On the resource allocation for radio access network slicing in cellular IoT with massive traffic

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*—The limited capacity of the random access channel (RACH) represents a challenge for adequate resource allocation in 5G radio access networks with network slicing. Furthermore, a fair division of scarce radio resources is required to simultaneously support many users with heterogeneous service requirements. In this work, we look at the problem of uplink radio resource allocation to slices on the radio interface of one cell in a non-stationary regime with mMTC, eMBB, and H2H traffic. We analyze four resource allocation policies for efficient random access to improve each slice's capacity in terms of successful access probability, the number of preamble transmissions, and access delay. Besides the number of available preambles in the RACH, we also consider the limitation of uplink grants in the radio access network.**

***Index Terms*—cellular systems; machine-type communications; RAN slicing; resource allocation; performance analysis.**

I. Introduction

Unrestricted access to information and services will soon be possible because of a vast number of linked gadgets. Most of these devices, collectively referred to as user equipments (UEs), send data sparsely over time using Internet of Things (IoT) applications. Cellular networks are the greatest option for UE interconnection because of their well-developed infrastructure.

In addition to building on the success of the 4G cellular network, the fifth-generation (5G) wireless technology is anticipated to enable a wide range of network services with various performance needs. One of the foundational technologies for 5G is the Network Slicing (NS) paradigm [1]. It can be viewed as a specially designed logical network made up of virtualized and dedicated resources used to meet the needs of a specific service [2]. It allows serving users from various verticals on the same physical infrastructure. Heterogeneous traffic types, their combined requirements and interactions, and NS in the Radio Access Network (RAN) are being studied from several angles [3], [4]. One of the most important issues to address is resource allocation, and as a result, several proposals are emerging.

Three macro classes have been established to categorize 5G services with distinct traffic patterns and needs: i) *enhanced mobile broadband* (eMBB), which comprises traffic mostly produced by multimedia services; it was common in previous generations, ii) *ultra-reliable and low-latency communications* (URLLC) that must adhere to strict latency and reliability standards, and iii) *massive machine-type communications* (mMTC or mIoT, indistinctly) the most capacity-intensive type of communication.

In this paper, we look at the problem of uplink (UL) radio resource allocation to slices on the radio interface of one cell in a non-stationary scenario of transient mMTC initial access. For this, we focus mainly on the coexistence of H2H, mMTC, and eMBB slices that use two uplink resources, namely preambles and UL grants, during the random access (RA) procedure. Regarding the URLLC slice, we assume it can use only dedicated resources (preambles) that are pre-allocated and fixed in time due to the stringent requirements of such applications. For the evaluation, we obtain the key performance indicators (KPI) defined by the 3GPP [5], namely, access success probability, number of preamble transmissions per access attempt, and access delay.

The rest of the paper is organized as follows. We review studies analyzing NS in Section II. Then, we describe the system model, RAN slicing policies, and the network configuration parameters used in this study in Section III, Section IV, and Section V, respectively. Our most relevant results are presented in Section VI, and finally, we present our conclusions in Section VII.

II. Related Work

Although several papers have focused on resource management and orchestration in 5G networks implementing NS, only a few have addressed resource allocation strategies at the RAN level, particularly in the random access channel (RACH). A significant problem is the coexistence of eMBB, mMTC, and URLLC services and applications in a 5G slice at the RAN

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00113

level. While there are already several pieces of research on the performance evaluation of 5G downlink (DL) use cases, there are few results on the UL [6].

In the RAN, the slicing is usually performed using orthogonal resource allocation. In [7], the performance of non-orthogonal slicing of RAN resources in the UL is investigated. The resources are shared by a set of service devices: eMMB, URLLC, and mMTC, with different reliability requirements. The RA procedure for resource allocation is not considered in the study. In an infrastructure with equivalent QoS requirements and different slicing configurations, it is concluded that, with non-orthogonal slicing, the UL presents a higher degradation than the DL in the RA process [8].

In [9] the authors propose prioritizing access to RACH through a segmentation of the preambles available in the system. It consists of a fixed separation of the preambles available for the RA procedure. For this, the preambles are divided into subsets. For example, the authors in [10] propose dividing the preambles into subsets to serve the HTC and MTC services in LTE. In these studies, the preamble allocation remains static regardless of the system load.

In [11], a preamble allocation model is presented based on the estimation of the system load and the priority given to each service class. Three classes of service, URLLC, eMBB, and mMTC, are considered. The load is estimated before each random access opportunity (RAO). Based on the arrival load estimate, the number of preambles allocated for each device class is updated before each RAO.

In [12], the RACH resource allocation in a 5G network implementing NS is studied. 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 performance in terms of blocking probability for each slice. It also compares an equal and proportional allocation of resources. An allocation of dedicated and shared preambles is performed. The evaluation is performed only for a network with two slices and includes neither the RA procedure nor the segmentation of UL grants.

The limited capacity of the RACH represents a challenge for adequate resource allocation. Furthermore, a fair division of scarce radio resources is required to simultaneously support many users with heterogeneous service requirements. This work seeks an efficient RA resource allocation policy considering preambles and UL grants.

III. System Model

A RAN with a set of $\mathcal{S} = \{1 \ldots S\}$ slices is considered. We concentrate on a cell-level resource allocation issue and study the allocation of UL resources used in the RA procedure. UEs are fully informed of the slice to which they belong. The base station (gNB in 5G) broadcasts system information about the access process and slice configuration. A slice policy (described in Section IV) determines how radio resources are distributed.

The RA can operate in two modes: contention-free and contention-based. The former is used for critical situations such as handover or positioning. The latter is the standard mode for network access; it is used by UEs to change the RRC state from idle to connected, to recover from radio link failure, to perform UL synchronization, or to send scheduling requests [13].

Random access attempts are allowed in predefined time/frequency resources, called RAOs. The gNB broadcasts the periodicity of the RAOs using a variable referred to as the PRACH Configuration Index. The periodicity varies between a minimum of 1 RAO every two frames (i.e., 1 RAO every 20 ms) and a maximum of 1 RAO per 1 sub-frame (i.e., every 1 ms) [14].

The physical RACH (PRACH) signals a connection request when a UE needs to access the RAN. It carries a preamble for initial access to the network. Up to $R = 64$ orthogonal preambles are available to the UEs per cell [14]. In contention-free mode, there is a coordinated assignment of preambles, so collision is avoided, but gNBs can only assign these preambles during specific slots to specific UEs. Hence, UEs can only use these preambles if assigned by the gNB and during specific slots. In the contention-based mode, preambles are selected randomly by the UEs, so there is a risk of collision; that is, there is a probability that multiple UEs in the cell pick the same preamble; therefore, contention resolution is needed. In the sequel, we focus on the contention-based random access mode.

A. Contention-based Random Access Procedure

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 [15]. If *Msg1* has sufficient transmission power, it will be decoded by the gNB [15]–[17]. 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 [15], [17]. 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) [15], [17].

The gNB transmits *Msg4* in response to *Msg3*. *Msg4* also uses the HARQ process. If the UE does not receive *Msg4* within the contention resolution time, the connection is declared failed, and a new access attempt is planned by increasing the transmission power. If a UE reaches the limit of unsuccessful re-transmissions, the network is declared unreachable, terminating the RA procedure [15]. 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 $\mathcal{U}(0, BI)$ ms before starting a new preamble transmission in a new RAO. BI is the backoff indicator, defined by the gNB and sent to the UEs in the *Msg2* [17], [18].

IV. RAN Slicing Policies

5G networks implementing NS require defining the allocation of RAN resources among the different slices. We analyze the allocation of the preamble and UL grants between the gNB and UEs statically and adaptively. In both cases, we consider i) a full isolation level between slices (Fully-sliced) in which preambles and UL grants are reserved for each slice; and ii) a medium isolation level between slices (Partially-sliced) in which the UL grants are not reserved but shared by all slices.

1) Fully-sliced Static Policy: Since the number of preambles assigned to each slice has a high impact on the probability of collision [19], in this proposal, we assume a fixed allocation in which the number of allocated preambles and UL grants are proportional to each other. This number is determined by the priority of the service using the slice. A cell with S slices is considered; the gNB performs a fixed allocation of subsets of different preambles and UL grants to each slice. Doing so allows additional QoS requirements to be handled with isolation between slices.

We consider three services: mIoT, eMBB, and H2H. Each service accesses a slice with different priorities (high, medium, low). For example, the mIoT service serves a hefty load of access requests from applications with machine-type devices, requiring a high-priority slice. On the other hand, eMBB requires a medium priority slice to serve a moderate number of access requests with high bandwidth requirements [9]. Finally, H2H traffic in which few accesses (compared to expected mIoT [15]) requires a low-priority slice.

To calculate the number of preambles assigned to each slice, we define a weight $\{w_i | \sum_{i=1}^{S} w_i = 1\}$ for high, medium, and low priority slices, respectively. Thus, the slice s (mIoT, eMBB, or H2H) receives a percentage of the total number of preambles available in the system calculated as

$$r_i = \begin{cases} \lceil R * w_i \rceil, & i = 1, \ldots, S-1 \\ R - \sum_{j=1}^{S-1} r_j, & i = S. \end{cases} \tag{1}$$

In addition, to ensure the isolation of each slice, an allocation of the available UL grants θ is performed by

$$g_i = \begin{cases} \lceil \theta * w_i \rceil, & i = 1, \ldots, S-1 \\ \theta - \sum_{j=1}^{S-1} g_j, & i = S. \end{cases} \tag{2}$$

2) Fully-sliced Adaptive Policy: The probability of successful access to a slice depends on the number of devices accessing and competing for system resources. Therefore, a static preamble allocation policy based on priorities alone will not be efficient. Ideally, it should be combined with the number of active requests in the RACH at each RAO [20]. Unfortunately, the number of active requests in the RACH is time-varying, composed of requests for new accesses and those requests that collided and are attempting again. Therefore, we need an algorithm that considers the number of active devices at each RAO to assign preambles to each slice.

We consider a slice with dedicated preambles for each type of traffic mIoT, eMBB, and H2H. In addition, we reserve a set of preambles shared by traffic flows of the dedicated slices that pass the conditions explained below. As indicated in Eq. (3), out of a total of R preambles available in the system, r_i preambles are reserved for the ith slice, and all slices share r_s preambles.

$$R = r_s + \sum_{i=1}^{S} r_i. \tag{3}$$

A higher number of collisions occur when a slice does not have enough preambles allocated. In addition, the gNB has a limited number of UL grants θ to respond to successfully detected preambles. Therefore, when the number of preambles detected by the gNB in a RAO is greater than θ there will be preambles that do not receive a UL grant. UEs that do not receive a UL grant should perform a new access attempt [19].

We then propose using the r_s subset as an alternative way to serve accesses with a high probability of failure if they use the preambles dedicated to their slice. This way, we mainly prevent these accesses from causing a collision and affecting other UEs. Access attempts using r_s contend for preambles other than those assigned to their slice. In this work, since we are considering collision detection in *Msg1*, having UL grants reserved for the shared preambles is unnecessary. Only detected and non-collided accesses using the r_s subset will require UL grants reserved to their slice.

To determine the percentage of shared preambles and UL grants assigned to each slice, we use the coefficient δ in Eq. (4). In high-traffic scenarios, the higher the level of sharing, the higher the collision probability is [12], [20].

$$r_s = \lceil \delta * R \rceil. \tag{4}$$

We calculate the subset of preambles assigned to each slice from the remaining preambles. The initial configuration of the proposal considers that the gNB will reserve some dedicated preambles for each slice equally; this number is calculated as

$$r_0 = \frac{R - r_s}{S} = (1 - \delta) R / S. \tag{5}$$

The number of preambles and UL grants assigned to each slice will be dynamically updated by the gNB using the *SIB2* message, which allows the gNB to transmit the configuration parameters to the UEs with a periodicity of $80\,\text{ms} = 16\,\text{RAOs}$ [17]. The number of preambles and UL grants assigned in each period is calculated based on the number of active devices in the ith slice N_i per RAO and is obtained as follows

$$r_i = \frac{\overline{N_i}}{-\ln(w_i) - \ln(x)}, \tag{6}$$

where w_i is the weight assigned to the ith slice, $\overline{N_i}$ is the average number of active devices in the ith slice per *SIB2* update period, and x represents a proportionality factor in ensuring that the available resources (preambles or UL grants) of the RACH are not exceeded; it is tuned to satisfy

$$R = r_s + \sum_{i=1}^{S} \frac{\overline{N_i}}{-\ln(w_i) - \ln(x)}. \tag{7}$$

The number of active devices for the ith slice N_i that access the RACH and wait for the preamble assignment varies in each RAO. Moreover, the gNB has no way of knowing this information; this value is estimated using the process reported in our previous works [17], [18], [21].

Bearing in mind that the maximum number of successful attempts is obtained when the number of contending UEs at the ith slice is approximately the number of preambles assigned to that slice (i.e., $r_s \approx N_i$) [15], we define thresholds for each service traffic. Requests from active nodes that exceed the corresponding threshold will use the r_s subset of preambles. In this way, we ensure each slice's maximum capacity, avoiding excess of collisions and retransmissions. Requests using r_s will attempt to complete the RA procedure with a lower successful access probability. Those UEs attempts that do not collide in the transmission of *Msg1* and are correctly detected by the gNB will wait for a UL grant to finish the procedure successfully. In contrast, the UEs attempts that used the r_s and failed will be able to make their next attempt once the backoff time has elapsed.

3) Partially-sliced scheme for Static and Adaptive Policies: We also analyzed a variation to the fully-sliced scheme in both Static and Adaptive policies where UL grants are not reserved for each slice. Instead, the UL grants are shared and available to access attempts that complete the *Msg1* and are correctly detected by the gNB regardless of the subset of slice preambles they used. It is evident that the access attempts will constantly utilize all UL grants in high traffic. A disadvantage of this variation is the partial loss of isolation using slice resources. That is, this scheme isolates preambles but not UL grants.

V. Network Configuration Parameters

A discrete-event simulator of the 5G RAN with NS has been developed in C++ to evaluate the proposals. Additionally, these results were corroborated with MATLAB simulations independently. The system accommodates three types of traffic in each simulation: mIoT, eMBB, and H2H, with different access request intensities. The distribution and parameters used by each traffic model are described in Table I. The contention-based RA procedure described in Section III-A is replicated with the parameters listed in Table II. Simulations were run j times until the difference of computing the corresponding metric in the j th simulation run differs from the one computed in the $j-1$ th simulation run by less than 1%, considering a minimum value for j such as 10^3. The simulator provides the flexibility of choosing the parameters of interest, including the type of traffic, number of devices, timing, processing and channel parameters such as the number of available preambles, number of slices, priorities, and backoff window size.

Table I
Traffic Models for 5G NS RACH Evaluation

Characteristics	Traffic Model mMTC	Traffic Model eMBB	Traffic Model H2H
Arrival distribution	Beta(3,4) over T	Poisson(5) over T	Uniform over T
Number of devices	$2500, \dots, 30000$	1000	33000
Distribution period (T)	10 seconds	10 seconds	60 seconds

Table II
General RACH slicing configuration

Parameter	Setting
Number of slices	3
PRACH Configuration Index	6
RA Periodicity (RAO)	5 ms
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
Preamble detection probability for kth preamble transmission	$Pd = 1 - \frac{1}{e^k}$
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

A. Performance Metrics

The three KPIs for the purpose of RACH capacity evaluation with each slicing policy are the following [5]:

1) Access success probability P_s is the probability of successfully completing the random access procedure within the maximum number of preamble transmissions.
2) Statistics of the number of preamble transmissions per access attempt K.
3) Statistics of access delay D defined as the time elapsed between the arrival of a UE and the successful completion of its RA procedure.

B. Static-sliced Policies

We define the vector $w = [0.64, 0.32, 0.04]$ for the high, medium, and low priority slices, respectively. We find the number of preambles assigned to each slice r_i using Eq. (1). It remains constant throughout the simulation and is reserved for use by the UEs of each slice. With the same logic, we use Eq. (2) for reserving the UL grants of each slice g_i.

C. Adaptive-sliced Policies

To evaluate these policies, prior to the start of the RA procedure, we find the number of shared preambles r_s. For this, we assume a $\delta = 10$ in Eq. (4) since it is the factor that maximizes performance in a high-traffic scenario, as observed

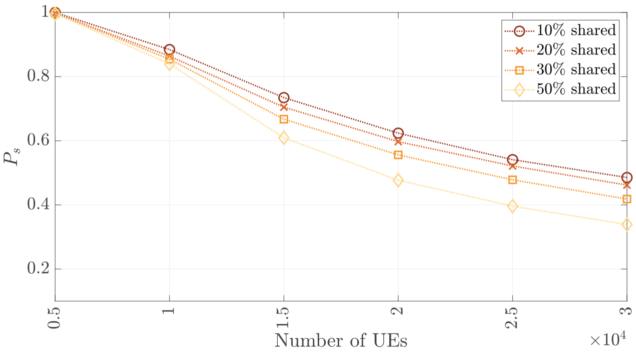

Figure 1. Succesfull access probability of mIoT traffic for each δ

in Fig. 1. The remaining preambles will be assigned using Eq. (5) to each slice dedicated to mIoT, eMBB, and H2H services.

In the following RAOs, the allocation of preambles and UL grants to each dedicated slice will be performed dynamically using Eqs. (6) and (7) each *SIB2* period. In addition, we define a priority vector $w = [0.57, 0.29, 0.14]$ for mIoT, eMBB, and H2H services.

VI. Results

In the following, we detail the results for each service according to the traffic models detailed in Table I and the network configuration described in Table II. The eMBB and H2H services are evaluated when mIoT traffic varies from light (2500 access requests) to heavy load (30000 access requests). We consider each scenario's eMBB service with medium-load (1000 access requests) and H2H service as background traffic. For the sake of comparison, we also evaluate a scenario without implementing network slicing, called *Unsliced*.

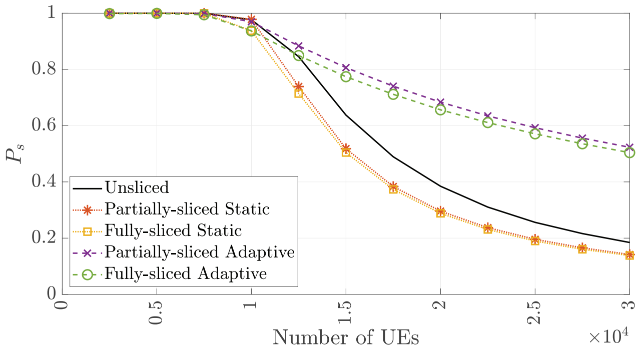

Figure 2. mIoT slice. Successful access probability P_s

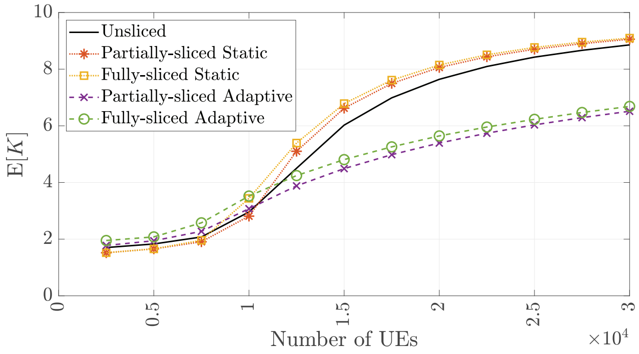

Figure 3. mIoT slice. Average number of preamble transmissions required for successful access $\mathbb{E}[K]$

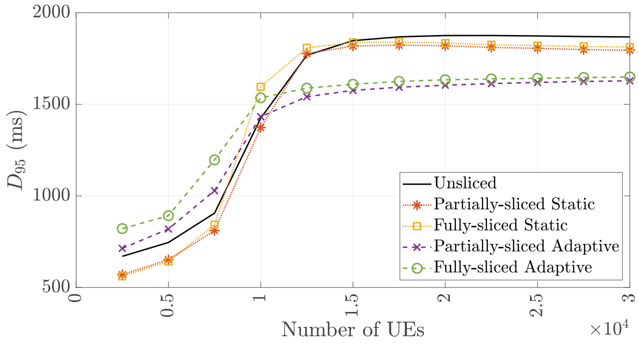

Figure 4. mIoT slice. 95th percentile of access delay D_{95}

A. mIoT service

Fig. 2 illustrates the P_s as a function of the number of mIoT UEs. As expected, P_s decreases as the number of mIoT UEs increases. The Adaptive-sliced policies maintain a higher value of P_s than the Unsliced and Static-sliced configurations. For light load scenarios (i.e., less than 10000 UEs), all slice policies present a high P_s value; it is evident that as the number of UEs competing for access in the RA procedure increases, the P_s drops drastically. Moreover, it is observed that the fully-sliced adaptive performance is very close to that of the partially-sliced adaptive, where the UL grants are not reserved but available for any service. The advantage of these policies is that the isolation level is improved (i.e., any flow changes in a slice can affect the performance of the remaining slices in a lesser way) since resource allocations are made dynamically with the evolution of active accesses. Fig. 3 depicts the average number of preamble transmissions required for successful access. Unsliced and static-sliced policies require a higher K than the adaptive-sliced ones. Finally, Fig. 4 illustrates the 95th percentile of the access delay D_{95}. We observe that it increases with the number of UEs in all cases. The Adaptive-sliced policies achieve a smaller D_{95} in heavy load conditions.

B. eMBB service

Fig. 5 illustrates the behavior of P_s. The Partially-sliced static policy performs better for light and heavy traffic conditions than other policies. From 10000 UEs onwards, both static-sliced policies provide higher P_s. Concerning K, Adaptive-sliced and Unsliced policies perform similarly in light load conditions as observed in Fig. 6. The Static-sliced policies perform uniformly for all network load conditions; in particular, the partially-sliced static requires fewer preamble

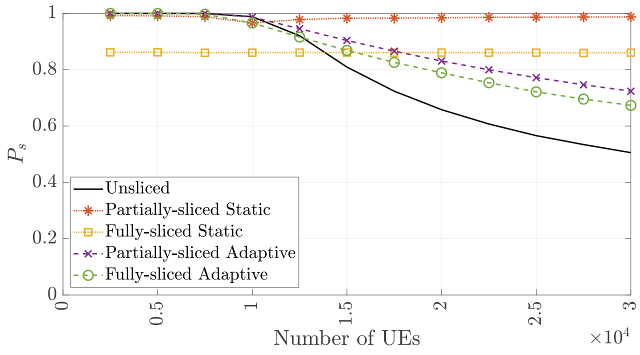

Figure 5. eMBB slice. Successful access probability P_s

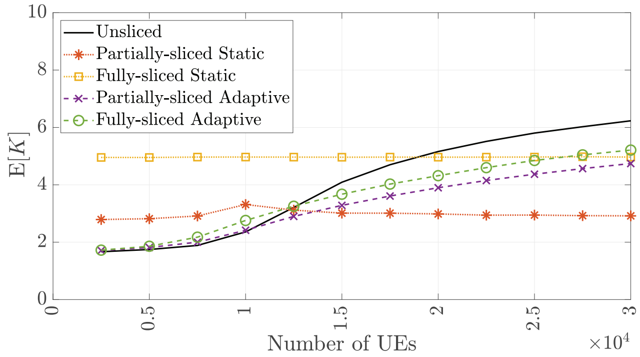

Figure 6. eMBB slice. Average number of preamble transmissions required for successful access $\mathbb{E}[K]$

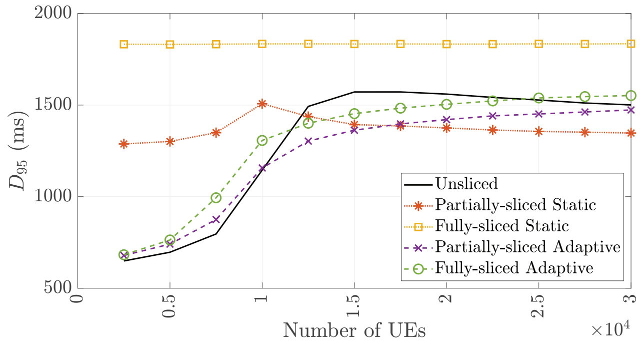

Figure 7. eMBB slice. 95th percentile of access delay D_{95}

transmissions in heavy load conditions for successful access. Regarding D, smaller values for light loads (less than 10000 UEs) can be obtained with the adaptive-sliced policies, as observed in Fig. 7. In heavy load scenarios, all policies present a similar behavior, particularly the partially-sliced static policy shows a smaller D.

C. H2H service

Figs. 8, 9, and 10 present the results of the evaluation of the H2H service and the effect it suffers with a variation of the number of mIoT UEs. Fig. 8 indicates that the best performance is obtained with the Static-sliced policies. Both

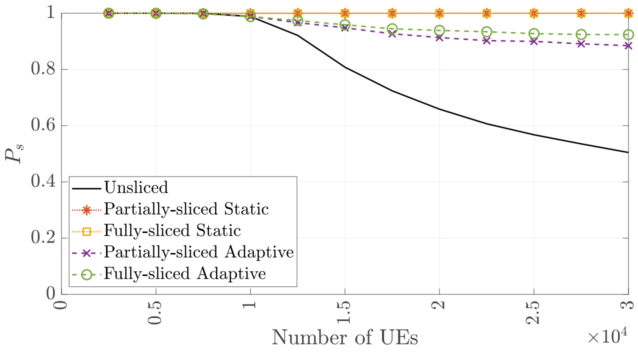

Figure 8. H2H slice. Successful access probability P_s

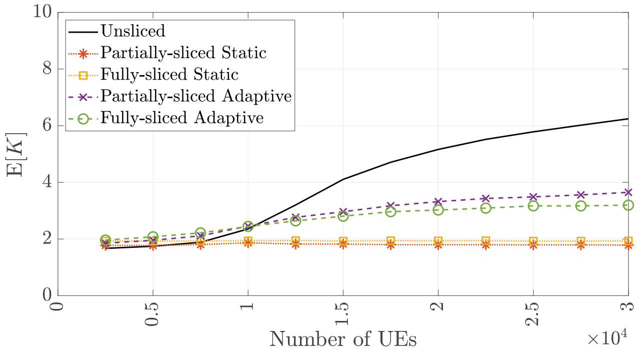

Figure 9. H2H slice. Average number of preamble transmissions required for successful access $\mathbb{E}[K]$

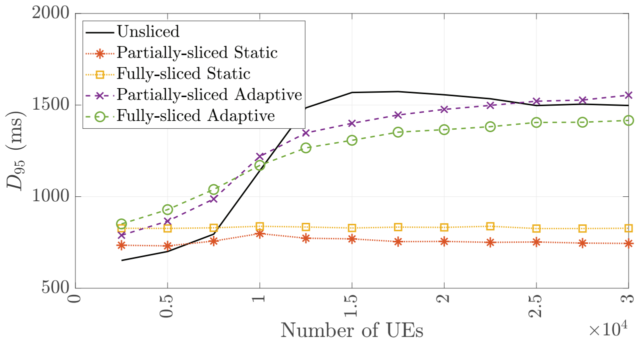

Figure 10. H2H slice. 95th percentile of access delay D_{95}

Adaptive-sliced policies perform considerably better than the Unsliced one. A value of P_s above 90 % is guaranteed with all slicing policies. The number of transmitted preambles and the access delay is illustrated in Figs. 9 and 10, respectively. When the number of UEs exceeds 7500, the Static-sliced policies provide the lowest values in the two metrics since they have reserved resources for efficient performance. The H2H slice performance is not affected by increasing the number of mIoT accesses with the Static-sliced policies. Comparing the Adaptive-sliced policies and the Unsliced one, they show similar behavior in terms of D and the Adaptive-sliced outperforms the Unsliced in terms of K.

VII. Conclusions

Implementing network slicing in 5G radio access networks achieves isolation between different services hosted by different slices. Traffic variations generated in one slice minimally affect the other slices. We verified that limiting the accesses to the maximum capacity of each slice allows for maximizing the utilization of the RACH by increasing the probability of successful access of UEs, isolating each slice from the congestion produced by the different services. A shared preamble subset serves connection requests that exceed capacity.

In the fully-sliced policies, a segmentation of preambles and UL grants is performed, which means that any congestion issue in one slice will not be propagated to the rest. In the partially-sliced policies, a segmentation of preambles and not of UL grants are performed; complete isolation is not reached, but an efficient occupation of the available UL grants is achieved since unused UL grants by slices with a light load can be exploited by access request from other slices.

The partially-sliced static policy can improve the performance of eMBB and H2H slices in heavy-load mIoT scenarios due to a constant allocation of resources. For mIoT services, the adaptive-sliced policies provide better performance.

Acknowledgment

This work was supported in part by Meta 2022 AI4AI Research, 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] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60–65, 2018.

[2] N. Alliance, "Description of network slicing concept," *NGMN 5G P*, vol. 1, no. 1, pp. 1–11, 2016.

[3] K. Samdanis, S. Wright, A. Banchs, A. Capone, M. Ulema, K. Obana *et al.*, "5G Network Slicing: Part 1–Concepts, Principales, and Architectures [Guest Editorial]," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 70–71, 2017.

[4] K. Samdanis, S. Wright, A. Banchs, A. Capone, M. Ulema, and K. Obana, "5G network slicing–part 2: Algorithms and practice," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 110–111, 2017.

[5] 3GPP, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Sept 2011.

[6] 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.

[7] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[8] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "Resource sharing efficiency in network slicing," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 909–923, 2019.

[9] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5G and future mobile networks: mobility, resource management, and challenges," *IEEE communications magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[10] C. Kalalas, F. Vazquez-Gallego, and J. Alonso-Zarate, "Handling mission-critical communication in smart grid distribution automation services through LTE," in *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2016, pp. 399–404.

[11] H. Althumali, M. Othman, N. K. Noordin, and Z. M. Hanapi, "Priority-based load-adaptive preamble separation random access for QoS-differentiated services in 5G networks," *Journal of Network and Computer Applications*, vol. 203, p. 103396, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804522000558

[12] 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.

[13] 3GPP, *TS 36.321, Medium Access Control (MAC) Protocol Specification*, Sept 2012.

[14] ——, *TS 36.211, Physical Channels and Modulation*, Dec. 2014.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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, 2020.

[21] L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, J. Martinez-Bauset, V. Casares-Giner, and L. Guijarro, "Efficient random access channel evaluation and load estimation in lte-a with massive mtc," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1998–2002, 2018.