Modeling the resource allocation in 5G radio access networks with network slicing

Daniel Haro-Mendoza
[1]*Universidad Nacional de Chimborazo*
Ecuador
[2]*Universidad Nacional de La Plata*
Argentina

Luis Tello-Oquendo
[1]*Universidad Nacional de Chimborazo*
Ecuador
[2]*North Carolina State University*
United States

Vicent Pla
Communications Department
Universitat Politècnica de València
Spain

Jorge Martinez-Bauset
Communications Department
Universitat Politècnica de València
Spain

Luis A. Marrone
LINTI
Universidad Nacional de La Plata
Argentina

Shih-Chun Lin
iWN Lab, Dept. of Electrical and Computer Engr.
North Carolina State University
United States

***Abstract*—Network Slicing (NS) is one of the technologies considered a pillar of 5G networks. It allows the division of the physical infrastructure of a network into several isolated logical networks (slices). The slices can have different sizes and be offered to other use cases. We analyze the radio resource allocation problem through a random access channel model considering the radio access network (RAN) with NS in a steady state. We perform an in-depth study of the random access procedure (RAP) to optimize resource allocation in a 5G RAN with NS. We focus on assigning subsets of preambles for each slice depending on the service's priority. The main contributions of our work are the following: i) A model for a scenario of n slices; that is, it has no limitation for the number of use cases. ii) An efficient RAP resource allocation policy to maximize the probability of successful access by UEs in each slice.**

***Index Terms*—5G cellular systems; network slicing; analytic model, RAN slicing; resource allocation.**

I. Introduction

Today's many connected devices allow massive and unrestricted access to information. However, most of these devices, called user equipment (UE), send data sparsely over time, using Internet of Things (IoT) applications. The best interconnection alternative for UEs is cellular networks due to their widely deployed infrastructure. However, cellular technology was conceived to handle human-to-human (H2H) traffic and not many UEs interacting simultaneously, as with machine-to-machine (M2M) communications. This results in many devices trying to connect to the base station of a cellular network with the corresponding congestion problems that this causes.

Fifth-generation (5G) networks emerge as an alternative to satisfy wireless network users' high service and connectivity requirements. With the implementation of 5G networks, data rates are expected to reach 10 Gbps [1]. It is also estimated that 5G will reach a total of 4.4 billion subscribed devices, which will represent 49% of all mobile subscriptions in 2027 [2]. Besides, the vision of 5G is to provide extremely low latency, higher capacity, and better QoS perceived by users [3].

Unlike 4G, which was conceived to provide mobile broadband communications, the 5G infrastructure is expected to enable the evolution of sectors such as industry 4.0, automotive, e-medicine, and entertainment, among others [4]. Although the vision and benefits of 5G are precise, enabling technologies are an open field of research. One of the technologies considered a pillar of 5G networks is Network Slicing (NS). NS allows the division of the physical infrastructure of a network into several isolated logical networks (slices). The slices can have different features and be offered to other use cases. In [5], a slice is defined as a combination of network functions (NF) and radio access technologies (RAT) for a specific use case. So, NS is allocating a dedicated or shared portion of the network resources for each slice [6].

Table I
Slice Types for Use Cases

Slice / Service type	SST value	Characteristics
eMBB	1	5G enhanced mobile broadband
URLLC	2	Ultra-reliable low latency communications
mIoT	3	Massive communications IoT
V2X	4	Vehicle to everything V2X services
HMTC	5	High-Performance Machine-Type Communications

In the ETSI Technical Specification 123 501 update [7], a slice is identified by a Single Network Slice Selection Assistance Information (S-NSSAI). An S-NSSAI comprises i) a Slice/Service type (SST), which identifies the expected service in the NS, and ii) a differentiator segment that allows distinguishing several NSs belonging to the same type of service. These standardized values in the update allow categorizing five use cases for NS, described in Table I.

In the following, we analyze the problem of radio resource allocation through a random access channel (RACH) model considering the RAN with NS in a steady state. For this, we focus on n traffic flows that, during the Random Access Procedure (RAP), use the uplink resources (preambles and uplink grants). For evaluation purposes, we compute the two Key Performance Indicators (KPIs) defined by the 3GPP [8]: the probability of successful access and the number of preamble transmissions per access attempt.

A. Random access procedure with NS

All UEs needing resources to access service must execute the RAP. It starts when the base station (gNB) offers a random access opportunity (RAO) to the UEs [6]. The RAP execution

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00258

uses two physical channels: the PRACH for the transmission of preambles and the PUSCH for the data [9]. A preamble is a specific identifier that UEs transmit to indicate their presence in the cell to the gNB. The preamble signals are orthogonal (i.e., the gNB can distinguish preambles sent simultaneously by multiple UEs). However, the number of preambles in a 5G New Radio (NR) cell is limited to 64. UEs randomly select one of these preambles to start their network access [10].

In the time domain, the access system is divided into slots. Each slot represents a RAO; it occurs periodically, and the *prach-ConfigIndex* parameter determines its periodicity [11]. We consider a subframe length of 1 ms and a RAO periodicity of 5 ms, corresponding to the setting *prach-ConfigIndex* = 6.

The RAP can be performed in two ways: i) contention-free or ii) contention-based. The former allocates reserved preambles during specific intervals, and for specific UEs (collision-free) [9]. In the latter, the UEs choose preambles randomly; two or more UEs in the same cell could choose the same preamble for the same RAO, causing a collision. A high number of collisions will cause a low probability of success and an increased access delay. The 3GPP standard suggests using 54 preambles for contention-based RAP [12].

Before an access attempt, the gNB shares network parameters with UEs through Master Information Block (MIB) and System Information Blocks (SIB) [11] messages. Among the parameters received through the SIB Type 2 is the periodicity in time of the RAOs [13].

1) Contention-based RAP: Its operation is based on executing the four-message handshake between the gNB and the UEs. A UE initiates its access attempt by sending *Msg1* to the gNB. *Msg1* contains a preamble randomly chosen by the UE from a set of preambles. Due to preamble orthogonality, several UEs can access the gNB in the same RAO using different preambles. However, if two or more UEs transmit the same preamble, the transmitted preamble cannot be decoded by the gNB, i.e., an *Msg1* transmission collision occurs [13]. If *Msg1* has sufficient transmission power, it will be decoded by the gNB [9], [13], [14]. If it is not decoded, the UE will make a new attempt by increasing the transmission power.

The gNB responds with an *Msg2* to each successfully decoded *Msg1*. The *Msg2* includes identification information for the detected preamble, and the granting of reserved resources (UL grant) for the *Msg3* transmission [9], [13]. The UEs that do not receive the *Msg3* within the W_{RAR} time window will raise their power and perform retransmission by randomly choosing a new preamble. All UEs that receive an UL grant through *Msg2* will be able to transmit *Msg3*. The transmission of *Msg3* is guaranteed through the hybrid automatic repeat request (HARQ) [9], [13].

The gNB transmits *Msg4* in response to *Msg3*. *Msg4* also uses a HARQ scheme. If the UE does not receive *Msg4* within the contention resolution time, the attempt is declared failed, and a new access attempt is planned, and the transmission power is increased. If a UE reaches the maximum number of re-transmissions, the network is declared unreachable, terminating the RA procedure [13]. UEs that complete the RA procedure receive a block of time-frequency resources for communication. All UEs that fail their transmission must execute a backoff procedure, regardless of the reason for the failure or the slice to which they belong. In this procedure, the UE waits for a random time $\mathcal{U}(0, BI)$ before starting a new preamble transmission. BI is the backoff indicator, defined by the gNB and sent to the UEs in *Msg2* [9], [15].

The rest of the paper is organized as follows. We conduct a literature review regarding NS in Section II. Then, we describe the system and analytical models in Section III and Section IV, respectively. Our most relevant results are presented in Section V, and finally, the conclusions are presented in Section VI.

II. Related Work

Most studies have focused on the management and orchestration of resources instead of how to allocate these resources in the 5G radio access network. The limited number of preambles and UL grants available in the RACH represents a resource allocation problem. Another significant issue is the coexistence of eMBB, mMTC, and URLLC services and applications in a 5G segment at the RAN level. While there is much research on performance evaluation of 5G downlink (DL) use cases, there are few results for UL [16].

In [17], an algorithm for optimizing the allocation of radio resources to the slices of the cell of a network that implements NS is proposed. Its performance is evaluated by simulation. The study compares different priority levels assigned to each slice. The priority of each slice is defined through four techniques: i) searches for the order that meets an objective function; ii) performs a random ordering; iii) performs an ordering to maximize the assigned resources; and iv) a prioritization based on the granularity of each slice. Three resource allocation methods that ensure isolation in a RAN with NS are presented in [18]. Notably, a proportional fairness algorithm limits the number of RBs assigned to each slice. The authors show through simulation that the isolation between slices is guaranteed. The results report an improvement in system performance for the three methods: static allocation, allocation to ordered slices, and impartial allocation to slices. In these investigations, the problem of allocating resources and allocating procedures in the RACH access is not considered.

An optimization approach for allocating radio resources in the 5G RAN that implements NS is addressed in [19]. Two types of generic 5G services are considered: eMBB and mMTC. Each service can receive dedicated and shared subsets of RAN and RACH resources. The proposed model analyzes the system's performance in terms of blocking probability for each slice without analyzing the access delay nor the number of retransmissions for successful access. Their model considers that collisions occur in *Msg3* of the RAP and evaluate an equal and proportional allocation of resources for two slices. An optimal resource segmentation alternative based on the number of slices in the system is not presented. This proposal does not achieve complete isolation between slices since segmentation of RAP uplink grants (UL grants) is not performed.

In [20], non-orthogonal random access (NORA) is proposed to reduce the problem of congestion in 5G networks. NORA is based on eliminating collisions caused by accesses from UEs that use the same preamble in *Msg1*. It does this by identifying the difference in arrival time of various UEs with identical preambles. The analysis carried out by simulation shows higher performance in terms of preamble collision probability and access success probability.

This paper considers an in-depth study of the RAP in a 5G network with NS to improve resource allocation. We focus on assigning subsets of preambles for each slice depending on the

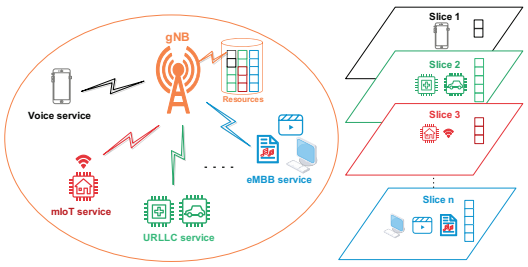

Figure 1. System model, 5G RAN with network slicing.

service's priority. The main contributions of our work are the following: i) a model for a scenario of n slices; that is, it has no limitation for the number of use cases, and ii) an efficient RAP resource allocation policy to maximize the probability of successful access by UEs in each slice.

III. System Model

We consider resource allocation at the RAP in a cell with S slices as illustrated in Fig. 1. Each slice serves users of a different service: eMBB, voice, mIoT, and URLLC. Each service is assigned a priority level. Each UE in a slice s must complete the *4-step RAP* to access the time-frequency resource blocks (RB) for data transfer. The RAP's physical resources (preambles) are allocated by the gNB to each slice, using a *resource allocation policy*.

We assume that arrivals are generated by a large population of independent users. Therefore, a Poisson process is appropriate to model the arrivals in each slice.

We consider that each of the slices is assigned a block of resources (preambles). Furthermore, we consider that preambles not assigned to any slice are shared and can be used by all slices. That is, we will have S slices and $S+1$ preamble blocks (number of slices plus the shared block). Finally, it is assumed that UL grants will not be reserved for any slice. In each RAO, all accesses correctly detected by the gNB and that have not collided will compete for the available UL grants regardless of the slice they come from.

A. Collision Model

There are two collision models when two or more UEs simultaneously transmit the same preamble [13]. First, the gNB cannot decode the preambles transmitted by multiple UEs, so all the collisions occur in the transmission of *Msg1*. Second, all *Msg1* are detected, and collisions occur in *Msg3*. In this work, we intend to study the behavior of the RACH in extreme operation scenarios; therefore, we assume that the collision detection is performed in *Msg1*. That is, only *Msg1*s that have been correctly decoded and have not collided will have the chance to receive a UL grant.

IV. Analytical Model

Unless otherwise stated or it is evident by the context, variables defined as "number of X" represent the average number of "X" per RAO.

Figure 2 illustrates how the resources are assigned for each slice. We are going to distinguish between slices and blocks

Table II
Notation used

Resources and system parameters	
Total number of preambles	L
Total number of UL grants	G
Number of preambles reserved in the block i	L_i
Total number of UL grants reserved in the block i	G_i
Maximum number of transmission attempts, *slice* s	k_s^m
Power ramping parameter, slice s	Δ_s
Traffic	
Number of new arrivals, slice s	a_s
Number of transmissions that are in the kth attempt, slice s	$a_s(k)$
Number of random access successfully completed, slice s	a_s^*
Number of transmissions in the block s	N_s
Average number for each random access, slice s	K_s
Probabilities	
Attempt k detection probability, slice s	$P_s^1(k)$
Probability of receiving a UL grant, slice s	P_s^2
Probability of no collision, slice s	P_s^{nc}
Probability of receiving a UL grant in the block s	p_s^2
Probability of no collision in the block s	p_s^{nc}
Successful probability of the kth attempt, slice k	$P_s(k)$
Successful probability	P_s

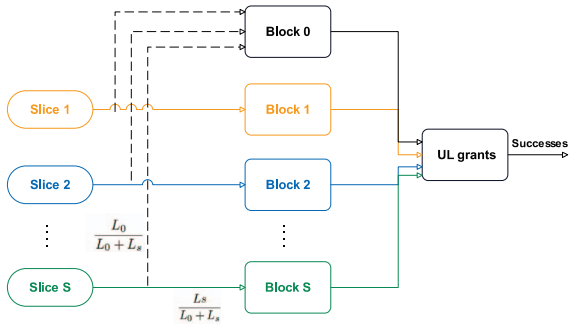

Figure 2. Slices and blocks of resources.

of resources. Each block is assigned many preambles. The distribution of access requests from each slice is proportional to the number of preambles assigned to each block. Thus, the fraction of accesses of slice s that use the shared block is given by

$$\frac{L_0}{L_0+L_s}, \tag{1}$$

whereas the fraction that used the reserved block is given by

$$\frac{L_s}{L_0+L_s}, \tag{2}$$

where L_s is the number of preambles reserved for each slice, and L_0 is the number of preambles reserved for block 0.

Let $a_s(k)$ be the number of transmissions of slice s that are in the kth attempt and k_s^m the maximum number of attempts. Taking into account the distribution between the shared common block and the reserved one, the average total number per RAO of preambles transmitted in each block s is obtained as

$$N_s = \frac{L_s}{L_s+L_0}\sum_{k=1}^{k_s^m} a_s(k), \qquad s=1,\ldots,S. \tag{3}$$

The average number of preambles that use the shared block per RAO is obtained by adding the contribution of each slice:

$$N_0 = \sum_{s=1}^{S} \frac{L_0}{L_s + L_0} \sum_{k=1}^{k_s^m} a_s(k) = \sum_{s=1}^{S} \frac{L_0}{L_s} N_s. \tag{4}$$

For *Msg1* to be successfully transmitted, three conditions must be met: i) *Msg1* is correctly detected by the gNB, ii) *Msg1* does not collide, and iii) detected and not collided preambles get a UL grant. Therefore, to determine the probability of successful accesses, we will calculate the probabilities of success in these situations.

Msg1 detection probability: the probability of success in detecting *Msg1* will depend on the number of previous *Msg1* transmissions in the same access attempt. This is due to the power ramping scheme. To insert an additional level of prioritization between slices, the factor Δ_s is introduced to the 3GPP specification. Therefore, the probability of detecting the kth transmission attempt of a slice s preamble will be given by

$$P_s^1(k) = 1 - e^{-k\Delta_s}. \tag{5}$$

If we multiply the detection probability corresponding to each block by the total number of preambles used, we obtain the total number of detected preambles:

$$N_s^1 = \frac{L_s}{L_s + L_0} \sum_{k=1}^{k_s^m} P_s^1(k) a_s(k), \qquad s = 1, \ldots, S \tag{6}$$

$$N_0^1 = \sum_{s=1} S \frac{L_0}{L_s + L_0} \sum_{k=1}^{k_s^m} P_s^1(k) a_s(k) = \sum_{s=1}^{S} \frac{L_0}{L_s} N_s^1. \tag{7}$$

Msg1 no collision probability: with the number of preambles of each block and the number of preambles detected by the gNB, we can calculate the *probability of no collision* of the transmitted preambles in the block s as

$$p_s^{\text{nc}} = \left(1 - \frac{1}{L_s}\right)^{N_s^1 - 1}. \tag{8}$$

Probability of getting a UL grant: the probability that a preamble transmitted in block s will get a UL grant (probability of success in *Msg2*) can be estimated as

$$p_s^2 = \min(1, \frac{G}{g_s}), \tag{9}$$

where G is the number of UL grants available, and g_s is the average number of UL grants needed for the block preambles s, which is calculated as the product of the number of detected preambles and the probability of not having a collision

$$g_s = N_s^1 p_s^{\text{nc}}. \tag{10}$$

From the probabilities of success of *Msg2* and of not colliding, using the proportion of attempts that go through block 0 and the proportion that goes through the block reserved for slice s, it is possible to obtain the probabilities of success in *Msg2* and no collision in slice s as follows:

$$P_s^2 = \frac{L_s p_s^2 + L_0 p_0^2}{L_s + L_0}, \qquad s = 1, \ldots, S \tag{11}$$

$$P_s^{\text{nc}} = \frac{L_s p_s^{\text{nc}} + L_0 p_0^{\text{nc}}}{L_s + L_0}, \qquad s = 1, \ldots, S. \tag{12}$$

From (5), (11), and (12), we get the success probability of the kth attempt in slice s:

$$P_s(k) = P_s^1(k) P_s^2 P_s^{\text{nc}}. \tag{13}$$

If the number of new arrivals (first attempt) in slice s is a_s, we have:

$$a_s(1) = a_s \tag{14}$$

$$a_s(k+1) = a_s(k)\big(1 - P_s(k)\big), \qquad k = 1, \ldots, k_s^m - 1. \tag{15}$$

To calculate the *throughput* (average number of successfully completed accesses per RAO) of slice s, we add the product of the number of transmissions a_s by the probability of success P_s of each attempt k

$$a_s^* = \sum_{k=1}^{k_s^m} a_s(k) P_s(k). \tag{16}$$

Finally, the probability of success is calculated as the ratio of successful transmissions to total transmissions:

$$P_s = \frac{a_s^*}{a_s}. \tag{17}$$

In addition, the average number of attempts (preamble transmissions) in slice s is calculated as the sum of the total number of transmissions per RAO divided by the number of new transmissions per RAO:

$$K_s = \frac{1}{a_s} \sum_{k=1}^{k_s^m} a_s(k). \tag{18}$$

V. Results

A. Model validation

The results of the analytical model have been validated with results obtained through computer simulation using MATLAB. For each numerical experiment, we set a basic load vector $a^0 = [a_1, \ldots, a_s]$, which establishes the load share of each slice, and then the total load is scaled by a factor f ranging from 0.2 to 2, while the load share of each slice is kept constant $a = f\, a^0 = f\,[a_1, \ldots, a_s]$. In the following, we detail the results according to the network configuration described in Table III.

Fig. 3 compares the results of the analytical model and the simulation. The horizontal axis represents each slice's initial load variation factor f. The initial load of each slice is the average number of RACH accesses per RAO. The results show a good match between the model and the simulation. The results for a low initial load are shown in Fig. 3. Fig. 4 depicts the results when we vary the initial load in one of the two slices. It is observed that the drop in performance of slice 2 is due to the increase in a load of access requests of slice 2. This is because there is no total isolation by having an assignment different from 0 in the subset of shared resources.

Table III
GENERAL RACH SLICING CONFIGURATION

Parameter	Setting
PRACH Configuration Index	6
Subframe length	1 ms
Total number of preambles	54
Maximum number of preamble transmissions	preambleTransMax = 10
RAR window size	$W_{RAR} = 5$
mac-ContentionResolutionTimer	48 sub-frames
Maximum number de UL grants per subframe	$N_{RAR} = 3$
Backoff Indicator	$BI = 20$ ms
HARQ re-transmission probability for *Msg3* and *Msg4* (non-adaptive HARQ)	10%
Maximum number of HARQ TX for *Msg3* and *Msg4* (non-adaptive HARQ)	5
Periodicity of RAOs	5 ms
Preamble transmission time	1 ms

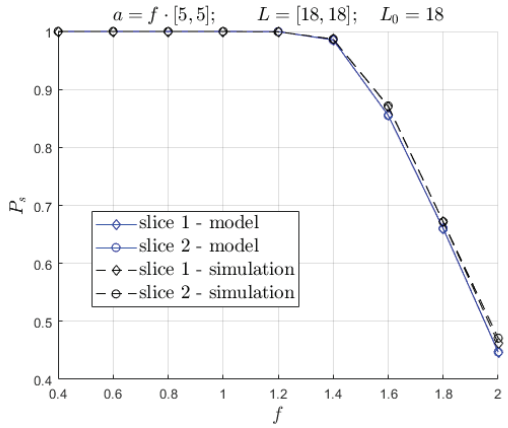

Figure 3. Successful access probability as a function of traffic load. Equitable allocation of resources for two slices in the RAN.

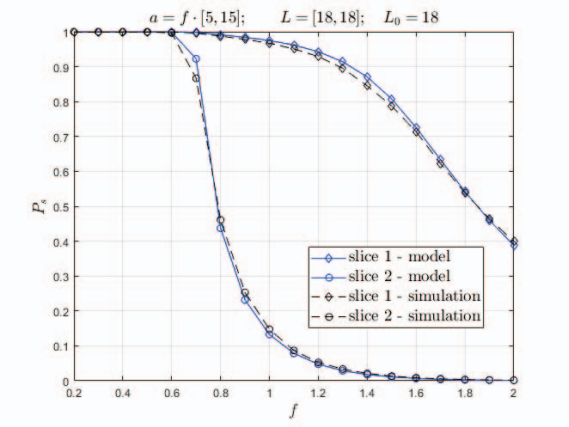

Figure 4. Successful access probability as a function of different traffic load per slice. Equitable allocation of resources for two slices in the RAN.

B. Equal sharing of resources

We analyze the behavior of the analytical model when an equal assignment of preambles is made for 2 slices.

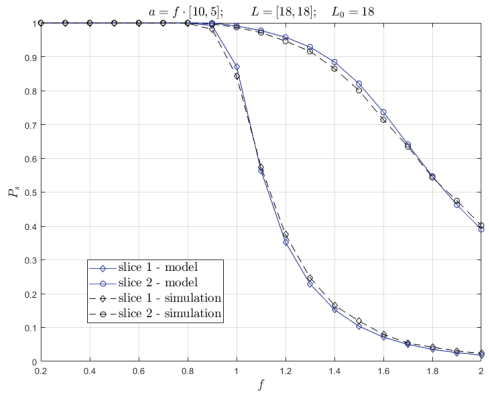

Figure 5. Successful access probability as a function of different traffic load per slice. Equitable allocation of resources for two slices in the RAN.

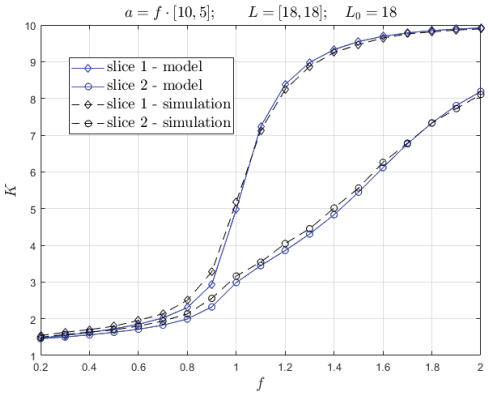

Figure 6. Average number of preamble transmissions as a function of different traffic load per slice. Equitable allocation of resources for two slices in the RAN.

We consider that the gNB will reserve several preambles for each block of resources equitably; this number is computed as

$$r_0 = \left\lceil \frac{R}{S+1} \right\rceil . \tag{19}$$

In Fig. 5, for an initial load $a_s = [10, 5]$, and resource blocks with an allocation $L = [18, 18]$ and $L_0 = 18$, a probability of successful access greater than 90% is obtained up to around $f = 1$ for slice 1 and $f = 1.4$ for slice 2, which represents an average of 10 and 7 access requests per RAO, respectively. Beyond this value, the RACH begins a drop in performance. As far as K is concerned, we can see in Fig. 6 that the number of retransmissions starts to increase significantly when $f > 0.8$ for slices 1 and 2. Reviewing these results, we can say that when $K > 3$, the performance of the RACH starts to drop.

C. Resource allocation proportional to load

To determine the percentage of available preambles allocated to the shared block, we use the coefficient δ as

$$L_0 = \lceil \delta \, R \rceil . \tag{20}$$

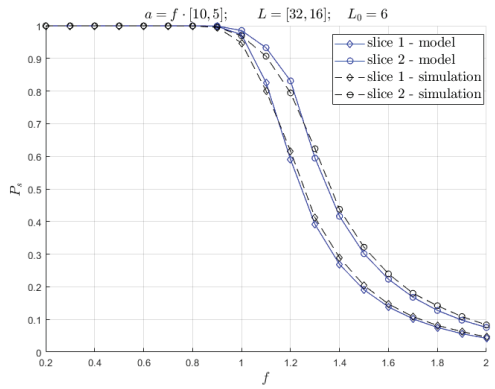

Figure 7. Successful access probability as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 2$.

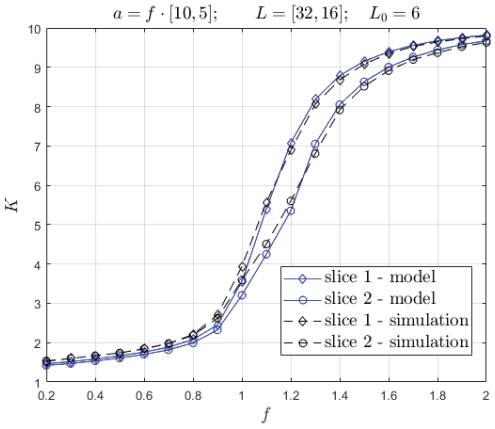

Figure 8. Average number of preamble transmissions as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 2$.

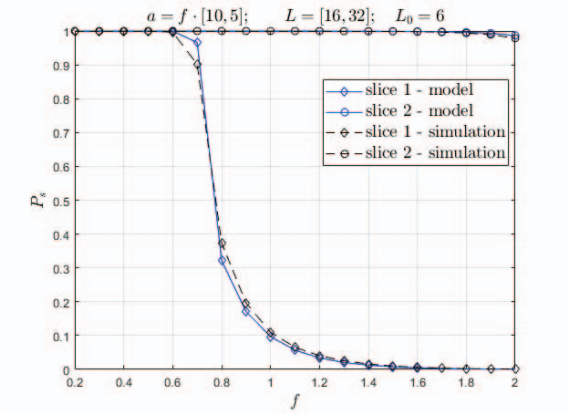

Figure 9. Successful access probability as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 0.5$.

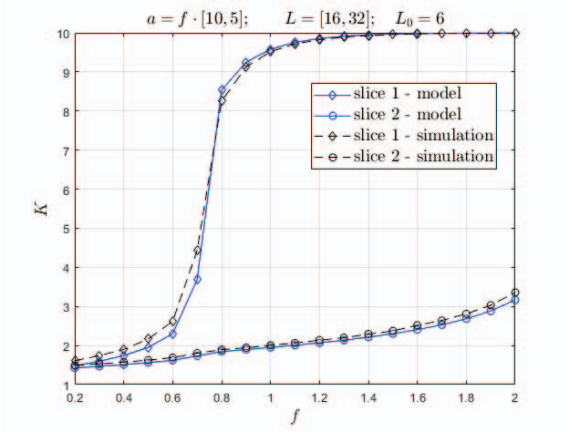

Figure 10. Average number of preamble transmissions as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 0.5$.

We calculate the number of preambles reserved for each slice from the remaining preambles. To do this, we use the proportion factor β as follows

$$L_1 = \beta L_2. \tag{21}$$

To avoid exceeding the number of available preambles, the maximum value that L_2 can take is 18 (when $\beta = 2$).

For the same initial load as in Fig. 5, in Fig. 7, we can observe the probability of success in the accesses for a scenario of 2 slices in which we set $\delta = 0.10$ and $\beta = 2$. A $P_s \geq 90\%$ is obtained up to around $f = 1.1$ for the 2 slices, averaging 11 and 5.5 access requests per RAO, respectively. Furthermore, we can see that slice 1 has a more pronounced drop in performance from this point on compared to slice 2. In Fig. 8, we can see again that at the inflection points of the curve, the mean number of retransmissions is approximately 3.

Figs. 9 and 10 illustrate the results when we set $\beta = 0.5$ and keep the remaining parameters the same. We can observe that the performance of slice 1 falls drastically due to the decrease in reserved resources, while the opposite occurs for slice 2.

As observed in Figs. 7 to 10, the differences between the model and the simulator for both P_s and K are almost indistinguishable. However, slight differences can be seen in certain areas due to multiple reserved blocks going into saturation simultaneously and competing with a higher load for shared resources. After this, the curves even overlap.

D. Increasing the number of slices

Finally, we will evaluate the analytical model when the number of slices exceeds 2. We assume a scenario with 5 slices with equitable allocation of resources serving services with different loads.

This scenario can represent a 5G network with 5 slices, each dedicated for each use case of Table I. As shown in Fig. 11, the performance drop in the P_s is related to the load of the slice. The higher the load, the faster the performance degrades. The same can be seen in Fig. 12, where those slices with a higher load carry out more retransmissions.

The results presented in Figs. 11 and 12 show that our model accurately represents the behavior of a 5G RAN with n slices.

VI. Conclusions

We have described an analytical model for the RAP of a 5G network implementing NS in detail. Our model can

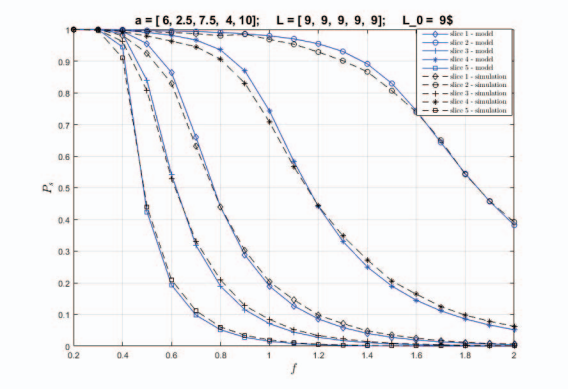

Figure 11. Successful access probability as a function of different traffic load per slice. Equitable allocation of resources for five slices in the RAN.

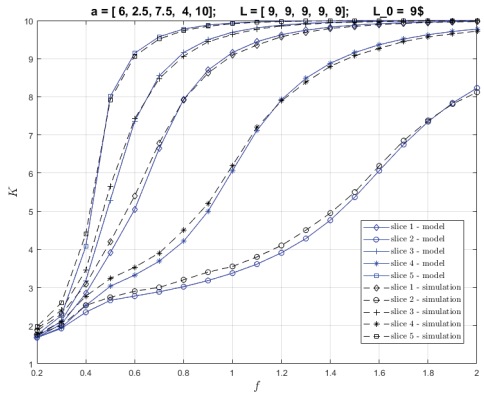

Figure 12. Average number of preamble transmissions as a function of different traffic load per slice. Equitable allocation of resources for five slices in the RAN.

be used to efficiently evaluate the performance of different resource allocation techniques to the 5G RAN slices. Furthermore, through evaluating performance indicators such as the probability of success in access and the number of necessary retransmissions, we have been able to analyze an equitable allocation of resources proportional to a load of each slice, the results of which have allowed us to validate our model. When performing segmentation of preambles and not of UL grants, we have observed partial isolation between slices. Resource isolation is one of the main applications of NS. It has been shown that the probability of success degrades significantly when the number of retransmissions is above three. Beyond this point, the RACH is severely congested. In future work, we plan to extend the model so that the case in which UL grants are reserved for each slice can be studied. This way, different network providers or tenants could use each slice virtually.

ACKNOWLEDGMENT

This work was supported in part by Cisco Systems, Inc., the NC State 2022 Faculty Research and Professional Development Program (FRPD), the NC Space Grant, the AC21 Special Project Fund (SPF), and the National Science Foundation (NSF) under Grant CNS-2210344. The work of J. Martinez-Bauset and V. Pla was supported by Grants PGC2018-094151-B-I00 and PID2021-123168NB-I00 funded by MCIN/AEI/10.13039/501100011033 and ERDF *A way of making Europe*.

REFERENCES

[1] J. Zhao, J. Liu, L. Yang, B. Ai, and S. Ni, "Future 5G-oriented system for urban rail transit: Opportunities and challenges," *China Communications*, vol. 18, no. 2, pp. 1–12, 2021.

[2] Ericsson, "Ericsson mobility report," https://www.ericsson.com/4ae6a5/assets/local/reports-papers/mobility-report/documents/2021/emr_november2021_screen_epsanol.pdf, 2021.

[3] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[4] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[5] N. Alliance, "5G white paper," *Next generation mobile networks, white paper*, vol. 1, no. 2015, 2015.

[6] V. Mancuso, P. Castagno, M. Sereno, and M. A. Marsan, "Modeling MTC and HTC radio access in a sliced 5G base station," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 2208–2225, 2020.

[7] Etsi.org, "5G System architecture for the 5G System (5GS) (3GPP TS 23.501 version 17.4.0 Release 17)," https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144, 2022.

[8] 3GPP, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Apr 2014.

[9] L. Tello-Oquendo, J.-R. Vidal, V. Pla, and L. Guijarro, "Dynamic access class barring parameter tuning in LTE-A networks with massive M2M traffic," in *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, 2018, pp. 1–8.

[10] S. Vural, N. Wang, P. Bucknell, G. Foster, R. Tafazolli, and J. Muller, "Dynamic preamble subset allocation for RAN slicing in 5G networks," *IEEE Access*, vol. 6, pp. 13 015–13 032, 2018.

[11] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7785–7799, 2017.

[12] J. Liu, M. Agiwal, M. Qu, and H. Jin, "Online Control of Preamble Groups With Priority in Massive IoT Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 700–713, 2021.

[13] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2017.

[14] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Communications Letters*, vol. 20, no. 2, pp. 372–375, 2015.

[15] J.-R. Vidal, L. Tello-Oquendo, V. Pla, and L. Guijarro, "Performance study and enhancement of access barring for massive machine-type communications," *IEEE Access*, vol. 7, pp. 63 745–63 759, 2019.

[16] V. Mancuso, P. Castagno, M. Sereno, and M. A. Marsan, "Serving HTC and Critical MTC in a RAN Slice," in *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 2021, pp. 189–198.

[17] C.-Y. Chang, N. Nikaein, and T. Spyropoulos, "Radio access network resource slicing for flexible service execution," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 668–673.

[18] D. Nojima, Y. Katsumata, T. Shimojo, Y. Morihiro, T. Asai, A. Yamada, and S. Iwashina, "Resource isolation in RAN part while utilizing ordinary scheduling algorithm for network slicing," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–5.

[19] O. Vikhrova, C. Suraci, A. Tropeano, S. Pizzi, K. Samouylov, and G. Araniti, "Enhanced radio access procedure in sliced 5G networks," in *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE, 2019, pp. 1–6.

[20] T. Nomu and R. Aravind, "Non-Orthogonal Random Access scheme in Spatial Group Based Random Access for 5G Networks," *International Journal of Innovative Research in Technology*, vol. 6, 2019.