

Real-time Dynamic Network Slicing for the 5G Radio Access Network

Massimiliano Maule*, Prodromos-Vasileios Mekikis*,
Kostas Ramantas*, John Vardakas*, and Christos Verikoukis†

**Iquadrat Informatica S.L., Barcelona, Spain*

†Telecommunications Technological Center of Catalonia (CTTC/CERCA), Castelldefels, Spain

Abstract—The 5G networks are expected to satisfy diverse use cases and business models with significant advancements in terms of capacity, reliability, and latency. The allocation and provisioning of network resources pose a challenge for this novel architecture to guarantee higher flexibility and quality of service. As a potential enabler, network slicing was proposed as an innovative approach for the control of the network resources. Although a static slicing approach can be suitable for the transport and core network, the stochastic behavior of the wireless channel requires fast and secure slicing techniques for resource allocation. In this paper, we propose a dynamic slicing approach for the radio access network, where the network resources are carefully assigned to guarantee the service level agreements and increase the number of served users. To prove the performance of our approach, we implemented a fronthaul testbed to emphasize the strength of our method in terms of throughput and resource utilization, compared to static slicing.

Index Terms—5G network, RAN slicing, Software-defined Network, Software-defined Radio, LTE, virtualization

I. INTRODUCTION

Expected to be deployed first in dense urban areas, 5G networks will cover over 20 percent of the world's population by the end of 2023 [1]. Mobile data traffic is expected to surge by eight times during the forecast period, reaching 110 exabytes per month by 2023 [1]. To achieve an ultra-high data transmission rate and an ultra-low response times (latency), 5G will be deployed through technologies, as Software Defined Networking (SDN) and Network Function Virtualization (NFV). SDN and NFV represent the key technologies for efficient 5G network management. With SDN, a network is capable to support the traffic management needs dictated by the new forms of distributed processing. Moreover, NFV introduces the concept of virtualization of Network Functions (NFs), translating hardware-based appliances into high volume servers, switches and storage.

A significant aspect of this new architecture concerns the provision of a wide range of services with different requirements. The control of this architecture is based on three classes of services (i.e., enhanced mobile broadband (eMBB), ultra-reliable and low-latency communication (URLLC) and massive machine-type communications (mMTC)) in which the flows are guided according to their requirements [2].

To perform these type of services on the new 5G architecture through SDN and NFV techniques, the concept of Network Slicing (NS) has been proposed. The term network slice refers to a network with specific resources and functions, which are

perceived by the user as if they were a dedicated physical network, isolated from other virtual systems [3]. A service-oriented slice approach represents the perfect technique for isolating specific functionalities, guaranteeing a certain type of resources and maintaining the continuity of the end-to-end service on top of the Physical Network (PN). To that end, there are two ways to perform slicing within a network. Static slicing consists in assigning a fixed and constant portion of physical resources to the virtual network for the duration of the service. Consequently, this approach can lead to a waste of resources, as the load in the network is variable and influenced by different factors (e.g., difference in use between day and night, flux variations by geographical area). On the other hand, dynamic slicing consists in assigning resources to a specific virtual network based on the type of service requested.

In the current state of the art, NS is mainly focused on a static approach since the current network architecture model does not fully support NFV and SDN techniques necessary for the correct establishment of the slices. In [4], a prototype of an end-to-end (E2E) NS testbed is proposed, where the importance of correct slice parameterization is discussed. Moreover, the work in [4] illustrates the limitations of static slicing and the need to introduce dynamic techniques.

Different models based on machine learning and artificial intelligence have been proposed for the derivation of the optimal dynamic slicing approach. In [5], the authors formulate the dynamic slicing problem as a Mixed Integer Linear Programming (MILP) problem. Also, in [6], authors focus on the Deep Q-Learning technique for slicing resource management while in [7] the authors propose a new dynamic NS scheme based on Baseband Unit (BBU) capacity allocation and Physical Resource Blocks (PRBs) management. According to our analysis, even though these approaches provide the optimal solutions for dynamic slice selection and configuration, they are far from a physical application in a real scenario, where the communication channel is subjected to sudden variations that can significantly compromise learning models based on the network evolution.

In this paper, we present a dynamic slicing approach whose objective is to minimize the waste of resources by guaranteeing the minimum requirements of each service requested by the users. The innovative part of our solution is the rapid slice parameterization changes according to the traffic requirements and the management of unpredictable variation due to the

aleatory behavior of the wireless channel. The obtained results illustrate how the correct resource allocation can maintain high Quality of Service (QoS) while increasing the number of users served.

Hence, our contribution is as follows:

- First, we provide an innovative dynamic model for slice management able to filter and analyze in real time the traffic parameters (packet received, packet transmitted, delay, buffer queue size, error rates) of each data flow, and setup the slice considering the minimum number of Resource Blocks (RBs) necessary to guarantee the perfect service supply.
- Second, we define a real scenario within our testbed, where two flows are appropriately parameterized to simulate a slice with multiple video traffic (streams with different video quality) and the other slice with mobile broadband traffic (messaging, email, updates).
- Third, we test our dynamic slicing model alongside a static slicing baseline model on top of our testbed over the same scenario, in order to compare their performance and extract useful insights.

The rest of this paper is organized as follows. In Section II, the system architecture and properties of NS approach are presented. Section III introduces our dynamic slicing algorithm through a flowchart representation and the testbed configuration. In Section IV, we evaluate the performance achieved, followed by the conclusion in Section V.

II. SYSTEM ARCHITECTURE

This section presents the standardized 3GPP architecture model for 5G systems [8]. Its main objective is the definition of an end-to-end platform based on NFV and SDN technologies for service implementation and management. In our work, the same architecture is employed with a precise focus to the fronthaul part that has been accurately replicated in our testbed in line with the 3GPP specifications. The highlights of this architecture can be defined as follows:

- A clear separation of Control Plane (CP) functions from User Plane (UP) functions.
- Scalability, flexibility and rapid establishment of NFs.
- Intelligent resource management that reduces the ability to reuse a network service for different tenants.
- A vision of the network as a single system, reducing the costs and management procedures for the interconnection of different NFs.
- A unique authentication system that allows a simple interaction between the network devices.
- Minimization of the private infrastructure concept. Access Network (AN) and Core Network (CN) belonging to different domains must be equipped with control and management systems capable of communicating with each other.
- Migration of NFs between AN and CN for service support and low latency and reduction of congested situations between them.

The 5G architecture is divided into three macro-sections [9]: i) the fronthaul, which is the network between the Remote Radio

Unit (RRU) and the Distributed Unit (DU), ii) the central unit between the DU and the CU, and iii) the network between the Centralized Unit (CU), the 5G CN, and other CUs. In the fronthaul part, the transformation from 4G LTE to 5G NR is mainly highlighted by the division of the BBU into three parts: CU, DU, and RRU. Although our system is based on the LTE standard, we have developed our testbed following the division of the BBU defined in 5G. This operation was possible thanks to the virtualization of every LTE entity and to the use of a Software Defined Radio (SDR).

Fig. 1 shows an example of end-to-end 5G non-roaming architecture. In this figure, each tenant belongs to a specific type of service (automotive for URLLC, mobile broadband for eMBB) with different network characteristics. Each node is interconnected with one or more network entities (server, datacenter, storage, router) equipped with different network functionalities, such as firewall, authentication service, proxy, virtual switch and virtual machines. The combination of NS with an SDN controller allows defining an optimal combination between the functions of each node and the service requirements. Moreover, at the fronthaul, additional nodes with computational resources are directly connected to the eNBs acting as Multi-access Edge Computing (MEC) nodes. This technology allows content caching close to the network edge, which can support the deployment of NS in the RAN part.

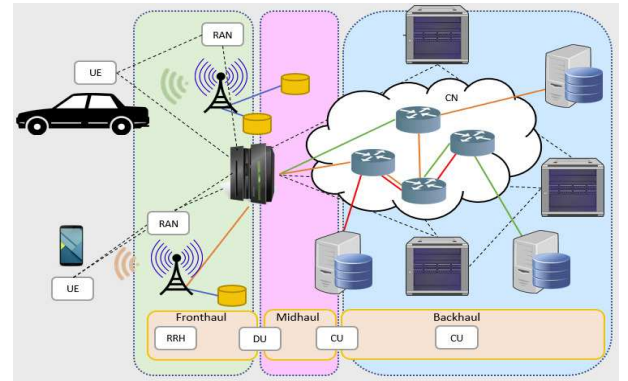


Fig. 1. 5G 3GPP End-to-End architecture

A. Network slicing properties

The concept of NS represents a key feature of our work and in general for 5G networks, to satisfy services with different needs in terms of latency, reliability, capacity and domain specific functionalities. NS is considered by the 3GPP standard organization towards defining a 5G system architecture. In order to be aligned with the latest 3GPP release on 5G NS, we developed our dynamic slicing solution following the 3GPP TR 28.801 slicing specification [10].

From this exhaustive investigation, the following requirements for NS have been defined [11]:

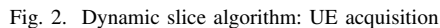
- Slice isolation: the transport network is by its nature used by a large number of users requesting different services. In order to guarantee the quality of a service, there are

- diversification: each slice has a large set of functions to use and share with other services
- deployment: an advanced resource management and migration allows a rapid assignment of functions belonging to different virtual networks to the same slice
- advanced management: the end user is offered a high level of abstraction for the network, allowing him to use functionalities and resources belonging to different operators.

III. REAL-TIME DYNAMIC SLICING APPROACH

initialization process, in which the operator registers the User Equipments (UEs) parameters (i.e., International Mobile Subscriber Identity, Public Land Mobile Network, Access Point Name, Operator key) in its Database (DB). The DB location is then communicated to the 5G-CN block responsible for the management of the tenant subscriber. Each block of the 5G-CN owns a specific functionality of the 5G control and data plane, which communicates among each other through specific network interfaces and protocols. Once they are active and synchronized, the 5G-RAN is initialized and connected to the 5G-CN. In our approach, at this point, the system communicates an initialization slice parameterization to the 5G-CN and 5G-RAN where the RBs are equally divided. The SDN controller, whose role is to monitor the status of the underlying network, is monitoring the RAN and CN parts, collects network statistics and posts new slice configurations when required. The gNodeBs (gNBs) are equipped with an SDN agent for the signaling between the SDN controller and the gNB CP. On top of the CP, the SDN controller and the SDN agent share the radio parameters such as frequency, number of RBs, transmission gain, etc., that are necessary for the configuration of the SDR.

When a UE sends a connection request to the gNB, the authentication procedure is activated. If its authentication parameters match the configuration of the 5G subscriber authentication function, the UE is accepted, otherwise, it is refused. As a first step, each accepted UE activates an initialization messaging phase with the gNB, where the main traffic requirements are communicated such as the traffic priority, average packet size, maximum packet delay, isolation restrictions and type of service (UE_req). This information is utilized by the 5G-RAN and 5G-CN to recognize the slice for the UE, and quantifies the resources needed (Slice_X_res) for the correct service supply. If the slice resources guarantee a proper service, the UE requirements are satisfied, otherwise, the system activates a reallocation resource procedure, as it will be explained in the following section.



This sub-section describes the 5G network architecture initialization and the acceptance procedure of the users, as it is illustrated in Fig. 2. In the beginning, there is an



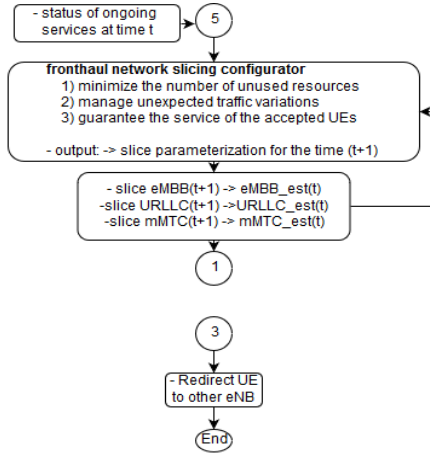


Fig. 4. Dynamic slice algorithm: system runtime optimization

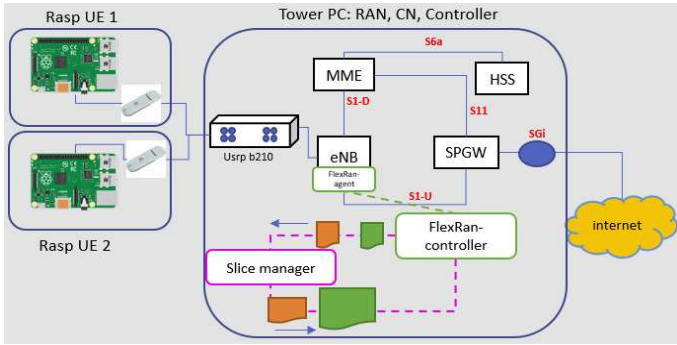


Fig. 5. Fronthaul Testbed Architecture

B. Statistic analysis and slice configuration

In order to reallocate the RBs among the slices, we designed a Slice Manager (SM) tool, which operating principle is illustrated Fig. 3. This reallocation procedure is activated by our SM when a slice does not fulfill the service requirements for the incoming UE. Moreover, the SM system communicates with the RAN and CN domains through the 5G CP specifications, ensuring a compliant with 3GPP standard solution. For each slice belonging to the gNB, the SM determines the amount of RBs (Slice_X). If at least one slice has available unused RBs (Slice_X_free_res), the SM monitors if the interslice functionality is enabled in the slicing parameterization, otherwise the UE is redirected to another gNB (point 3). When the interslice functionality is active, the SM extracts a certain amount of RBs from each slice according to a specific slice weight (Slice_X_w), until the sum of unused RBs from each slice is greater or equal to the minimum UE_req requirements. Another advanced functionality is the intraslice sharing. If active, the SM reconfigures the RBs and the scheduling procedure among the UEs served within the same slice.

The new parameterization obtained from the reallocation procedure is translated in a 5G CP-compliant file format, which is sent from the SM to the SDN controller. The SDN

controller communicates the new settings to the corresponding gNB SDN agent that applies the new system changes.

C. Background real-time dynamic system optimization

Our innovative method for the optimal parameterization of the slices is illustrated in Fig. 4. According to a specific granularity (i.e., one frame length), the output of this method is the optimal slice parameterization by taking into account: i) the ongoing services, ii) the unpredictable traffic variations, iii) the release of RBs from UEs that completed the service session, and iv) the changes of the service type (i.e. from eMBB to URLLC) during an ongoing transmission for the same UE. Moreover, this algorithm, which is implemented inside the SM, is executed in background mode in order to reduce the CPUs load. When the procedure is enabled (point 5 in Fig. 2), the SM adjusts the number of RBs until the estimated slice throughput is as close as possible to the measured real slice datarate. This technique guarantees the allocation, according to the system granularity, of the optimal amount of RBs to each slice (EMBB, URLLC, MMTC), without the isolation of unused RBs.

This method allows our system to be always equipped with the optimal configuration in line with the services evolution. As a consequence, the delay due to the RB reconfiguration when a new connection request arrives is reduced, and a homogeneous resource distribution is applied among the slices.

TABLE I
TESTBED PARAMETERS

Type	Value
OS	Ubuntu 16.04.6 LTS (Xenial Xerus) 64 bits
RAM	8 Gb
UE antenna	Huawei E3372
UE HW	Raspberry Pi
UE OS	Raspbian
eNB Radio	USRP B210 SDR - Dual Channel Transceiver (70MHz-6GHz) Ettus Research
Radio Splitter	LTE band7 Duplexer

IV. PERFORMANCE EVALUATION

In this section, we present the setup of our testbed and the obtained results from the conducted experiments. Due to hardware limitation, our dynamic slicing algorithm was tested on top of a LTE architecture, following as close as possible the virtualization principles and network implementation defined for 5G from 3GPP. To the best of our knowledge, our implementation does not present any limitation for a future migration to a 5G testbed.

A. Testbed description

In this section, we describe the main features of our testbed configuration and the motivation behind our system choice implementation. Since a substantial contribution for the management of resources through slicing techniques in the backhaul and midhaul networks already exists in the literature, we focused our attention on the fronthaul part, because RAN

slicing, in particular dynamic, has attracted its attention only in the last years. Fig. 5 shows the structure of our testbed used to perform dynamic slicing in the fronthaul network. Currently, there are several software solutions that implement the RAN and CN part of 4G and 5G systems. For our platform, we decided to use OpenAirInterface (OAI) [12]. The division of NFs into independent blocks further emphasized as NFV represents a dominant technology in future networks. In a real system, this allows the placement of each entity in a different node of the network, significantly reducing the installation and maintenance costs present in today's systems. To support multi-tenant systems, the synchronization of each entity is provided by standardized interface systems based on IPv4 and IPv6 protocols. The radio system used is Universal Peripheral Radio Software (USRP) B210, from Ettus Research [13]. This equipment provides a fully integrated USRP platform specifically designed for low-cost experimentation. Our test considers two UE based on Raspberry Pi platform with Linux operating system, each one equipped with a Hauwei Dongle antenna for the interface toward the eNB. In parallel to the CN, FlexRAN controller is used as an independent entity [14]. This module does not require synchronization with the CN as it manages slicing for the access network. FlexRAN represents a SDN controller belonging to the Mosaic5G project [15]. The project aims to transform radio access and CNs into an agile network service delivery platform to rapidly explore novel concepts as well as application and business needs. In particular, FlexRAN constitutes one tool of this platform, and represents a flexible and programmable platform for software-defined RAN [16]. Table I summarizes the hardware and software parameters of our testbed.

B. Simulation results and analysis

In this section, we evaluate the performance of our solution using our testbed in a scenario that includes two UEs. The Channel Quality Indicator (CQI) is considered 15 and the Modulation Coding System (MCS) to 28. The eNB is configured with a downlink and uplink band of 10 MHz, allowing a total of 50 RBs for each direction. To provide a realistic scenario, we defined a variable packet load (from 500 to 1200 bytes) and the variable type of traffic (UDP, ICMP), taking into consideration different types of interarrival packet distributions. The fundamental parameters of our implementation are illustrated in Table II.

Using the aforementioned implementation, we performed different types of experiments: i) first, we emphasize the importance of a correct configuration of the slices, and ii) then, we compare our algorithm against static slicing.

The first experiment is illustrated in Fig. 6. In this scenario, a user requests a service with constant traffic during the transmission. The RAN architecture is equipped with a single slice with 50 RBs DL and 50 RBs UL. In Fig. 6a, the maximum estimated throughput is compared with the real measured throughput, which presents a constant trend inline with the traffic interarrival distribution. Fig. 6b shows the

number of RBs assigned to the slice along with the packet delay metric.

TABLE II
EXPERIMENT PARAMETERS

Variable	Value
frame type	FDD
downlink frequency	2650 MHz
uplink frequency	2530 MHz
downlink max bandwidth	10 MHz
uplink max bandwidth	10 MHz
number of RB downlink	50
number of RB uplink	50
Number UE	2
packet size	500 - 1200 bytes
channel quality indicator	15
modulation coding scheme	28
Inter-departure packet distribution	Constant, Poisson
protocols	UDP, ICMP

With a configuration equal to 100 percent of the available resources (50 RBs), the service is abundantly guaranteed (around 94 Kbps) and greater than the minimum throughput required (8-9 Kbps). With this configuration, a significant number of RBs are assigned, but they are not fully used. To optimize the system, it is necessary to refine the slice RBs percentage so that the maximum estimated throughput is as close as possible to the real measured throughput.

The SM defines a "security threshold" variable in order to keep the estimated throughput always slightly higher than than the real throughput. This functionality is introduced to handle channel fluctuations and queues saturation during the service transmission. The importance of this parameter is highlighted in the red section of Fig. 6a: with a small security threshold, the estimated throughput is almost equal to the measured throughput, and the packet delay increases as there are not enough resources available to handle the queues. The configuration of resources between slices is particularly important in the case of scenarios where multiple slices coexist with time variant services.

Fig. 7 and 8 illustrate the behavior in the case of a scenario with two slices, one with constant low-bitrate traffic (blue line), and the other with a time-varying burst traffic (orange line). In order to reflect a realistic scenario, the low-bitrate slice emulates 5G broadband mobile traffic, while the burst slice emulates video streaming, where packet buffer is filled multiple times during the transmission. The dashed lines represent the ideal configurations to guarantee the services in the same scenario under the static slicing method. The zero flexibility of the static approach forces to configure each slice according to the maximum throughput requested by a service. This involves a high allocation of resources especially in the case of variable traffic in the network.

In Fig. 7, the application of the dynamic approach in the variable burst traffic slice guarantees the SLAs of the user's served, while reducing the use of RBs up to 23 percent. The granularity of the post configuration for a new slice parameterization depends from the bitrate of the served traffic. In this

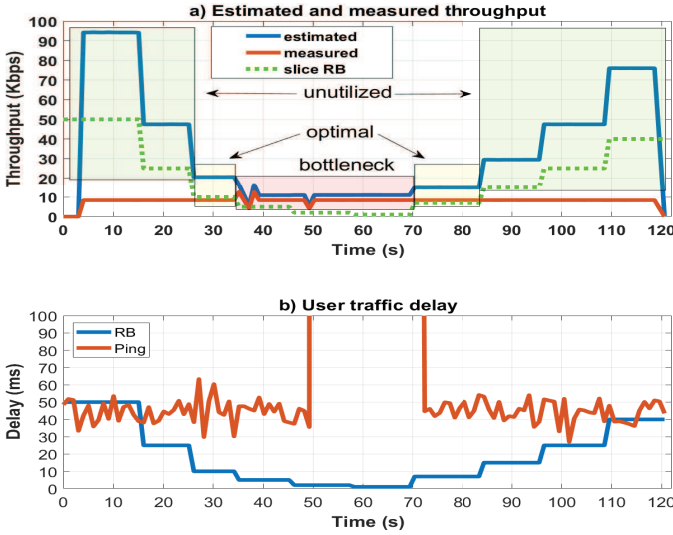


Fig. 6. Performance comparison: a) estimated throughput vs. real throughput and b) packet delay vs. number of resource blocks assigned

experiment, it was possible to assign a new parameterization every 10 milliseconds, like the LTE frame size. This level of granularity permits to optimally handle the traffic variation, avoiding congestion in the queues and reducing packet delay. In the case of constant low bitrate traffic (blue line), a dynamic approach does not bring greater benefits than the static method. This result shows a case where it is preferable to manage the traffic with the static system and to keep the dynamic approach for the management of variable and complex traffic.

To present the throughput performance of each dynamic slice, we present in Fig. 8 the throughput performance of each dynamic slice. The correct parameterization of the slices, explained in Fig. 7, is highlighted in this figure, where the measure throughput follows a trend proportional to the number of RBs assigned to each slice. Burst traffic is subject to multiple peaks (up to 9 Mbps) due to queue adjustments in the scheduler or retransmission of a large number of packets due to a channel nature issues. These unexpected variations do not damage the traffic of other users, as the resources are appropriately divided and independent among the slices.

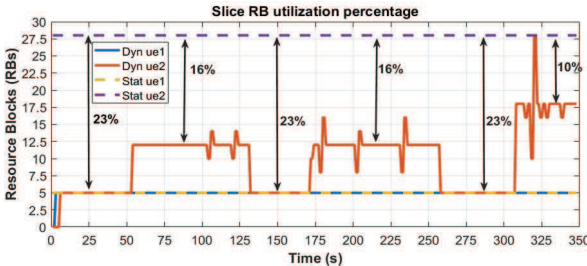


Fig. 7. Comparison static slicing versus dynamic slicing

V. CONCLUSION

In this paper, we presented a dynamic NS algorithm for the resource management of a 5G network, inline with the

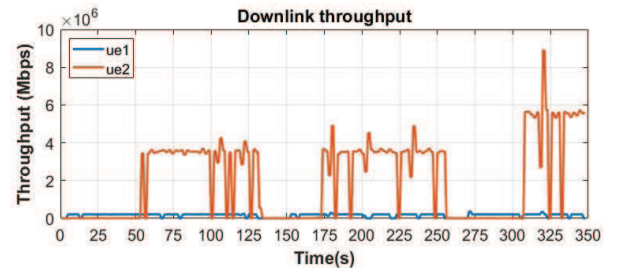


Fig. 8. Resource Blocks traffic-based assignment

3GPP architecture and slicing specifications. With the support of our testbed, we evaluated the algorithm on top of a real 5G scenario. Our solution illustrates a simple technique for network resources management optimization, enhancement of network QoS user performance, improved stability of the 5G CP, and an increment of the number of users served. Future developments for this platform will include an optimization of the hand-shaking phase between the user and the eNB for a more rapid convergence in the allocation of resources, and the introduction of a system of evaluation of the quality of the chosen parameterization.

REFERENCES

- [1] Ericsson. "Ericsson predicts 1 billion 5g subscriptions in 2023" White paper (2017).
- [2] 5G PPP Architecture Working Group. "View on 5G Architecture (Version 2.0)." 5GPPP White paper (2017).
- [3] NGMN Alliance, NGMN Network Slicing "Description of Network Slicing Concept", Available: <https://www.ngmn.org>
- [4] Gebremariam, Anteneh A., et al. "Towards E2E Slicing in 5G: A Spectrum Slicing Testbed and Its Extension to the Packet Core." 2017 IEEE Globecom Workshops (GC Wkshps). IEEE, 2017.
- [5] Raza, Muhammad Rehan, et al. "Dynamic slicing approach for multi-tenant 5G transport networks." Journal of Optical Communications and Networking 10.1 (2018): A77-A90.
- [6] Li, Rongpeng, et al. "Deep Reinforcement Learning for Resource Management in Network Slicing." IEEE Access 6 (2018): 74429-74441.
- [7] Lee, Ying Loong, et al. "Dynamic network slicing for multitenant heterogeneous cloud radio access networks." IEEE Transactions on Wireless Communications 17.4 (2018): 2146-2161.
- [8] ETSI. TS. "123 501 V15. 2.0 (2018-06) 5G." System Architecture for the 5G System (Release 15)
- [9] GSTR-TN5G, I. T. U. T. "Technical Report "Transport network support of IMT-2020/5G", Feb 2018."
- [10] 3GPP TR 28.801 V2.0.1."Study on management and orchestration of network slicing for next generation network", TSG-SA, (2017-09)
- [11] 3GPP, "Service requirements for next generation new services and markets", 3GPP TS 22.261 1.1.0, February 2017
- [12] Nikaein, N., et al. "OpenAirInterface: A flexible platform for 5G research." ACM SIGCOMM Computer Commun. Review 44.5 (2014): 33-38.
- [13] <https://www.ettus.com/all-products/UB200-KIT/>
- [14] Foukas, Xenofon, et al. "FlexRAN: A flexible and programmable platform for software-defined radio access networks." Proceedings of the 12th International Conference on emerging Networking EXperiments and Technologies. ACM, 2016.
- [15] Nikaein, N., et al. "Mosaic5G: Agile and flexible service platforms for 5G research." ACM SIGCOMM Computer Communication Review 48.3 (2018): 29-34
- [16] Ramantas, Kostas, et al. "Implementation of an SDN-Enabled 5G Experimental Platform for Core and Radio Access Network Support." Interactive Mobile Communication, Technologies and Learning. Springer, Cham, 2017.