Network Slicing for 5G Networks & Services

November 2016

TABLE OF CONTENTS

1.0 INTRODUCTION

5G cellular systems are expected to enable a major digital transformation that will provide people, businesses and governments with unprecedented capabilities to share information. The industry consensus is that 5G should be known not only for its cutting-edge radio access technologies, but also for the way it integrates cross-domain networks so operators can provide networks on a need-for-service basis.

5G will enable new verticals, new services and new business models that aren't possible or practical with 4G and other legacy mobile technologies. Examples includes wearables for advanced telemedicine applications, virtual/augmented reality, UHD video and machine-to-machine (M2M) applications that require single-digit-millisecond latencies such as driverless cars.

5G technologies aim to provide an end-to-end infrastructure capable of delivering a consistent user quality of experience in a heterogeneous environment across a wide variety of use cases. To achieve this goal, 5G must be able to support very high bit rates, high vehicular speeds, low latencies and high device densities. One example is Internet of Things (IoT) applications, where there may be thousands of 5G devices within a single acre, all requiring single-digit-millisecond latencies and other capabilities that legacy mobile technologies can't support well or at all.

Designing a network that can simultaneously support both a wide variety of use cases and demanding performance requirements, all with a single set of standard network functions, would be extremely complex and prohibitively expensive. The alternative "network slicing," which is considered to be key for meeting 5G's diverse requirements, including future-proof scalability and flexibility.

The network slicing concept enables the network elements and functions to be easily configured and reused in each network slice to meet a specific requirement. The implementation of network slicing is conceived to be an end-to-end feature that includes the core network and the RAN. Each slice can have its own network architecture, engineering mechanism and network provisioning.

Existing cellular network architectures are relatively monolithic, with a transport network that facilitates mobile traffic to end devices. They are not flexible enough to support wider range of performance and scalability requirements. Network slicing would allow 5G networks to be sliced logically into multiple virtual networks. Each slice can be optimized to serve a specific vertical application to efficiently support network services, thus providing high degree of flexibility in enabling several use cases to be active concurrently. This is already a well understood methodology in the wireless industry in some limited environments, such as software-defined core networks.

Network slicing leverages the latest innovations in cloud mobile access and core. Combining cloud technologies with the capabilities of software defined networking (SDN) and network function virtualization (NFV) provides the necessary tools to enable network slicing. Virtualization technologies provide a key foundation for network slicing by enabling use of both physical and virtual resources to create the service they are designed for. It is envisioned that this trend will continue, where virtualization technologies will be applied across RAN and portable services and across portable devices to wearable devices.

Network slicing implementation is end-to-end from the core through the RAN. In the core, NFV and SDN virtualize the network elements and functions in each slice to meet its own requirement. In the RAN, slicing can be built on physical radio resources (e.g., transmission point, spectrum, time) or on logical resources abstracted from physical radio resources.

The commercial introduction of NFV/SDN (Network Function Virtualization/Software Defined Network) based 4G EPC networks is now occurring and is expected to grow tremendously over the next several years. This will allow network slicing to enable much more flexible instantiations of networks that can be designed to meet the specific needs of the applications, services and operator business models. Both NGMN[1] and 3GPP[2] have been developing the definition and use cases for network slicing so that the SDOs (Standards Development Organizations) can provide detailed studies to understand the features and functionalities that will be required for network slicing beyond what is already defined in 3GPP Rel-13 and ETSI NFV.[3]

This white paper intends to describe the concept of network slicing, explore an end-to-end 5G system framework to build customized network slices, and discuss the application of network slicing to air-interface technologies. This paper initiates the discussion and development on the long-range technology roadmap and solutions for E2E network slicing in 5G and beyond. Furthermore, the white paper is organized as follows: section 2 discusses the role of network slicing, the basic 5G requirements for network slicing and the benefits to the operators. Section 3 provides a description of an end-to-end network slicing system architecture. The network slicing framework in CN and RAN are detailed in Sections 4 & 5 respectively. Section 6 provides a complete overview of the operational aspects of network slicing is presented. A discussion of the network slicing opportunities for operators is provided in section 7. Finally, section 8 summarizes the conclusions.

2.0 ROLE OF NETWORK SLICING IN 5G

As the work on 5G began in 3GPP, it quickly became evident that where the 4G network was tailored for the smartphone, the 5G network needs to support a variety of device types. In 4G networks, extra effort and functionality needed to be added to support, for example, very small M2M devices. The 5G system must be capable of supporting a diverse set of devices, from the very smallest to the most powerful, in an efficient manner. This means that it must be possible to configure the 5G system to match the unique needs of both each device and the applications running on that device. These needs are succinctly described in 3GPP's initial set of network slicing requirements.

2.1 5G REQUIREMENTS ON NETWORK SLICING

The requirements for 5G network slicing are currently in a proposed state and are listed in clause 5.2.3 of 3GPP Technical Report 22.891.[4] These requirements are considered very stable at this point and will eventually be placed in a Technical Specification. Although there are only eight requirements listed as of version 14.0.0, they provide a strong, high-level view and direction for the next generation (NexGen) work of 3GPP with respect to network slicing:

- The operator shall be able to create and manage network slices that fulfil required criteria for different market scenarios
- The operator shall be able to operate different network slices in parallel with isolation that for example, prevents one slice's data communication to negatively impact services in other slices
- The 3GPP system shall have the capability to conform to service-specific security assurance requirements in a single network slice, rather than the whole network

[1] Description of Network Slicing Concept by NGMN Alliance. Public version.
[2] 3GPP TS 22.891 v14.0.0, "Feasibility Study on New Services and Markets Technology Enablers; Stage 1", Release 14.
[3] ETSI GS NFV 002 v1.2.1 (2014-12).
[4] Ibid.

- The 3GPP system shall have the capability to provide a level of isolation between network slices to confines a cyber-attack to a single network slice
- The operator shall be able to authorize third parties to create, manage a network slice configuration (e.g., scale slices) via suitable APIs, within the limits set by the network operator
- The 3GPP system shall support network slice elasticity in terms of capacity with no impact on the services of this slice or other slices
- The 3GPP system shall be able to change the slices with minimal impact on the ongoing subscriber's services served by other slices: specifically, of new network slice addition, removal of existing network slice or update of network slice functions or configuration
- The 3GPP System shall be able to support end-to-end resource management for a network slice

What these requirements describe is an efficient, powerful, flexible system that operators can tailor to their specific needs. Using building block functions that can be assembled in a variety of ways, operators can assemble the functionality needed to efficiently support different needs, such as IoT devices that never move, smartphones and corporate devices requiring very secure VPN services.

2.2 HOW NETWORK SLICING CAN FULFILL THESE REQUIREMENTS

Network slicing, in its simplest description, is the ability to tailor a set of functions to optimize use of the network for each mobile device. All of the functionality needed, but only the functionality needed, is assembled in a way that optimizes that device's ability to find the correct network, access the network efficiently and securely, and be attached to the core network with the set of functionality needed by that device.

A device that is simple and requires only access to a single slice—for example, a water meter reader—can find a radio interface that is tailored to very small, infrequent messages. That RAN attaches the device to a simple, efficient authentication process that results in the device being connected to a single core network slice that handles just water meter reader devices. As such, the connected functionality provides all that the device needs to accomplish its work, without the overhead of unneeded functions and features.

A device that is more complex and provides rich functionality, such as a smartphone, can find the radio interfaces that it needs, as well. The device can be authenticated and attached to a diverse set of network slices that are each tailored to a specific purpose: streaming video, voice calls, internet browsing, chatting and so on. Such devices may need to carry on high-quality voice and video calls, while at the same time continuing background browsing and upload/download functions. The independence of network slices indicated in the 3GPP requirements ensures that the high-quality video call that it will not be interrupted nor impacted by other background functions. Such services as video calling can be established with high security, while the security applied to other activities may use completely different methods.

But at the same time that these network slicing features are supporting very specific needs of devices and their users, they are also providing the ability to the network operators to deploy only the specific functions needed to satisfy their customers' needs. This ability to focus on 'what is needed' reduces investment in unnecessary features, saving operator's money and making them more competitive.

2.3 BENEFITS TO OPERATORS

An immediately obvious benefit to operators is the ability to deploy only the functions needed to support particular customers and particular market segments. Additional functionality not needed for the particular customer or market segment need not be deployed. This results directly in savings compared to being

required to deploy full functionality to support devices that will use only a part of that functionality. And a derivative benefit is the ability to deploy 5G systems more quickly because fewer functions need to be deployed, enabling faster time-to-market. Sections 7 and 8 provide an in-depth discussion of these and other operator benefits.

3.0 NETWORK SLICING SYSTEM ARCHITECTURE (E2E NETWORK SLICING)

Figure 1 illustrates the system architecture used in network slicing.

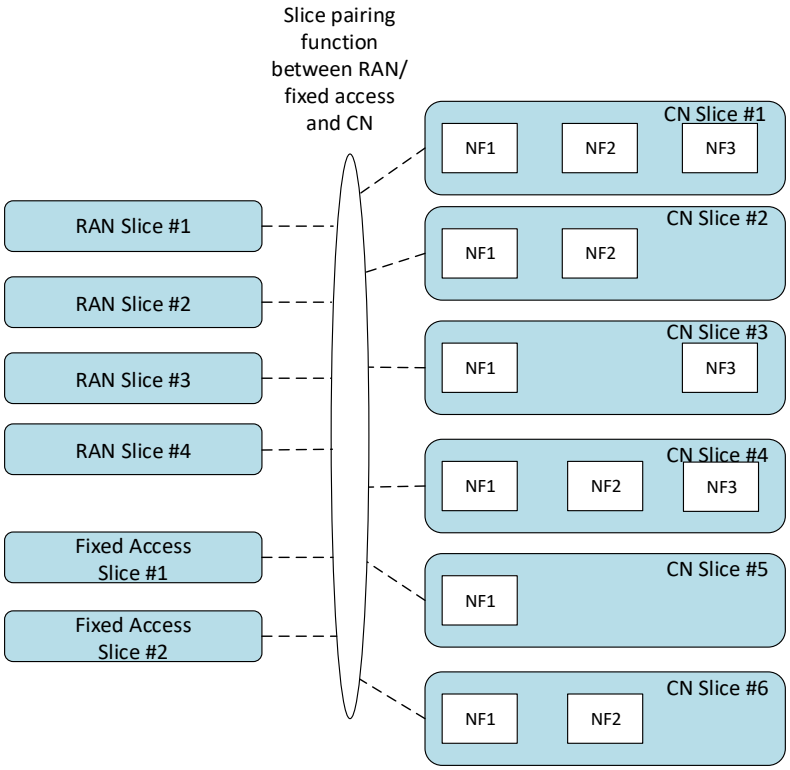

Figure 1. Network Slicing Architecture.

The network slicing architecture contains access slices (both radio access and fixed access), core network (CN) slices and the selection function that connects these slices into a complete network slice comprised of both the access network and the CN. The selection function routes communications to an appropriate CN slice that is tailored to provide specific services. The criteria of defining the access slices and CN slices include the need to meet different service/applications requirements and to meet different communication requirements.

Each CN slice is built from a set of network functions (NFs). An important factor in slicing is that some NFs can be used across multiple slices, while other NFs are tailored to a specific slice.

The mapping among devices, access slices and CN slices can be 1:1:1 or 1:M:N, as Figure 2 illustrates. For example, a device could use multiple access slices, and an access slice could connect to multiple CN slices.

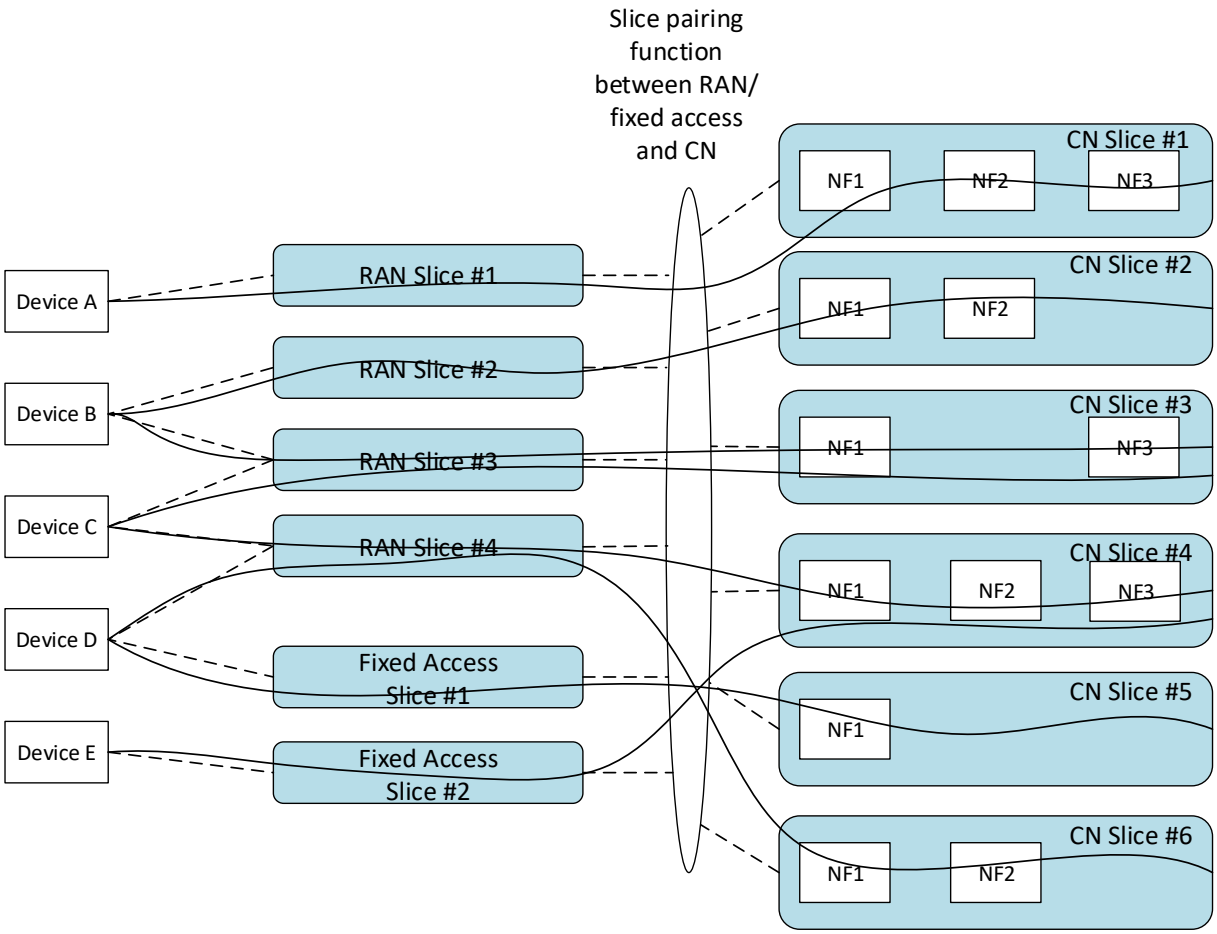

Figure 2. UE Connection with the Network Slices.

The pairing between access slices and CN slices can be static or a semi-dynamic configuration to achieve the required network function and communication needs. A network slice would last throughout the intended service lifetime and would provide full network function support to the devices connected with the network slice.

Examples of network slices include:

- A slice serving a utility company
- A slice serving remote control for a factory
- A slice serving a virtual operator
- A slice optimized for streaming video

Figure 3 illustrates some examples:

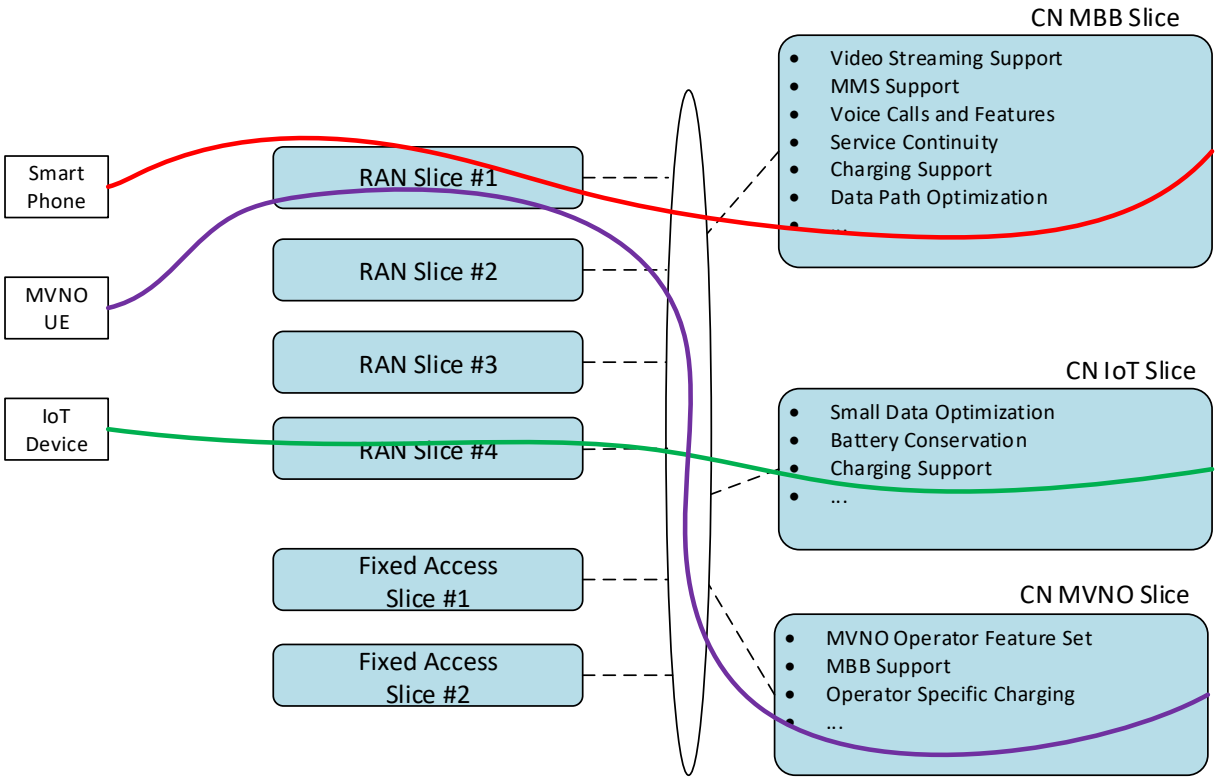

Figure 3. Network Slice Examples.

As Figure 3 shows, some slices can share an access slice, as is shown in the example for the "Smart Phone" and "MVNO UE" devices. For instance, slices serving different utility companies can all be connected to the access slice tailored for supporting massive machine type communication (mMTC), while the slice optimized for streaming video can share the access slice for mobile broadband (MBB). Some slices would have a dedicated access slice and CN slice. For example, a slice for factory remote control may need to have dedicated access and CN functions to provide guaranteed performance such as latency and reliability. A fixed broadband access slice could have a dedicated CN slice tailored for the fixed access.

Later sections of this paper discuss enabling technologies in access and core network slices. Slicing at the end-user device level is currently out of scope of this document.

4.0 NETWORK SLICING IN CN

CNs traditionally have been designed as a single network architecture serving multiple purposes, addressing a range of requirements and supporting backward compatibility and interoperability. This one-size-fits-all approach has kept costs at a reasonable level because one set of vertically integrated nodes provides all functionality. The current technological evolution towards virtualization, NFV, SDN and advanced automation and orchestration makes it possible to build networks in a more scalable, flexible and dynamic way. Such capabilities allow today's network designers to contemplate the core in a radically different way, providing greater possibilities for tailored and optimized solutions. Evolving standards work in 3GPP is focused on allowing network architectures to develop in revolutionary ways.

Network slicing allows core networks to be logically separated, with each slice providing customized connectivity, and all slices running on the same, shared infrastructure or on separate infrastructures as the operator requires. This is a much more flexible solution than a single physical network providing a maximum level of connectivity. It is likely that networks will need to be deployed using different hardware technologies, with different feature sets placed at different physical locations in the network, depending on the use case. To support a specific set of services efficiently, a network slice should have access to different types of resources, such as infrastructure—including VPNs, cloud services and access—as well as resources for the core network in the form of VNFs. The flexibility of 5G core networks will improve significantly by supporting a full separation of control plane and user plane, and through adopting selected SDN principles and technologies.

Supporting the separation of the control and user-plane functions is one of the most significant principles of the 5G core-network architecture. Separation allows:

- Control- and user-plane resources to be scaled independently
- Distribution of the user plane to sites closer to the user device
- Selective choice of the user plane functions needed for different slices
- Decomposition of the user plane into smaller functions
- Support for migration to cloud-based deployments

The control plane, illustrated in Figure 4, can be agnostic of many user-plane aspects, such as physical deployment, and L2 and L3 transport specifics. Typical control-plane functionality includes capabilities such as the maintenance of location information, policy negotiation and session authentication.

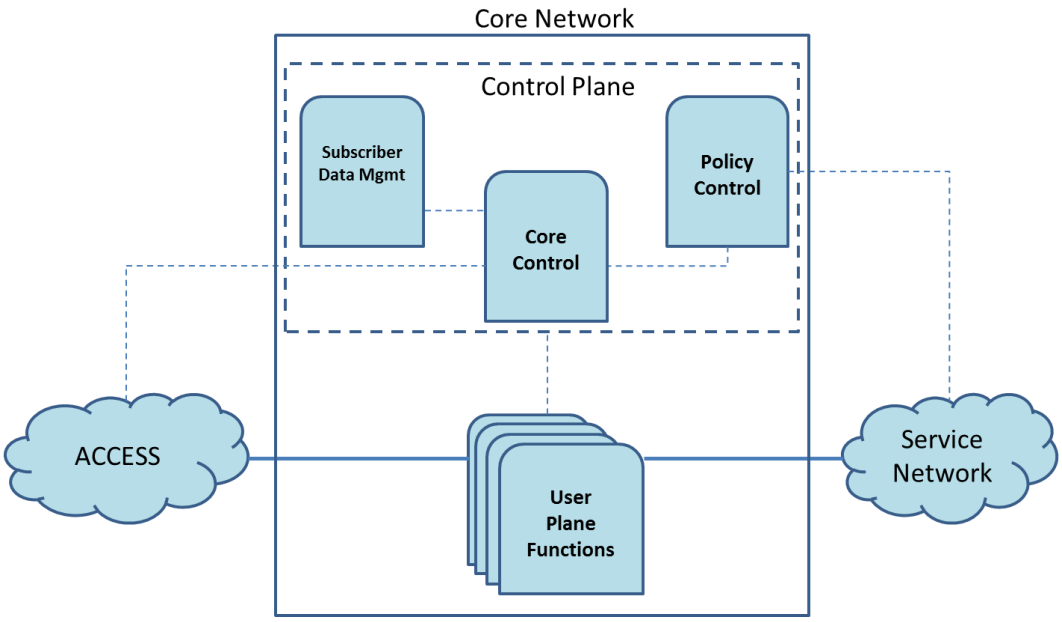

Figure 4. CP-UP Separation.

User-plane functionality can be deployed to suit a specific use case. The connectivity needs of each use case varies, so the most cost-efficient unique deployment can be created for each scenario. For example, an M2M service with a small payload volume and low mobility has connectivity needs quite different from those of an MBB service with high payload volume and high mobility. An MBB service can be broken down into several sub-services, such as video streaming and web browsing, which can in turn be implemented by separate feature sets within the network slice. Such additional decomposition within the user plane domain further increases the flexibility of the core network. The strict separation of the control and user planes enables different execution platforms to be used for each. Similarly, different user planes can be deployed with different execution platforms, even within a user plane, all depending on which solution is most cost-efficient. In the above MBB example, one sub-chain of services may run on general-purpose CPUs, whereas another sub-chain of services that requires simple user-data processing can be executed on low-cost hardware.

By separating user- and control-plane resources, the planes may also be established in different locations. For example, the control plane can be placed in a central site, which makes management and operation less complex. And the user plane can be distributed over a number of local sites, bringing it closer to the user. This is beneficial, as it shortens the round-trip-time between the user and the network service, and reduces the amount of bandwidth required between sites. Content caching is a good example of how locating functions on a local site reduces the required bandwidth between sites.

Special attention must be paid to the design of the user plane to meet requirements for high bandwidth, which may apply on an individual subscriber basis or as an aggregated target. For example, in some use cases, the majority of user-plane traffic may require only very simple processing, which can be run on low-cost hardware; whereas the remainder of the traffic might require more advanced processing. Cost-efficient scaling of the user plane to handle the increasing individual and aggregated bandwidths is a key component of a 5G core network.

5.0 NETWORK SLICING IN THE RAN

When considering how network slicing is supported in the RAN, it's important to consider two aspects:

1. The radio access type (RAT) that supports the network services provided by the slice
2. The configuration of RAN resources to appropriately interface with and support the network slice.

5G must meet multi-dimensional KPI goals[5]. Some of the goals cannot be met at the same time. For example, low latency and reliability often come at the cost of spectral efficiency. For example:

- Optimization for low latency and high reliability
- Optimization for massive connection
- Optimization for high spectrum efficiency
- Optimization for high data rate
- Optimization for wide coverage

New features were added in 3GPP LTE later releases for supporting services such as narrow band internet of things (NB-IoT) and machine type communications (MTC). In addition, configuration parameters were added to be able to create direct device-to-device (D2D) communications. The LTE radio architecture was primarily designed for MBB services, so co-existence with legacy LTE radio is required when adding new RATs, which imposes constraints on the new RAT design. For example, as LTE physical downlink common control channel (PDCCH) is transmitted over the whole system bandwidth, the MTC radio has to skip the symbols occupied by PDCCH (e.g., the first 3 symbols of each sub-frame). This limits the number of symbols MTC can use within each sub-frame (e.g., can only use 11 out of the 14 symbols per sub-frame). In 4G LTE, adding a new radio configuration requires patches on top of the legacy LTE radio. By comparison, the 5G radio must be designed to be forward compatible so that future RATs and configurations can be nicely fit into the 5G radio framework.

RAN slicing is a mapping of slice-ID to a set of configuration rules for the RAN. So rather than separate control and user plane RAN functions on a per-slice basis, there are RAN configuration rules associated with each slice to accomplish the network services supported by the slice.

Some design and operational requirements on RAN slicing are:

- Each slice is supported in the RAN by applying a set of configuration rules to the RAN control and user plane functions
- Some network functions are common to several slices (e.g., mobility management)
- Common control functions coordinate RAN resource usage among the slices. This is to improve the performance and efficiency of the entire 5G system

The following design aspects can be considered in RAN slicing:

1) **Resource management.** Radio slices in the RAN can share the radio resources (time, frequency, space) and the corresponding communication hardware (digital baseband processing components, analog radio components) in a dynamic or static manner according to the configuration rules for

[5] Recommendation ITU-R M.2083: IMT Vision - "Framework and overall objectives of the future development of IMT for 2020 and beyond", September 2015.

the network slice. With dynamic resource sharing, each of the slices obtains use of resources based on its demand and priority

The radio resource sharing among the slices can be done by scheduling or by contention. If done by scheduling, each of the slices submits resource requests to a central scheduler, such as the scheduler in a base station or a central RAN controller. The scheduler then allocates radio resources to the slices based on factors such as the quantity of resources requested by the slice, the priority of the work being performed by the slice and the overall traffic load. In a contention-based system, each of the slices autonomously acquires radio resources following some pre-defined rules. With static resource assignment, a slice is pre-configured to operate in a dedicated resource throughout its operation time. Static resource sharing provides guaranteed resource allocation to the slices. Dynamic resource sharing allows overall resource usage optimization.

Figure 4 illustrates an example in which dedicated radio resources are allocated for each slice. The slices share the radio resources using frequency division multiplexed (FDM) and time division multiplexed (TDM) methods. Due to different dynamic resource configuration for different radio access types, the appearance of RAN configuration to the radio access types could be different.

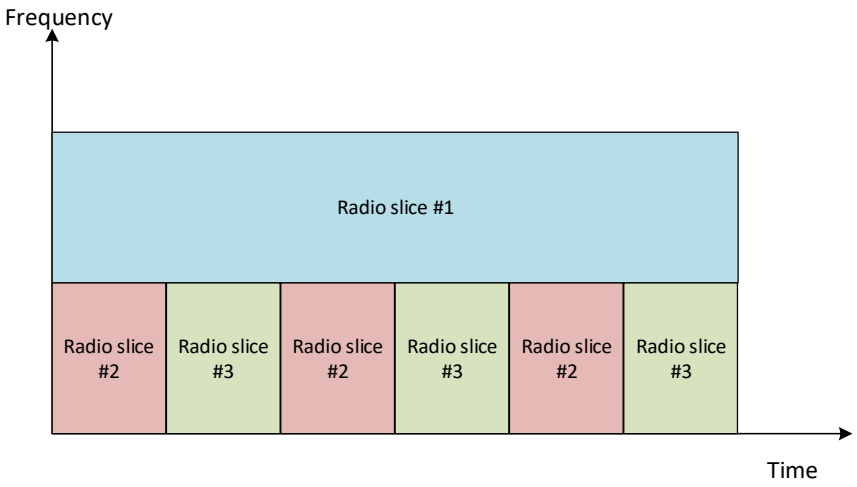

Figure 4. Example of Multiple Slices Sharing the Radio Resource in FDM and TDM.

2) **Control plane and user plane.** In the L2 C-plane (also known as the Access Stratum), common C-plane functions are defined, such as common C-plane functions in device idle mode, and slice specific configuration rules for the C-plane functions tailor those C-plane functions for each of the slices.

Note that not all slices will need the use of all C-plane functions, and their RAN configuration rules would indicate the exact functions needed. In the L2 U-plane, the slices could each have a configured U-plane protocol stack. For example, a slice supporting different QoS classes could be configured to use separate packet segmentation and scheduling layers (similar to the RLC and MAC layers in LTE). Slices with a single QoS class and small packets could be configured to use a single L2 U-plane to reduce header overhead. A hierarchical media access control (MAC) layer can be configured that includes an outer layer MAC for intra-slice scheduling, and an inner layer MAC for inter-slice scheduling.

3) **Slice-specific admission control.** This is needed to meet the initial access requirements of various network slices. RAN configuration rules supply the ability to meet the access needs of a variety of network slices. For example, a network slice serving mission-critical services must get guaranteed low-latency access. With slice-specific admission control configuration rules, a UE operating in one slice may not be admitted to an access point if the slice is not active in the access point. For slices that have not been activated in an access point, or for slices that do not require a slice-specific admission control, the common admission control can be used. In this case, the common admission control also be used as a way of activating a slice in the RAN.

4) **UE awareness on the RAN configurations.** A UE may or may not be aware of the RAN configurations for the different services. In one scenario, the UE requests access accesses to a service defined by one core network slice, and the RAN assigns proper radio resource to support the service. The RAN resource configuration can be transparent to the UE. In another scenario, the UE needs to receive the RAN configuration priori to access the service, such as services that are not one of the default services supported by the RAN and need specified RAN configuration and explicit system information signaling.

Note that the concepts described in this section are still in progress in 3GPP. New developments are expected in the coming months.

6.0 OPERATIONAL ASPECTS OF NETWORK SLICING

6.1 NETWORK SLICES MANAGEMENT FRAMEWORK

NGMN defines network slicing architecture as composed of three layers:[6]

- **Service instance layer:** End user services or business services are supported by the network. Such services can be provided by the network operator or by a third party
- **Network Slice instance layer:** The network slice instance provides the network characteristics required by a service instance
- **Resource layer:** The actual physical and virtual network functions used to implement a slice instance

NGMN also defines a network slice blueprint, which is a complete description of the structure, configuration and the plans/work flows for how to instantiate and control the network slice instance during its life cycle.

The E2E slicing architecture described in section 4 of this document represents a logical decomposition of the network slice instance layer, taking into account specific network domain functions such as CN and radio network domains.

From the operational perspective, and based on the 3GPP requirements, network operators shall have the following basic requirements for network slicing operations:

- Create and manage network slices instances
- Create and manage services to slices pairing function
- Create network slices blueprints
- Deploy and operate fault, configuration, accounting, performance and security (FCAPS) functions for the defined network slices. Configuration management (CM), fault management (FM) and performance management (PM) capabilities should be provided per slice so operators can monitor end-to-end service health, including relevant information about the slice's performance relative to the required QoS from the slices, and also about the NFs' performance
- Create and manage slice components pairing function (e.g., CN slice to RAN slice pairing rules)
- Modification of network slices will be possible, such as adding, deleting and modifying network slices
- Slices are end to end (e.g., RAN, CN)
- Resource management for a network slice may cross operator domains, and so may require cooperating resource management domains
- Slice-as-a-service concept allows a third party to create and manage a network slice configuration (e.g., scale slices) via suitable APIs, within the limits set by the operator

Figure 7 illustrates the network management and orchestration (NMO) plane, which provides management and orchestration functions for the three layers composing the sliced network architecture. NMO functions need to allow for the orchestration and management in a per-slice level.

---

[6] *Splendid Isolation: A Slice Abstraction for Software-Defined Networks*, Stephen Gutz, Alec Story, Cole Schlesinger, Nate Foster, ACM proceedings. August 2012.

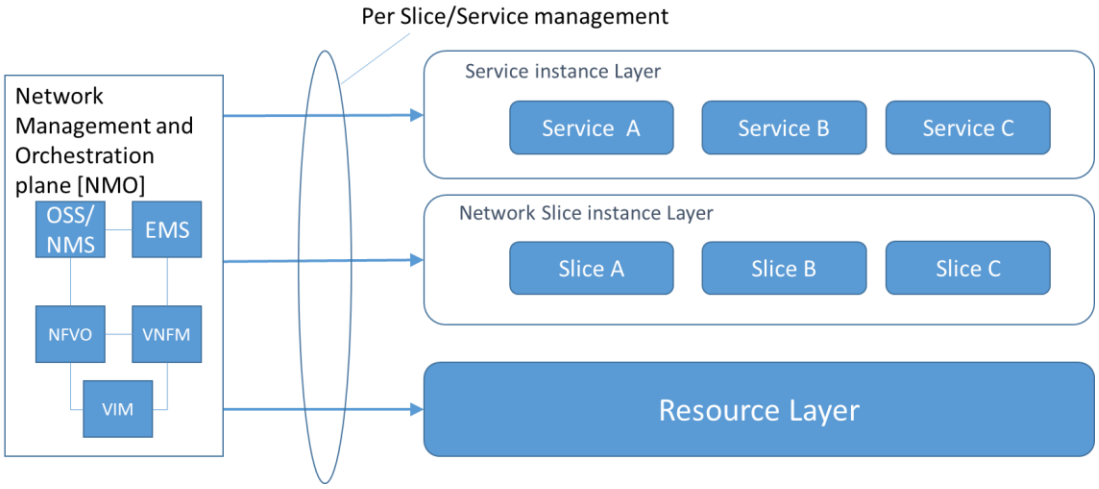

Figure 7. Network Slice Management.

As mentioned above, network slicing management needs to take into account the relevant mapping functions used to build a slice. It also need to account for the rules that the service provider uses to define initial state of this mapping and life cycle decisions based on network and usage dynamics.

Figure 8 presents the relevant pairing functions that need to be managed for network slicing.

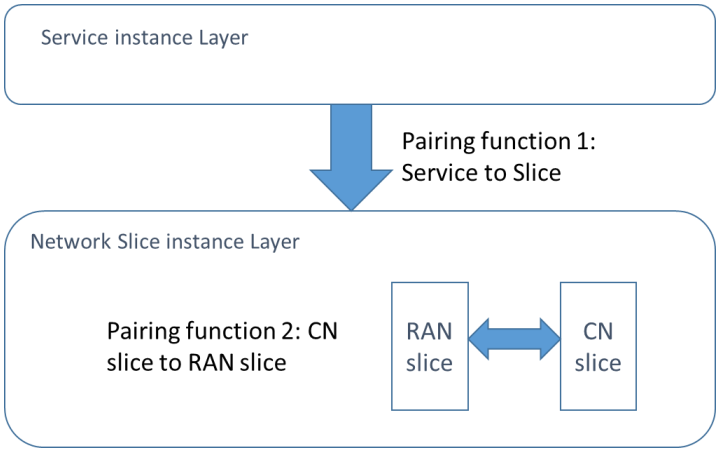

Figure 8. Service to Network Slice Pairing.

Figure 9 shows how services can be mapped onto different slices, according to operator policies and business strategies, which are represented as Pairing Function 1 in Figure 8. Each slice has a certain set of attributes for example end to end low latency or high bandwidth, which can make it more appropriate for certain type of services.

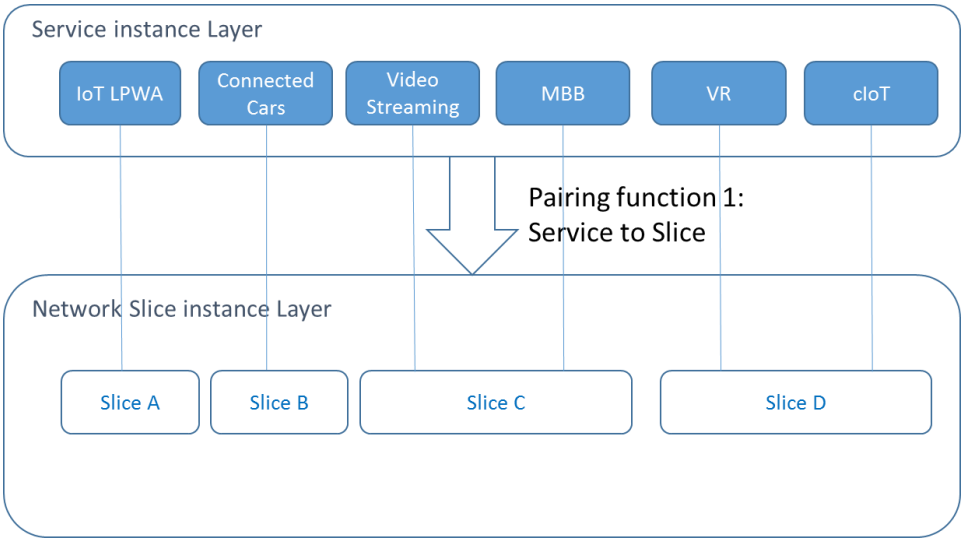

Figure 9. Service Instance to Network Slice Instance.

From the operational perspective, the management system shall provide operators with the capability of mapping services to slices, such as:

- The operator has to map an offered service to particular network slice
- A network slice provides a set of network capabilities and performance level/SLA that may be suitable for a set of service types. As a result, multiple services may be mapped to the same network slice. For example, an eMBB-oriented slice may support video streaming, music downloads and large file transfers. Meanwhile, another low latency-oriented slice may support VoIP and real-time messaging

Users with same service type may be associated with different slices. One example is when there is a need for tenant isolation, where each slice is dedicated to a group of users such as local users and roaming users). Another example is geographic separation, where users are directed to a slice that is closest to them to reduce latency and backhaul requirements.

A network slice may be dedicated to a particular service type. For example, IoT communications may be directed to a slice whose low overhead and minimal features are designed to provide the best performance for IoT devices.

Due to the nature of a network slice, multiple administrative domains may be involved. This encompasses both wired and wireless network resources, virtualized or otherwise. For example, roaming is one scenario where a network slice instance may traverse more than one administrative domain. Another example is the support for business verticals, where additional capabilities may be provided by other administrative domains. Therefore, the services, network service instances or network slice instances need to be operated across multiple administrative domains. In this context, a network slicing orchestration function, responsible for managing network slicing, is needed at the service layer (customer facing service according to the TM Forum). At the resource layer (resource facing service according to the TM Forum), network slice management is performed by the resource orchestrator, composed of NFVO (for the VNF part) and of application resource configurators, where applications are 3GPP services, transport, etc.

The following approaches may play a role in operator networks to enable building, deploying and managing network slices with different sets of attributes:

- For the virtualized part of the network, network service lifecycle management operations defined by ETSI NFV and performed by the management and orchestration (MANO) systems

- QoS for RAN and QoS for CN, along with policy and charging control, to ensure end-to-end QoS and QoE, all as defined by 3GPP, applied to each slice
- Network slice selection based on UE type, similar to 3GPP's "Dedicated Core Networks" for different use cases and users

Figure 10 illustrates slicing governance using the Os-Ma interface as an example.

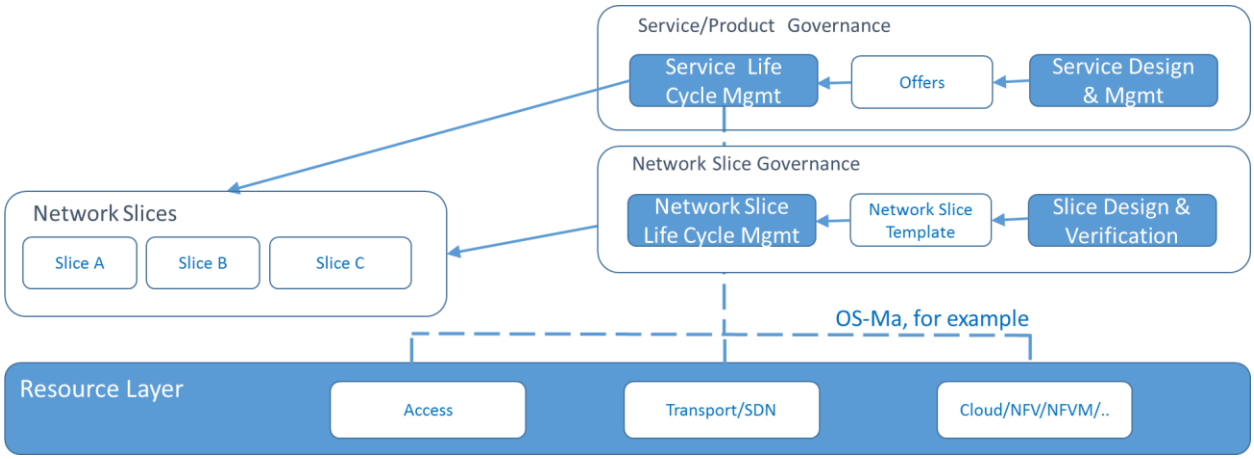

Figure 10. Network Slice Management and Resources.

- The Os-Ma interface is described in ETSI GS NFV 002 v1.2.1 (2014-12) [3]. OSS/BSS - NFV Management and Orchestration (Os-Ma) reference point is used for: Requests for network service lifecycle management. Requests for VNF lifecycle management. Forwarding of NFV-related state information
- Policy management exchanges
- Data analytics exchanges
- Forwarding of NFV related accounting and usage records
- NFVI capacity and inventory information exchanges

6.2 NETWORK SLICING MANAGEMENT AND ETSI NFV MANO

A fundamental component of network slicing is a MANO system that enables full lifecycle management of network slices and associated network resources. ETSI NFV[7] standard defines lifecycle management of "Network Services" ("NSs") that can be reused in the lifecycle management of network slicing. 3GPP SA5 hhas recently started a new study item: "Study on management and orchestration of network slicing for next generation network" (FS-MONETS). One of the objectives of FS-MONETS is to investigate and make recommendations about the relationship between network slice MANO concepts developed in the study and the MANO defined by ETSI NFV.

It is possible to derive and find similarities between 3GPP network slicing concept and the ETSI NFV concepts of ETSI NFV Network Service (NS). An ETSI NFV Network Service (NS) is composed of VNFs and PNFs, along with the virtual links or interconnections between them. The NS Descriptor (NSD) includes the information on the VNFs the NS is comprising, the interconnection topology between the VNFs and PNFs and the information used for the lifecycle management of NSs. As an ETSI NFV MANO component, the NFVO is responsible for the LCM of NS. ETSI NFV MANO also includes VNF Manager (VNFM) and Virtual Infrastructure Manager (VIM). A VNF is typically provided by a vendor to a network operator, while

[7] Meaning of the term "Network Service" in ETSI NFV is different from the meaning of this term in 3GPP; with the latter usually understood as the service supplied to specific user

the NS would be designed by the network operator. Note that PNF in ETSI NFV are typically PNF end points that VNFs can connect to via virtual links. ETSI NFV does not manage PNFs, nor physical links between PNFs.

Using existing LTE architecture elements as an example, a slice could consist of the well-known network entities (eNB, MME, SGW, PGW, PCRF, HSS, etc.) required to provide an LTE service to a population of UEs. In ETSI NFV terms, the eNB, at least in part, is a PNF, and the others could be either PNFs or VNFs. These VNFs and PNFs could be part of NSs as defined by ETSI NFV NFVO.

This LTE slice, which could be composed of ETSI NFV NSs, could be described in terms of NS Descriptors (NSD), and these ETSI NFV NSs would be managed (instantiated, scaled up/down, healed, etc.) by ETSI NFV MANO. The VNFs themselves would be managed by one or more VNFMs, in the usual way.

The Network Slice Life Cycle Management would typically interact with ETSI NFV MANO for these virtualized network services and functions, as shown in Figure 11, with an open Nsl-Ma interface, with Nsl standing for “network slice.”

However, the PNF would be managed outside of MANO, as well as subnetwork parts composed of connection of PNF. This would require the Network Slice Life Cycle Management to also interface with a non-virtualized life cycle management & operation environment, as shown on Figure 11, with an open Nsl-PN interface, with PN standing for “physical network.”

It is important to note that creation of a slice is not limited to NFV operations as PNFs are not under control of MANO, but also because some PNFs are shared between different slices such as PNF part of RAN. There will be network management operations to configure in certain ways the network entities (NEs) that are not under control of the NFVO and their interconnections, to allocate some (non-NFV) network resources, and/or set policies for their allocation. This part, referred as “non-virtualized life cycle management’ in Figure 11, potentially quite significant, is yet to be specified in 3GPP SA5.

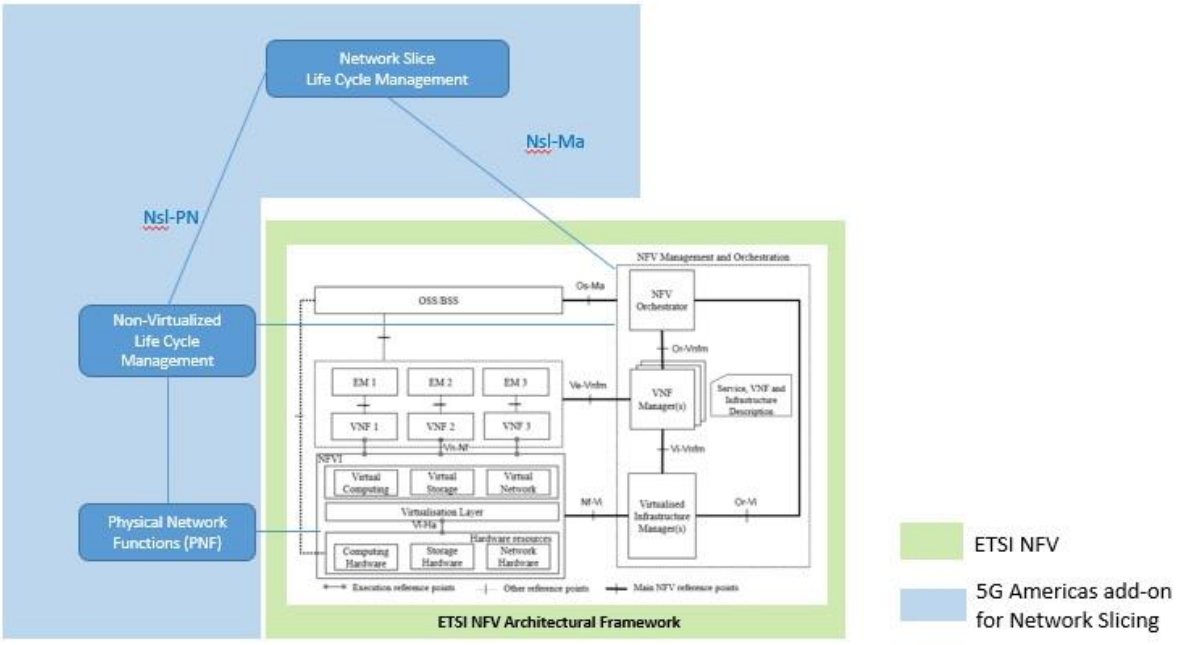

Figure 11. Network Slice Management and ETSI NFV MANO.

As shown in Figure 11, this may introduce some new management systems (i.e., Network Slice Life Cycle Management and some new interfaces).

According to current IFA013: "VNF instances can be shared by several NS instances. In some cases, the OSS/BSS FB will indicate to the NFVO that a VNF instance (already being used by at least one NS instance) shall be used in another NS instance."

6.3 NETWORK SLICE MANAGEMENT

Automation for network slices is a key strategy to achieve operational goals. It starts with transferring the software for network slices, which may come from vendors, to the operator, followed by initial testing of the slices in the operator's network. Various suppliers (i.e., software developers) electronically deliver updates to different components that make up a network slice to the operator, where they are automatically integrated with any operator updates or configuration changes. Automation then continues with deploying the network slices in the production network and further controlling the scaling of slices and performing longer term maintenance on the slices.

6.3.1 NETWORK SLICE LIFE CYCLE MANAGEMENT

This section describes five elements in Network Slice Life Cycle Management.

6.3.1.1 SLICE TEMPLATE CREATION

Network slicing must take advantage of cloud technologies and cloud-based principles. One fundamental telecom cloud principle is the management of network function based on blueprints or descriptors. All resources that make up a network slice need to be defined and managed by the blueprint. The management methods must be uniform across all types of VNFs, services and vendors. The operator's ability to manage slices using blueprints allows an efficient allocation of network slices against the set of physical and virtual network resources at the operator's disposal. This also means that it must be possible to manage different vendors' VNFs within a single slice with a common blueprint and MANO system.

A network slice template defines the setup of the slice, including the components that need to be instantiated, the features to enable, configurations to apply, resource assignments and all associated workflows, including all aspects of the life cycle (such as upgrades and changes). The template contains machine-readable parts, similar to OASIS TOSCA models, which support automation.

Leveraging cloud technologies means that multiple slices with different service properties are still able to run over a common set of resources. Furthermore, all VNFs execute over a common pool of hardware and common infrastructure/platform software. The platform includes the hypervisor or container system, virtual infrastructure manager such as OpenStack, virtual storage system and networking such as an SDN controller and switches.

6.3.1.2 SLICE INSTANTIATION

In the instantiation phase, a network slice instance is created, or instantiated. The virtualized part of the network slice instance would be instantiated by an ETSI NFV NFVO, while the non-virtualized part would be created and configured by the non-virtualized management environment.

For the virtual part, instantiation may include an initial feasibility check phase, in which the NFVO checks if there are sufficient resources to instantiate the VNFs composing the network slice requested. In addition, the NFVO can check if the requested input parameters, and NSD for the new slice instance, make sense. Upon successful completion of this phase, the NFVO instantiates the VNFs, which comprise the network slice as defined in the template, via VNFM and allocates required resources via VIM.

As a network slice is comprised of a cooperating set of VNFs, there is a need for network function discovery. VNF instances, allocated to a slice according to the slice blueprint, need to be able to find one another during slice instantiation phase in order to work together automatically. This shall be driven by the VNF management and service orchestration mechanisms. This shall not be dependent on service/VNF implementation as this discovery is a general requirement and must be uniform across all types of workloads and vendors VNFs. If VNFs and services perform this discovery in duplicate but different ways, it will be inefficient and result in fragmenting the network resources.

6.3.1.3 SCALING

Once a network slice is in full production, scaling of network slices is a crucial capability. But for operators to fully exploit this, the scaling process has to be automated. The processing of monitoring data and triggering of scale in/out/up/down events is too dynamic to require human intervention and must be programmable. Scaling can occur at different levels: at virtualization level for a particular VNF (or within some VNF components) or a particular network service (collection of VNFs). Alternatively, triggers can occur at the application level. For EPC, for example, it may be the number of gateway sessions or number of traffic flows processed.

While it is understood that network slices may include non-virtualized resources, automatic scaling using virtualized resources is a preferred model for network slices, especially for resources that require dynamic scaling and automation to cope with peak of traffic and resource need.

6.3.1.4 NETWORK SLICE ISOLATION

Elasticity of a network slice will be provided in terms of capacity with minimal impact on the services of this slice or other slices. Operators shall be able to define the maximum amount of resources allocated to any slice in the NFVI components. The non-virtualized infrastructure components and the management system will be able to scale out slices, per need, up to the maximum allocated capacity. In addition, it should be possible, by management system, to isolate the allocated resources to different slices, so that scaling of resources for a certain slice will not come at the cost of resources to another slice.

There are a number of dimensions related to this isolation topic that impact management of the slices:

- **Management of the isolation of traffic.** While network resources are being shared among virtual network slices, the network slicing mechanism should guarantee that data flowing through a given slice does not move to another slice.
- **Management of the isolation of bandwidth.** Any given slice and flow going through a slice is allocated certain bandwidth and should not steal bandwidth from another slice. This includes bandwidth on the links and on the actual network nodes CPU/storage/network capacity.
- **Management of the isolation of processing.** While network/cloud resources are being shared among virtual network slices, the processing of packets on a slice should be independent of all other slices. Packet processing for multiple slices must be separated by virtualization/container environments to run independently if on the same physical server.

- **Management of the isolation of storage.** Data related to a given slice should be stored independently from data of another slice, whether memory or disk storage is being used.

Different mechanism may be envisioned for that: either a low-level mechanism such as VLAN, SDN or programming isolated slices as investigated in [5] for traffic isolation, or leveraging other NFV mechanisms for other virtual resource and management isolation. Constant monitoring of the network slices, with closed loop mechanism to fix the potential issues and misconfiguration, is also envisioned to ensure isolation is not violated.

6.3.1.5 ROUTINE MAINTENANCE

The routine maintenance of network slices also require automation. Slices may be taken out of service for various reasons, whether it's an excessive load or administrative reasons (i.e., hardware/software maintenance). In these cases, taking down an active slice will require all end users to be redirected to another slice instance while minimizing service interruption (e.g., signal all UEs to change slice).

6.3.2 CONFIGURATION MANAGEMENT

This section explains operator policies and the isolation of management of the network slices.

6.3.2.1 POLICIES

Some of the slice operations must be governed by operator policies and rules per user subscription. For example, it must be possible to assign UEs to particular slices based on their subscription. Also, the applications that can be invoked may be controlled by policies that take into account factors such as the network conditions and UE capabilities.

6.3.2.2 ISOLATION OF MANAGEMENT

One important aspect of network slicing management is applying isolation of management of the network slices, which might belong to different organizations, users or departments. It is not desirable that information related to the operational aspects of a network slice used by one organization will be exposed to another organization by error or intentionally.

While management of the operator's resources is under the operator administrator responsibility, the lifecycle management and configuration of a network slice is often restricted to a network slice administrator. Therefore, access right restrictions shall be applied to limit the level of control of specific users and/or administrators from one organization. This approach ensure that they handle only the slices that belong to their organization.

6.3.3 PERFORMANCE MANAGEMENT

This section explains SLA Management and Service Assurance and Programmability.

6.3.3.1 SLA MANAGEMENT

Each slice is defined to meet different service/application requirements, which are represented in a certain QoS level. A QoS level can be defined by certain performance descriptors such as delay, jitter, packet loss and throughput. As part of the performance management (PM) capabilities provided by the NMS, the NMS will expose the actual QoS metrics, which are achieved by the measured slice/service. Operators will use this information to understand the extent to which the network delivers the expected QoS to the devices utilizing the slice.

The operator will define a slice based on required QoS for users/applications/devices using the slice. As a result, it would be beneficial to get the monitoring/exposure capabilities per network slice, per reporting period sample, as described in Table 1.

Table 1. Template for Monitoring Capabilities per Network Slice.

QoS metric	T1	T2	T3	…	Tn
Metric 1 - measured					
Metric 1 - % of target					
Metric 2 – measured					
Metric 2 - % of target					
Metric 3 – measured					
Metric 3 - % of target					

SLA management per slice should be automated to account for varying network conditions and changing user patterns on the different slices. Automated SLA management per slice can be achieved by leveraging network programmability, which is described in next section, and a controlling function that implements automation, orchestration and optimization algorithms based on the exposed network, slices and users' information.

6.3.3.2 SERVICE ASSURANCE AND PROGRAMMABILITY

Service assurance in the classical sense is the ability to manage performance, faults, experience and so on. That capability will to continue to be utilized for network slices, as well their sub-components. However, in order to achieve the defined objective for automation, and increase services agility, networks of the future must be programmable to provide the right performance, at the right time. Programmability will lead to networks that are not only aware but also adaptable, meaning they're able to create possibilities to expose capabilities and data to users, partners and third parties. While there is an ongoing debate about how programmability is created, emerging technologies such as SDN and NFV will play an important role in increasing the possibilities of programmable networks. A tiered architecture for managing programmable networks may contain:

Exposure: Allows for external applications to use the network and interact with it in a simple way

Network state: Keeps track of the network state, and how a specific consumer is, and should be, treated from a quality perspective

Policy: Contains the rules about how different situations and customer roles should be treated, and sends instructions to the execution layers. It is crucial that the policies applied in this layer are maintained and organized in a centralized manner

Execution: The technical infrastructure in the networks, such as the network orchestrator, needs to be dynamic so it can adapt and change to certain conditions, and execute orders—for example from policy controller—in the hierarchy

Network programmability is thus the ability of the network not only to orchestrate new services but also to automatically adjust the network resources, based on policies, in order to meet the performance and business objectives. Programmability will also allow for dynamic changes in the slice topology that can occur due to dynamic changes in resource layer or in the network state.

6.3.4 CI/CD AND NEW SERVICE ROLLOUT

Network slicing can be used as enabling technology for new service rollout in a secure and reliable manner. Both suppliers and operators will establish their respective pipelines for software development, various kinds of testing (e.g., unit, integration, field) and deployment. Such pipelines can now be interconnected to create an extended, near continuous delivery process across both types of organizations. The suppliers develop new features and new configurations suitable for a particular customer and perform the initial testing and bug fixing. This software is then delivered in an automated fashion to operator's DevOps system. Here the new software is automatically integrated and tested in operator networks under controlled conditions. Bug reports generated during operator testing are returned back to each supplier to form a closed testing loop.

The operator programmatically controls the integration of new software with existing components and even adjusts the configuration parameters of existing elements. Controlled testing can be achieved through the use of an isolated/protected slice in the production network. After passing such tests, the software can be tested in an experimental slice where some real customers/traffic are placed in the real production network. However, the traffic loads may be light and scenarios simple. As confidence is gained, the software is finally rolled out into the production network in an automated manner. By completely automating this process, operators can move toward a sort of near-continuous integration cycle. Exactly how continuous the cycle is, will depend on each operator's requirements. Critical or important updates can be rolled into production quickly, while more revolutionary features may be introduced more slowly.

7.0 OPPORTUNITIES INTRODUCED BY NETWORK SLICING

7.1 EVOLUTION OF SLICING TECHNOLOGY

Slicing technology is evolving rapidly. In the process, it's drastically changing the architecture and nature of communications. Many use cases are emerging that have diverse requirements in terms of speed, number of connections, availability, reliability, battery life and latency. These use cases may be broadly categorized as MBB, MTC and critical MTC.

For example, a service like virtual telepresence or a hologram will require low latency and high bandwidth capable of running a high-definition or UHD video stream in both uplink and downlink. An application like a building climate control system will utilize a massive number of sensors and actuators. Those sensors need to have a long battery life. As another example, a connected car application requires low end-to-end latency while maintaining high security, reliability and availability. A traditional network tailored for a specific type of communication will not be able to cater above demands. Hence operators will benefit from a more flexible sliced network design to meet these requirements.

7.2 SERVICES TO VIRTUAL OPERATORS

Network as a service (NaaS) is another topic that comes to mind with the introduction of network slicing. Operators get the benefit of providing third parties with access to slices, and allowing them to operate the slices independently within certain rules laid out by the operator. This NaaS model creates new business opportunities for operators. Running dedicated slices for an MVNO would be a great example in which operators can utilize NaaS.

An operator may benefit from exposing different control levels to third party tenants. It is possible for a mobile operator to deploy and operate services and grant third parties access for monitoring purposes by exposing dashboard-like web services. Also, it is possible to provide limited access to third parties to compose services according to use case requirements. However, the service deployment and operation are still performed by the MNO. On the other hand, an operator may provide full control to a third party according to the IaaS model[8] by exposing resources to the third party. The third party may operate the MANO stack and the NMS.

7.3 FURTHER BENEFITS OF NETWORK SLICING

With the introduction of slicing, operators get the opportunity to differentiate through service customization. In the case of connected cars, for example, different car vendors may be operated in different slices. Operators get the ability to allocate dedicated virtual resources to different car manufacturers and operate them independently. Each car manufacturer may be equipped with different services such as traffic efficiency, traffic safety, infotainment, security and emergency support. Service differentiation allows mobile operators to create different pricing strategies.

Monetizing QoE is another area where operators can benefit from network slicing. In 5G, it is possible to define QoS based on each application's unique requirements. For example, an operator may create different slices for different MVNOs, and each slice may have different QoE values, thereby monetizing the experience.

[8] *Network functions virtualization and software management*, Ericsson, whitepaper. December 2014.

One of the biggest challenges that operators encounter today is that a few misbehaving devices can clog the network, impacting availability and reliability of the network resources to other legitimate devices. Network slicing opens up another opportunity for operators where certain types of traffic are contained within a slice and the other slices are not impacted by the behavior of that network segment. A misbehaving sensor or an actuator in a network slice will not impact a critical service, such as public safety, running in another slice.

Network slicing, as discussed in this paper, segments one big broadband network into multiple virtual networks to serve verticals and applications in a more cost efficient manner. Network slicing can be an efficient solution to address one of the challenges faced by the wireless industry: how to expand the market to support vertical services and industries. Moving forward, the wireless industry will need to continue to improve user experience, augment device capability and support traffic scaling. To address these challenges, slicing technology can be used. The computational resources at the network infrastructure and in mobile devices can be sliced and integrated to form virtual computation platforms via the air interface. Based on the virtual computation platform, communication and computation help each other achieve the capacity scaling and device capability augmentation goals. In one aspect, techniques such as edge computation can terminate traffic at the network edge and therefore relax the capacity scaling demand in the deeper network infrastructure. (Studies have showed that higher capacity scaling is easier to obtain at the network edge by densification and radio resource reuse.) In another aspect, communication enables computation offloading from low-capability devices, such as wearable devices, to high-capability devices, such as portable devices, or network infrastructure to augment the capability of the low-capable device beyond its physical limitation.

8.0 CONCLUSIONS

The success of 5G technology is predicated by the ability to meet a wide range of diverse requirements with varied requirements for latency, throughput and availability while delivering multiplicity of use cases. The network has to support low-cost bandwidth at one end of the spectrum while providing low-power, low-speed IoT connections and low latency, high-speed, ultra-reliable connections on the other end. Today's "one-size-fits-all" approach to wireless networks for a highly connected world with different types of devices everywhere is not viable. The key to this shift lies in how end-to-end 5G networks will be designed, architected, implemented and operated.

The system architecture needs to be highly adaptable to meet the performance expectations to serve new and legacy use cases, services, business models, infrastructure usage approaches and radio access needs that will emerge with 5G. The proposed network slicing solution capitalizes on the latest advancements in SDN and NFV enables to configure the 5G system to match the needs of the device and the applications of that device.

Network slicing enables operators to create and support multiple virtual networks by slicing the network into multiple virtual networks running on a common network infrastructure that includes the RAN, backhaul and the CN to support different service types. This approach offers several benefits by enhancing the ability of operators to deploy only the specific functions needed to support specific use cases and customers. In the process, network slicing enables operators to provide service differentiation, which is most desirable because 5G must be able to serve an unprecedented diversity of applications, users, verticals and business models.

Although network slicing holds major promise in optimizing 5G networks in addressing a wide variety of use cases, a lot of questions need to be addressed to enable it, such as network slicing criterion and granularity, the air interface and protocols, the slice-specific RAN, CN operations and the coordination and co-existence of the slices. The 3GPP work on network slicing in relation to 5G is currently in progress, with finalized standards expected in early 2017.

APPENDIX A: FOOTNOTE REFERENCES AND SUMMARY OF EXISTING WORKS

[1]. Description of Network Slicing Concept by NGMN Alliance. Public version. https://www.ngmn.org/uploads/media/160113_Network_Slicing_v1_0.pdf. January 2016

[2]. 3GPP TS 22.891 v14.0.0, "Feasibility Study on New Services and Markets Technology Enablers; Stage 1", Release 14.

[3]. ETSI GS NFV 002 v1.2.1 (2014-12) http://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_nfv002v010201p.pdf

[4]. Recommendation ITU-R M.2083: IMT Vision - "Framework and overall objectives of the future development of IMT for 2020 and beyond", September 2015.

[5]. Splendid Isolation: A Slice Abstraction for Software-Defined Networks, by Stephen Gutz, Alec Story, Cole Schlesinger, Nate Foster, ACM proceedings, August 2012. http://frenetic-lang.org/publications/splendid-isolation-hotsdn12.pdf

[6]. Ericsson - Network functions virtualization and software management – whitepaper – December 2014. https://www.ericsson.com/res/docs/whitepapers/network-functions-virtualization-and-software-management.pdf

[7]. Ericsson –5G system – Whitepaper January 2015. https://www.ericsson.com/res/docs/whitepapers/what-is-a-5g-system.pdf

[8]. NGMN 5G Whitepaper, by NGMN Alliance, February 2015. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[9]. Nokia – Programmable 5G multi-service architecture – press release – September 2015 http://networks.nokia.com/news-events/press-room/press-releases/nokia-networks-unveils-its-programmable-5g-multi-service-architecture-networksperform, September 2015.

[10]. ATIS, "5G Reimagined: A North American Perspective", white paper, https://access.atis.org/apps/group_public/download.php/27373/ATIS-I-0000050.pdf, November 2015.

[11]. Network Store: Exploring Slicing in Future 5G Networks, by Eurecom, September 2015. http://www.eurecom.fr/en/publication/4641/download/cm-publi-4641.pdf

[12]. Network slicing management & prioritization in 5G mobile systems, by Menglan Jiang, Massim Condoluci and Toktam Mahmoodi, King's College of London, 2016.

While definition of 5G architecture and topics such as Network Slicing are still being discussed, a number of vendors, industry consortia, standard organizations and academic research projects have started to publish documents that provide some initial definitions and proposals towards these new concepts, as listed in references. Following is a description of the works:

Ericsson – 5G system – Whitepaper Jan 2015 [8]

The paper addresses the key challenge of today's networks on how One-size-fits-all network architectures with purpose-built systems for single-service subscriber networks with predictable traffic and growth has made it difficult to scale telecom networks, adapt to changing subscriber demands and meet the requirements of emerging use cases. It discusses how the 5G systems will be built to enable logical network slices, which will help operators to provide networks on an as-a-service basis and meet the wide range of use cases that the 2020 timeframe will demand. In general, it describes 5G systems as "One network – many industries, many network slices". It also highlights how the key enablers like Cloud/SDN/NFV and orchestration can help in slicing the network and then can chain them together – both programmatically and virtually – to suit the services being offered and scaled within each network service. It talks about how different industries going through digital transformation can take advantage of the network slices offered

from 5G systems. It discusses the benefit of network slices in key use cases like Mobile broadband, Media delivery, Machine Type Communication (Massive MTC, Critical MTC, Remote Machinery) and ITS.

Ericsson – Network functions virtualization and software Network functions virtualization and software management – whitepaper – December 2014 [6]

The paper talks about how network slicing opens up a new way of achieving in-service software management at the network level. It discusses the operator challenges in upgrading complex network with many different network elements; the total time it takes to upgrade a network can be as high as the total sum of upgrading all of the network elements one by one in a row which is often the limiting factor for adding new functionalities to the network. With software upgrades based on network slicing, a parallel network can be set up with a new software version of the involved functionality, followed by step-by-step migration of sessions – lowering the risk for the operator and ensuring minimal disruption to subscribers. It also brings an interesting perspective of how network slicing can help in eliminating lab network validation in operator environments and use a slice of production network to do software upgrade and testing/validation.

NGMN 5G Whitepaper - Feb 2015 [9]

The paper has positioned the demands and business requirements beyond 2020 and introduced the network service deployment concept of Network Slicing with some high level examples. The paper lists network slicing as one of the key principles for building 5G systems. NGMN 5G White paper gives very little specific information on functional requirements on the core network part. Main focus is on network KPI performance targets of the 5G systems. As a follow-up to the 5G systems white paper, NGMN has published "Description of Network Slicing Concept" document in January 2016 which describes further details of Network Slicing including the Network slicing concept and definitions. This document describes the high level architecture for the network slicing concept consisting of 3 layers: 1) Service Instance Layer, 2) Network Slice Instance Layer, and 3) Resource layer.

Nokia – Programmable 5G multi-service architecture – press release – September 2015 [10]

Building a separate system to meet the future requirements and use cases is not an option. The real opportunity is in developing 5G as a system of systems that can meet all requirements and provide a seamless service from the user's perspective. Nokia's architecture uses a 'system of systems' approach to integrate and align the many different and independent parts of a network to achieve higher performance and more functionality as compared to today's networks. Nearly all network functions will become software-defined, cognitive technologies will automatically orchestrate the network, and content and processing will be distributed across the network close to where they are needed.

ATIS – 5G Reimagined: A North American Perspective – white paper – November 2015 [11]

The scope of this white paper is to understand, define, and advance North American requirements for 5G. It describes use cases which show, from a North American perspective, possible scenarios for 5G networks. These use cases include both commonly recognized baseline requirements and also more disruptive service examples representing a more challenging conception of aspects of 5G. The scope of the use cases is not limited to just the narrowly defined mobile network. Many of these cases include interactions with other elements, including some not normally standardized, such as content provider applications/ networks, operational systems within a carrier network and traffic scheduling and steering algorithms. Based on the documented use cases, the white paper identifies unique characteristics of the North American network and regulatory requirements. Although the focus is on North American requirements, these are considered

in a global context to leverage synergies wherever possible, and to identify new requirements only where necessary.

Among those different papers, some definitions of network slicing have been used, such as the following:

- "A network slice, namely "5G slice", supports the communication service of a particular connection type with a specific way of handling the C- and U-plane for this service. To this end, a 5G slice is composed of a collection of 5G network functions and specific RAT settings that are combined together for the specific use case or business model. Thus a 5G slice can span all domains of the network … as well as configuration of the 5G device" (NGMN)
- The network slice concept consists of 3 layers:
    1. Service Instance Layer
    2. Network Slice Instance Layer
    3. Resource Layer" (NGMN)
- "A slice is composed of a collection of logical network functions that support the communication service requirements of particular use case(s)… The network slicing primarily targets a partition of the core network, but it is not excluded that the RAN may need specific functionality to support multiple slices or even partitioning of resources for different network slices." (ATIS)
- "Multiple independent and dedicated virtual sub-networks (network instances) are created within the same infrastructure to run services that have completely different requirements for latency, reliability, throughput and mobility." (Nokia)
- "Slicing allows multiple service providers to share a common infrastructure; and supports many policies and business models for cost sharing. We propose four requirements for slicing home networks: bandwidth and traffic isolation between slices, independent control of each slice, and the ability to modify and improve the behavior of a slice." (Stanford University)

Eurecom - Network Store: Exploring Slicing in Future 5G Networks [12]

This paper presents a vision for 5G networks, where SDN programs wireless network functions, Mobile Network Operators (MNO), Enterprises, and Over-The-Top (OTT) third parties provide the NFV-ready Network Store. The Network Store proposed here serves as a digital distribution platform of programmable Virtualized Network Functions (VNFs) that enable 5G application use-cases. Currently existing application stores, such as Apple's App Store for iOS applications, Google's Play Store for Android, or Ubuntu's Software Center, deliver applications to user specific software platforms. Ideas are presented where a digital marketplace, gathering 5G enabling Network Applications and Network Functions runs on top of commodity cloud infrastructures, connected to remote radio heads (RRH).

The initial design of Network Functions presented in this paper is aimed at encouraging 3rd party vendors to implement Network Applications in the same way as application stores that motivate software development of mobile platforms, such as iOS and Android.

A slice is defined as a composition of adequately configured network functions, network applications, and underlying cloud infrastructures that are bundled together to meet the requirement of a specific use case or business model. This paper presents a design of a 5G-ready architecture and a NFV-based Network Store that can serve as a digital distribution platform for 5G application use-cases. The proposed Network Store framework aims to achieve two complementary goals:

1. Service-oriented network architecture that is programmable and extensible in terms of infrastructure, network services, and applications;

2. Network slicing to operate virtual networks on top of physical infrastructures, with virtual resource isolation and virtual network performance guaranties.

The purpose of the Network Store is to provide programmable code that resources, deploys and runs the necessary software components, configure and program network elements according to the SDN and NFV paradigms, and provide the end-user with a 5G slice that perfectly matches the demands. This paper further studies, designs, and prototypes a Network Store along with a network slicing architecture for 5G systems.

To meet these demands, this paper advocates 5G system to migrate from slow-moving proprietary and expensive hardware/software platforms (vertical approach) towards open software-defined functions on top of logical resources, leveraging the commodity hardware (horizontal approach). The proposed approach is targeted to enable the delivery of the network as a service and provide the flexibility needed to provision network resources on-demand and to tailor network slices to particular use cases. To achieve this goal, the Infrastructure is used as a Service (IaaS) cloud. Moreover, the network services operate over virtual and/or physical cloud resources under the control of the service manager and service orchestrator.

King's College of London – Network slicing management & prioritization in 5G mobile systems [13]

5G mobile network is expected to serve flexible requirements hence dynamically allocate network resources according to the demands. Network slicing is considered as a key paradigm where network resources are packaged and assigned to set of users according to specific requirements. Network slicing has a twofold impact in terms of user/traffic prioritization as it dictates the simultaneous management of the priority among different slices (i.e., interslice) and the priority among the users belonging to the same slice (i.e., intra-slice). In this paper introduces a novel heuristic based admission control mechanism that dynamically allocates network resources to different slices to maximize the satisfaction of the users while meeting the requirements of the slices. Simulations studies are presented that demonstrates the following:

(i) higher user experience in individual slices,
(ii) increased utilization of network resources and
(iii) higher scalability when the number of users in each slice increases

The proposal in this is based on the idea that network slices communicates to an admission control entity with a desired QoS. The admission control mechanism is based on the priority of the slice. The virtual network allocates the physical radio resources to the UEs of the admitted slices according to the inter- and intra-slice priority. According to the decision of the admission control, the resource allocation task is performed with the aim to maximize the quality of experience (QoE) of the users within each slice, by considering the inter-slice priority.

In this paper, the QoE is measured by considering the effective throughput experienced by the users, normalized according to their maximum requested data rate. If necessary, the resources allocated to a slice with low priority could be reduced to a minimum amount to meet the basic QoS requirements to admit new slice(s) with higher priority. This dynamically changes the amount of network resources allocated to network slices according to the traffic load without affecting the QoE of the users while improving the network utilization.

APPENDIX B: NETWORK SLICING IN 4G

As the 5G core network Architecture is being defined and standardized in 3gpp, utilizing the advancements in virtualization/SDN/automation capabilities of NFV, orchestrated overlay virtual network services can be

achieved using the existing 4G network functions. There are multiple existing methodologies to experiment with network slicing using the 4G architecture. Some of them include

- Service Specific PLMN
- MOCN
- Service Specific APNs
- GWCN
- DECOR
- EDECOR

Overlay network selection can be achieved by DÉCOR/EDCOR methodologies.

Service Specific PLMNs:

Network slicing can be realized today using dedicated PLMNS for certain type of devices and an overlay EPC network to service those UEs. Example of such an implementation can be using overlay EPC network for IoT (Internet of things) or Critical MTC as a separate PLMN.

MOCN:

The Multi-Operator Core Network (MOCN) feature allows two or more operators to share a radio network, while keeping their core networks separated. Traditionally, using these feature two operators can share a common radio network, which saves costs compared to the alternative of having two independent radio networks. At the same time MOCN makes it possible to have separate core networks, where each operator can implement proprietary user services. Each of the separate core networks works as independent network slices using a common RAN. MOCN is specified in 3GPP 23.251 Network sharing Architecture and functional specification. The same can be used to create multiple slices of core networks using multiple PLMN ids sharing the same RAN network.

Service Specific APNs:

Having multiple APNs on the PGW helps to isolate multiple services and consider them as slices. Ex: IMS APN, Roaming partner APNs, MVNO partner APNs etc. Dedicated PGWs and PCRFs for certain APNs can also be used to create slices of networks catering to certain MVNOs or other industry verticals like enterprises/M2M etc. even today.

GWCN:

Gateway Core Network (GWCN) feature allows two or more operators to share a radio network and MME while keeping their own Gateways separated. Using GWCN, operators can provide differentiated offering to their customers while reusing common RAN and MME. GWCN is specified in 3GPP 23.251 Network sharing Architecture and functional specification.

DÉCOR:

DECOR (Dedicated Core) is a 3GPP feature that allows an operator to deploy multiple Dedicated Core Networks (DCNs) within a single PLMN. The Dedicated Core Network consists of one or multiple core network entities and aligns quite well with the “Network Slicing” concept. DECOR is currently being standardized in 3GPP Release 13. The DECOR architecture is specified by updates to existing specifications, most notably 23.401 and 23.060. Requirements and architecture aspects of DECOR are specified in 3GPP CR S2-152107 which provides detailed information on how the DECOR features is

implemented using the 3GPP architecture for LTE, but DECOR can be deployed for GERAN and UTRAN as well.

Slice selection in DECOR is based on an operator configured subscriber parameter (UE Usage Type) provided by the HSS to the MME or SGSN, or also configuration in MME/SGSN. The MME or SGSN evaluates this parameter and if needed the UE is re-directed to an MME or SGSN that is part of the DCN. The re-direction is conducted before authentication and registration is performed in the core network. The DECOR feature does not require any modification or configuration of the UE. Not requiring updates or configuration of the UE makes DECOR a good mechanism to use when introducing slicing to a network with legacy terminals that are not aware of Network Slicing.

The current DECOR work in 3GPP does not include for example support of Wi-Fi or Circuit Switched network. Current DECOR feature requires IMSI authenticated UEs, making it unsuitable for non-SIM based UEs. DECOR will also have a negative impact on access times at initial access due to new evaluation steps being introduced in the core network for some procedures. There will also be increased signaling due to additional request to the HSS as well as possible re-direction messages between core network and RAN.A mechanism for selecting the Core Network instance in a specific Network Slice instance needs to be able to support full isolation between Network Slice instances and shall not require IMSI as subscriber identification. Thus enhancements beyond current DECOR are needed.

EDCOR:

Objective of the EDCOR is to “Improve DCN selection mechanism by providing assistance information from the UE.” It also addresses improved isolation of network slices and reduced signaling to accomplish slice selection. UE provided input for Core Network selection enables improved isolation & signaling optimization - an evolution of DECOR. It also addresses some of the key challenges with DÉCOR related to increased delay and signaling. Since EDECOR requires UE support and it cannot be used with legacy devices. In addition, it also adds functionality to RAN and increases the dependency on the core network. EDECOR is currently being standardized in 3GPP Release 14.

APPENDIX C: ACRONYNMS AND DEFINITIONS

3GPP	Third Generation Partnership Project – The international standards body that produces standards for cellular networks.
5G	Fifth Generation radio system.
CM	Configuration Management)
CN	Core Network

CP	(also C-Plane) Control Plane
E2E	End to end
EPC	Evolved Packet Core – The core network used in 4G LTE networks.
ETSI	European Telecommunications Standards Institute
FCAPS	Fault, Configuration, Accounting, Performance, Security management
FDM	Frequency Division Multiplexed
FM	Fault Management)
HD	High Definition
ICN	Information Centric Networking
IoT	Internet of Things.
KPI	Key Performance Indicator
L2	Layer 2
LTE	Long Term Evolution – The system architecture for 4G LTE networks.
M2M	Machine to Machine
MAC	Medium Access Control
MBB	Mobile Broadband
MBMS	Multimedia Broadcast Multicast Service
mMTC	massive Machine Type Communication
MNO	Mobile Network Operator
MOCN	Multiple Operator Core Networks – A feature allowing multiple mobile network operators to use the same radio infrastructure.
MTC	Machine Type Communications
MVNO	Mobile Virtual Network Operator
NB-IOT	Narrowband Internet of Things – The 3GPP radio interface tailored for machine type communications for devices sending small, infrequent data.
NDN	Named Data Networking
NF	Network Function
NFV	Network Function Virtualization
NGMN	Next Generation Mobile Network

NMO	Network Management and Orchestration
NMS	Network Management System
OASIS	Organization for the Advancement of Structured Information Standards
PHY	Physical radio layer
PM	Performance Management
QOE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network formed based on the defined RAT
RAT	Radio Access Technology
RF	Radio Frequency
RLC	Radio Link Control
RRC	Radio Resource Control
SDN	Software Defined Network
SDO	Standards Development Organization
TDM	Time Division Multiplexed
TOSCA	Topology and Orchestration Specification for Cloud Applications
UP	(also U-Plane) User data Plane
VNF	Virtual Network Function
VPN	Virtual Private Network

ACKNOWLEDGEMENTS

The mission of 5G Americas is to advocate for and foster the advancement and full capabilities of LTE wireless technology and its evolution beyond to 5G throughout the ecosystem's networks, services, applications and wirelessly connected devices in the Americas. 5G Americas' Board of Governors members include América Móvil, AT&T, Cable & Wireless, Cisco, CommScope, Entel, Ericsson, HPE, Intel, Kathrein, Mitel, Nokia, Qualcomm, Sprint, T-Mobile US, Inc. and Telefónica.

5G Americas would like to recognize the significant project leadership and important contributions of project co-leaders Rao Yallapragada and Clara Li of Intel as well as Sabareesan Soundarapandian of Ericsson, and representatives from member companies on 5G Americas' Board of Governors who participated in the development of this white paper.

The contents of this document reflect the research, analysis, and conclusions of 5G Americas and may not necessarily represent the comprehensive opinions and individual viewpoints of each particular 5G Americas member company.