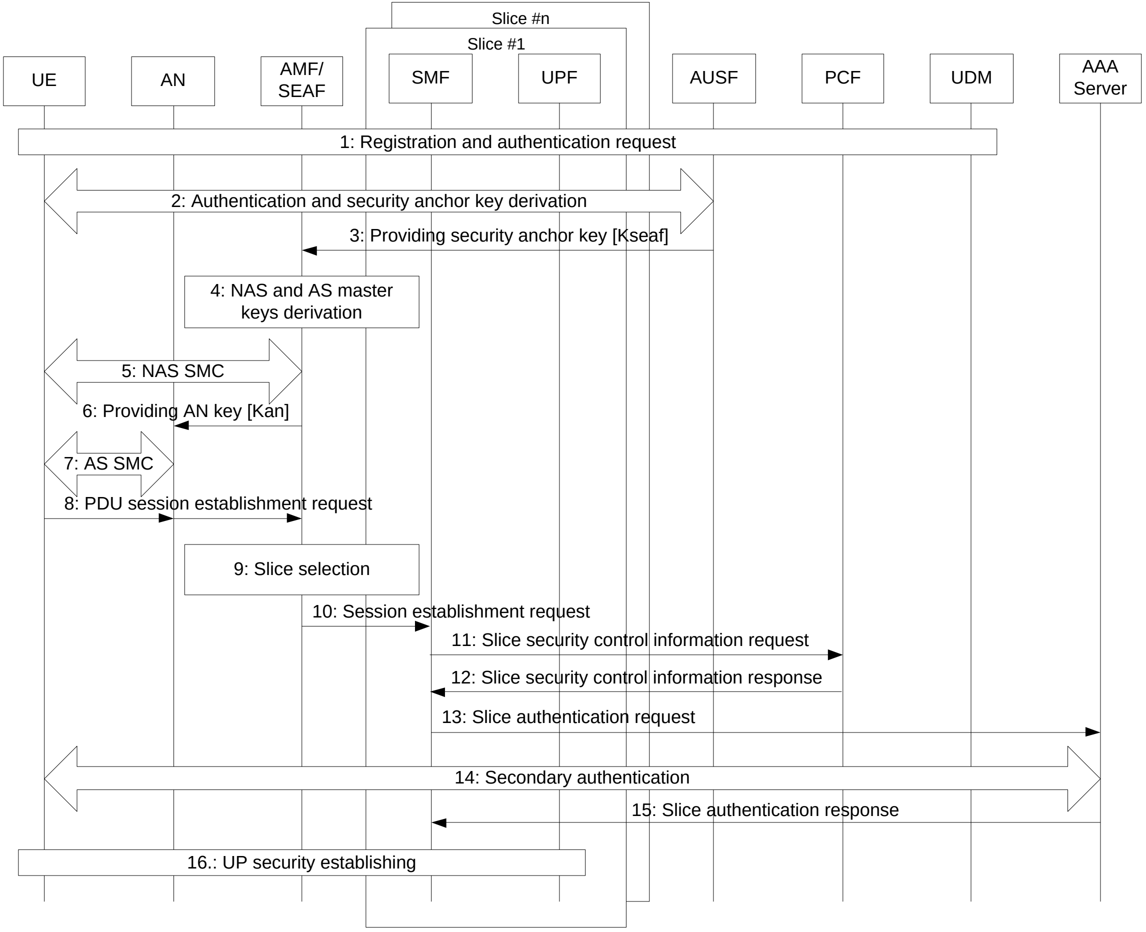

Slice #n
Slice #1
UE
AN
AMF/ SEAF
SMF
UPF
AUSF
PCF
UDM
AAA Server
1: Registration and authentication request
2: Authentication and security anchor key derivation
3: Providing security anchor key [Kseaf]
4: NAS and AS master keys derivation
5: NAS SMC
6: Providing AN key [Kan]
7: AS SMC
8: PDU session establishment request
9: Slice selection
10: Session establishment request
11: Slice security control information request
12: Slice security control information response
13: Slice authentication request
14: Secondary authentication
15: Slice authentication response
16.: UP security establishing

Figure 8-2: Network slice security architecture

1. The UE sends a registration request to the AN. The AN selects an AMF and then forwards the registration request to the AMF. The AMF selects an AUSF and then sends an authentication request to the AUSF. The AUSF retrieves an authentication vector from the UDM.
2. The AUSF and UE perform mutual authentication. The exact details of the flows here depend on the authentication method being used. A successful authentication will result in a security anchor key (Kseaf) at UE and AUSF.

3. The AUSF provides the security anchor key (Kseaf) to the AMF.
4. The AMF (SEAF) derives the NAS security master key (Knas) and AN security master key (Kan).
5. AMF and UE perform NAS Security Mode Command (SMC) procedure in order to establish NAS security.
6. The AMF provides the AN security master key (Kan) to the AN.
7. AN derives RRC master key (Krrc) and UP master key (Kup). AN and UE perform RRC SMC procedure in order to establish RRC security.
8. The UE sends a PDU session establishment request to the AMF.
9. The AMF selects a network slice instance based on the UE's PDU session request.
10. The AMF interacts with the NSSF for selecting a network slice for the UE, The AMF sends a session establishment request to the SMF in the selected slice instance.
11. The SMF sends a slice security control policy request to the PCF.
12. The PCF retrieves the slice security control policy applicable to the UE and returns it to the SMF.
13. The SMF checks the slice security control policy. In the case where the UE needs to be authenticated in the slice instance by a 3rd party, the SMF sends an authentication request to an AAA Server. Otherwise goes to step 16.
14. The AAA Server and UE perform a secondary authentication. The exact details of the flows here depend on the secondary authentication method being used.
15. The AAA Server acknowledges the secondary authentication result to the SMF.
16. In case UP security is not established yet, the AN and UP perform a UP SMC procedure to establish UP security between them. After that a PDU session is established between UE and the UPF.

9 Summary

This paper discussed the concept, scenarios and requirements of E2E network slicing, and outlined the objective and a system framework for 5G network with E2E network slicing function. To meet the requirements of network slicing on the core network, the RAN, the transport network and the devices, key technologies and related procedures, as well as potential security solutions were proposed respectively.

However, implementing an E2E network slicing is a challenging goal. It needs well-developed enabling technologies, global standards and mature ecosystem. It is important to establish common understanding and tight cross-industry collaboration among operators, vendors, and vertical industries and to help create a network slicing based 5G ecosystem.

Reference

[1] NGMN Alliance: "Description of Network Slicing Concept", Version 1.0, January 13, 2016

[2] 3GPP TR 22.864: "Feasibility Study on New Services and Markets Technology Enablers - Network Operation"

[3] 3GPP TS 22.261: "Service requirements for the 5G system: stage I"

[4] 3GPP TR 23.799: "Study on Architecture for Next Generation System"

[5] 3GPP TS 23.501: "System Architecture for the 5G System"

[6] 3GPP TR 38.801: "Study on New Radio Access Technology: Radio access architecture and interfaces"

[7] 3GPP TR 38.804: "Study on New Radio Access Technology: Radio interface protocol aspects"

[8] White Paper: "5G Network Slicing for Vertical Industries", Global mobile Suppliers Association, Sept., 2017.

[9] CMCC report: "5G C-RAN Wireless Cloud Network Technical Report", Sept., 2017.

[10] White Paper: "5G Performance-Guaranteed Network Slicing Service", CMCC, Huawei, March, 2017

[11] Qian (Clara) Li, Geng Wu, Apostolos (Tolis) Papathanassiou, Udayan Mukherjee, "An end-to-end network slicing framework for 5G wireless communication systems." https://arxiv.org/abs/1608.00572.

[12] 3GPP TR 38.300: "NR and NG-RAN Overall Description; Stage 2"

Glossary

· **AF**:	Application Function
· **AMF:**	Core Access and Mobility Management Function
· **AUSF:**	Authentication Server Function
· **CN:**	Core Network
· **CU/DU**:	Centralized Unit/Distributed Unit
· **DSM**:	Domain Slice Manager
· **DSS**:	Domain Slice Support System
· **E2E:**	End to End
· **FlexE:**	Flex Ethernet
· **MNO**:	Mobile Network Operator
· **NEF:**	Network Exposure Function
· **NF**:	Network Function
· **NRF:**	NF Repository Function
· **NSI:**	Network Slice Instance
· **NSM:**	Network Slice Management
· **NSMF**:	Network Slice Management Function
· **NSSAI**:	Network Slice Selection Assistance Information
· **NSSF**:	Network Slice Selection Function
· **NSSMF:**	Network Slice Subnet Management Function
· **PCF:**	Policy Control Function
· **O&M:**	Operation and Maintenance

· **OS**:	Operation Systems
· **QoE**:	Quality of Experience
· **QoS** :	Quality of Service
· **RT/NRT**:	Real Time/non Real Time
· **SLA**:	Service Level Agreement
· **SMF:**	Session Management Function
· **SDSF:**	Structured Data Storage network function
· **SSS**:	Slice Support System
· **TSN:**	Time Sensitive Network
· **TTM:**	Time to Market
· **UDSF:**	Unstructured Data Storage network function
· **UDM:**	Unified Data Management
· **UPF:**	User plane Function
· **VM**:	Virtual Machine

Acknowledge

Grateful thanks to the following contributors for their wonderful work on this white paper:

Editor:	Chih-Lin I , PAN Chengkang
Contributors:	
CMCC:	Chih-Lin I, CUI Chunfeng, PAN Chengkang, SUN Junshuai, WANG Yingying, CHEN Yami, WANG Ke, SUN Xiaowen, CHEN Wei, ZUO Min
CATT:	LI Guoqing, ZHOU Wei
Huawei:	ZHOU Liping, YANG Xudong, ZOU Lan, ZHU Fenqin, YU Fang, HAN Feng, LOU Chong, JIA Xiaoqian, LI Yan, NI Hui
Intel:	RUAN Leifeng, Hiremath Chetan, ZHAO William, YAN Feng, ZHANG Xu
Fiberhome:	WEI Xueqin, HAN Zhen, CHEN Yunchang
Xidian:	ZHAO Liqiang, ZHANG Yaoyuan

未来移动通信论坛
FuTURE MOBILE COMMUNICATION FORUM

WIRELESS WORLD
RESEARCH FORUM