JULY 2022

A 5G Americas White Paper

COMMERCIALIZING 5G NETWORK SLICING

Contents

Executive Summary

One of the major features of 5G networks is the capability to deliver connectivity across the vastly different quality of service (QoS) requirements required by modern applications. As opposed to the “one size fits all” network paradigm employed by previous generations of mobile networks, 5G networks can address sophisticated and varied challenges from applications like machine-type communication (MMTC), ultra-reliable low latency communication (URLLC), and enhanced mobile broadband (EMBB) delivery.

This strength of 5G networks is managed via a concept called “Network Slicing,” which is the main topic of this paper.

The Network Slicing framework is one of the major additions in the 5G Standalone standards. While it is still in its infancy, it is also one of the critical pillars for enabling operators to create and deliver new services. Network slicing is by no means a simple feature, as it touches nodes in all the wireless network domains – device, core, and radio - and one could possibly compare its complexity with the introduction of IP Multimedia System (IMS) and voice in packet switched networks more than a decade ago.

In a technical sense, network slicing creates independent logical mobile networks - network slices - on the same physical mobile network infrastructure. Each network slice is an isolated end-to-end network tailored to fulfil diverse requirements required by specific applications or customers. Traditional mobile networks offer one ‘data pipe’ for everything and everyone, with very few mechanisms that allow them to treat specific type of traffic different than regular data. Network slicing addresses the need to use the same spectrum and hardware but accommodate different combinations of network attributes for specific scenarios and use-cases.

While a lot of documentation and literature has already been produced aimed at educating engineers and customers about what network slicing is for, how it works, and what it can potentially deliver, this whitepaper tackles a more complex domain. We delve more deeply into the central compromises of business and technology with the goal of launching a first commercial product based on network slicing.

Chapter 1 starts with an introduction of the basics of network slicing to establish – without claiming to be exhaustive - a basic understanding of some of the fundamental concepts needed as well as with an overview of the main standard bodies that support them. Chapter 2 then builds a structure of the main concepts and building blocks for some of the most discussed network slicing-based services, with the goal of enabling a decision-making process that considers as many factors as possible. These include technical capabilities, existing business impact and ecosystem availability, and ultimately add some clarity into the complexity of initial commercialization possibilities. Finally, Chapter 3 highlights some of the future improvements for architectures and platforms with good prospects of adding to and strengthening the network slicing domain.

1. Introduction

This chapter will introduce you to some key concepts relating to network slicing, standards bodies which are developing specifications and capabilities for this technology, and technology architecture frameworks around which network slicing is based. This chapter is focused on a discussion revolving around the first commercialization of network slicing services, while the next chapters will bring to the forefront some notable aspects of current development in this domain, without claiming to be exhaustive.

1.1 Network Slicing 101

The purpose of network slicing in 5G is to give operators the capability of tailoring network resources so they can meet the requirements of a wide range of customers.

In earlier generations of cellular networks like 4G and 3G, network resource allocation (e.g., in RAN or Core networking) was managed by using identifiers like Quality of Service (QOS), Public Land Mobile Network (PLMN), Access Point Name (APN) etc. However, these identifiers have limited granularity, are only useful in managing specific areas of the network, and do not address emerging customer requirements.

5G introduces a network slicing framework integrated into the fundamental architecture, which specifies a slice identifier that can be used across all network elements. 3GPP specifies a network slicing framework for 5G Standalone based on the Network Slice Selection Assistance Information (NSSAI) tag. Figure 1.1 below shows different mechanisms and slicing parameters that can be configured on the (RAN/CN) and Radio Access Technology (RAT), such as Long-Term Evolution (LTE) or New Radio (NR).

Network slices can be created by transforming the physical network into multiple logical networks on top of a shared infrastructure. Logical networks are created and tailored to serve defined business purposes, as they contain specific and customized network resources and configuration. Depending on the configuration and customer requirements, each network slice may belong to a single customer/group or may be shared among multiple customers. The characteristics of network slices may include low latency, high bandwidth, ultra-reliability and/or advanced security features. The combination of network slice characteristics will empower operators to implement specific customer requirements in consumer, enterprise, and new verticals marketplace.

Figure 1.2 shows an example of various slices and profiles that can be associated with a service.

As network slicing attempts to solve some of the more complex service requirements, full scale automation helps efficient deployment and management of network slices. The network slice automation requires new slice orchestration elements, updated or modern OSS/BSS and inventory, as well as an upgrade of the operator specific tools used in managing legacy networks.

Figure 1.1 Network Resourcing parameters in LTE/NR

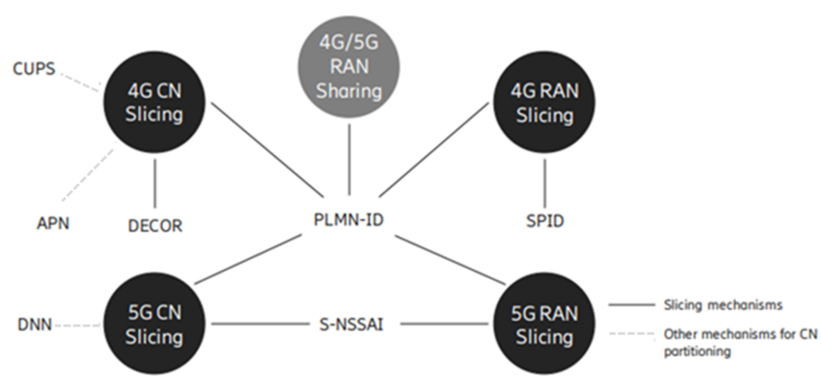
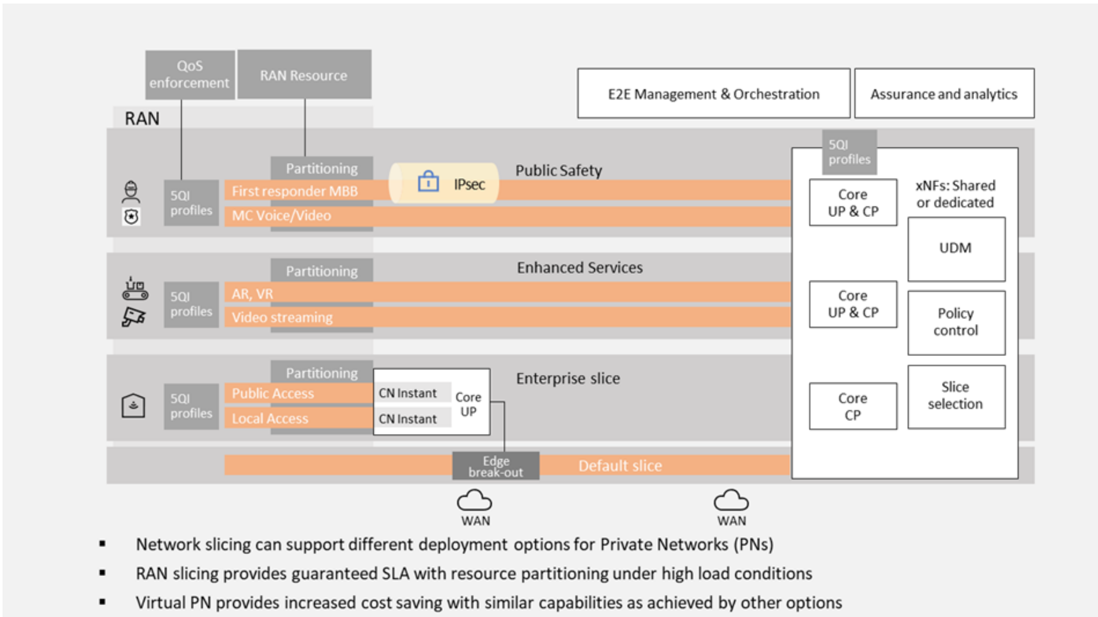


Figure 1.2 Network Slice Profile Examples & E2E View



1.2 Standards and Industry Forum Frameworks for Network Slicing

A number of different standards bodies and technical forums have come to the forefront of the work involved in developing network slicing specifications. These groups represent significant interests from the wireless cellular industry in moving forward network slicing as a technology for 5G networks.

1.2.1 3GPP (Third Generation Partnership Project)

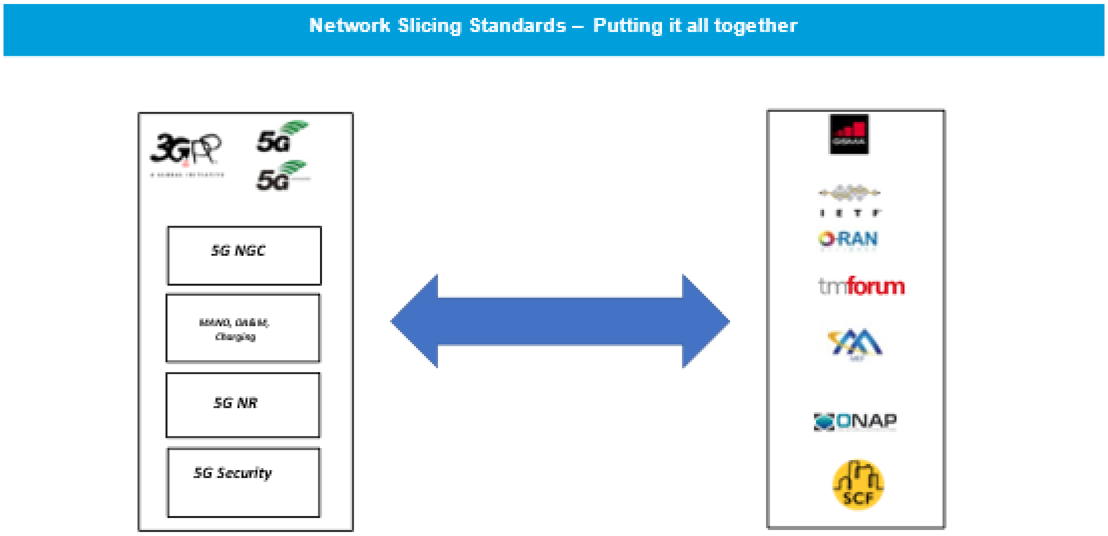
For some time, the 3GPP and other standards forums have discussed the concept of network slicing in mobile networks. Pre-5G, for example, 3GPP developed a concept of overlay 4G networks where devices could connect to Dedicated Core Networks (otherwise known in the 3GPP as (e)DÉCOR) which allowed some level of customization. However, the real effort to support slicing in mobile networks started from 5G 3GPP Release 15 when this functionality was defined as a basic network capability.

The Network Slice Selection Function (NSSF) and other new network functions were introduced to enable the selection of network slices. Additional work in Release 16 also defined access to specific network slices, which are authorized and authenticated through additional User Identifiers (UI), such as additional third-party authentication/authorization before accessing a network slice. Interworking for network slicing between 4G Evolved Packet Cores (EPC) and 5G Cores (5GC) was enhanced in Release 16 as well. At the time of the writing of this white paper, Release 17 development is ongoing. 3GPP continues to enhance network slicing capabilities with an expectation of completion in mid-2022. These Release 17 network slicing enhancements include:

- *Support of network slice related quota on the maximum number of User Equipment (UE)s*
- *Support of network slice related quota on the maximum number of Protocol Data Unit (PDU) Sessions*
- *Limitation of data rate per network slice in (upload) UL and (download) DL per UE*
- *Support for network slice quota event notification in a network slice*
- *Dynamic adjustment to meet the limitation of data rate per network slice in UL and DL*
- *Constraints on simultaneous use of the network slices*
- *Support of 5GC assisted cell selection to access network slices*

Beyond 3GPP, several industry forums and standards bodies have been working on network slicing, with many of these efforts feeding their output into 3GPP work or enabling network slicing separate from the 3GPP. These industry forums and standards bodies include the Internet Engineering Task Force (IETF), GSM Association (GSMA), Metro Ethernet Forum (MEF), O-RAN ALLIANCE, TM Forum (TMF), and Small Cell Forum (SCF).

Figure 1.3 Network Slicing Standards and Industry Forums



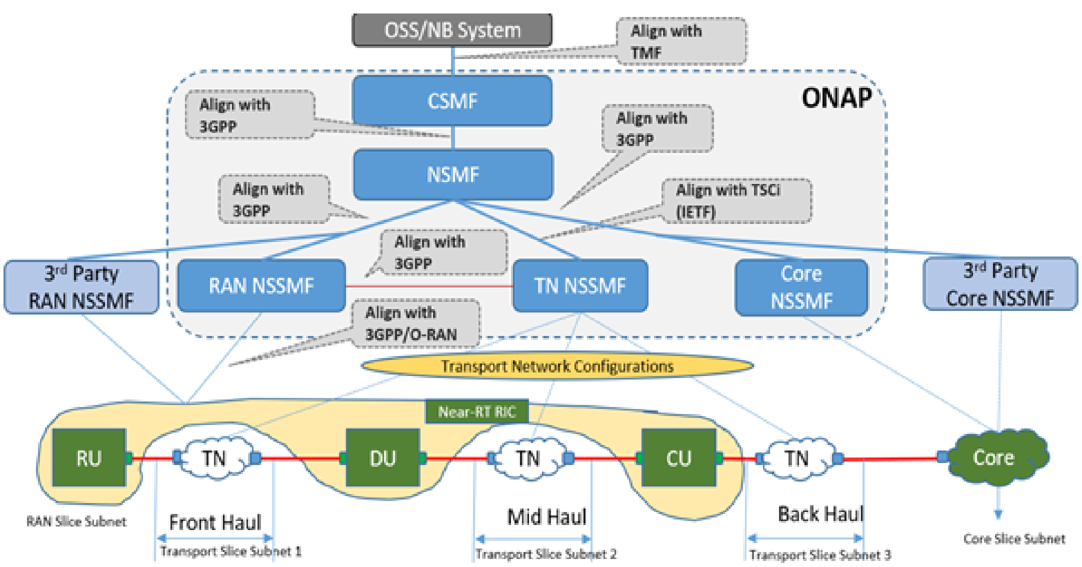
1

1.2.2 Open Network Automation Platform (ONAP)

ONAP is an open-source software platform for orchestrating, managing, and automating network services enabled by physical and virtual network functions. ONAP has made significant progress in supporting network slicing capabilities and is supporting the implementation of a variety of network slicing use cases.

Illustrated in Figure 1.4, the ONAP framework [1] allows for mobile network customers and operators to order slice-based services, create end-to-end network slices, and operate over the entire lifecycle of a network slice. The ONAP framework supports network slice design and creation, activation, deactivation, and termination across the RAN, core network, and transport network (TN) domains.

Figure 1.4 ONAP E2E Network slicing – Alignment with SDOs



The ONAP framework comprises two main parts called ‘environments,’ that support network slicing in the following ways:

- *Design-time environment is a development environment with all the functions and libraries needed for the development of new capabilities. This includes design of Network Slice Templates (NST) as well as Network Service (NS) definitions. These definitions form the basis for network slice management in the ONAP framework.*
- *Execution-time environment executes slice selection and optimization policies with the help of rules prepared in the design-time environment and handles tasks such as performance monitoring and service orchestration for network slices.*

The ONAP framework aligns with different standard bodies- 3GPP, TMF, European Telecommunications Standards Institute (ETSI), IETF and O-RAN - to enable end-to-end framework based on open Application Program Interface (API)s and interfaces.

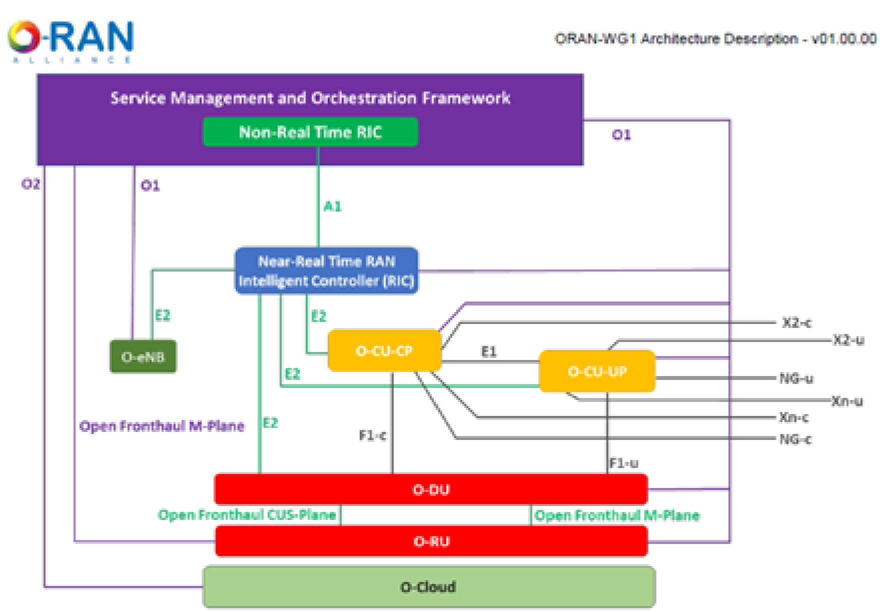
1.2.3 O-RAN ALLIANCE

The O-RAN (Open RAN) ALLIANCE specifies solutions for an open RAN to support end-to-end network slicing for 3GPP-specified 5G systems. A network slice spans from the UE to the Data Network (DN), supported by open RAN functions (O-RU, O-DU, O-CU: see figure below), Transport Networks (TN), and the 5G Core. O-RAN defines network slicing related technical use cases, requirements, and architecture that can be realized with open RAN functions and interfaces [2]. For instance, the O-RAN Service and Management Orchestration Framework (SMO) configures and orchestrates the open RAN functions and related TNs to realize a network slice. The O-RAN network slicing technical use cases cover RAN lifecycle, optimization, and other areas where the RAN play a role in network slicing, and some of them include:

- *Slice Subnet Management and Provisioning Use Case including Slice Subnet Instance Creation, Activation, Modification, Deactivation, Termination and Configuration.*
- *Multi-vendor Slices Use Case where operators select appropriate vendor (Open Distributed Unit) O-DU and O-CU (Open Core Unit) functions to meet service requirements with consideration of RAN sharing among different operators.*
- *NSSAI Resource Allocation Optimization Use Case to optimize the resource allocation for network slice instances to improve user experiences and maintain optimal network usage efficiency.*
- *RAN Slice SLA (Service Level Agreement) Assurance Use Case for fine-tuned RAN behavior to assure RAN slice SLAs dynamically, and assurance mechanisms to detect and prevent SLA violations.*

The “O-RAN Working Group 1 Slicing Architecture” document specifies the use cases and architectures for network slicing and the “O-RAN Working Group 1 Study on O-RAN Slicing” document considers network slicing needs from a RAN and potential solutions [3].

Figure 1.5 Logical architecture of O-RAN



2. Network Slicing Use Cases: Path to Initial Commercialization

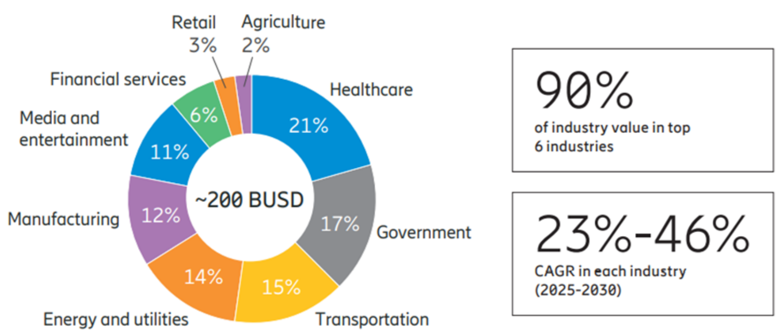
In this chapter, we investigate the opportunities of network slicing, as 5G networks (and 5G standalone networks in particular) continue to proliferate around the globe. As of March 2022, GSA (Global Mobile Supplier Association) reported 488 operators investing in 5G worldwide, with 103 of them focused on 5G Standalone [4]. In a different report, the same source catalogued 20 operators as having deployed or launched 5G Standalone in public networks [5]. As 5G standalone continues to grow around the world, network slicing opportunities will grow right alongside them.

2.1 Introduction

Many industry analysts and technology pundits believe that network slicing could be a key factor in monetizing 5G investments. With the introduction of slicing capabilities in 5G Standalone networks, as more are launched worldwide, communication service providers (CSPs) are looking to create new services and applications that will be enabled by these new capabilities.

The most immediate comparison to today's situation occurred in the early 2000s when 3G was introduced with higher data rate capability and CSPs were looking for a 'killer application' that would ignite demand for it. Such new demand, if backed by new revenues, would create and fund a natural expansion of the new 3G networks, and eventually help replace the outdated GSM networks. Ultimately the real explosion of data had to wait until the iPhone launched in 2007, and although it was a GSM(EDGE)-only device, it created the 'Internet in your hand' experience that would evolve in the smartphones we have today.

Figure 2.1 The industries with the highest revenue potential for network slicing for CSPs (Network slicing: A go-to-market guide to capture the high revenue potential – Ericsson 2022)



In the quest for identifying the best use case opportunities to address with network slicing, vendors and consultants have published a few reports, such as "Network Slicing: Top 10 use cases to target" and "Network slicing: A go-to-market guide to capture the high revenue potential" [6]. As the basis for these reports, an international management consulting firm (Arthur D Little) analyzed more than 70 external market reports with the goal of finding revenue potentials of most promising industry segments. After reviewing more than 400 digital use cases in 70 industries, this research indicated that six industries have about 90% of addressable revenue potential, and one or two use cases in each industry will account for most of the addressable revenue.

Additionally, the report detailed the most promising use cases in each of these industries and identified common denominators requirements across those industries from an implementation standpoint:

- *High throughput video (for ex. in telehealth and remote assessment in healthcare, or tele-operated vehicles, remote inspection in energy/utilities etc.)*
- *(Ultra) Low latency (for ex. in command/control of tele-operated vehicles or remote surgery, AR/VR)*
- *High availability – in times of congestion*
- *Improved security (as some of these services have critical impact)*

Notably, high throughput and low latency capabilities in the RAN domain carry a high cost in terms of spectrum. This is where network slicing functionalities will show their value, as they would allow selective enablement for those features only for specific groups of users. The slicing benefit will also expand to core and transport, allowing CSPs to create even more custom-tailored logical slices to address the new services. Having established the demand and potential for new services and possible associated revenues, the decision on initial commercialization can then go through different directions of assessment, such as customer appeal, and public relations impact - all the way down to economics.

In the next subchapters we will try to define a structure for the 'main ingredients' that can aggregate into creating a new slicing service that might address a specific customer use-case. We will propose a categorization of use-cases based on commonalities of implementations and a differentiation based on monetization approaches. Following up, we'll identify a few examples of real-life scenarios that will help articulate such a structure for the purpose of facilitating a decision on initial commercialization service or use case.

The chapter will close with an overview of the slicing ecosystem status, for the purpose of determining the confines of such decision based on availability near to mid-term, and a final consideration section.

2.2 Use Cases Realization: Categories

There are various categories of slicing use cases identified in the current literature that demonstrate how CSPs and other entities can create a strategy for deployment or a portfolio of sliced based services. Our focus is to help flesh out the initial commercialization view and address CSP concerns around the economics. From a planning/engineering view some of the important differentiations are:

- *Localization categories:*
  - » ***Localized use cases**: slice service is confined to a small area (one site and/or its immediate neighbors). For example, creating reserved/dedicated resources for an enterprise or a fixed wireless user.*
  - » ***Area wide uses cases:** slice service is enabled in an area (such as a campus or airport) that would need multiple sites to be configured with specific slice configurations*
  - » ***Nationwide use cases:** slice service to be configured across the entire network. For instance, creating a nationwide gaming slice that allows users to experience the gaming slice benefit wherever they go.*
- *Time/duration categories:*
  - » ***Temporary use cases:** where a slice service can be created for a limited duration (to minimize the impact for existing services) and destroyed at the end of the offering. Example - upselling a latency boost for a limited time for a user*
  - » ***Long term/Permanent use case:** the slice for the specific service is created permanently. For example, of a gaming slice or a video streaming slice that can be accessed by premium users all the time*

The impact of localization and temporal differentiation of use cases can play an important role in deciding the commercial viability of a new service. Note that the slicing framework does not create new resources, and, specifically in the RAN realm, the finite nature of resources translates any additional requirements into capacity impact, which can incur additional costs by triggering network expansion. For example, creating a permanent slice that reserves resources will reduce capacity for existing traffic and can trigger additional considerations for increasing capacity.

From this perspective combining the advantages of temporary/localized use cases can create an attractive model for initial commercialization from several perspectives:

- ***Temporary use cases**: for example upselling a temporary boost in throughput (for better video performance) or latency (for a gamer) – can discount the capacity impact on the network, which is usually a major concern for CSPs. These cases can benefit from the flexibility of orchestration in creating/deleting network elements in Core/RAN and network configurations*
- ***Localized use cases:** could have a similar advantage as above and would also introduce less network configuration complexity compared with enabling a nationwide (for example gaming) always available slice service*

Another important differentiation in implementation is based on the capabilities of network and devices in terms of the mechanisms and also the granularity of allocating and managing the slices:

- ***Subscription-based slicing:***
  - *Network can establish one slice (for each of the DNNs the device is using) based only on the subscription profile (DNN and NSSAI combinations)*
- ***Policy-based slicing:***
  - *The device and network can use slicing policies (containing traffic descriptors matched with NSSAI slice tags) to create/manage slices*
- ***Application-based slicing:***
  - *A subset of the policy-based slicing where the device is also able to match individual applications and 'route' them to specific slices*

Subscription-based slicing can be used with virtually any 5G Standalone (SA) capable device, as it does not need any advanced slice related capabilities. The subscription profile will have the slice information and can be as granular as per Data Network Name (DNN), so when device attaches the provisioned NSSAI will be applied to RAN and Core and can be used to customize a specific treatment of user-plane data in, for example, higher prioritization or reserved resources. Since devices usually use one common DNN for regular data traffic, this approach will apply the slice configured treatment to the entire DNN. The network is unable to distinguish between different applications or types of traffic and will do it for the entire life of the data session.

Policy-based slicing assumes the network can communicate to the device slice policies via the UE Route Selection Policy (URSP), as defined by 3GPP, that can be processed by the device to initiate/modify PDN sessions with specific slice IDs (NSSAIs) for the packets that match the traffic descriptors contained in the URSP table. As the URSP table has many elements that can be used for implementation, we will consider this a broader category that includes even devices with limited capabilities in this realm.

Application-based slicing is a subset of the ***Policy-based slicing*** that is complemented by the ability of device and network to match specific (or unique) slices to individual applications, or type of applications. Such devices will be able to map a specific application to a specific slice offered by the network in the URSP table, and the network can further manage it through policy changes.

As we will discuss later in the ecosystem status, the application-based slicing is a more complex scenario that requires further developments and agreements between CSPs and device manufacturers and OS providers. From a practical perspective this is the case of a smartphone that can carry applications (video, gaming, AR/VR etc.) that can benefit from a different and individual treatment for each of these applications.

From a potential economic perspective and technological approach to each, possible examples between use cases could perhaps include areas such as:

- ***Upsell use-case****: where the user pays a one-time fee for a premium slice-enabled treatment, such as in the previously mentioned example of a temporary boost in performance for video or gaming*
- ***Premium service use-case:*** *where the user pays in a subscription model and will be able to use such premium treatment whenever it needs*

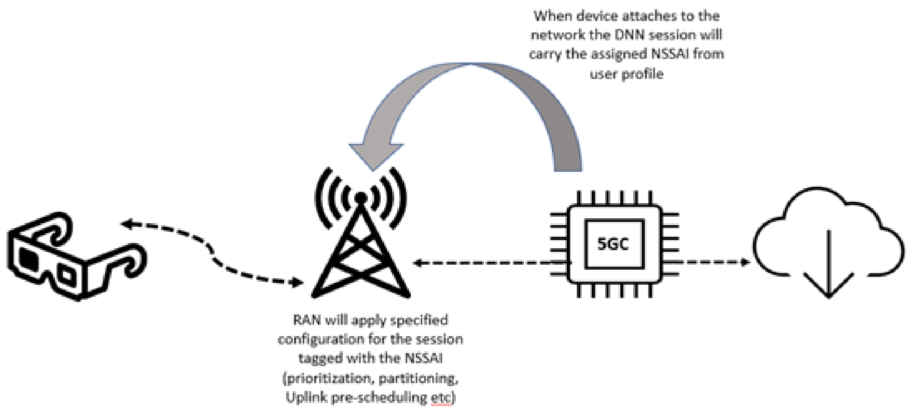
2.3 Use Cases Realization: Building Blocks

Creating a new service using network slicing starts with identifying the new attributes required by it, and then using the slicing framework to 'insert' the new service into the network, making sure it can only be used by specific customers. In some cases, it can also be required to use slicing to isolate such new service from the regular traffic to manage the impact to and from such traffic. Sometimes the mobility aspect of the new service may have additional requirements, so it does not cease to function in areas not supporting the slice.

The simplest approach is obviously to start with the use case requirements based on customer demand and then identify the network configurations that are needed to fulfill them. In the next step, the network nodes apply such configurations for just that specific slice to ensure that only subscribers that attach with such slice will benefit from them.

Let us take for example the use case of AR/VR glasses (see figure 2.2), which would typically require the network to provide a lower latency and higher throughput on downlink. The operator can create a dedicated AR/VR slice which in the RAN node as an example will activate higher priority on downlink for the rendered video stream - similar with how it does today for video calling - as well as pre-allocate uplink grants so it can keep end-to-end latency low. Since the AR/VR glasses will generate only this type of traffic on a single session, the operator can use the Subscription-based slicing and configure the user profile with the specific NSSAI to allow a device to attach to the network DNN session that will be carrying it. The RAN and Core nodes will then be configured to support such NSSAI and apply the special treatment defined for it to all the packets carried by the device session.

Figure 2.2 An example of AR/VR service slice realization



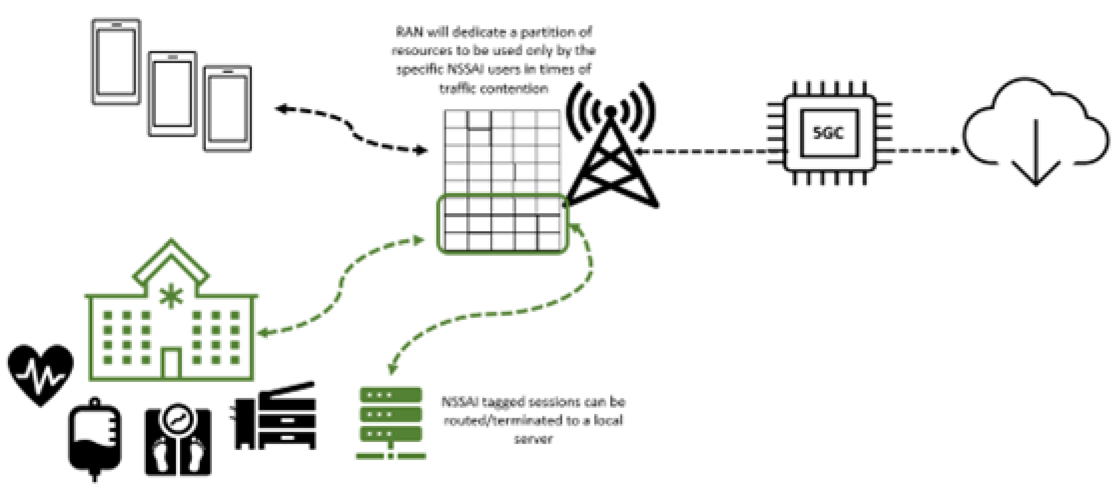
By assigning a specific configuration only for a specific NSSAI, a virtual network has been created for these subscribers. Using this implementation, the operator can define a special treatment that would only be applied to packets carried over the session tagged with the specific NSSAI, such as:

- *Schedule downlink rendered video packets with higher priority than regular data*
- *Pre-allocate uplink grants for faster delivery of uplink video*
- *If desired, allocate a separate partition of radio-resources that would only be used by the AR/VR device in times of traffic contention or congestion*
- *The NSSAI can also be used to assign a higher priority for packets over backhaul, or to route the traffic to the closest video server (or a dedicated gateway).*

The slicing framework in this case can offer a level of service control, in the sense that the operator can restrict on what cells such service is available. This can be useful to make sure the service is only offered on sites that have enough resources to ensure a good experience.

Another useful example to examine is a private network use case (see figure 2.3), where a customer might require dedicated wireless network resources on its premises - for example, when a hospital that wants to connect various devices using an operator wireless network. While it can also do so by using regular data service offered by the CSP, in this case the demand could be for a premium network availability/stability even in cases of traffic congestion. By using the slicing approach and an existing site that covers the customer premise, an operator can avoid building a dedicated site. The virtual network it creates by using a specific NSSAI will ensure dedicated resources and could additionally route the traffic of specific NSSAI subscribers to a local server, if requested, using a separate and secured backhaul.

Figure 2.3 An example of slice enabled private network using radio resource partitioning



For IoT types of devices such as printers and ventilators (or other device), a Subscription-based slicing model can be easily implemented so when they attach to the network, their entire session would be tagged with a specific NSSAI and benefit from the special treatment configured for it.

A more complex scenario could apply for smartphones owned by the hospital that might require specific applications to be able for example to connect to a local hospital server. This can be solved by using a Policy-based slicing approach. This particular example is also called "Enterprise-slicing," in which the device receives an URSP table that maps the specific hospital applications with a specific NSSAI at the moment of registration, which can then be used by the Core to route-select the NSSAI tagged session to the local server.

One of the implications of using this approach though comes from having to handle the mobility aspect of the slicing. While the IoT devices are presumably confined to the premises and only be served by the site configured with their assigned NSSAI, the smartphones could move to another site that would not have the hospital's NSSAI configured. In such case, any devices associated with a non-supported NSSAI would be dropped from the network and further attempts to restore them would be rejected, resulting in connectivity loss for the specific applications. One way to address this mobility aspect of an NSSAI 'island' is to configure all the sites in the same Tracking Area (TA) to support the NSSAI but without enabling the dedicated attributes (partitioning or prioritization) that have been applied on the 'hospital site'. This will ensure seamless handover of the specific NSSAI-tagged sessions to the surrounding sites. Ultimately, when the hospital's smartphones would cross over the TA border, the Core will be able to detect it and send an updated URSP table that would remove the hospital's NSSAI.

Another approach for services that can be enabled through slicing capabilities could start with the operator identifying certain network features that can improve specific characteristics of the data service and then packaging them for a dedicated slice. A good example of this would be features that focus on reducing end-to-end latency, such as prioritization (in the RAN and transport) or uplink grant prescheduling. The next step would be to configure the network nodes to activate such features only for a specific NSSAI, that can then be offered to subscribers as a Temporary use case. In such an example, the CSP can create a unique NSSAI that enables low-latency features and upsell it as a gaming slice for subscribers that demand – and are willing to pay a premium for - such service for improved experience in FPS (first person shooter) games.

Implementing the above-mentioned gaming slice requires all the other network building blocks mentioned in the previous examples:

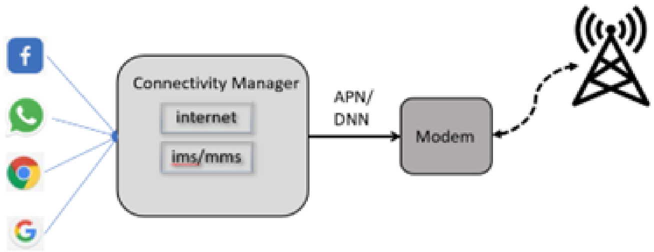
- *RAN nodes must have the capability to enable the features required for low latency only for specific NSSAI*
- *Core must be able to on-demand setup and manage the URSP policies that would indicate the subscriber the required NSSAI, or use Subscription-based slicing*

In the subscription-based approach, setting the subscriber with a specific NSSAI for a data PDN session will result in all of data traffic using the session to be tagged with it, as opposed to an Enterprise slicing example, where such a session would be used only by the enterprise managed applications. Of course, neither of these examples would be useful in the gaming slice scenario because allowing all mobile traffic to use low latency features could degrade the experience. The enterprise slice approach requires the use of an MDM (Mobile Data Management) system that can designate which applications could request a specific slice (MDM cannot be used on consumer devices w/o consent of the subscriber). As such, for our gaming slice example (or "category slicing offer") to work, it would require the device to support Application-based slicing, which relies on Policy-based slicing but additionally allows individual applications – in our case gaming apps – to request a specific slice.

To understand this better, see Figure 2.4 below for a simplified view of the Android OS (operating system) management of data connectivity for applications. Applications invoke the connectivity manager (CM) process and request network capabilities (such as internet, IMS, mms etc.) that are then used by the CM to decide what APN (or DNN in 5GS) with which to ask the modem to establish a session.

The URSP capability designed by 3GPP enables the modem to initiate PDN sessions with specific NSSAI for CM requests with specific traffic descriptors such as DNNs, appID+OSid (application ID and operating system ID) or IP-tuples. In our gaming slice example, assuming the network already indicated to the modem a URSP rule containing the gaming NSSAI route that offers low latency treatment, a mechanism that would allow the OS/CM to pass a matching traffic descriptor from our gaming application would be needed. Such mechanism is currently not available on Android or iOS smartphones, and CSPs and OS vendors are working together to identify a solution that would satisfy many considerations such as data privacy, data/service

Figure 2.4 Android Connectivity manager simplified view



abuse policing and others. Until then, the category slicing (gaming, AR/VR etc.) can only work using subscription-based approaches, so only with dedicated devices (non-smartphones).

One aspect we have hinted at but not discussed in detail is the on-demand enablement of slicing in the upsell scenario. While the Premium use case approach can be implemented by configuring the network with allowed NSSAI-s and modifying the user profile to use specific NSSAI when accessing the network, the Upsell approach requires communication between subscriber's app, or the OS in case it intermediates the payment, and the Core for the purpose of dynamically managing the slicing activation/deactivation, as well as the monetization aspect. An example of upsell in our gaming slice scenario involves asking the user to purchase the gaming slice treatment for a desired duration at a premium slice, at which point a network API would be accessed to enable the user to access the gaming slice. In this example, it would trigger the sending of a new URSP table that contains the rule and route for the gaming NSSAI.

In summary, here are some of the building blocks that are required for the realization of a new slice-enabled service:

- *RAN node capability of enabling specific features for specific NSSAI - for example, resource partitioning, scheduler prioritization, uplink prescheduling etc. in other words the attributes of the customized service*
- *Core Network support for NSSAI management (like creation, deletion etc.) and API support for end-to-end commercialization*
- *Device capabilities to accept and follow slicing policies as directed by the network (in the case of Policy-based slicing)*

While some use cases can be implemented using a subscription-based slicing approach and have a very strong chance to become the initial commercialization of slicing due to simplicity, they can also raise concerns on capacity impact and profitability. From this standpoint, we should not ignore the policy-based slicing use-cases in combination with upsell functionalities (based on network APIs) with the advantage of lower impact to existing spectrum, flexible scalability and management, as well as more appeal to the customer in terms of monetization (Upsell vs Premium).

2.4 Ecosystem Status

This section provides an overview of the current state of capabilities and technical features available in the ecosystem across domains.

2.4.1 Radio Access Network (RAN)

By assessing the status of the RAN in terms of the capability to support introduction of slice-based services, a clear distinction must be made between features needed for specialized treatment (such as features that improve latency or throughput) and features that make possible to offer such specialized treatment only to individual slices (or a group of them). The latter are commonly named 'Service-aware' because they can be enabled based on the service they are meant to differentiate. Without the Service-aware capabilities, the operator would not be able to selectively deliver enhanced features as needed for a new service, and so would not be able to monetize or even deploy them.

Without going into technical details related to 'specialized treatment' type of features, most of them would fall in areas such as radio resource management, observability, mobility management, transport differentiation and others.

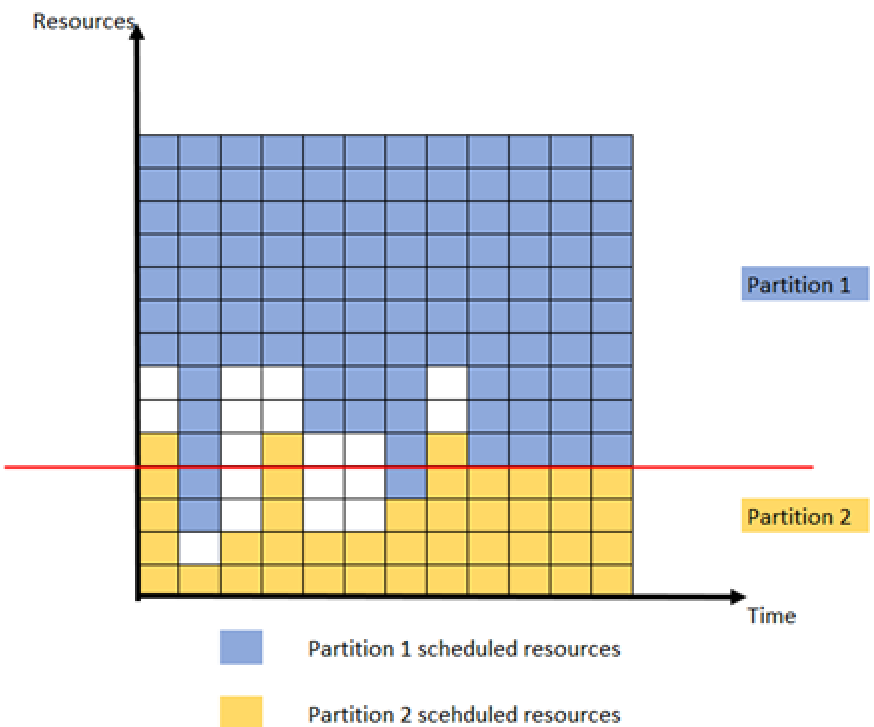
RAN resource management (RRM) enables the orchestration of RAN resources and QoS policies in a way that can guarantee fulfillment of SLAs, while maintaining access to available resources for different service categories. This is achieved by providing slice awareness in link adaptation, scheduler configuration, radio resource partitioning and admission control, and being able to customize all these aspects for a specific slice is the main pillar for slice enabled services.

Service-aware scheduler prioritization can assign specific scheduling weights based on a standardized 5G QoS Identifier (5QI) and NSSAI, which is a powerful tool in designing different special treatments required by different services. For example, in offering a video prioritization type of service using NSSAI differentiation, an operator can configure a specific prioritization for video packets to improve the streaming performance. Other techniques that can reduce the uplink latency, such as pre-scheduling or grant-free, can also be enabled for specific NSSAI for the same purpose.

Radio resource partitioning (RRP) is one of the features that falls in both categories mentioned above, in the sense that it offers a specialized treatment by separating resources and allocates them differently based on NSSAI. Making RRM slice-aware allows the RAN domain to create logical and virtual radios that can be isolated in times of traffic contention but also shared when usage is low.

In the example shown in Figure 2.5, this capability can address the private networks use case. Traditionally a request by a small business to augment capacity/coverage on their premises was addressed by an operator through installing a dedicated radio. With RRP, in the case where the business premise is already covered by an existing site, the customer can be offered dedicated and isolated resources, and additional backhaul security if required. It would be a virtual dedicated radio, but one that can be shared with regular subscribers when the business does not use it.

Figure 2.5 A simplified view of Radio resource partitioning



RAN slicing observability provides the monitoring of relevant Performance Measurements and calculation of KPIs per slice, which is essential in providing insights and assurance to the tenants of each network slice. It also helps the RRM to optimize the RAN resources for the serving slices, under varying traffic and radio conditions. The observability should be filterable with multiple criteria to gain a good observation. Some example criteria are counters per S-NSSAI (per slice), PLMN ID, and QoS(5QI).

Service aware mobility is important in network slicing because a user needs service continuity for its slice with mobility. Some of the features provided by RAN vendors in this domain include service-aware handover, which allows the handover between nodes that do not have the same NSSAI support to accept only bearers with supported NSSAI – and service-aware traffic steering, which operators can use to trigger mobility of specific NSSAI to desired layers either in idle or connected mode. As part of the network slicing framework, neighboring RAN nodes signal to each other the set of network slices supported for each tracking area and this information can be used to steer users towards the cells where the slices in use by a terminal are supported. In addition to the radio conditions and load, mobility management functions consider the supported network slices in allocating a candidate target cell. It is important to note that slice information cannot be carried from 5G Standalone to 4G in inter-RAT (radio access technology) mobility.

In the transport domain the RAN can offer specific virtual local area networks (VLAN)s and even specific ports to be used for configured slices, as well as enabling IPsec on a slice basis, allowing the offering of services with enhanced security on the backhaul.

Another aspect related to security is slice isolation through radio resource partitioning, which ensures protection against DOS (denial of service) attacks, as well as against resource exhaustion. The other security threats that network slicing faces include monitoring, traffic injection, and impersonation attacks. RAN provides observability in terms of proper alarm generation along with the root cause analysis. Security functions detect network anomalies based on security analytics and provide adaptive security by looping back analytics to the policy automation

2.4.2 Core Network

The slicing feature set on the core network can be broadly divided into three categories:
Provisioning of Slice, Slice Aware Policy/Feature capabilities, and Management and Orchestration.

The provisioning of slice use cases which require a profile update with new slice ID's and DNN (Data Network Name) based on static requirements (e.g., high throughput, low latency) are the most straightforward to implement and operate (e.g., Subscription-based slicing). These use cases can be supported with a standard feature set on the core network, as it does not require traffic classification and slice selection at application level. The provisioning effort typically requires a one-time update of the subscriber profile and network configuration. This architecture model is ideal for Nationwide and Areawide use cases. Assuming best effort traffic belongs to default slice (Slice/Service Type-1, Slice Differentiator 1), two new slices designed for high throughput (Slice/Service Type-1, Slice Differentiator 2) and low latency (Slice/Service Type-1, Slice Differentiator-3) can be enabled by network configuration and subscriber profiles accordingly.

The provisioning of Policy/Application-based slicing use cases requires tight interworking with policy control function (PCF) and operator provisioning tools that update subscriber profiles and network config. This framework supports more advanced slice use cases, requiring slicing at the application, location, and time granularity. This is achieved by provisioning and communicating UE Route Selection policy (URSP) rules to the device. URSP rules contain Traffic Descriptors and Route Selection Descriptor components as listed in Tables 2.1 and 2.2, respectively. The traffic descriptor determines when the rule is applicable, whereas the Route selection Descriptor provides parameters needed to initiate a request for a slice.

For example, consider a smartphone with application requirements that include best-effort traffic (e.g., regular internet), high throughput, and low latency (e.g., AR/VR).

To support these, PCF may provide three URSP rules each indicating traffic descriptors and route slice selection parameters, as per tables 2.1 and 2.2. As soon as the UE detects a new application, the UE will evaluate the URSP rules in the order of Rule Precedence and determine whether the traffic descriptor of any URSP rule matches the application. If a URSP rule applies to an application, the UE determines the appropriate Route Selection Descriptor in the order of Route Selection Descriptor Precedence within that URSP rule.

2.4.2.1 Slice Based Policy and Deployment Model

Besides provisioning-related capabilities highlighted in Figure 2.7, the core network plays a vital role in supporting slice use cases requiring traffic isolation, security, and latency. Traffic isolation, Security, and latency-related use cases can be supported by deploying slice-specific dedicated 5G core nodes. Although 3GPP standards support the deployment of 5GC nodes at the slice level, from a practical implementation perspective, dedicated packet core nodes (e.g., SMF/UPF/AMF) would address most of the slice requirements. For example, a low latency slice can be supported by deploying a dedicated UPF at the edge while retaining all other control plane functions in a centralized location in a shared environment.

Since the network functions are cloud native in a 5GC network, creating or deleting them as part of slice life-cycle management is a straightforward process when scaling across various use cases. A cost-effective option would be to only create/delete on-demand dedicated UPF for customized user plane requirements, such as security and latency if deployed at a local facility/datacenter, while control functions can be reused. Applying the same approach to the control plane would only be needed for redundancy or if customers require more isolation. For the 5GC the network functions such as Access and Mobility Function (AMF), Session Management Function (SMF), User Plane Function (UPF) and Policy Control Function (PCF) can be dedicated or shared per slice as illustrated in Figure 2.8. If the slices are owned by an operator and not in a Multiple Operator Core Network (MOCN) or Multiple Operator Radio Access Network (MORAN) type of architecture, it can be assumed that a deployment would have dedicated network functions while the subscriber profile database is shared.

Figure 2.6 Examples of Network Slice Types

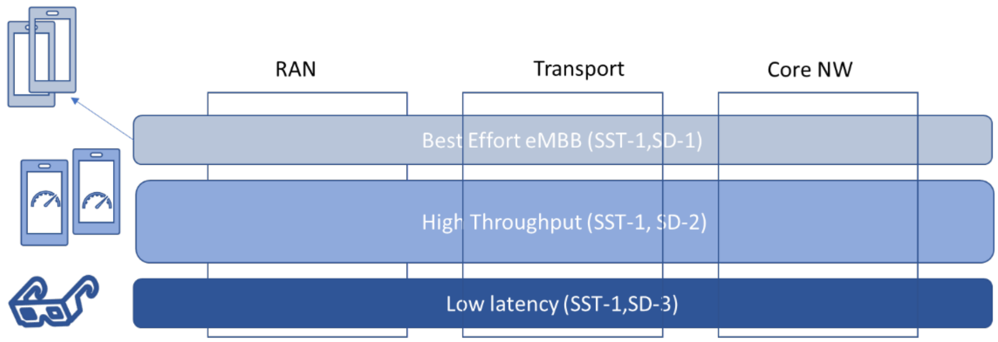


Table 2.1 URSP Rule-Traffic Descriptor

Information name	Description
Rule Precedence	Determines the order the URSP rule is enforced in the UE
Traffic descriptors	Defines the traffic descriptor components for the URSP rule
Application descriptors	Consists of OSId and OSAppId(s)
IP descriptors	Destination IP 3 tuple(s) (IP address or IPv6 network prefix, port number, protocol ID of the protocol above IP)
Domain descriptors	Destination FQDN(s) or a regular expression as a domain name matching criteria
Non-IP descriptors (NOTE 5)	Desciptor(s) for destination information of non-IP traffic
DNN	Matched against the DNN information provided by the application
Connection Capabilities	Matched against the information provided by a UE application when it requests a network connection with certain capapbilities

Figure 2.7 Example of Multiple Slices using URSP

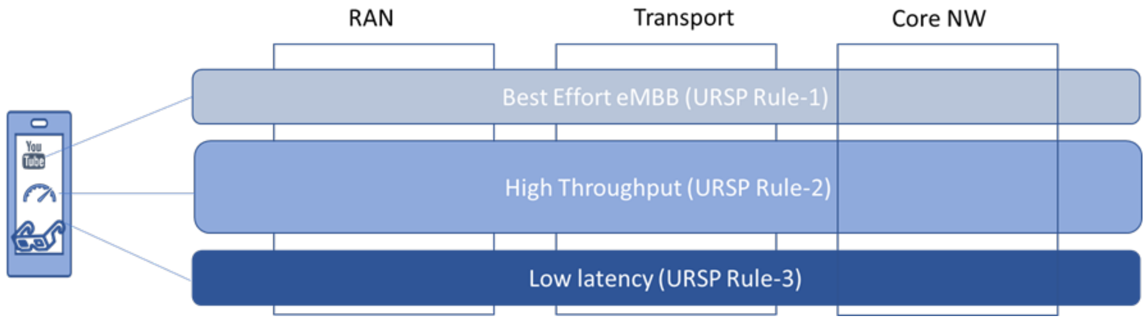
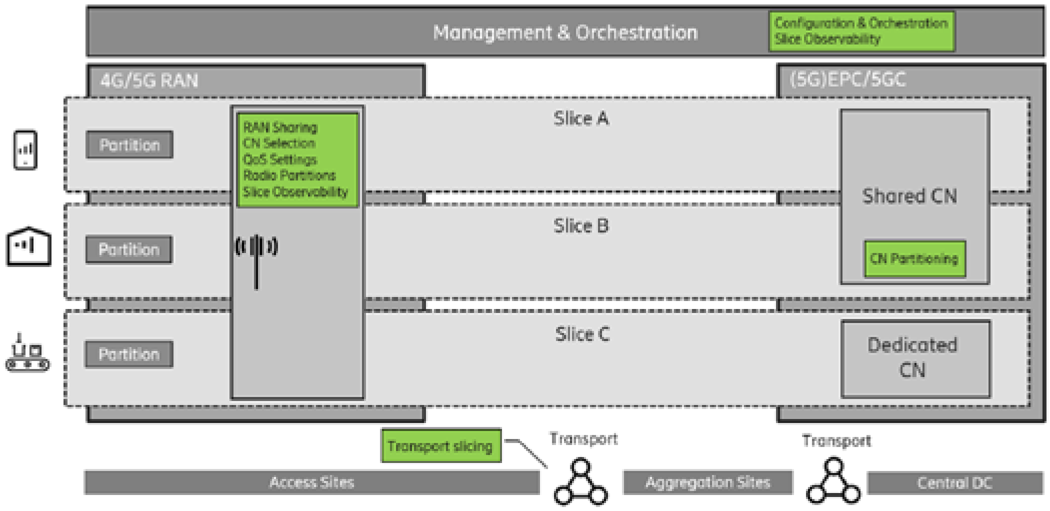


Table 2.2 URSP Rule-Route Selection Descriptor

Information name	Description
Route Selection Descriptor Precedence	Determines the order the Route Selection Descriptor are to be applied
Route selection components	Defines the route selection compenents
SSC Mode Selection	One single value of SSC mode
Network Slice Selection	Either a single value or a list of vales of S-NSSAI(s)
DNN Selection	Either a single value or a list of values of DNN(s)
PDU Session Type Selection	One single value of PDU Sesssion Type
Non-Seamless Offload indication	Indicates if the traffic of the natching application is to be offloaded to non-3GPP access outside of a PDU Session
Access Type preference	Indicates the preferred Access Type (3GPP or non-3GPP or Multi-Access) when the UE establisdhes a PDU Session for the matching application
Route Selection Validation Criteria	Defines the Route Validation Criteria components
Time Window	Time window when the matching traffic is allowed. The RSD is not considered to be valid if the current time is not in the time window
Location Criteria	UE location where the matching traffic is allowed. The RSD rule is not considered to be valid if the UE location does not match the location criteria

Figure 2.8 Shared or Dedicated Core Resources



2.4.3 Transport Network

In the end-to-end scheme of network slicing, the transport network plays key role in meeting end customer slice requirements. The scope of 3GPP slice identifiers is not extended into IP domain, this makes it difficult to slice transport resources at the same granularity as 5G Core or Radio Network. However, IP traffic management solutions available at underlay and overlay could be used to meet necessary slice requirements. For the transport network to provide adequate transportation for network slicing, it needs these functions: inter-domain packet mapping, transport resource partitioning, resource isolation, optimal path selection, and quality assurance.

Generally, data packets are mapped from RAN and Core domains to the transport domains. This could be done in uplink based on VLAN ID, destination IP address or physical port from the RAN node and in downlink based on source or destination IP address. The standard practice to prioritize traffic on transport is to mark packets using different DSCP (Differentiated Service Point Code) values based on traffic service requirements. If a slice or group of slices have several traffic flows with individual requirement on transport characteristics, then extension of DSCP (DSCP or 802.1p bits) at slice and 5QI granularity will provide capability to get a proper treatment for that slice traffic. Mapping of 5QI to DSCP and 802.1p can be done individually per slice or group of slices. For isolation and quality assurance, VPNs, IPsec, GRE, Segment Routing, MPLS can be used.

In today's network slicing ecosystem, the industry is mainly focused on RAN and Core, while relying on DSCP to differentiate network slices in transport. However, it is very likely that in the future, new capabilities will be needed to improve optimal path selection and quality assurance in transport. For example, areas that would create such demand include remote inspection or even remote surgery over long distances. A possible direction to address this challenge could be using SDN (software defined network)/SD-WAN (software defined wide area network) technologies that can deliver better managed latency, which are more suitable for such future use cases. Integration of a central SDN/SD-WAN controller with network slicing orchestration management might be needed to enable a more granular end-to-end slicing coordination. It is likely other possible transport enhancements will be reviewed and debated when the need for such use cases arise.

2.4.4 Device

To enable 5G Network slicing, the device plays a very important role in the end-to-end (E2E) architecture. UE is responsible for creating a new network slice connection when the conditions are met such as a new DNN connection is initiated, or a specified application is launched. Device slicing architecture can be generally divided into Operating System-Centric or Modem-Centric.

Modem-Centric architecture is suitable for 5G devices that do not rely on Operating System (OS) to interface between modem and application. For example, an IoT device may have an application program to directly communicate with the modem over command. Based on the slicing information predefined on the device, provisioned in the SIM, or provided by network during

5G registration, the modem will determine application traffic and slicing mapping. This is relatively straightforward architecture and deployment for use cases that does not require support for multiple slices.

OS-Centric architecture is suitable for 5G devices that rely on Operation system to bridge between modem and application for network connection (see 2.3 for details). Applications on smartphone Android and iOS do not communicate with the modem directly for network connectivity request. OS manages all the connectivity requests for different applications and manipulate data flow on each network slice. In an OS-centric architecture, OS is responsible to map a data stream from an application to a corresponding slice NSSAI, it gathers policy information (URSP) and other relevant information such as designated applications provide by a trusted entity such as mobile device management (MDM) and determine the slice and application mapping.

URSP is essential for application-based slicing (see 2.4.2 for details on URSP), as the traffic descriptor in the URSP defines rules for matching applications or category of applications with the specific network slice. The OS ingests the URSP policy from modem over API and uses it for traffic steering. Other traffic descriptors in the URSP such as the Destination Fully Qualified Domain Name (FQDN) and IP address can be used for domain specific use cases in FWA (Fixed Wireless Access) and IoT (Internet of Things) applications.

The current device ecosystem already has several generations of 5G capable chipsets, and from the slicing support perspective here is a summary:

- *All 5G Standalone capable devices can be used in a Subscription-based slicing use-case. The NSSAI is transparent to the device, since it is applied based on the UDM (Unified Data management) profile per DNN, with the drawback that this limits the 'granularity' of slicing in the sense that all the traffic with a particular DNN will use the same NSSAI.*
- *Policy-based slicing is supported by almost all of the current 5G Standalone smartphones, but since no commercial services have been launched with it a conservative assumption would be only basic traffic descriptors (such as DNN, appID+Osid) are supported, and more testing would be needed for example for advanced features such as location (TAC, Cellid) based rules*
- *Enterprise slicing – announced late 2021 with Android 12 – allows the use of MDM systems to tag certain applications with a specific appID+Osid tag, which can be used in a URSP rule to associate them with a distinct NSSAI, but the use of mobile device management (MDM) is usually associated with non-consumer devices, such as the one owned and controlled by the enterprise*
- *Application-based slicing still needs cooperation between OS vendors and CSPs, and would probably not become ready for commercialization before 2023*

2.5 Slice Orchestration and Automation

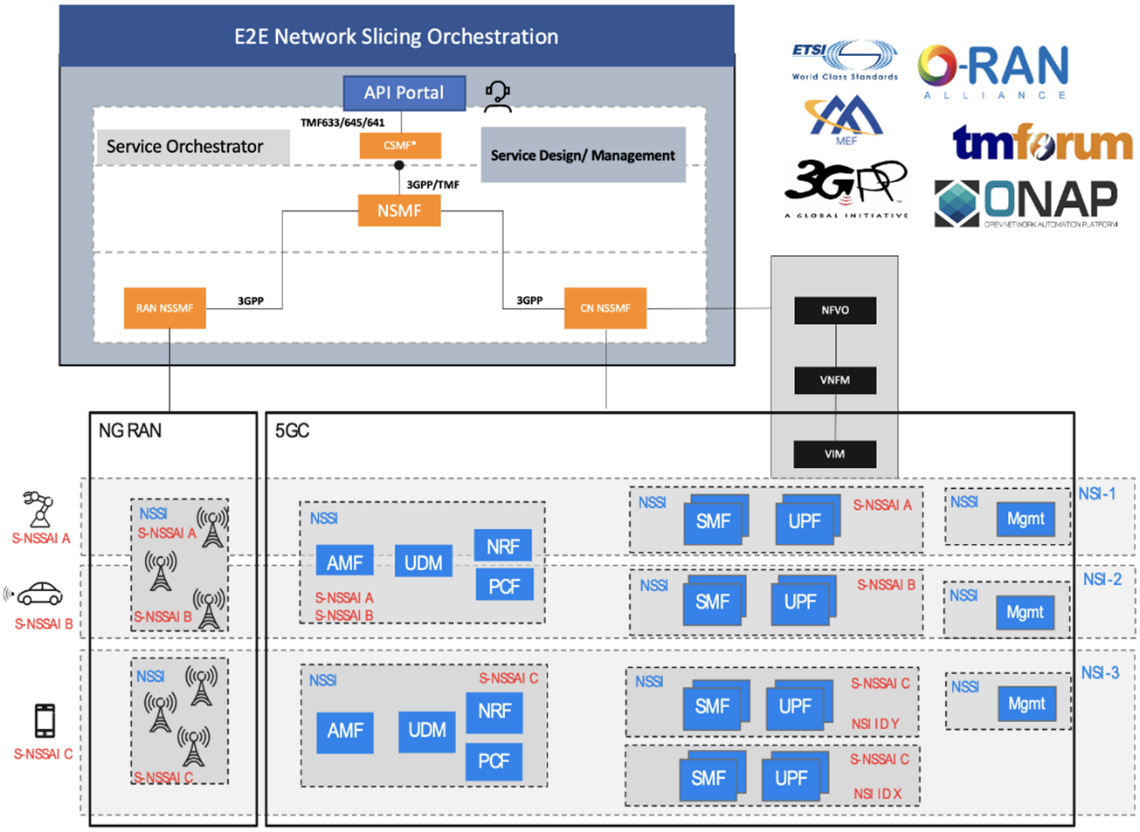
Telecom service providers have traditionally delivered services with limited flexibility and control. However, this traditional approach is not suitable for supporting new-generation requirements, which are complex, dynamic, and often require automated service feedback. The network slicing in 5G could advance these complex requirements to the next level. It is highly recommended to have a full-scale slice orchestration solution to handle complex slice requirements and have better agility in service offering. This section provides an overview of E2E network slice orchestration, standards framework, high-level flow for slice life cycle, and connection with NFV.

2.5.1 Orchestration and Automation Overview

An End-to-End (E2E) network slice orchestration landscape involves several standard bodies (e.g., 3GPP, ONAP, TMFORUM). At the heart of the solution are three logical entities specified by 3GPP, namely Communication Service Management Function (CSMF), Network Slice Management Function (NSMF), and Network Slice Subnet Management function (NSSMF). Integration of these new nodes with other network components like OSS/BSS, Inventory, RAN, Core Network, and Transport network elements will provide slice orchestration capabilities.

Figure 2.9 shows the high-level architecture for slicing Management and Orchestration (MANO).

Figure 2.9 End-to-End Slicing Orchestration



3GPP Network Slicing Management Functions:

- *Communication Service Management Function (CSMF), which acts as the interface towards service order management and Operations Support Systems (OSS)/Business Support System (BSS).*
- *Network Slice Management Function (NSMF), which manages the life cycle of the end-to-end slice across the network domains: Radio Access Network (RAN), 5G Core network, and the transport network.*
- *Network Slice Subnet Management Function (NSSMF), which manages the lifecycle of the Network Slice subnets within a network domain and applies the NSMF's life cycle management commands (For example, instantiate, scale, heal, terminate). There can be more than one NSSMF in a network.*

The Network Slicing management layers enable users to plan, design, and instantiate end-to-end network slices across the Radio Access Network (RAN), the mobile core network and transport domains.

The cross-domain network slice orchestration (NSMF) must support multi-vendor capabilities, it should be able to integrate with any 3rd party domain level slice orchestrators (NSSMF). NSSMF must be vendor agnostic. Core network slices can be created as either shared or dedicated according to the slice policy.

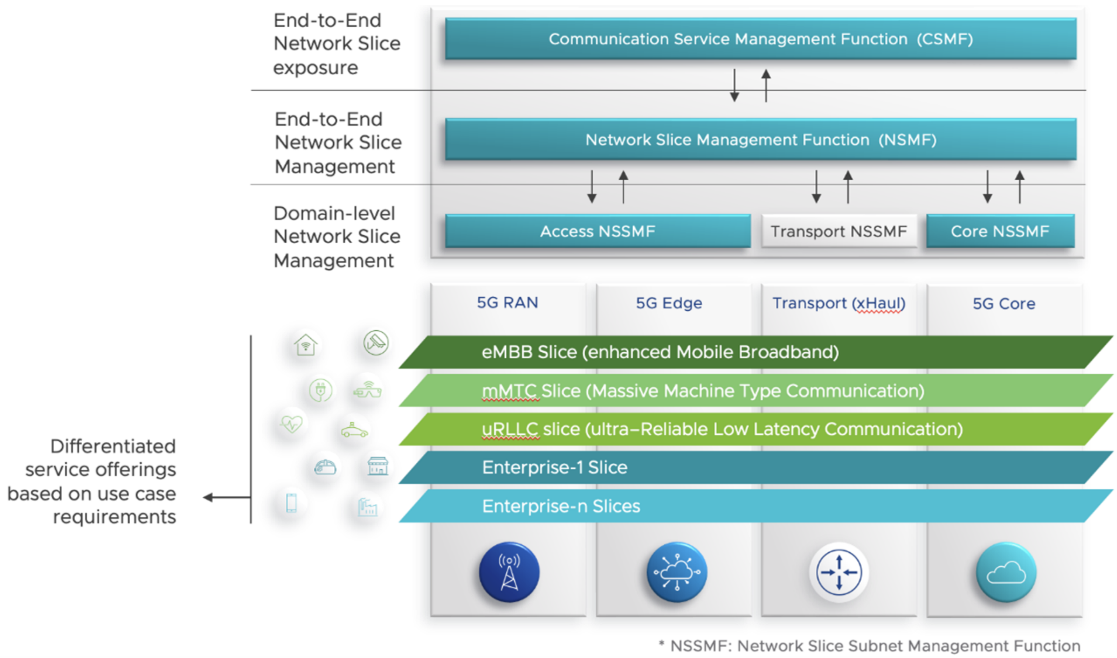
There will be a mix of infrastructure in the network, including Physical Network Function (PNF)s, Virtual Network Function (VNF)s, Cloud-Native Network Function (CNF)s. The orchestration solution should support various permutations of network functions, so the network slicing will not depend significantly on a specific component of the underlying network.

The orchestration framework supports open APIs to provide abstraction at domain and cross-domain level. The API driven model helps easy integration with 3rd party software components hence enables the best of breed orchestration.

Adding automation to the network slicing orchestration is very essential. It enables service providers to unify domains automation and close the gap between the endpoint consumed services and the required network resources, from a physical or a cloud infrastructure. Both enterprises (slice consumers) and service providers (slice providers) need a highly efficient way to deliver and manage network slices at scale including ordering, instantiating, provisioning, monitoring, and managing the lifecycle of network slices.

The dynamic network slice managed by the end-to-end orchestration is provided by Topology and Orchestration Specification for Cloud Applications (TOSCA) template model for reusable tasks and closed-loop control in network

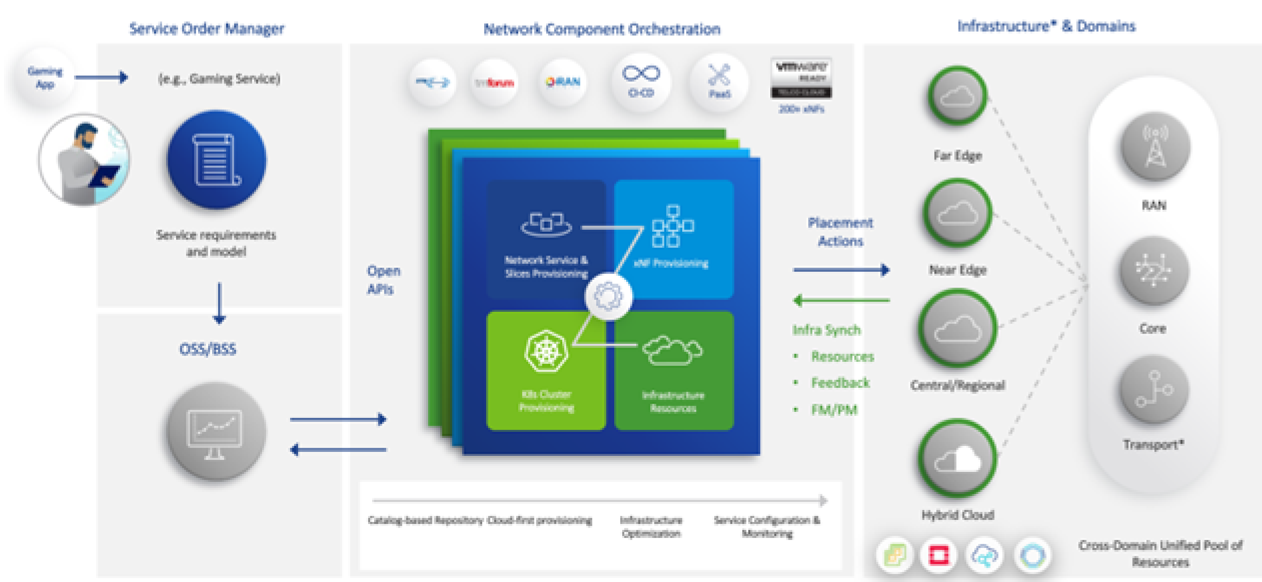
Figure 2.10 3GPP Standard-compliant Network Slicing Management Architecture [7]



slicing. To instantiate or modify the network slice, the network slice template model needs to be created. To realize the automation, a catalog of network slice template is needed which helps quick instantiation and provisioning. The catalog also provides an overview of network services, functions, resources, etc.

Figure 2.11 is one example of automation of cross-domain orchestration from VMware.

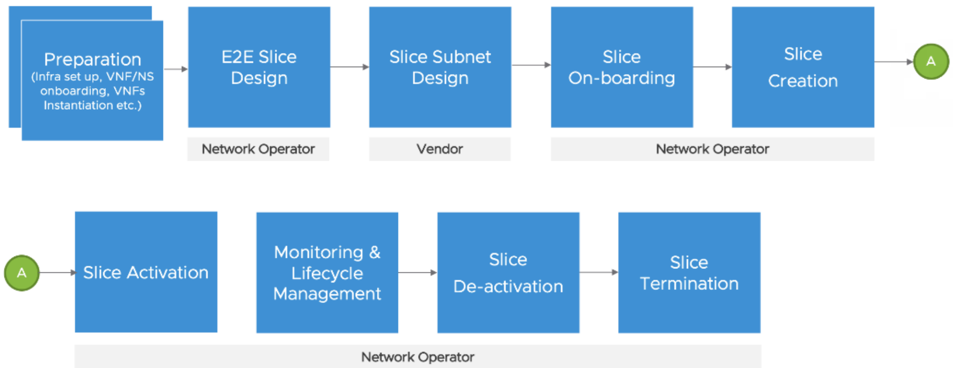
Figure 2.11 Automation Workflow of Cross-domain Orchestration [7]



The orchestration and automation equip user with a streamlined process for the whole slicing cycle. Figure 2.14 shows the steps in user’s point of view. To prepare an E2E network Slice, the VNFs and NSs should be onboarded to the platform and instantiated, the underlying resources are allocated in NFVI. The design of an E2E network slice includes setting up slice profile template, SLA, slice configuration, subnet design (I.e. RAN, Core and transport), etc. The Orchestrator prepares the required template for network slice manager to serve the NSs and VNFs. The profile template is specific for the slice type like eMBB, uRLLC to differentiate the service offerings. The slice configuration includes the configuration of KPIs, such as, throughput (per slice / per UE), delay tolerance, connections per slice, packet size, user density, etc.

Figure 2.12 shows the steps involved from a user perspective:

Figure 2.12 Network Slicing User Journey



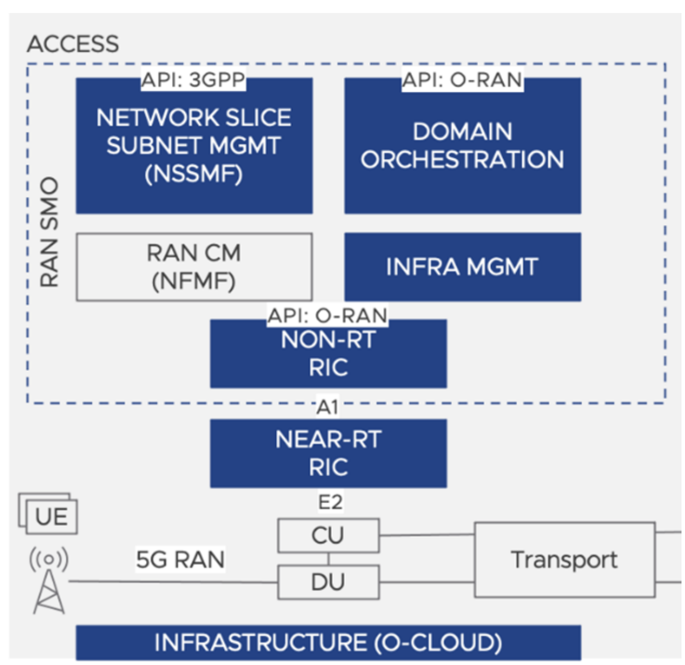
Furthermore, the orchestration prepares for the Network Slice service delivery. NSaaS (Network Slicing as a Service) service delivery model allows the operators to provision customized network slices to individual customers, and eventually enable these customers to gain access to some network slice management capabilities. It is up to the operator to decide on which specific management capabilities are made available to each customer, typically exposed through customer-facing APIs. Verticals can adopt the network slicing through NSaaS platform model, which can help alleviate operational complexities that may be faced by verticals when employing network slicing.

The RAN can have its own slice orchestration which enables automation of the slice Life Cycle Management (LCM) for RAN. In the case of O-RAN, the RAN domain orchestrator can be placed in SMO (Service and Management Orchestration), along with the non-RT (non-realtime) RAN Intelligent Controller (RIC), as shown in the example below. The RAN orchestrator can provide the function of multi-vendor RAN Element Management System (EMS). It is responsible for fault, configuration, accounting, performance, and security (FCAPS) over O-RAN and proprietary interfaces. It also exposes RAN data via APIs for use by apps or other entities, helps provisioning RAN configuration changes, supports policies, as well as other general SMO functions.

The fundamental role of the Non-RT RIC is to gather long term slice related data through interaction with the SMO framework and apply AI/ML based approaches interworking with the Near-RT RIC to provide innovative RAN slicing use cases. For this purpose, the Non-RT RIC should be aware of RAN slices and their respective SLAs through SMO. In addition, the Non-RT RIC may retrieve enrichment information from 3rd party applications enabling advanced RAN slicing technology to be applied in O-RAN framework.

The Near-RT RIC enables near-real-time RAN slice optimization through execution of slicing related xApps (applications designed to run on the near-RT RIC, which may be provided by third parties) and communicating necessary parameters to O-CU and O-DU through the E2 interface. Deployed xApps may utilize either AI/ML based models or other control schemes which can further be guided by A1 policies that are generated by Non-RT RIC. While RIC provides intelligent operation, the RAN orchestrator facilitates the actions, for example, when CM parameters need to be changed.

Figure 2.13 Example of Network Slicing Building Blocks for O-RAN, adapted [7]

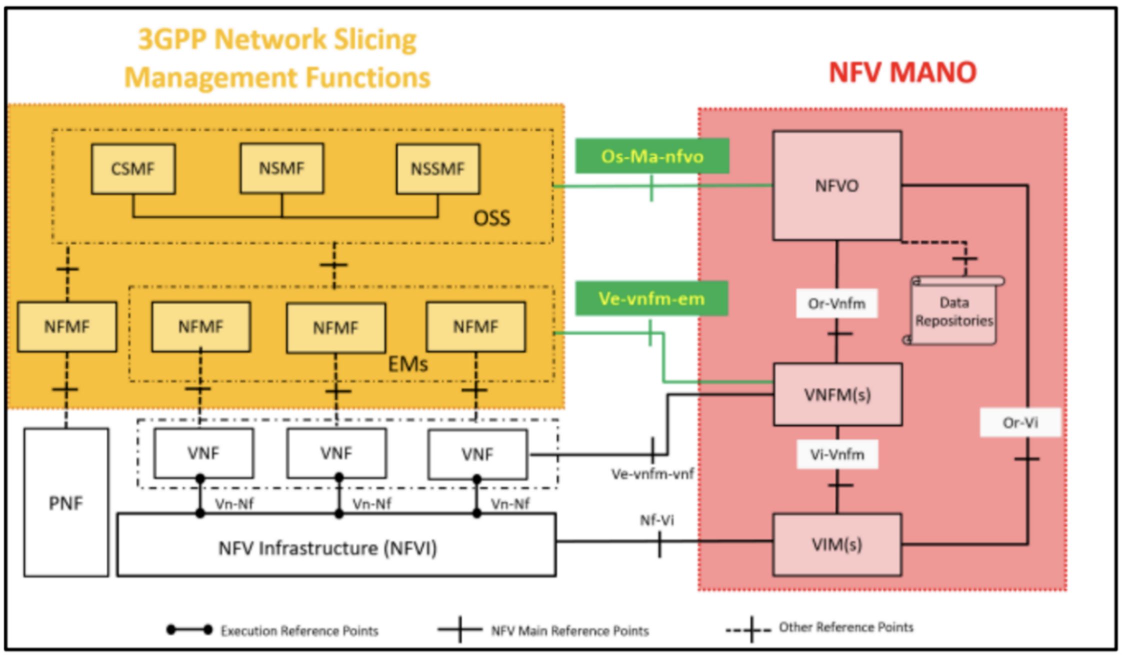


2.6 NFV in Network Slicing

As is commonly realized, 5G is evolving on the basis of virtualization. Network Function Virtualization (NFV) architecture advances are leading this effort and is the enabler for Network Slicing. NFV has a wide range of potential advantages like reduced CapEx/OpEx, virtualization, orchestration, dynamic scaling, and automation which are considerations for network slicing commercialization. The orchestrator in the NFV framework helps Network Slicing creation and life cycle management.

NFV is an architectural framework that is developed by the ETSI NFV Industry Specification Group. There can be different ways vendors design or integrate slicing management with OSS/BSS and EMS. Figure 2.14 below shows how Network Slicing Management, OSS/BSS and NFV framework work together:

Figure 2.14 Example of Network Slice Management in an NFV Framework [8]



In this NFV framework, NFVI provides the physical compute, storage, and networking hardware that hosts the VNFs. Each NFVI block can be thought of as an NFVI node and many nodes can be deployed and controlled geographically. The hypervisor provides a virtualization layer that allows for workloads that are agnostic to the underlying hardware.

The Management and Orchestration (MANO) functional block is responsible for the management of all the resources in the infrastructure along with the orchestration and life cycle management of VNFs and it has three components: NFV Orchestrator, Virtualized Infrastructure Manager (VIM), VNF Management/Manager. The details of these three components can be found in ETSI document.

Virtual Network Functions (VNFs) are software applications that deliver network functions, and they are deployed as virtual machines (VMs) or containers. They have often been considered to be the next step for telecommunications providers in their digital transformation from the physical network functions (PNFs) of legacy network appliances.

MANO in NFV framework interacts with slicing management. Network slicing requirements are gathered and addressed at the top layer of the NFV architecture, such as with OSS/BSS. MANO has an Orchestrator who works with Network slice manager to address the requirements and manages the dynamic aspects of infrastructure and services. A NS (Network Service) can be created by composing a group of VNFs.

After the preparation work is done and the slice design is finished, the E2E slice is created and ready to be activated. Once the network slice is activated, various operations are included, such as monitoring, scaling, upgrade, update, healing, termination, etc. MANO (NFVO, VNFM and VIM) is involved in the slice life cycle management. ETSI document contains detailed workflow of those procedures.

NFVO uses VNFM for the life cycle management of slices. NFVO translates the actual infrastructure or application requirements. NFVO sends request to VNFM. VNFM translates the request from NFVO into actual commands that VIM understands. VIM executes the command to manage the NFVI.

In the slice operation phase, slice instances can be shown to consumers, the slice performance is monitored. A slice can be deactivated or decommissioned, at that time, all VNFs, NSs are deactivated and NFVI resources are deallocated. Through the whole slice cycle, the resources for the slice consumer are allocated dynamically based on the needs and deallocated once the cycle finishes. NFV framework helps achieve that goal.

2.7 Final Considerations on Initial Commercialization

In the previous chapters, we attempted to build a ‘cookbook’ for network slicing based new services and applications that can be commercialized in 5G SA. One of the main ingredients is simply the differentiation on air interface, and features like prioritization and resource partition offer good mechanisms for operators to create various new services and applications. However, it is worth mentioning the challenges associated with assessing the economics, impact as well as market appeal, and we can list here some important ones:

- *creating radio resource partitions impacts existing capacity and must be accounted for in economic assessments, and additionally some of the use cases using partitioning (for ex. for nationwide category slices) will be difficult to scale*
- *(higher) prioritization delivers noticeable improvements only in traffic contention situations, which is not an issue yet on the new 5G infrastructure, so customers might not consider it a value added service*
- *Without Application-based slicing, at least until OS vendors and CSPs agree on a viable path, slicing mechanisms can only be offered per entire DNN (currently common for all data services), or for an enterprise group of applications (only for business owned devices)*

Other ‘recipes’ can add increased security on radio, backhaul and core, which could target segments such as government services or businesses. Complemented by network slicing capability to connect to a dedicated (private) core, and possibly including the radio differentiation mentioned previously, such combinations could become a good target for initial commercialization for businesses, specifically for the industries that, according to the Arthur D. Little study mentioned in the introduction section, could have the highest revenue potential for slicing application: healthcare and government services. In other words, the ones that would be willing to pay for such ‘preferential treatment’.

Ultimately, launching a slicing capable network, even with a limited scale orchestration functionality is a significant investment for a CSP so it should be a business decision that has to rely on a technological and economic assessment. An alternative though would be to launch a slice-based capability for the purpose of ‘testing the waters’ or enticing the market to slicing services. As mentioned in the introduction, this is a time where a ‘killer’ slicing application could perhaps be needed to ignite the demand.

Considering all the various elements described in this white paper, network slicing remains in the early stages. Network slicing business realization is obviously a complex topic. With more developments in progress, or already happening, new business demands will continue to feed the discussion on commercialization. In the end we can only hope the information presented so far, even if it might not accurately predict a first launch, would at least help structure the conversation around the topic, and facilitate the best business and engineering decisions towards it.

3. Evolution of Network Slicing Technology

There are many efforts to address some of the challenges the current framework for slicing in the context of the growing number of use cases. In this chapter, we select a few that we consider important or with interesting potential to add strength to this new technology.

3.1 Standards Development

Standards development is a process that involves many individuals, companies, and often governments. A standards development organization (SDO) is an organization focused on developing, publishing, or disseminating technical standards to meet the needs of an industry or field. SDOs may use consensus-based or non-consensus-based processes and these efforts are critical to the development of drafts, frameworks, and specifications, that lead to widespread adoption of new technologies.

3.1.1 3GPP

In 2022, the 3GPP started an architectural study of remaining network slicing issues for Release 18:

- *Enhancements to roaming support by providing UE information on available network slices in the roamed to network*
- *Enhancements in network control of slice usage*
- *Considerations for non-overlapping areas of service with Tracking Areas or when they have a limited lifetime*
- *Enhancements in support of more than one Network Slicing Admission Control functions involved in enforcing a shared maximum allowed number of the UEs or Sessions for a network slice*
- *Improvements to service continuity when existing slice can no longer serve a session*
- *Registration to a network slice in a new tracking area after a rejection in a different tracking area of same registration area.*

One other useful development is the NWDAF (Network Data Analytics Function) in the control plane, which is responsible for collecting data on the load level of a NF or a network slice, playing a very similar role to that of the MDAF (management data analytic function) in the orchestration domain. NWDAF uses the data to produce a forecast that be used by the PCF to optimize its policies, such as: 1) the QoS parameters (for those services that can be provided at different QoS levels); 2) the access and mobility policies; or 3) the UE Route Selection Policy (URSP).

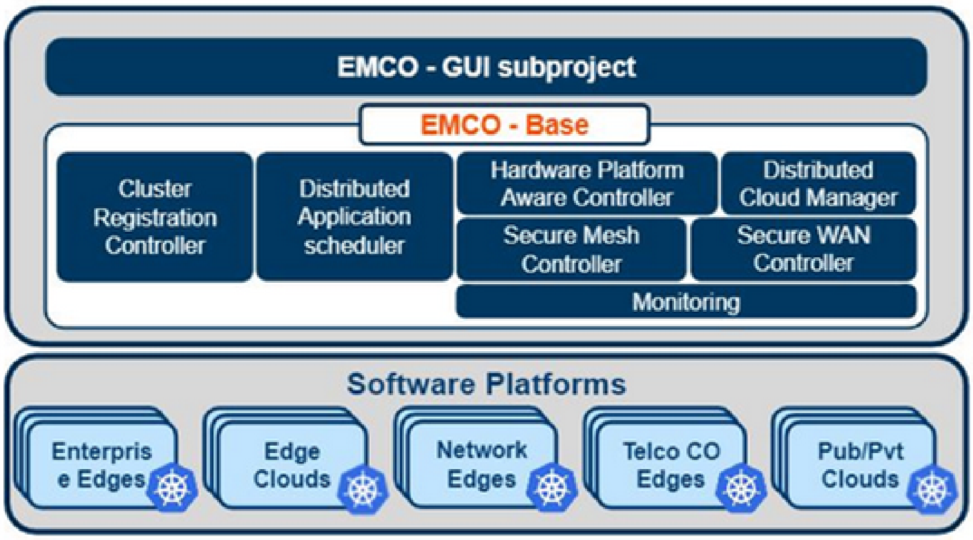
While the NWDAF module has been designed for the network core, a similar approach can be applied to the RAN. Although 3GPP has not yet proposed modules equivalent to NWDAF in the RAN, other initiatives such as the O-RAN alliance have taken this path. In the ORAN architecture, the SMO collects load information of flows or flow aggregates at the RAN level, the RAN Intelligent Controller (RIC) enables near real-time control of RAN elements/resources, and the RAN resource orchestrator handles the overall resources at the base station level. In this case, these short-term forecasts can be leveraged by the RIC to perform the optimization of the radio resources at a fine time granularity (in the order of hundreds of milliseconds) and by the RAN resource orchestration to update the resource and bandwidth allocation at larger timescales (up to the order of minutes).

3.1.2 EMCO - Linux Foundation

The latest project launched in Oct 2021 by Linux Foundation, Edge Multi-Cluster Orchestrator (EMCO) aims to create a universal control plane that helps organizations to securely connect and deploy workloads across public clouds, private clouds and edge locations, with end-to-end inter-application communication enabled [9].

Edge computing remains one of the most attractive models for enabling 5G promise of running applications at the edge. Since this is a shared infrastructure from application perspective, application of network slicing at edge is being considered for providing secure, agile and centrally managed environment for workload management. The EMCO is a Geo-distributed application orchestrator that intelligently places a tenant's workload onto one or more clusters. These clusters can be a public cloud cluster, an enterprise's private cloud cluster, an IoT edge site cluster, a 5G cell cluster, or a telco's CO cluster. The workload can be a simple application or a complex application that is composed of multiple simple applications, or it can be a network function in the form of container or VM.

Figure 3.1 Linux Foundation EMCO [9]



EMCO operates at a higher level than Kubernetes. It makes a decision on which clusters a workload should run, then it interacts with the Kubernetes API server and hands over that workload to the Kubernetes control plane. EMCO Supports multiple placement constraints: Affinity and Anti-Affinity; Platform capabilities; Latency; Cost. Through auto-configuring service mesh and security policy (NAT, firewall), it enables the communication between deployed workloads or between a deployed workload and an external service within or across clusters with/without mutual TLS. It provides application life cycle management including upgrades as well as comprehensive status monitoring of deployed applications.

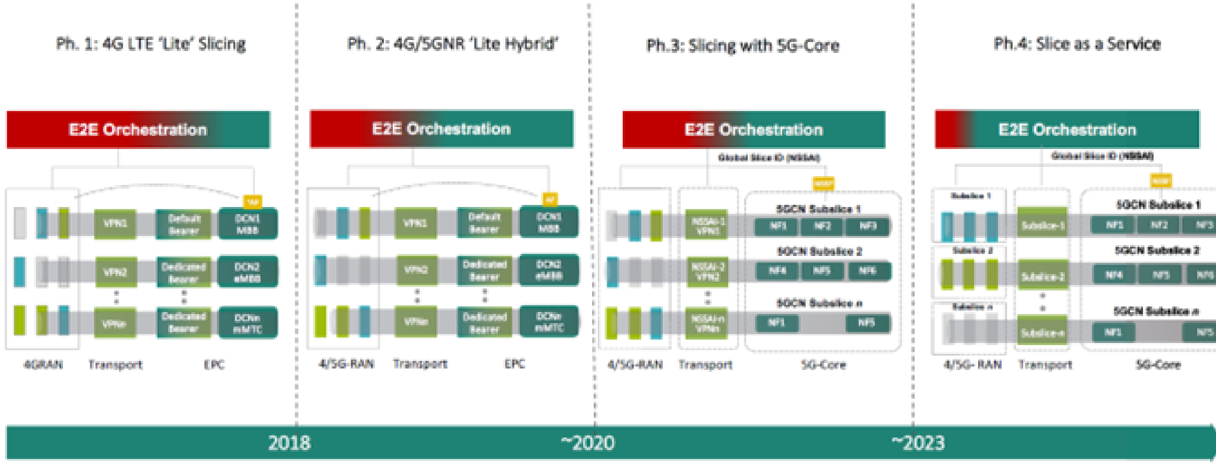
Main use cases targeted by EMCO range from enablement of dynamic network slicing, connectivity to edge applications and subscriber traffic steering. While EMCO does not itself contain a slice manager function (CSMF/NSMF), it works in tandem with slice managers to extend slicing capabilities to the network edge.

3.1.3 Telecom Infra Project (TIP)

The Telecom Infra Project (TIP) is an engineering-focused initiative driven by operators, suppliers, developers, integrators, and start-ups to disaggregate the traditional network deployment approach. The collective aim of the TIP community is to collaborate on new technologies, examine new business approaches and spur new investments in the telecom space.

The End-to-End Network Slicing (E2E-NS) Project Group in TIP is aimed at defining a commercially viable end-to-end network slicing ecosystem that can be deployed over fixed and mobile operator networks. The long-term objective of the E2E-NS Project Group is to identify use cases to be researched, developed, and demonstrated as proof-of-concept addressing key challenges for 4G and emerging 5G network slicing arenas. It aims to seize opportunities afforded by TIP membership to involve different players (telecom operators, equipment vendors, orchestration suppliers, application providers, and network integrators and technologists) within the ecosystem. Further, the group looks to build on identified use cases to help carrier trials and deploy real-world network slicing offerings at scale.

Figure 3.2 E2E-NS architecture roadmap - TIP

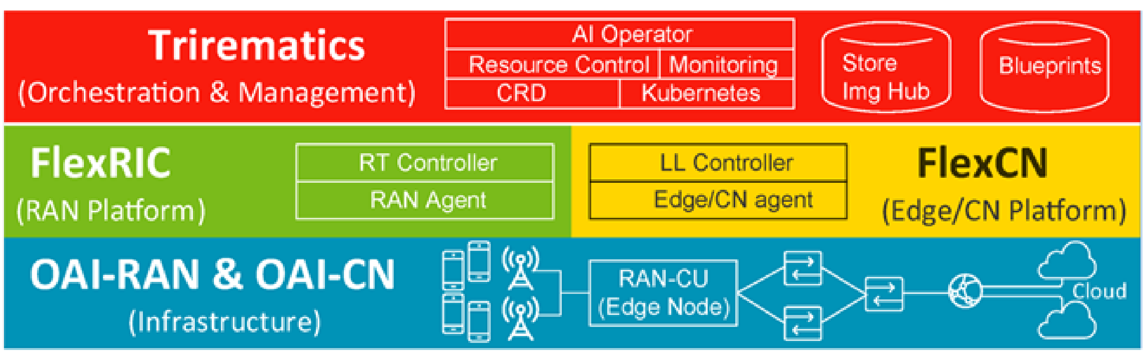


3.1.4 MOSAIC5G

The newly created MOSAIC5G (M5G) PROJECT GROUP within the Open Air Interface (OAI) consortium aims to transform RAN and core networks into agile and open network-service delivery platforms. Such a platform allows for exploring new use-cases of interest to different vertical industries. Mosaic5G introduces the world's first ecosystem of 5G R&D open source platforms ranging from the centralized network control to the mobile edge network deployment.

MOSAIC5G is developing a set of extendable control and orchestration frameworks with extendable APIs on the top of Open Air Interface (OAI) RAN and CN to enable fine-grained monitoring and programmability of the underlying network infrastructure.

Figure 3.3 Open Air Interface's MOSAIC5G



One of the components being developed is FlexRIC, which is a software suite that has a built-in Service Model for monitoring and network slicing, that can be easily customized and extended to fulfill the diverse 5G use cases. FlexRIC's Application Protocol and Service Model are encoding and decoding agnostic. It also supports the creation of new Service Models "à la carte" to satisfy specific, yet not standardized, use cases. FlexRIC has two main components, a RAN agent that allows for interfacing with the radio stack, and a real-time (RT) controller [10].

3.2 From Vertical Slicing to Horizontal Slicing

The E2E network slicing discussed so far has involved vertical slicing, as it transforms one big network into multiple parallel vertical E2E slices, like shown in Figures 2.10 and 2.12. Vertical slicing allows resource sharing between different services and applications to ensure QoS.

There is another kind of slicing called 'horizontal slicing' which is a step further, as it enables resources (including over-the-air resources) sharing among network nodes and devices, such as high capability network nodes/devices that share their resources to enhance the capabilities of less capable network nodes/devices. It has numerous use cases and offers incredible benefits to both consumers and enterprises [11] [12] [13] [14] [15] [16]. For example, AR, VR and head-mounted displays (HMDs) can have enormous applications with horizontal network slicing supported by borrowed processing powers from tethered connection to a robust device.

In 5G networks, as more user traffic is generated at the edges than core, more capacity scaling is needed at the edges. Horizontal slicing is designed to accommodate capacity scaling and enable Edge computing and offloading. MEC can be an enabler of horizontal slicing, as it enhances the processing capacity of a mobile device independent of its physical limitations. This will eventually build a fresh generation of moving underlay networks.

Table 3.1 summarizes the comparison of vertical slicing and horizontal slicing.

Table 3.1 Comparison of Vertical and Horizontal Slicing

	Vertical Slicing	Horizontal Slicing
Resource sharing	Among different services and applications	Among network nodes and devices
Traffic flow transition	Between the Core network and the terminal devices	Locally between two ends of the slices
Function on nodes	Similar functions on each of the network ntoes along slices	New functions can be created at a network node

Vertical slicing could be the early implementation of slicing but preparing for horizontal slicing as the next phase is needed. Both approaches can be implemented simultaneously, and they can work together.

Figure 3.4 is a picture to show the concept of horizontal slicing. The slices contain tiers. Each tier is comprised of a group of computation nodes. For example, tier1 nodes can be very low power wearable devices, tier2 nodes can be personal computers, tier3 nodes can be Edge nodes or Base Stations, etc. There can be a different number of tiers, and nodes in each tier can vary in different situations. The idea is that the higher tier has stronger computation power than the lower tier.

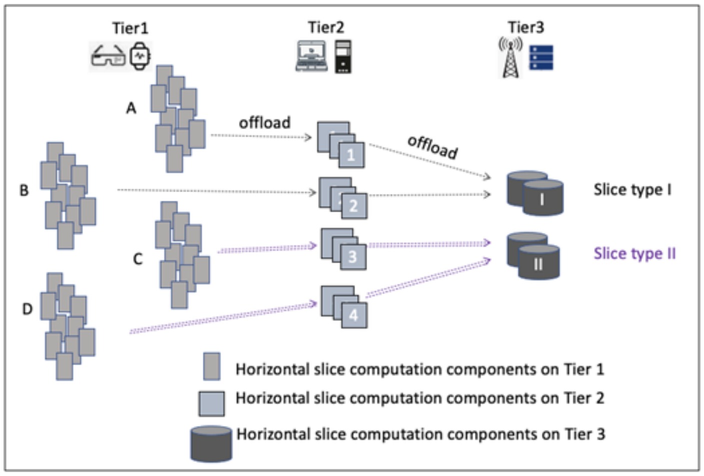
The computation from one tier can be offloaded to the next tier when the traffic demand is beyond what its own computation capability can handle. On the same tier, each slice performs a specific function, the computation capacity can scale up or scale down based on the traffic demand. A new function should be added to a node when a new slice is created.

A slice is based on the function. In Figure 3.4 there are two slice types and same slice type refers to the same slice template. For instance:

- *A-1-I, belongs to slice type I;*
- *B-2-I, belongs to slice type I;*
- *C-3-II, belongs to slice type II;*
- *D-4-II, belongs to slice type II.*

The components here in tier 1,2,3 can be physical nodes, or logical nodes (I.e., virtual machines) which perform a specific function.

Figure 3.4 An example of Horizontal Slicing Concept Illustration



3.3 Platform evolution

5G is cloud native and as the technology evolves, CSPs and enterprises are increasingly moving their workloads to the cloud. In the cloud migration path, many companies will use multi-cloud environments to distribute their workloads. The adoption rates of multi-cloud deployments are expected to continue rising.

However, developing applications in multi-cloud has many challenges. For instance, applications can run on different versions of Kubernetes, services from different cloud vendors create complexity, each cloud has unique APIs, operators have fragmented infrastructure, different monitoring, and management environments, etc. Building Network Slicing on top of a multi-cloud environment is even more challenging.

Evolving the platform to support multi-cloud is the way to solve the above problems, thus helping CSPs unlock new revenue streams. On top of that, in Network Slicing orchestration, adding an additional level of flexibility for the service providers to be able to automate across multi-cloud will make more scenarios and use cases than ever possible.

Successful network slicing orchestration supports lifecycle management for the creation, modification, and termination of individual services while also promoting the assignment of underlying resources.

The multi-cloud support platform offers a single abstraction across all domains and clouds which provides crucial value proposition for CSPs working with distributed, multi-cloud networks. Open APIs are supported in various interfaces which makes the VNFs onboarding, integration, and interworking of various components of E2E slices smooth. It enables a "single pane of glass" slice assurance for the whole slice life cycle. It helps reduce the complexity of multi-cloud operation. From a business point of view, the muti-cloud orchestration platform helps CSPs bridge different clouds with cost-saving and embrace best-of-breed technologies from different clouds including hardware and software.

Network Slicing supported on multi-cloud enables nearly unlimited use cases, which can include (but are certainly not limited to) autonomous vehicle operations, remote healthcare, smart energy, sports, gaming, and many more.

3.4 Network Slicing for Shared Networks and Multi-Tenancy

Network sharing is a cost-effective solution towards the deployment of 5G mobile networks. The step forward is full multi-tenancy thanks to NFV which helps broaden the road [17] [18].

3.4.1 Network Slicing for Shared Networks

Network slicing can be used by any or all involved service providers that are sharing a network in a multi-operator core network (MOCN) or a multi-operator RAN (MORAN) configuration. The service providers sharing a network may deploy a different set of network slices.

When multiple operators share RAN equipment, they typically use their own dedicated core networks. This makes RAN as the main entity which is shared and must provide for the different operator's network slicing services. Without network slicing, PLMN-ID alone controls the selection and steering of traffic to different core networks. With network slicing, the combination of PLMN-ID and Single Network Slice Selection Assistance Information (S-NSSAI) is used for identifying and steering traffic inside RAN. A shared RAN configuration provides for the separate configuration or parameter settings taking each service providers and slice into account.

3.4.2 Network Slicing for Multi-Tenancy

The evolution of network sharing towards full multi-tenancy, which is more on-demand, relies on virtualization mechanisms, such as NFV. This kind of network slicing will be realized by allocating not only network capacity, but also VNFs, computing resources, per slice tailored control/user-plane splits, shared network functions across different slices and RAT setting.

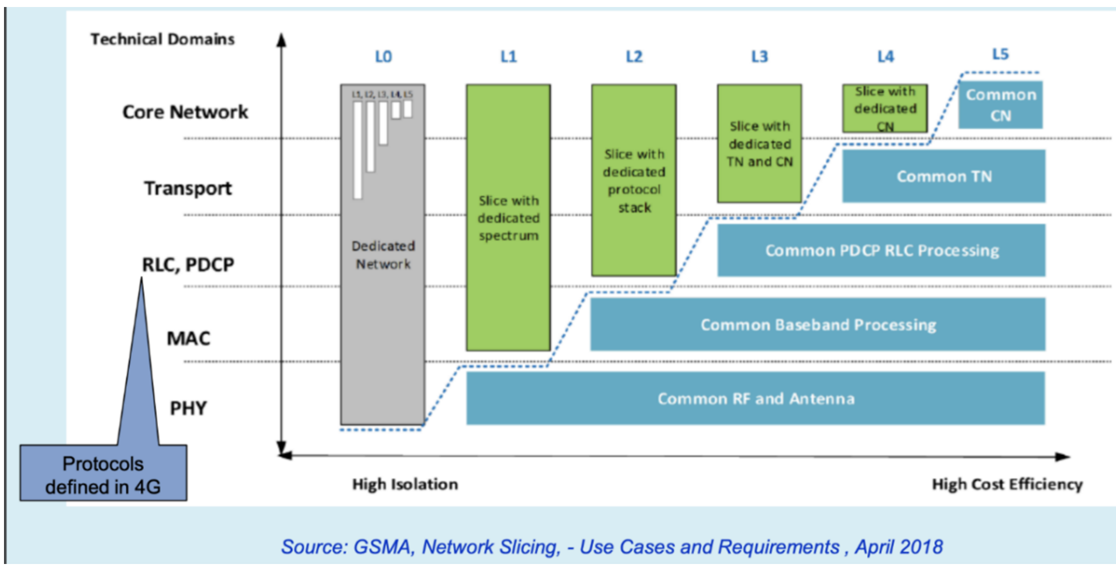
In multi-tenancy scenario, the tenants can be different customers, or can be different organizations with the same customer. The tenants share the same platform. A multi-tenant architecture is based on central administration, and involves a common code application, operating common instance(s) of applications for multiple tenants. In addition, it also secures private data for each tenant from others.

There must be some coordinator who is responsible for the aggregation of different slice requests from individual tenants, to facilitate resource allocation, perform admission control, monitors performance, usage, handles charging, etc. Research has been started in this direction, and the concept of Network Slice Broker was introduced.

Multi-tenancy is cost-effective for both consumers and vendors. Especially for small business consumers, the cost-saving benefit is very attractive. Besides that, multi-tenancy has other benefits. As SaaS-based applications continue to grow, multi-tenancy architecture allows vendors to make system updates for all tenants at once instead of taking care of each tenant individually. Multi-tenancy architecture is easy scalable.

There are different levels of isolation for multi-tenancy network slicing, in terms of subnet, radio spectrum, protocol stack, etc. Different levels of isolation will have different costs. Below Figure 3.5 shows the cost comparison. Most expensive mode will be dedicated RAN according to GSMA.

Figure 3.5 Cost Comparison of Network Slicing with Shared Network [12]



In a multi-tenancy environment, there are multiple ways to achieve tenant isolation. For example, data access control, separate security domain, VMs or containers/namespaces, are just a few.

Below is one example of using VMs to separate two tenants. The two tenants both need to access VNF-x and VNF-y and they share the same slice with the same slice configuration. However, they are allocated different VMs for these VNFs, and two vLANs, vLAN1 and vLAN2, are created for them separately, so the tenants can access the same VNFs through different vLANs.

The implementation of multi-tenancy involves distribution and allocation of slicing instances among tenants. There are multiple approaches of multi-tenancy, illustrated in figure 3.6.

- *Multiple Tenants accessing multiple network slices. The example on the top in below picture 'INDUSTRY' sector, where different slice instances are allocated for multiple tenants' use cases. There can be cases when the multiple tenants access the same slice.*
- *A tenant accessing multiple network slices. The example in the middle in below picture 'HEALTH' sector, where different slice instances are allocated to a single tenant's use cases.*
- *Each tenant accessing respective network slice. The example at bottom in below picture 'TRANSPORT' sector, where one slice instance is allocated for one tenant.*

Figure 3.6 scenarios can be configured based on the customer requirements.

Figure 3.7 scenarios will be designed and as per the infrastructure, requirements and demands of the tenants. Proper Orchestration will serve the needs of scale up, scale down, scale in, and scale out.

Figure 3.6 Example of Using vLANs to Isolate Tenants

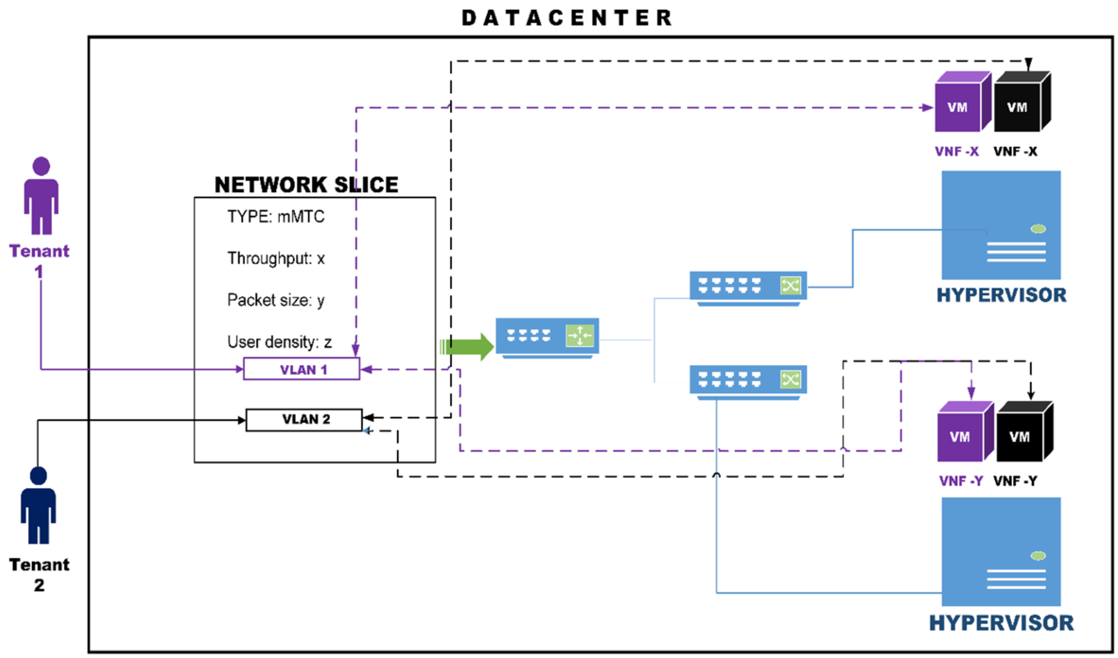
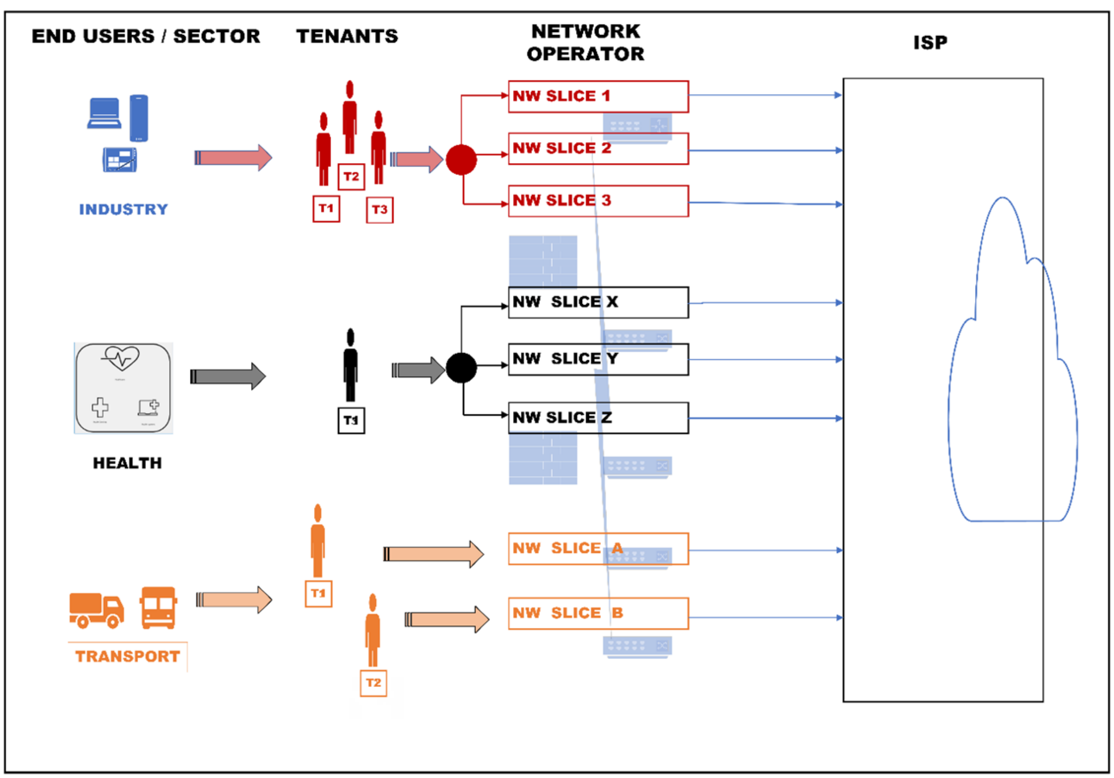


Figure 3.7 Multi-tenancy Deployment Approaches



Conclusion

After the introduction of Network Slicing in 3GPP standards in Release 15, more developments and refinements were added in the next release, and the on-going Release 17 will add even more techniques and mechanisms to enhance its capabilities and make it more mature. Recognizing the importance of network slicing, many industry projects and alliances are building on it, creating implementation frameworks that will expand this technology and make it more available to operators.

This white paper brings together the currently available technical aspects of network slicing into practical scenarios and use cases and applications that could be deployed by this new generation of wireless with the purpose of supporting the business making decision process. As we pointed out, there are many factors and variables that must be considered when deciding on the initial commercialization. But overall, there is no doubt that network slicing will be an important tool in enabling 5G networks to deliver on the new demands in the communication industry of today.

In the next few years as 5G Standalone networks proliferate globally, it is possible that slicing will become a fundamental building block that could create new services and enable new experiences. As more developments occur in the slicing domain, the added synergy between slicing and virtualization will create even more advanced capabilities and help to enable the realization of the fourth industrial revolution.

Acronyms

3GPP: 3rd Generation Partnership Project

BSS: Business Support System

CM: Configuration Management

CNF: Cloud-native Network Function

CSMF: Communication Service Management Function

CU: Centralized Unit

DECOR: Dedicated Core

DSCP: Differentiated Services Code Point

DU: Distributed Unit

EMCO: Edge Multi-Cluster Orchestrator

ETSI: European Telecommunications Standards Institute

FWA: Fixed Wireless Access

GSMA: Global System for Mobile Communications

IETF: Internet Engineering Task Force

LCM: Life Cycle Management

MANO: Management and Orchestration

MDM: Mobile Device Management

MEC: Multi-access Edge Computing

MOCN: Multi-Operator Core Network

MORAN: Multi-Operator Radio Access Network

NF: Network Function

NFV: Network Function Virtualization

NFVI: Network Function Virtualization Infrastructure

NS: Network Service

NSaaS: Network Slicing as a Service

NSMF: Network Slice Management Function

NSSAI: Network Slice Selection Assistance Information

NSSMF: Network Slice Subnet Management function

NST: Network Slice Template

O-RAN: Open Radio Access Network

OSS: Operations Support System

PLMN: Public Land Mobile Network

PNF: Physical Network Function

QOS: Quality Of Service

RRM: Radio Resource Management

RU: Radio Unit

S-NSSAI: Single Network Slice Selection Assistance Information

TIP: Telecom Infra Project

TOSCA: Topology and Orchestration Specification for Cloud Applications

URSP: UE Route Selection Policy

VNF: Virtual Network Function

VONR: Voice Over New Radio

References

[1] “ONAP Framework,” [Online]. Available: https://www.onap.org/wp-content/uploads/sites/20/2020/12/ONAP_Technical-Overview_120720.pdf.

[2] “O-RAN Alliance,” [Online]. Available: https://www.o-ran.org/.

[3] “O-RAN Specifications,” [Online]. Available: https://www.o-ran.org/specifications.

[4] G. Report, “GSA Inforgraphic, End of March 2022,” [Online]. Available: https://gsacom.com/paper/gsa-infographic-end-march-2022.

[5] G. Report, “5G Snapshot, March 2022,” [Online]. Available: https://gsacom.com/paper/5g-snapshot-march-2022-member-report.

[6] Ericsson, “Network Slicing,” [Online]. Available: https://www.ericsson.com/en/network-slicing.

[7] “Network Slicing for VMware Telco Cloud Automation: End-to-End Service Orchestration and Network Slicing,” VMware, [Online]. Available: https://blogs.vmware.com/telco/network-slicing/.

[8] C. Tranoris, “Openslice: An opensource OSS for Delivering Network Slice as a Service,” Electrical Engineering Dept., University of Patras, [Online]. Available: https://arxiv.org/pdf/2102.03290.pdf.

[9] EMCO, “Edge Multi-Cluster Orchestrator,” [Online]. Available: https://project-emco.io/.

[10] OperAirInterface, “OpenAirInterface MOSAIC5G Project Group,” [Online]. Available: https://openairinterface.org/mosaic5g/.

[11] A. R. Dimitris Mavrakis, “The Evolution of Network Slicing,” Intel, [Online]. Available: https://builders.intel.com/docs/networkbuilders/the-evolution-of-network-slicing.pdf.

[12] E. Borcoci, “End-to-end, multi-domain and multi-tenant aspects in 5G network slicing,” IARIA, [Online]. Available: https://www.iaria.org/conferences2018/filesICSNC18/EugenBorcoci_Tutorial_E2EMultiTenantMultiDomain.pdf.

[13] M. Mazumdar, “5G Network Slicing: Growth and Evolution,” STL, [Online]. Available: https://www.stl.tech/blog/5g-network-slicing-the-growth-and-evolution/.

[14] O. Dharmadhikari, “Network Slicing: Building Next-Generation Wireless Networks,” [Online]. Available: https://www.cablelabs.com/blog/network-slicing-building-next-generation-wireless-networks.

[15] G. W. A. (. P. U. M. Qian (Clara) Li, “An end-to-end network slicing framework for 5G wireless communication systems,” Intel, [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1608/1608.00572.pdf.

[16] S. A. M. N. Asrar Ahmed Baktayan, “Fog Computing for Network Slicing in 5G Networks: An Overview,” Journal of Telecommunications System & Management, [Online]. Available: https://www.researchgate.net/publication/329094055_Fog_Computing_for_Network_Slicing_in_5G_Networks_An_Overview.

[17] X. C.-P. V. S. Konstantinos Samdanis, “From Network Sharing to Multi-tenancy: The 5G Network Slice Broker,” NEC Europe Ltd, Germany, [Online]. Available: https://arxiv.org/pdf/1605.01201.pdf.

[18] Parallel Wireless, “Multi-Tenant and Sharing (MORAN and MOCN),” [Online]. Available: https://www.parallelwireless.com/products/multi-tenant-and-sharing

Acknowledgments

5G Americas' Mission Statement: 5G Americas facilitates and advocates for the advancement and transformation of LTE, 5G and beyond throughout the Americas.

5G Americas' Board of Governors members include Airspan Networks, Antel, AT&T, Ciena, Cisco, Crown Castle, Ericsson, Intel, Liberty Latin America, Mavenir, Nokia, Qualcomm Incorporated, Samsung, Shaw Communications Inc., T-Mobile USA, Inc., Telefónica, VMware and WOM.

5G Americas would like to recognize the significant project leadership and important contributions of group leaders Helen Zeng of VMware, Adrian Singereanu and Suresh Thanneeru of T-Mobile, along with many representatives from member companies on 5G Americas' Board of Governors who participated in the development of this white paper.