IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 14, 2020, accepted March 6, 2020, date of publication March 13, 2020, date of current version March 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980685

Characterization of Radio Access Network Slicing Scenarios With 5G QoS Provisioning

IRENE VILÀ, JORDI PÉREZ-ROMERO, (Member, IEEE), ORIOL SALLENT, AND ANNA UMBERT

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain

Corresponding author: Irene Vilà (irene.vila.munoz@upc.edu)

This work was supported in part by the Spanish Research Council and FEDER funds through SONAR 5G Grant under Grant TEC2017-82651-R, and in part by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia under Grant 2019FI_B1 00102.

ABSTRACT 5G systems are envisaged to support a wide range of application scenarios with variate requirements. To handle this heterogeneity, 5G architecture includes network slicing capabilities that facilitate the partitioning of a single network infrastructure into multiple logical networks on top of it, each tailored to a given use case and provided with appropriate isolation and Quality of Service (QoS) characteristics. Network slicing also enables the use of multi-tenancy networks, in which the same infrastructure can be shared by multiple tenants by associating one slice to each tenant, easing the cost-effective deployment and operation of future 5G networks. Concerning the Radio Access Network (RAN), slicing is particularly challenging as it implies the configuration of multiple RAN behaviors over a common pool of radio resources. In this context, this work presents a Markov model for RAN slicing capable of characterizing diverse Radio Resource Management (RRM) strategies in multi-tenant and multi-service 5G scenarios including both guaranteed and non-guaranteed bit rate services. The proposed model captures the fact that different radio links from diverse users can experience distinct spectral efficiencies, which enables an accurate modeling of the randomness associated with the actual resource requirements. The model is evaluated in a multi-tenant scenario in urban micro cell and rural macro cell environments to illustrate the impact of the considered RRM polices in the QoS provisioning.

INDEX TERMS Markov processes, radio access networks, RAN slicing, radio resource management, quality of service.

I. INTRODUCTION

The forthcoming Fifth Generation (5G) systems target the simultaneous support of a wide variety of application scenarios and vertical industries (e.g. automotive, utilities, smart cities, high-tech manufacturing) with distinct and variate requirements (i.e. high data rates, low latency, high mobility) [1]. 5G will enable both the evolution of the current business models and the emergence of new ones. Partnerships at multiple-layers will be established, ranging from sharing the infrastructure to exposing specific network capabilities as an end to end service and integrating partners' services into the 5G system through a rich and software oriented capability set [2].

In this context, 5G systems need to be provided with the flexibility and configurability required to satisfy the foreseen diversity. With this purpose, one key feature of the 5G system architecture is network slicing, which is based on Software-Defined Networking (SDN) and Network Function Virtualization (NFV) technologies [3] and allows the sharing of a common infrastructure among diverse end-to-end (self-contained) logical networks (i.e. network slices), each tailored for a given use case [4]. Each network slice can be provided with the appropriated isolation and optimized characteristics for a particular application. Therefore, network slicing enables the use of multitenancy [5], in which multiple tenants, e.g. communication providers or mobile virtual network operators (MVNO), can share the common infrastructure to provide services to their own users, resulting in reductions of capital and operational costs.

Network slicing support is especially challenging for the Radio Access Network (RAN) [6], [7], which is the most resource-demanding (and costliest) part of the mobile network. The reason is that the radio spectrum is a limited

The associate editor coordinating the review of this manuscript and approving it for publication was Usama Mir.

resource and the level of isolation required by the different slices can compromise the efficiency of the radio resources usage. Consequently, one of the major research problems in the field is the definition of novel Radio Resource Management (RRM) strategies, or the adaptation of existing ones, that allow both the implementation of slicing at the RAN and the fulfilment of the users' Quality of Service (QoS) requirements [8].

The first steps in the standardization process of the system architecture and functional aspects to support network slicing in both the 5G Core Network (5GC) and in the Next-Generation RAN (NG-RAN) have been conducted by 3GPP [9], [10]. This includes the definition of the 5G New Radio (NR) interface and the 5G QoS model. Moreover, implementation aspects of network slicing in the NG-RAN have been studied from multiple angles, ranging from virtualization techniques and programmable platforms with slice-aware traffic differentiation and protection mechanisms [11]–[13] to algorithms for dynamic resource sharing across slices [14]. Similarly, [15] analyses the RAN slicing problem in a multi-cell network in relation to RRM functionalities and [16] proposes an adaptation algorithm for resource allocation, which is based on the deviations from requirements. In turn, [17] proposes a set of vendor-agnostic configuration descriptors intended to characterize the features, policies and resources to be put in place across the radio protocol layers of a NG-RAN node for the realization of concurrent RAN slices. Also, [18] proposes a procedure to establish the level of centralization of different RRM functions while [19] presents an adaptive inter-slice TDD allocation algorithm, where resources are assigned by minimizing costs and interferences. Some other works focus on the network slice admission control for slices requests that need to support a given number of users for a certain time, such as [20], [21], which target to optimize the infrastructure providers' revenue, or [22], which optimizes the network utilization by incorporating traffic forecasting capabilities. Moreover, frameworks for the realization and study of network slicing have been implemented and tested such as in [23], which designs a system for the management of RAN slices and the provisioning of radio resources to slices based on their requirements in Long Term Evolution (LTE) technology, considering its extension to 5G NR.

In the above context, this paper tackles the RAN slicing problem from a modeling perspective by proposing and developing a Markov model characterization of RAN slicing in multi-tenant and multi-service scenarios. Markovian approaches have been widely used to characterize the utilization of resources in many fields, such as in mobility [24], cloud computing [25], Call Admission Control (CAC) scheme for 3G [26] or for heterogeneous networks Radio Access Technologies (RAT) policies [27]. More recently, works in the field of 5G exploit Markov modeling to approach a proactive resource allocation scheme in highly mobile networks [28], the management of Admission Control (AC) for handoff requests between small cell and macro cell domains [29], the computation of the estimated spectrum requirement [30] and the management of slices' creation [31]. Markov chain models have also been considered in [32] for spectrum sharing schemes and primary/secondary scenarios [33]–[35]. Our recent works [36], [37] introduced a first approach to the use of Markov chains models to characterize different RRM policies for RAN slicing at different layers of the protocol stack. Similarly, [36], [37] considered the 5G QoS model, which embraces prioritization among traffic flows. This paper constitutes a step forward in the establishment of a wider range of relationships among the different dimensions of the RAN slicing problem, including aspects of the radio environment, services types and configurations, traffic scenarios, etc.

Specifically, this paper includes several novelties and advances: (i) The model allows the definition of both Guaranteed Bit Rate (GBR) and non-Guaranteed Bit Rate (non-GBR) services in terms of its corresponding 5G QoS parameters such as the Allocation and Retention Priority (ARP) and the 5G QoS Identifier (5QI) parameters. (ii) RRM policies at the different radio protocol layers have been properly characterized to support both types of services (i.e. GBR and non-GBR) and to provide slicing capabilities, so that isolation between slices is achieved both in the admission of users and the allocation of resources. (iii) In terms of the resource allocation, a new statistical model has been developed that allows capturing 5G scenarios with variate radio propagation conditions (i.e. diverse spectral efficiencies can be perceived by the different users in the system). In particular, the presented statistical model for layer 2 allows deriving the probability density function (pdf) of both the required resources and the assigned resources in multi-tenant and multi-service scenarios based on the QoS requirements. From these pdfs, different performance indicators are extracted. (iv) The proposed model is evaluated in a scenario comprised of two tenants providing both GBR and non-GBR services over urban micro cell and rural macro cell environments, where different performance metrics of interest (e.g. blocking probability, occupancy, throughput) are assessed and the relationships between the different parameters are analyzed.

The rest of the paper is organized as follows. Section II presents the system model, describing the analytical Markov chain approach considered. In Sections III and IV the slicing-aware AC policy performed at layer 3 and the radio resource allocation procedure at layer 2 are characterized, respectively. Section V presents different performance metrics that can be extracted from the presented model. Section VI describes two example scenarios considered for 5G RAN slicing for the subsequent validation of the model and analysis of the performance results at state and system levels, as well as a discussion of the impact of introducing a new tenant. Finally, Section VII summarizes the conclusions.

II. SYSTEM MODEL

A scenario comprised of a common radio infrastructure shared by N tenants is assumed. Each tenant operates in a

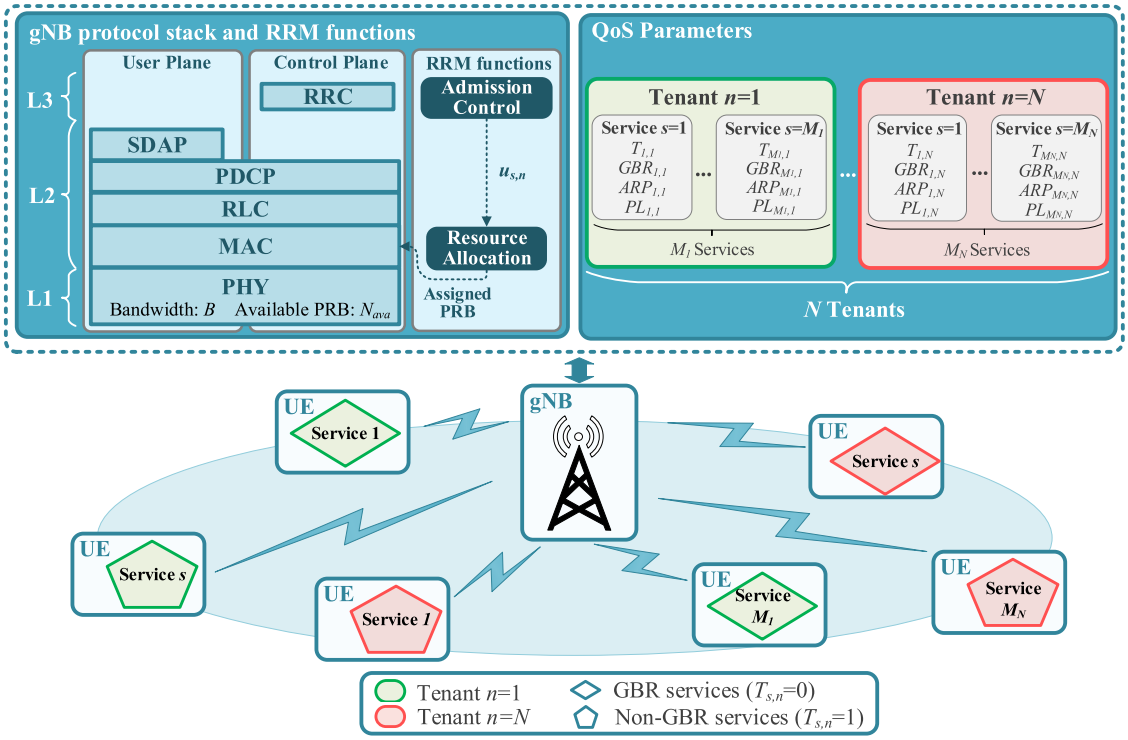

FIGURE 1. System model conceptual scheme.

RAN slice, which has already been created and deployed by the Operations, Administration and Management (OAM) system among the RAN infrastructure by means of the Network Slice Subnet Management Function (NSSMF) [38]. The n-th tenant provides M_n service types, which can be either GBR or non-GBR services. The service type indicator $T_{s,n}$ takes value 0 if the s-th service of the n-th tenant is GBR and 1 if it is non-GBR. The QoS profile is given by the GBR value (i.e., the bit rate to be provided to the user of a GBR service, also referred to as Guaranteed Flow Bit Rate (GFBR) in 5G 3GPP's terminology), the ARP indicator [9], which defines the relative importance of the service requesting for resources and starts from 1 (highest priority) onwards (for successive lower priority services), and the priority level associated with the 5QI. Therefore, for GBR services, the QoS profile is characterized by the guaranteed $GBR_{s,n}$, the $ARP_{s,n}$ and the 5QI priority level $PL_{s,n}$ for $s = 1, \ldots, M_n$ and $n = 1, \ldots, N$. In the case of non-GBR services, the $GBR_{s,n}$ is set to 0, as no data rate is guaranteed.

The considered scenario is depicted in Fig. 1. It is comprised of a gNB, which is the NG-RAN node operating the 5G NR interface, composed of a cell with a certain bandwidth subdivided in Physical Resource Blocks (PRB) of bandwidth B. Hence, the cell has a number of available PRBs N_{ava} at layer 1 to serve the User Equipment (UE) traffic demands.

Fig. 1 also illustrates the different layers of the radio interface protocol stack at the gNB that determine how the user plane information and the control plane signaling is transferred between the UE and the gNB. Specifically, the transfer of the user plane information (e.g. IP packets associated to the services of the UE) is carried out through the Service Data Adaptation Protocol (SDAP), Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), Medium Access Control (MAC) and physical (PHY) layers. In turn, the control plane signaling can be either generated at the Radio Resource Control (RRC) layer (e.g. for measurement reporting) or at upper layer Non Access Stratum (NAS) protocols for signaling between the UE and the 5GC network (e.g. for session establishment). In both cases, signaling is transferred between UE and gNB using the PDCP, RLC, MAC and PHY layers. Further details on the functionalities of each layer can be found in [39].

In order to efficiently use the radio resources and ensure the QoS requirements of the users in the system when transferring the information through the different layers of the protocol stack, the gNB includes a set of RRM functionalities, namely the AC function at Layer 3 (L3) and the resource allocation at Layer 2 (L2), which are the focus of this paper and are briefly described in the following.

Whenever a new session of a service (i.e. a new QoS flow in 3GPP terminology) is established for transferring user data through the 5G system for a given UE, the 5GC network requests the gNB to set up resources to support this QoS flow at the radio interface. The 5GC network provides the gNB with the slice identifier of the tenant, i.e. the Single Network Slice Selection Assistance Information (S-NSSAI) [9], and the QoS parameters of the service. At the gNB, the SDAP layer maps the QoS flow to a Data Radio Bearer (DRB) that enables data transfer through the different layers of the radio

interface protocol stack according to the expected QoS [10]. For this reason, and to make sure that the cell will have sufficient resources to support the requested QoS, the AC function at L3 is required to determine the acceptance or rejection of the new DRB in accordance with its QoS parameters and the available capacity.

Let us consider that each UE is provided with a single DRB and that, as a result of the AC function, the number of admitted DRB, and thus users, of the s-th service of the n-th tenant in the cell is $u_{s,n}$. Then, the resource allocation function is associated to the MAC layer at L2 and is in charge of dynamically assigning the N_{ava} available PRBs in the cell among the admitted DRBs, thus determining how the data of these DRBs travels through the physical layer.

In order to configure the multi-tenant behaviour of the AC and resource allocation functions, the OAM provides the RRM policy information to the gNB [38], including guidance for the split of radio resources among the different RAN slices. Specifically, this paper assumes that the OAM provides each gNB with the per-tenant parameters required to configure the RRM functionalities, as described in the detailed models for the L3 admission control and L2 resource allocation functions that are given in Sections III and IV, respectively.

Assuming that users generate sessions of exponential duration according to a Poisson arrival process, the dynamic evolution of the number of admitted users of each service type and tenant can be characterized in general by a Continuous Time Markov Chain (CTMC) with $(M_1 + M_2 + \ldots + M_N)$-dimensional states. Let us define $S_{(u_{1,1},\ldots,u_{M_1,1},u_{1,2},\ldots,u_{M_2,2},\ldots,u_{1,N},\ldots,u_{M_N,N})}$ as the state in which $u_{1,1},\ldots,u_{M_1,1},u_{1,2},\ldots,u_{M_2,2},\ldots,u_{1,N},\ldots,u_{M_N,N}$ users are admitted in the system. The state space is defined as:

$$S = \{S_{(u_{1,1},\ldots,u_{M_N,N})} | u_{s,n} \leq U_{\max,s,n}\} \tag{1}$$

where $U_{max,s,n}$ is the maximum allowed number of users of the s-th service of the n-th tenant, which is established for hardware limitation purposes (processor, memory, power). It is worth mentioning that the AC can further restrict the number of users per service to a value lower than $U_{max,s,n}$ depending on how the AC policy is specified.

Transitions between states occur due to session arrivals or session departures. In this respect, it is considered that session arrivals are generated with rate $\lambda_{s,n}$ for the s-th service of the n-th tenant, while the average session duration of this service is $1/\mu_{s,n}$. Moreover, since AC in L3 is in charge of admitting or rejecting users' requests depending on the system occupation, it also affects the transitions between states. In this respect, let us define $AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})}$ as the binary AC indicator for the arrivals of the s-th service and n-th tenant, taking the value 1 if the new service request is accepted and 0 otherwise. It is worth mentioning that, since transitions can only occur due to the admission of a new session or the finalization of an existing session, they can only increase or decrease the number of admitted users in one unit, meaning that transitions are only possible between neighboring states. Therefore, the transition rate $q_{x,y}$ from a state x to another state $y \neq x$ is given by:

$$q_{x,y} = \begin{cases} \lambda_{s,n} AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})} & \text{if } x = S_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})} \text{ and} \\ & y = S_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{M_N,N})} \\ u_{s,n}\mu_{s,n} & \text{if } x = S_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})} \text{ and} \\ & y = S_{u_{1,1},\ldots,u_{s,n}-1,\ldots,u_{M_N,N})} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where transitions to states with one more user (i.e. first condition in (2)) depend on the $AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})}$ indicator and $\lambda_{s,n}$ while the transitions to a state with one less user (i.e. second condition in (2)) depend on the number of users of the service that decreases and $\mu_{s,n}$. The rest of transitions from/to states with more or less than one user are not allowed (i.e. third condition in (2)).

Fig. 2 illustrates a state transition diagram with all the possible transitions to/from a given state $S_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}$ when increasing or decreasing one user of each of the services. From the state transition diagram, the general Steady-State Balance Equation (SSBE) is given by:

$$\begin{aligned} &P_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{\mathrm{M}_N,N})}[\sum_{s,n} u_{s,n}\mu_{s,n} \\ &\quad + \sum_{s,n \atop S_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{\mathrm{M}_N,N})} \in S} \lambda_{s,n} AC^{s,n}_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{\mathrm{M}_N,N})}] \\ &= \sum_{s,n} P_{(u_{1,1},\ldots,u_{s,n}-1,\ldots,u_{\mathrm{M}_N,N})} \lambda_{s,n} AC^{s,n}_{(u_{1,1},\ldots,u_{s,n}-1,\ldots,u_{\mathrm{M}_N,N})} \\ &\quad + \sum_{s,n \atop S_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{\mathrm{M}_N,N})} \in S} P_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{\mathrm{M}_N,N})}(u_{s,n}+1)\mu_{s,n} \end{aligned} \tag{3}$$

where $P_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}$ corresponds to the steady-state probability of being in state $S_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}$. When the SSBEs are obtained for all the states, the steady-state probabilities can be computed by using numerical methods capable of solving the system of equations composed by the different SSBEs and the following normalization constraint:

$$\sum_{S_{(u_{1,1},\ldots,u_{\mathrm{M}_N,N})} \in S} P_{(u_{1,1},\ldots,u_{\mathrm{M}_N,N})} = 1 \tag{4}$$

III. ADMISSION CONTROL AT LAYER 3

The AC function at L3 decides on the acceptance or rejection of users initiating new sessions depending on their requested QoS and the already admitted users in the cell. Since the QoS requirements have different nature for GBR and non-GBR services, the considered AC policy behaves differently in each case.

For GBR services ($T_{s,n} = 0$) the admission or rejection decision of a new user from the s-th service of the n-th tenant considers its requested $GBR_{s,n}$ and $ARP_{s,n}$ parameters and

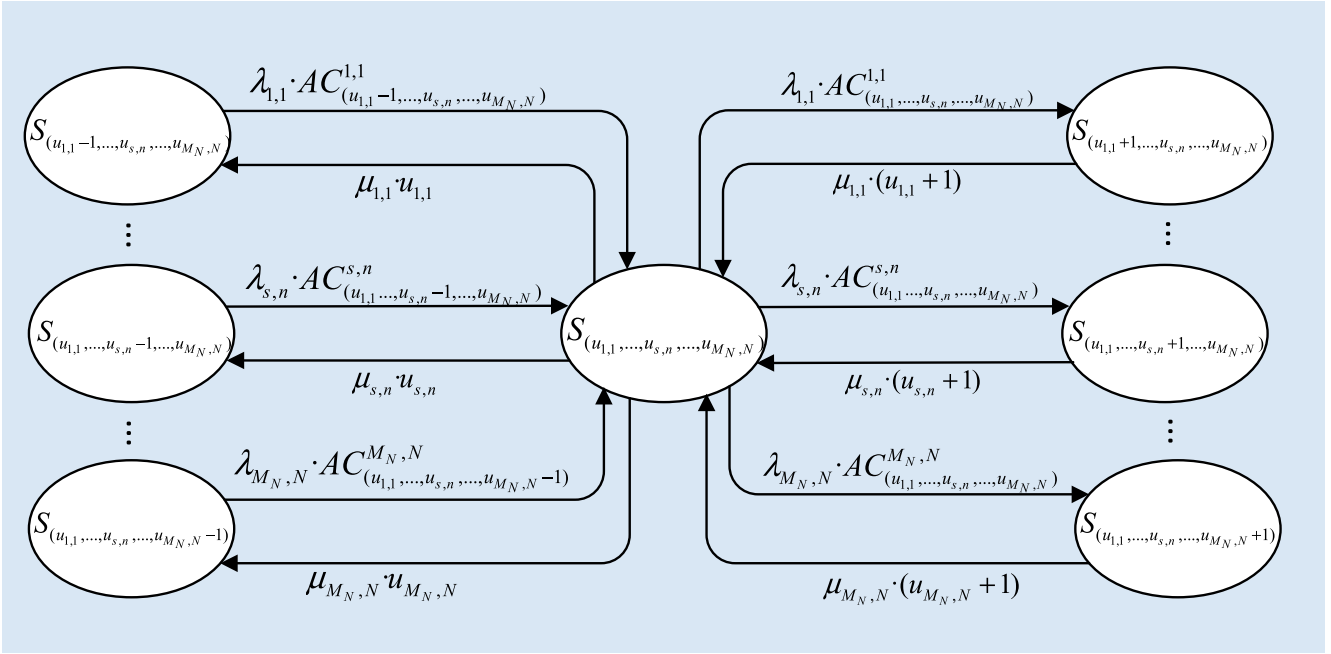

FIGURE 2. State transition diagram with all the transitions from/to a given state $S_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}$ with $u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N}$ users.

those of the already admitted GBR users of the same tenant, together with a per-tenant capacity threshold $C_{max,n}$ defined as the maximum aggregate GBR that can be admitted for tenant n. This threshold $C_{max,n}$ is provided by the OAM as the RRM policy information to be enforced by the AC. Correspondingly, the new GBR user can be admitted if the aggregate GBR considering both the new user and the already accepted users of tenant n with higher or equal priority than the new user (i.e. with ARP lower or equal than $ARP_{s,n}$) does not exceed the threshold $C_{max,n}$. It is worth noting that, in practice, $C_{max,n}$ could be dynamically adjusted, e.g. through a Self-Organizing Network (SON) function [40], to account for the spectral efficiency conditions experienced by the users in the cell.

In contrast, for non-GBR services ($T_{s,n} = 1$) the system is not committed to guarantee any GBR value. Therefore, the AC in this case only checks that the number of admitted users does not exceed the maximum threshold $U_{max,s,n}$.

Based on these considerations, the AC decision for a new user of the s-th service of the n-th tenant is given by:

$$AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})} = \begin{cases} 1 & \text{if } (T_{s,n}=0 \text{ and } \sum_{\substack{s'=1 \\ ARP_{s',n}\le ARP_{s,n} \\ T_{s',n}=0}}^{M_n} u_{s',n}\cdot GBR_{s',n} + GBR_{s,n} \le C_{\max,n}) \\ & \text{or } (T_{s,n}=1 \text{ and } (u_{s,n}+1) \le U_{\max,s,n}) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Note that the considered AC function is performed independently for each tenant, in the sense that the admission of a new user belonging to a certain tenant only depends on this tenant's parameters and occupation. Also, notice that the time scale at which AC is triggered depends on the session generation rate of the users $\lambda_{s,n}$, which will typically be in the order of seconds.

IV. MODEL OF RADIO RESOURCE ALLOCATION AT LAYER 2

From the perspective of the Markov model, admissions at L3 and session finalizations generate state transitions. In turn, within the sojourn time of a given state, the resource allocation at L2 defines how the N_{ava} available PRBs in the cell are assigned to the admitted users in a given state. In practice, this is performed by the Packet Scheduler (PS), which dictates in a short time basis (millisecond time scale) the user allocation of PRBs to transmit data packets by considering the QoS constraints and the actual radio conditions associated with each user. In this paper, the PS behavior is characterized in accordance with the sojourn time of the Markov model states (i.e., typically few seconds) and, consequently, all the short time scale components of the PS (e.g. fast fading) will be averaged out in the resource allocation model.

Fig. 3 shows the considered resource allocation process. It is performed independently for each tenant n and assuming a maximum number of PRBs, $N_{th,n}$, to be allocated to this tenant, which is also determined by the OAM and provided to the gNB as RRM policy information. As depicted in Fig. 3, the process firstly allocates PRBs among the admitted users of GBR services. For this purpose, let us denote as $arp(1,n)$, $arp(2,n)$,..., $arp(M_n,n)$ the list of ARP

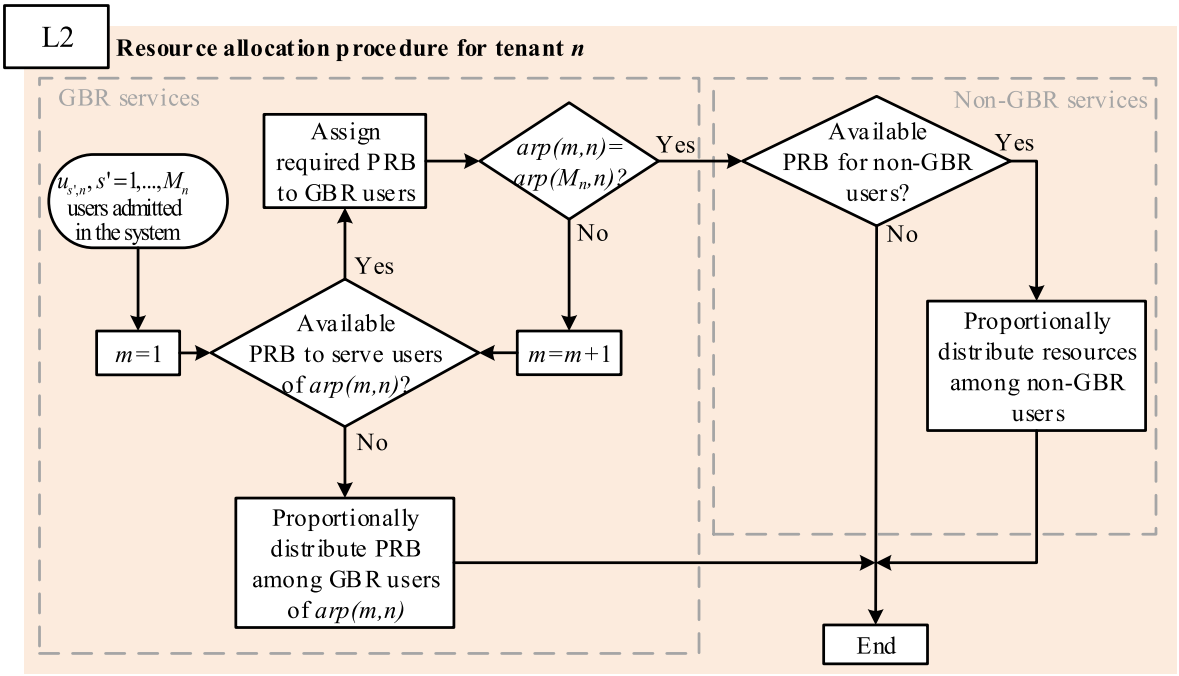

FIGURE 3. Resource allocation procedure.

values in increasing order for the GBR services of tenant n, i.e. starting by $arp\,(1,n) = min_{s|T_{s,n}=0} ARP_{s,n}$ and ending with $arp\,(M_n,n) = max_{s|T_{s,n}=0} ARP_{s,n}$. Then, the process firstly allocates the PRBs required by all the GBR services with $arp(1,\ n)$, then with $arp(2\ ,n)$, and so on. As long as there are available PRBs to serve the users of an ARP value, resources are assigned. Instead, when not enough PRBs are available (i.e. there is congestion), the procedure proportionally distributes the available PRBs among the users of that ARP value. Other criteria might be adopted for handling persistent congestion situations such as the use of congestion control functions. However, they are out of the scope of this work. After this process, the remaining PRBs are allocated among the admitted users of non-GBR services. Both phases are further developed in the following.

A. RESOURCE ALLOCATION TO GBR SERVICES

For a given state $x = S_{(u_{1,1},\ldots,u_{M_N,N})}$, the target here is to model the aggregate number of assigned PRBs to the users of a given GBR service s of tenant n, denoted as $a_{x,s,n}$. Provided that the $u_{s,n}$ users of this service experience variable propagation conditions, this is a random variable that should be characterized statistically through its pdf, denoted as $f_{a_{x,s,n}}(k)$. In order to obtain $f_{a_{x,s,n}}(k)$, the first stage is to formulate the statistical distribution of the number of required PRBs by the $u_{s,n}$ users. Then, the procedure of Fig. 3 is needed to allocate the $N_{th,n}$ available PRBs among the M_n service types, considering that the required PRBs of each service can be different and that the ARPs establish a prioritization among services.

The number of required PRBs by a user of the GBR service s of tenant n is given by:

$$N_{req,s,n} = \frac{GBR_{s,n}}{S_{eff} \cdot B} \tag{6}$$

where S_{eff} is the spectral efficiency, measured in b/s/Hz, and B is the PRB bandwidth. Since S_{eff} fluctuates depending on the propagation conditions that users experience when moving around the cell, it is treated as a random variable. Therefore, based on measurements collected from the different users, it is possible to derive the pdf of the random variable $Y = 1/(S_{eff} \cdot B)$, denoted as $f_Y(y)$. This pdf is obtained by gathering samples of the wideband Channel Quality Indicator (CQI) distribution [41]. The CQI is an integer index that indicates the modulation and coding scheme available to user in accordance to its experienced propagation conditions. The wideband CQI distribution is computed by the gNB from the CQI reports provided by the different users, and it is reported to a Management and Data Analytics Function (MDAF) [42] with a given periodicity (e.g. 15 minutes). The MDAF gathers samples of this distribution and averages them for a longer time period in order to get the adequate statistical validity to be representative of the cell conditions. Then, the averaged distribution of the CQI indices can be directly mapped to the distribution of the S_{eff} according to the tables in section 5.2.2 of [43]. From the distribution of the spectral efficiency, $f_{S_{eff}}\ (y)$, the pdf of Y, $f_Y(y)$ can be extracted based on $f_{S_{eff}}(y)$ as:

$$f_Y\ (y) = f_{S_{eff}} \left(\frac{1}{y \cdot B}\right) \frac{1}{y^2 B} \tag{7}$$

Correspondingly, the number of required resources $N_{req,s,n}$ is another random variable, whose pdf is:

$$f_{N_{\text{req},,s,n}}(k) = f_Y\left(\frac{k}{GBR_{s,n}}\right)\frac{1}{GBR_{s,n}} \tag{8}$$

Assuming that each user experiences independent propagation conditions, the pdf of the aggregate number of required PRBs $r_{x,s,n}$ of the s-th service of the n-th tenant in state x can be computed as:

$$f_{r_{x,s,n}}(r) = \left(\frac{1}{GBR_{s,n}}\right)^{u_{s,n}} \cdot \underbrace{f_Y\left(\frac{r}{GBR_{s,n}}\right) * \ldots * f_Y\left(\frac{r}{GBR_{s,n}}\right)}_{u_{\text{s,n}}} \tag{9}$$

where $*$ represents the convolution operator.

As seen in Fig. 3, the allocation of resources to GBR users is performed in accordance with the priorities established by the ARP values of each service in the tenant, treating all the services with the same *arp(m,n)* together. Correspondingly, let us denote as $\mathbf{R_{arp(m,n)}} = [r_{x,s',n} \mid ARP_{s',n} = arp(m,n),\ T_{s',n} = 0]$ the vector of required resources for the GBR services of tenant n with the same *arp(m,n)*.Considering independence between requirements of different services, the joint pdf of the $r_{x,s,n}$ variables in $\mathbf{R_{arp(m,n)}}$ is:

$$f_{\mathbf{R_{arp(m,n)}}}\left(\mathbf{R_{arp(m,n)}}\right) = \prod_{\substack{s=1 \\ ARP_{s,n}=arp(m,n) \\ T_{s,n}=0}}^{M_n} f_{r_{x,s,n}}(r_{x,s,n}) \tag{10}$$

In turn, let us denote as $Z_{arp(m,n)}$ the random variable representing the aggregated number of PRBs already assigned to the GBR services in tenant n of ARP lower than *arp(m,n)* and $f_{Z_{arp(m,n)}}(z)$ its pdf. For *arp*(1 ,*n*), since no resources have been yet assigned, $f_{Z_{arp(1,n)}} = \delta(z)$, where $\delta(\cdot)$ is the Dirac delta function.

The pdf $f_{a_{x,s,n}}(k)$ of the s-th service of the n-th tenant with $ARP_{s,n} = arp(m,n)$ can be computed by conditioning the value of assigned resources $a_{x,s,n}$ to the requirements $\mathbf{R_{arp(m,n)}}$ of the services with the same *arp(m,n)* and to the already assigned resources by GBR services $Z_{arp(m,n)}$. Then, by applying the law of the total probability, this yields to:

$$\begin{aligned} f_{a_{x,s,n}}(k) = &\int_0^\infty \cdots \int_0^\infty f_{a_{x,s,n}|\mathbf{R_{arp(m,,n)}},Z_{arp(m,n)}}(k|\mathbf{R_{arp(m,n)}}, z) \\ &\cdot f_{Z_{arp(m,n)}}(z) \cdot f_{\mathbf{R_{arp(m,n)}}}(\mathbf{R_{arp(m,n)}}) \cdot \\ &\cdot dz \cdot \prod_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} dr_{x,s',n} \end{aligned} \tag{11}$$

where $f_{a_{x,s,n}|\mathbf{R_{arp(m,n)}},Z_{arp(m,n)}}(k|\mathbf{R_{arp(m,n)}}, z)$ is the pdf of $a_{x,s,n}$ conditioned to $\mathbf{R_{arp(m,n)}}$ and $Z_{arp(m,n)}$, which is given by:

$$f_{a_{x,s,n}|\mathbf{R_{arp(m,n)}},Z_{arp(m,n)}}(k|\mathbf{R_{arp(m,n)}}, z) = \begin{cases} \delta(k - r_{x,s,n}) \\ \qquad \text{if} \sum_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} r_{x,s',n} \le N_{th,n} - z \\ \delta(k - \alpha_{arp(m,n)} r_{x,s,n}) \\ \qquad \text{if} \sum_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} r_{x,s',n} > N_{th,n} - z \end{cases} \tag{12}$$

The first condition in (12) considers the case that the number of available resources for the tenant is higher than the required resources given by $\mathbf{R_{arp(m,n)}}$ vector. Then, each user gets the required resources, i.e. the aggregate resources are $a_{x,s,n} = r_{x,s,n}$. In turn, the second condition in (12) considers the case that there are not sufficient resources for the tenant to fulfil the requirements of all the users of *arp(m,n)* (i.e. there is congestion). In this case, the assigned resources are proportionally distributed among these users as $a_{x,s,n} = \alpha_{arp(m,n)} \cdot r_{x,s,n}$ with:

$$\alpha_{arp(m,n)} = \max\left(\frac{N_{th,n} - z}{\sum_{\substack{s=1 \\ ARP_{s,n}=arp(m,n) \\ T_{s,n}=0}}^{M_n} r_{x,s,n}}, 0\right) \tag{13}$$

Once the computation of $f_{a_{x,s,n}}(k)$ for each of the services with *arp(m,n)* is completed, the pdfs for the subsequent ARP value *arp(m + 1 ,n)* need to be derived. Then, the pdf $f_{Z_{arp(m+1,n)}}(z)$ is computed based on $f_{Z_{arp(m,n)}}(z)$ and the aggregate resources that have been assigned to the services with *arp(m,n)*, denoted by the random variable $A_{arp(m,n)}$. This yields to:

$$f_{Z_{arp(m+1,n)}}(z) = \int_0^{+\infty} f_{Z_{arp(m,n)}}(z') \cdot f_{A_{arp(m,n)}|Z_{arp(m,n)}}(z - z'|z') \cdot dz' \tag{14}$$

where $f_{A_{arp(m,n)}|Z_{arp(m,n)}}(t|z)$ is the pdf of $A_{arp(m,n)}$ conditioned to $Z_{arp(m,n)}$ given by:

$$\begin{aligned} f_{A_{arp(m,n)}|Z_{arp(m,n)}}(t|z) = &f_{K_{arp(m,n)}}(t) \cdot H(t, N_{th,n} - z) \\ &+ \delta(t - (N_{th,n} - z)) \cdot \int_{N_{th,n}-z}^{\infty} f_{K_{arp(m,n)}}(k) dk \end{aligned} \tag{15}$$

$$\mathrm{H}(x, y) = \begin{cases} 1 & x < y \\ 0 & x \ge y \end{cases} \tag{16}$$

where $f_{K_{arp(m,n)}}(z)$ is the pdf of the aggregate required resources for all the GBR services of $arp(m,n)$, given by:

$$f_{K_{arp(m,n)}}(t) = \underbrace{f_{r_{x,1,1}}(t) * \ldots * f_{r_{x,s,n}}(t)}_{s=1,\ldots,M_n, ARP_{s,n}=arp(m,n), T_{s,n}=0} \quad (17)$$

Note that, for the case $m = M_n$, which corresponds to the maximum ARP value among the GBR services in tenant n, expression (14) gives the pdf of the random variable $Z_{GBR,n}$ that represents the aggregate PRBs assigned to all the GBR services in that tenant, i.e.

$$f_{Z_{GBR,n}}(z) = \int_0^{+\infty} f_{Z_{arp(M_n,n)}}(z') \cdot f_{A_{arp(M_n,n)}|Z_{arp(M_n,n)}}(z - z'|z') \cdot dz' \quad (18)$$

B. RESOURCE ALLOCATION TO NON-GBR SERVICES

As seen in Fig. 3, the remaining PRBs after the allocation to GBR users of tenant n are distributed among the users of non-GBR services of the tenant. This distribution is carried out proportionally based on the 5QI priority level and the number of admitted users $u_{s,n}$ of each service, with the proportional constant $\sigma_{s,n}$, defined as:

$$\sigma_{s,n} = \frac{u_{s,n} \cdot (1/PL_{s,n})}{\sum_{\substack{s'=1 \\ T_{s',n}=1}}^{M_n} u_{s',n} \cdot (1/PL_{s'n})} \quad (19)$$

Consequently, the assigned PRBs to the non-GBR service s of tenant n are given by $a_{x,s,n} = (N_{th,n} - Z_{GBR,n}) \cdot \sigma_{s,n}$. The pdf of $a_{x,s,n}$ can be obtained by using the pdf of the assigned resources after GBR allocation $f_{Z_{GBR}}(z)$ as:

$$f_{a_{x,s,n}}(k) = \frac{1}{\sigma_{s,n}} f_{Z_{GBR,n}}(N_{th,n} - \frac{k}{\sigma_{s,n}}) \quad (20)$$

V. PERFORMANCE METRICS

Based on the steady-state probabilities, this section develops the different performance metrics of interest for the evaluation of the considered RRM strategies.

A. BLOCKING PROBABILITY

Blocking states are those in which the acceptance of a new user of a given service is not possible. Specifically, the set of blocking states for users of the s-th service of the n-th tenant is denoted as $S^b_{s,n}$, defined as:

$$S^b_{s,n} = \{S_{(u_{1,1},..,u_{M_N,N})} \in S | AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})} = 0\} \quad (21)$$

Similarly, the set of blocking states for the n-th tenant, S^b_n, are those states in which the acceptance of one user from any of the services of this tenant is not possible. Therefore, it is defined as the intersection of the sets of blocking states for the services of this tenant, i.e. $S^b_n = S^b_{1,n} \cap S^b_{2,n} \cap \ldots \cap S^b_{M_n,n}$. Similarly, the set of all blocking states in the system S^b is expressed as the intersection of the set of blocking states of each tenant/service.

Based on the blocking states, the blocking probability computed per service and per tenant is given by:

$$P^b_{s,n} = \sum_{S(u_{1,1},\ldots,u_{M_N,N}) \in S^b_{s,n}} P_{(u_{1,1},\ldots,u_{M_N,N})} \quad (22)$$

This can be easily extended to compute the blocking probability per tenant or the global blocking probability by considering S^b_n or S^b, respectively, in the summation of (20).

B. OCCUPANCY METRICS

Given the steady-state probabilities $P_{(u_{1,1},\ldots,u_{M_N,N})}$, it is also possible to compute different metrics that provide information about the occupancy of the system. The average number of admitted users $\overline{U_{s,n}}$ of the s-th service of the n-th tenant is given by:

$$\overline{U_{s,n}} = \sum_{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S} u_{s,n} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \quad (23)$$

The average number of admitted users per tenant can be computed by adding the average number of users per service, i.e. $\overline{U_n} = \overline{U_{1,n}} + \overline{U_{2,n}} + \ldots + \overline{U_{M_n,n}}$. Similarly, the global system average number of admitted users $\overline{U}$ would be computed as the sum of the average number of users for all services and tenants.

Another system occupancy metric that can be obtained from the model is the average PRB utilization $\overline{a_{s,n}}$ aggregated per each service. The average aggregated PRB utilization in a given state $x = S_{(u_{1,1},\ldots,u_{M_N,N})}$ for the s-th service and the n-th tenant $\overline{a_{x,s,n}}$ is:

$$\overline{a_{x,s,n}} = \int_0^{\infty} k \cdot f_{a_{x,s,n}}(k) \cdot dk \quad (24)$$

Then, the system average aggregated PRB utilization per service $\overline{a_{s,n}}$ can be computed by considering all the states $\overline{a_{x,s,n}}$ and the steady-state probabilities as:

$$\overline{a_{s,n}} = \sum_{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S} \overline{a_{S_{(u_{1,1},\ldots,u_{M_N,N})},s,n}} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \quad (25)$$

Accordingly, the average PRB utilization per tenant would result from $\overline{a_n} = \overline{a_{1,n}} + \overline{a_{2,n}} + \ldots + \overline{a_{M_n,n}}$ while the global system PRB utilization $\overline{a}$ can be computed by adding the average PRB utilization of each tenant. Then, the average normalized PRB utilization of the s-th service of the n-th tenant $\overline{\omega_{s,n}}$ is expressed as:

$$\overline{\omega_{s,n}} = \frac{\overline{a_{s,n}}}{N_{ava}} \quad (26)$$

Similarly to the other metrics, the average normalised PRB utilization per tenant $\overline{\omega_n}$ and the global average normalized PRB utilization $\overline{\omega}$ result from adding the average normalized PRB utilization of the tenant's services or all the services, respectively.

C. AVERAGE AGGREGATED THROUGHPUT

For a state $x = S_{(u_{1,1},\ldots,u_{M_N,N})}$, the average aggregated throughput for the s-th service of the n-th tenant $\overline{Th_{x,s,n}}$ is computed differently for GBR and non-GBR services. In the case of GBR services, $\overline{Th_{x,s,n}}$ is computed as:

$$\overline{Th_{x,s,n}} = \int_0^\infty \cdots \int_0^\infty k \cdot f_{Thr_{x,s,n}|\mathbf{R_{arp(m,n)}},Z_{arp(m,n)}}(k|\mathbf{R_{arp(m,n)}}, z) \cdot f_{Z_{arp(m,n)}}(z) \cdot f_{\mathbf{R_{arp(m,n)}}}(\mathbf{R_{arp(m,n)}}) \cdot dz \cdot \prod_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} dr_{x,s',n} \cdot dk \quad (27)$$

where $f_{Th_{x,s,n}|\mathbf{R_{arp(m,n)}},Z_{arp(m,n)}}(k|\mathbf{R_{arp(m,n)}}, z)$ is the pdf of the aggregated throughput $\overline{Th_{x,s,n}}$ for the GBR service s from tenant n conditioned to $\mathbf{R_{arp(m,n)}}$ and $Z_{arp(m,n)}$, given by:

$$f_{Th_{x,s,n}|\mathbf{R_{arp(m,n)}},Z_{arp(m,n)}}(k|\mathbf{R_{arp(m,n)}}, z) = \begin{cases} \delta(k - u_{s,n} \cdot GBR_{s,n}) \\ \quad \text{if} \sum_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} r_{x,s',n} \leq N_{th,n} - z \\ \delta(k - \alpha_{arp(m,n)} \cdot u_{s,n} \cdot GBR_{s,n}) \\ \quad \text{if} \sum_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} r_{x,s',n} > N_{th,n} - z \end{cases} \quad (28)$$

Expression (28) follows the same principle as (12), considering that, when the number of available PRBs is higher than the number of required PRBs (first condition in (28)), all users get the required PRBs and thus the throughput of each user is $GBR_{s,n}$. Instead, when there are not sufficient resources (second condition in (28)), the assigned PRBs are just a fraction $\alpha_{arp(m,n)}$ of the required ones and thus the throughput of each user is $\alpha_{arp(m,n)} \cdot GBR_{s,n}$.

Regarding non-GBR services and considering that the number of assigned PRBs to the non-GBR users depends on the spare PRBs after the allocation to GBR users, the average aggregated throughput $\overline{Th_{x,s,n}}$ for non-GBR users is independent of its actual spectral efficiency and is given by:

$$\overline{Th_{x,s,n}} = \overline{a_{x,s,n}} \cdot \overline{S_{eff}} \cdot B \quad (29)$$

where $\overline{S_{eff}}$ is the average of the spectral efficiency S_{eff}, derived from measurements in a similar way as variable Y.

The system average aggregated throughput $\overline{Th_{s,n}}$ for the s-th service of the n-th tenant can be computed by considering $\overline{Th_{x,s,n}}$ for all services and the steady-state probability as:

$$\overline{Th_{s,n}} = \sum_{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S} \overline{Th_{S_{(u_{1,1},\ldots,u_{M_N,N})},s,n}} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \quad (30)$$

The average aggregated throughput for the n-th tenant $\overline{Th_n}$ and the average global system aggregated throughput $\overline{Th}$ result from the summation of the average aggregated throughputs of the tenant's services or all the services in the system, respectively.

D. DEGRADATION PROBABILITY

Degradation occurs when congestion is reached and some GBR admitted users cannot be assigned with their required resources to provide $GBR_{s,n}$. Instead, they are assigned with a lower number of resources according to the considered resource allocation criteria.

For a given state x, the degradation probability of the GBR service s-th of the n-th tenant with $ARP_{s,n} = arp(m,n)$ can be obtained by the following expression:

$$p_{\deg,x,s,n} = \int_0^\infty \cdots \int_0^\infty \delta(k - \alpha_{arp(m,n)} r_{x,s,n}) \cdot H(N_{th,n} - z, \sum_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} r_{x,s',n}) \cdot f_{Z_{arp(m,n)}}(z) \cdot f_{\mathbf{R_{arp(m,n)}}}(\mathbf{R_{arp(m,n)}}) \cdot dz \cdot \prod_{\substack{s'=1 \\ ARP_{s',n}=arp(m,n) \\ T_{s',n}=0}}^{M_n} dr_{x,s',n} dk \quad (31)$$

where $H(\cdot)$ is defined in (16) and allows considering in the computation only those cases in which there are not enough resources to satisfy the service requirements (i.e. the second condition in (12)).

Based on the degradation at state level and the steady-state probabilities, the degradation probability of the s-th of the n-th tenant at system level is given by:

$$P_{s,n}^{\deg} = \sum_{x \in S} p_{\deg,x,s,n} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \quad (32)$$

Note that the degradation probability is not computed for non-GBR services, as $GBR_{s,n}$ is not established for them.

VI. PERFORMANCE EVALUATION

This section presents a performance analysis of the proposed analytical model. Firstly, the detailed characteristics of the considered scenarios are presented. Then, the model validation, the analysis of the scenarios at both state and global system levels and a discussion of the impact of introducing a new tenant follow.

A. CONSIDERED SCENARIO

The scenario under test considers $N = 2$ tenants, referred to *Tenant 1* and *Tenant 2*. Tenant 1 provides $M_1 = 3$ different services while Tenant 2 provides $M_2 = 2$ services. For each of the services, the QoS parameters summarised

TABLE 1. Services per tenant.

Tenant id (n)	Service id (s)	Type	GFBR	ARP	PL
1	1	GBR	5 Mb/s	1	30
	2	GBR	2 Mb/s	2	40
	3	Non-GBR	-	3	70
2	1	Non-GBR	-	3	80
	2	Non-GBR	-	3	60

in Table 1 have been specified according to the QoS model of [9]. The included parameters consist of the ARP, the PL and the GFBR, which specifies the GBR value to be provided to a QoS flow.

The NG-RAN is composed by one gNB with a single cell. Two different radio environments are considered: a Urban Micro-cell (UMi) and a Rural Macro-cell (RMa) according to [44]. The configured parameters used to obtain the different results included in this section are summarized in Table 2. Considering the defined 5G NR numerologies and channel bandwidth constraints for the different 5G NR bands [45], the largest channel bandwidth allowed to the considered RMa environment operating in the 700 MHz band is 20 MHz. Then, in order to ease the comparison of results in both RMa and UMi, the same 20 MHz bandwidth has been selected for both environments, which is composed by 51 PRBs [46], each one of $B = 360$ kHz corresponding to 12 Orthogonal Frequency-Division Multiple Access (OFDMA) subcarriers with subcarrier separation of 30 kHz. Notice that, for both of the environments, the spectral efficiency samples of the different users are generated by mapping the Signal to Interference Ratio (SINR) of users to spectral efficiency values according to the link-level model of section. 5.2.7 of [47], which includes a parameter α to model modem implementation losses and link conditions such as the effects of error rates and retransmissions.

B. MODEL VALIDATION

The proposed model has been validated by comparing its results to the ones obtained with a custom made system-level simulator. The results from the Markov model have been extracted by developing it on Matlab and employing the Gauss-Seidel method [48] to solve the SSBE equation in (3) for all the states [37].

The system-level simulator used for the validation allows defining 5G multi-tenant and multi-service scenarios with diverse cell deployments and with different QoS requirements for each service. Sessions generation follow a Poisson distribution with exponentially distributed session duration. The implemented RRM functionalities in terms of AC and resource allocation in the simulator follow the same principles as in the Markov model (i.e. Sections III and IV of this paper) and the adopted validation scenarios correspond to those considered in Table 1 and Table 2. The simulation

TABLE 2. Model configuration parameters.

Environment	UMi	RMa
ISD (Inter-Site distance)	200 m	1735 m
gNB height	10 m	35 m
UE height	1.5 m	1.5 m
Minimum gNB-UE distance	10 m	35 m
Path Loss and Shadowing model	Model of sec. 7.4 of [44]	
Shadowing standard deviation in Line of Sight (LOS)	4	4
Shadowing standard deviation in Non-Line of Sight (NLOS)	7.82	8
Frequency	3.6 GHz	704 MHz
Total gNB transmitted power	37 dBm	45 dBm
gNB antenna Gain	Omnidirectional antenna with 5 dBi gain	
UE noise figure	9 dB	
Link-level model to map SINR and bit rate	Model in section. 5.2.7 of [47] with maximum spectral efficiency of 5.97 b/s/Hz (corresponding to SINR= 30 dB) and minimum SINR= -10 dB with α=0.6.	
Number of spectral efficiency samples for $f_Y(y)$ generation	10^7 samples obtained by simulating users at different random positions uniformly distributed within the cell and measuring their spectral efficiency.	
Average Spectral Efficiency	5.7 b/s/Hz	5.1 b/s/Hz
Cell available PRBs (N_{ava})	51 PRBs	
PRB Bandwidth (B)	360 kHz	
Cell maximum number of users	$U_{max,s,n}$=50 users for s,n=1,2	
Cell total capacity	104 Mb/s	94 Mb/s
L3 Maximum Capacity Threshold Tenant 1 ($C_{max,1}$)	62.4 Mb/s (corresponds to the 60% of the cell capacity)	56.4 Mb/s (corresponds to the 60% of the cell capacity)
L3 Maximum Capacity Threshold Tenant 2 ($C_{max,2}$)	N/A (only non-GBR services)	
L2 Maximum number of resources per tenant ($N_{th,1}$)	35.7 PRBs (corresponds to the 70% of the total cell PRBs)	
L2 Maximum number of resources per tenant ($N_{th,2}$)	15.3 PRBs (corresponds to the 30% of the total cell PRBs)	
Average session generation rate Tenant 1	varied from 0.001 to 0.1 sessions/s (corresponds to 0.27 Mb/s to 27.6 Mb/s of traffic offered load by GBR services)	
Average session generation rate Tenant 2	0.04 sessions/s	
Average session duration	120 s	
Session generation distribution Tenant 1	30% of session arrivals are of service 1, 40% of service 2 and 30% of service 3.	
Session generation distribution Tenant 2	40% of session arrivals are of service 1 and 60% of service 2.	

time has been set to 10^8 s to ensure a proper convergence of the results.

Table 3 contains the absolute percentage of error of the simulator results with respect to the Markov model results for all the services in the system for two session generation rates of Tenant 1. The small error percentages obtained show the suitability of the analytical model to evaluate RAN slicing

TABLE 3. Comparison of results between Markov model and simulator.

Tenant id (n)	Service id (s)	Tenant 1 Session generation rates	UMi			RMa		
			% Error ($P^b_{s,n}$)	% Error ($\overline{a_{s,n}}$)	% Error ($\overline{Th_{s,n}}$)	% Error ($P^b_{s,n}$)	% Error ($\overline{a_{s,n}}$)	% Error ($\overline{Th_{s,n}}$)
1	1	0.02	0%	2.6%	3.3%	0%	1.9%	3.3%
		0.1	2.6%	0.6%	0.4%	2.0%	1.3%	0.5%
	2	0.02	0%	2.5%	1.5%	0%	2.2%	1.5%
		0.1	1.4%	1.7%	0.8%	0.2%	0.3%	0.7%
	3	0.02	-	0.0%	0.1%	-	0%	0.2%
		0.1	-	0.7%	0.6%	-	1.6%	1.8%
2	1	0.02	-	0.1%	0.0%	-	0.1%	0.1%
		0.1	-	0.2%	0.3%	-	0.7%	0.6%
	2	0.02	-	0.1%	0.1%	-	0.1%	0.2%
		0.1	-	0.0%	0.2%	-	0.4%	0.6%

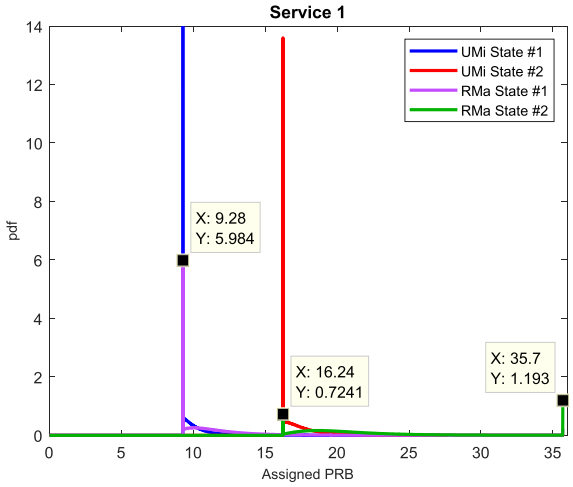

FIGURE 4. Pdf of assigned resources to service 1 for State #1 and State #2 in UMi and RMa scenarios.

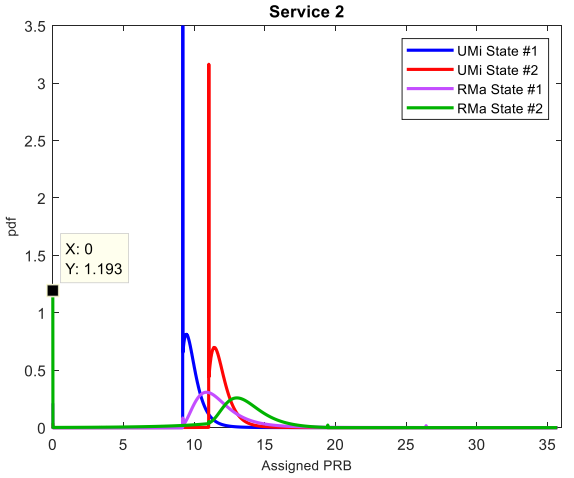

FIGURE 5. Pdf of assigned resources to service 2 for State #1 and State #2 in UMi and RMa scenarios.

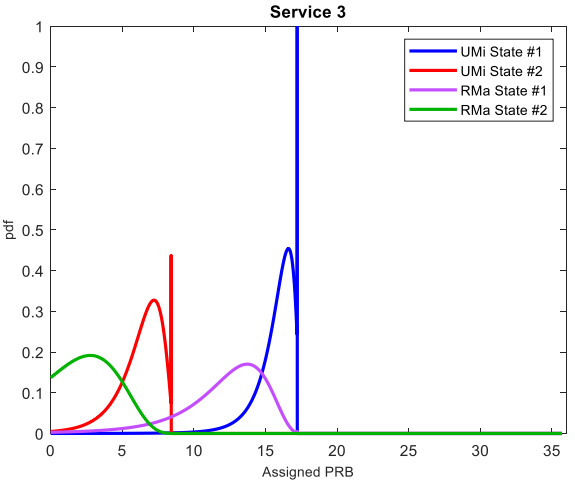

FIGURE 6. Pdf of assigned resources to service 3 for State #1 and State #2 in UMi and RMa scenarios.

scenarios with multiple tenants and services when considering diverse environments in terms of cell deployment (cell radius, transmitted power, etc.) and different traffic loads.

C. PERFORMANCE RESULTS

This section includes the performance results analysed both from state and system level perspectives. In addition, the analysis of the impact of introducing a new tenant in the system is provided.

1) STATE LEVEL ANALYSIS

In order to analyze the behavior of the proposed radio resource allocation procedure in L2, two states have been selected and the procedure of Section IV has been followed. State #1 comprises $u_{1,1} = 4$, $u_{2,1} = 10$ and $u_{3,1} = 12$ users belonging to Tenant 1 and $u_{1,2} = 6$ and $u_{2,2} = 6$ belonging to Tenant 2. State #2 is composed of $u_{1,1} = 7$, $u_{2,1} = 12$ and $u_{3,1} = 12$ users belonging Tenant 1 and $u_{1,2} = 9$ and $u_{2,2} = 6$ users belonging to Tenant 2. Notice that State #2 has higher GBR requirements than State #1. Both states have been evaluated in both UMi and RMa environments, defined in Table 2.

Regarding the resource allocation of Tenant 1, Fig. 4, Fig. 5 and Fig. 6 include the pdfs of the assigned resources $fa_{x,s,n}(k)$ in both states and environments for all services of Tenant 1. The impact of the environment can be appreciated: the pdfs for the UMi environment are narrower than the ones for the RMa environment. This reveals that the UMi has better radio propagation conditions than the RMa and less variability in terms of spectral efficiency given the more confined coverage area. In this sense, the peak observed in the obtained pdfs, which is produced by those users having the maximum spectral efficiency available in the system, is much prominent for the pdfs belonging to the UMi scenario.

The effect of increasing the number of GBR users can be observed in Fig. 4 and Fig. 5 when comparing the pdfs of States # 1 and # 2 of services 1 and 2, respectively (i.e. $u_{1,1} = 4$ and $u_{2,1} = 10$ in State # 1 and $u_{1,1} = 7$ and $u_{2,1} = 12$ in State # 2). As expected, the pdfs in State #2 are

TABLE 4. Average aggregated assigned PRBs and throughput.

Environment		UMi		RMa	
		$\overline{a_{x,s,n}}$ (PRB)	$\overline{Th_{x,s,n}}$ (Mb/s)	$\overline{a_{x,s,n}}$ (PRB)	$\overline{Th_{x,s,n}}$ (Mb/s)
State #1					
Tenant 1	Service 1	10.05	19.99	11.94	19.99
	Service 2	9.98	19.88	11.85	19.88
	Service 3	15.67	32.02	12.09	22.31
Tenant 2	Service 1	6.55	13.39	6.55	12.09
	Service 2	8.74	17.86	8.74	16.13
State #2					
Tenant 1	Service 1	17.58	35	20.83	34.94
	Service 2	11.94	23.95	12.92	22
	Service 3	6.24	12.74	3	5.53
Tenant 2	Service 1	8.1	16.55	8.1	14.95
	Service 2	7.2	14.7	7.2	13.28

shifted to a higher number of assigned PRBs, as a result of having a greater resource demand. Moreover, a peak at 35.7 PRBs appears for service 1 in State # 2 in the RMa environment, reflecting that there exists a probability that all the resources available for Tenant 1 (i.e., 35.7 PRBs) are assigned to service 1. In this case, given that service 1 has the lowest ARP (i.e. the highest priority), this means that no resources can be assigned to service 2, which is evidenced by the peak at 0 assigned PRBs that appears in Fig. 5.

In the case of service 3, as it is a non-GBR service, it is provided with the remaining resources after the allocation to GBR services (i.e. service 1 and service 2), as shown in Fig. 6. This explains the fact that in State # 1 the pdf is centred in a higher number of resources than in State # 2. Furthermore, the peaks in the pdfs appear at the right hand side instead of the left hand side observed for GBR services, reflecting that service 3 is granted with lower priority and is assigned with the spare radio resources.

Table 4 presents the state average aggregated PRB utilization, $\overline{a_{x,s,n}}$, and the state average aggregated throughput, $\overline{Th_{x,s,n}}$, for States # 1 and # 2 and all the services in the system in both UMi and RMa environments. In the case of Tenant 2, given that both services are non-GBR, no requirements in terms of resources are established, so the pdf of the assigned PRBs does not depend on the spectral efficiency or the environment. Consequently, the same values in terms of $\overline{a_{x,s,n}}$ are observed for UMi and RMa environments. From these values, the effect of the proportional sharing constant $\sigma_{s,n}$ is appreciated, which depends on both the PL and the number of users. For State # 1, service 1's average aggregated PRB utilization is lower than the one obtained for service 2, as the latter has a lower PL (higher priority) and both services have the same number of users. The contrary case is obtained in State # 2, where service 1's average aggregated PRB utilization is higher than the one obtained for service 2 as a result of the higher number of users of service 1. In terms of the average aggregated throughput of Tenant 2's services, higher throughput is achieved in the UMi scenario thanks to its better propagation conditions.

Furthermore, in order to get a deeper insight into the behavior of the resource allocation procedure from a multi-state perspective, Fig. 7 presents the average aggregated PRBs utilization $\overline{a_{x,s,n}}$ in a RMa environment for each of the services in the different states, given by the number of users $u_{s,n}$ of each tenant/service. Regarding Tenant 1, Fig. 7a shows that the resource allocation for service 1 is performed independently of the number of users of service 2. This is because service 1 is the first provided with resources as it has the lowest ARP value (i.e. higher priority). Differently, the resource allocation for service 2 depends on the number of users of service 1 (Fig. 7b), as its PRB allocation is performed after the allocation to service 1. Therefore, the PRB utilization of service 2 increases when reducing the number of users of service 1. Nevertheless, service 2 PRB utilization values are lower than in the case of service 1, as the GBR requirement is also lower. Moreover, Fig. 7c proves that service 3 is provided with the remaining resources after the allocation of PRBs of services 1 and 2, so the highest PRB utilization of service 3 is achieved when the number of users of services 1 and 2 is low. This happens because service 3 is the one with lowest priority (i.e. highest ARP) and is a non-GBR service. Besides, Fig. 7d and Fig. 7e show how the $N_{th,2}$ PRBs available to Tenant 2 are distributed among the non-GBR services 1 and 2 according to the proportional sharing constant $\sigma_{s,n}$. In this sense, both graphs are complementary to each other, i.e. in states with no users of service 1, all the PRBs are allocated to service 2, while in states with no users of service 2 all the PRBs are allocated to service 1.

2) SYSTEM LEVEL ANALYSIS

This section discusses the global system performance in both RMa and UMi scenarios, focusing on the behavior of the RRM procedures considered for L3 and L2 and the capability of the overall system to adopt diverse configurations. Results have been generated according to the configuration in Table 2, by varying the session generation rate of Tenant 1, which also implies the variation of its GBR services' traffic offered load, and keeping constant the rate of Tenant 2. Traffic offered load by the GBR services of the n-th tenant is defined as:

$$\theta_n = \sum_{\substack{s \\ T_{s,n}=0}}^{M_n} GBR_{s,n}\lambda_{s,n} \cdot (1/\mu_{s,n}) \tag{33}$$

In order to analyse the RRM at L3, Fig. 8 represents the blocking probability of the GBR services in the system (i.e. services 1 and 2 from Tenant 1) in both considered environments as a function of their offered load. For high offered loads, blocking probabilities for service 2 are slightly higher than for service 1 in both environments, exhibiting that it has a higher ARP (i.e. less priority) than service 1. Anyway, the blocking probabilities remain in low values (i.e. less than 1.5%) for the considered offered loads, which are lower than the maximum capacity threshold in both UMi

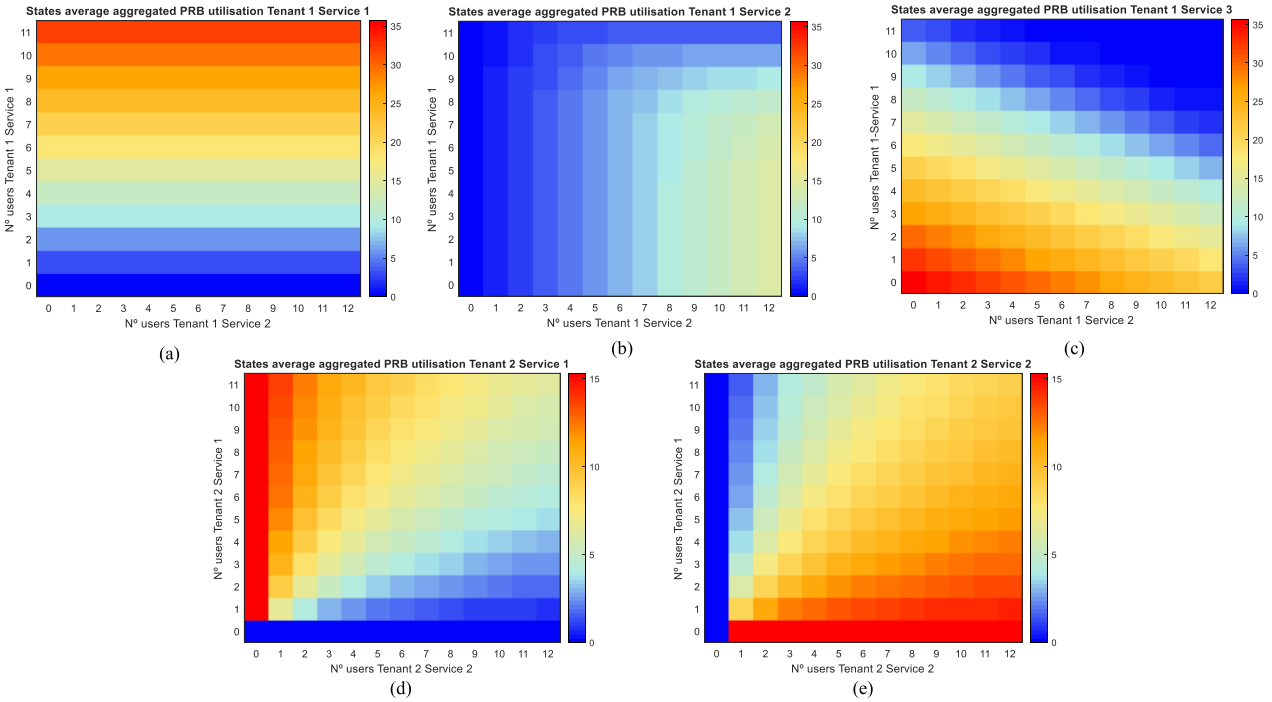

FIGURE 7. Average aggregated PRB utilization in the different states for (a) Service 1 of Tenant 1, (b) Service 2 of Tenant 1, (c) Service 3 of Tenant 1 (d) Service 1 of Tenant 2 and (e) Service 2 of Tenant 2.

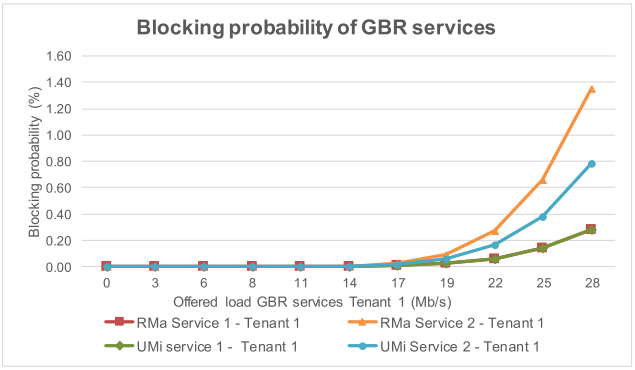

FIGURE 8. Blocking probability of GBR services in RMA and UMi environments.

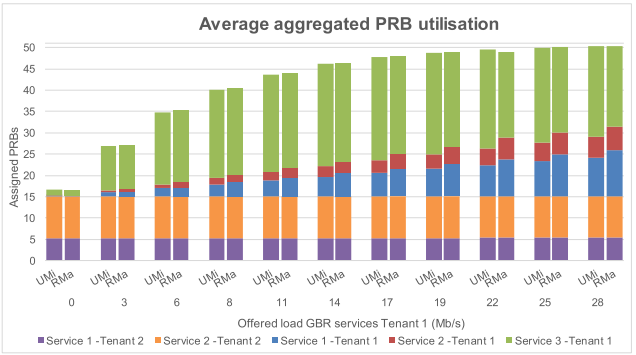

FIGURE 9. Average assigned PRBs aggregated per service in the RMa and UMi environments.

and RMa environments (i.e. $C_{max,1} = 62.6$ Mb/s for UMi and $C_{max,1} = 56.2$ Mb/s in RMa). In addition, higher blocking probabilities of service 2 are found for the RMa environment. In this case, the admission of a user implies a higher number of required resources, so a lower number of users can be admitted in the system. It is worth pointing out that this effect is not noticed for service 1 due to its higher priority. For non-GBR services, all users are admitted as the maximum number of users $U_{max,s,n}$ is not reached, which is the only constraint considered for them.

The RRM at L2 is analyzed in terms of the system average PRB utilization, $\overline{a_{s,n}}$, and throughput $\overline{Th_{s,n}}$, both aggregated per service. Fig. 9 represents the system average aggregated PRB utilization by each of the services for the RMa and UMi environments, as a function of Tenant 1's GBR services offered load. Similar to the previous section, the average aggregated PRB utilization of GBR services is greater in the RMa scenario as a result of worse propagation conditions, which leads to a higher resource demand. Moreover, the PRB utilization of GBR services (i.e. service 1 and 2 of Tenant 1) is consistent with the configured GBR and ARP values. For instance, the average aggregated PRB utilization of service 1 of Tenant 1 is greater than the one for service 2 of the same tenant, as service 1 has a larger GBR requirement (i.e. $GBR_{1,1} = 5$ Mb/s vs $GBR_{2,1} = 2$ Mb/s) and lower ARP (i.e. more priority). Apart from this, when varying the traffic offered load of Tenant 1, the PRB utilization of services belonging to Tenant 2 remain constant, showing the isolation capability of the resource allocation procedure included in the model.

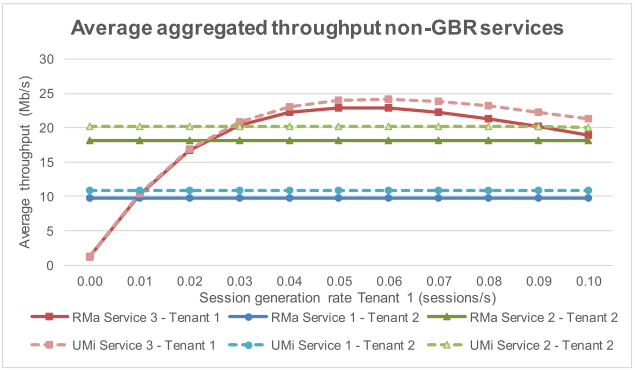

FIGURE 10. Average aggregated throughput to non-GBR services in the RMa and UMi environments.

In terms of the system average aggregated throughput, $\overline{Th_{s,n}}$, GBR services present equivalent results to the ones obtained for the PRB utilization. For GBR services, the obtained average aggregated throughput is proportional to the GBR value and the average number of users in the system. This proportionality is achieved thanks to the low degradation probabilities that are obtained in the considered scenarios (maximums of 0.001% for UMi and of 2% in the RMa).

However, the mentioned proportionality is not found for non-GBR services, so its throughput dynamics deserve a deeper analysis. Fig. 10 shows the behavior of the average aggregated throughput in both RMa and UMi scenarios for non-GBR services, this time as a function of Tenant 1's session generation rate, which embraces all Tenant 1's services. For all non-GBR services, higher average aggregated throughput is given in the UMi scenario provided its better propagation conditions. This occurs in spite that the average PRB utilization of all the services of Tenant 2 in Fig. 9 is the same for both environments and, in the case of service 3 of Tenant 1, it is larger for the RMa environment. This reflects that the throughput metric for non-GBR does not only depend on the number of assigned resources but also on the propagation conditions in each scenario.

Besides, as a result of the isolation capability achieved at layer 2, Fig. 10 also shows that the aggregated average throughput of Tenant 2 services remains constant when increasing the session generation rate of Tenant 1. The effect of the PL is also noticeable as higher throughput is provided to service 2, which has the lowest PL (i.e. higher priority). Instead, the throughput of service 3 of Tenant 1 experiences a different trend depending on the session generation rate of this tenant. Specifically, the average throughput of this service increases with the session generation rate up to approximately 0.05 sessions/s. The reason for this is that increasing the session rate raises the probability of having at least one user of this service that can exploit all the PRBs left by GBR users. However, for traffic generation rates higher than 0.05 sessions/s, the average throughput starts to decrease as there are less available resources to service 3, which is caused by a higher resource requirement of GBR services.

TABLE 5. Slicing thresholds for the proposed solutions to include Tenant 3.

Tenant id (n)	N_{ava}	RMa		UMi	
		$C_{max,n}$	$N_{th,n}$	$C_{max,n}$	$N_{th,n}$
Solution A: Increase PRB availability					
1	65 PRBs	62.4 Mb/s	35.7 PRBs	56.4 Mb/s	35.7 PRBs
2		N/A (only non-GBR services)	15.3 PRBs	N/A (only non-GBR services)	15.3 PRBs
3		20.8 Mb/s	10 PRBs	18.8 Mb/s	10 PRBs
Solution B: Re-configure slicing thresholds					
1	51 PRBs	62.4 Mb/s	35.7 PRBs	56.4 Mb/s	35.7 PRBs
2		N/A (only non-GBR services)	5.1 PRBs	N/A (only non-GBR services)	5.1 PRBs
3		20.8 Mb/s	10 PRBs	18.8 Mb/s	10 PRBs

3) ANALYSIS OF THE INTRODUCTION OF A new tenant

One key feature of network slicing is the flexibility to create, re-configure and release network slices. The creation of a network slice implies the allocation of a suitable capacity to support the traffic of a new tenant. In case that the existing tenants do not consume all the capacity in the system, the L3 maximum capacity $C_{max,n}$ and L2 maximum number of PRBs $N_{th,n}$ of the new slice can be configured to provide the required capacity from the remaining part. If no capacity is available, as it is the case in the example of previous section, where Tenant 1 and Tenant 2 already consume 100% of the PRBs of the gNB, the simplest approach would be to extend the system capacity e.g., through extending the assigned bandwidth. In case that this is not feasible, due to e.g. hardware constraints or spectrum limitation, the re-configuration of the existing slices so that the actual capacity is redistributed among all the tenants could be explored. In order to illustrate these approaches, this section discusses the addition of a new tenant in the scenario described in Section VI.A.

Let us assume that the new tenant, denoted as *Tenant 3*, provides two GBR services with $GBR_{1,3} = 1$ Mb/s and $GBR_{2,3} = 0.5$ Mb/s, both of them with the same priority $ARP_{1,3} = ARP_{2,3} = 2$. The average session generation rate of this tenant is 0.07 sessions/s, the average session duration is 120s, and 70% of sessions belong to service 1 and 30% to service 2.

The first solution to create the new slice, denoted as *Solution A,* increases the PRB availability in the gNB, by increasing the cell bandwidth from 20 to 25 MHz, which results in an increase of available PRBs from $N_{ava} = 51$ to $N_{ava} = 65$ PRBs [45]. Then, Tenant 3 is configured to use part of this additional capacity by configuring the values of the thresholds L3 maximum capacity $C_{max,3}$ and L2 maximum number of PRBs $N_{th,3}$ as in Table 5. The corresponding thresholds of Tenants 1 and 2 remain unchanged with respect to Section VI.A. In contrast, the second solution, denoted as *Solution B*, re-configures the values of $C_{max,n}$ and $N_{th,n}$ for

TABLE 6. Blocking probability of GBR services for *Solution A* and *B*.

Tenant (n)	Service (s)	Blocking probability $P^b_{s,n}$ (%)			
		Solution A		Solution B	
		RMa	UMi	RMa	UMi
1	1	0.282	0.282	0.282	0.282
	2	1.351	0.788	1.351	0.788
3	1	0.014	0.002	0.014	0.002
	2	0.011	0.001	0.011	0.001

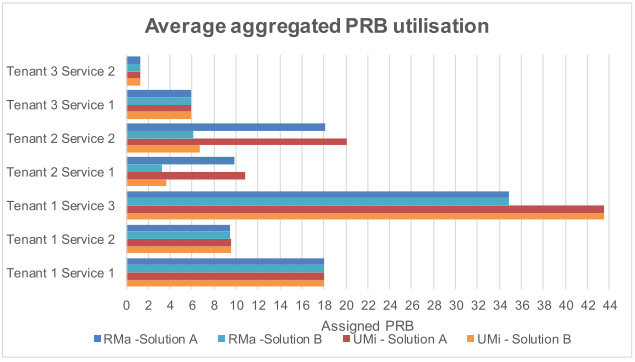

FIGURE 11. Average assigned PRBs aggregated per service in the RMa and UMi environments for Solution A and B.

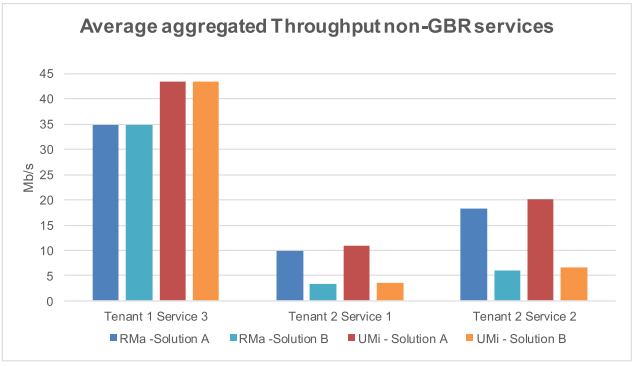

FIGURE 12. Average aggregated throughput to non-GBR services in the RMa and UMi environments considering Solution A and B.

TABLE 7. Average number of users and average aggregated throughput for GBR services in the system.

Tenant id (n)	Service id (s)	RMa		UMi	
		$\overline{U_{s,n}}$ (users)	$\overline{Th_{s,n}}$ (Mb/s)	$\overline{U_{s,n}}$ (users)	$\overline{Th_{s,n}}$ (Mb/s)
1	1	3.60	18.01	3.60	18.02
	2	4.74	9.41	4.74	9.53
3	1	5.91	5.89	5.92	5.91
	2	2.52	1.26	2.53	1.27

the existing tenants. Specifically, given that Tenant 2 only carries non-GBR services, the selected re-configuration consists in reducing the value of $N_{th,2}$ without modifying the slicing thresholds of Tenant 1, as seen in Table 5. For the new Tenant 3 the same thresholds as in *Solution A* are used.

Regarding RRM L3, the blocking probability $P^b_{s,n}$ of all the GBR services in the system has been obtained for *Solution A* and *Solution B* when the session generation rate of Tenant 1 is 0.1 sessions/s. Results are presented in Table 6. Since the value of $C_{max,1}$ and $C_{max,3}$ is the same in both solutions, the obtained blocking probabilities for all GBR services of Tenants 1 and 3 are also the same. It is observed that the blocking probabilities of Tenant 1 do not change with the introduction of Tenant 3, i.e. they are the same as in Fig. 8. This shows that the AC algorithm provides isolation in the admission of GBR users. Regarding Tenant 3, although blocking probabilities are low for both UMi and RMa, the lowest blocking probabilities are achieved once again in the UMi environment. Moreover, blocking probabilities of service 1 are higher than those of service 2 although both services have the same ARP value. The reason is that the traffic offered for service 1 of Tenant 3 is higher than that of service 2.

The analysis of the proposed solutions for including Tenant 3 has also been analysed in relation to RRM at L2 by means of the PRB utilization $\overline{a_{s,n}}$ and the average aggregated throughput $\overline{Th_{s,n}}$, also for a session generation rate of Tenant 1 equal to 0.1 sessions/s. Fig. 11 contains the comparison of the PRB utilization values obtained for each of the services for *Solutions A* and *B* in the RMa and UMi environments. It is observed that the PRB utilization of Tenant 1 and Tenant 3 is the same for both solutions, because the thresholds $N_{th,1}$ and $N_{th,3}$ of these tenants are the same. Instead, when comparing the PRB utilization of services 1 and 2 of Tenant 2 in *Solution A* and *B*, significant differences are obtained. In fact, the PRB utilization for *Solution B* is 66.6% lower than the one obtained for *Solution A* in both RMa and UMi, which matches the reduction of Tenant 2 $N_{th,2}$ threshold (i.e., from $N_{th,2} = 15.3$ PRBs in *Solution A* to $N_{th,2} = 5.1$ PRBs in *Solution B*). In terms of the PRB utilization of the Tenant 3 services, higher utilization is obtained for service 1 than for service 2, due to the fact that the GBR requirement of the former is twice the GBR requirement of the latter and the session generation of service 1 is also higher.

The decrease of Tenant 2 PRB utilization has as a consequence a reduction in its aggregated throughput $\overline{Th_{s,n}}$, as shown in Fig. 12, which shows the average aggregated throughput for non-GBR services with *Solutions A* and *B*. The performance of service 3 of Tenant 1 is the same for both solutions given that Tenant 1 thresholds remain unchanged, while Tenant 2 services present a 66.6% reduction with *Solution B* in comparison to *Solution A*, which is the same decrement observed in the PRB utilization. Therefore, the realization of *Solution B* may be subject to a renegotiation of the Service Level Agreement terms with Tenant 2.

Regarding GBR services, the aggregated average throughput is shown in Table 7. In this case, as GBR values are provided almost always and low degradation probabilities are achieved for all services, the resulting throughput $\overline{Th_{s,n}}$ is approximately the product of the average number of users $\overline{U_{s,n}}$ of each GBR service and the GBR value. Since the $C_{max,n}$ of Tenant 1 and Tenant 3 have not been changed for

Solutions A and *B*, the obtained $\overline{U_{s,n}}$ and $\overline{Th_{s,n}}$ of each service is the same, so Table 7 does not present separate results for each solution. Besides, it is also observed that very small differences are obtained in the results of the UMi and RMa environments.

VII. CONCLUSION

This paper has presented a Markov model that characterizes the resource sharing in RAN slicing scenarios, where multiple tenants provide GBR and non-GBR services. The model can include diverse RRM functions in terms of admission control at layer 3 and radio resource allocation at layer 2. In terms of admission control, which determines the transition probabilities between the different states in the model, a slice-aware admission control policy has been selected. The model has also been provided with a slicing-aware resource allocation procedure that considers variable propagation conditions by deriving the pdfs of both the required and assigned resources according to the service's QoS parameters.

The effect of the considered RRM functions has been studied by evaluating different performance metrics (blocking probability, degradation probability, throughput, occupation, etc.) in a scenario considering two tenants providing GBR and non-GBR services in both urban micro cell and rural macrocell environments. Based on the considered scenario, the analytical model's suitability has been validated given the low percentage errors obtained (i.e. maximum relative errors of 3%) when comparing the model's results with the ones obtained with a system level simulator. The performance analysis conducted at state and system levels, as well as the analysis of the introduction of a new tenant, has revealed that (i) the considered admission control policy and resource allocation procedure are able to achieve isolation between the different slices, so that overload situations in one slice do not affect the performance of GBR users of the other slice while preserving the maximum capacity allowed to each of the slices; (ii) ARP priorities are respected by providing better performance to those GBR services with lower ARP (higher priority); (iii) GBR services are provided with negligible degradation rates, which implies that the requested GBR values are provided to the admitted users in the system; (iv) non-GBR services are provided with the remaining resources after the allocation to GBR services according to its priority level, so that better performance is given to those non-GBR services with higher priority (lower priority level); (v) The model is able to capture the radio propagation effects, enabling the analysis of different performance metrics in 5G environments of interest; (vi) The introduction of new tenants into the system can be performed by re-configuring the maximum capacity and maximum PRBs to be provided to each of the tenants, although the re-configuration of these values may impact on the performance of already operating tenants if the total amount of PRBs is not modified.

Given the potential of the proposed model, a possible future extension of the resource allocation procedure at layer 2 can be the joint optimization of the performance of different tenants by distributing the unused PRBs of one tenant among the existing users of other tenants.

REFERENCES

[1] R. Hattachi, Ed., "5G White paper, version 1," NGMN Alliance, Boston, MA, USA, White Paper, Feb. 2015. [Online]. Available: https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf

[2] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: Enabling enterprises' own software-defined cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 146–153, Jul. 2016.

[3] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 80–87, May 2017.

[4] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 84–91, May 2016.

[5] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[6] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.

[7] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

[8] D. M. Gutierrez-Estevez, O. Bulakci, M. Ericson, A. Prasad, E. Pateromichelakis, J. Belschner, P. Arnold, and G. Calochira, "RAN enablers for 5G radio resource management," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Helsinki, Finland, Sep. 2017, pp. 1–6.

[9] *System Architecture for the 5G System; Stage 2 (Release 15)*, document 3GPP TS 23.501 V16.0.2, Apr. 2019.

[10] *NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)*, document 3GPP TS 38.300 V15.5.0, Apr. 2019.

[11] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[12] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[13] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[14] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Pérez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[15] O. Sallent, J. Pérez-Romero, R. Ferrús, and R. Agustí, "On radio access network slicing from a radio resource management perspective," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 166–174, Oct. 2017.

[16] B. Khodapanah, A. Awada, I. Viering, D. Oehmann, M. Simsek, and G. P. Fettweis, "Fulfillment of service level agreements via slice-aware radio resource management in 5G networks," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Porto, Portugal, Jun. 2018, pp. 1–6.

[17] R. Ferrús, O. Sallent, J. Pérez-Romero, and R. Agustí, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, May 2018.

[18] E. Pateromichelakis and C. Peng, "Selection and dimensioning of slice-based RAN controller for adaptive radio resource management," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[19] E. Pateromichelakis and K. Samdanis, "A graph coloring based inter-slice resource management for 5G dynamic TDD RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[20] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.

[21] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, and X. Costa-Perez, "A machine learning approach to 5G infrastructure market optimization," *IEEE Trans. Mobile Comput.*, vol. 19, no. 3, pp. 498–512, Mar. 2019.

[22] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.
[23] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2017, pp. 127–140.
[24] J. Epperlein and J. Marecek, "Resource allocation with population dynamics," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2017, pp. 1293–1300.
[25] S. Jagannatha, N. S. Shravan, and S. Kavya, "Cost performance analysis: Usage of resources in cloud using Markov-chain model," in *Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Coimbatore, Indian, Jan. 2017, pp. 1–8.
[26] H. Y. Ng, K. T. Ko, and K. F. Tsang, "3G mobile network call admission control scheme using Markov chain," in *Proc. 9th Int. Symp. Consum. Electron. (ISCE)*, Jun. 2005, pp. 276–280.
[27] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 10, pp. 1257–1270, Oct. 2008.
[28] V. V. Paranthaman, Y. Kirsal, G. Mapp, P. Shah, and H. X. Nguyen, "Exploring a new proactive algorithm for resource management and its application to wireless mobile environments," in *Proc. IEEE 42nd Conf. Local Comput. Netw. (LCN)*, Singapore, Oct. 2017, pp. 539–542.
[29] S. Al-Rubaye, A. Al-Dulaimi, J. Cosmas, and A. Anpalagan, "Call admission control for non-standalone 5G ultra-dense networks," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1058–1061, May 2018.
[30] Y. Kim and S. Park, "Analytical calculation of spectrum requirements for LTE-A using the probability distribution on the scheduled resource blocks," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 602–605, Mar. 2018.
[31] B. Han, D. Feng, and H. D. Schotten, "A Markov model of slice admission control," *IEEE Netw. Lett.*, vol. 1, no. 1, pp. 2–5, Mar. 2019.
[32] M. N. Patwary, R. Abozariba, and M. Asaduzzaman, "Multi-operator spectrum sharing models under different cooperation schemes for next generation cellular networks," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Toronto, ON, Canada, Sep. 2017, pp. 1–7.
[33] S. Lin, L. Kong, Q. Gao, M. K. Khan, Z. Zhong, X. Jin, and P. Zeng, "Advanced dynamic channel access strategy in spectrum sharing 5G systems," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 74–80, Oct. 2017.
[34] K. B. Ali, M. S. Obaidat, F. Zarai, and L. Kamoun, "Markov model-based adaptive CAC scheme for 3GPP LTE femtocell networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 6924–6928.
[35] A. Jee, S. Hoque, B. Talukdar, and W. Arif, "Analysis of link maintenance probability for cognitive radio ad hoc networks," in *Proc. 5th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Noida, India, Feb. 2018, pp. 385–389.
[36] I. Vila, O. Sallent, A. Umbert, and J. Pérez-Romero, "Guaranteed bit rate traffic prioritisation and isolation in multi-tenant radio access networks," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Barcelona, Spain, Sep. 2018.
[37] I. Vila, O. Sallent, A. Umbert, and J. Pérez-Romero, "An analytical model for multi-tenant radio access networks supporting guaranteed bit rate services," *IEEE Access*, vol. 7, pp. 57651–57662, Apr. 2019.
[38] *Management and orchestration; Provisioning; (Release 16)*, document 3GPP TS 28.531 v16.2.0 Jun. 2019.
[39] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2018.
[40] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, "Self-optimised admission control for multitenant radio access networks," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.
[41] *Management and Orchestration; 5G Performance Measurements (Release 16)*, document 3GPP TS 28.552 v16.2.0 Jun. 2019.
[42] *Management and Orchestration; Architecture Framework (Release 16)*, document 3GPP TS 28.533 v15.1.0, Jun. 2019.
[43] *NR; Physical Layer Procedures for Data (Release 15)*, document 3GPP TS 38.214 v15.6.0, Jun. 2019.
[44] *Study on Channel Model for Frequencies From 0.5 to 100 GHz (Release 15)*, document 3GPP TS 38.214 v15.6.0, Mar. 2019.
[45] *NR; Base Station (BS) Radio Transmission and Reception (Release 15)*, document 3GPP TS 38.104 v15.5.0, Apr. 2019.
[46] *NR; Physical Channels and Modulation (Release 15)*, document 3GPP TS 38.211 v15.3.0, Sep. 2018.
[47] *Study on New Radio Access Technology: Radio Frequency (RF) and co-Existence Aspects (Release 14)*, document 3GPP TR 38.803 v14.2.0, Sep. 2017.
[48] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*. Princeton, NJ, USA: Princeton Univ. Press, 1994.

IRENE VILÀ received the B.E. degree in telecommunication systems engineering and the M.E. degree in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2015 and 2017, respectively. She is currently pursuing the Ph.D. degree with the Mobile Communication Research Group (GRCM), Department of Signal Theory and Communications (TSC), UPC, supported with an FI AGAUR Grant by the Government of Catalunya. In 2018, she joined the Mobile Communication Research Group (GRCM), Department of Signal Theory and Communications (TSC), UPC. Her current research interests include RAN Slicing, radio resource management, software defined networking (SDN), and network function virtualization (NFV), concepts to be included in new 5G technologies.

JORDI PÉREZ-ROMERO (Member, IEEE) is currently a Professor with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. He has been working in the field of wireless communication systems, with particular focus on radio resource management, cognitive radio networks, and network optimization. He has been involved in different European Projects and in projects for private companies. He has published more than 200 articles in international journals and conferences.

ORIOL SALLENT is currently a Professor with the Universitat Politècnica de Catalunya (UPC). He has participated in a wide range of European projects with diverse responsibilities as the Workpackage Leader and a Coordinator partner and contributed to standardization bodies, such as 3GPP, IEEE, and ETSI. He has published more than 200 articles mostly in IEEE journals and conferences. His research interests include cognitive management in cognitive radio networks, self-organizing networks, radio network optimization, and QoS provisioning in heterogeneous wireless networks.

ANNA UMBERT received the Engineering and Ph.D. degrees in telecommunications from the Universitat Politècnica de Catalunya (UPC), in 1998 and 2004, respectively. In 2001, she joined UPC as an Assistant Professor, and became an Associate Professor, in 2017, which is her current status. Since 1997, she has been participating in several projects founded by both public and private organizations. She has published more than 50 articles in international journals and conferences. Her research interests are focused in radio resource and QoS management in the context of heterogeneous wireless networks, cognitive management in cognitive radio networks, dynamic spectrum access and management, self-organized networks, and network optimization.