IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 29, 2019, accepted April 16, 2019, date of publication April 25, 2019, date of current version May 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913323

An Analytical Model for Multi-Tenant Radio Access Networks Supporting Guaranteed Bit Rate Services

IRENE VILÀ, ORIOL SALLENT, ANNA UMBERT, AND JORDI PÉREZ-ROMERO
Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain

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

This work was supported in part by the EU funded H2020 5G-PPP project 5G ESSENCE under Grant 761592, in part by the Spanish Research Council and FEDER funds under SONAR 5G Grant (ref. 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 2018FI_B_00412.

ABSTRACT Network slicing is a key feature of forthcoming fifth generation (5G) systems to facilitate the partitioning of the network into multiple logical networks customized according to different business and application needs. Network slicing is a fundamental capability for enabling a cost-effective deployment and operation of 5G, as it allows the materialization of multi-tenant networks in which the same infrastructure is shared among multiple communication providers, each one using a different slice. This paper proposes a Markovian approach to characterize the resource sharing in multi-tenant scenarios with diverse guaranteed bit rate services by considering a slice-aware admission control policy. After describing the Markov model and its implementation and discussing its suitability, the model is applied to study the performance attained in a scenario with two different slices, one for enhanced mobile broadband communications and the other for mission critical services. The system is analyzed under standard and disaster situations, thus illustrating the capability to properly manage the different multi-tenant and multi-service traffic loads.

INDEX TERMS Admission control, Markov processes, mobile communication, multi-tenancy, radio access networks, RAN slicing.

I. INTRODUCTION

Fifth Generation (5G) systems target the simultaneous support of a wide range of application scenarios and business models (e.g., automotive, utilities, smart cities, high-tech manufacturing) [1]. Partnerships will be established on multiple layers 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.

The sharing of mobile network infrastructure among multiple communication providers, denoted as ''tenants'', is one of the main characteristics of the future architectures of mobile networks, since the sharing process will reduce capital and operational costs [2]. Multi-tenancy can be materialized through network slicing capabilities [3], which enable logical networks/partitions to be created (i.e., network slices). In this way, self-contained networks considered as individual networks are conformed and provided with appropriate isolation and optimized characteristics to serve a particular purpose or service category (e.g., applications with different access and/or functional requirements) or even individual customers (e.g., enterprises, third party service providers). This is especially relevant for the Radio Access Network (RAN), which is the most resource-demanding (and costliest) part of the mobile network and the most challenged by the support of network slicing [4].

System architecture and functional aspects to support network slicing in the 5G Core Network (5GC) and in the Next-Generation RAN (NG-RAN) have already been defined by 3GPP [5], [6]. Moreover, the 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 [7], [8] to algorithms for dynamic resource sharing across slices [9]. Similarly, [10] analyses the RAN slicing problem in a multi-cell network in relation to Radio Resource Management (RRM) functionalities.

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

In turn, [11] 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. Some other works focus on the network slice admission control for slice requests that need to support a given number of users for a certain time in the RAN, such as [12], [13], which target to optimize the infrastructure provider's revenue, or [14], which optimizes the network utilization by incorporating traffic forecasting capabilities.

In this context, this paper proposes a Markov model to characterize and assess the performance of RAN slicing in multi-service and multi-tenant scenarios. Markovian approaches have been widely used to characterize the utilization of resources in many fields, such as mobility [15], cloud computing [16], Call Admission Control (CAC) for 3G [17] and 4G femtocells [18] or for a heterogeneous network's Radio Access Technologies (RAT) policies [19]. More recently, works in the field of 5G exploit Markov modeling to approach a proactive resource allocation scheme in highly mobile networks [20] and the management of admission control for handoff requests between small-cell and macrocell domains [21]. Markov chain models have also been considered in [22], [23] for characterizing spectrum sharing schemes.

Nevertheless, none of the above papers have considered the use of Markov chain models to study the performance of different Admission Control (AC) policies performed at user level for RAN slicing, which constitutes one of the novelties of this work, as a difference from e.g. [12]–[14], which have dealt with the admission of slices. In this respect, this paper presents an analytical Markov chain model approach considering multi-tenant and multi-service scenarios that allows assessing the behavior of AC policies applied for RAN slicing. Another novelty is that the model includes the definition of services in terms of its Guaranteed Bit Rate (GBR) and its Allocation and Retention Priority (ARP) indicator that are part of the 5G QoS (Quality of Service) profile of New Radio (NR). These QoS parameters play a fundamental role in the definition of the AC and the resource allocation functions included in the model. A first version of this model was presented in our recent work [24], while in this paper, several advances and novel contributions are provided with respect to this prior work: (1) The model's implementation is elaborated and described in detail, (2) The theoretical model is validated against a wide range of scenarios that include different propagation conditions, mobile speeds and traffic loads, which are essential to delineate the model's suitability, (3) The resource allocation procedure is elaborated and described in detail, (4) A relevant use case envisaged for 5G is addressed through the presented theoretical framework, considering enhanced mobile broadband (eMBB) and mission critical (MC) services provided by different tenants and analyzing its behavior under standard and disaster traffic conditions. Supported by the definition of additional metrics to be extracted from the analytical model, the evaluation conducted enables highlighting the potentials and usefulness of RAN slicing in multi-tenant 5G scenarios.

The rest of the paper is organized as follows. Section II presents the proposed system model, describing the analytical Markov chain approach and introducing the considered slicing-aware AC policy and resource allocation criterion. Section III presents different performance metrics that can be extracted from the model. Section IV presents the example scenario considered for 5G RAN slicing, describes the model implementation and validation and provides the performance results. Finally, Section V summarizes the conclusions.

II. SYSTEM MODEL

A multi-sliced RAN scenario comprised of N tenants is assumed, each of them operating in a RAN slice of a common infrastructure and sharing the same resources. The n-th tenant provides M_n service types, each one with specific QoS requirements. This paper assumes GBR services, whose QoS profile is given by the GBR value (i.e., the bit rate to be provided to the user, also referred to as Guaranteed Flow Bit Rate (GFBR) in 5G NR terminology) and the ARP indicator [5], which defines the relative importance of the service requesting for resources and starts from 1 (highest priority) onwards (for successive lower priority services). Therefore, let us denote as $GBR_{s,n}$ and $ARP_{s,n}$ the GBR and ARP values, respectively, of the s-th service of the n-th tenant for $s = 1, \cdots, M_n$ and $n = 1, \cdots, N$.

Let us assume 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. Then, when a user generates a new session, an AC mechanism is needed to decide whether the new request can be accepted in the system or not, depending on the available capacity, the GBR requirements and the corresponding ARP. The capacity is defined on a per-tenant basis, where $C_{max,n}$ is the established capacity for tenant n and is measured as the maximum aggregate GBR that can be admitted for all the users of this tenant. The number of admitted users in the system for the s-th service of the n-th tenant is denoted as $u_{s,n}$.

Assuming that users generate sessions according to a Poisson arrival process and these sessions have an exponential duration, 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 $(M_1 + M_2 + \cdots + M_N)$-dimensional states, considering all the services of the N tenants in the system. 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. Transitions between the different states within the Markov Chain 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 is in charge of admitting or rejecting user requests depending on the system occupation, it also affects the transitions between states.

In this respect, let us define $AC^{s,n}_{(u_{1,1}...,u_{M_N,N})}$ 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.

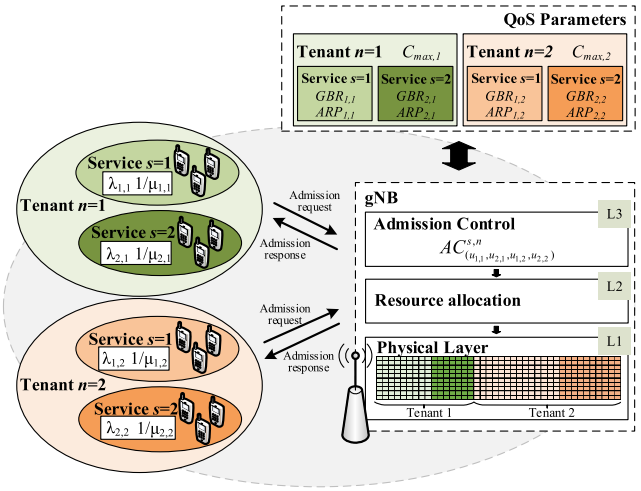

FIGURE 1. System model conceptual scheme.

To describe the considered system in terms of involved functionalities, corresponding protocol layers and mapping to physical entities, Fig. 1 illustrates an example scenario with $N = 2$ tenants, each one with 2 services (i.e., $M_1 = 2$ and $M_2 = 2$).

Whenever a user generates a new session, an "Admission Request" is sent to the gNB. The AC function is associated with the layer 3 (L3) and implemented at the gNB. As a result of the AC decision making process, the gNB replies to the user with an "Admission response" message (i.e., accepted or rejected). At layer 2 (L2), the resource allocation function is in charge of assigning the available PRBs in layer 1 (L1) among the admitted users in accordance with their expected GBR value.

Based on the above, the following subsections present the proposed Markov chain model as well as the admission control and resource allocation procedures.

A. MARKOV CHAIN MODEL DEFINITION

The states that compose the state space are defined as:

$$S = \{S_{(u_{1,1},\ldots,u_{M_N,N})} | \; u_{s,n} \leq \min\left(\left\lfloor \frac{C_{\max,n}}{GBR_{s,n}} \right\rfloor, U_{\max,s,n}\right) \quad \forall s,n\} \tag{1}$$

This definition considers that the number of users of each service is limited by the tenant's maximum capacity $C_{max,n}$, the requested $GBR_{s,n}$ and $U_{max,s,n}$, which is the maximum allowed number of users in a cell for the s-th service of the n-th tenant, established for hardware limitation purposes (i.e., processor, memory, power).

Given the state space, transitions between states occur due to the admission of a new session's arrival by the AC function or the finalization of a session of an admitted user. Therefore, a given state can only change by increasing or decreasing a single user, meaning that transitions are only possible between neighboring states. Fig. 2 depicts an illustrative state transition diagram for the particular case of $N = 2$ tenants, each of them providing 2 different services (i.e., $M_1 = 2$ and $M_2 = 2$).

From the state transition diagram, the general Steady-State Balance Equation (SSBE) is given by:

$$\begin{aligned} &P_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}\Big[\sum_{s,n} u_{s,n}\mu_{s,n} + \\ &\sum_{s,n|S_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{M_N,N})}\in S} \lambda_{s,n} AC^{s,n}_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}\Big] \\ &= \sum_{s,n} P_{(u_{1,1},\ldots,u_{s,n}-1,\ldots,u_{M_N,N})}\lambda_{s,n}\, AC^{s,n}_{(u_{1,1},\ldots,u_{s,n}-1,\ldots,u_{M_N,N})} \\ &+ \sum_{s,n|S_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{M_N,N})}\in S} P_{(u_{1,1},\ldots,u_{s,n}+1,\ldots,u_{M_N,N})} \\ &\times(u_{s,n}+1)\mu_{s,n} \end{aligned} \tag{2}$$

where $P_{(u_{1,1},\ldots,u_{s,n},\ldots,u_{M_N,N})}$ corresponds to the steady-state probability of being in the 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 normalization constraint:

$$\sum_{S_{(u_{1,1},\ldots,u_{M_N,N})}\in S} P_{(u_{1,1},\ldots,u_{M_N,N})} = 1 \tag{3}$$

B. ADMISSION CONTROL

The proposed model can adopt different AC policies to determine the acceptance of a user according to its QoS parameters. In this paper, a slicing-aware AC policy has been selected, which provides isolation in the admission of users of different tenants by guaranteeing a proportion of the available radio resources to each of the tenants so that the admission of users from one tenant does not impact on the other tenant. With this purpose, the admission or rejection decision of a new user from the s-th service of the n-th tenant considers its maximum capacity threshold $C_{max,n}$, the priority $ARP_{s,n}$ indicator and the number of admitted users in the system $(u_{1,1}, \ldots u_{M_N,N})$, according to:

$$AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})} = \begin{cases} 1 & \text{if } \sum_{s'|ARP_{s',n}\leq ARP_{s,n}} u_{s',n}\cdot GBR_{s',n} + GBR_{s,n} \leq C_{\max,n} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

This expression considers that a new user of service s and tenant n can be admitted if the aggregate requested bit rate considering both the new user and the already accepted users of any service s' from the n-th tenant with higher or equal priority than the new user (i.e., with ARP lower or equal to $ARP_{s,n}$) does not exceed the capacity threshold $C_{max,n}$.

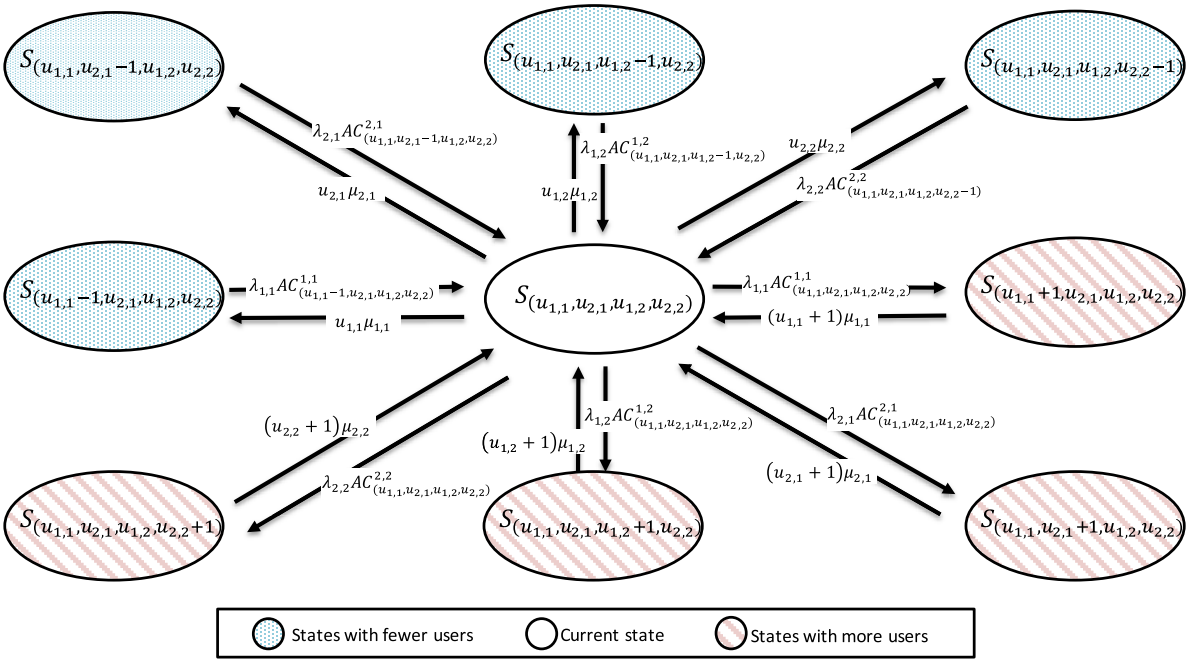

FIGURE 2. State transition diagram in the case of $N = 2$ tenants, each of them providing 2 different services (i.e., $M_1 = 2$ and $M_2 = 2$).

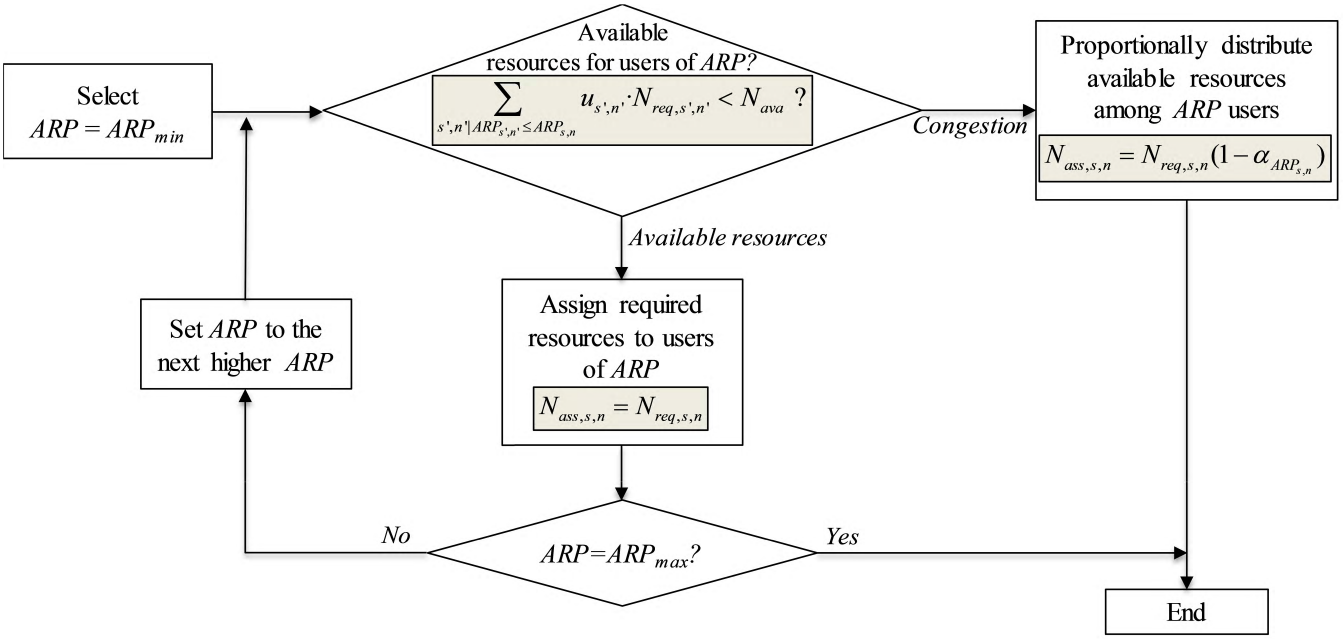

FIGURE 3. Resource allocation procedure.

C. RESOURCE ALLOCATION MODEL

The considered resource allocation model intends to compute the number of assigned PRBs, $N_{ass,s,n}$ to each admitted user of service s and tenant n in a given state in accordance with the requested bit rate $GBR_{s,n}$. Since the resource allocation process is implementation dependent, different criteria can be assumed. In that sense, the considered resource allocation model is illustrated by the procedure of Fig. 3. It takes into account the number of available PRBs in the cell N_{ava}, the $ARP_{s,n}$ indicator and the number of requested resources per user $N_{req,s,n}$ for GBR services, which is given by:

$$N_{req,s,n} = \frac{GBR_{s,n}}{B \cdot S_{eff}} \tag{5}$$

B being the PRB bandwidth and S_{eff} the spectral efficiency associated with the considered radio connection. In general, S_{eff} will be a random variable with a certain statistical distribution. In that respect, for the purpose of this paper and

considering that the resource allocation is modeled on average terms, S_{eff} is considered as the average spectral efficiency that users would observe in the scenario. The implications and accuracy of this assumption are discussed in Section IV.C, as part of the model validation.

For a given number of admitted users in the system, the procedure in Fig. 3 is iteratively done starting by the users of lower ARP to the ones with higher ARP. As long as there are available resources to serve the users of a given ARP, each user gets the required resources $N_{req,s,n}$ and the available number of resources are reduced accordingly before moving to the next ARP. If ARP_{max} is reached, the spare resources remain unused. Instead, when there are not sufficient available resources to serve all the users of a given ARP (i.e., there is congestion), the number of assigned resources $N_{ass,s,n}$ to each user of this ARP is obtained by proportionally reducing $N_{ass,s,n}$ according to the obtained ARP resource excess $\alpha_{ARP_{s,n}}$, given by:

$$N_{ass,s,n} = N_{req,s,n}(1 - \alpha_{ARP_{s,n}}) \tag{6}$$

$$\alpha_{ARP_{s,n}} = \frac{\left(\sum_{s',n'|ARP_{s',n'} \leq ARP_{s,n}} u_{s',n'} \cdot N_{req,s',n'}\right) - N_{ava}}{\sum_{s',n'|ARP_{s',n'}=ARP_{s,n}} u_{s',n'} \cdot N_{req,s',n'}} \tag{7}$$

When congestion occurs for a certain ARP value, the users with this ARP level up to ARP_{max} remain in the system with degraded quality in their connections.

III. PERFORMANCE METRICS

Based on the steady-state probabilities, this section develops the different performance metrics of interest for the evaluation of the considered slicing-aware AC mechanism.

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, denoted as $S^b_{s,n}$, is defined as:

$$S^b_{s,n} = \{S_{(u_{1,1},\cdots,u_{M_N,N})} \in S | AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})} = 0\} \tag{8}$$

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.

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

$$P^b_{s,n} = \sum_{S(u_{1,1},\ldots,u_{M_N,N}) \in S^b_{s,n}} P_{(u_{1,1},\ldots,u_{M_N,N})} \tag{9}$$

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 (9).

B. DEGRADATION PROBABILITY

Another subset of states are the so-called degraded states, in which congestion is reached and some admitted users cannot be assigned with their required resources $N_{req,s,n}$ to provide $GBR_{s,n}$. Instead, they are assigned with a number of resources $N_{ass,s,n} < N_{req,s,n}$, according to the considered resource allocation criteria. The set of degraded states for the s-th service of n-th tenant is expressed as:

$$S^{\deg}_{s,n} = \{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S | N_{ass,s,n} < N_{req,s,n}\} \tag{10}$$

The set of degraded states for the n-th tenant $S^{\deg}_n$ are those states in which the users of at least one service of the tenant are degraded. Therefore, $S^{\deg}_n$ is defined as the union of the degraded states for the services of the n-th tenant, i.e., $S^{\deg}_n = S^{\deg}_{1,n} \cup S^{\deg}_{2,n} \cup \ldots \cup S^{\deg}_{M_n,n}$. Equivalently, the global system degraded states $S^{\deg}$ would be computed as the union of the degraded states of each of the tenants.

By using the previous definitions, the degradation probability per service and tenant is defined as:

$$P^{\deg}_{s,n} = \sum_{S(u_{1,1},\ldots,u_{M_N,N}) \in S^{\deg}_{s,n}} P_{(u_{1,1},\ldots,u_{M_N,N})} \tag{11}$$

This expression can be easily extended to compute the degradation probability per tenant or the global degradation probability by considering $S^{\deg}_n$ or $S^{\deg}$ in the summation, respectively.

C. OCUPANCY 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:

$$\overline{U_{s,n}} = \sum_{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S} u_{s,n} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \tag{12}$$

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 $\overline{N_{ass\_all,s,n}}$ aggregated per service, which can be computed by considering the number of assigned PRB per user $N_{ass,s,n}$, the number of users $u_{s,n}$ and the steady-state probabilities as:

$$\overline{N_{ass\_all,s,n}} = \sum_{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S} u_{s,n} \cdot N_{ass,s,n} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \tag{13}$$

Accordingly, the average PRB utilization per tenant would result from $\overline{N_{ass\_all,n}} = \overline{N_{ass\_all,1,n}} + \overline{N_{ass\_all,2,n}} + \ldots + \overline{N_{ass\_all,M_n,n}}$ while the global system PRB utilization $\overline{N_{ass\_all}}$ 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 $\overline{\omega_{ass\_all,s,n}}$ may be expressed as:

$$\overline{\omega_{ass\_all,s,n}} = \frac{\overline{N_{ass\_all,s,n}}}{N_{ava}} \tag{14}$$

Similarly to the other metric computations, the average normalized PRB utilization per tenant $\overline{\omega_{ass\_all,n}}$ and the global average normalized PRB utilization $\overline{\omega_{ass}}$ result from adding the average normalized PRB utilization of the tenant's services or all the services, respectively.

D. AVERAGE AGGREGATED THROUGHPUT

The average aggregated throughput $\overline{R_{s,n}}$ for the s-th service of the n-th tenant can be computed by considering the steady-state probabilities as:

$$\overline{R_{s,n}} = \sum_{S_{(u_{1,1},\ldots,u_{M_N,N})} \in S} u_{s,n} \cdot N_{ass,s.n} \cdot B \cdot S_{eff} \cdot P_{(u_{1,1},\ldots,u_{M_N,N})} \tag{15}$$

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

IV. PERFORMANCE EVALUATION

This section presents a performance analysis using the proposed Markov model in different scenarios based on practical use cases. It also includes details on the model implementation and on the model validation by contrasting the Markov model results with a system-level simulator.

A. CONSIDERED SCENARIO

The scenario under test considers a commercial operator that has deployed a NG-RAN in order to provide eMBB services to its users. Meanwhile, the commercial operator leases its infrastructure to a public safety operator serving MC communications. Therefore, the NG-RAN serves two different tenants, referring the commercial operator as *Tenant* 1 and the public safety operator as *Tenant* 2. In addition, each operator provides two different GBR services: Tenant 1 includes two video profiles, a basic profile with a standard quality and a premium profile with High Definition (HD) quality, whereas Tenant 2 provides two MC services, namely, MC Video and MC Push to Talk (MC PTT).

For each of the services, the QoS parameters summarized in Table 1 have been specified according to the QoS model of [5]. The included parameters consist of the ARP and the GFBR, which specifies the GBR value to be provided to a QoS flow.

TABLE 1. Services per tenant.

Tenant	Tenant id (n)	Service	Service id (s)	Type	ARP	GFBR
Commercial operator	1	Premium – Video HD	1	GBR	2	10 Mb/s
		Basic - Video	2	GBR	3	3 Mb/s
Public safety operator	2	MC Video	1	GBR	2	5 Mb/s
		MC PTT	2	GBR	1	48 kb/s

The NG-RAN is composed by one gNB, with a single cell operating in a 20 MHz channel composed by 51 PRBs [25], each one of $B = 360$ kHz corresponding to 12 Orthogonal Frequency-Division Multiple Access (OFDMA) subcarriers with subcarrier separation of 30 kHz, which is one of the numerologies defined for 5G NR [26]. The configured parameters used to obtain the different results included in this section are summarized in Table 2.

Two different scenarios are defined in terms of load: a standard scenario, in which the traffic coming from the public safety operator is low while the traffic from the commercial operator is progressively increased, and a scenario representing a situation when an emergency or disaster has occurred and therefore the public safety traffic generated is varied until reaching high values whereas the traffic from the commercial operator remains moderate. Different maximum tenant capacity thresholds $C_{max,n}$ are configured: for the standard scenario $C_{max,1}$ of the commercial operator is higher whilst in the disaster scenario it is reduced, allowing more resources to be devoted to the public safety operator (i.e., $C_{max,2}$ higher).

B. MODEL IMPLEMENTATION

Regarding the implemented method to compute the state probabilities and solving the SSBE, the model has been approached as a CTMC, so the following equation needs to be solved:

$$\underline{\pi}^T \cdot \mathbf{Q} = \underline{\mathbf{0}} \tag{16}$$

where $\underline{\pi}$ is a column vector containing all the steady-state probabilities $P_{(u_{1,1},\ldots,u_{M_N,N})}$, the superscript T denotes the transposed operation, and $\underline{\mathbf{0}}$ is a row vector with all elements equal to 0. The matrix $\mathbf{Q}$ is the state transition rate matrix, considering that rows define the origin state x and the columns the destination state y.The elements of $\mathbf{Q}$, referred to $q_{x,y}$, are defined as follows:

$$q_{x,y} = \begin{cases} \lambda_{s,n} AC^{s,n}_{(u_{1,1},\ldots,u_{M_N,N})} & \text{if } x = S_{(u_{1,1},\ldots u_{s,n}\ldots,u_{M_N,N})} \text{ and} \\ & y = S_{(u_{1,1},\ldots u_{s,n}+1\ldots,u_{M_N,N})} \\ u_{s,n}\mu_{s,n} & \text{if } x = S_{(u_{1,1},\ldots u_{s,n}\ldots,u_{M_N,N})} \text{ and} \\ & y = S_{(u_{1,1},\ldots u_{s,n}-1\ldots,u_{M_N,N})} \\ -\sum_{z \in S, z \neq x} q_{x,z} & \text{if } x = y \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

TABLE 2. Model configuration parameters.

Parameter	Value	
Number of available PRBs (N_{ava})	51 PRB	
PRB Bandwidth (B)	360 kHz	
Spectral Efficiency (S_{eff})	8.5 b/s/Hz	
Data rate per PRB	3 Mb/s/PRB	
Cell Maximum number of users	$U_{max,s,n}$=50 users for s,n=1,2	
Cell total capacity	156 Mb/s	
Scenario 1: Standard		
Average session generation rate	Tenant 1	varied from 0.001 to 0.12 sessions/s (corresponds to a variation of 0.6Mb/s to 82.8 Mb/s)
	Tenant 2	0.05 session/s (corresponds to 6.2 Mb/s)
Average session duration	120 s	
Maximum capacity threshold	Tenant 1	$C_{max,1}$=93.6Mb/s (corresponds to the 60% of the total capacity)
	Tenant 2	$C_{max,2}$=31.2 Mb/s (corresponds to the 20% of the total capacity)
Tenant generation distribution	Tenant 1	40% of session arrivals are of service 1 and 60% of service 2
	Tenant 2	20% of session arrivals are of service 1 and 80% of service 2
Scenario 2: Disaster		
Average session generation rate	Tenant 1	0.04 session/s (corresponds to 17.76 Mb/s)
	Tenant 2	varied from 0.001 to 0.25 sessions/s (corresponds to a variation of 0.36 Mb/s to 88 Mb/s)
Average session duration	120 s	
Maximum capacity threshold	Tenant 1	$C_{max,1}$=46.8 Mb/s (corresponds to the 30% of the total capacity)
	Tenant 2	$C_{max,2}$=78 Mb/s (corresponds to the 50% of the total capacity)
Tenant generation distribution	Tenant 1	10% of session arrivals are of service 1 and 90% of service 2
	Tenant 2	60% of session arrivals are of service 1 and 40% of service 2

The Gauss-Seidel method [27] has been selected to solve the SSBE system of equations (16) and compute the steady-state probabilities. This method avoids the discretization of the CTMC transition rate matrix and provides a good compromise between accuracy and complexity in comparison to other methods. By employing the Gauss-Seidel method, expression (16) is transposed, leading to $\mathbf{Q^T}\cdot\underline{\boldsymbol{\pi}} = \underline{\mathbf{0}}^{\mathbf{T}}$ and then the matrix $\mathbf{Q^T}$ is decomposed as:

$$\mathbf{Q^T} = \mathbf{D} - (\mathbf{L} + \mathbf{U}) \tag{18}$$

where $\mathbf{D}$ matrix is the diagonal of $\mathbf{Q^T}$ while $\mathbf{L}$ and $\mathbf{U}$ are, respectively, the strictly lower and upper triangular matrices of $\mathbf{Q^T}$. Then, the iterative matrix $\mathbf{H_{GS}}$ can be constructed according to:

$$\mathbf{H_{GS}} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U} \tag{19}$$

Based on this matrix, the Gauss-Seidel method applies an iterative method that computes the state's probabilities in each iteration until reaching convergence. Specifically, the state probability at the k-th iteration, denoted as $\underline{\pi}^{(\mathbf{k})}$, is computed as:

$$\underline{\pi}^{(\mathbf{k+1})} = \mathbf{H_{GS}}\underline{\pi}^{(\mathbf{k})} \tag{20}$$

where the initial probability vector $\underline{\pi}^{(\mathbf{0})}$ is set randomly. Convergence is achieved at the first iteration k that fulfills the following condition based on the norm of successive iterates:

$$\left\| \underline{\pi}^{(\mathbf{k})} - \underline{\pi}^{(\mathbf{k-1})} \right\| < \varepsilon \tag{21}$$

where ε is the desired accuracy, which in our case is set to 10^{-5}.

C. MODEL VALIDATION

The results of the Markov model implementation have been compared with the output of a system-level simulator that allows defining realistic scenarios where different environments can be configured in terms of cell deployment, propagation conditions and mobility. User's sessions are generated following a Poisson distribution, while session durations are modeled by an exponential distribution. The AC and resource allocation procedures implemented in the simulator follow the same principles as in the Markov model, thus placing the focus on assessing the impact that propagation, velocity and traffic load have in terms of predicted performance metrics. The simulator operates on a discrete-time basis with time steps of 1 s. Simulation statistics are measured by averaging discrete samples taken during the simulation time.

To study the validity of the Markov model, the simulator has been configured according to the parameters of the scenario 1 described in Table 2. Three different environments are considered: Urban Micro-cell (UMi), Urban Macro-cell (UMa) and Rural Macro-cell (RMa) [28]. The configuration of each environment is detailed in Table 3. Additionally, pedestrian (3 km/h), urban (30 km/h) and high-speed (120 km/h) mobility patterns have been studied, where the User Equipment (UE) position is updated every 5 s following a random-walk model. The simulation duration has been set to ensure the observation of at least 100 blocking events for each service along a simulation (except for the MC PTT service that does not experience blockings in any of the considered cases). In Table 4, the results obtained through both the simulator and the Markov model (values in parenthesis) are compared.

In terms of blocking probability, very few differences are found between the Markov model and the simulator for all the studied environments. Indeed, the main assumption affecting the accuracy of the Markov model is, as discussed in

TABLE 3. Environment's cell configuration.

Environment	UMi	Uma	RMa
ISD (Inter-Site distance)	200 m	500 m	1735 m
gNB height	10 m	25 m	35 m
UE height	1.5 m	1.5 m	1.5 m
Minimum gNB-UE distance	10 m	35 m	35 m
Path Loss and Shadowing model	Model of sec. 7.4 of [28].		
Shadowing standard deviation in Line of Sight (LOS)	4	4	4
Shadowing standard deviation in Non-Line of Sight (NLOS)	7.82	6	8
Frequency	3.6 GHz		
Total gNB transmitted power	44 dBm	49 dBm	52 dBm
gNB antenna Gain	Omnidirectional antenna with 5 dBi gain		
UE noise figure	9 dB		
Link-level model to map Signal to Interference and Noise Ratio (SINR) and bit rate	Model in section A.F of [29] with maximum spectral efficiency of 9.96 b/s/Hz (corresponding to SINR=30 dB) and minimum SINR=-10 dB		
Average Spectral Efficiency	8.5 b/s/Hz	6.9 b/s/Hz	6.3 b/s/Hz
Cell total capacity	156 Mb/s	126.7 Mb/s	115.6 Mb/s
Maximum Capacity Threshold Tenant 1 ($C_{max,1}$)	93.6 Mb/s	76 Mb/s	69.4 Mb/s
Maximum Capacity Threshold Tenant 2 ($C_{max,2}$)	31.2 Mb/s	25.3 Mb/s	23.1 Mb/s
Simulation duration λ_1 =0.05	$15 \cdot 10^6$ s	$2.5 \cdot 10^6$ s	$8 \cdot 10^5$ s
Simulation duration λ_1 =0.1	$7.5 \cdot 10^6$ s	$1.25 \cdot 10^6$ s	$4 \cdot 10^5$ s

Section II.C, the consideration of S_{eff} as the average spectral efficiency that users would observe in the environment. Correspondingly, the comparison in terms of PRB utilization shows more accurate results for those environments with lower variations on the spectral efficiency (i.e., lower cell range such as, e.g., UMi). Similarly, for the aggregated throughput per slice, the biggest discrepancies are found for the RMa environment where users can experience high spectral efficiency variability.

From the mobility point of view, under a given environment higher speeds present a better match in the throughput and PRB utilization results obtained with the simulator and the Markov model. This is because in environments with higher speed, more variated samples in terms of user position are obtained, which contributes to reach more averaged results, presenting more similarities to the ones obtained through the Markov model.

Bearing the above results in mind, it is concluded that the theoretical model provides accurate performance results in a good number of representative realistic environments. For environments with larger fluctuation on S_{eff}, the model could be further extended by considering S_{eff} as a random variable with a certain statistical distribution. Due to the non-negligible complexity of this extension, this is left for future work.

D. PERFORMANCE RESULTS

This section includes the comparison between the standard and disaster scenarios specified in Table 2, based on the analysis of different performance metrics obtained by means of the Markov model. Specifically, the impact of the ARP indicator and the tenant capacity thresholds of the selected AC policy on the system performance under different load conditions is discussed. According to the configured parameters, during standard conditions the traffic associated with

TABLE 4. Comparison of results obtained via the system-level simulator in different scenarios and the markov model (in parentheses).

Environment	Speed (km/h)	λ Tenant 1	PRB utilisation(%)				Blocking probability (%)				Throughput per slice (Mb/s)	
			eMBB-Premium video	eMBB-Basic Video	MC-Video	MC-PTT	eMBB-Premium video	eMBB-Basic Video	MC-Video	MC-PTT	eMBB	MC
UMi	3	0.05	16.56 (15.36)	7.45 (6.92)	4.13 (3.84)	0.16 (0.15)	0.06 (0.06)	0.21 (0.21)	0.12 (0.13)	0 (0)	34.96 (34.78)	6.25 (6.23)
		0.1	32.02 (29.79)	13.19 (12.23)	4.17 (3.85)	0.16 (0.15)	3.12 (3.15)	11.67 (11.68)	0.15 (0.13)	0 (0)	65.83 (65.58)	6.3 (6.25)
	30	0.05	15.76 (15.36)	7.09 (6.92)	3.93 (3.84)	0.15 (0.15)	0.06 (0.06)	0.21 (0.21)	0.12 (0.13)	0 (0)	34.96 (34.78)	6.25 (6.23)
		0.1	30.46 (29.79)	12.56 (12.23)	3.96 (3.85)	0.15 (0.15)	3.12 (3.15)	11.67 (11.68)	0.15 (0.13)	0 (0)	65.83 (65.58)	6.3 (6.25)
UMa	3	0.05	21.49 (18.79)	9.63 (8.42)	5.47 (4.7)	0.21 (0.18)	0.96 (0.83)	1.36 (1.37)	0.98 (0.86)	0 (0)	34.58 (34.48)	6.3 (6.18)
		0.1	38.71 (33.79)	15.27 (13.21)	5.31 (4.7)	0.21 (0.18)	10.57 (10.81)	22.52 (22.58)	0.85 (0.87)	0 (0)	59.99 (59.54)	6.12 (6.19)
	30	0.05	16.56 (15.36)	7.45 (6.92)	4.13 (3.84)	0.16 (0.15)	0.06 (0.06)	0.21 (0.21)	0.12 (0.13)	0 (0)	34.96 (34.78)	6.25 (6.23)
		0.1	32.02 (29.79)	13.19 (12.23)	4.17 (3.85)	0.16 (0.15)	3.12 (3.15)	11.67 (11.68)	0.15 (0.13)	0 (0)	65.83 (65.58)	6.3 (6.25)
RMa	120	0.05	15.76 (15.36)	7.09 (6.92)	3.93 (3.84)	0.15 (0.15)	0.06 (0.06)	0.21 (0.21)	0.12 (0.13)	0 (0)	34.96 (34.78)	6.25 (6.23)
		0.1	30.46 (29.79)	12.56 (12.23)	3.96 (3.85)	0.15 (0.15)	3.12 (3.15)	11.67 (11.68)	0.15 (0.13)	0 (0)	65.83 (65.58)	6.3 (6.25)

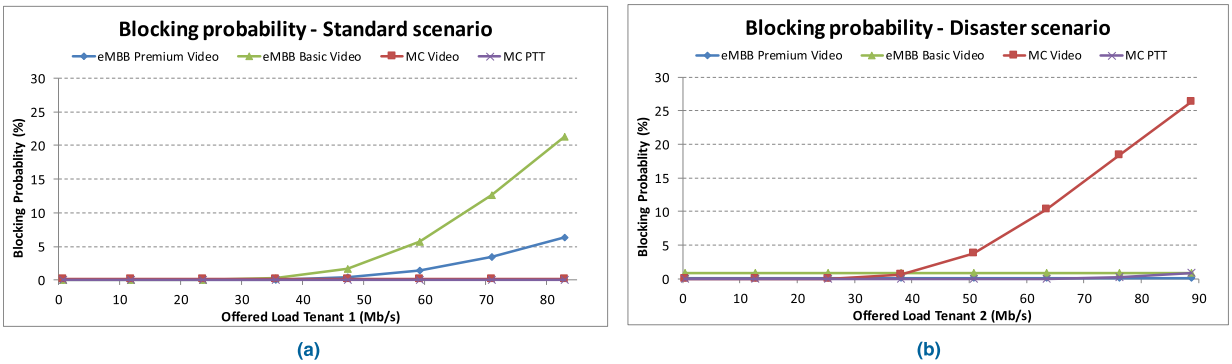

FIGURE 4. Blocking probability of each service in (a) standard and (b) disaster scenarios.

public safety is relatively low and the system is configured to devote higher capacity to commercial traffic (i.e., $C_{max,1} = 93.6$ Mb/s and $C_{max,2} = 31.2$ Mb/s). In turn, during the emergency, both the public safety and the commercial traffic are significant, and the system is configured to devote higher capacity to public safety (i.e., $C_{max,1} = 46.8$ Mb/s and $C_{max,2} = 78$ Mb/s).

Fig. 4 depicts the blocking probabilities for the different services in the standard (Fig. 4a) and the disaster (Fig. 4b) scenarios as a function of Tenant 1 and Tenant 2 offered loads, respectively, understood as the generated traffic load for each of the tenants. For the standard scenario, it can be observed how eMBB services blocking probabilities grow when Tenant 1's offered load is increased. Higher blocking probabilities are observed for the eMBB Basic Video compared to the Premium Video due to the higher ARP value (i.e., lower priority) of the former. In turn, MC services present low blocking probabilities as a result of the low offered load associated with Tenant 2. Specifically, the MC PTT attains almost 0% blocking because the service is granted with the highest priority (i.e., ARP = 1). It is also worth noting that, for public safety services, the blocking probabilities remain constant as the eMBB traffic increases, thus reflecting that the AC policy, which establishes capacity thresholds on a per-tenant basis, provides isolation between Tenant 1 and Tenant 2.

In the disaster scenario, and given that Fig. 4b considers variable load for Tenant 2, the contrary case is found: blocking probability grows for MC services while eMBB services remain invariable. The low blocking probabilities observed for Tenant 1 result from the moderate offered loads, which do not exceed its maximum capacity (i.e., $C_{max,1} = 46.8$ Mb/s). Regarding the public safety services, the MC Video can reach substantial blocking in case the traffic grows significantly, whereas the MC PTT service is able to maintain low blocking probabilities.

Blocking probability results can be contrasted by analyzing the state probability distribution when setting a fixed load for each Tenant. In particular, Fig. 5 shows the steady-state probability distribution for each of the tenants when

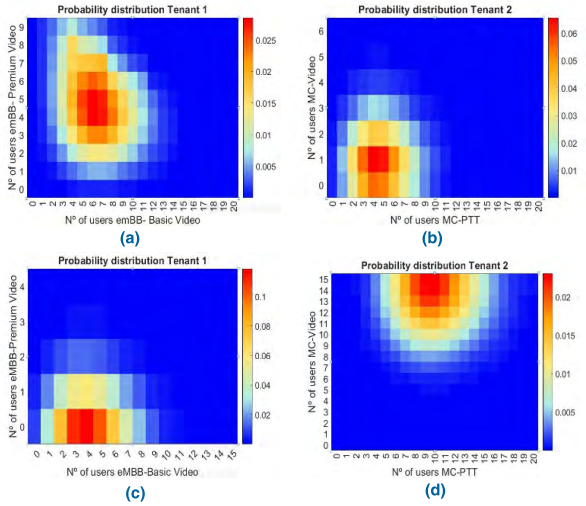

FIGURE 5. Probability distribution for each of the tenants in the standard scenario (Tenant 1 traffic generation rate set to 0.12 sessions/s) and disaster scenario (Tenant 2 traffic generation rate set to 0.25 sessions/s). (a) Tenant 1- Standard. (b) Tenant 2- Standard. (c) Tenant 1- Disaster. (d) Tenant 2- Disaster.

Tenant 1's traffic generation rate is set to 0.12 sessions/s (which corresponds to 82.8 Mb/s) in the standard scenario and Tenant 2's traffic generation is set to 0.25 session/s (which corresponds to 88 Mb/s) in the disaster scenario. The axes of the graphs reflect the number of users of each tenant for each represented state probability. For the standard scenario and Tenant 1 (Fig. 5a), the states with highest probability are mostly those with an intermediate number of users for the eMBB Premium video service, while for Tenant 2 (Fig. 5b), the highest probabilities correspond to states with a low number of users in relation to its maximum. This explains that the blocking probability for Tenant 1 is relatively high in Fig. 4a, while for Tenant 2, it is lower.

For the disaster scenario, the state space for Tenant 1 (Fig. 5c) is reduced in comparison to the standard scenario case because the $C_{max,1}$ threshold has been set to a lower value (i.e., $C_{max,1} = 46.8$ Mb/s compared to $C_{max,1} = 93.6$Mb/s in the standard scenario). Therefore, the highest

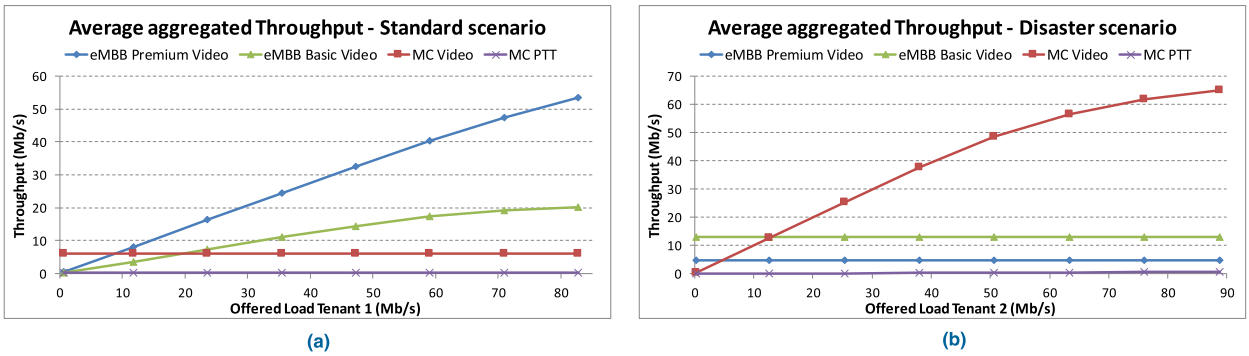

FIGURE 6. Average aggregated throughput and offered load of each service in (a) standard and (b) disaster scenarios.

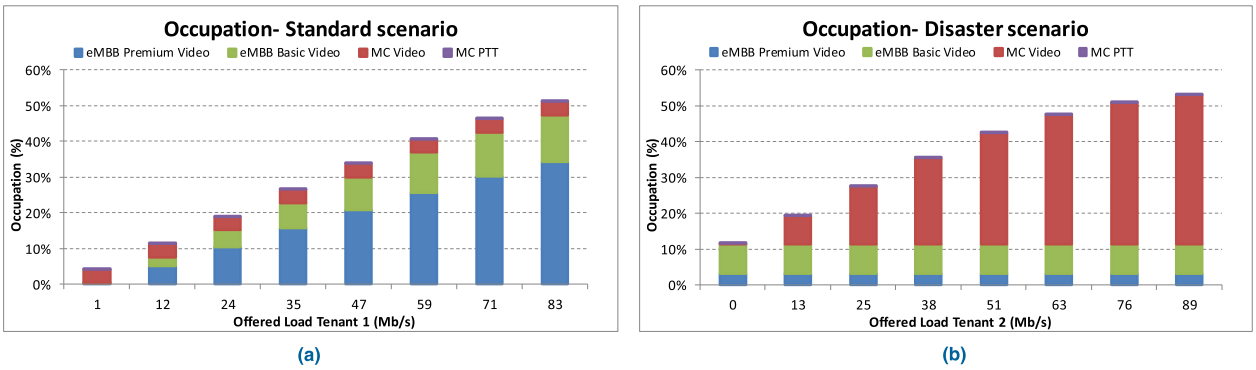

FIGURE 7. Average occupation per tenant and service for the standard and disaster scenarios. (a) Standard Scenario. (b) Disaster Scenario.

probability is concentrated in states with a low number of users, which is consistent with the low blocking probabilities found in Fig. 4b for Tenant 1. Regarding Tenant 2, the range of possible states is increased as the maximum tenant capacity threshold has been increased to $C_{max,2} = 78$ Mb/s. In that case, the highest probability concentrates in states with a large number of users for MC Video service and low number of users for MC PTT service, which results in the large Tenant 2 blocking probability found in Fig. 4b for MC Video and low probabilities for MC PTT.

Although not included graphically, the analysis of the degradation probability reveals that both of the studied scenarios present negligible degradation probabilities (i.e., lower than $3 \cdot 10^{-7}$ for the standard scenario and $2 \cdot 10^{-11}$ for the disaster scenario), implying that the admitted users in the system are provided with its GBR requirements satisfactorily most of the time.

Additionally, the average aggregated throughput per service has been analyzed in Fig. 6. For the standard scenario (Fig. 6a), the offered load of eMBB services is increased, so its average throughput for eMBB services grows until achieving high loads, where the system starts overloading. This is directly related to the high blocking probabilities observed for those services in high load conditions in Fig. 4a. In contrast, MC services average throughput remains constant for all loads, as a result of its constant offered load and the low blocking probabilities achieved in the standard scenario. In the case of the disaster scenario (Fig. 6b), eMBB services and MC-PTT throughput stay invariable as its corresponding low blocking probabilities (Fig. 4b) whereas MC Video throughput grows until reaching high loads, when the throughput reduces its growing speed as a consequence of the high blocking probabilities obtained for MC-Video in this scenario. By observing the graphs and considering that negligible degradation probabilities are obtained, it can be considered that the increase of offered load in one tenant does not affect the other tenant's throughput.

Finally, an analysis in terms of the system occupation is conducted for each of the scenarios by considering the average normalized PRB utilization. For the standard scenario (Fig. 7a), higher PRB utilization is found for eMBB services than for MC services, which are barely affected when Tenant 1's load is varied. In contrast, higher PRB utilization is achieved for MC services than for eMBB services in the disaster scenario (Fig. 7b). The reason for this can be found in the variation of the maximum capacity thresholds per tenant for each of the scenarios, as in the standard scenario, higher occupation is allowed to Tenant 1, and in the disaster scenario, the reverse situation is configured. Another relevant effect regarding the results is that the PRB utilization of MC PTT is truly low, given its low GFBR requirements.

V. CONCLUSIONS AND FUTURE WORK

This paper has presented a Markov Model approach for characterizing the resource sharing in RAN slicing scenarios, where multiple-tenants provide multiple GBR services. The model is able to capture different admission control policies, which determine the transition probabilities between the different states in the model. In particular, a slice-aware admission control policy has been studied by evaluating different performance metrics (blocking probability, degradation probability, throughput, occupation, etc.).

The Markov model has been validated by contrasting the predicted performance with that achieved through system level simulations under different environments, user speeds and load levels. The results have shown that the model is suitable to obtain reliable results to study RAN slicing in a good number of representative scenarios with QoS requirements in terms of GBR and ARP. Afterwards, the model has been exploited to obtain performance results in a relevant use case envisaged for 5G, which considers enhanced mobile broadband and mission critical services provided by different tenants. The analysis, conducted under standard and disaster traffic conditions, has revealed that (i) the proposed admission control policy is able to achieve isolation between the different slices, so that overload situations in one slice do not affect the acceptance of users of the other slice, while preserving the maximum capacity allowed to each of the slices; (ii) ARP priorities are respected by providing lower blocking probability to those services with lower ARP (higher priority); and (iii) the proposed admission control policy provides negligible degradation rates, which implies that the requested GBR values are provided to the admitted users in the system.

Based on the potential of the proposed analytical model, different future research directions are envisaged. First, the model can be extended to include non-GBR services, which require elaboration of the resource allocation procedure included in the model. Second, the development of RAN slicing mechanisms at the packet scheduling level according to different policies can also be studied by modifying the resource allocation process. Third, the model accuracy can be upgraded to cope with scenarios with highly variable spectral efficiency by considering the aggregate statistical distribution of the spectral efficiency associated with the number of users in each state of the Markov chain.

REFERENCES

[1] R. Hattachi, Ed., "5G white paper," NGMN Alliance, Boston, MA, USA, White Paper, Feb. 2018. [Online]. Available: https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf

[2] K. Samdanis, X. C. 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.

[3] P. Rost *et al.*, "Mobile network architecture evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 84–91, May 2016.

[4] R. Peter *et al.*, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.

[5] *System Architecture for the 5G System; Stage 2 (Release 15)*, document 3GPP TS 23.501 V15.4.0, Dec. 2018.

[6] *NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)*, document 3GPP TS 38.300 V15.4.0, Dec. 2018.

[7] 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.

[8] 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.

[9] P. Caballero *et al.*, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[10] 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.

[11] 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. 184–193, May 2018.

[12] 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*, Atlanta, GA, USA, May 2017, pp. 1–9.

[13] 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.*, to be published.

[14] 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*, Atlanta, GA, USA, May 2017, pp. 1–9.

[15] J. Epperlein and J. Mareček, "Resource allocation with population dynamics," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2017, pp. 1293–1300.

[16] 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, India, Jan. 2017, pp. 1–8.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] S. Lin *et al.*, "Advanced dynamic channel access strategy in spectrum sharing 5G systems," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 74–80, Oct. 2017.

[24] I. Vilà, 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, pp. 1–6.

[25] *NR; Base Station (BS) Radio Transmission and Reception (Release 15)*, document 3GPP TS 38.104 v15.4.0, Dec. 2018.

[26] *NR; Physical Channels and Modulation (Release 15)*, document 3GPP TS 38.211 v15.4.0, Dec. 2018.

[27] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton, NJ, USA: Princeton Univ. Press, 1994.

[28] *Study on Channel Model for Frequencies From 0.5 to 100GHz (Release 15)*, document 3GPP TR 38.901 v15.0.0, Jun. 2018.

[29] *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.

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. In 2018, she joined the Mobile Communication Research Group (GRCM), Department of Signal Theory and Communications (TSC), UPC, where she is currently pursuing the Ph.D. degree, supported with an FI AGAUR grant by the Government of Catalunya. Her current research interests include RAN Slicing, software defined networking (SDN), and network function virtualization (NFV), concepts to be included in new 5G technologies.

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 Workpackage leader and Coordinator partner and contributed to standardization bodies such as 3GPP, IEEE, and ETSI. He has published over 200 papers 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. She joined UPC, as an Assistant Professor, in 2001, where she is currently an Associate Professor. Since 1997, she has participated in several projects founded by both public and private organizations. She has published more than 50 papers in international journals and conferences. Her research interests include 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.

JORDI PÉREZ-ROMERO (S'98–M'04) is currently a Professor with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. He has been involved in different European projects and in projects for private companies. He has published more than 200 papers in international journals and conferences. He has been working in the field of wireless communication systems, with particular focus on radio resource management, cognitive radio networks, and network optimization.