# 5G RAN slicing: dynamic single tenant radio resource orchestration for eMBB traffic within a multi-slice scenario

Massimiliano Maule*, John Vardakas*, and Christos Verikoukis †
**Iquadrat Informatica S.L., Barcelona, Spain*
*†University of Patras, ATHENA/ISI, and Iquadrat Informatica S.L., Barcelona, Spain*

***Abstract*—Emerging 5G systems will need to seamlessly guarantee novel types of services in a multi-domain ecosystem. New methodologies of network and infrastructure sharing facilitate the cooperation among the operators, exploiting the core and access sections of the system architecture. Network Slicing (NS) is the operators' best technique on how to build and manage a network. Without NS, the 5G requirements in terms of flexibility, optimal resource utilization, and investment returns cannot materialize. Before slicing is commercially available, different sections of the 5G architecture should be modified to include NS. In this work, we present a novel dynamic Radio Access Network (RAN) slicing resource sharing method aimed to guarantee the optimal Service Level Agreements (SLAs) through the monitoring of each slice tenant Key Performance Indicators (KPIs). The experiments are conducted following the 3GPP specifications, and the solution is validated using a testbed based on the main 5G functionalities.**

***Index Terms*—5G network, 5G-NR, RAN slicing, Software-defined network, Virtualization**

## I. Introduction

The evolution of Radio Access Networks (RANs) in the last years is now at 5G. From the previous 4G technology, the Third-Generation Partnership Program (3GPP) has added multiple functionalities to the RAN and Core Network (CN) to support new network elements and applications. Even though the Long-Term Evolution (LTE) technology has introduced significant RAN improvements, the exponential growth of connected devices and innovative services during the last decade pushes Mobile Network Operators (MNOs) to investigate new methods for improving the coverage, increase their systems capacity and reduce the services latency in mobile and fixed networks. As a consequence, in the next years, the main mobile infrastructure challenges for the MNO will be to [1]: i) Enabling 80-100 MHz of spectrum for the operator at 3.5 GHz, and around 1 GHz per operator at 26-28 GHz [2], ii) Definition of new market and business models through the deployment of new infrastructures and software solutions, iii) Massive network expansion and upgrade of private solutions, iv) Intensive infrastructure deployments, advanced indoor and outdoor distributed antenna system, and v) Advanced Hardware and Software (HW/SW) solutions for new services, with a smart energy consumption and accurate system management.

The aforementioned novel capabilities are an opportunity for operators to move further the simple connectivity concept and explore new architectural solutions such the coexistence with 4G networks and other access/core technologies to provide an unlimited, fast, reliable and enhanced broadband experience to the society, reaching data rates up to 1 Gigabit per second and latency less than 4 milliseconds. These novel system performance permit the operators to increase their portfolio with cloud and Artificial Intelligence (AI) solutions, massive Internet of Things (IoT) deployments, and advanced industrial applications for new vertical markets.

Earmarked as a prominent feature of 5G for enabling the aforementioned technological capabilities, the concept of NS has been introduced. This key technology enabler permits to multiple vertical industries the execution of their solutions on top of a shared infrastructure, where the Service Provider (SP) customize the network capabilities (security, connection, processing power, storage, etc.) [3]. From a SP point of view, this solution represents a new type of business model, known as NS as a Service (NSaaS), which will be the future network trend of the next years. MNOs, using an unique network slice type, are able to arrange the needs of multiple customers, as well as different services belonging to multiple network slices may be gathered together and supplied as a single slice to a customer with diverse requirements. Typical example of vertical market using this solution is autonomous driving, where telemetry and infotainment services are simultaneously provided in a single package.

Due to the stochastic behavior of networks, static allocation of network resources does not represent an efficient approach. In particular, when applied to wireless resources, due to burst of traffic, user mobility, and time-varying channel, only through resource overprovisioning techniques we are able to deal with multiple services, without exploiting the flexibility of short-medium term fluctuations in terms of resource requirements [4]. For this reason, dynamic NS represent the leading approach for the future networks. With this technique, network resources are dynamically assigned to meet tailored performance requirements (e.g. capacity, latency, priority, security) through a seamless and virtually continuous network propagated across the 5G networks architecture [5]. This methodology will help operators to optimize resource usage, cost reduction and accelerate time to market for innovative new service offerings.

To this end, this work illustrates a novel real-time end-to-end NS framework where multiple network slice instances are dynamically modelled according to the MNO's SLA and users' Quality of Service (QoS) to meet the optimal radio resource sharing and service performance. The concept of network slicing has initially been proposed for the 5G CN, while only in the last few years 3GPP study items started to investigate the impact of slicing in the RAN part of the network.

To the best of our knowledge, our solution represents the first application of dynamic NS solution through a joint evaluation of the slice tenant owner SLA and the real-time service performance from the user perspective. The main contributions of our solutions includes:

- Creation of a network traffic generator tool for testing different services inline with the testbed capabilities. The traffic requirements are proportionally scaled according to the performance of our real testbed.
- Highly customizable configuration of the SP traffic requirements, slices isolation policies, processing resources for new slice instances, and logging.
- Real-time analysis of the slice performance and resources configuration through a joint evaluation of the user traffic trend and slice SLA.
- Implementation of the dynamic NS solution using platform independent interfaces for advanced scalability with third party software and hardware.

The remainder of this article is organized as follows: Section II gives an overview of the concept of NS and its properties. Section III illustrates the architecture and the optimization algorithm of our solution. Section IV illustrates the experimental features and results, while concluding remarks are given in Section V.

## II. NS features, Techniques, and Development

This section illustrates some basic considerations necessary for the SPs when a new slice-based system is instantiated, and a description of two NS implementation methodologies will provide a better understanding of our approach. To motivate the proposed system architecture of this work, a brief discussion behind the RAN advancements is performed.

### *A. Requirements*

This subsection illustrates the principal requirements and considerations adopted by a SP when a new slice instance should be instantiated. In our solution, similar analysis is conducted for defining the initial configuration of each deployed slice.

For each tenant, the SP configures a network slice instance following some baseline principles [6]:

- The network section: RAN, Transport Network (TN), and CN. Each section presents different requirements and performance which should be carefully analyzed before instantiating a service.
- The SLAs: latency, Guaranteed Bit Rate (GBR), Non-Guaranteed Bit Tate (N-GBR), availability, and packet loss are some parameters to evaluate when the slice should be defined.
- The type of vertical market application: industry 4.0, Vehicle-to-everything (V2X), smart cities, Internet-of-Things (IoT), etc.
- The deployed access technology: cellular and fixed networks (i.e., LTE, WiFi, optical access solutions) present different technological performance and limitations. Since a slice could potentially exploit all the architecture, the SP should consider how each access technology affects the global service performance.
- The type of service provided: enhanced Mobile BroadBand (eMBB), Ultra-Reliable and Low Latency Communication (uRLLC), machine Massive Type Communications (mMTC). According to the service requirements, the traffic flows belong to different types of services, facilitating the management and service supervision.
- Cross domain services: services across multi-provider domains such as L2, L3, and VPN services.

The customization of multiple slice granularities introduce many challenges, especially in terms of network management and orchestration. For this reason, the SP must have a complete vision of the network capabilities necessary to perform optimal network management and orchestration of the resources.

### *B. Multi-provider and Multi-domains NSaaS*

Since the main principles of 5G are scalability and flexibility, multiple SPs may share the same physical infrastructure, and activating different services on top of multi-domain ecosystem. Moreover, with the growth of Software Defined Network (SDN) and Network Function Virtualization (NFV) technologies, the sharing of network resources became a common practice among SPs [6].

While in a single domain scenario, a SP is aware of the topology and available network resources, in a multi domains scenario does not exist any management tool for sharing the topologies and resources informations across the SPs. Therefore, 5G introduces the exchange of information across these providers through a series of specific interfaces, as standardized in [6]. Using IF1 interface, the tenant sends request for the deployment of a service or a slice. To manage operations on top of a multi SP system, IF2 interface is specific for the communication among the orchestrators. Finally, IF3 interface facilitates the management of multi-domain networks through the separation of the technological solution from the vendor specific infrastructure.

This solution facilitates the creation of intelligent 5G network management systems across different domains and connected optimization functions. This principle improves network management operations for multiple markets as well as the vision of an unique transparent infrastructure to the final service tenant.

### *C. Static vs Dynamic slicing*

The major element underlying NS is the mechanism for resource allocation amongst slices. During the first trials on NS, 3GPP suggested that the base station resources are accurately divided based on predefined network policies [7]. Multiple providers share the same infrastructure, and the resources are allocated according to the QoS requirements.

Resource overprovisioning is utilized against SLA violation, introducing as side effect a reduction of the system performance due to the possible allocation of unusable resources. Moreover, the stochastic behavior of the network medium introduces complexity when it comes to allocate the resources of a new slice instance.

With radio dynamic NS, multiple tenants adjust their network capacities during different time periods, according to service and system variations. Machine learning and/or traffic forecasting techniques can be deployed to assist the slice provider to handle unexpected network situations and traffic pattern fluctuation which would involve SLAs violation. Concurrently, Reinforcement Learning (RL) solution such as Q-Learning can efficiently determine the optimal slice admission policy that maximizes the MNO's revenue [8]. Even though the Q-Learning method is capable to be executed in an online learning fashion with a much more reasonable computation cost, the training and decision phases may not be processed in time considering the complexity of 5G frame structure patterns (ETSI TS 138-211 V15.3.0), implying congestion and SLA's violation in the RAN part and modest service quality. To overcome this issue, our solution introduces an extra resource tolerance pool to each slice able to amortise unexpected traffic peaks, as it will be explained in Section III-B.

### D. RAN slicing evolution

The RAN has seen an exponential evolution over the last years and a considerable effort will be needed as we enter a new phase of the mobile industry.

The future RAN network will be more intricate and diffused, reinforcing well-known solutions such as network hypervisors, Virtual Machines (VMs), containers, while simultaneously exploiting novel technologies such as SDN, NFV, virtualization (vRAN), and Cloud (C-RAN) [9]. Network hypervisors are the network elements that abstract the physical infrastructure into logically isolated virtual network slices. VM enables the virtualization of a physical resource where each client can execute its own Operating System (OS), and resources such as computing, storage, memory, and network are shared among VMs. From the combination of the aforementioned tools, containers are light-weight alternatives to hypervisor-based VMs. A physical server in containers is virtualized such that standalone applications and services can be instantiated on a isolated servers. vRAN applies the features of NFV by virtualizing Network Functions (NFs), providing higher degree of flexibility for the RAN section. A virtual RAN consists of a centralized pool of baseband units (BBUs), virtualized RAN control functions and service delivery optimization platforms [10]. Furthermore, vRAN permits to a shared CN interface to assist multiple 5G New Radios (5G-NRs), facilitating the deployment of 5G street macro and small cells in areas characterized by different density. C-RAN represents the network architecture which can be used to activate virtualized networks. It requires a high capacity and low latency access network to manage fronthaul traffic.

The combination of vRAN and C-RAN enables the division of the Control Plane (CP) and Data Plane (DP), reducing the decoupling complexity of the NFs from private hardware, and increasing the level of versatility of MNOs networks needed for the commercialization of 5G. These elastic and scalable access and connectivity related functionalities are provided as-a-service to customers in a given geography area using 3GPP standard-based technologies.

## III. System Design and Management

### A. System architecture

The proposed dynamic RAN slicing solution architecture is illustrated in Fig.1. Our solution consists of three macro layers (fetch, management, execution), nestled together following a bottom-up workflow. It is important to remark that our proposed solution is backward compatible with existing 3GPP 5G stack, since it utilizes standardized protocols and functionalities to communicate with the main network architecture elements. Moreover, even though this work is focused on RAN slicing, using a suitable set of input data and output configuration files settings, it is possible to extend our approach to other network sectors (midhaul, backhaul), without downgrade the optimization capabilities of our model.

*1) Fetch phase:* In this subsection, all the information from the data acquisition blocks are collected, sorted, and skimmed to define a network data model utilized as input for the upper layers. The *NIC traffic* block filters the incoming traffic according to the MNOs slices requirements, using variable window granularities. Different type of filters can be customized based on the operator's policies: packet size, colored traffic, IP subnet, etc.

The *CP acquisition* block monitors the behavior of the served users to evaluate if the SLAs and custom user traffic requirements are guaranteed. As the current served users have the priority, this block can dynamically modify the incoming services acquisition rules. Through this methodology, new incoming users are rejected in the case the slice resources are saturated, or no resource sharing policy are available. Moreover, if the slice tenant decides to modify the SLA requirements after the service is instantiated, this block is responsible to reconfigure the acquisition criteria for the new and served users, while maintaining a seamless service.

To conclude this subsection, the *Channel Acquisition block* monitors the access and fronthaul parts to identify physical variations of the medium, which would affect not only the service acquisition rate, but also the configuration of the slices in the upper layers. For example, if a channel degradation is identified, the amount of resources for users must be redefined, together with the slices acceptance rates.

*2) Management phase:* This part illustrates the leading block of the presented approach. As central block, the *Manager* synchronizes and coordinates the tasks of each other blocks. Its implementation can be centralized or distributed, as most of the system components are virtualized. The correct placement of the *Manager* improves the system tolerance against failures, and minimize traffic overloading in sensitive network nodes. For this reason, in this work we assume the proposed framework installed in the edge part of the network, which represents a strategic point for the management of

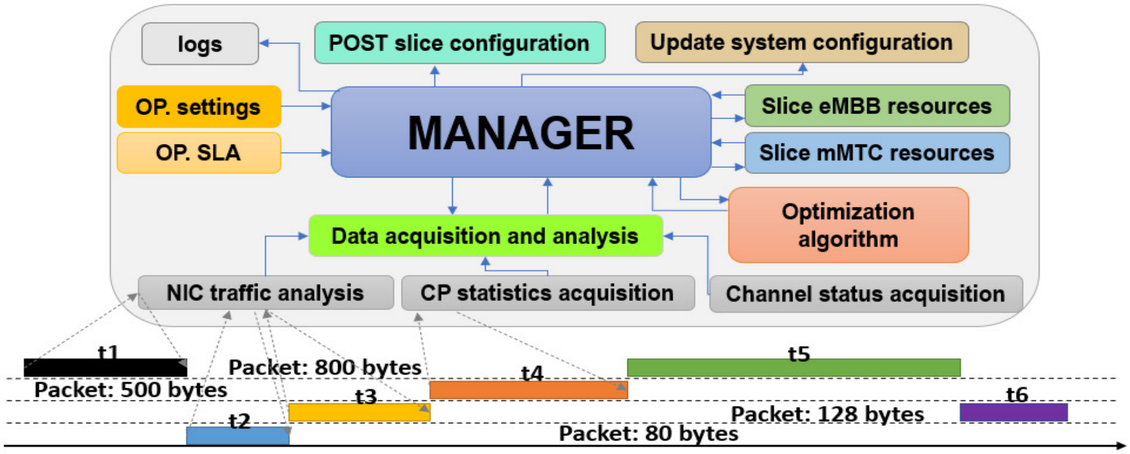


Fig. 1. Framework structure of our solution for dynamic RAN slicing

multiple RAN aspects. As minor task of the *Manager*, it is responsible to instantiate or delete the slices, according to the slice owner decision policies and/or the system performance.

As first step, the *Manager* receives in input the real-time data scenario model from the fetch layer, the SP specific settings, and per slice operator SLA. These information are encapsulated following a specific pattern, and forwarded to the *Optimization Algorithm* block, which return back the optimal slices parameterization for the next system processing window. For each slice of the MNO is defined a container with the amount of available resources, and predefined scheduling policies. In Fig.1, only two slice containers are represented, one eMBB and one mMTC, in order to keep our explanation aligned with the subsequent testing part.

For the identification of RAN resources changes, the *Manager* estimates a slice parameterization every time new data are received from the fetch phase. If the novel slice configuration differs from the current one in terms of required resources per slice, the *Manager* reconfigures the amount of resources assigned to each slice according to the service's need. From a practical perspective, this procedure corresponds to shift portion of Resource Blocks (RBs) among the slices, while preserving the SLA and QoS of each user. Following a tunable granularity, this operation can be executed dynamically, in accordance with the Transmission Time Interval (TTI) of our system. The new slices parameterization is structured following the JavaScript Object Notation (JSON) syntax, and forwarded to the top tier of the architecture.

*3) Execution phase:* The top system layer implements a set of RESTful-based Application Programming Interfaces (APIs) for the exchange of control information between the system and third-party radio software. Once a new slice configuration is posted to the RAN, the *POST* block sends back to the *Manager* an Acknowledgement (ACK) with the outcome of the operation.

As final operation, the *Manager* calls the *Update System Statistics* block, which is responsible to update the system variables required for the optimal processing of the forthcoming slices configurations. Optionally, a log file can be provided to trace potential issues during the entire workflow.

### B. Optimization algorithm

Inside the *Optimization Algorithm* block, the system partitions the physical resources of each slice into three modalities, each one with a specific role, as illustrated in Fig. 2:

- *Support*: the slice accepts incoming service requests, and part of its resources can be shared with other slices.
- *Conservative*: the slice prioritizes its own traffic by disabling sharing of resources with others.
- *Critical*: the amount of allocated resources does not guarantee complete SLA. The system evaluates if exists one or more slices in support mode able to share part of their resources to the critical slice.

The set of thresholds for each slice is defined considering the type of traffic, isolation policies, custom configurations, slice SLA, etc. Given specific KPI for each slice, the goal of the optimization block is to guarantee the optimal performance and SLA through the real-time sharing and balancing of the physical resources in the RAN part. For each slice, the tenant defines the SLA in terms of maximum guaranteed Bit Error Rate (BER), minimum and average guaranteed data rate, maximum tolerable latency, and maximum percentage of rejected requests. For the estimation of radio resources per user, a joint reverse-engineering approach with the slice SLA is performed by the *Manager* to estimate the amount of physical resources needed for a given user data rate. Once the user is accepted, its assigned resources are further refined according to the system evolution.

When the analysis of the real-time user traffic and slice SLA requires an amount of resources that exceed the second threshold, the slice shifts to critical mode, and the algorithm activates the resource sharing procedure.

The definition of a critical mode represents the innovation behind our dynamic management of the radio resources. Since the wireless channel is characterized by a stochastic behavior, the definition of a pool of extra resources handles unexpected traffic peaks, and the incoming users can anyway be accepted and served, while the algorithm performs resource balancing among the competing slices. This principle would be highly complex and not accurate using a machine learning solution, since the training sets are difficult to model for a real-time unpredictable RAN traffic, and the elaboration of a optimal slice resource configuration may not be computed in time according to the frame and subframe structure.

Through the illustrated slice modes division, the *Manager* is always aware of the traffic served by each slice, and the amount of available resources, since through the comparison of the occupied radio resources from the served users with the slice maximum capacity and threshold values, it is possible to always estimate the current mode for each slice. This principle facilitates the identification of the slice which will most probably require additional amount of resources, without impacting the SLA.

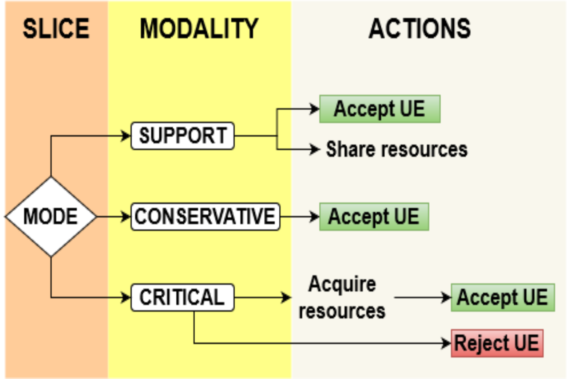


Fig. 2. Slice mode division

## IV. Single Tenant, Two Slices: a Case Study

### A. System scenario

In order to assess the performance of our solution, a HW/SW experimental platform is deployed, as illustrated in Fig.3. We focus our experiments in downlink traffic and for the proof of concept we use two types of slices, eMBB and mMTC, as part of a single tenant scenario. Since we want to evaluate our solution under heavy traffic condition, significant eMBB traffic is injected [11], while the primary role of the mMTC slice is to assist the supply of eMBB radio resources along the simulation. It is important to highlight that our system applies inter-slice resource prioritization only if negotiated during the SLA definition process. Otherwise, *First Come First Served (FCFS)* policy is applied. Due to the current limitation in open source 5G Standalone (SA) platforms, in this work, the performance of the dynamic NS solution are evaluated on top of a LTE-based testbed, equipped with 5G virtualization functionalities. Following, the main HW/SW tools of our testbed are described, while an exhaustive illustration of all the system elements was presented in our previous work [12]. Our platform is based on OpenAirInterface (OAI) [13], a flexible solution for 5G research implementing the 3GPP Cellular stack on general-purpose processor architectures. The OAI radio section presents a series of interfaces for the interconnection of different 3rd party RF modules. Our scenario utilizes the USRP B210 Software Defined Radio (SDR), provided by Ettus Research. This SDR guarantees at 2.5 GHz a bandwidth of 10 MHz in downlink, which corresponds to 50 RBs in the RAN. From the client side, a Raspberry Pi 4 Model B is equipped with a 4G/LTE HAT board, provided by Sixfab GmBH.

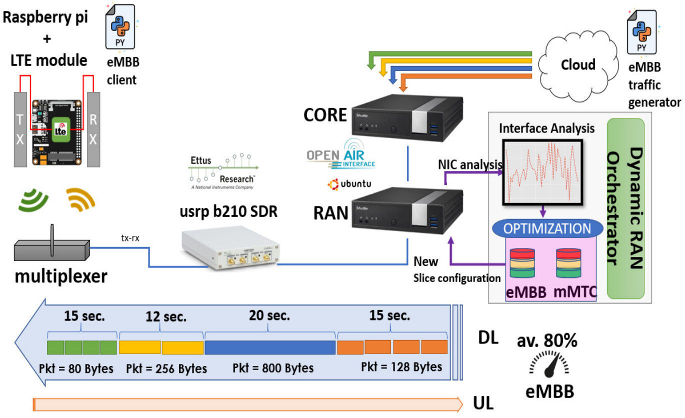


Fig. 3. testbed scenario and system architecture

As shown in Fig.3, we separated in two different machines the RAN and CORE parts. This implementation choice balances the processing workload of the system, allows flexibility in terms of centralized or distributed management, and the deployment of multiple split options as standardized for 5G [14].

Table I summarizes the main parameters of the tested scenario. The initial slice division represents the percentage of RBs assigned to each slice given the total system capacity, while the thresholds refer to the maximum capacity of each slice mode.

TABLE I
SYSTEM AND SIMULATION PARAMETERS

<table>
<tr><td colspan="3">Total duration</td><td colspan="3">150 sec</td></tr>
<tr><td colspan="3">Nr. slices</td><td colspan="3">2</td></tr>
<tr><td colspan="3">Bandwidth</td><td colspan="3">10 MHz (50 RBs)</td></tr>
<tr><td colspan="3">MCS DL</td><td colspan="3">28</td></tr>
<tr><td colspan="3">MCS UL</td><td colspan="3">8-20</td></tr>
<tr><td colspan="3">Granularity</td><td colspan="3">3 sec.</td></tr>
<tr><td colspan="3">Protocol</td><td colspan="3">UDP</td></tr>
<tr><td colspan="3">Direction</td><td colspan="3">Downlink</td></tr>
<tr><td>Nr. input flows</td><td>Duration (sec)</td><td>Packet size 80</td><td>Packet size 128</td><td>Packet size 500</td><td>Packet size 800</td></tr>
<tr><td>10</td><td>5</td><td>1</td><td>2</td><td>1</td><td>1</td></tr>
<tr><td>12</td><td>4</td><td>1</td><td>1</td><td>0</td><td>2</td></tr>
<tr><td>15</td><td>5</td><td>2</td><td>0</td><td>0</td><td>3</td></tr>
<tr><td>20</td><td>6</td><td>2</td><td>1</td><td>2</td><td>1</td></tr>
<tr><td colspan="3">Support th. eMBB</td><td colspan="3">20%</td></tr>
<tr><td colspan="3">Conservative th. eMBB</td><td colspan="3">50%</td></tr>
<tr><td colspan="3">Support th. mMTC</td><td colspan="3">20%</td></tr>
<tr><td colspan="3">Conservative th. mMTC</td><td colspan="3">50%</td></tr>
<tr><td colspan="3">Initial slice division DL</td><td colspan="3">50%</td></tr>
<tr><td colspan="3">Initial slice division UL</td><td colspan="3">50%</td></tr>
</table>

### B. Performance analysis

Using the aforementioned scenario as baseline of our experiments, we evaluate the efficiency of the proposed dynamic NS solution when a high load eMBB traffic is injected, and as consequence the SM selects radio resources from the mMTC slice to balance the system until the SLAs are not fulfilled.

This condition is confirmed through the evaluation of the final slice eMBB mode occupancy percentage (8% support, 24% conservative, and 68% critical), where the high presence of critical calls force the SM to perform resource migration among the slices during the entire simulation time.

Fig.4 illustrates the behavior of the eMBB slice during the entire simulation time. To correctly interpret the achieved results, the reader should take into account that the maximum downlink capacity using OAI-based testbed is around 30 Mbps.

The blue line indicates the injected eMBB high load traffic (average 20.62 Mbps, standard dev. 3.83, variance 14.74), while the red line represents the slice eMBB capacity growth trend until the SLA are not reached. For every variation of the traffic flow, the SM evaluates if a new slice configuration must be applied. This procedure is displayed with the bar plot, where each bar represents the amount of RBs required by the incoming service to reach the optimal slice eMBB SLA. When the bar sequence has a growing trend, the SM increases the amount of RBs for the slice eMBB in the next slice configuration. Conversely, a decreasing bar sequence trend indicates that the current slice configuration correctly match the served users traffic, and the SM can also decide for a partial release of RBs.

As stress downlink test for the slice eMBB, a series of flows are generated from the cloud network towards the users, with an average traffic load equals to 80% the total system capacity. As expected, the consecutive resource optimization calls of the SM brings a continuous increment of the eMBB slice capacity (from 25 RBs of the initial setup, until a final slice capacity of 44 RBs), with an occupation of 88% of the total system capacity before the end of the simulation.

With our solution, using a SM decision granularity of 3 seconds, after approximately 60 seconds the system reaches a steady resource balancing, optimally configured for the type of injected traffic. Moreover, as the sharing policies take into account the performance of the complete set of slices within the system, the SLAs are guaranteed for both the slices during the entire process, until a final stable slice configuration.

With an average amount of received packets of 20.50 Mbps, this experiment presents a Packet Error Ratio (PER) equals to 0.005. The dynamic approach of our solution, as expected, slighly affect the PER, which is nevertheless acceptable if compared with the standardized eMBB requirements [15].

For every iteration, the impact on the slice eMBB capacity proportionally affects all the modes, as illustrated in Fig. 5. This proportional scaling of the RBs of each mode reduces the ping-pong effect among the slices, which appears when a highly variable traffic is injected, and a continue scale up and down of the resources impacts the PER and management complexity. Without the division of each slice in different modes, the system would continue moving radio resources among the slices with the objective to balance the maximum capacity of each slice, defining a new slice configuration even for irrelevant traffic variations.

The results of Fig. 5 brings to light an intrinsic principle regarding the initial parameterization of the *Support* and *Conservative* eMBB slice modes. Allocating a tiny amount of RBs in *Support* (20% of the initial capacity), pushes the system in *Conservative* mode even when a small amount of traffic is injected in the system. This condition protects the eMBB RBs, since the RBs sharing functionality is disabled, as explained in Section III-B. Moreover, even a reduced *Conservative* threshold benefits the eMBB slice, since the slice is more inclines to enter the *Critical* mode, requesting other slices to share part of their RBs. This observation highlights how the initial slice modes setup should be carefully investigated taking into account the type of slice traffic, the slice isolation policies, and specific tenant's requirements.

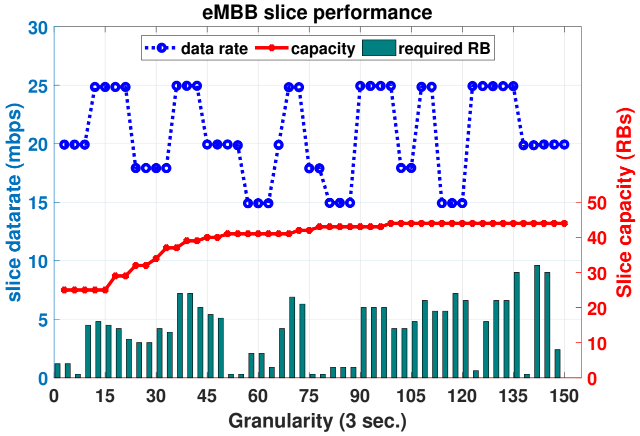


Fig. 4. eMBB data rate and slice capacity variation

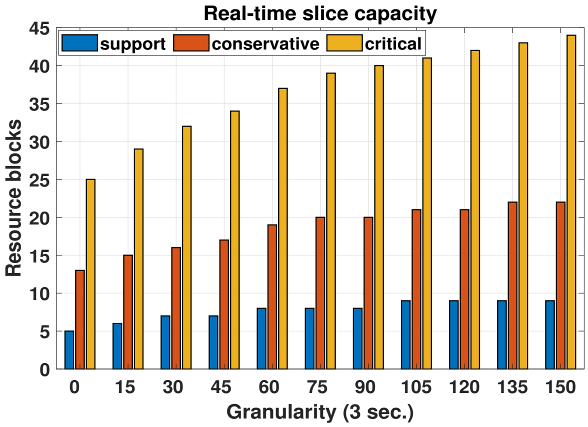


Fig. 5. eMBB slice capacity mutation

To conclude this section, an average jitter of 0.206 ms confirms the optimal configuration of the tested scenario, ensuring high order modulation scheme during the entire simulation. The high customization degree of the SM parameters permits to tackle multiple scenarios, and simultaneously optimize different type of slices.

A trade-off between slice resource assignation accuracy and a new slice configuration processing time should be carefully inspected. Depending on the type of traffic, this feature may have an impact on the slice throughput, latency and responsiveness to real-time traffic pattern changes.

## V. Conclusions

In this paper, we presented a complete solution for dynamic RAN slicing resource allocation where the optimal slice configuration is computed through a joint evaluation of the slice SLAs and the real-time evolution of the served users' traffic. Even though the experimental section is conducted using a single tenant scenario with two slices (eMBB and mMTC), the proposed method can be extend to multi-tenant and multi-slice environments.

The obtained results show how our solution is able to autonomously remap the radio resources in few seconds, while keeping a PER of 0.005 under heavy traffic scenario. We have proved that a suitable configuration of the slice policies and system parameters guarantees optimal performance for different type of traffic, matching the scalability and flexibility properties of the 5G networks.

As future work, we will define an exhaustive formulation of the optimization problem where the probability of accepting/rejecting an incoming user connection request is done combining a state-independent model of the multi-slice scenario with the analysis of the system capabilities in terms of real-time resources availability and SLAs. The stochastic model outcomes will be used by the *Manager* to further improve the resource management and accelerate the decision process, while simultaneously reducing the risk of resource saturation of the system. Moreover, the testing of our solutions on top of a SA E2E open-source 5G testbed is also one of our core priorities.

## VI. Acknowledgement

This work is supported by the research projects 5GROUTES (GA 951867), MSCA ITN 5G STEPFWD (GA 722429), SPOT5G (TEC2017-87456-P) and AGAUR (2017-SGR-891).

## References

[1] Telefonica, Ericsson. ”Cloud-ran architecture for 5G.” White Paper (2015).

[2] GSM Association. ”5G Spectrum-Public Policy Position.” GSMA, London, UK, White Paper (2016).

[3] GSA White Paper, ”5G Network Slicing for Vertical Industries”, Sept. 2017.

[4] Jiang, H., Gui, G. (2020). Channel Modeling in 5G Wireless Communication Systems. Springer.

[5] Zeydan, Engin, and Omer Narmanlioglu. ”Dynamic slicing for mobile network infrastructures: Challenges, opportunities and business aspects.” EuCNC. IEEE, 2017.

[6] Makhijani, K., et al. ”Network slicing use cases: Network customization and differentiated services.” draft-netslices-usecases-02 (2017).

[7] 3GPP TS 22.101 V15.1.0, ”Study on Radio Access Network (RAN) sharing enhancements”, Jun. 2014.

[8] Han, Bin, and Hans D. Schotten. ”Machine Learning for Network Slicing Resource Management: A Comprehensive Survey.” arXiv, 2020.

[9] SdxCentral, ”How Is the RAN Network Evolving?”, Jan. 2018.

[10] Wind, an Intel company, white paper, ”vRAN: The Next Step in Network Transformation”, Nov. 2017.

[11] Qualcomm and Nokia, white paper, ”Making 5G a reality: Addressing the strong mobile broadband demand in 2019  beyond”, Sept. 2017.

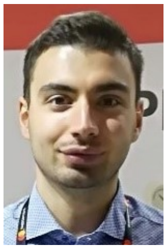
[12] Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J.,  Verikoukis, C. (2019, December),”Real-Time Dynamic Network Slicing for the 5G Radio Access Network”, GLOBECOM, 2019.

[13] Lu, Z., Hu, Z., Han, Z., Wang, L., Knopp, R.,  Zhang, Y. (2020). An Artificial Intelligence Enabled F-RAN Testbed. IEEE Wireless Communications, 27(2), 65-71.

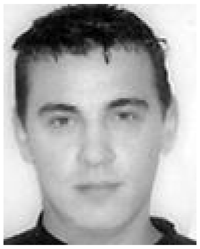
[14] Nomor research GmbH, ”3GPP 5G Adhoc:Any Decisions on RAN Internal Functional Split?”, Munich, Germany, January 26, 2017.

[15] Chandramouli, Devaki, Rainer Liebhart, and Juho Pirskanen, eds. 5G for the Connected World. John Wiley  Sons, Incorporated, 2019.

## VII. Authors



**Mr. Massimiliano Maule** received his B.Sc. degree in Information Engineering from the University of Padova, Padua, Italy, in 2014, and the M.Sc. degree in Telecommunication Engineering from the University of Trento, Trento, Italy, in 2017. From 2017 to 2018, he worked for Nokia as a Radio Researcher in Espoo, Finland and as a Network and System Architecture Engineer in Vimercate, Italy. Since July 2018, he is an Early Stage Researcher at Iquadrat Informatica S.L, Barcelona, Spain, in the context of the MSCA ITN 5G STEP FWD program.



**Mr. John S. Vardakas** (SM'20) received the Dipl.-Eng. in Electrical and Computer Engineering from the Democritus University of Thrace, Greece, in 2004 and his Ph.D from the Electrical Computer Engineering Dept., University of Patras, Greece. His research interests include performance analysis and optimization of communication networks and smart grids. He is a senior member of the IEEE and a member of the Technical Chamber of Greece (TEE).

**Mr. Christos Verikoukis** (SM'07) received the Ph.D. degree in broadband indoor wireless communications from UPC, Barcelona, Spain, in 2000. He is currently an Associate Professor with CEID, University of Patras, Greece. He has authored more than 138 journal articles, over 200 conference articles, 3 books, 14 book chapters, and 2 patents. He has participated in more than 30 competitive research projects, while he has supervised 19 Ph.D. students and 5 postdoctoral researchers. Dr. Verikoukis received the Best Paper Award at the IEEE ICC 2011  2020, the IEEE GLOBECOM 2014 and 2015, the EuCNC 2016, and the EURASIP 2013 Best Paper Award of the Journal on Advances in Signal Processing. Currently he is the IEEE ComSoc EMEA Director and Member-at-Large of IEEE ComSoc GITC.