Analysis of the Implementation of Network Slicing in 5G Radio Networks

Sérgio Marinheiro
Luis M. Correia
Instituto Superior Técnico / INESC-ID
University of Lisbon
Lisbon, Portugal
sergio.marinheiro9696@gmail.com
luis.m.correia@tecnico.ulisboa.pt

Ricardo Dinis
NOS SGPS
Lisbon, Portugal
ricardo.dinis@nos.pt

***Abstract* - The main goal of this thesis is to analyse one of the main technologies used in 5G radio networks, network slicing. To achieve this, several scenarios, resulting from making variations of a reference scenario, are deployed and further analysed. This work focuses on how the data rate should be allocated among different slices and users to maximise the cell capacity. Several types of services, SLAs, and priorities are taken into consideration when studying this problem. Alongside the parameters regarding the scenario with all the slices and slices specifications, the model takes as input the necessary parameters to compute the maximum achievable cell data rate. The model considers six parameters to evaluate the data rate allocation including the percentage of total assigned data rate, the VRRM capacity share, the total data rate of each service, the percentage of served users, the data rate of each user, and the users' satisfaction. The results show that the models' serving weights enable slice isolation and proper data rate allocation according to its services. Also, one can confirm that the total assigned data rate is always 100% independently of the scenario and number of users/services.**

Keywords-5G, Network Slicing, VNO, SLA, eMBB, URLLC, mMTC.

I. INTRODUCTION

There are four generations of mobile networks and we are currently getting closer to the fifth. With this comes a lot of expectations regarding 5G capabilities, like very high peak data rates, ultra-low latency, and being able to support massive amounts of devices connected to the network at the same time. Several applications are expected to appear or be enhanced with 5G capabilities, such as autonomous driving, remote surgery, smart power grids, smart cities, industrial automation, and many more. To support all these demands, new technologies need to be used.

The clear need for improvement demands a network with higher scalability flexibility and with a more efficient usage of its resources. Network slicing appears to tackle these problems and it is one of the most important and innovative technologies in 5G. This technology enables operators to deploy full networks in the form of logically separated and independent slices, each slice with its on configuration and customized according to one specific use case. This way, instead of trying to optimise one network for all users and all services, operators can focus on its time optimising slices for specific services [1].

Current solutions exploited over 4G systems are not able to perform resource allocation in slicing environments, which means that mobile phones receive resources from the same traffic with equal priority levels. This problem arises from the way resource allocation is managed in 4G, which is by associating a priority to a service requested by the user. In contrast, in 5G multiple users can be associated to multiple slices, where each slice has a designated priority, resulting in a priority that considers not only the priority of the slice as well as the priority of the service the user is requiring. Hence, higher user experiences are achieved in 5G compared to 4G due to traffic being satisfied by different slices [2].

The paper is organized as follows. Section II presents the state of the art. Section III contains the model development, presenting the model parameters, following by the data rate calculation method, the VRRM optimisation, the model implementation, and finally the presentation of the model assessment. Section IV contains the results analysis, where the reference scenario is described, as well as the variations made to it. Next, each of these variations are studied. In Section V, the most important conclusions of this work are presented.

II. STATE OF THE ART

The authors of [2] aim to tackle future conflicting demands that will appear in slice allocation. Network slicing will have conflicts regarding traffic prioritisation in the sense that it demands simultaneous management for priority among different slices and for priority among the users in the same slice. To solve this problem and with maximum user satisfaction as a goal, the authors propose a novel heuristic-based admission control mechanism that is capable of dynamically allocate network resources to different slices, while guaranteeing each slice meets its requirement. The model created has four main elements: the service slice layer, the virtual network layer, the physical resources, and the admission control manager. The service slice layer contains different services that need resources. The virtual network layer gives an abstraction of the physical network resources. The physical resources dictate what radio resources are available for the virtual network. Finally, the admission control manager is in

charge of new incoming slices and optimising resource allocation. The authors demonstrate that their model achieves higher user experience in individual slices, optimises network resource usage, and provides higher scalability when the number of users per slice increase.

Similar to the work described above, a solution for optimising resource allocation in network slices is implemented in [3]. This greatly benefits infrastructure providers (InPs) due to simplifying the work of implementing new admission control policies each time a request is made by a network slice, depending on their SLA. InPs can use traffic forecasting techniques to make the proper resource allocation to each slice and still meet the slice's SLAs. Three main blocks are described in [3], these being: traffic analysis and prediction per network slice, admission control decisions for network slice requests, and adaptive correction of the forecasted load based on measured deviations.

Network slices can be quickly deployed by recurring to SDN and NFV technologies. This will result in simplified management and optimal resource allocation, however knowing how to efficiently allocate, manage and control the network slice resources will prove to be a challenging task. The algorithmic side of these challenges is focused on [4].

The isolation challenges of network slicing are addressed in [5]. One of the main advantages of network virtualisation and network slicing is the capability of deploying multiple logical networks using the same physical infrastructure. Guaranteeing the proper resource usage of each slice and ensuring that each slice is isolated from the other is not an easy task that, if not done properly, may lead to congestion between slices. [5] discuss these challenges in the context of a wireless system having a time-varying number of users that need reliable low latency and self-managed network slices. A novel control framework for stochastic optimisation based on the Lyapunov drift-plus-penalty method is proposed in [5]. The main advantages of this framework are minimising the power consumption of the system, slice isolation and providing low latency end-to-end communication for RLL slices. The InP provides to the service provider (SP) one network slice with E2E requirement in terms of rate, maximum tolerable delay and reliability number of the users. It is in the SPs best interest and responsibility to control the number of users per slice because exciding a certain threshold would result in a QoS drop. The model described in the paper takes into consideration the number of physical RBs and the number of users per slice during a certain time interval.

III. Model Development

A. *Model Overview*

The purpose of this thesis is to develop a model capable of analysing the implementation of network slicing in 5G radio networks with service differentiation, by defining the resources allocated to each slice and the corresponding SLA in order to achieve near-optimum performance. Figure 1 presents an overview of the model developed and both the inputs and outputs.

The model inputs consist of two classes: cell and network. Each of these classes have different parameters that are worth looking at. Regarding the first class, Cell, these inputs are used in the equation for throughput calculation for NR provided by 3GPP. The remaining equation parameters are not listed as inputs because they are calculated with the ones that are listed as inputs. For example, to know how many RBs are available, one needs to know which numerology is being used and which is the available bandwidth.

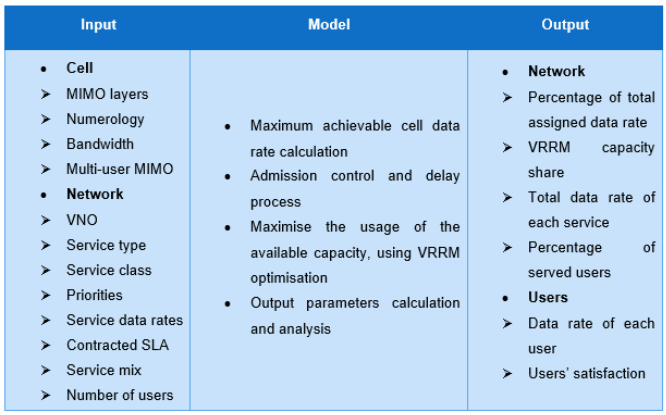


Figure 1. Model overview

The second class, Network, is composed of several parameters that define one slice and are used as inputs for the VRRM. VNO is the identification tag of VNO. Service Type refers to which type of service is being provided (e.g., voice). Service class is the class in which the service is inserted (e.g., conversational). Priorities refers to the priority level given to the service. This parameter is composed of two variables, one regarding the priority given by the VNO and the other given by the InP. The Service data rates, which is the acceptable range of data rate for a given service. The Contracted SLA is the type of SLA defined between VNO and InP (GB, BG or BE). The Service mix, which is the percentage of users that are assigned to each service. Finally, the Number of users that is the total number of users allocated to the slice.

The model consists of four main stages. First the maximum achievable cell data rate calculation. This stage uses the cell input parameters detailed above, to make this calculation. The output serves as an input of the third stage, the VRRM optimisation. The second stage is the admission control and delay process, which is essential to guarantee that the VRRM problem is possible to solve. When the network is congested, and the total minimum thresholds of guaranteed demands get higher than the available VRRM capacity, it is necessary to delay the BE users because they do not have a guaranteed bit rate associated with their SLA. Then, low priority users is delayed one by one until the capacity is enough. This stage uses the capacity calculation of the first stage and network input parameters, specifically it makes the summation of the minimum demanded capacities of all services, in order to compare the two and make the admission. The third stage is solving the VRRM problem, which aims to maximise the usage of available capacity in order to satisfy the contracted SLAs to the highest possible level, considering services' priority, while

distributing capacity in a fair manner, subject to some constraints, including maximum achievable capacity, predefined SLA thresholds [6]. This stage uses the calculation of the maximum achievable cell capacity provided by the first stage and the network input parameters.

The model outputs also divide into two different classes: network and users. The first one has parameters that are important from the network viewpoint, and the second class contains parameters that are important from the users' viewpoint: Each parameter and respective expression are detailed in the respective subsection.

B. Maximum Achievable Cell Data Rate

In order to calculate how the data rate distribution among the services is done, first one needs to know the maximum achievable data rate. For that purpose, one uses an adapted expression provided by 3GPP that takes into consideration several parameters, such as number of MIMO layers and modulation, making the calculation for the DL or UP data rate [7].

$$R_{cell\ [\mathrm{Mbps}]} = \sum_{j=1}^{J} \left(v_{Layers}^{(j)}\ Q_m^{(j)}\ f^{(j)}\ R_{max}\ \frac{12\ N_{PRB}^{BW(j),\mu}}{T_{s\ [\mathrm{s}]}^{\mu}}\ \left(1 - O_h^{(j)}\right)\ M_{UMIMO} \right) \quad (1)$$

where:

- R_{cell}: Achievable data rate in one cell.
- J: Number of aggregated component carriers in a band or band combination.
- R_{max}: Maximum coding rate.
- For the j-th component carrier:
  - $v_{Layers}^{(j)}$: Maximum number of supported MIMO layers.
  - $Q_m^{(j)}$: Maximum supported modulation order.
  - $f^{(j)}$: Scaling factor.
  - $N_{RB}^{BW(j),\mu}$: Maximum number of RB allocation in bandwidth $BW^{(j)}$ with numerology μ.
  - T_s^{μ}: Average OFDM symbol duration in a subframe for numerology μ.
  - $O_h^{(j)}$: Overhead.
  - M_{UMIMO}: Multi-user MIMO factor

The added parameter, M_{UMIMO}, takes into consideration Multi-user MIMO, which was lacking from the original expression, which allows the base station to serve multiple users, that are close to it, with the same RB making a much more efficient usage of resources.

5G uses a scalable OFDM in the sense that each subcarrier is separated 15×2^n kHz ensuring it is possible to provide a variety of services on a wide range of frequencies. As such, the available numerologies for FR1 are 0, 1, and 2 that correspond to a subcarrier spacing (SCS) of 15 kHz, 30 kHz and 60 kHz respectively. With μ one can calculate T_s^{μ} as shown in expression (2).

$$T_{s\ [\mathrm{s}]}^{\mu} = \frac{10^{-3}}{14 \times 2^{\mu}} \quad (2)$$

C. VRRM Optimisation

VRRM aims to maximise the usage of aggregated capacity and then allocate the resources to the different services of the VNOs by considering the set of available radio resources. This model was developed by [6] and its explanation is given bellow.

The model is formulated as a constrained concave optimisation problem. The objective function balances the efficiency and fairness when allocating resources to a multi service network. This function is a measure of efficiency and quantifies the expected users' satisfaction. The chosen utility function, in order to formulate the problem as a convex optimisation, is the logarithmic one. This function copes with the criterion of proportional fairness, by being an effective function to maximise the average long-term users' data rate while sharing the capacity among users in accordance with predefined weights [6]. The logarithmic utility function makes the balance between two competing interests, these being providing the maximum capacity to a user and at the same time providing all users with the minimal level of service. Therefore, the problem of the objective function of VRRM, $f_{VRRM}(\boldsymbol{R^{srv}})$, is formulated as the logarithm of the normalised weighted sum of the total data rate for different services.

$$Max\ f_{VRRM}(\boldsymbol{w^{usr}}) = Max \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log \left(\sum_{v_s=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} \frac{R_{v_s\ [\mathrm{Mbps}]}^{srv_{max}}}{R_{cell\ [\mathrm{Mbps}]}} \right) \quad (3)$$

where:

- $\boldsymbol{w^{usr}}$: Vector of users' weights, to obtain the long-term average data rate of users, which can be written as $\boldsymbol{w^{usr}} = \left[w_{1,1}^{usr}, \dots, w_{N_1^{usr}}^{usr}, \dots, w_{N^{srv},1}^{usr}, \dots, w_{N^{srv},N_{N^{srv}}^{usr}}^{usr} \right]$
- N^{srv}: Number of services.
- $N_{v_s}^{usr}$: Number of users performing service s, from VNO v.
- $R_{vs\ [Mbps]}^{srv_{max}}$: Maximum assignable data rate to the user of service s, from VNO v.
- $w_{v_s,i}^{usr}$: Assigned weight to user i, performing service s, from VNO v, ranging in [0,1].
- λ_{v_s}: Tuning weight associated with service s, provided by VNO v, to prioritise data rate assignment.

There are two constraints associated with the problem of VRRM and the objective function has to be solved respecting these constraints. The first constrain considers that the average long-term data rate assigned to each user has to fall within this acceptable data rate interval due to VNO policies:

$$R_{v_s\ [\mathrm{Mbps}]}^{srv_{min}} \leq w_{v_s,i}^{usr} R_{v_s\ [\mathrm{Mbps}]}^{srv_{max}} \leq R_{v_s\ [\mathrm{Mbps}]}^{srv_{max}} \quad (4)$$

where:

- $R_{vs\ [Mbps]}^{srv_{min}}$: Minimum assignable data rate to the user of service s, from VNO v

The second constrain is a logical constraint, which indicates that the whole bandwidth allocated to all users cannot exceed

the total aggregated cell capacity. Therefore, the entire VRRM bandwidth assigned to all users are subject to an upper bound defined by the InP:

$$\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{[\text{Mbps}]}^{cell} \tag{5}$$

D. Model Output Parameters

It is important to know if the overall user experience and network performance are working as intended. Several evaluation metrics are defined with the goal of measuring these performance requirements. The metrics used for network performance evaluation are here described.

1) Percentage of total assigned data rate

One of the most important metrics, showing the total network throughput in terms of data rate allocation, the values closer to 100% obviously leading to a better VRRM performance:

$$p_{VRRM[\%]}^{tot} = 100 \ \frac{\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}}}{R_{[\text{Mbps}]}^{cell}} \tag{6}$$

2) VRRM capacity share

The percentage of capacity allocated to each VNO, out of the total available VRRM one, being a key performance metric from VNOs and VRRM perspective

$$R_{VRRM[\%]}^{VNO_v} = 100 \ \frac{\sum_{v_s=1}^{N_{VNO_v}^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}}}{R_{[\text{Mbps}]}^{cell}} \tag{7}$$

3) Total data rate of each service

It shows the total data rate assigned to each service slice of a VNO, being important from both VRRM and VNOs' viewpoints:

$$R_{v_s[\text{Mbps}]}^{srv_{tot}} = \sum_{i=1}^{N_k^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \tag{8}$$

4) Percentage of served users

The percentage of served users performing a specific service out of the total number of users from that service, which is an essential metric for VNOs:

$$p_{v_s[\%]}^{usr_{net}} = 100 \ \frac{N_{v_s}^{usr}}{N^{usr_{tot}}} \tag{9}$$

5) Data rate of each user

The data rate allocated to a user is an important QoS metric from both users' and VNOs' viewpoints, having a direct impact on the satisfaction level of the served users:

$$R_{v_s,i\,[\text{Mbps}]}^{usr} = w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \tag{10}$$

6) Users' satisfaction, $S_{v_s,i}^{usr}$

It is important from a VNO perspective because it reflects the user satisfaction. This metric is measured differently according to the service class and/or type. For voice, one uses AMR-WB codecs with the respective voice quality mean opinion score (MOS) values provided by NOS. For the remaining services, a similar way of classification is used. Depending on the service type, it is defined five levels of user satisfaction, where just like MOS one represents bad quality and five represents excellent quality. For Background there is no expression defined due to the service nature.

E. Model Implementation

Figure 2 illustrates the flowchart of the model. The input parameters are loaded from an excel data sheet and the maximum achievable cell capacity is calculated with the cell input parameters.

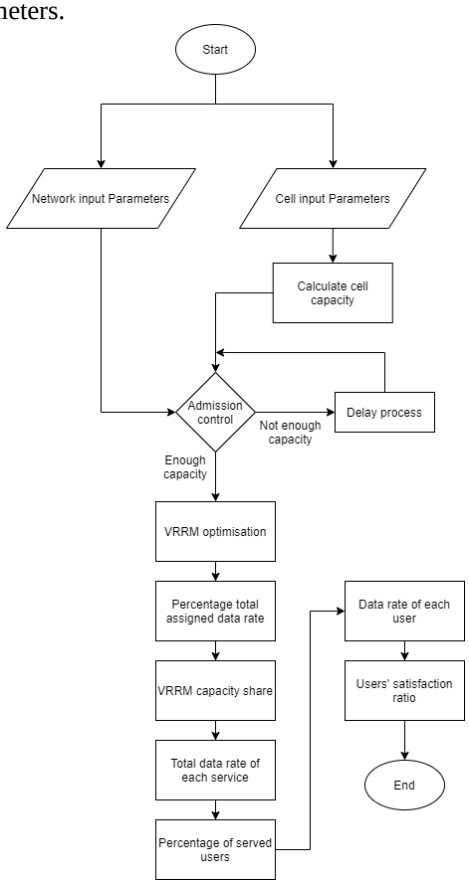


Figure 2 – Model flowchart.

After calculating the maximum achievable cell capacity, the admission control and delay process algorithm take place. There are two possible scenarios at this point.

- First scenario, the previously calculated cell capacity is enough to serve all users, with at least the minimum service requirements. In this scenario the algorithm admission control and delay process algorithm are not executed.
- Second scenario, the cell capacity is not enough to serve all users. This last scenario is the reason why it is necessary to implement a delay process algorithm and ensure that higher priority services are served with the minimum demanded capacity.

Once it is verified that all users can be served with at least the minimum demanded capacity for their service, the VRRM optimisation algorithm is initiated. After running the VRRM optimisation, the programme computes the several output parameters that reflect the overall network performance and user satisfaction: percentage of total assigned data rate, VRRM capacity share, total data rate of each service, percentage of served users, data rate of each user and the users' satisfaction.

F. Model Assessment

The purpose of the model assessment is to validate the implemented model. This is done by defining a set of tests in which the output was previously calculated. Valid outputs will confirm the model validation. Table 1 shows the set of chosen tests.

Table 1 - Module Assessment Tests.

Number	Description.
1	Validation of the input file read, by verifying if the type of variable is correct.
2	Validation of the input variables, by verifying if the parameters are correctly stored in memory.
3	Validation of the computation of the cell capacity: • Check the computation of T_s^{μ}. • Check if the overhead, the number of resource blocks, and the modulation order are correct. • Check if the computation R_{cell} is correct.
4	Validation of the admission control and delay process: • Check if the available capacity is not enough to serve all users with minimum demanded data rate. • Check if all BE users are delayed. • Check if the rest of delayed users were delayed based on their slice and service priorities. Check if the new minimum demanded capacity is equal or approximate to the previously calculated cell capacity.
5	Validation of the VRRM optimisation: • Check if the size and values of the created array with the minimum demanded capacity for all users are correct. • Check if the CVX status is solved. • Check if the computed values of the users' weights give the optimal solution to the problem. • Check if the imposed restrictions are being complied with.
6	Validation of the output files, by checking if they are correctly printed and plotting the output results.

IV. Results Analysis

A. Scenarios

The scenario, chosen by NOS and presented in Table 2, represents a case study with multiple VNOs sharing the resources of a cell. Each VNO is characterised with an SLA and an assigned priority level, γ_s, provided by the InP. Within the VNO there are multiple services that the VNO intends to provide, each characterised by the class of service it belongs to, the minimum and maximum required data rate to provide the service, and the priority assigned to the service, δ_s, by the VNO. The last parameter, Mix, shows how the users are distributed among the several services of all VNOs.

The parameters for the maximum achievable data rate computation are the following:

- MIMO layers, $v_{Layers}^{(j)}$: 4.
- Numerology, μ: 1.
- Bandwidth [MHz]: 100.
- Multi-user MIMO, M_{UMIMO}: 2.

With the intent of better analysing and representing other 5G cases, some variations are made to the reference scenario. Besides the cell input parameters, each modification has a dedicated section intended for its study.

Table 2 – Reference, URLLC and mMTC scenarios.

VNO	Service	γ_s	δ_s	SLA	Mix RS [%]	Mix URLLC [%]	Mix mMTC [%]		Users
							Relative	Global	
GB Real-time (RT)	Voice	30	10	GB	20	18	20	0.83	100
	Video		8		35	30	35	1.45	
	Music		9		5	5	5	0.21	
BG Interactive (IA)	SN	20	5	BG	10	8	10	0.41	
	Web		6		10	8	10	0.41	
	FS		4		3	3	3	0.12	
BG RT	VR/AR	40	7	BG	2	3	2	0.083	
	RTG		6		2	5	2	0.083	
BE Background (BkG)	Email	10	3	BE	3	3	3	0.12	
	IoT		2		10	8	10	0.41	
GB URLLC (L)	FA	40	10	GB	-	2	-	-	
	RSI		11		-	5	-	-	
	RS		12		-	2	-	-	
BE uMMTC (MM)	SM	1	1	BE	-	-	100	95.85	2311

The cell input parameters changes are studied in the subsection reserved for the study of the reference scenario. These inputs are the BW and the Multi-user MIMO. BW assumes three different values that correspond to the allocated BW for the frequencies used in 5G. Multi-user MIMO also assumes three values but only when the BW is 100 MHz BW as it is the only case where this factor is relevant. The variation on the Number of MIMO layers and numerology intend to increase the scientific coherence of this

study (a comparison between the results is not relevant). MIMO layers changes according to the BW. For 10 MHz and 20 MHz, MIMO layers takes a value of 2, whilst for 100 MHz, it takes a value of 4. The numerology is maintained at 1, except in the URRLC scenario where it changes to 2. Table 3 shows the summary of the variations performed to the input parameters.

Table 3 – Reference scenario variations.

MIMO layers, $v_{Layers}^{(j)}$	{2, 4}
Numerology, μ	{1, 2}
Bandwidth [MHz]	{10, 20, 100}
Multi-user MIMO, M_{UMIMO}	{1, 2, 3}
Number of VNOs	[4, 8]
Priority, γ_s	[10, 100]
Service mix	[0, 100]
Number of Users, $N_{v_s}^{usr}$	[50, 1100]

The next set of variations is to the mix and number of users. First it is necessary to, maintaining the mix, increase the number of users from 50 to 100, 150, 200, and so on until the programme reaches its limit. It is beneficial to understand not only how the model reacts to an increasing number of users allocated to a specific cell but also determine the maximum number of users one cell can serve simultaneously, for a given mix. From this point on, all scenarios are subjected to this variation in the number of users. As for the mix variations, two new scenarios were conceived and are presented in Table 4.

Table 4 – Service mix variation scenarios.

VNO	Service	γ_s	δ_s	SLA	Mix Football [%]	Mix Hospital [%]	Users
GB RT	Voice	30	10	GB	5	30	100
	Video		8		1	5	
	Music		9		1	5	
BG IA	SN	20	5	BG	25	3	
	Web		6		21	10	
	FS		4		1	8	
BG RT	VR/AR	40	7	BG	30	8	
	RTG		6		5	3	
BE BkG	Email	10	3	BE	1	10	
	IoT		2		10	8	
GB L	FA	40	10	GB	-	0	
	RSI		11		-	0	
	RS		12		-	10	

These new scenarios are direct variations of the reference and URLLC scenarios: the first representing a football game scenario and the second representing a hospital scenario. This way one can study how the programme responds to people's needs in several different locations.

The final variation is related to the priorities assigned from the InP to the VNO, γ_s. This subsection has two goals, the first being the study of two similar VNOs where one has a much higher priority than the other, creating a "premium" and "low-cost" slices type scenario, this scenario being presented in Table 5, and the second goal is to increase the priority assigned to one VNO, specifically VNO GB L, to study the behaviour of the model and determine the ideal value for this priority.

Table 5 – Premium and low-cost slices scenario.

VNO	Service	γ_s	δ_s	SLA	Mix [%]	Users
GB RT	Voice	30	10	GB	10	100
	Video		8		17	
	Music		9		3	
GB RT 2	Voice	15	10	GB	10	
	Video		8		17	
	Music		9		3	
BG IA	SN	20	5	BG	5	
	Web		6		5	
	FS		4		2	
BG IA 2	SN	10	5	BG	5	
	Web		6		5	
	FS		4		2	
BG RT	VR/AR	40	7	BG	1	
	RTG		6		1	
BG RT 2	VR/AR	20	7	BG	1	
	RTF		6		1	
BE BkG	Email	10	3	BE	1	
	IoT		2		5	
BE BkG 2	Email	5	3	BE	1	
	IoT		2		5	

B. *Reference scenario study*

For the reference scenario, $R_{cell\,[\mathrm{Mbps}]} = 1777.30$ and the minimum demanded data rate by all users is 238.21 Mbps, which means that the cell can serve all users. Figure 3 depicts the results obtained by varying M_{UMIMO}, with BW fixed to 100 MHz.

In this scenario, using $M_{UMIMO} = 3$ would be relevant because it enables VNO BG RT to provide its services with better user satisfaction, with all services having the best score in user satisfaction except VR with a score of good at 164.06 Mbps. Thus, using $M_{UMIMO} = 3$ would be more useful when the number of users and the data rate demands increase. On the other hand, using $M_{UMIMO} = 1$ is not enough to achieve good user satisfaction among all services. VNO BG RT can only achieve a fair and good user satisfaction for virtual reality and real time gaming, respectively.

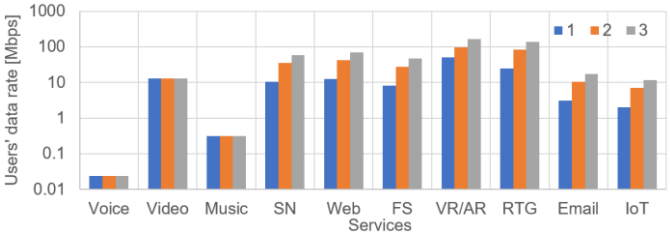


Figure 3 - Comparison between users' data rates for different M_{UMIMO} values.

When BW = 10 MHz, it is not possible to serve all users due to the lack of resources, Figure 4. As such, the delay process algorithm is executed delaying all users from VNO BE BKG,

which has a BE SLA, and then starts delaying users one by one based on priorities and SLAs. Because the algorithm was executed, the VRRM optimisation did not happen, therefore $\boldsymbol{w}^{usr}$ has no assigned values. VNO BG IA is the first to be delayed, followed by VNO BG RT, seeing as both have BG SLA. The last VNO, VNO GB RT, with a GB SLA has all its users served with minimum demanded data rate for each service. One can infer that the lack of resources is due to VNO BG RT, which includes services with high demands of data rate. Due to the high priority of γ_s set between the VNO and the InP, VNO BG IA is affected with all its users not being served, which is why only 59.87% of the cell resources are being used. Because all services are being served with minimum demands, the user satisfaction is bad except for Voice being classified with a fair value according to MOS. The 700 MHz frequency is going to be used mainly in rural areas where this problem will not have impact.

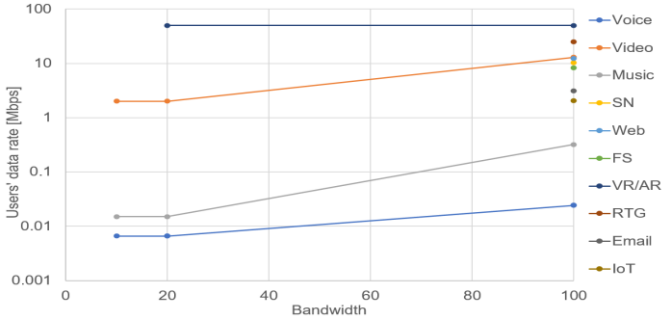


Figure 4 - Comparison between users' data rates for different BWs.

C. *Influence of Incrementing the Number of users*

Figure 5 illustrates the impact of increasing the number of users in the reference scenario.

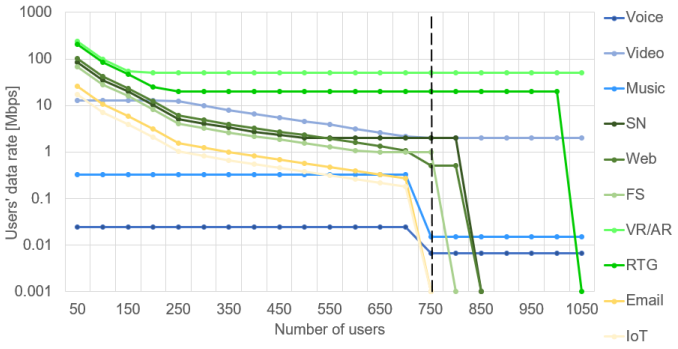


Figure 5 – Impact of the increment of number of users in the reference scenario.

The model behaves as expected and the decision making for the allocated data rate for each service follows some principles. Firstly, the data rate range, which is particular to each service, defined by the VNO. This data range value, together with the priority, λ_{v_s}, work together to decide how much data rate is allocated to a specific user. Secondly, the SLA type that each VNO has, which is a key factor in deciding what services get delayed first if there are not enough resources to serve all users. With the increased number of users, the allocated data rate of these services starts to decrease while the GB ones remain stable. VR/AR and RTG are the first services to reach the minimum threshold of allocated data rate, due to their very high demanded values, with VR/AR starting at 200 users and RTG at 250. The other BG services will not reach their respective minimums until 500, when SN does. The only GB service that is also affected by these numbers of users is video, due to its data rate requirement. Marked by the dotted line, in Figure 5, is the beginning of the delay process algorithm. At this point, all services are being served with the minimum defined data rate and the cell does not have enough capacity to serve more users. Both email and IoT are delayed first since they do not meet the minimum demands associated with their service. Next, by order of the lowest priority service from the lowest priority VNO, users are delayed one by one until the cell has enough capacity to serve the remaining users.

Figure 5 alone is not enough to study the scenario because it only shows the data rate of the served users, meaning it does not encompass the full spectrum of users that are getting delayed and not having their needs fulfilled, so Figure 6 is presented as well.

Figure 6 details the exact percentage of the served users from each service. The first "wave" of delays is for email, IoT and FS, where the first two are completely delayed and the last needs to delay 40% of its users. The second "wave" of delays, affects all FS users and around 75% of web users. This process repeats itself like the figure is showing. Voice, video, and music are the only services that are never delayed due to the combination of the three principles mentioned at the beginning of the subsection. Both are within a GB VNO, have the highest priorities among the services in that VNO, and have low data rate demands.

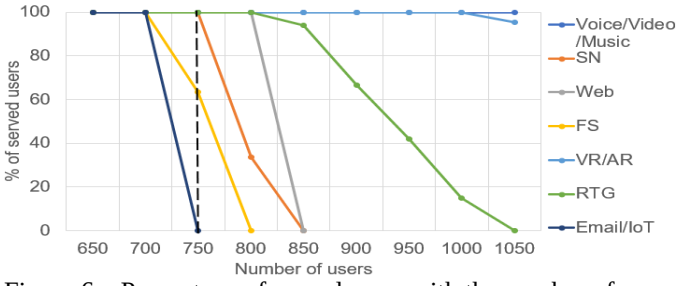


Figure 6 – Percentage of served users with the number of users increasing.

Presented in Figure 7, GB VNO sees an increase in its capacity share until the point it starts reducing the allocated data rate for its users. More specifically, when the number of users surpasses 250, video users no longer are allocated 13 Mbps but rather a lower value that continues decreasing with the increment of the number of users. The capacity share of VNO GB RT continues to decrease until the number of users is 750, point where CVX optimisation stops and the delay algorithm starts delaying users. BE VNO users are all delayed, and BG VNO users start being delayed one by one. The natural consequence is to see an increase of VNO GB RT that maintains all its users with minimum demanded data rate. VNO BG RT appears to behave differently, which is true to some extent, but most of its behaviour is similar to VNO GB RT with a shift in the curve. The start is different because VNO BG RT has a BG SLA that does not limit the amount of data rate allocated to its service. This means that with the increase of the number of

users, the allocated data rate per user starts decreasing until it reaches the minimum demanded by the service. For the specific case of VNO BG RT, the minimum occurs when the number of users is 200 and all VR/AR users are being allocated 50 Mbps. From here, it assumes the same behaviour that VNO GB RT did for 750 users, with all users at a fixed minimum capacity and with the number of users increasing its natural that the capacity share also increases. The second point worth mentioning is at 850 users when all VNO BG IA users have been delayed and the first RTG users start to be delayed, resulting in the decrease of VNO BG RT capacity share.

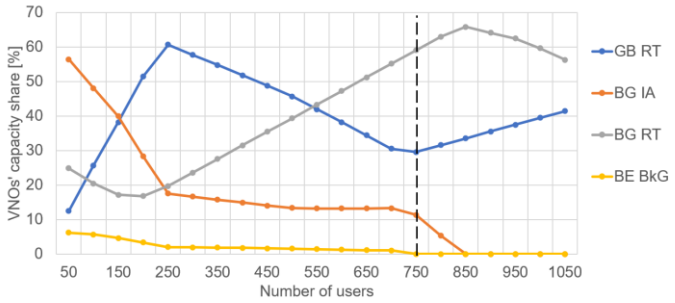


Figure 7 – VNOs' capacity share with the number of users increasing.

D. Impact of the URLLC VNO

Comparing it to the reference scenario, one can observe a similar behaviour, but more crowded due to the added VNO that brings new services. As a result, the delay process starts at 450 users, as opposed to 750 users. This figure, however, does not represent the percentage of allocated users to each service. For example, between 450 and 550 RTG users' data rate maintains constant, but the percentage of served users is dropping, which results in an overall less data rate allocated to RTG. Figure 9 illustrates how the VNOs compete for the capacity with the number of users increasing. The two GB VNOs behave in a similar fashion, starting with a small percentage of the cell capacity and increasing up until the point where one of its services starts getting allocated less data rate than its maximum. For GB RT peaks at 200 users while GB L peaks at 100 users. Comparing with the precious scenario, GB RT has roughly less 20% of the capacity share because of the new added VNO also using the cell resources. Both VNOs gradually decrease their capacity share until their services start serving with minimum demands of data rate and start increasing again from this point onwards.

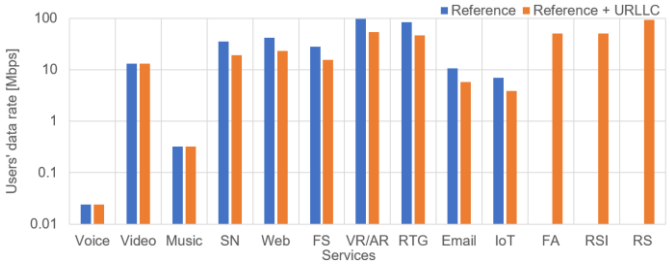


Figure 8 - Comparison between users' data rates for the reference scenario and the URLLC scenario.

BG RT manages to peak at 62%, being the VNO with more capacity allocated at a given point, the reasoning being the high demand services it has, even when being served minimum demands. From this point, there is a drop on its share resulting from the delay of RTG users. BG IA and BE BkG see a gradual decrease in the data rate allocated to them, due to the nature of their SLA and priorities. When the delay process starts, at 450 users, BG IA has only 2.28% of the cell capacity.

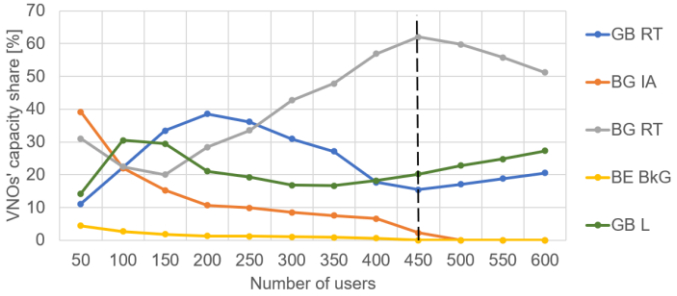


Figure 9 - VNOs' capacity share with the number of users increasing.

E. Impact of the mMTC VNO

In practice, this scenario leads to results similar to the reference scenario. Even with thousands of devices connected to the cell, their demands are so low that the outputs of the model are almost the same. In the end, all the data rate that gets allocated to this VNO is 46.22 Mbps which is roughly equivalent to one VR/AR user. Even with such low data rate requirements, the devices are not always served with maximum data rate, because the priority associated to this service is 1, i.e., 300 times lower than the priority of voice for example. Therefore, the impact these devices have in the maximisation of the objective function is much lower. Resulting from the explanation above, the difference between the VNOs' capacity share from the reference scenario and this one is computed, which uses

$$R_{VRRM[\%]}^{VNO_v} = R_{VRRM[\%]}^{VNO_{v\_Ref}} - R_{VRRM[\%]}^{VNO_{v\_mMTC}} \quad (11)$$

Regarding the VNO GB MM's capacity share, not only it never occupies a big part in the division of data rate among VNOs (highest percentage being 2.6%), but it also starts decreasing when the number of users is greater than 450 (lowest percentage 0.8%). That is due to the priority set to this VNO, as explained above. Further important conclusions, concerning the remaining VNOs' capacity share, are the following. VNO GB RT is only different when the number of users ranges from 250 to 750. This is because the allocation of data rate for video is being optimised, by the CVX solver, in this range of values. The result is not the same because of the existence of VNO GB MM in one of the scenarios. Outside this range of values, the difference is zero because all services are being served with either maximum or minimum data rate. For the other VNOs the same logic applies. When all the services belonging to a VNO do not suffer changes to their allocated data rate, the value maintains constant, whilst when there is optimisation or delaying being done the difference between VNOs capacity share changes. All these differences are almost overlooked with the biggest difference, outside the new added VNO, being 2%.

F. Influence of Service Mix

1) Football Scenario

The obtained results show that this is a very demanding scenario where most services are being served with low data rate, which results in a worse user satisfaction. The main service for this scenario VR/AR is being served with minimum demanded data rate, 50 Mbps, because there are a lot of users allocated to this service. Also, when studying the impact of adding more users, the problem becomes automatically impossible to solve due to the lack of capacity in the cell. The delay process starts in the first iteration, where the number of users is 150. VNOs should take into consideration that increasing the number of people using this type of service requires more capacity than usual due to its high demands. Using higher levels of M_{UMIMO} and carrier aggregation are some possibilities that will help improve scenarios like this one.

2) Hospital Scenario

In Figure 10, the VNOs' capacity share evolution is presented, with the increase of served users. For the reference number of users, the most important VNO of this scenario, GB L, is being allocated the biggest percentage out of all the VNOs. At the point of delay, BG RT has almost 80% of the cell capacity, because the minimum demands of data rate of its services and number of users are much greater than the other services. This result is inevitable, because it does not depend on the priority associated with the service. Nevertheless, it is possible for VNO GB L to have a bigger share of the cell capacity prior to this threshold by increasing the priority γ, meaning paying more to the respective InP.

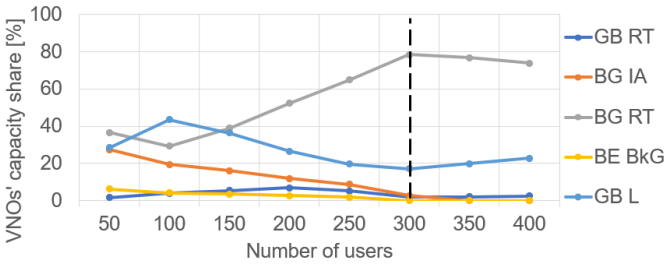
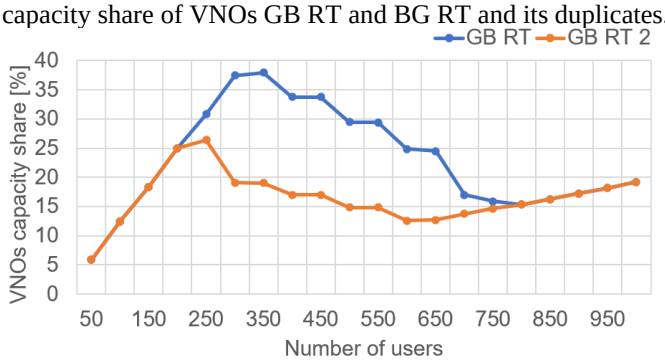


Figure 10 – VNOs' capacity share with the number of users increasing for the hospital scenario.

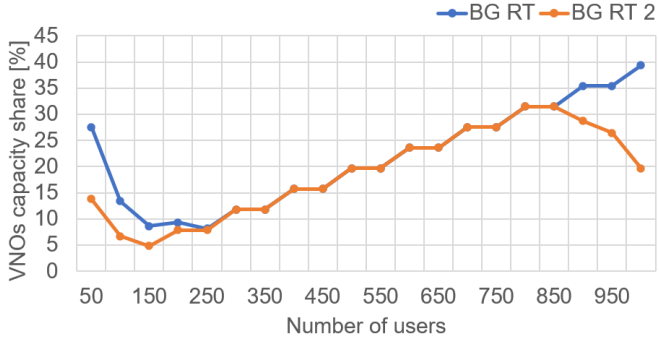
G. Influence of the Priorities

1) VNO Duplication Analysis

VNO GB RT, that comprises voice, video and music, sees that both voice and music do not have any difference regarding the data rate serving its users. Both have the same minimum and maximum demand and one has half the priority of the other. This happens because these services are extremely low data rate demanding so the cell has no problems serving both the same way. On the other hand, the same does not apply to video. The higher priority service has a significant amount of more data rate allocated to him, between 200 and 800 served users. In this zone, the cell needs to manage its resources and prioritises the service of the VNO that paid more. Outside this range both services have the same allocated data rate because either there is enough capacity to serve both at maximum demanded data rate (<200 users) or there are so many users that both services need to operate at the minimum demanded data rate (>800 users). As for the BG RT VNO and its duplicate lower quality VNO, the main difference is verified prior to 200 users, for VR/AR, and 300 users, for RTG. These are high demanding services, so it is natural that they start being served with minimum demanded data rate sooner than other services. Regarding the evolution of the other services, this is the same across all of them: the duplicate VNO services are allocated half the data rate of the main VNO until they converge to the minimum demand or zero. Figure 11 illustrates the VNOs' capacity share of VNOs GB RT and BG RT and its duplicates.



(a) GB RT VNO



(b) BG RT VNO

Figure 11 - VNOs' capacity share with the number of users increasing of VNOs GB RT and BG RT.

With the analysis of these results the VNO can determine whether it should or should not invest in a certain slice depending on the target audience. For example, if the average number of users of a certain cell is 350, VNOs that want to provide video should invest in better quality slices because the difference in the capacity share is around 18%, the biggest gap verified.

2) γ_s Impact Analysis

Figure 12 shows the evolution of the users' allocated data rate with changing γ_s and a constant number of 100 users. FA and RSI converge to their respective maximums when $\gamma_s = 30$, but only when $\gamma_s = 50$, the same is true for RS, because of the higher maximum demanded data rate of the service. The other services also see some changes resulting from more data rate being allocated to VNO GB L, with the ones being more affected being services from BG RT.

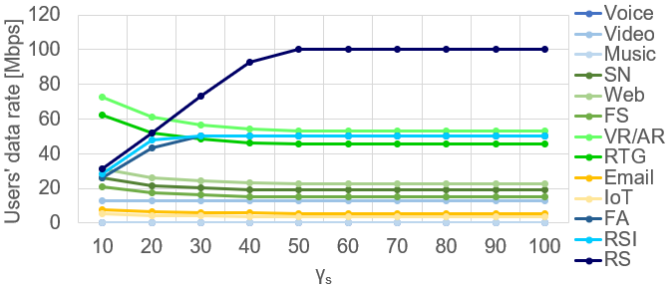


Figure 12 – Users' data rate with γ_s increasing.

V. CONCLUSIONS

The results from increment of the number of users to the reference scenario test confirm what was to be expected regarding the behaviour of user's allocated data rate. GB VNOs have their users served with maximum data rate, for the respective service, and BG and BE VNO have the maximum possible data rate according to their priorities. When the number of users increases first one can see a decrease in the data rate of BG and BE VNOs and only then the GB VNO. After 750 users the cell can no longer support the increasing rate of users, so it starts delaying them. The VNO GB RT never has any user delayed because of its SLA and priority. In this situation one can also see the isolation factor between slices that protects VNO GB RT from the other users in the sense that when users from other VNOs are delayed this does not affect the users from the first VNO.

Concerning the URRLC VNO, this added VNO brings users with considerable amounts of data rate. The results show the impact this VNO has in the reference scenario, BG and BE VNO services are allocated much less data rate. Considering 100 users, VR/AR has a 43.89 Mbps decrease and RTG 37.8 Mbps decrease. The delay point now is at 450 users as opposed to 750 users. Something to be noted is that the allocated data rate to the URLLC VNO, VNO GB L, is always above the minimum threshold due to its high priority level, only reaching it at 400 users.

Afterwards, the mMTC scenario results may be surprising as not much has changed from the reference scenario. There is a slight variation to the data rate allocated to each user, slightly less, but not enough to have an impact. This happens because even though thousands of devices are connected, their data rate requirements are so low, they end up not having an impact on the network traffic. One can conclude that if the network is prepared to connect thousands of devices, as a 5G network should be, they will offer no difficulties or problems to other existent VNOs.

In the football scenario the cell has barely enough resources to serve all users because the incredibly amount of VR/AR users are consuming almost all the resources. It is not even possible to make a study incrementing the number of users because the model automatically enters the delay process algorithm which means the cell is overloaded. This scenario proves that in order to serve high demanding data rate services such as VR/AR, not only the 100 MHz BW needs to be used, but also other techniques that increase the data rate such as carrier aggregation, multi-user MIMO and so on.

In the hospital scenario, the users from RS, the most important service in this scenario, are allocated 76.29 Mbps which corresponds to a "fair" users' satisfaction just shy of "good" at 80 Mbps. The service is always being allocated data rate even when the delay process starts. The remaining services are also allocated data rates such that its users fall between the classification of "fair" and "excellent", which is also very positive. Regarding the capacity share, for 100 users VNO GB L, is being allocated the biggest percentage out of all the VNOs. At the point of delay, BG RT has almost 80% of the cell capacity. Nevertheless, it is possible for VNO GB L to have a bigger share of the cell capacity prior to this threshold by increasing the priority γ, meaning paying more to the respective InP.

Concerning the VNO duplication analysis, the evolution of the capacity share of the VNOs show that for GB RT the main difference between higher and lower priority VNOs is when the number of users ranges from 200 and 800 users due to the service video allocated data rate being optimised. This is the opposite of what happens for the VNO BG RT where between 250 users and 850 capacity share is the same due to its services being served with minimum demands of data rate. With this information, VNOs can choose what contract to do with the InP, based on the average number of users and the type of traffic they intend to serve.

Regarding the influence of γ_s, one can see an increase in allocated data rate with the increase of the priority level, as to be expected, and a saturation of the allocated data rate when the VNO priority is 50. Thus, all users from VNO GB L experiencing maximum data rate and are fully satisfied.

REFERENCES

[1] B. Han, S. Tayade and H. D. Schotten, "Modeling profit of sliced 5G networks for advanced network resource management and slice implementation," in Proc. of 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 576-581

[2] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network Slicing Management & Prioritisation in 5G Mobile Systems," Euro. Wireless 2016, Oulu, Finland, 2016, pp. 1–6. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7499297

[3] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia and A. Banchs, "Mobile traffic forecasting for maximising 5G network slicing resource utilisation," in Proc. of IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, GA, 2017, pp. 1-9.

[4] S. Vassilaras et al., "The Algorithmic Aspects of Network Slicing," IEEE Communications Magazine, vol. 55, no. 8, pp. 112-119, Aug. 2017.

[5] A.T.Z. Kasgari and W. Saad, "Stochastic optimisation and control framework for 5G network slicing with effective isolation," in Prof. of 52nd Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 2018, pp. 1-6. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8362271&isnumber=8362188

[6] B. Rouzbehani, On-demand RAN Slicing Techniques for SLA Assurance in Virtual Wireless Networks, Ph.D. Thesis, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2019.

[7] 3GPP, User Equipment (UE) radio access capabilities, TS 38.3063, Mar. 2020.