Afolabi, Ibrahim; Taleb, Tarik; Samdanis, Konstantinos; Ksentini, Adlen; Flinck, Hannu

Network Slicing & Softwarization

Published in:
IEEE Communications Surveys and Tutorials

DOI:
10.1109/COMST.2018.2815638

Published: 01/01/2018

Document Version
Peer reviewed version

Please cite the original version:
Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A., & Flinck, H. (2018). Network Slicing & Softwarization: A Survey on Principles, Enabling Technologies & Solutions. *IEEE Communications Surveys and Tutorials*, *20*(3), 2429 - 2453. https://doi.org/10.1109/COMST.2018.2815638

Network Slicing & Softwarization: A Survey on Principles, Enabling Technologies & Solutions

Ibrahim Afolabi, Tarik Taleb, Konstantinos Samdanis, Adlen Ksentini, and Hannu Flinck

***Abstract*—Network slicing has been identified as the backbone of the rapidly evolving 5G technology. However, as its consolidation and standardization progress, there are no literatures that comprehensively discuss its key principles, enablers and research challenges. This paper elaborates network slicing from an end-to-end perspective detailing its historical heritage, principal concepts, enabling technologies and solutions as well as the current standardization efforts. In particular, it overviews the diverse use cases and network requirements of network slicing, the pre-slicing era, considering RAN sharing as well as the end-to-end orchestration and management, encompassing the radio access, transport network and the core network. This paper also provides details of specific slicing solutions for each part of the 5G system. Finally, this paper identifies a number of open research challenges and provides recommendations towards potential solutions.**

***Index Terms*—Network Slice, 5G, Network Softwarization, Orchestration, Network Management, NFV, SDN, Cloud, Mobile Network, MANO and Open Source.**

I. Introduction

The emerging 5G mobile system is expected to build on the success of the current 4G technology offering support for a plethora of network services with diverse performance requirements. 5G era is touted as the generation of mobile networks that will support dedicated use-cases and provide specific types of services to satisfy simultaneously various customer demands. Unlike the "one-fit-all" type of the 4G architecture, 5G is anticipated to consider diverse business demands with often conflicting requirements encouraging service innovation and programmability through the use of open sources and open interfaces that allow access to third parties. By allowing different parties to instantiate and run a software-based architecture, 5G becomes inherently a multi-tenant ecosystem, whereby a tenant refers to a user or group of users with specific access rights and privileges over a shared resource. Hence, 5G networks offer multi-tenancy support and service-tailored connectivity, providing a top-notch Quality of Service (QoS) which will ultimately result in a long lasting Quality of Experience (QoE) with a truly differentiated service provisioning on top of a shared underlying network infrastructure.

5G networks are also expected to create new service capabilities relying on recent advancements in the Internet of Things (IoT) area. In particular, analysts forecast that by 2025 the number of IoT devices could grow to a stunning figure of about 100 billion devices [1], supporting a wide range of services spanning from low-cost sensor-based metering services and delay-tolerant vehicle services to critical communications including e-health, e-business and automotive. For mobile operators, IoT does not mean only support for many more devices and massive connectivity, but also defines a promising opportunity for offering novel services and business solutions within the IoT value chain beyond simple connectivity. To this end, 5G enables open interfaces to support vertical segments, i.e. third parties not owning network infrastructure and requiring networking services with specific needs, as well as new business solutions, e.g., AT&T digital life customized to the needs of users. The automotive industry defines one of the most significant 5G vertical segments. It requires efficient networking capabilities combined with IoT and edge-cloud to facilitate a number of services including autonomous driving, bird eye view, and real-time assessment of road conditions, just to mention a few.

5G should leverage the benefits of network virtualization to accommodate flexibility in providing carrier-grade differentiated mobile network services and ubiquitous coverage, unifying heterogeneous radio and networking technologies supporting the existing 3G (3rd Generation), LTE (Long Time Evolution) and Wi-Fi technologies, and efficiently interworking them with the emerging 5G new radio and fixed access networks. The notion of network virtualization concentrates on the concept of a software-based representation of both the hardware and software resources considering both data and/or control-plane functions. It is the main foundation of (i) network softwarization and (ii) network slicing. Network softwarization[1] is the concept of designing, architecting, deploying and managing network components, primarily based on software programmability properties [2]. It enables flexibility, adaptability, and even total reconfiguration of a network on the fly based on timely requirements and behaviors by considering cost and process optimization in the overall maintenance of the network lifecycle. Network slicing on the other hand seeks to assure service customization, isolation and multi-tenancy support on a common physical network infrastructure by enabling logical as well as physical separation of network resources.

Softwarization is expected to impact several aspects of network development and services such as Content Delivery Networks (CDN) or video accelerators [3] [4]. It has shown huge potential in revolutionizing the deployment and operations of mobile networks, by simply untying network functions from

Ibrahim Afolabi is with Aalto University, Espoo, Finland. Tarik Taleb is also with Aalto University, Espoo, Finland and Sejong University, Seoul, Korea. Konstantinos Samdanis is with Huawei European Research Center, Munich, Germany. Adlen Ksentini is with Eurecom, Nice, France. Hannu Flinck is with Nokia Bell Labs, Espoo, Finland. Emails:{firstname.lastname}@aalto.fi, konstantinos.samdanis@huawei.com, adlen.ksentini@eurecom.fr, hannu.flinck@nokia-bell-labs.com.

[1]Softwarization encompasses orchestration.

proprietary hardware and enabling them to run on commercial off-the-shelf (COTS) hardware computers or data-centers. This possibility makes it doable to deploy software programs as feature updates/upgrades to the necessary network parts to enable newer network functions or simply fixing bugs in existing ones. The feasibility to plug newer functions into the network through programmable interfaces makes the network more flexible, scalable, elastic and perhaps reactive through the use of technologies such as artificial intelligence [5].

Network slicing has been recently gaining momentum among an ever-growing community of researchers from both academia and industry. It has been also the focus of different standardization bodies (e.g., 3GPP (3rd Generation Partnership Project), IETF (Internet Engineering Task Force) and ITU-T (International Telecommunication Union - Telecommunication Standardization Sector)). This concept can be traced back to the idea of Infrastructure as a Service (IaaS) cloud computing model, whereby different tenants share computing, networking and storage resources, in order to create different isolated fully-functional virtual networks on a common infrastructure. In the context of 5G and beyond, a network slice is a unification of virtual resources (e.g., VMs) wherein a set of Virtual Network Functions (VNF) are instantiated and connected via a virtual network, e.g., Virtual Local Area Network (VLAN) and Virtual Private Network (VPN). The possibility to create, on demand and in a programmable fashion, cost-efficient end-to-end network slices and dedicate them for the dynamic provisioning of diverse services is seen as an important feature of 5G. In this vein, efforts are ongoing towards developing a 5G mobile system capable of deploying network slices of varying sizes and structures.

To the best knowledge of the authors, there are no detailed surveys on the topic of network slicing in the literature, except those presented in [6], [7], [9]. In [6], the authors presented a brief analysis of the state of the art of network slicing in 5G using a framework to evaluate the maturity state of the identified projects and their relevance towards the 5G network evolution. Likewise, in [7], the authors describe how wireless network resources can be adequately allocated to network slices without the resources of one slice negatively impacting the quality of others and with a focus on the Radio Access Network (RAN). In [9], the authors study the notion of network slicing in 5G following an architecture model that consists of the infrastructure layer, network function layer and service layer providing an overview and the corresponding challenges. In [8], the authors focus on how Multi-access Edge Computing (MEC) is advantageous to the RAN, especially with respect to latency, its enabling technologies, and orchestration options. Other relevant articles in light of network slicing and its enablers such as Network Functions Virtualization (NFV) and Software Defined Networking (SDN) are also presented in [10], [11], [85] and [12]–[14], respectively.

In comparison to these surveys, the contributions of this paper are manifold. First, this paper presents an extensive and exhaustive review of the principal concepts and enabling technologies for facilitating end-to-end network slicing encompassing all aspects of the 5G and networking technologies from the access and core networks, to the transport networks in a holistic manner. It also overviews the key business drivers and major ongoing research projects in line with the automation and orchestration of end-to-end network slices as well as its lifecycle management. This survey also delves into the roots of network slicing as well as its emerging technologies detailing their various impacts in the architectural evolution of 5G networks with particular focus on slice orchestration and management. Finally, we identify and discuss a number of open research challenges relevant to security [15] as well as network slice resource allocation.

The rest of this paper is organized in the following fashion. Section II presents the 5G service and business requirements, focusing mainly on 5G service requirements and business drivers for emerging markets as enumerated by IMT-2020 (International Mobile Telecommunication system - beyond 2020) and the European Commission (EC) 5G Infrastructure Public Private Partnership (5GPPP). Section III provides an overview of the network slicing concept and presents its main use cases with a glimpse into pre-5G network slicing. Section IV provides an overview of 3GPP network sharing, highlighting how it started, discussing its relevance to network slicing, and presenting its variants. As enablers of network slicing, different virtualization technologies are detailed in Section V. Section VI describes network slice orchestration and management, with particular focus on network slice orchestration architecture, broker, capacity provisioning policy, lifecycle management and explains how network slices can be federated across multiple administrative domains. Section VII analyzes RAN slicing, explaining the RAN slicing requirements, slice resource management and isolation, RAN programmability, and RAN functional split as well as the fronthaul/backhaul slice transport options. Section VIII sheds light on core network slicing, discussing separately the slicing principles of the Evolved Packet Core (EPC) and the new 5G core network. Section IX elaborates pending research problems and challenges relevant to network slicing. The paper concludes in Section X.

II. 5G Service & Business Requirements

A. 5G Service Requirements

5G networks are anticipated to revolutionize the user experience introducing new requirements to shape network platforms for launching new innovative services. These services have diverse requirements, involving higher data traffic volumes and potential number of devices. The initial roll out of 5G is expected by 2020 in order to meet the emerging business and consumer demands. The IMT-2020 vision assists the development of various industry sectors [16], introducing the following targets for research and innovation:

- low latency, i.e. 1ms over-the-air, and high reliability,
- user density with area traffic capacity of 10 Mbps/m^2,
- peak data rate of 10 Gbps with particular scenarios supporting up to 20 Gbps,
- service continuity under high mobility with 500km/h,
- connection density with 10^6 devices per km^2,
- 100 Mbps user experienced data rates for wide area coverage,

- three times higher spectral efficiency (i.e., in comparison to 4G),
- 100 times more energy-efficient networking, and
- energy lifetime for sensors to be greater than 10 years.

5GPPP [17] that encourages research towards 5G, brought light into the requirements of 5G through the flagship project METIS introducing the following target network capabilities [18]:

- *Amazingly fast:* A feature that shall enable instantaneous network connectivity for all applications by providing 10Gbps data rates.
- *Great service in a crowd:* A feature that shall enable a broadband experience, regardless of the user density, assuring a traffic volume of 9 Gbytes/h and a data rate up to 20 Mbps per user.
- *Best experience follows you:* A feature that allows a fixed line network experience for users on the move, with at least 100 Mbps in the downlink and 20 Mbps in the uplink.
- *Super real-time and reliable connection:* A feature that aims at supporting mission-critical machine type communications, ensuring 99.999% reliability and less than 5ms end-to-end latency.
- *Ubiquitous things communicating:* A feature that aims at providing wireless connectivity for sensors and actuators, supporting 300,000 devices per cell, while prolonging the battery lifetime of devices in the order of a decade.

These main service requirements are encouraging rapid time-to-market for launching new services (e.g., deployment time shorter than 90 minutes) and at reducing network management Operational Expenditures (OPEX) by 80%. IMT-2020 and 5GPPP have identified the need for enhanced security and privacy, without explicitly quantifying them.

B. Business Drivers & Emerging Markets

5G is expected to facilitate a business ecosystem, enabling innovative services and networking capabilities not only for consumers, but also for new industry stakeholders. Hence, 5G needs to adopt new partnerships and business models for different types of customers, being the key asset for enabling vertical industries and contributing to the fourth industrial revolution impacting multiple sectors [19]. Verticals can facilitate the development of new products and services, while network operators can create partnerships to accelerate network service roll-outs or to create customized services to vertical industries. The business roles that the 5G architecture would facilitate through virtualization and slicing are the following:

- *Infrastructure providers:* offer the physical network infrastructure and are responsible for upgrades and maintenance. Currently, network operators take the role of the infrastructure providers. However, in the emerging 5G, third players can provide networking hardware and connectivity in private or community indoor areas, e.g., in a stadium or shopping mall.
- *Cloud providers:* facilitate third parties with computation and storage resources and potential cloud services, e.g., platform-as-a-service such as Linux's Openstack, Amazon web service's Elastic Compute Cloud (EC2), Google's Kubernetes, and Microsoft's Azure.
- *Virtual network operators:* lease resources from an infrastructure provider to either complement their own capacity and/or coverage (e.g., Lycabmobile, Lebara, and Virgin Mobile), or gain network coverage in case they lack physical infrastructure. Such leased resources can help against complex and lengthy processes for site acquisition in urban areas as well as to enhance network coverage with low risk in remote areas.
- *Service broker:* interacts with the physical network, collects abstracted resource information and acts as a mediator mapping the service requests originated from virtual network operators, application providers and verticals, to the mobile network operator's resources. A service broker can be a component of the infrastructure provider, mobile network operator or an independent third party.
- *Application providers:* offer, with best-effort performance, services operating on top of a network belonging to an operator. 5G applications with a high data consumption may push application providers (e.g., Netflix and Hulu), to buy network resources from operators, in order to encourage end-users to consume their services without being charged per data volume usage. In addition, application with stringent requirements may pre-define a Service Level Agreements (SLA) set of requirements with operators to ensure a satisfying user experience.
- *Verticals:* offer a variety of services to a non-telecom specific industry, exploiting network and cloud resources from network operators and cloud providers. Most of the new growth is anticipated in taking place through digitalization of the vertical industries such as factories, transportation and health care.

Partnerships between different business players can be established over networking and cloud resources, network capabilities exposure, value-added services and network context information as well as on providing 5G services as a programmable and software oriented capability set. Network slicing is a key technology and business enabler for 5G, facilitating multi-tenancy and enhanced network coverage for third parties in a flexible way, assuring an extra revenue means for operators, infrastructure and cloud providers. Network slices can be established on a permanent basis or on-demand, either opportunistically or periodically, with network and cloud resources belonging to a single or multiple operators or to a mix of different business players.

III. Network Slicing Concept & Use-Cases

A. Early pre-5G Network Slicing

Network slicing rely on virtualization concepts, which have been around as far back as the 1960s [20] [21] when the first operating system (CP-40) was developed by IBM [22]. The design of the CP-40 on IBM system 360/40 supported time-sharing and virtual memory, introducing a breakthrough in computing by accommodating up to fifteen users simultaneously [23] with the illusion of working individually on a complete set of hardware and software [20] [23]. The idea of

virtualization, i.e. creating a virtual form of a physical entity through software methods and processes, formed the vision of virtual systems spanning across computing platforms, network resources, and storage devices [22]. Virtualization was widely adopted for data centers in the 70s and by the early 80s, it was applied into networking, for connecting remote sites securely with controlled performance through the Internet.

The introduction of overlay networks in the late 80s that consist of nodes connected over logical links forming a virtual network over a network composed of physical infrastructure can be seen as an early form of network slicing, combining heterogeneous resources over various administrative domains. Overlay networks provide QoS guarantees in a service-oriented fashion. They are flexible in nature but not automated nor programmable. By 2000, the first-generation platforms for verifying and evaluating new network protocols were established based on overlay networks. PlanetLab [24] [25] adopted a common software package called MyPLC enabling distributed virtualization by allowing users to obtain isolated application specific slices. A slice was defined as a unit component with allocated resources such as computing power/storage on servers or resources existing in namespaces. However, such overlay platforms had limitations in underlay network controls.

In 2008, the GENI project, a US National Science Foundation (NSF) [26] initiative, pushed forward the development of a testbed based on network virtualization technologies for promoting research on a clean slate network, while considering federated resources and mobile network environments [27]. GENI offers instrumentation and measurement tools used for carrying out both active and passive measurements and for visualizing and analyzing measurement results [28]. By 2009, Software Defined Network (SDN) technologies enabled researchers to run their experiments in a slice of existing campus networks allowing programmability via open interfaces [29].

B. The 5G Network Slice Concept & Principles

Network slicing in the context of 5G is a new defined concept introduced by NGMN (Next Generation Mobile Network) in [30]. Network slicing facilitates multiple logical self-contained networks on top of a common physical infrastructure platform enabling a flexible stakeholder ecosystem that allows technical and business innovation integrating physical and/or logical network and cloud resources into a programmable, open software-oriented multi-tenant network environment. 3GPP defines network slicing as a technology that "enables the operator to create networks, customized to provide optimized solutions for different market scenarios which demand diverse requirements, e.g. in terms of functionality, performance and isolation [31]. For ITU-T, network slicing is perceived as Logical Isolated Network Partitions (LINP) composed of multiple virtual resources, isolated and equipped with a programmable control and data plane [32].

Network slicing enables value creation for vertical segments, application providers and third parties that lack physical network infrastructure, by offering radio, networking and cloud resources, allowing a customized network operation and true service differentiation. The VNFs, which constitute a network slice, may vary drastically depending on the service requirements of that particular slice. The type of service associated with a network slice would determine the resources and service treatment the network slice would receive, e.g. a real-time communication network slice would receive the appropriate resources and service treatment to meet ultra low latency demands [33]. Network slicing builds on top of the following seven main principles that shape the concept and related operations:

- *Automation:* enables an on-demand configuration of network slicing without the need of fixed contractual agreements and manual intervention. Such convenient operation relies on signaling-based mechanisms, which allow third parties to place a slice creation request indicating besides the conventional SLA which would reflect the desired capacity, latency, jitter, etc., timing information considering the starting and ending time, and duration or periodicity of a network slice.
- *Isolation:* is a fundamental property of network slicing that assures performance guarantees and security (to defend network openness to third parties) for each tenant even when different tenants use network slices for services with conflicting performance requirements. However, isolation may come at the cost of reducing multiplexing gain, depending on the means of resource separation for explicit use, which may result in inefficient network resource utilization. The notion of isolation involves not only the data plane but also the control plane, while its implementation defines the degree of resource separation. Isolation can be deployed (i) by using a different physical resource, (ii) when separating via virtualization means a shared resource and (iii) through sharing a resource with the guidance of a respective policy that defines the access rights for each tenant.
- *Customization:* assures that the resources allocated to a particular tenant are efficiently utilized in order to meet best the respective service requirements. Slice customization can be realized (i) in a network wide level considering the abstracted topology and the separation of data and control plane, (ii) on the data plane with service-tailored network functions and data forwarding mechanism, (iii) on the control plane introducing programmable policies, operations and protocols and (iv) through value-added services such as big data and context awareness.
- *Elasticity:* is an essential operation related with the resource allocated to a particular network slice, in order to assure the desired SLA under varying (i) radio and network conditions, (ii) amount of serving users, or (iii) geographical serving area because of user mobility. Such resource elasticity can be realized by reshaping the use of the allocated resources by scaling up/down or relocating VNFs and value-added services, or by adjusting the applied policy and re-programing the functionality of certain data and control plane elements. Elasticity can also take the form of altering the amount of initially allocated resources by modifying physical and virtual

network functions, e.g. by adding a different RAN technology or a new VNF, or by enhancing the radio and network capacity. However, this process requires an inter-slice negotiation since it may influence the performance of other slices that share the same resources.

- *Programmability:* allows third parties to control the allocated slice resources, i.e. networking and cloud resources, via open APIs that expose network capabilities facilitating on-demand service-oriented customization and resource elasticity.
- *End-to-end:* is an inherent property of network slicing for facilitating a service delivery all the way from the service providers to the end-user/customer(s). Such a property has two extensions, (i) it stretches across different administrative domains, i.e. a slice that combines resources that belong to distinct infrastructure providers, and (ii) it unifies various network layers and heterogeneous technologies, e.g. considering RAN, core network, transport and cloud. In particular, an end-to-end network slicing consolidates diverse resources enabling an overlaid service layer, which provides new opportunities for efficient networking and service convergence.
- *Hierarchical abstraction:* is a property of network slicing that has its roots on recursive virtualization, wherein the resource abstraction procedure is repeated on a hierarchical pattern with each successively higher level, offering a greater abstraction with a broader scope. In other words, the resources of a network slice, allocated to a particular tenant, can be further traded either partially or fully to yet another third player, which relates to the network slice tenant facilitating in this way another network slice service on top of the prior one. For example, a virtual mobile operator who acquired a network slice from an infrastructure provider, offers a partial amount of such resources to enable a utility provider that uses its virtual network to form an IoT slice.

According to [34], the network slicing process is broadly broken down into three main layers, namely the service instance layer, the network slice instance layer, and the resource layer, as illustrated in Fig.1. Each service instance reflects a service provided by a vertical segment, application provider or mobile network operator. The network slice instance represents a set of resources customized to accommodate the performance requirements of a particular service and may contain none, one or a number of different sub-network instances, being isolated or shared. A sub-network instance can be a network function, e.g. IP Multimedia Subsystem, or sub-set of network functions or network resources realizing a part of a network slice instance.

Each network slice instance is established end-to-end and may contain different sub-networks of distinct administrative and/or technology domains being fully or partly, logically and/or physically isolated from another network slice instance. In particular, the resources associated with a sub-network can be used in an isolated, disjunctive or shared manner following the network slice instance specific policies and

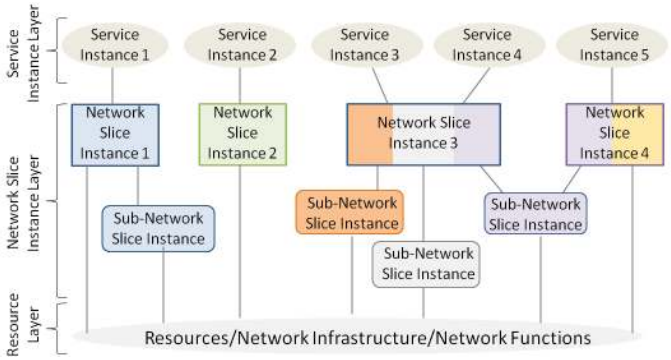

Figure 1: The NGMN network slicing concept.

configuration arrangements. A network slice instance, in turn, can be exclusively used by a service instance or shared among difference service instances, typically of the same type. Common abstractions of relevant resources and open programmable interfaces allow dynamic control and automation of network slice instances reflecting dynamic service demands.

C. Network Slicing Use Cases

The development of 5G networks was initially shaped by NGMN through the introduction of an industrial vision for 5G as summarized in [30] that addresses emerging service and business demands within the time target of 2020 and beyond. The main objective of 5G is to enable an end-to-end ecosystem that provides a fully mobile and consistent experience that empowers a socio-economic transformation in multiple ways, many of which are yet to come. NGMN also anticipates a number of emerging 5G use cases focusing on:

- *Enhanced broadband access everywhere:* envisions a minimum amount of bandwidth, at least 50Mbps, ensuring a connected global society, via high speed Internet. This asset can serve a default general purpose usage.
- *Enhanced broadband access in dense areas:* provides broadband access with up to 10Gbps bandwidth in densely populated areas, e.g. stadiums or open-air festivals, enabling multimedia services, e.g. ultra high definition video streaming.
- *High user mobility:* offers broadband support for mobile users in extremely fast moving vehicles such as high speed trains.
- *Massive Internet of Things:* supports broadband access for ultra-dense networks of sensors and actuators, considering devices in need of super low-cost, long range and low power consumption, e.g. providing utility measurements.
- *Extreme real time communication:* assuring ultra-low latency connectivity, e.g. for interactive tactile Internet.
- *Ultra reliable communication:* provides ultra-low latency [35], reliability and availability of network connectivity supporting, e.g. autonomous driving.
- *Lifeline communication:* supports connectivity in case of natural disasters and emergencies capable to accommodate flexibly a sudden tremendous traffic increase, while assuring resilient connectivity.

- *Broadcast-like service:* provides network connectivity for broadcasting, e.g. news or firmware updates for instance to improve the breaking system of a car or braise up a detected security hole in cars [36].
- *Light-weight communication [37]:* provides network connectivity for supporting essential instantiation, configuration and maintenance service information.
- *Multi-connection:* assures network connectivity for users with different smart devices, e.g. smart glass and smartphones, using multiple access technologies.

Some of these use cases impose diverse and often conflicting performance requirements, while others can be easily combined since the service requirements are similar. Use cases with diverse performance requirements can be realized by different network slices. Typically, service and security requirements determine the type of network slice, while further administrative aspects, e.g. charging, can further distinct slices considering the business model. 3GPP initiated a study named New Services and Market Enablers (SMARTER) [38] in the 3GPPP Services Working Group SA 1. This study detailed new market segments and business opportunities that could be launched with the roll out of 5G, specifying more than 70 use cases, which were later grouped into the following four categories:

- *Enhanced Mobile Broadband (eMBB)* [39] mainly focuses on an umbrella of use cases with service requirements that cannot be met with the Evolved Packet System (EPS). eMBB aims at facilitating support for high data rates including the uplink direction, accommodate high data traffic volumes and User Equipment (UE) connectivity per area, provide wide area connectivity and coverage, while considering fixed mobile convergence and high user mobility as well as support devices with highly variable data rates. Hence, eMBB requires high network capacity, low latency and high network availability.
- *Critical Communications (CriC)* [40] supports the need of services requiring ultra reliability and low latency communications with packet loss as low as 1 packet out of every 10,000 packets and 1ms latency. CriC aims at facilitating mission critical services, industrial automation and control, Augmented Reality/Virtual Reality (AR/VR), tactile Internet, public safety, and disaster and emergency response. CriC needs to facilitate isolation, prioritization, rapid communication setup, very low jitter, location precision and support local content, applications, and services.
- *Massive Internet of Things (MIoT)* [41] facilitates connectivity for high density of devices, typically stationary with non-time critical service requirements, but in need of security, as well as configuration and operational simplicity allowing a long battery lifetime. MIoT is expected to enable smart wearables, e-Health and sensor networks that allow smart home/city [42], farming and smart utilities. MIoT should provide a common communication and interworking framework for various devices, supporting diverse connectivity and edge computing for scalability.
- *Enhanced Vehicular to Everything (eV2X)* [44] focuses on (i) safety-related services such as autonomous driving, platooning (i.e. closely linked vehicles), teleoperated support (i.e. remote control), bird eye view, situation awareness, and cooperative driving allowing direct vehicular communications, and (ii) comfort services including entertainment, mobile hot-spot, and map updates. eV2X should allow high bandwidth, low latency up to 1ms and ultra high reliability, integrating network and cloud-based information for supporting image, video and a range of proximity services considering pedestrians and high density vehicular scenarios. eV2X needs to assure service continuity regardless of speed, even with no network coverage, and support inter-system mobility facilitating high position accuracy. Different slices are recommended for distinct V2X services.

IV. An Overview of 3GPP Network Sharing

The earliest 3GPP network sharing can be traced back to Rel.99 where network sharing was introduced for the very first time in the UMTS (Universal Mobile Telecommunications System) mobile networks. This section provides an overview of the business requirements, passive and active network sharing and network sharing management.

A. Network Sharing Service Requirements

The initial GSM (Global System for Mobile Communications) mobile network design did not support sharing, but it became soon apparent that such a feature was needed, especially with the arrival of UMTS, which required new network deployments. 3GPPP Services Working Group SA1 initially captured in [45] the service and user requirements that should be fulfilled to enable network sharing in a standardized way, specifying the following five main business scenarios:

- *Multiple core networks sharing a common RAN*, where operators share RAN elements by connecting to the same Radio Network Controller (RNC), but not the spectrum.
- *Operator collaboration to enhance coverage*, where two or more operators with individual frequency licenses and RANs that cover different regions, provide together coverage for the entire country.
- *Sharing coverage on specific regions*, allowing an operator to share its network coverage with the subscribers of another operator(s). Outside such region, coverage is provided by each operator independently.
- *Common spectrum sharing* considering the following two variations: (i) an operator has a frequency license and shares its spectrum and (ii) a number of operators decide to pool their individual spectrum together in order to form a large spectra and share it.
- *Multiple RANs share a common core network*, with each RAN and spectrum belonging to different operators.

A network operator should be able to differentiate its services from the services of other operators in the shared network and assure service continuity while network sharing being transparent to end users.

B. Passive & Active Network Sharing

The first generation of network sharing involved simple commercial solutions, concentrating on passive sharing and network roaming. Passive network sharing is characterized by the sharing of site locations and supporting RAN equipment without the need for active coordination between sharing partners. Passive sharing initially concentrated on site sharing to ease site acquisition considering the space and optionally sharing shelters, power supply, air conditioning and other supporting facilities, but with separate installations of masts, antennas and backhaul equipment. Mast sharing followed, enabling mobile operators to additionally share the antenna frame, but keeping their own RAN equipment, offering separate coverage.

The succeeding active RAN sharing concentrated on sharing base stations, antennas, core network and mobile backhaul equipment, also enabling operators to share spectrum resources based on fixed term contractual agreements. Three types of active network sharing are considered in [46], (i) concentrating on the RAN only, (ii) stretching beyond the RAN towards the core and (iii) roaming. 3GPP Architecture Working Group SA2 defined two distinct types of active network sharing architectures as documented in [47] including the following:

- *Multi-Operator Core Network (MOCN)* concentrates on sharing the RAN, including the spectrum, wherein each participating operator maintains a separate EPC. Shared base stations are connected to the EPC of each participating operator via a separate S1 interface. This allows operators to customize RAN operations on the allocated resources.
- *Gateway Core Network (GWCN)* allows operators to share the RAN and additionally the Mobility Management Entity (MME). Such an approach is cost efficient since more network equipment can be shared, but is less flexible, i.e. it is specific to LTE, restricting mobility towards different RAN technologies and impacting the legacy circuit switching fallback for voice traffic.

Network sharing is transparent to UE, which can distinguish up to six sharing operators based on the broadcasted Public Land Mobile Network Identifier (PLMN)-id via the Uu interface. Such PLMN-id is used to enable UEs to obtain connectivity, i.e. select the desired network, and perform a handover as specified in [48]. The S1 interface in turn supports the exchange of operator specific PLMN-ids between eNBs and MMEs, assisting the selection of the corresponding core network [49], while the X2 interface supports an equivalent PLMN-id exchange among neighboring eNBs for handover purposes [50]. Roaming is an alternative means of active network sharing, which does not involve the sharing of network equipment, but the hosting operator provides a network service to visiting users based on a contractual policy agreement and in return of an extra charging fee [46].

C. Network Sharing Management

The notion of network sharing management with respect to performance, sharing agreements and policies is studied by 3GPPP Services Working Group SA1 in [51], introducing the following use case scenarios:

- *RAN sharing monitoring* considering (i) the measurements that should be shared with a participating operator based on the host's regulation policies, (ii) the information that enable the participating operator to manage the allocated resources, e.g. in case of an alarm, and (iii) harvesting RAN coverage quality information from UEs
- *Flexibility in capacity allocation* with respect to (i) revenue, e.g. for fixed or specified duration, first come first served, (ii) asymmetric or unequal resource allocation and controlling per participant, (iii) load balancing in shared RAN, while respecting the agreed shares, (iv) automated capacity brokering for participating operators upon request, (v) dynamic RAN sharing varying the allocated resources during different times and granularity, e.g. at the radio sector level.
- *RAN sharing charging* in terms of (i) event triggering to generate charging records, e.g. a UE enters or exits shared RAN, and (ii) charging reconciliation enabling the host to independently verify the usage of the shared RAN and generate charges per amount of data for each QoS level.
- *RAN sharing broadcast capability* enabling (i) UEs to select their home PLMN based on host's guidance and (ii) public warning in designated coverage areas to alert UEs of urgent conditions related to public safety.

Such a study lead the 3GPP Telecoms Management Working Group SA5 towards extending the legacy network management paradigm to accommodate network sharing requirements in [52] considering the corresponding network sharing architectures. This network sharing management paradigm introduces the Master Operator (MOP) as a single actor, which is responsible for the shared infrastructure deployment, including the spectrum and related operations. In particular, the MOP provides network management services to Participating Operators (POPs), via the means of an enhanced management system referred to as MOP-Network Manager (MOP-NM). The MOP-NM configures which POP shares a network element or a shared RAN/core network domain manager.

The MOP-NM provides notifications to the corresponding POP in case of an alarm and allows POPs to activate a signaling-based radio coverage quality information from subscribers residing on the shared RAN, i.e. a process known as Minimization of Drive Tests (MDT) [52] [53]. MOP-NM communicates with the POP Network Manager (POP-NM) via the Type 5 interface, i.e. the roaming interface, to provide network management services on the shared network. For managing the shared network the MOP-NM employs the Itf-N interface to communicate with the RAN and core domain manager or directly with the equipment element manager, which is enhanced to distinct POPs and support configuration, alarm and performance monitoring. The communication between the RAN/core domain manager and network equipment is performed via the Itf-B interface.

V. Network Slice enabling Technologies

The advent of virtualization technologies has shown tremendous disruptive advantages and opportunities in terms of multi-

tenancy, programmability and flexibility to both networking and computing business initiatives. Virtualization technologies are a key enabler for network slicing. This section overviews such fundamental virtualization technologies from the prism of network slicing.

A. Hypervisor

The concept of virtualization has introduced an additional layer between a regular physical infrastructure and the operating system running on the top. This layer is responsible for producing, controlling and managing virtual machines and it is referred to as Virtual Machine Monitor (VMM) or otherwise hypervisor. The hypervisor is a firmware, which provides a virtual platform for guest operating systems allowing applications and/or other services to be executed. Hypervisors enable and supervise the sharing of hardware resources between network slice instances. There are two main types of hypervisor, namely the type-1, referred to as native or bare metal hypervisor, and the type-2, also known as hosted hypervisor. The type-1 hypervisors are called bare metal hypervisors because they are installed directly on the host machine's hardware, i.e. the same way an operating system is installed. Examples of type-1 hypervisor include the Oracle OVM server for SPARC [54], XEN [55], VMware ESX/ESXi [56], and Linux's Kernel-based Virtual Machine (KVM) [57]. The type-2 or hosted hypervisors are installed on top of the host's operating system, similarly to other computer applications. Hypervisors of type-2 category encompass Oracle Virtual Box [58], VMware fusion [59], VMware Workstation player [60], and Oracle VM for x86 [61]. Besides the fundamental two types of hypervisor, a third one, known as Operating System-level (OS-level), virtualizes multiple servers [21] running in isolated containers. OS-level hypervisors support only the OS similar to that of the host since the virtualized servers, also known as Virtual Private Servers (VPS) [62], share the host's kernel.

B. Virtual Machines & Containers

In computing platform virtualization, the creation of a Virtual Machine (VM) provides the effect of a physical resource that runs its own OS. The actual hardware virtualization takes place on the host machine, while the guest machine is the VM. Current cloud platforms are capable of hosting multiple VMs, running simultaneously and executing different applications concurrently. Each VM shares resources such as computing, storage, memory and network, while its operation is completely isolated from that of the host and fellow guest VMs. From the hardware perspective, there are two types of virtualization, namely the full virtualization and paravirtualization. In full virtualization, also known as native virtualization [21], the complete emulation of the host hardware is enabled, wherein software applications such as guest operating systems can be installed [63] [62]. On the other hand, in paravirtualization, the host hardware environment is not emulated but the installed guest OS is modified to make software applications run on isolated domains [64] [65]. Depending on the underlying hypervisor, VMs known as system virtual machines [66], can be instantiated to offer full support for a complete OS executing multiple processes.

Containers are created based on the idea of an OS-level virtualization, where a physical server is virtualized to enable multiple instances of isolated servers to run as standalone applications. Containers are light-weight alternatives to hypervisor-based VMs, using the OS-level abstraction to partition the system resources creating multiple isolated user-space server instances [67]. Examples of container-based virtualization include Linux-Vserver [68], OpenVZ [69], Solaris Container [70], and Docker [71]. Both containers and VMs are capable of running VNFs, which can be chained together for delivering a particular network service in a flexible manner, forming the base functionality for network slicing. However, while VMs may offer full logical isolation for operating VNFs in a network slice, the light-weight nature of containers can efficiently support network slices with highly mobile users.

C. Software Defined Networking

Software Defined Networking (SDN) simplifies network management, introducing programmability and open network access by decoupling the control plane from the data plane and via logically centralizing network intelligence. SDN provides key characteristics such as flexibility, service-oriented adaptation, scalability, and robustness [72], which are essential for enabling network slicing. A SDN controller facilitates third parties with an abstracted network view, i.e. via the virtualizer, and through the means of an agent, it allows multi-tenancy support [73]. Each tenant is assigned a policy that governs its capabilities to program the underlying data layer using the data control plane function.

The SDN paradigm is elaborated from the service perspective in [74], considering closed control loop means for maintaining the desired performance, e.g. in terms of latency, while a network slicing analysis is considered in [75] with every SDN client context representing a potential slice as shown in Fig. 2. By providing a set of network resource abstraction, resource groups, which also support control plane logic, form a functioning network slice. The SDN controller manages network slices effectively by applying rules when necessary and in accordance with the corresponding network policy. This architecture greatly impacts 5G network slicing from the perspectives of both the control and data planes considering the flexibility offered through the use of context. In particular, a SDN controller can maintain a distinct slice client context originating from different sources. This allows a SDN controller to dynamically manage network slices through grouping of slices belonging to the same context and maintaining a global map between the corresponding server-client context.

Some of the popular SDN solutions, which can benefit network slicing, include the following:

1) Open Network Operating System (ONOS): [76] is deployed as a service on a cluster of servers enabling rapid failure recovery and service scaling. ONOS allows potential tenants to easily create new network services without the

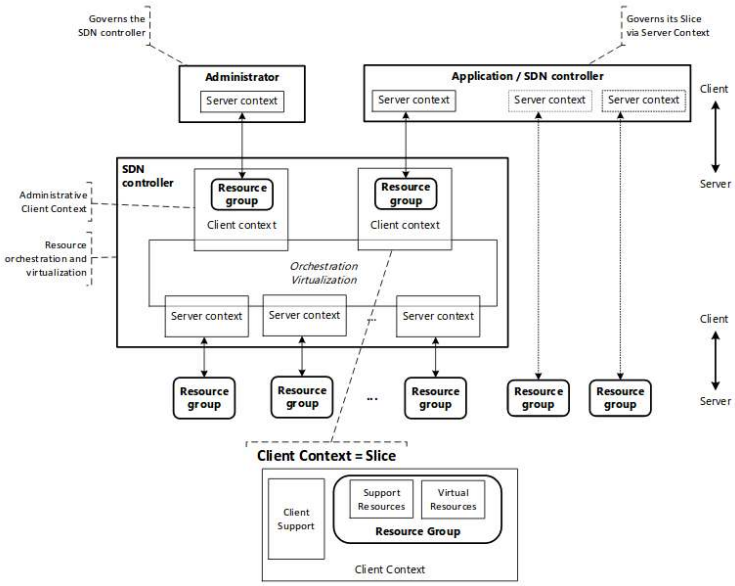

Figure 2: The SDN Architecture for Network Slicing [75].

need to alter the data plane systems and offers powerful northbound abstractions for DevOps. In the context of network slicing, ONOS can offer VNF composition in central office environment and VPN connectivity, e.g. via segment routing.

2) Mobile Central Office Re-architected as a Datacenter (M-CORD): [77] is a cloud-native solution based on CORD leveraging the benefits of SDN, NFV and agility of cloud computing to deliver cost effective mobile networking and operator specific services. M-CORD lays the foundations for 5G end-to-end slicing, by providing virtualization and programmability of RAN and mobile core, i.e. vEPC, enabling a service-oriented network arrangement, which can be scaled dynamically taking advantage of edge computing, real-time monitoring and analytics [78].

3) OpenDayLight (ODL): [79] is a modular open SDN platform offering network programmability, while facilitating customization and automation for networks of any size and scale. ODL comprises the foundation for commercial SDN solutions addressing: (i) automated service delivery considering also cloud and virtualization services, (ii) network resource optimization based on network load and state and (iii) network visibility and control towards third parties. ODL relies heavily on its micro-service architecture to provide dynamic, agile and programmable SDN services for optimizing existing networks to fit the needs of continuously evolving service demands.

D. Network Function Virtualization

NFV allows the deployment of originally hardware-based proprietary network functions on virtual environments leveraging the cost efficiency and time-to-market benefits of cloud computing [80]. VNFs are deployed on VMs, which can be chained together in a co-located or distributed cloud environment, offering network or value added services [85]. The NFV architectural framework [81] defines:

- VNFs that are software implementations of network functions deployed on virtual environments.
- NFV Infrastructure (NFVI), which comprises the logical environment's building blocks, i.e. storage, compute, network and their respective assisting hardware components.
- Management & Orchestration (MANO) that is responsible for managing and orchestrating VNFs and the NFVI.

In the context of network slicing, the NFV framework enables service chaining [83], capacity and latency-oriented VNF embedding as well as management of VNFs. In particular, NFV has introduced a flexible way of chaining VNFs by using a dynamic establishment of a Network Function Forwarding Graph (NFFG) that offers an efficient control on network functions [84]. The use of a NFFG enables an on-the-fly deployment of network services considering a diverse set of network functions, which may be virtualized or non-virtualized depending on the network service needs. The NFV orchestration [86], defined as the automation, management and operation of the distributed NFVI, is in charge of the network wide orchestration and management of both hardware and software and for delivering NFV services. To actualize the NFVI and service delivery, the MANO architecture defines the following three main functional components [82]:

- NFV Orchestrator (NFVO) is responsible for (i) network resource orchestration, i.e. NFVI orchestration with the help of the Virtual Infrastructure Manager (VIM) (which could be more than one), within an administrative domain, (ii) validating and authorizing NFVI resource requests from the VNF manager and (iii) network service life-cycle management. In coordination with the VNF manager, NFVO performs the orchestration and life cycle management of VNF service chains in a network slice.
- VNF Manager (VNFM) is responsible for the life cycle management of a single or multiple VNF instances of the same or different types, running in a network slice. Such a process includes VNF configuration and instantiation considering a network slice template, scaling in/out and up/down, and collecting NFVI performance information related to network slice instances.
- Virtualized Infrastructure Manager (VIM) controls and manages the NFVI associated resources usually belonging to a single network operator. Depending on its setup, a VIM may be dedicated, controlling a specific type of NFVI resource, e.g. a compute resource, or for managing multiple NFVI resources. For a network slice, the VIM allocates NFVI resources and manages their association, i.e. service chain, and traffic steering.

An overview of different orchestrator solutions is provided in [10] and [87], while two of the most comprehensive ones, that can be applied for network slicing, are the following:

1) ECOMP: [88] driven principally by AT&T. It is an open platform providing Enhanced Control, Orchestration, Management and Policy. It extends the ETSI (European Telecommunications Standards Institute) MANO architecture by launching distributed controllers to assure flexibility and reliability, alongside an end-to-end network service orchestrator called the Master Service Orchestrator (MSO). ECOMP introduces an infrastructure controller for computing resources,

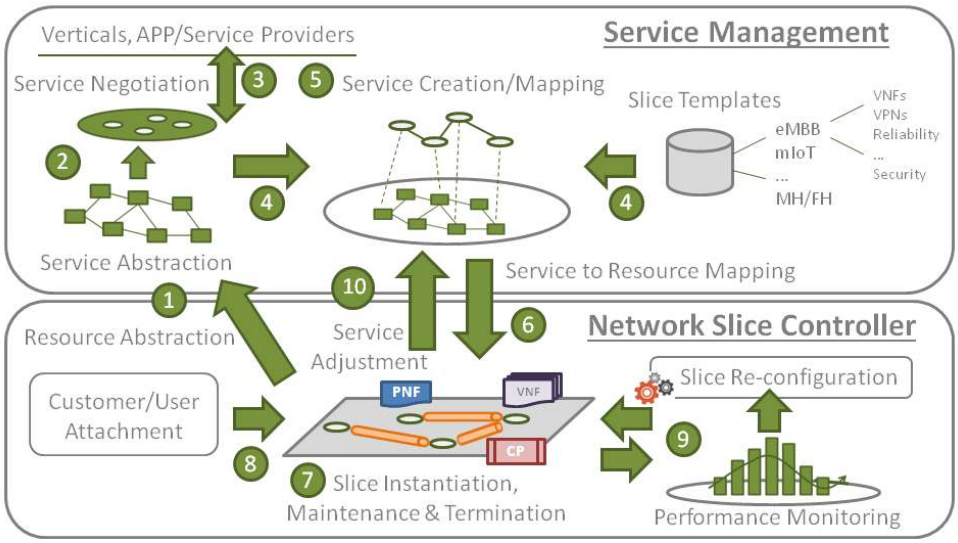

Figure 3: Service management and network slice control.

a network controller in charge of the network configuration and an application controller to take care of the application specific components. ECOMP uses the AT&T Service Design and Creation (ASDC) component to collect metadata about network services, supporting a data input format such as TOSCA and YANG. ECOMP is broadly divided into two major execution environments: (i) the design that defines and programs necessary system parameters, and (ii) the execution which specifies the logic for executing a closed-loop policy.

2) Open-O: [89] supported by the LINUX foundation. It leverages the benefits of a hierarchical placement of three orchestrators namely: the Global service orchestrator, NFVO and SDN to establish an end-to-end service orchestration platform. OPEN-O supports composite network services over both virtualized and physical network resources leveraging the benefits of the combined orchestration capabilities enabling agility to automate the orchestration of end-to-end services across multiple administrative domains. Open-O enhances service interoperability and scalability while it shortens the time to market for emerging services adopting industry-wide data models such as TOSCA, YANG, and REST APIs.

The merge of ECOMP and Open-O created the ONAP (Open Network Automation Platform) solution [90] under the Linux foundation leverages the benefits from both ECOMP and Open-O. ONAP creates a flexible, resilient and cost efficient end-to-end network service orchestration platform.

E. Cloud & Edge Computing

Cloud and edge computing offers storage, computational, and networking facilities within a single or multiple platforms for enabling a network slice [91]. Such basic infrastructure services can be offered by separate service providers such as MapReduce [92] and GoogleFS [93] respectively, or as an all-in-one higher infrastructure service, e.g. by Amazon, OpenStack, and Rackspace. Edge computing enables computing applications, data management and analytics, as well as service acquisition in close proximity to end users allowing a form of edge-centric networking, which facilitates data proximity, assuring ultra-low latency, high data rates, and intelligence and control as advocated by [94] [8]. Edge computing has numerous realizations. The ETSI Multi-access Edge Computing (MEC) [95] and Fog computing [96] are two of the most popular ones. MEC considers stand-along platforms focusing originally on RAN and later on fixed access networks, with the main object being the development of a northbound interface, which enables 3rd parties to instantiate and control various services from the network edge. Fog computing is introduced by Cisco to enable data transfer between connected wireless devices in an IoT system. It is a networking paradigm that considers a hierarchy of platforms that cooperate together to serve specific application needs.

VI. Network Slice Orchestration & Management

A. Service Management & Network Slice Control

Network slicing provides an end-to-end connectivity [97], allowing the co-existence of different network technologies on top of a common infrastructure [34] and relies on a continuous closed-loop process that analyzes the service requirements to assure the desired performance [98]. This process consists of:

- *Service management layer* handles service operations such as (i) abstraction, negotiation, admission control and charging for verticals and 3rd parties (steps 2 and 3) and (ii) service creation once a slice request is accepted based on the slice requirements in combination with the appropriate slice template (steps 4 and 5). The desired service combines VNFs, Physical Network Functions (PNFs), value added services, data/control plane, and security mechanisms, which are exposed to the underlying network (step 6).
- *Network slice control layer* provides resource abstraction to service management (step 1) and handles network slice resource management as well as control plane operations, including (i) instantiation of the slice resources based on

the service mapping (step 7), (ii) performance maintenance via monitoring, analysis and slice re-configuration procedures (step 9) and (iii) slice selection, attachment and support for multi-slice connectivity (step 8).

An overview of service management and slice control is shown in Fig.3. When resource re-configuration is not sufficient or latency cannot be fulfilled for assuring the desired service, the network slice controller contacts the service manager requesting an adjustment, i.e. an inter-slice re-configuration (step 10). Such adjustment may include modifications of service specific parameters and/or allocation of more topology, links or cloud resources.

B. Network Slice Orchestration Architecture

Network slices consist of VNFs, PNFs, value added services, network and cloud resources from dedicated or shared software and hardware in the RAN, transport and core networks, combining different technologies. A representative example of the network slice orchestration architecture considering the 5GPPP approach [99] is illustrated in Fig.4. The network slice orchestration architecture consists of:

- *End-to-end service management and orchestrator:* receives network slice requests from verticals and third parties, and creates a slice by performing slice brokering, admission control, policy provisioning and service mapping, considering the desired SLA, customization and slice template. It creates a network service graph that is passed on to the virtual resource orchestrator. This layer is also responsible for multi-domain slicing.
- *Virtual resource orchestration:* is responsible for VNF embedding and instantiation of the virtual network service graph and for performing all related MANO operations taking care of the life-cycle management of VNF instances as well as value-added services.
- *Network resource programmable controller:* facilitates flexible VNF service chaining, QoE control and resource programmability separating the control and data planes. The programmable controller can be (i) a network infrastructure's PNF offering network programmability on behalf of third parties or (ii) a VNF allowing third parties to directly program the allocated slice resources. A programmable resource can be a dedicated one or shared among different tenants. In the later case, the programmable controller enables short-term decisions (e.g. scheduling) and performs resource coordination (e.g. spectrum management) considering the assigned policies.
- *Life-cycle management:* performs legacy management, i.e. Operations, Administration and Management (OAM), element management, and policy provisioning.

Such an architecture considers a flexible separation of the control and data planes across a shared and dedicated network segments. Network operators and service providers can determine and offer a certain set of principles for establishing network slice orchestration considering the type of services. A realization of network slice orchestration architecture, focusing on the RAN and on distributed core network, is considered in [100] [101]. The architecture relies on the principles of multi-service and multi-tenancy support. A network slicing architecture for integrated 5G communications is analyzed in [102], which demonstrates its realization for LTE considering different orchestration and control technologies.

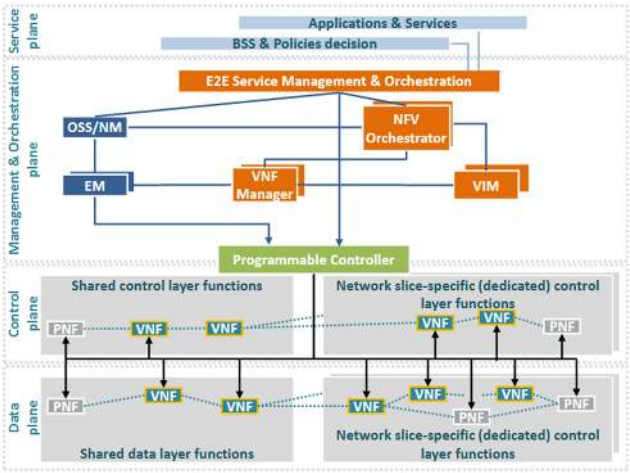

Figure 4: 5GPPP network orchestration architecture [99].

C. Network Slice Broker

For supporting cost-efficiency and ensuring good performance, network slicing uses a mechanism referred to as Network Slice Broker (NSB) [103], which facilitates an on-demand allocation of network resources performing admission control, resource negotiation and charging. The NSB considers a global network view based on the combination of network monitoring and traffic forecasting in order to assure resource availability, latency and resiliency for the duration of a slice request. It takes care of the inter-slice resource allocation and selects a configuration policy to guide the allocated resources. An economic analysis for allocating network slice requests, considering the maximum cost benefit from a mobile network operator's perspective based on optimal stopping theory, is elaborated in [104].

Network slice blueprints/templates are used to create a Network Slice Instance (NSI), which provides the network characteristics required by a Service Instance. A NSI may be dedicated or shared across multiple Service Instances. A network slice blueprint is a complete description of the structure, configuration and work flows for instantiating and controlling a NSI reflecting certain network characteristics (e.g. ultra-low latency and ultra-reliability) [34]. It refers to the required physical and logical resources, and the sub-networks. Network slices are designed to reflect the characteristic building blocks, encompassing the structure and configuration of the intended network service(s) defining a complete network architectural description guiding the instantiation and control of a NSI during its life-cycle [105]. On the other hand, a network slice template is a logical representation of network function(s) and the corresponding resource requirements to facilitate the required services and network capabilities [106].

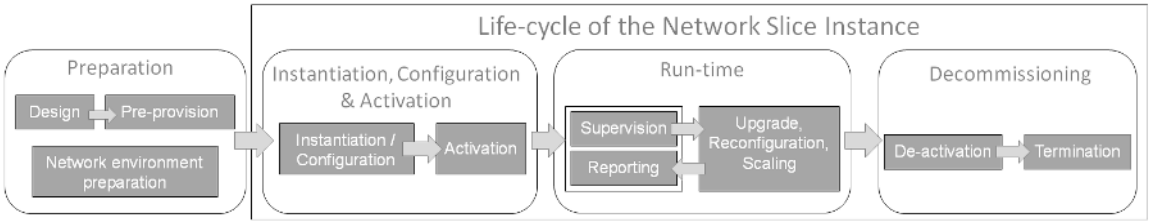

Figure 5: Network Slice Instance Life-cycle Management [113]

The resulting network slice architecture is then an abstraction of a completely functioning end-to-end mobile network, consisting of carefully connected PNFs and VNFs hosted in a virtual environment with the potential to be independent of the infrastructure provider. Following a standard slice blueprint, based on the ETSI NFV framework architecture, the resulting slice template should be independent of the deployment environment or virtualization service provider. Certain resource provisioning policies have to be agreed upon and enacted between the network slice owners and the network infrastructure provider based on SLA. Network services can be successfully carried out once certain network resources, i.e. computing and networking resources, are provisioned with an optimal level of consistency for the network slice duration.

Since network resources may be limited, especially in particular coverage areas or cloud platforms. It is therefore important to optimally provision the network and cloud capacity with respect to service performance needs. To this end, different resource provisioning and policy schemes have been proposed including both static and dynamic ones. In [107], an on-demand network capacity provisioning is elaborated with the admission control based on traffic forecasting considering also different traffic classes, i.e. guaranteed or best-effort service, and user mobility. A study on traffic forecasting methods for admission control with respect to network slicing has been considered in [108], wherein it is shown that Holt-Winters exponential smoothing suits better short-term prediction scenarios.

With respect to cloud resources, the work in [109] introduced the notion of sustained max, which enables a number of virtual instances for the entire life of a slice, and a set of dynamic policies including an on-demand and flexible resource provisioning, average queued time and multi-cloud optimization policy. For cloud providers, resource provisioning policies need to reflect both cost-awareness [110] and failure-awareness [111], while the respective algorithms should offer an on-demand and flexible resource dispatching and scheduling. For example, the work in [112] suggests the use of a three-step algorithm involving resource brokering, dispatch sequences and resource scheduling considering both deterministic and probabilistic methods, while utilizing checkpointing techniques for fault management.

D. Orchestration & Life-cycle Management

3GPP in the study document [113] decouples the life-cycle management of a NSI from the corresponding on-the-top service instance, which uses it. This allows scalability in provisioning network slices independently from service instances and facilitates efficient sharing among multiple services. To effectively manage a NSI, the following management procedures are considered: (i) fault management, (ii) performance management, (iii) configuration management and (iv) policy management. Service management takes place in the service provider's domain, while the management of network slices is performed in the network operator's domain. Such a paradigm requires a business to business interface, e.g. SLA. Based on this business relationship, a network operator can offer various levels of control to the service provider such as monitoring only, limited control to compose slices from a catalog or an extended control, where the service provider instantiates its own VNFs and MANO stack.

The life-cycle management of a NSI includes the following phases: (i) preparation, (ii) instantiation, configuration and activation, (iii) run-time and (iv) decommissioning as shown in Figure 5, which are administrated by the network operator. The preparation phase takes care of series of pre-NSI processes, which include the preparation of the network for the instantiation and support of a network slice that will be created via the respective network slice template. The instantiation, configuration and activation phase is broadly divided into the instantiation/configuration sub-phase wherein the necessary resources, both shared and dedicated including network functions, are configured and instantiated but not yet used, while in the activation sub-phase, the NSI becomes active handling network traffic and user context. The run-time phase focuses on data traffic supporting different types of communication services, while guiding and reporting the network service performance, which involves a closed loop management process that may suggest NSI re-configurations or scaling depending on the evolving needs. Finally, the decommissioning phase includes the deactivation and termination of the NSI reclaiming the allocated resources [113].

E. Federated Network Slicing

The nature of network slicing is end-to-end with slice segments potentially stretching across different administrative domains that follow different control and data planes. Irrespective of such federated nature of network slicing, third parties require a unified control plane and service abstraction with standardized APIs that make transparent the diversity of different administrative domains [114]–[117]. To achieve a unified control on top of a federated infrastructure, there is need for exchange points that perform the resource negotiation between different administrative domains. Currently two

distinct architectures exist, (i) the hierarchical architectures wherein different administrative domains are linked towards a higher layer slice broker capable to acquire and negotiate resources from different underlying administrative domains [118] and (ii) the flat or cluster-based architectures that directly link slice brokers from different administrative domains [119].

The resources that form a federated slice may include cloud resources, which can facilitate VNFs and value-added services in a single domain or across different domains as well as PNFs and network links forming a multi-domain structure [120]. Concave SLA parameters, such as bandwidth, can be handled easier, while delay or jitter, which have an additive nature, need to be divided among different domains. To allow a tighter control assuring that each domain does not surpass a recommended delay and/or jitter limit, a minimum/maximum per domain indication combined with a set of corrective measures, e.g. enhancing packet prioritization or re-routing, is essential. As different administrative domains adopt distinct abstraction structures and life-cycle management means, the notion of a self-contained slice segment within each administrative domain that represents abstracted resources and the associated management can assist the efficient service creation and control for end-to-end network slices [34]. Hence, a multi-domain slice simply combines such self-contained slice segments providing an over-the-top service control, while the underlying slice control is performed independently within each domain.

VII. RAN Slicing

A. Slicing Requirement at RAN Domain

The notion of network slicing in the RAN domain requires elasticity, efficient resource sharing and customization. These properties are needed at the RAN in order to adequately manage the scarce and limited frequency spectrum resources. In particular, RAN requirements with respect to network slicing include the following:

- *Dynamic resource management*: enables efficient resource sharing using sophisticated MAC scheduling functions, considering different Key Performance Indicators (KPIs) for each slice (e.g. an eMBB slice seeks high bandwidth, while an Ultra-Reliable Low-Latency Communications - URLLC - slice needs very low latency). As slice allocation can be performed on demand, even with a short duration, the corresponding resource management procedures should be flexible and programmable leveraging the benefits of open APIs.
- *Resource isolation and sharing*: end-to-end slices enable logical self-contained networks with the appropriate degree of isolation. Critical communication slices have stringent requirements for spectrum isolation due to latency and security assurance reasons that can be satisfied only when a hard spectrum slicing is employed. However, in the RAN domain, a hard spectrum isolation may prove to be a bottleneck due to limitations in multiplexing gains. Hence, there are different isolation requirements at the RAN level domain that consider specific resource management means to meet the corresponding KPIs.
- *Functional requirements*: each network slice may need a different control plane/user plane functional split, and a distinct VNF placement to ensure an optimal performance.

B. Slice Resource Management & Isolation

There are different slice resource management models depending on the level of resource isolation, which may handle frequency spectrum as a dedicated medium per slice or shared resource among specific slices. In the dedicated resource model, a RAN slice consists of isolated resources in terms of the control and user plane traffic, MAC scheduler and spectrum. Each slice has access to its own RRC/PDCP/RLC/MAC (Radio Resource Control, Packet Data Convergence Protocol, Radio Link Control, Medium Access Control) instances and a percentage of dedicated PRBs (Physical Resource Blocks) or a subset of channels. Although the dedicated resource model ensures committed resources per slice, i.e. assuring delay and capacity constraints, it reduces resource elasticity and limits the multiplexing gain. Indeed, the dedicated resource model restricts the slice owner to modify the amount of resources (i.e. PRB) committed to a slice during its life-cycle, even if they are not utilized. On the other hand, the shared resource model allows slices to share the control plane, MAC scheduler and spectrum. In particular, the PRBs belonging to the shared spectrum are managed by a common scheduler that allocates resources to slices according to a specified policy and other business criteria. Whilst this solution exploits statistical scheduling of physical resources ensuring elasticity, it lacks the support of strict QoS guarantees and traffic isolation.

The resource sharing model exploits the experience acquired from numerous a-priori studies on RAN sharing, which address spectrum sharing among Mobile Virtual Network Operators (MVNO) by providing modifications on the MAC scheduler. A preliminary approach for virtualizing an LTE eNB via the means of introducing a hypervisor is described in [121], taking into account radio conditions and traffic load. In advancing the basic eNB virtualization, the work in [122] introduces the notion of Network Virtualization Substrate (NVS), which operates on top of the MAC scheduler. Its objective is to flexibly allocate shared resources modifying the MAC scheduler to reflect MVNO's traffic needs considering the corresponding SLA. A network-wide radio resource management framework for RAN sharing is detailed in [123]. A mixture of reserved and shared resources by modifying the MAC scheduler is proposed in [124], introducing enhanced flexibility in allocating resources. Arguing that most of the MAC schedulers for RAN sharing consider only SLA-based resource sharing, an application-oriented RAN sharing solution, referred to as AppRAN, is proposed in [125]. The aim is to adapt the RAN sharing mechanism to the applications' needs in terms of QoS.

Such proposed schemes fall into the category of joint scheduling where optimization models with multi-objective functions try to satisfy a number of heterogeneous slice

requirements, such as latency and throughput, with the optimal solution being NP hard. To relax this constraint, the work in [126] introduces a two-level scheduler to share PRB among network slices. The first level, referred to as Slice Resource Manager (SRM), is a slice-tailored scheduler, which allocates virtual RB (vRB) to UEs belonging to a slice. The second level is an inter-slice scheduling process, called Resource Manager (RM), which aims at translating the SRM allocation (i.e. vRB) to PRB. A designated policy at the RM ensures that a slice is not exceeding its allowed resources. A similar concept is also presented in [127], considering a RAN controller that guides the inter-slice resource allocation based on a predetermined policy. The RAN slicing problem in a multi-cell network in relation to the RRM (Radio Resource Management) functionalities, that can be used as a support for splitting the radio resources among the RAN slices, is analyzed in [128] considering different approaches in terms of the granularity in the assignment of radio resources, isolation and customization. A joint scheduling and backhauling multi-objective problem considering latency, throughput and resiliency is analyzed in [129], introducing a heuristic solution that relies on a slice-tailored resource allocation, scheduling, and path selection using adaptive routing. A multi-objective solution, considering capacity allocation per slice and load balancing among different slices, is elaborated in [130]. A capacity broker paradigm taking into account a range of capacity and spectrum sharing options is studied in [131].

C. RAN Programmability

RAN programmability, also referred to as Software Defined RAN (SD-RAN), is a key attribute of RAN slicing abstracting the underlying RAN resources and facilitating open APIs towards third parties via the means of a Service Orchestrator entity that dynamically manages the resources dedicated to a network slice. Among the earliest studies, SoftRAN [132] presents the idea of a big base station abstraction, aiming at managing dense network deployments via the means of separating the control plane from the data plane. In particular, the control plane is centralized (e.g., centralizing mobility management), with the exception of time-critical functions, such as downlink scheduling, which are distributed at base stations. Unlike SoftRAN, the work in [133] proposed a hierarchical architecture, called Connectivity Management as a Service (CMaaS), which consists of a four-layer hierarchy wherein each layer is responsible for a different operation considering time criticality. The lower layer consists of a UE controller, which manages the radio access technology selection for a user constrained by the local network status and applied policies. The succeeding layer contains the base station controller, which manages the time-sensitive radio resource management and scheduling with a local network view. The RAN controller layer controls a set of base stations with a regional view, while on the top, the network controller with a global network view manages services such as QoS, routing, and mobility management, and instructs the lower layer controllers.

Recently, the FlexRAN protocol [134] has defined and implemented a SD-RAN architecture based on the OpenAir-Interface (OAI) tool. FlexRAN performs RAN abstraction providing an open API and allowing RAN programmability, while defining a southbound API to translate third parties' instructions into a set of suitable configurations for the OAI eNB. FlexRAN relies on the concept of agents, wherein a central FlexRAN master controller communicates via the southbound APIs with a set of distributed agents hosted by the eNBs that perform time-critical control functions, e.g. a UE scheduling process. In [126], a new architecture is introduced to enforce RAN slices, using FlexRAN modifying the eNB by integrating a two-level MAC scheduler (as described in subsection VII-B) to share the RAN resources among different slices. The authors implemented such a two-level scheduler employing the OAI tool and demonstrated RAN slice's life-cycle management via FlexRAN. A SDN-based RAN programmability paradigm reflecting service requirements of third parties is elaborated in [135], allowing elastic resource sharing among Frequency Division Duplex (FDD) and Time Division Duplex (TDD) systems enhancing the flexibility of network resource management. An extension of such a scheme considering a service-oriented slice allocation is analyzed in [136], [137], wherein each slice can adopt a different UL/DL rate reflecting best the service requirements within a reserved amount of radio resources for the duration of the service, in order to avoid inter-slice interference at the cost of multiplexing gain.

xRAN (extensible RAN) [138] is an initiative formed by an industrial consortium to bring into light a new RAN architecture for designing the future programmable wireless dense networks. The objective of xRAN is to decouple the service from hardware by devising a generic and programmable substrate to realize a flexible multi-service network. The main components of xRAN include: (i) a decoupled data plane from control plane and via an open API that allows a programmable user plane, (ii) a software defined control plane to manage complex network arrangements and (iii) a slicing plane to share the physical infrastructure among multiple applications with customized network stacks. Recently, xRAN and CORD joined efforts to create a carrier grade open reference implementation of xRAN in the context of M-CORD. Besides control plane solutions, OpenRadio [139] and PRAN [140] deal with data plane programmability and allowing the deployment of new wireless protocols on-the-fly.

D. Flexible RAN Virtualization & Functional Split

RAN virtualization is based-on the notion of base station softwarization, which allows certain RAN functions to run at remote cloud platforms. Such a paradigm gained momentum within the emergence of the Cloud-RAN (C-RAN) concept [141] [142], where RAN functions are split between the Base Band Unit (BBU), hosted in the cloud, and Remote Radio Headers (RRHs) that provide antenna equipment and radio access. The initial C-RAN deployments considered a high capacity fronthaul network, typically based on optical technology to connect the BBU, that provides the corresponding RAN functionalities, with several RRHs. However, the wide

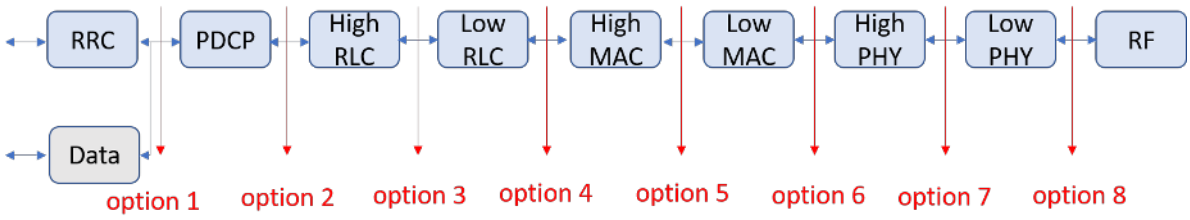

Figure 6: A high-level overview of the different functional split options.

availability of high speed optical links, especially in urban small cell environments, is questionable with many works such as the work in [143] studying more flexible solutions exploring the RAN functionality that is viable to move to a cloud environment. This gave rise to the concept of flexible functional split considering a range of different C-RAN deployment options based on the fronthaul capacity, latency and the time-critical nature of certain RAN functions taking into account the users' load and environmental conditions. Figure 6 provides an overview of the different functional split options. In the following, the most common ones are detailed further:

- PHY-layer option (option 6): this option provides the highest centralization and can be realized only with an ideal fronthaul, i.e. a high data rate and low-latency optical fiber.
- MAC-layer option (option 4): The MAC layer and the layers above it are virtualized and run on a BBU with real-time scheduling performed aggregately for multiple RRHs. This option leverages the benefits of connecting distributed RRH physical layers to a common MAC, which allows coordinated scheduling and dynamic point selection, i.e. Coordinated Multi-Point (CoMP). However, this option requires a low-latency fronthaul as some of the MAC procedures are time-critical (e.g. UE scheduling) and need to generate a configuration at the TTI (Transmission Time Interval) level.
- RLC-layer option (option 3): The RLC layer and other layers above it are virtualized at the BBU allowing multiple MAC entities to be associated with a common RLC entity. This option reduces the fronthaul latency constraints as real-time scheduling is performed locally in the RRH.
- PDCP-layer option (option 2): This option is non-time critical. It runs the PDCP functions at the BBU and may use any type of fronthaul network. The main advantage of this option is the possibility to have an aggregation of different RRH technologies (e.g. 5G, LTE, and WiFi).

The IEEE NGFI (Next Generation Fronthaul Interface) group studies the functional split from the fronthaul performance perspective, introducing in [144] a number of functional splits considering the interface bandwidth and latency requirements. 3GPP RAN3 study group introduced a study in [145] discussing the 5G RAN architecture that will have a major impact on mobile network architecture. In this study, the BBU consists of two new entities, named the Central Unit (CU), which may host time-tolerant functions, such as PDCP, and the Distributed Unit (DU) that holds time-critical functions, such as MAC and/or part of the physical layer functions. CU is envisioned to cover an area of 100-200 km radius, while DU should operate in an area of 10-20 km. The RAN3 group is also discussing: (i) the split between the CU and DU; (ii) the fronthaul split towards the RRH and (iii) the RAN internal split of the user plane and the control plane.

This flexible functional split can highly impact the performance of network slicing and the optimal split largely depends on the characteristics of the target service. For example, a uRLLC slice may require most RAN functions to run on DU in order to fulfill latency requirements, while in an eMBB slice, a higher centralization can enhance the throughput by aggregating RRHs (e.g., enabling Coordinated Multipoint - CoMP). In the context of network slicing, certain RAN functions can be also shared among different slices as elaborated in [127]. For example, each network slice may have its own instance of RRC (configured and tailored user plane protocol stack), PDCP, and RLC (non-real time functions), while the low RLC (real-time function), MAC scheduling (inter-slice scheduler) and physical layer can be shared. Some network slices may also have their own intra-slice application scheduler or tailor the RLC and PDCP functions to the specific slice type. For example, in a network slice supporting low latency, the header compression may not be used and RLC transparent mode may be configured, while a service requiring high QoS/QoE may activate an acknowledged RLC mode.

E. 5G Slicing & Fronthaul/Backhaul Integration

The emerging 5G networks introduce a heterogeneous fronthaul/backhaul landscape that consists of various technologies such as optical, millimeter-wave, Ethernet, and IP. Currently, network virtualization in the mobile backhaul relies on dedicated and overlay networks over a shared infrastructure [146], converging distinct transport network services into a unified infrastructure [147]. The catalyst for enabling such scalable multi-service mobile backhaul is Multi-Protocol Label Switching (MPLS) that supported the progressive adoption of different transport layer technologies unifying 2G Time Division Multiplexing (TDM) and High-level Data Link Control (HDLC) transport, 3G Asynchronous Transfer Mode (ATM) and Frame Relay as well as 4G Ethernet and IP [148] [149].

The stringent 5G RAN requirements, in terms of device and load density, and high mobility, are expected to shape the transport network layer facilitating enhanced capacity, high availability and an agile control. For the fronthaul/backhaul,

this means multi-path connectivity, tighter synchronization, coordination of both radio and transport layers and software defined control. To address latency, jitter and availability, Deterministic Networking (DetNet) [150] considers: (i) packet prioritization and buffer allocation to assure a maximum latency limit and jitter avoidance, (ii) congestion protection by synchronizing the rate between network nodes to avoid packet loss, and (iii) the use of multiple paths replicating data packets related to services that require high resiliency and availability.

Initial efforts to address such requirements considered: (i) small cell enhancements for the mobile backhaul focusing on scalable connectivity and various coordination types with the macro, i.e. tight or loose [151], and (ii) RAN centralization using the Common Public Radio Interface (CPRI) interface [152], which allows an ideal optical fiber fronthaul. A flexible SDN based architecture that can program the fronthaul network associating flexibly RRHs to a BBU and considering the dynamics of user and network performance is introduced in [153]. However, CPRI is expensive and difficult to deploy. Hence, alternative transport layer solutions, such as eCPRI (Enhanced CPRI) [154], are considered to relax such requirements enabling a flexible base station split.

In principle, different flavors of base station split can offer a particular service performance, requiring a distinct capacity and delay from the transport network layer. An integrated fronthaul/backhaul architecture, i.e. offering fronthaul/backhaul services on common links, can assure the desired performance by allowing a different centralization of the control and data planes for each service, while optimizing the network resource efficiency. Network slicing can assure isolation and performance guarantees between the different logical networks that employ a different fronthaul/backhaul flavor according to the corresponding base station functional split. Such an integrated fronthaul/backhaul architecture is based on a unified control plane and on a data plane that relies on network nodes capable of integrating different transport technologies for fronthaul and backhaul via a common data frame [155] [156]. A novel SDN-enabled access network architecture based on a smart gateway (Sm-GW) between the small cells and conventional backhaul gateways is introduced in [157]. It efficiently controls the resource allocation in backhaul links and provide multi-tenancy support facilitating sharing within the small cell network.

VIII. Core Network Slicing

The mobile core network has gone through a significant evolution during this last decade. Starting with LTE, which introduced a full IP core network, passing via the softwarization and the virtualization of the core network elements and ending with 5G and network slicing. Indeed, the need for more flexibility and elasticity has led to the consideration of SDN and NFV as the key enablers for more dynamic EPC networks, paving the way to a network of capabilities. To this end, the 3GPP has reshaped completely the core network, by defining a more modular architecture, wherein the main EPC entities have been divided into granular network functions. In addition, 3GPP has adapted solutions, such as eDecor, and built specifications (e.g. network slice discovery and selection, and network function sharing) to enable network slicing through the creation of core network instances for different types of services.

A. EPC Virtualization

The mobile core network is a major part of the mobile network serving as the flagship of a mobile network provider. The advent of network softwarization and advancements in NFV in general has served as major enabling technologies in virtualizing the core network. Core network entities such as MME, Home Subscriber Server (HSS), Packet data network GateWay (PGW), Serving GateWay (SGW), and Policy and Charging Rules Function (PCRF) can now be deployed on sophisticated virtual platforms, thanks to the ever progressive standardization activities going on in the area of NFV [80]–[82]. The fact that these building block entities of the mobile core network can be deployed as virtual instances brings more flexibility, elasticity and QoS assurance to the service provisioning techniques of the EPC network. The flexibility and elasticity in service provisioning implies that a mobile network operator can now deploy multiple instances of the EPC, all at the same time, to serve different categories of users based on their service requirements. Moreover, while some services may need all the components that constitute the EPC, others do not (e.g., a mIoT service with limited mobility does not need MME in the EPC). Therefore, the notion of core network slicing is centered around the possibility to deploy multiple instances of virtual EPC (vEPC) running in parallel in order for each to fulfill different service demands, e.g. a delay-sensitive service may require a distributed vEPC closer to the end user.

Thanks to NFV and SDN, vEPC can be orchestrated and managed throughout its lifetime over cloud platforms. In fact, different orchestration schemes can be utilized offering an efficient management and operation of the EPC entities. Both the control plane (MME and HSS) and user plane (SGW and PGW) entities of the EPC can now be provided as a service on commodity servers, which brings new dimensions to the operating models of the MVNO's market. The EPC as a Service (EPCaaS) [158] work leverages the cost efficiency offered by deploying vEPCs on the cloud to introduce two major virtualization approaches in providing the EPCaaS model and four different architectural implementation scenarios. The EPCaaS framework firstly proposed a full virtualization approach where both the control and user plane entities of the EPC are virtualized. For the second approach, only the control plane entities are virtualized while the data plane components are deployed on proprietary hardware and that is to ensure high throughputs and to enable the implementation of traffic inspection policies. The suggested implementation scenarios include a 1:1 mapping; wherein each EPC functional component runs on a VM, 1:N mapping; wherein each EPC functional component runs on multiple VMs, N:1 mapping; wherein all EPC functional components run on a VM, and N:2 mapping; wherein the control plane and data plane components of the EPC run on a VM each.

Similarly, the carrier cloud work in [159] demonstrates not only how the EPC can be offered as a service on a

commercial off-the-shelf server, but also how to expand and enlarge the service model of cloud infrastructure providers from only providing computing and storage capabilities as a service in data centers to also enabling end-to-end mobile connectivity as a service. The carrier cloud work explained how LTE, EPC and even PDN can all be offered as a service i.e (LTEaaS), (EPCaaS) and (PDNaaS), in fact anything can be offered also as a service [160] over the cloud by introducing four important stakeholders who are primary enablers in the cloud service providers' architecture, namely the service provider, carrier cloud service platform, virtual infrastructure and Physical infrastructure. All of these works are performed towards enabling the support of diverse service requirements, which would allow the realization of 5G network slicing.

B. Dedicated Core Networks

Aiming at providing a Dedicated Core Network (DCN) set of EPC functions to a specific group of users or services, e.g. IoT, 3GPP introduced the DECOR concept in Release 13. DECOR constitutes an early form of network slicing, wherein eNBs can select the appropriate core network functions for the control and user planes of specific UEs. The DECOR network selection procedure assumes that a UE first connects to a default MME, which redirects the Non-Access Stratum (NAS) connection setup, based on HSS, to the appropriate MME. Such a process introduced a lot of signaling for redirecting UEs' traffic. As a remedy to this issue, the enhanced DECOR (eDECOR) [161] was developed. eDECOR relies on the involvement of UEs to reduce signaling. During the connection setup, i.e. RRC, a UE integrates a DCN selection parameter received by the network, which is operator specific and can map to a certain DCN.

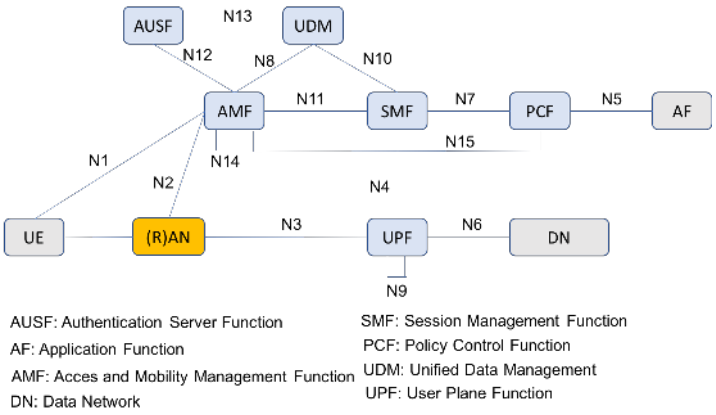

Figure 7: NG core architecture: point-to-point model.

eNBs are able to select the appropriate DCN based on such parameter sent by the UE in combination with a pre-configured logic. During the S1 setup procedure, MME provides the supported NAS type and the DCN selection assistance information to eNB as part of the S1 setup response message. In [126], the authors used eDECOR to enforce network slicing at the core network level and proposed replacing the DCN selection parameter by a slice ID. The latter could be hard encoded in UEs (i.e. UMTS Subscriber Identity Module - USIM) or encoded through the PLMN. As for eDECOR, UEs communicate the slice ID during the RRC connection procedure as well as in the NAS procedure, which allows eNBs to assign UEs to the requested slice(s) and handle their traffic according to the specified SLA.

C. Next Generation 5G Core network

The need for more flexibility, elasticity and scalability of the mobile core network has motivated 3GPP in [162] to introduce a new core network architecture (namely the Next Generation (NG) core or 5G system architecture), which separates the current EPC functions into more fine-granular network functions. The reshape of the EPC functions into more granular functions has been first suggested in [163], where the authors broke down the monolithic core network functions into more modular functions that constitute a control plane service. In a recent work [164], the same authors show how the modular functions could be composed to build a control plane service tailored to a network slice. Within the NG core architecture, some Network Functions (NFs) have their equivalents in LTE, while others are entirely new. Notably, the access control and session management are combined in EPC, but separated in NG core to better support fixed access and ensure scalability and flexibility. The most prominent NFs as defined in NG core are:

- Access and Mobility Management Function (AMF) - handles access control and mobility amongst others. AMF also integrates network slice selection functionality as part of its basic set of functions. In case of fixed access, the mobility management functionality becomes not required in AMF.
- Session Management Function (SMF) - is setup according to the network policy to handle user sessions.
- Policy Control Function (PCF) - corresponds to PCRF in LTE. This function integrates a policy framework for network slicing.
- User Plane Function (UPF) - can be deployed based on the service type, in several configurations and locations.
- Unified Data Management (UDM) - is similar to HSS in LTE. However, it is envisioned to integrate subscriber information for both fixed and mobile accesses in NG core.
- NF Repository Function (NRF) - is a new function. It provides registration and discovery functionality allowing NFs to discover each other and communicate via open APIs.

The NG core architecture is assumed to be deployed within two phases. In the first phase, the different components are connected using a point-to-point connection, based on reference interfaces, similar to the current LTE architecture. Fig. 7 illustrates the NG system architecture using the reference point representation. Similar to a traditional 3GPP architecture, the proposed architecture connects the different 5G core NFs together and with UEs as well as the Access Network (AN) via reference interfaces. However, this way of defining the NG

core network architecture may introduce complexity to add new network elements/instances, as it requires the operator to reconfigure multiple end-to-end interfaces.

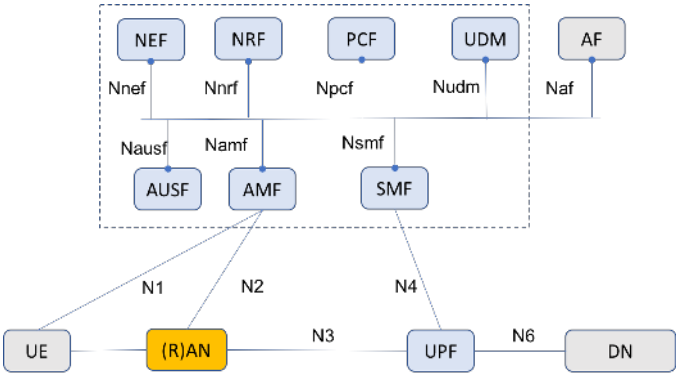

Figure 8: NG core architecture: service-oriented model.

The point-to-point oriented architecture is expected to evolve in the second phase to a service-oriented architecture. Whilst the latter incorporates the same functional elements and the same user-plane processing path between the UE and external data networks, it differs in the control plane (Fig. 8). In place of the predefined interfaces between elements, a service model is used in which components query a NRF to discover and communicate with each other. In the new model, the network functions use the concept of producer/consumer, where a NF may register for specific events provided by another NF via an API. Thus, either 1:1 or 1:N communication is possible.

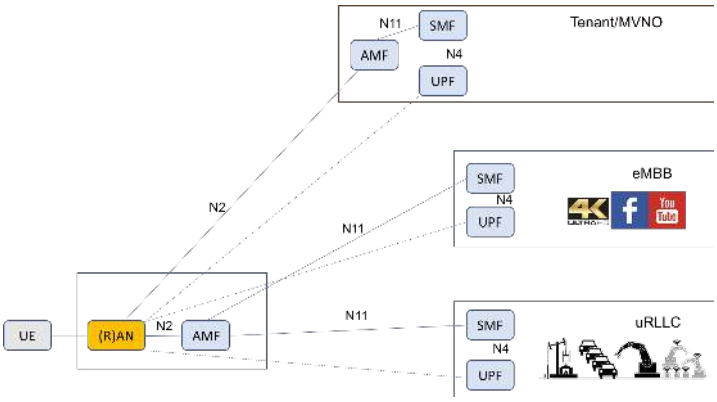

Figure 9: Example of AMF sharing among network slices.

For instance, the AMF service exposes to other NFs information regarding the mobility related events and statistics. Similarly, PCF provides all operations related to policy rules to other NFs. The service-oriented NG core architecture could be easily deployed in a software-based/virtualized environment (e.g. VM or container), wherein libraries of functions may be requested from a VNF catalog and composed into end-to-end service chains on demand. In addition, the composition of specific function would enable tailoring 5G core network to network slice specific.

Besides a tailored NF composition, one of the advantages of having fine-granular NFs for network slicing is the possibility to share some NFs among slices aiming at: (i) reducing the management complexity of network slices by sharing the Authentication Server Function (AUSF) and Unified Data Management (UDM), mobility management procedure (i.e. AMF); (ii) reducing the signaling over the air – the higher the number of shared core network control plane NFs are, the lower the signaling load is; and (iii) managing a common hardware, in the case of NFs which cannot be deployed in a software environment (i.e. PNFs such as eNBs). Accordingly, it is important to identify common NFs that need to be shared by several end-to-end network slices, benefiting from the reshape of RAN and core network functions (i.e. fine granular) thanks to the functional split and the NG core architecture.

In [31], three groups are discussed. Each group has a set of common NF among network slices. The first group (A) is similar to the concept of eDECOR, where the RAN NFs are common to all network slices, while each network slice has a completely dedicated core NFs. Thus, a UE may obtain services from different network slices and different core network instances, which allows logical separation and isolation of the core NFs. In this case, the dedicated core NFs could be subscription management, mobility management or session management (e.g. AUSF, AMF, SMF, and UDM). This solution may ease isolation between the core NFs, but it may increase the signaling overhead. The second group (B), which is also used in [127], assumes that some NFs are common between the network slices (e.g. RLC low, MAC scheduler, AUSF, UDM, and AMF), while others are slice specific (e.g. PDCP, RRC, UPF, SMF). The third group C considers that the control plane management is common between the slices, while the user plane(s) are handled as different network slices (e.g. PDCP and UPF). Figure 9 illustrates the share of AMF function among the eMBB and uRLLC slices, while a MVNO has its own AMF function.

D. Network Exposure Function

The relationship between a network slice provider and a slice consumer is governed by a SLA between the two parties. This agreement should define the terms and conditions of service, the level of exposure and the amount of control the slice consumer/tenants would have over the operation of the network slice. Mobile networks may offer four essential functionalities including communications (e.g., voice communication service and SMS), context (i.e., real-time user info such as users' locations and profiles), subscription (e.g., subscription identity) and control (e.g., policy and security), which can be exposed to third parties such as MVNOs, application providers and verticals via secured network open APIs. In an initial study, 3GPP introduced the Service Capability Exposure Function (SCEF) [165], which resides in the MNO's secure domain to expose the aforementioned network services while providing vital network functions such as authentication/authorization and secure network access.

The Network Exposure Function (NEF) [166] is a SCEF evolution and defines an API layer related with the 5G NG core, which can effectively control network slices. The level of control may depend on the network slice types, based in turn on the functional and non-functional requirements the network slice should embody. The network slice consumers should be exposed to a multifaceted levels of control for a successful delivery of network services by their respective network slices. The multifaceted levels of slice control exposure have to be properly managed between different network slices by the network slice provider. This may imply that the network slice provider needs to implement sophisticated network slice management interfaces which will create abstraction layers of interfaces, which can be easily mapped to the different levels of network slice control exposed to the slice owners/tenants. In the context of this article, the following are common levels of exposure and control a slice tenant could have over the allocated network slice:

- *Basic or passive level control*: In this level of control, a slice tenant can view and monitor the performance of a network slice, e.g. via a web interface. Through the same web interface, the slice owner can place a network slice configuration request from a pre-defined slice catalog, which may be made available by the network slice provider. This level of control only allows the network slice owner to carry out a passive form of control.
- *Extended control and management*: In this level of control, a slice tenant does not only monitor or carry out passive control, but can also update (e.g., scale up or down) existing network slice's functionalities by introducing different slice configurations based on both the available and desired network functions, deployable by the network slice provider. Using interfaces provided by the network slice provider, the slice consumer can change the composition of a network slice to suit the service needs. This form of control is a step ahead of the previous level in the sense that it gives a slice tenant additional privileges to perform an active form of network control, hence, its naming as active network control.
- *Full control and management*: In this level of control, a slice tenant has the highest level of network slice control and privileges. This is because the SLA allows the slice consumer to operate and manage his/her own virtualization platform and other related management support systems, and hence can deploy any form of network functions as needed to deliver the desired set of network services. However, the slice consumer is not allowed to redesign or change the composition of the network functions available from the network operator, which can be deployed on top of the allocated virtual platform.

By leveraging the benefits of NEF, 5G networks will be able to cope with and handle the business requirements and service demands of third parties. Based on the SLA policy and network service adaptation provided by the network, vertical industries and third parties can request the creation of customized service capabilities for their applications, and optimally utilize these exposed network capabilities to efficiently exploit the allocated network resources.

E. Slice Discovery & Selection

The network slice selection and discovery functionality consists of associating a user with the appropriate slice instance. In [162], 3GPP has defined a first set of procedures to enable network slice selection and discovery. The proposed procedures rely on the introduction of the Network Slice Selection Assistance Information (NSSAI), which is sent by the UE during the RRC and NAS-message registration procedures. NSSAI is a vector of max 8 single-NSSAI (S-NSSAI) values that are used to identify and select slice instances. Fundamentally, NSSAI is composed of: (i) a slice/service type (SST) ID, which corresponds to the expected network slice behavior in terms of features and services; and (ii) an optional information, namely Slice Differentiator (SD) to allow further differentiation for selecting a network slice from multiple network slice instances, fulfilling the indicated slice/service type functionality. S-NSSAI can have standard values or PLMN-related values. The S-NSSAI standard values are about mainly the SST part. The objective beneath standard SST values is to ensure a global interoperability for slicing, allowing roaming for the most commonly-used slices/services, such as eMBB, uRLLC and mIoT. For example, the eMBB slice type has SST=1, the uRLLC slice type has SST=2, and SST=3 corresponds to the mIoT slice type.

Once the access network (i.e., RAN or fixed - (R)AN) receives the RRC connection establishment message, it extracts the NSSAI information and relays the NAS messages to an AMF entity associated with NSSAI. The contacted AMF starts the slice selection using NRF services and sets up a Protocol Data Unit (PDU) session with the proper user plane functions for the slice. In fact, AMF selects SMF in a network slice to establish a PDU-session upon the reception of SM message. The SMF discovery/selection is either based on NRF or S-NSSAI and Data Network Name (DNN). The selected SMF establishes a PDU-session using S-NSSAI and DNN. As a single UE may be served by at most 8 network slices at a time, it is required that AMF should logically belong to each network slice, hence AMF shall be common to the network slices serving a UE. Figure 9 illustrates the case of shared AMF among network slices. If the (R)AN is unable to select an AMF based on NSSAI, it routes the NAS signaling to a default AMF from a set of AMFs. The use of the default AMF is also required, when the UE does not indicate an NSSAI value in the RRC message. The PDU session establishment in a network slice enables data transmission in a network slice. Usually, a Data Network (DN) is associated with an S-NSSAI and a DNN. If a UE is associated with different network slices, the operator may provision the UE with Network Slice Selection Policy (NSSP). The NSSP consists of a set of rules (at least one rule), where each rule attributes an application with a certain S-NSSAI. A default rule may exist that matches all applications to an S-NSSAI.

F. Multiple Slice Connectivity

It is envisaged that UEs can potentially connect to multiple network slices. Depending on the capacity and design capabilities of these UEs, they will leverage this possibility to obtain the best network connection experience and network services. Based on this vision, 3GPP in [31] considered designing the NG core to suit the needs of the following three UE categories:

- UEs can obtain connectivity from multiple slices belonging to different core networks with a logical separation from each other on both the control and user plane, i.e. without sharing any NFs. Such logical separation would enable isolation, which guarantees inter-network slice security, but at the cost of additional signaling overhead.
- UEs can connect to a single core network at a time. They therefore can only access slices belonging to the same core network. For this category, even though the user plane of the individual slice instance is separated, some control plane functions are still shared. Thus, network slices only have a partial level of isolation depending on the number of shared NFs.
- UEs share the same characteristics with the previous category except that all slices share all NFs on the control plane. Hence, the probability of network slice compromise is the highest.

IX. Challenges and Open Research Problems

Similar to other emerging technologies, there is no doubt that network slicing brings forward a significant potential, but also introduces several technical and business challenges. Although network slicing is currently undergoing a standardization phase, there are still numerous open research problems and implementation challenges to be addressed. Some of which include, but are not limited, to techno-economics, slicing architecture, security, inter-working and capability exposure APIs, slice optimality, and UE slicing.

A. Network Slicing Techno-economics Aspects

The planning and development of network slicing is seen as a key enabler in launching new business opportunities, while driving down both OPEX and CAPEX in a mobile network. Network slicing profits concentrate on offered network capabilities, such as NFs and performance assurance, while value-added services include big data, localization, and edge computing [167]. However, charging models regarding network capabilities, dynamic slice usage, value-added services, and management and orchestration are still open. Focusing on the impact on CAPEX, the authors in [168] showed that the deployment of mobile networks on the foundation of SDN and NFV could save up to 13.85% of the cost of capital investment of network operators. Such savings on the investment cost could even rise to as high as more than half of the total cost of investment, when adopting active resource brokering in the form of network slicing. However, most, if not all, of these conclusions are based on simulated results and mathematical models. Understanding a practical insight would be an interesting challenge in need of further investigation.

B. RAN Slicing & Traffic Isolation

One remaining challenge regarding RAN slicing is the virtualization of the physical channel (i.e., RAN traffic isolation). Indeed, unless no beamforming is used, statically assigning PRB to UEs allows ensuring RAN traffic isolation on one hand but limits multiplexing gain. Traffic prediction-based slice adjustments can enhance multiplexing gains at the cost of network scalability, while a policy-based PRB allocation considering the entire spectrum may enhance further multiplexing gains but at the cost of security since there is no hard spectrum isolation. 5G is also expected to rely on beamforming; hence new solutions to virtualize the physical channel and enable RAN slices to benefit from beamforming are needed. Such solutions should create a specific physical channel for a slice, requiring that each physical layer should process the common in-phase quadrature phase (I/Q) flow (e.g., orthogonal frequency-division multiplexing (OFDM) modulation/demodulation in LTE) before being able to decode the traffic dedicated to UE from one slice; which is very resource consuming and inefficient. However, user-specific physical layer processing (e.g., turbo coding/decoding in LTE) can be virtualized (i.e., shared), depending on the required level of isolation.

C. Slice Security

Security concerns in virtual or cloud platforms are well determined [169] [170]. However, with regards to the security challenges on network slicing, the threats are multifaceted. First is the issue of inter-slice security threats and in case of a federated architecture, the second is the security challenges of network resource harmonization and amalgamation between inter-domain slice segments. Further security challenges raise due to the different layers of interaction between the technology domains providing the slice resources, the tenants sharing the resources and the different levels of exposure that exist between the network slices and the hosted tenants. Based on the current network slicing architectures, different levels of interactions and layers of isolations exist between distinct slices depending on the business and functional requirements. Network slices use these interaction interfaces to exchange certain vital information about the network state and network resources. The higher the number of VNFs that network slices may have in common or share, the higher the level of security vulnerabilities which may exist between them. In order to cater for the variabilities in the levels of security that may exist between slices, the SDD (Service Description Document) project [171] for network slicing proposed the use of additional quantitative or qualitative parameters to distinguish the levels of security required by individual slices.

Slices, though with different requirements, may be orchestrated using resources across multiple technological domains. Each of these domains have different levels of abstractions for their respective underlying physical infrastructures, which may have been implemented using different state of the art security detection and possibly prevention techniques. Orchestrating a network slice across such different proprietary virtual platforms with their respective unique security properties will

expose the resulting network slice(s) to different forms of both intra-slice security threat i.e., security threats within the orchestrated slices, and inter-domain security i.e., security between the domains from where the slice resources are orchestrated. In the case of network sharing involving slices of different tenants, the security vulnerabilities are even more complex. The fact that each of the operating tenants may share resources alongside their isolated network slices with different security parameters exposes their individual network slices to a form of intra-tenant security threat.

D. Slice Optimality/Pareto Optimality

One of the most difficult challenges faced in network slicing is how to optimally describe, define, and dynamically adapt network slice templates (i.e., dynamic slice resource allocation and optimal VNF placement). Traditionally, virtual networks are created by statically allocating a set of expected amount of network resources i.e. networking, processing/computing and storage. This method is often non-optimal, in the sense that, network resources are usually either over-provisioned or under-provisioned. This is due to the fact that network slices, especially those which are not deployed for a known number of users, often face the challenge of insufficient resources as the number of users increases, thereby resulting into poor network performance. To mitigate this challenge, efficient network slice resource allocation algorithms have to incorporate a form of Pareto optimality techniques, whereby network slice resources can be dynamically scaled up/down or in/out to optimally serve the total number of slice consumers without leaving any other active network slices with insufficient amount of resources i.e., negatively impacting the performance of any active network slices. Technically, developing such flexible optimal slice resource allocation is very challenging, particularly when considering a set of functional requirements that should be known before hand in order to be fulfilled. In addition, it would be even more challenging and complex if unprecedented network behaviors are taken into account in the algorithm's design.

E. UE Slicing

In 4G, UEs are not differentiated from each other in terms of service demands and functional requirements. Hence, the network treats them in the same way. Conversely, 5G networks are expected to handle UEs based on their individual characteristics or usage-class types. This implies that every UE shall be connected not to a single size-fits-all network anymore but to a customized slice, which is specifically created for that type of UE. For instance, a UE belonging to the CriC usage-class type will be connected to the CriC network slice, similarly, a UE belonging to the eV2X usage-class type will be connected to the eV2X network slice, thereby guaranteeing a high degree of QoS.

Bearing in mind the aforementioned 5G network features, some further research efforts introduce a new concept of slicing in the UE, in particular, the portable and smart devices such as phones, tablets, pads and potentially laptops. The UE slicing is supposed to bring about more freedom, customization and huge range of application availability to the end users. UE slicing considers the smart devices as a commodity hardware platform having a pre-installed middleware (similar to a hypervisor on general-purpose servers or PCs), which can accommodate, manage and schedule resources between multiple mobile OS entities. These OSs shall be installed on logical container partitions created by the middleware on the smart hardware commodities and shall manage the resources between them. These logical partitions of the OSs will present the UE as a platform where slices of different OSs are running. The OSs would retain the customizable features allowing them to install and run their respective applications, thereby providing end users with a rich source of applications running on slices of the UE.

F. Network Slicing Architecture Evolution & APIs

Current research and standardization activities around 5G pointed out the need for an enhanced mobile and transport network architecture to support new radio technologies such as 5G New Radio (NG) and millimeter wave. In addition, such 5G architecture should support a timely launch of new services and resource sharing via virtualization, enabling in this way 5G network slicing. To achieve this, a tighter integration of networking and cloud computing technologies is needed offering a flexible degree of customization for network functions and value-added services. Although several ongoing works address particular architectural issues related with 5G, it is still a challenge to provide a comprehensive solution considering new radio, cloud and service aspects, which can support the tight requirements of all emerging 5G use-cases.

Current 4G networks rely on a statically pre-configured transport layer network that is responsible to carry the traffic related to different GPRS Tunneling Protocol (GTP) bearers. With the introduction of network slicing on the mobile network layer, new requirements related to on-demand, autonomous and dynamic network configuration are also raised for the transport network layer. To efficiently enable this, there is need for a new interface between the mobile network and the transport network as identified by the 3GPP Network Management Working Group SA5 in [113]. Such an interface is envisioned to connect the management system of the mobile network with the transport network controller. Its main objective is to facilitate the capability exposure of the transport network and the mapping of a mobile network slice to the underlying transport network resources. The main challenges for achieving this are the development of: (i) data models that shall reflect the network slicing requirements related to the transport network, (ii) the processes and network functions responsible for mapping the requirements of a mobile network slice to the transport network resources considering the underlying transport technology and (iii) a network slice database to keep track of the resources related with a transport network slice for assisting the mapping process, considering both common and dedicated resources.

X. Conclusions

This article presents a comprehensive survey regarding the current maturity state of network slicing in 5G. It provides insights into the historical heritage of network slicing from network sharing in the early days and its evolution as the major bedrock of the 5G technology. It also discusses the main concept and principles, introducing the enablers of network slicing such as the NFV, SDN and cloud technologies. This paper presents the 5G services and business drivers as well as the impact of network slicing across the RAN, the core network and the transport network; thereby making it a survey of an end-to-end network slicing. By presenting relevant projects with regards to the orchestration of end-to-end network slices and management, we successfully reflect the importance of network slicing as a major enabler for 5G. With regards to end-to-end network slicing, this survey describes how network slicing can be achieved by slicing the RAN and core networks, while describing practical examples. Finally, we observe and discuss open research challenges in line with the realization of end-to-end network slicing in 5G mobile networks.

Acknowledgment

This work was partially funded by the Academy of Finland Project CSN - under Grant Agreement 311654 and also partially supported by the European Union's Horizon 2020 research and innovation program under the 5G!Pagoda project and the Global5G.org project with grant agreement No. 723172 and No. 761816, respectively.

References

[1] Two million devices an hour to be connected in 2025, Iteuropa.com, 2017. [Online]: Available: http://www.iteuropa.com/?q=two-million-devices-hour-be-connected-2025.
[2] A. Manzalini, et al., "Towards 5G Software-Defined Ecosystems: Technical Challenges, Business Sustainability and Policy Issues", IEEE Future Directions White paper, Jul. 2016.
[3] P. Frangoudis, L. Yala, A. Ksentini, and T. Taleb, "An architecture for on-demand Service Deployment over a Telco CDN", IEEE ICC, Kuala Lumpur, May 2016.
[4] S. Retal, M. Bagaa, T. Taleb, and H. Flinck, "Content Delivery Network Slicing: QoE and Cost Awareness", IEEE ICC 2017, Paris, May 2017.
[5] T. Taleb, B. Mada, M. Corici, A. Nakao, and H. Flinck, "PERMIT: Network Slicing for Personalized 5G Mobile Telecommunications", IEEE Communications Magazine, Vol. 55, No. 5, pp. 88–93, May 2017.
[6] X. Foukas, G. Patounas, A. Elmokashfi and M. Marina, "Network Slicing in 5G: Survey and Challenges", IEEE Communications Magazine, Vol. 55, No. 5, pp. 94-100, May 2017.
[7] M. Richart, J. Baliosian, J. Serrat and J. Gorricho, "Resource Slicing in Virtual Wireless Networks: A Survey", IEEE Transactions on Network and Service Management, Vol. 13, No. 3, pp. 462-476, Sep. 2016.
[8] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "Towards Mobile Edge Computing: A Survey of the Emerging 5G Network Edge Architecture", IEEE Communications Surveys & Tutorials, Vol.19, No.3, pp. 1657-1681, 3rd Quarter 2017.
[9] X. Foukas, G. Patounas, A. Elmokash, and M.K. Marina, "Network Slicing in 5G: Survey and Challenges", IEEE Communications Magazine, Vol.55, No.5, pp. 94-100, May 2017.
[10] R. Mijumbi, et al., "Network Function Virtualization: State-of-the-Art and Research Challenges", IEEE Communications Surveys & Tutorials, Vol.18, No. 1, pp. 236-262, 1st Quarter 2016.
[11] J. Gil Herrera and J. Botero, "Resource Allocation in NFV: A Comprehensive Survey", IEEE Transactions on Network and Service Management, Vol. 13, No. 3, pp. 518-532, Sep. 2016.
[12] V. Nguyen, A. Brunstrom, K. Grinnemo and J. Taheri, "SDN/NFV-based Mobile Packet Core Network Architectures: A Survey", IEEE Communications Surveys & Tutorials, Vol. 19, No. 3, pp. 1567-1602, 3rd Quarter 2017.
[13] A. Blenk, A. Basta, M. Reisslein and W. Kellerer, "Survey on Network Virtualization Hypervisors for Software Defined Networking", IEEE Communications Surveys & Tutorials, Vol. 18, No. 1, pp. 655-685, 1st Quarter 2016.
[14] Y. Li and M. Chen, "Software-Defined Network Function Virtualization: A Survey", IEEE Access, vol. 3, pp. 2542-2553, Dec. 2015.
[15] S. Lal, T. Taleb, and A. Dutta, "NFV: Security Threats and Best Practices", IEEE Communications Magazine, Vol. 55, No. 8, pp. 211-217, May 2017.
[16] ITU - RM.2083-0, IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and beyond, Sep. 2015.
[17] 5G Vision: The 5G Infrastructure Public Private Partnership: The Next Generation of Communication Networks and Services, Feb. 2015.
[18] A. Osseiran, et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project", IEEE Communications Magazine, Vol.52, No.5, pp. 26-35, May 2014.
[19] 5GPPP, 5G Empowering Vertical Industries, 2016.
[20] S. Nanda and T. Chiueh, "A Survey on Virtualization Technologies", [Online]: http://www.ecsl.cs.sunysb.edu/tr/TR179.pdf.
[21] S. Meier, "IBM Systems Virtualization: Servers, Storage, and Software", IBM Redbook, May 2008.
[22] R. Goldberg, "Survey of Virtual Machine Research", IEEE Computer, Vol. 7, No. 6, pp. 34-45, Jun. 1974.
[23] A. Lindquist, R. Seeber and L. Comeau, "A Time-Sharing System using an Associative Memory", IEEE Proceedings, Vol. 54, No. 12, pp. 1774-1779, Dec. 1966.
[24] L. Peterson, and T. Roscoe. "The design Principles of PlanetLab", ACM SIGOPS Operating Systems Review, Vol. 40, No. 1, pp. 11-16, Jan. 2006.
[25] PlanetLab: http://www.planet-lab.org
[26] NSF: https://www.nsf.gov
[27] M. Berman, et al., "GENI: A Federated Testbed for Innovative Network Experiments", Computer Networks, Vol.61, pp. 5-23, Mar. 2014.
[28] GENI: http://www.geni.net
[29] N. McKeown, et al., "Openflow: Enabling Innovation in Campus Networks", ACM SIGCOMM Computer Communication Review, Vol. 38, No. 2, pp. 69-74, Apr. 2008.
[30] NGMN Alliance, NGMN 5G WHITE PAPER, Feb. 2015.
[31] 3GPP TR 23.799, Study on Architecture for Next Generation System, Rel.14, Dec. 2016.
[32] ITU-T Y.3011, Framework of Network Virtualization for Future Networks, Next Generation Network - Future Networks, Jan. 2012.
[33] S. Shenker, "Fundamental Design Issues for the Future Internet", IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, pp. 1176-1188, Sep. 1995.
[34] NGMN Alliance, Description of Network Slicing Concept, NGMN 5G P1 Requirements & Architecture, Work Stream End-to-End Architecture, Version 1.0, Jan. 2016.
[35] I. Farris, T. Taleb, H. Flinck, and A. Iera, "Providing Ultra-Short Latency to User-Centric 5G Applications at the Mobile Network Edge," in Trans. on Emerging Telecomm. Technologies (ETT) Mar. 2017.
[36] Lookup Safety Recalls & Service Campaigns by VIN, Toyota.com, 2017.
[37] T. Taleb, A. Ksentini, and A. Kobbane, "Lightweight Mobile Core Networks for Machine Type Communications," in IEEE Access Magazine, Vol 2, Oct. 2014. pp.1128-1137.
[38] 3GPP TR 22.891, Study on New Services and Markets Technology Enablers, Rel.14, Sep. 2016.
[39] 3GPP TR 22.863, Feasibility Study on New Services and Markets Technology enablers for enhanced Mobile Broadband, Rel.14, Jun. 2016.
[40] 3GPP TR 22.862, Feasibility Study on New Services and Markets Technology enablers for Critical Communications, Rel.14, Jun. 2016.
[41] 3GPP TR 22.861, Feasibility Study on New Services and Markets Technology enablers for Massive Internet of Things, Rel.14, Jun. 2016.
[42] T. Taleb, S. Dutta, A. Ksentini, I. Muddesar, and H. Flinck "Mobile Edge Computing Potential in Making Cities Smarter," in IEEE Communications Magazine, Vol. 55, No. 3, Mar. 2017, pp. 38-43.
[43] AT&T Digital Life - Home Security & Automation Systems, https://www.att.com/digital-life.
[44] 3GPP TR 22.886, Study on enhancement of 3GPP Support for 5G V2X Services, Rel.15, Dec. 2016.
[45] 3GPP TR 22.951, Service Aspects and Requirements for Network Sharing, Rel.10, May 2011.
[46] GSMA, Mobile Infrastructure Sharing, Sep. 2012.

[47] 3GPP TS 23.251, Network Sharing; Architecture and Functional Description, Rel. 12, Mar. 2015.
[48] 3GPP TS 36.331, Radio Resource Control (RRC); Protocol specification, Rel.14, Jan. 2018.
[49] 3GPP TS 36.413, S1 Application Protocol (S1AP), Rel.14, Jan. 2018.
[50] 3GPP TS 36.423. X2 Application Protocol (X2AP), Rel.14, Jan. 2018.
[51] 3GPP TR 22.852, Study on RAN Sharing Enhancements, Rel.13, Sep. 2014.
[52] 3GPP TS 32.130, Telecommunication management; Network Sharing; Concepts and requirements, Rel.12, Dec. 2014.
[53] 3GPP TR 32.851, Study on Operations, Administration and Maintenance (OAM) aspects of Network Sharing, Rel.12, Dec. 2013.
[54] VM Server for SPARC — Virtualization — Oracle, Oracle.com, 2017.
[55] Xen Project's Lars Kurth, xenproject.org, 2017.
[56] vSphere ESXi Bare-Metal Hypervisor, Vmware.com, 2017.
[57] KVM, Linux-kvm.org, 2017.
[58] Oracle VM VirtualBox, Virtualbox.org, 2017.
[59] Mac Virtualization for Everyone: VMware Fusion, Vmware.com, 2017.
[60] Workstation for Windows - VMware Products, Vmware.com, 2017.
[61] VM Server for x86 — Virtualization — Oracle, Oracle.com, 2017.
[62] J.P. Walters, V. Chaudhary, M. Cha, S. Guercio Jr. and S. Gallo, "A Comparison of Virtualization Technologies for HPC", IEEE AINA, Okinawa, Mar. 2008.
[63] A. Carvalho, et al., "Full Virtualization on Low-end Hardware: A Case Study", IEEE IECON, Florence, Oct. 2016.
[64] S. Han and H. Jin, "Full Virtualization based ARINC 653 Partitioning", IEEE/AIAA DASC, Seattle, Oct. 2011.
[65] P. Barham, et al., "Xen and the Art of Virtualization", ACM SOSP, Bolton Landing, Oct. 2003.
[66] J. Smith and R. Nair, "The Architecture of Virtual Machines", Computer, Vol. 38, No. 5, pp. 32-38, May 2005.
[67] M. Xavier, et al., "Performance Evaluation of Container-Based Virtualization for High Performance Computing Environments", IEEE PDP, Belfast, Feb. 2013.
[68] Linux-VServer, Linux-vserver.org, 2017.
[69] OpenVZ Virtuozzo Containers Wiki, Openvz.org, 2017.
[70] Solaris Container, Oracle.com, 2017.
[71] Docker, docker.com, 2017.
[72] W. Xia, Y. Wen, C. Foh, D. Niyato and H. Xie, "A Survey on Software-Defined Networking", IEEE Communications Surveys & Tutorials, Vol. 17, No. 1, pp. 27-51, 1st Quarter 2015.
[73] ONF TR-504, SDN Architecture Overview, Issue 1, Nov. 2014.
[74] ONF TR-521, SDN Architecture Overview, Issue 1.1, 2016.
[75] ONF TR-526, Applying SDN Architecture to 5G Slicing, Issue 1, Apr. 2016.
[76] ONOS : An Overview - ONOS - Wiki, Wiki.onosproject.org, 2017.
[77] L. Peterson, et al., "Central office Re-architected as a Data Center", IEEE Communications Magazine, Vol. 54, No. 10, pp. 96-101, Oct. 2016.
[78] M-CORD: Mobile CORD - Enable 5G on CORD, opencord.org, 2017.
[79] Platform Overview — OpenDaylight, Opendaylight.org, 2017.
[80] ETSI NFV, Network Functions Virtualization: An Introduction,Benefits, Enablers, Challenges & Call for Action, White paper, Issue 1, SDN & OpenFlow World Congress, Daramstadt, Oct. 2012.
[81] ETSI NFV, Network Functions Virtualisation (NFV); Architectural framework , GS NFV 002, Oct. 2013.
[82] ETSI NFV, Network Functions Virtualisation (NFV); Management and Orchestration , GS NFV-MANO 001, Dec. 2014.
[83] A.M. Medhat, T. Taleb, A. Elmangoush, G. Carella, S. Covaci, and T. Magedanz, "Service Function Chaining in Next Generation Networks: State of the Art and Research Challenges," in IEEE Communications Magazine. Vol. 55, No. 2, Feb. 2017, pp. 216-223.
[84] A. Fischer, J. Botero, M. Beck, H. de Meer and X. Hesselbach, "Virtual Network Embedding: A Survey", IEEE Communications Surveys & Tutorials, Vol. 15, No. 4, pp. 1888-1906, 4th Quarter 2013.
[85] J. Gil Herrera and J. Botero Vega, "Network Functions Virtualization: A Survey", IEEE Latin America Transactions, Vol. 14, No. 2, pp. 983-997, Mar. 2016.
[86] ETSI NFV, Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV , GS NFV 002, Oct. 2013.
[87] R. Mijumbi et al., "Management and Orchestration Challenges in Network Functions Virtualization", IEEE Communications Magazine, Vol. 54, No. 1, pp. 98–105, Jan. 2016.
[88] ECOMP (Enhanced Control, Orchestration, Management & Policy) Architecture White Paper, AT&T Inc., 2016.
[89] OPEN-O Project at a Glance, open-o.org, 2017.
[90] ONAP (Open Network Automation Platform), onap.org, 2017.
[91] A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm, "What's inside the cloud? An Architectural Map of the Cloud Landscape", IEEE ICSE CLOUD, Vancouver, May 2009.
[92] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Communications of the ACM, Vol.51, No.1, Jan 2008.
[93] S. Ghemawat, H. Gobioff and S-T. Leung, "The Google File System", ACM SOSP, Bolton Landing, Oct. 2003.
[94] P.G. Lopez, et al., "Edge-centric Computing: Vision and Challenges", ACM SIGCOMM Computer Communication Review, Vol. 45, No. 5, pp. 37-42, Oct. 2015.
[95] ETSI NFV, Mobile-Edge Computing – Introductory Technical White Paper, Issue 1, Sep. 2014.
[96] M. Abdelshkour, "IoT, from Cloud to Fog Computing", blogs@Cisco - Cisco Blogs, 2017.
[97] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-End Network Slicing for 5G Mobile Networks," in J. Information Processing, Vol. 25, No. 1, Jan. 2017.
[98] S. Sharma, R. Miller, and A. Francini, "A Cloud-Native Approach to 5G Network Slicing", IEEE Communications Magazine, Vol.55, No.8, Aug. 2017.
[99] 5GPPP, View on 5G Architecture, 5G PPP Architecture Working Group, EuCNC, Athens, Jul. 2016.
[100] A. Banchs, et al., "A Novel Radio Multiservice Adaptive Network Architecture for 5G Networks", IEEE VTC-Spring, Glasgow, May 2015.
[101] P. Rost, et al., "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks", IEEE Communications Magazine, Vol.55, No.5, pp. 72-79, May 2017.
[102] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini and T. Braun, "Network Slices toward 5G Communications: Slicing the LTE Network", IEEE Communication Magazine, Vol.55, No.8, pp. 146-154, Aug. 2017.
[103] K. Samdanis, X. Costa-Perez and V. Sciancalepore, "From Network Sharing to Multi-tenancy: The 5G network Slice Broker", IEEE Communications Magazine, vol. 54, no. 7, pp. 32-39, Jul. 2016.
[104] D. Bega, et al., "Optimising 5G infrastructure markets: The Business of Network Slicing", IEEE Infocom, Atlanta, May 2017.
[105] ITU, Draft Terms and Definitions for IMT-2020, FG-IMT2020, Dec. 2016.
[106] ITU, Draft Technical Report: Report on Application of Network Softwarization to IMT-2020, FG-IMT2020, Dec. 2016.
[107] V. Sciancalepore, et al., "Mobile Traffic Forecasting for Maximizing 5G Network Resource Utilization", IEEE Infocom, Atlanta, May 2017.
[108] G. Tseliou, K. Samdanis, F. Adelantado, X. Costa-Perez, and C. Verikoukis, "A Capacity Broker Architecture and Framework for Multi-tenant support in LTE-A Networks", IEEE ICC, Kuala Lumpur, May 2016.
[109] P. Marshall, H. Tufo and K. Keahey, "Provisioning Policies for Elastic Computing Environments", IEEE IPDPSW, Shanghai, May 2012.
[110] R. Tripathi, S. Vignesh and V. Tamarapalli, "Cost-aware Capacity Provisioning for Fault-tolerant Geo-distributed Data Centers", IEEE COMSNETS, Bangalore, Jan. 2016.
[111] A. Giorgetti, et al., "Failure-aware Idle Protection Capacity Reuse", IEEE Globecom, St. Louis, Nov. 2005.
[112] B. Javadi, P. Thulasiraman and R. Buyya, "Cloud Resource Provisioning to Extend the Capacity of Local Resources in the Presence of Failures", IEEE HPCC-ICESS, Liverpool, Jun. 2012.
[113] 3GPP TR 28.801, Study on Management and Orchestration of Network Slicing for Next Generation Network, Rel.15, May 2017.
[114] GEC 14 Slice Around the World Demo, groups.geni.net, 2017.
[115] G. Roberts, et al., "Network Services Framework v1.0", OGF NSI GFD.173, Dec. 2010.
[116] OpenStack API, api.openstack.org, 2017.
[117] J. Schoenwaelder, Common YANG Data Types, IETF RFC 6021, Oct. 2010.
[118] ITU-T Y.3011, Framework of Network Virtualization for Future Networks, Jan. 2012.
[119] C.J. Bernardos et al., "5GEx: Realising a Europe-wide Multi-domain Framework for Software-Defined Infrastructures", Transactions on Emerging Telecommunications Technologies, Vol. 27, No. 9, pp. 1271–1280, Sep. 2016.
[120] I. Afolabi et al, "Towards 5G Network Slicing over Multiple-Domains", IEICE Transaction on Network Virtualization, Network Softwarization and Fusion Platform of Computing and Networking, Vol. E100.B, No. 11, Nov. 2017, pp. 1992-2006.
[121] Y. Zaki et al., "LTE Wireless Virtualization and Spectrum Management", IEEE/IFIP WMNC, Budapest, Oct. 2010.

[122] R. Kokku, R. Mahindra, H. Zhang and S. Rangarajan, "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks", IEEE/ACM Transactions on Networking, Vol. 20, No. 5, pp. 1333-1346, Oct. 2012.
[123] R. Mahindra, M. Khojastepour, H. Zhang and S. Rangarajan, "Radio Access Network sharing in cellular networks", IEEE ICNP, Goettingen, Oct. 2013.
[124] T. Guo and R. Arnott, "Active LTE RAN Sharing with Partial Resource Reservation", in IEEE VTC Fall, Las Vegas, Sep. 2013.
[125] J. He, and W. Song, "AppRAN: Application-oriented Radio Access Network Sharing in Mobile Networks", IEEE ICC, London, Jun. 2015.
[126] A. Ksentini and N. Nikaein, "Towards Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction", IEEE Communications Magazine, Vol.55, No. 6, pp. 102-108, Jun. 2017.
[127] P. Rost et al. "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks", IEEE Communications Magazine, Vol. 55, No.5, pp. 72-79, May 2017.
[128] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective", IEEE Wireless Communications, Vol.24, No.5, pp. 166-174, Oct. 2017.
[129] E. Pateromichelakis, K. Samdanis, Q. Wei and P. Spapis, "Slice-tailored Joint Path Selection & Scheduling in mm-Wave Small Cell Dense Networks", IEEE Globecom, Singapour, Dec. 2017.
[130] P. Caballero Garces, X. Costa-Perez, K. Samdanis, A. Banchs, "RMSC: A Cell Slicing Controller for Virtualized Multi-tenant Mobile Networks", IEEE VTC-Spring, Glasgow, May 2015.
[131] J. Panchal, R. Yates, and M. Buddhikot, "Mobile Network Resource Sharing Options: Performance Comparisons", IEEE Transactions on Wireless Communications, Vol. 12, No. 9, pp. 4470-4482, Sep. 2013.
[132] A. Gudipati , et al., "SoftRAN: Software Defined Radio Access Network", ACM HotSDN, Hong Kong, Aug. 2013.
[133] V. Yazıcı, et al., "A New Control Plane for 5G Network Architecture with a Case Study on Unified Handoff, Mobility, and Routing Management", IEEE Communications Magazine, Vol. 52, No.11, pp. 76-85, Nov. 2014.
[134] X. Foukas et al., "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks", ACM CoNEXT, Irvine, Dec. 2016.
[135] R. Shrivastava, et al., "An SDN-based Framework for Elastic Resource Sharing in Integrated FDD/TDD LTE-A HetNets", IEEE CloudNet, Luxembourg, Oct. 2014.
[136] S. Costanzo, et al., "Service-oriented Resource Virtualization for Evolving TDD Networks Towards 5G", IEEE WCNC, Doha, Apr. 2016.
[137] E. Pateromichelakis and K. Samdanis, "Graph Coloring based Inter-Slice Resource Management for 5G Dynamic TDD RANs', IEEE ICC, Kansas City, May 2018.
[138] xran.org
[139] M. Bansal, et al., "OpenRadio: A Programmable Wireless Dataplane", ACM HotSDN, Helsinki, Aug. 2012.
[140] W. Wu, L.E. Li, A. Panda, and S. Shenker, "PRAN: Programmable Radio Access Networks", ACM Hot topics in Networks, Los Angeles, Oct. 2014.
[141] J. Wu, Z. Zhang, Y. Hong and Y. Wen, "Cloud radio access network (C-RAN): A Primer", IEEE Network, Vol. 29, No. 1, pp. 35-41, Jan. 2015.
[142] China Mobile Research Institute, C-RAN The Road Towards Green RAN White Paper, Version 2.5, Oct. 2011.
[143] A. Checko, et al., "Evaluating C-RAN Fronthaul Functional Splits in terms of Network Level Energy and Cost Savings", Journal of Communications and Networks, Vol. 18, No. 2, pp. 162-172, Apr. 2016.
[144] IEEE NGFI P1914, https://standards.ieee.org/develop/wg/NGFI.html
[145] 3GPP TSG RAN2 Tdoc R21700637, Summary of RAN3 Status on CUDU Split Option 2 and Option 3, and QuestionsIssues for RAN2, RAN3 NR Rapporteur (NTT DoCoMo, Inc.), Spokane, Jan. 2017.
[146] N.M.M.K.Chowdhury and R. Boutaba, "A Survey of Network Virtualization", Computer Networks, Vo.54, No.5, pp. 862-876, Apr. 2010.
[147] A. Malis, "Converged Service over MPLS", IEEE Communication Magazine, Vol,44, No.9, pp. 150-156, Sep. 2006.
[148] BBF TR-221, Technical Specifications for MPLS in Mobile Backhaul Networks, Oct. 2011.
[149] NGMN, LTE Backhauling Deployment Scenarios, White paper, Jul. 2011.
[150] N. Finn, P. Thubert, B. Varga and J. Farkas, Deterministic Networking Architecture, IETF Internet-Draft, Oct. 2017.
[151] BBF TR-221 Amendment 1, Technical Specifications for MPLS in Mobile Backhaul Networks, Nov. 2013.
[152] Common Public Radio Interface (CPRI); Interface Specification, CPRI Specification V7.0, Oct. 2015.
[153] M.Y. Arslan, K. Sundaresan, and S. Rangarajan, "Software-Defined Networking in Cellular Radio Access Networks: Potential and Challenges", IEEE Communication Magazine, Vol.53, No.1, pp. 150-156, Jan. 2015.
[154] Common Public Radio Interface: Requirements for the eCPRI Transport Network; Requirements Specification, eCPRI Transport Network D0.1, Aug. 2017.
[155] A. De la Oliva, et al., "XHaul: Towards an Integrated Fronthaul/Backhaul Architecture in 5G Networks", IEEE Wireless Communications, Vol.22, No.5, pp. 32-40, Oct. 2015.
[156] L. Xi, et al., "5G-Crosshaul Network Slicing Enabling Multi-Tenancy in Mobile Transport Networks", IEEE Communication Magazine, Vol.55, No.8, pp. 128-137, Aug. 2017.
[157] A.S. Thyagaturu, Y. Dashti, and M. Reisslein, "SDN-Based Smart Gateways (Sm-GWs) for Multi-Operator Small Cell Network Management", IEEE Transcations on Network and Service Management, Vol.13, No.4, pp. 740-753, Dec. 2016
[158] T. Taleb, et al., "EASE: EPC as a Service to Ease Mobile Core Network," IEEE Network Magazine, Vol.29, No.2, pp. 78–88, Mar. 2015.
[159] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, & Solutions," IEEE Wireless Communications Magazine, Vol.21, No.3, pp. 80-91 Jun. 2014.
[160] T. Taleb, A. Ksentini, and R. Jantti, "Anything as a Service for 5G Mobile Systems", in IEEE Network Magazine, Vol. 30, No. 6, Dec. 2016.
[161] 3GPP TR 23.711, Enhancement of Dedicated Core Networks Selection Mechanism, Rel.14, Sep. 2016.
[162] 3GPP TS 23.501, System Architecture for the 5G System, Rel. 15, Sep. 2017.
[163] H-J. Einsiedler, et al., "System Design for 5G Converged Networks", EUCNC, Paris, Jun. 2015.
[164] K. Mahmood, et al., "On the Integration of Verticals Through 5G Control Plane", EUCNC, Oulu, Jun. 2017.
[165] 3GPP TR 23.708, Architecture enhancement for Service Capability Exposure, Rel.13, Jun. 2015.
[166] 3GPP TS 23.502, Procedures for the 5G System, Rel. 15, Sep. 2017.
[167] GSMA, An Introduction to Network Slicing, 2017.
[168] B. Naudts, et al., "Techno-economic Analysis of Software Defined Networking as Architecture for the Virtualization of a Mobile Network", IEEE EWSDN, Darmstadt, Oct. 2012.
[169] C. Modi, D. Patel, B. Borisaniya, A. Patel, and M. Rajarajan, "A Survey on Security Issues and Solutions at different Layers of Cloud Computing", Journal of Supercomputing, Vol. 63, No. 2, pp. 561–592, Feb. 2013.
[170] F. Gens, "IT Cloud Services User Survey, Pt. 2: Top Benefits & Challenges", IDC eXchange, 2008.
[171] V. Choyi, A. Abdel-Hamid, Y. Shah, S. Ferdi and A. Brusilovsky, "Network Slice Selection, Assignment and Routing within 5G Networks", IEEE CSCN, Berlin, Oct. 2016.