GTI 5G Network Slicing White Paper

GTI

http://www.gtigroup.org

5G Network Slicing White Paper

V 1.0

Version	V1.0
Deliverable Type	□Procedural Document √Working Document
Confidential Level	□ Open to GTI Operator Members □Open to GTI Partners √Open to Public
Program Name	5G eMBB
Working Group	
Project Name	Architecture
Source members	ZTE, Nokia
Support members	
Editor	
Last Edit Date	12-02-2018
Approval Date	DD-MM-2018

Confidentiality: The GTI documents may contain information that is confidential and access to the documents is restricted to the persons listed in the Confidential Level. This document may not be used, disclosed or reproduced, in whole or in part, without the prior written authorisation of GTI, and those so authorised may only use this document for the purpose consistent with the authorisation. GTI disclaims any liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice.

Document History

Date	Meeting #	Revision Contents	Old	New
2018/02/12		The 1st version		V1.0

Executive Summary

The 5G network system can build on-demand network slices to satisfy different service requirements. Network slicing enables the operator to create networks customized to provide optimized solutions for different market scenarios which demand diverse requirements, e.g. in the areas of functionality, performance and isolation. For each scenario, the 5G network can provide appropriate network functionalists, to achieve the goal of building on-demand network and get better performance. Multiple 5G slices can concurrently operate on the same infrastructure.

Table of Contents

1. Introduction

1.1. Main challenges in the network and problems to be solved

Increasingly, more and more vertical services would be supported by 5G. There are identified three types of use cases: MBB, Massive MTC and URLLC. These use cases do not exclude any new types of service in future. The performance requirements for these use cases are quite different. For MBB (Mobile Broad band), the scalable control plane, high performance user plane and high speed of mobility is required. For Massive MTC, it needs optimized handling for small data transmission, and the control plan is relatively simpler compared to MBB. In addition, mMTC demands extreme power saving mechanism. For URLLC, it needs highest reliability on control plane and user plane, meanwhile its user plane is required to locate as closely as possible to the edge in order to achieve lowest latency.

On the one hand, traditional mobile devices, such as smartphones and tablets, and related services require support for high mobility, high aggregate traffic capacities, and stringent delays. In contrast, MTC requires support for typically stationary devices with low per-device throughput, relatively loose delays and high geographic device densities.

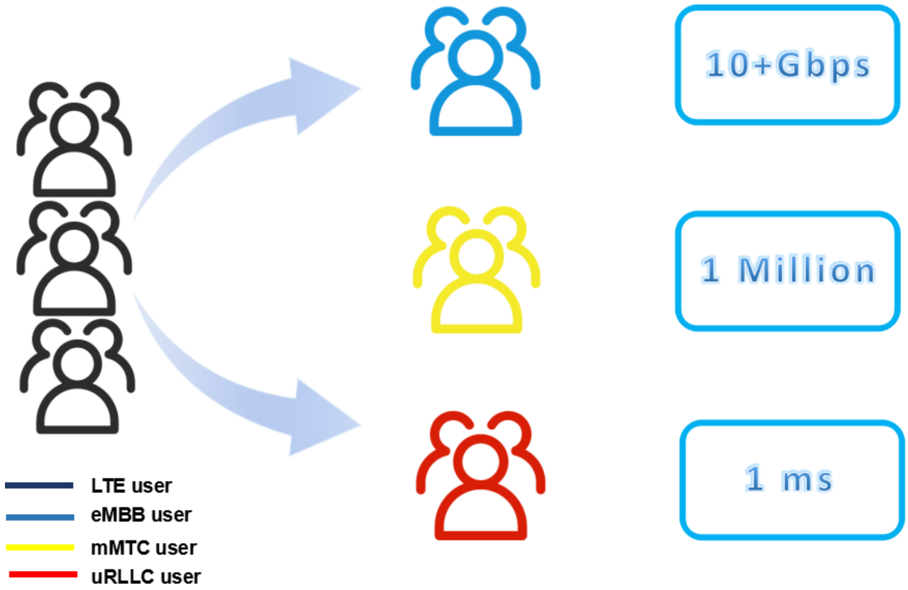


Figure 1-1 main challenges

1.2. Network Slicing overview

Network slice is composed of a set of network functions (e.g., potentially from different vendors) , the resources to run these network functions as well as policies and configurations and specific RAT settings that are combined together for the specific use case or business model.

End-to-end network slicing, including Access, Core and transport network, with physical resources dedicated to either this slice or being shared with other slices. The slices shall be isolated against each other, but sharing a common infrastructure layer for efficiency reasons.

The network slicing concept consists of 3 layers: 1) Service Instance Layer, 2) Network Slice Instance Layer, and 3) Resource layer.

The Service Instance Layer represents the services (end-user service or business services) which are to be supported. Each service is represented by a Service Instance. Typically services can be provided by the network operator or by 3rd parties. In line with this, a Service Instance can either represent an operator service or a 3rd party provided service.

A network operator uses a Network Slice Blueprint to create a Network Slice Instance. A Network Slice Instance provides the network characteristics which are required by a Service Instance. A Network Slice Instance may also be shared across multiple Service Instances provided by the network operator.

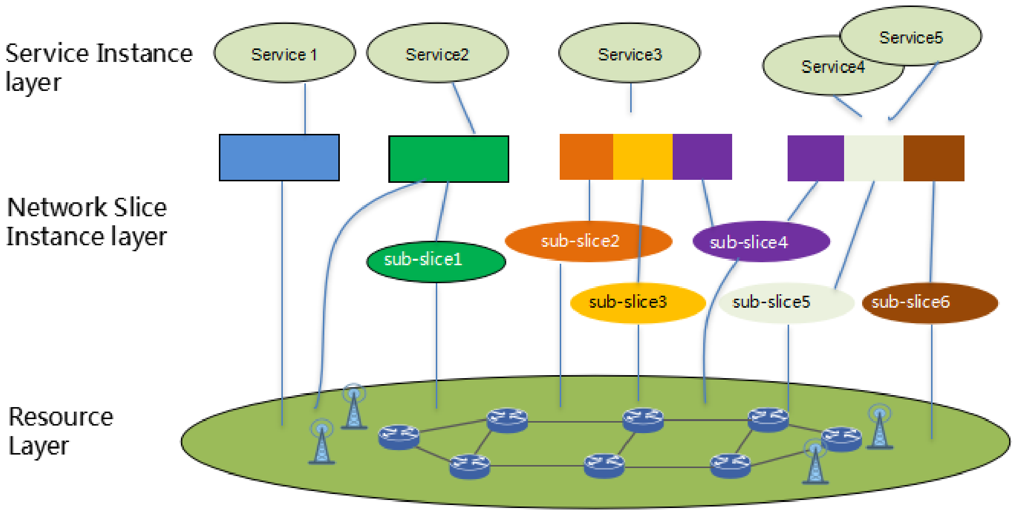


Figure 1-2 network slicing overview

2. Use cases and main requirements for 5G

The services foreseen in the 5G era fall into three typical scenarios: enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC). eMBB focuses on services characterized by high data rates, such as high definition (HD) videos, virtual reality (VR), augmented reality (AR), and fixed mobile convergence (FMC). URLLC focuses on latency-sensitive services, such as self-driving, remote surgery, or drone control. mMTC focuses on services that have high requirements for connection density, such as those typical for smart city and smart agriculture use cases. Each scenario requires a completely different network service and poses requirements that are radically different, sometimes even contradictory.

For example, 5G networks will connect the factory of the future and help create a fully automated and flexible production system. In the healthcare industry, hospitals will be able to arrange remote robotic surgeries as if the surgeon were physically present next to the patient. At the same time, 5G connected healthcare chips will constantly monitor vital signs, prevent conditions from becoming acute, and adapt medication to meet changing conditions. While on the roads, self-driving cars and smart infrastructures enabled by 5G networks will reduce accidents and save millions of lives every year.

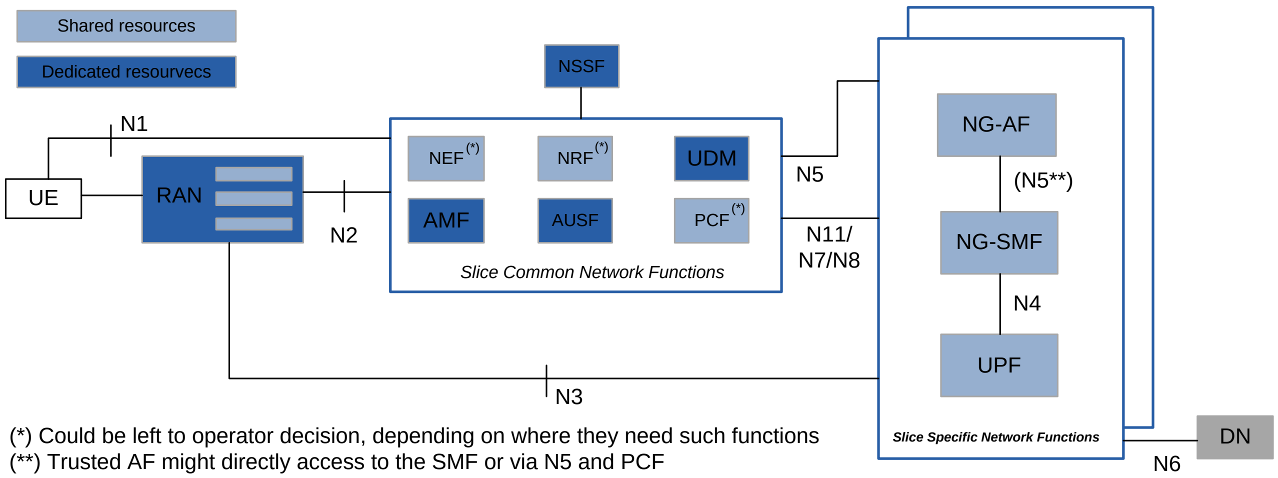
The 5G network is designed to support very diverse and extreme requirements for latency, throughput, capacity and availability. Network slicing offers a solution to meet the requirements of all use cases in a common network infrastructure. The same network infrastructure can support, for example, smartphones, tablets, virtual reality connections, personal health devices, critical remote control or automotive connectivity. With network slicing, different end-to-end logical networks with isolated properties are provided and operated independently. These enable operators to support different use cases, with devices able to connect to multiple slices simultaneously.

Figure 2-1 network slices for s a variety of use cases

3. Solutions for 5G network slicing

3.1. Network Slicing Architecture

Enabling network slicing in 5G requires native support from the overall system architecture. Which contains access slices (both radio access and fixed access), core network (CN) slices and the selection function that connects these slices into a complete network slice comprised of both the access network and the CN. The selection function routes communications to an appropriate CN slice that is tailored to provide specific services. The criteria of defining the access slices and CN slices include the need to meet different service/applications requirements and to meet different communication requirements. Each CN slice is built from a set of network functions (NFs). An important factor in slicing is that some NFs can be used across multiple slices, while other NFs are tailored to a specific slice.



(*) Could be left to operator decision, depending on where they need such functions
(**) Trusted AF might directly access to the SMF or via N5 and PCF

Figure 3-1 5G Network Architecture for Network Slicing

3.2. Network Slice Instance Selection and Association

Slice selection refers to the mechanisms or set of mechanisms to identify the NSIs (Network Slice Instance) of a UE. The subscription of the UE to NSIs can be determined via a slice ID (S-NSSAI) in the subscription database. The S-NSSAI is used to identify a slice and, thus, to assist the 5G network in selecting a particular NSI.

When a UE registers with a PLMN, the UE shall provide the network in RRC and NAS layer a Requested NSSAI containing the S-NSSAI(s) corresponding to the slice(s) to which the UE wishes to register or requires, in addition to the 5G-S-TMSI if one was assigned to the UE.

The RAN shall route the NAS signalling between this UE and an AMF selected using the Requested NSSAI obtained during RRC Connection Establishment. If the RAN is unable to select an AMF based on the Requested NSSAI, it routes the NAS signalling to an AMF from a set of default AMFs.

The AMF verifies whether the S-NSSAI(s) in the Requested NSSAI are permitted based on the Subscribed S-NSSAIs and check whether it can serve the UE. If the AMF can't serve the UE or if the UE context in the AMF does not yet include an Allowed NSSAI, the AMF queries the NSSF, with Requested NSSAI, the Subscribed S-NSSAIs, PLMN ID of the SUPI, location information, and possibly access technology being used by the UE. The NSSF selects the NSIs to serve the UE and determines the NRF(s) to be used to select NFs/services within the selected NSI(s).

When establishing a PDU session associated to an S-NSSAI and a DNN, a UE that is registered in a PLMN and has obtained an Allowed NSSAI, shall indicate in the PDU Session Establishment procedure the S-NSSAI and, if available, the DNN the PDU Session is related to.

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

The AMF queries the NRF to select an SMF in a NSI based on S-NSSAI, DNN and other information e.g. UE subscription and local operator policies, when the UE triggers the establishment of a PDU Session. The selected SMF establishes a PDU Session based on S-NSSAI and DNN.

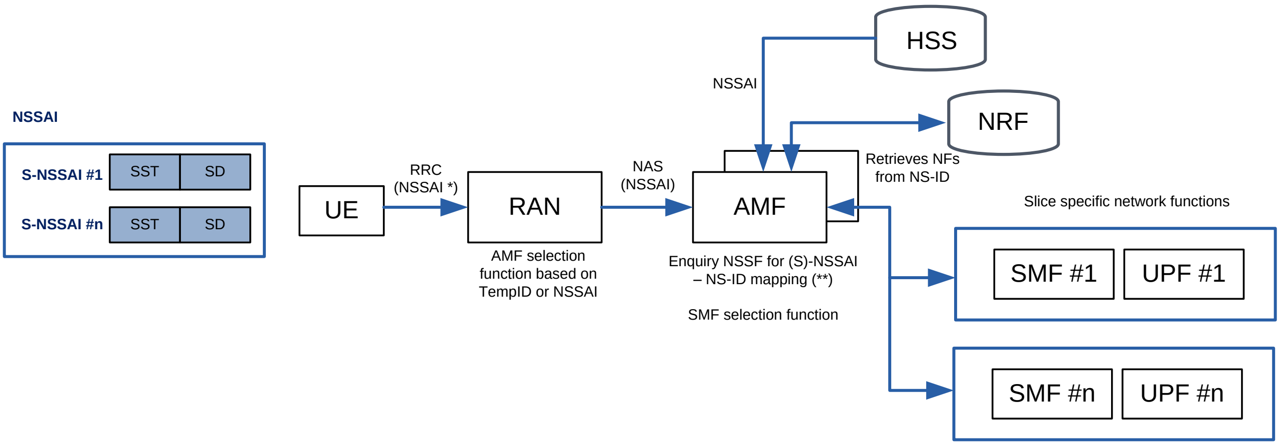


Figure 3-2 Network Slice selection

3.3. Network Slicing resource management

The overall architecture supports the isolation of NSIs, including resource isolation, O&M isolation, and security isolation. NSIs can be either physically or logically isolated at different levels.

Resource isolation enables specialized customization and avoids one slice affecting another slice. E.g. RAN needs to provide and enforce differentiation, and maintain isolation between slices where resources are constrained including RF resource, backhaul transport resource and computing resource.

Hardware/software resource isolation is up to implementation. Each slice may be assigned with either shared or dedicated radio resource up to RRM implementation and SLA (Service Level Agreement). The logical isolation of resources may require resource multiplexing. The amount of allocated resources can be scaled up or down for higher utilization efficiency depending on the traffic load of each NSI. Co-existence of shared and dedicated functions allows time- and frequency-domain resource isolation without sacrificing resource efficiency.

To enable differentiated handling of traffic for network slices with different SLA:

- NG-RAN is configured with 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.

To guarantee security for the network services provided by an NSI, it requires embedding the security mechanism and security provisioning entity (e.g. security anchors and security functions) into the logical network architecture of the NSI.

A terminal should be authenticated and authorized to access a specific NSI. The communication between the terminal and the allocated NSIs should be protected against attacks. In addition, terminals may require different levels of security protection associated with different NSIs.

3.4. Network Slicing mobility support

3.4.1. Cell selection and reselection

Like in LTE, UE behavior in IDLE state includes cell selection and re-selection. As in LTE, UE can prioritize a frequency based on service, Cell broadcasts (e.g. in minimum SI) the services supported by it, and network slices can use these technologies.

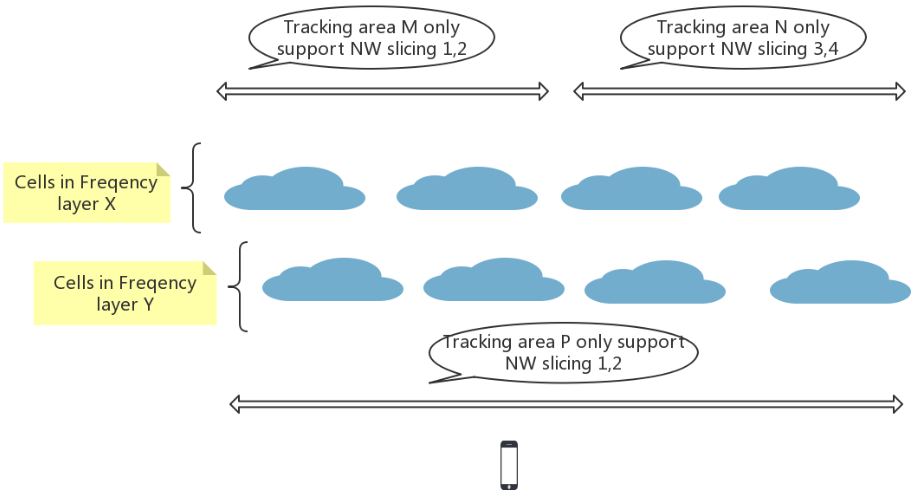


Figure 3-3 cell selection and reselection for network slicing

Therefore the issue is whether Slice based prioritization is also used for NR and what is the impact of this function on NR idle behavior.

It was agreed that for intra-freq cell reselection the UE try to always camp on the best cell. Additional functionality for RACH resource isolation/differentiated treatment will not be supported for slicing for Rel-15.

3.4.2. Network slicing roaming support

Considering the support of network slicing in roaming scenarios, there are two scenarios:

Home routed roaming case: The UE's traffics are transferred via the UP functions in both VPLMN and HPLMN. The NFs in the VPLMN cooperate with the NFs in the HPLMN to provide the end-to-end services for the roaming UE.

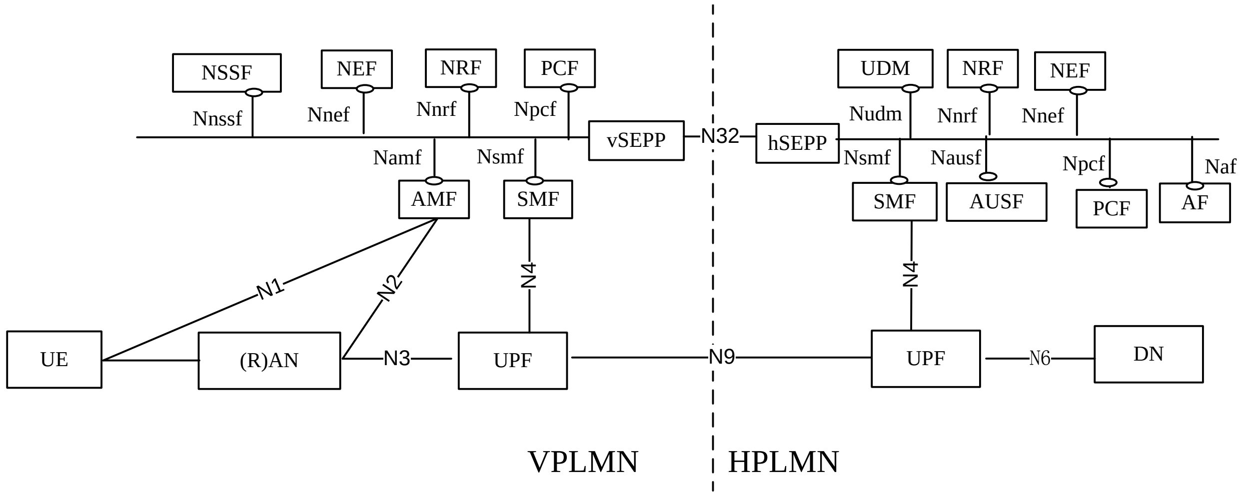


Figure 3-4 network slicing roaming architecture-home routed scenario

Local breakout roaming case: The UE's traffics are transferred via UP functions within VPLMN, and the HPLMN may provide policy control function for the roaming UE.

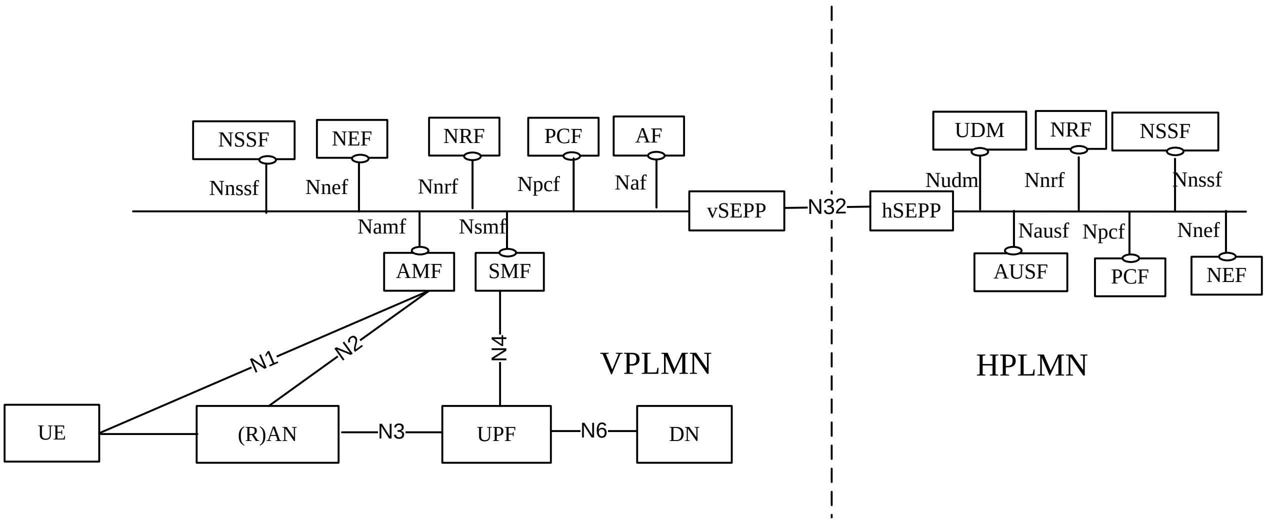


Figure 3-5 network slicing roaming architecture-local breakout scenario

If the VPLMN and HPLMN have an SLA to support non-standard S-NSSAI values in the VPLMN, the NSSF of the VPLMN maps the Subscribed S-NSSAIs values to the respective S-NSSAI values to be used in the VPLMN. The S-NSSAI values to be used in the VPLMN are determined by the NSSF of the VPLMN based on the SLA.

Depending on operator's policy and the configuration in the AMF, the AMF may be allowed to decide the S-NSSAI values to be used in the VPLMN and the mapping to the Subscribed S-NSSAIs. The NSSF in the VPLMN determines the Allowed NSSAI without interacting with the HPLMN. When the UE constructs Requested NSSAI, it may also provide the mapping of each S-NSSAI of the Requested NSSAI to the S-NSSAIs of the Configured NSSAI for the HPLMN. Upon successful completion of a UE's Registration procedure, the UE obtains an Allowed NSSAI, which includes one or more S-NSSAIs, from the AMF, possibly associated with mapping of Allowed NSSAI to Configured NSSAI for the HPLMN.

In a PDU Session Establishment procedures, the UE includes a Subscribed S-NSSAI based on the NSSP (HPLMN S-NSSAI), and the related S-NSSAI from the Allowed NSSAI (VPLMN S-NSSAI). For home routed case, the V-SMF send the PDU Session Establishment Request message to the H-SMF along with the HPLMN S-NSSAI. When a PDU Session is established, the CN provides to the AN the VPLMN S-NSSAI corresponding to this PDU Session.

3.4.3. Network slicing handover support

Some slices may be available only in part of the network. The RAN and the CN are responsible to handle a service request for a slice that may or may not be available in a given area. Admission or rejection of access to a slice may depend by factors such as support for the slice, availability of resources, support of the requested service by other slices.

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.

Slice-aware supported in the cells of its neighbouring gNBs may be beneficial for mobility in connected mode. It is assumed that the slice configuration does not change within the UE’s registration area.

Solutions for how slice availability may be handled during mobility include: Neighbours may exchange slice availability on the interface connecting two nodes, e.g. Xn interface between gNBs, and the core network could provide the RAN a mobility restriction list. This list may include those TAs which support or do not support the slices for the UE.

3.5. Network Slicing QoS support

3.5.1. QOS for network slicing

In LTE, the bearer in CN and RB in RAN are of one to one mapping. One PDN connection can contain multiple EPS bearers, and one EPS bearer can contain multiple IP flows. In the RAN side, the EPS bearer is mapped to the radio bearer with one-to-one mapping relationship.

The 5G QoS framework can provide the wide range of existing and future emerging use cases/services. The 5G QoS model is based on QoS Flows. NG-RAN and 5GC ensure quality of service (e.g. reliability and target delay) by mapping packets to appropriate QoS Flows and DRBs. The QoS Flow is the finest granularity of QoS differentiation in the PDU Session.

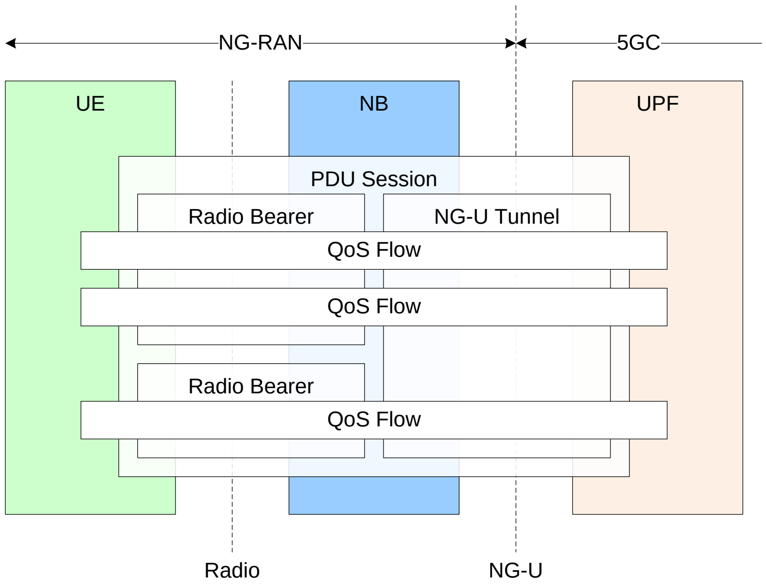


Figure 3-6 QoS architecture

The PDU sessions from different CN slices are just different service flows with different QoS requirements and implicitly mapping the CN slice to RAN via QoS characteristic and session characteristic (e.g. which slice's UP entity the session connects to). That is beneficial to realize some specific slices, such as V2X, which have diverse traffic type (entertainment, traffic safety) with different QoS requirements.

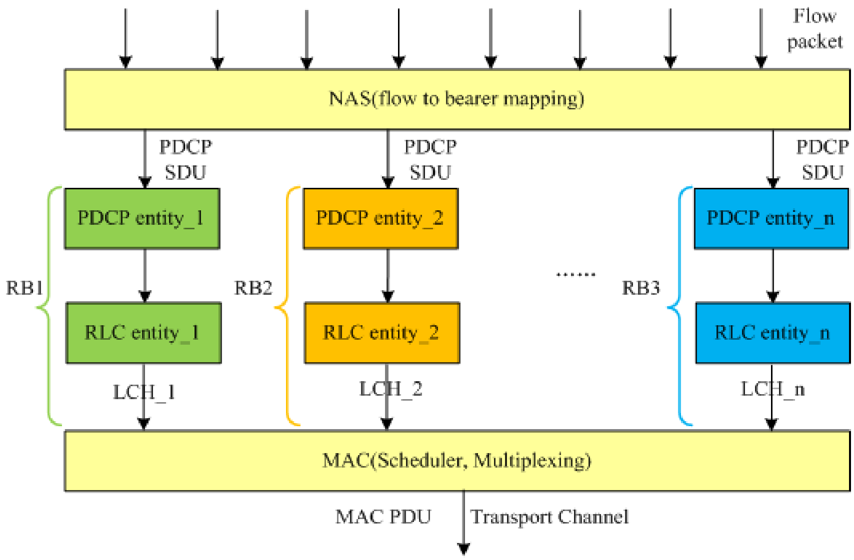


Figure 3-7 QOS mapping for network slicing

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 SLA represented by the network slice, while still being able to apply (for example) differentiated QoS within the slice. And network slicing related parameters are included in Initial Context Setup signaling and establishment/modification/release of a PDU session signaling.

3.5.2. Support for UE associating with multiple network slices simultaneously

A single UE can simultaneously be served by one or more Network Slice instances via a 5G-AN. A single UE may be served by at most eight Network Slices at a time. The AMF instance serving the UE logically belongs to each of the Network Slice instances serving the UE, i.e. this AMF instance is common to the

Network Slice instances serving a UE. If network slices are isolated with each other, the UE should not include the isolated S-NSSAIs in the Request NSSAI and will not access such network slices simultaneously.

4. Abbreviations

AF	Application Function
AMF	Access and Mobility Management Function
AUSF	Authentication Server Function
NEF	Network Exposure Function
NRF	Network Repository Function
NSSF	Network Slice Selection Function
PCF	Policy Control Function
SMF	Session Management Function
UDM	Unified Data Management
UPF	User Plane Function
URLLC	Ultra-Reliable and Low Latency Communications