End-to-End Network Slicing in the 5G System

EFORT
http://www.efort.com

1. Introduction

A key benefit of 5G is the ability to deliver dedicated and differentiated Quality of Service (QoS) to different groups of users and applications over the same network infrastructure, using dynamic on demand network slicing. With use cases broadly categorized into eMBB (enhanced Mobile Broadband), uRLLC (ultra-Reliable and Low Latency Communications) and mMTC (massive Machine Type Communications), communication service providers can monetize 5G by means of network-as-services as well as offer differentiated QoS/SLA required by enterprises and industry.
The service provider for a network slice instance will need all the advanced SLA monitoring capabilities to provide real-time SLA status, monthly SLA reports and customer SLA portal. It is also about predicting potential SLA violations before they are actually occurring and to react to the issue so that the end-user is not affected. It is important to consider “SLA management by design“, meaning that SLA management should be part of the design of a network slice instance.
The goal of this tutorial is to explain what a network slice instance is, what the SLA associated to a network slice instance means and the tools needed to monitor the performance KPIs/SLAs.

2. Network slice definition

5G is a network of networks: Enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) (also referred to as mission-critical communications). These are called network slices.

A succinct characterization of these network slices can be put forward as follows: eMBB supports stable connections with very high peak data rates; mMTC supports a massive number of Internet of Things (IoT) devices, which are only sporadically active and send small data payloads; uRLLC supports low-latency transmissions of small payloads with very high reliability from a limited set of terminals, which are active according to patterns typically specified by outside events, such as alarms.

5G service provider offers a network slice instance as a network-as-a-service for the enterprise market with dedicated and differentiated QoS. The network slice instance is tailored based on the specific requirements adhered to a Service Level Agreement (SLA) agreed between Network Slice Customer and Network Slice Provider.

A service provider becomes a telco cloud provider to offer the network-as-a-service with differentiated SLA.

A large number of KPIs related to each network slice instance are measured by a network data analytic function called NWDAF.

5G is complex to deploy. 1st step : NSA mode where 5G Radio interfaces to 4G core network to handle eMBB use case mainly. 2nd step ; SA mode where 5G Radio interfaces to 5G core network to handle all use cases and network slicing.

A network slice could span across multiple network domains used by a service provider (e.g. access network, core network and transport network). A network slice consists of dedicated and/or shared resources, e.g. in terms of functionality, processing power, storage, and bandwidth. Dedicated resources can be isolated from other network slices.
Network Slice Instance: E2E instance of a network slice which is instantiated in service provider's network for specific customer and specific service type.
Transport Slice: It is also called Transport Sub-Slice. A set of connections between various network functions with deterministic SLAs.
RAN Slice: It is also called RAN Sub-Slice. The context and personality created on RAN network functions for each E2E network slice.
Core Network Slice: It is also called Core Network Sub-Slice. The context and personality created on Core network (CN) functions for each e2e network slice.

management aspects of network slicing, consists of four phases : Preparation, Commissioning, Operation, Decommissioning.

In the preparation phase the Network Slice instance does not exist. The preparation phase includes network slice design, network slice capacity planning, on-boarding and evaluation of the network functions, preparing the network environment and other necessary preparations required to be done before the creation of an Network Slice instance.

In the commissioning phase, network slice instantiation and configuration are performed.

In the operation phase, network slice is activated and the status of the network slice is observed, and analytics guarantee that SLAs are met.

The decommissioning phase involves the deleting of a slice when the service is no longer in use.

Figure 1 presents three E2E Network Slice Instances (red, green and blue lines). Each E2E network slice instance consists of a RAN Slice, a Transport Network Slice and a Core Network Slice

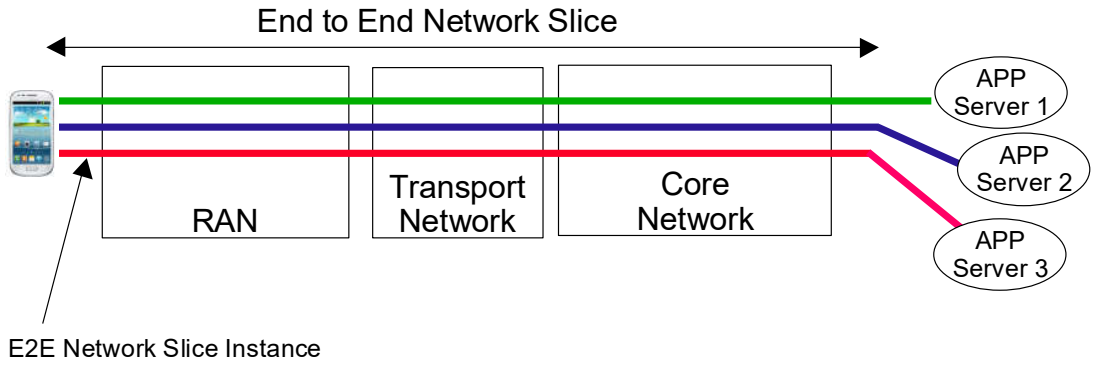

Figure 1 : End-to-End Network Slice Instance

2.1. Important issues

Mobile data networks are connection oriented. The objective is to get a data connection associated with an IP address to access to its applications.
The configuration (QoS, IP address type, etc) of the data connection is associated to the APN (Access Point Name). Typically, an eMBB user needs two data connections : one to access to the Internet (Internet APN) and one to access to the IMS (IMS APN).

In 4G, a UE activates an APN and opens a data connection; an IP address is assigned to the UE. Then, the UE can exchange its service data flows.
In 5G, a UE activates an APN (called DNN, Data Network Name) in a network slice instance (called S-NSSAI, Single Network Slice Selection Assistance Information) and opens a data connection.
The goal of the service provider is to provide the most appropriate data connection which meets the data flow requirements (throughput, latency, mobility, redundancy for reliable delivery, security, etc).
A network slice instance consists of data connections (PDU sessions) with appropriate behavior compliant with user data flow characteristics.
For the mMTC slice, there are many types of data connections that may be established each offering a different behavior. For example, Non-IP data connection, IP data connection, control plane data connection, user plan data connection, etc.
When an mMTC network slice instance is created, based on customer requirements, only the most appropriate data connection type(s) will be supported by this network slice instance.
A network slice instance is a customized network instance for a particular usage. IT IS ALL ABOUT DATA CONNECTIONS with appropriate behavior!

2.2. PDU session within a network slice instance

A PDU session is a data connection established within a network slice instance.
A PDU session transports data flows. Each data flow is characterized by a given QoS.
A broadband user with a smartphone may way to establish two PDU sessions within an eMBB slice instance. One PDU session to access to the Internet applications and one PDU session to access to IMS services (e.g., VoIP) (Figure 2).
The Internet PDU session will handle all the Internet flows that will all share the same best effort QoS.
The IMS PDU session will handle the SIP signaling flow and VoIP media flow each with its own QoS. The SIP QoS is interactive, the VoIP QoS is conversational audio.

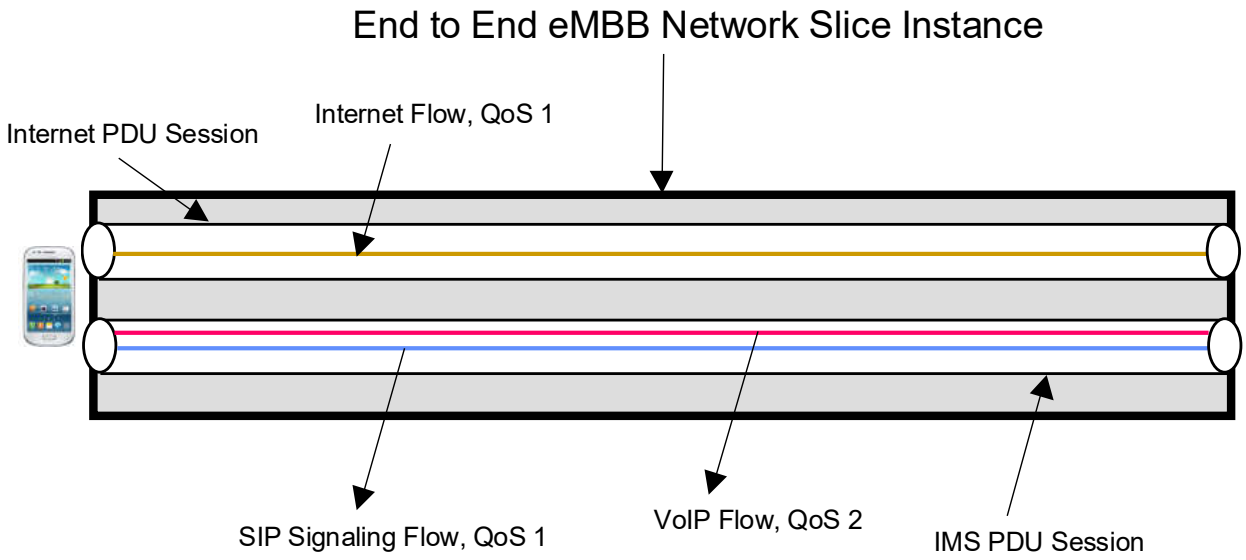

Figure 2 : eMBB Network Slice Instance with two PDU sessions per User Equipment

2.3. eMBB Network Slice Instance example

A car manufacturer wants to promote a car WiFi mobile broadband plan associated to each car being sold.
The car manufacturer gets an eMBB slice instance for running its business of car WiFi mobile broadband plan. This is a B2B2C business model. This slice instance is dimensioned for 1 million devices (cars), and one PDU session per device with 100 Mbit/s maximum bit rate UL and DL per device, with national coverage of the network slice instance and fair use depending on the plan (bronze, silver, gold).
The car manufacturer promotes a car WiFi mobile broadband connection plan when it sells a car.
The car manufacturer may reconfigure anytime its eMBB slice instance if the number of cars which connect to that slice instance exceeds 1 million.
As can be seen, the overall opportunity with network slicing is that it opens up for new types of service offerings and support different enterprise business models, in a flexible way with a high service deployment velocity. It is an enabler to generate more revenues for the service provider.
The eMBB slice instance will have the following characteristics (GSMA in its NG.116 document provides the standardised list of attributes that can characterise a type of network slice in the form of Generic network Slice Template):

- Coverage: National
- Availability: Very high: >99,999%
- Downlink maximum throughput per UE: 100 Mbit/s
- Uplink maximum throughput per UE: 50 Mbit/s
- Maximum number of PDU sessions: 1 000 000
- Maximum number of UEs: 1 000 000
- MMTel support (Telephony service support): Not supported
- Session and Service Continuity support: 3 (support of edge computing to provide low latency for the internet flows)
- Slice QoS parameters (Internet flow QoS):
    - 5G QoS Class Identifier: 8
    - Priority level (ARP): 6
- Device velocity: 150 km/h
- Network functions owned by Network Slice customer: Policy Control Function and Charging Function are part of the car manufacturer ecosystem to manage the fair usage policy associated with the car WiFi mobile broadband plan. AAA Server is also part of the Network slice customer to authenticate the user equipment (i.e., car) before it can access to the network slice instance.

2.4 mMTC Network Slice Instance example

In 2013, a train derailment occured in France simply because some bolts on the railway were not tighten enough. 7 people died and 70 people were injured.
French railway company with the help of a startup designed a connected bolt (Eatwork) which measures its tightening level.
The goal is to replace hundreds of thousands of traditional bolts by connected bolts on the rail infrastructure
The objectives for the french railway company is to get a smart infrastructure slice instance with the appropriate scalability, which enables its bolts to open their PDU session and send data reports when the tightening level of the bolt becomes too low.
Then the maintenance team will need to intervene to tighten again the bolt.

The smart infrastructure slice instance will have the following characteristics:

- Coverage: National
- Availability: Very high: >99,999%

- Delay tolerance: Supported
- Downlink maximum throughput per UE: 10 kbit/s
- Uplink maximum throughput per UE: 5 kbit/s
- MMTel support (Telephony service support): Not supported
- NB-IoT support: Supported
- Maximum number of PDU sessions: 100 000
- Maximum number of UEs: 100 000
- Supported device velocity: 0 km/h: Stationary
- Support of non-IP traffic: Supported

NIDD (Non-IP Data Delivery) enables IoT devices to transmit data without allocation of an IP address to these devices.
Two benefits are cited:

- better security by not using IP in transmission since the device is not accessible via the Internet;
- better efficiency by not requiring header information. There is just “less of everything” available in a connected device: less memory, less processing power, less bandwidth, etc., and of course, less available energy.

The maximum size of a NIDD message is 2K Bytes. NIDD is bidirectional data transfer service. If the device is not reachable and an incoming data message is being received, store and forward applies. A data connection is established in a network slice instance on the control plane instead of user plane for the transfer of NIDD traffic.

2.5 5G Network Slice examples in a 5G System

- The service provider creates a network slice instance for its internal need. One eMBB slice instance accessible by all its mobile broadband residential customers with their smartphone.
- The service provider creates an eMBB slice instance for a light MVNO. It is a B2B2C business model. The light MVNO on its turn sells mobile broadband subscriptions to end customers.
- The service provider creates an eMBB slice instance for a car manufacturer. It is B2B2C business model. The car manufacturer on its turn sells WiFi mobile broadband data plan to car owners.
- The service provider creates two mMTC slice instances for smart grid companies. The smart meters of these companies will create their PDU session with appropriate behavior in their network slice and send data reports periodically.
- The service provider creates two mMTC slice instances for smart grid companies. The smart meters of these companies will create their PDU session with appropriate behavior in their network slice and send data reports periodically.
- The service provider creates two mMTC slice instances for two smart cities for the connectivity of their IoT devices.
- The service provider creates a uRLLC slice instance for a car manufacturer to enable its cars to exchange V2V and v2I traffic.
- A car may have two use cases and may open PDU sessions within two different slice instances at the same time: eMBB slice instance and uRLLC slice instance. The 5G standard allows a UE to open PDU sessions within up to 8 network slice instances.

The 5GS has a common network function called NWDAF (Network Data Analytics Function) which measures KPIs per network slice instance.

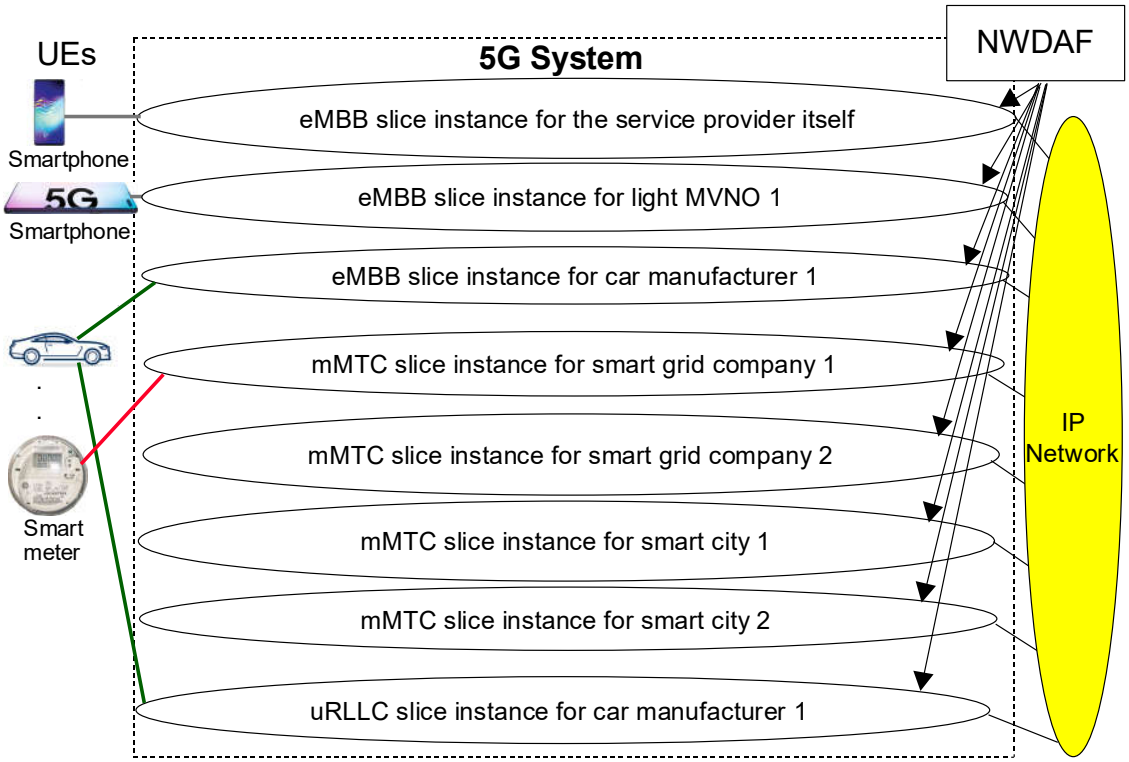

Figure 3: 5G Network Slice examples in a 5G System

3. NWDAF

A Key requirement is to include Network Performance Monitoring and Service Assurance to network slice delivery. This allows maintaining the committed Service Level Agreement (SLA).
During the Network Slice Instance (NSI) performance supervision, the end to end KPIs provided by NWDAF should be used for SLA fulfilment evaluation. According to the evaluation result, NSI modification can be triggered for network slice performance assurance. Service assurance requires its own dedicated slice that consists of all the functions that collect monitoring data needed by NWDAF and contribute to service assurance. VIAVI is an example of vendor providing such assurance slice.

NWDAF collects various types of network and subscriber data, applies an “analysis” to these data, and offers the results to other Network Functions through a service-based interface. NWDAF measures end-to-end KPIs (3GPP TS 28.554) per network slice instance as well as access and core network KPIs (3GPP TS 28.552).

The end-to-end KPI categories are:

- *Accessibility KPIs*
  - Registration success rate of one single network slice instance (RegSR)
  - PDU session Establishment success rate of one network slice (PDUSessionEstSR)
- *Integrity KPIs*
  - Average e2e uplink delay for a network slice (DelayE2EUlNs)
  - Average e2e downlink delay for a network slice (DelayE2EDlNs)
  - Upstream throughput for a single Network Slice Instance (UpstreanThr)
  - Downstream throughput for Single Network Slice Instance (DownstreamThr)
- *Utilization KPIs*

    - Mean number of PDU sessions of Single Network Slice Instance (PDUSesMeanNbr)
    - Virtualised resource utilization of single network slice instance (VirtualResUtilizaiton)
    - PDU session establishment time related to one single network slice (PDUEstTime)
- *Retainability KPIs*
    - QoS flow Retainability (QoSRetain). A measurement that shows how often an end-user abnormally loses a QoS flow during the time the QoS flow is used.
    - DRB Retainability (DRBRetain) : A measurement that shows how often an end-user abnormally loses a DRB (Data Radio Bearer) during the time the DRB is active.
- *Mobility KPIs*
    - NG-RAN Handover Success Rate (GRANHOSR). A KPI that shows how often a handover within NR-RAN is successful, regardless of the reason handover was made
    - Mean Time of Inter-gNB handover of one single network slice (InterGNBHOMeanTime)
    - Successful rate of mobility registration updates of Single Network Slice (MobilityRegUpdateSR).