Guaranteed Bit Rate Traffic Prioritisation and Isolation in Multi-tenant Radio Access Networks

I. Vilà, O. Sallent, A.Umbert, J. Pérez-Romero
Universitat Politècnica de Catalunya (UPC)
irene.vila.munoz@upc.edu, [sallent, annau, jorperez]@tsc.upc.edu

***Abstract*— Network slicing is a key feature of forthcoming 5G systems to facilitate the partitioning of the network into multiple logical networks customised according to different operation and application needs. Network slicing allows the materialisation of multi-tenant networks, in which the same infrastructure is shared among multiple communication providers, each one using a different slice. The support of multi-tenancy through slicing in the Radio Access Network (RAN) is particularly challenging because it involves the configuration and operation of multiple and diverse RAN behaviour over a common pool of radio resources while guaranteeing a certain Quality of Service (QoS) and isolation to each of the slices. This paper presents a Markovian approach to model different QoS aware Admission Control (AC) policies in a multi-tenant scenario with Guaranteed Bit Rate (GBR) services. From the analytical model, different metrics are defined to later analyse the effect of AC mechanisms on the performance achieved in various scenarios. Results show the impact of priorities for services of different tenants and isolation between tenants when different AC polices are adopted.**

Keywords—Multi-tenancy; Network slicing; RAN slice; Admission Control; Markov model

I. INTRODUCTION

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 future architectures of mobile networks, since the sharing process will reduce capital and operational costs [2]. Multi-tenancy can be materialised through network slicing capabilities [3], which enable logical networks/partitions to be created (i.e. network slices) with appropriate isolation and optimised 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 5G Core Network (5GC) and Next Generation RAN (NG-RAN) have already been defined by 3GPP [5][6]. Moreover, implementation aspects of network slicing in the NG-RAN have been studied from multiple angles, ranging from virtualisation 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. In turn, [11] proposes a set of vendor-agnostic configuration descriptors intended to characterise the features, policies and resources to be put in place across the radio protocol layers of a NG-RAN node for the realisation of concurrent RAN slices.

In this context, this paper addresses one of the fundamental problems in RAN slicing, which is dealing with the trade-off between ensuring isolation among tenants (i.e., traffic overload from one tenant should not negatively impact another tenant) and at the same time achieving an efficient usage of radio resources. The main novelty is that the radio resource sharing problem is tackled through the use of a Markov chain model.

Markovian approaches have been widely used to characterise the resource sharing paradigm in many fields, such as in mobility [12], cloud computing [13], Asynchronous Transfer Mode (ATM) dynamic capacity allocation [14] as well as in cellular networks (see e.g., [15] for a CDMA radio channel emulator, [16] for a Call Admission Control (CAC) scheme for 3G or [17] for heterogeneous networks Radio Access Technologies (RAT) policies). More recently, works in the field of 5G exploit Markov modelling for high mobility networks [18][19]. Markov chain models have also been considered in [20] for spectrum sharing schemes and primary/secondary scenarios [21][22]. Nevertheless, none of the above papers have considered the use of Markov chains models to study the performance of different Admission Control (AC) policies into RAN slicing scenarios. Thus, this paper presents an analytical Markov chain model approach considering multi-tenant and multi-service scenarios in the context of NG-RAN, which enables studying the impact of AC mechanisms on the performance of different multi-sliced scenario configurations.

The rest of the paper is organised as follows. Section II presents the proposed system model, describing its analytical Markov chain approach and introducing different AC polices that consider the concepts of priority and isolation. Section III proposes different performance metrics to evaluate the analytical model. Section IV presents the example scenario considered for 5G RAN slicing and provides performance results. Finally, Section V summarises the conclusions.

II. SYSTEM MODEL

A multi-sliced RAN scenario comprised of N tenants is assumed, each one 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 Quality of Service (QoS) requirements. Without loss of generality, this paper assumes Guaranteed Bit Rate (GBR) services, whose QoS profile is given by the GBR (i.e., the bit rate to be provided to the user) and the Allocation and Retention Priority (ARP) indicator, which defines the relative importance of the service requesting for resources and starts from 1 (highest priority) onwards (for successive lower priority services). Therefore, the QoS profile of the s-th service of the n-th tenant is specified in terms of the guaranteed $GBR_{s,n}$ and the $ARP_{s,n}$ for n=1…N and s=1,… M_n.

Let assume a cell with a certain bandwidth subdivided in resource units (e.g. the Physical Resource Blocks (PRB) in the case of Long Term Evolution (LTE) or Fifth Generation New Radio (5G NR)). The number of resource units required by each type of service $N_{req,s,n}$ depends on $GBR_{s,n}$. 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 availability of resource units, the GBR requirements and the corresponding ARP. The AC establishes a maximum system occupation threshold ω_{max}, measured as a fraction of the total number of available resource units N_{ava} in the cell.

Assuming that the users generate sessions according to a Poisson arrival process and have an exponential duration, the dynamic evolution of the number of users of each service type/tenant can be characterised in general by a Markov Chain with (M_1+M_2+…+M_N)-dimensional states.

In this paper, a 4D Markov chain is considered, accounting for N=2 Tenants (referred to as Tenant 1 and Tenant 2), each of them providing 2 different services (i.e., M_1=2 and M_2=2). Let denote as i and j the number of admitted users of services 1 and 2 of Tenant 1, respectively, and as k and l the number of admitted users of services 1 and 2 of Tenant 2, respectively. Let define $S_{(i,j,k,l)}$ as the state in which i, j, k and l users are admitted to 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 according to a Poisson process with rate $\lambda_{s,n}$ for the s-th service of the n-th tenant. The session duration follows an exponential distribution with mean $1/\mu_{s,n}$.

Moreover, since AC is in charge of admitting or rejecting users' requests depending on the system's occupation, it also affects the transitions between states. In this respect, let define $AC^{s,n}_{(i,j,k,l)}$ as the binary AC indicator for arrivals of the s-th service and n-th tenant, taking the value 1 if the new service request is accepted and 0 otherwise.

Based on the above, the Markov chain model is characterised in the following subsections.

A. State Space

In order to properly define the Markov model, (1) defines the set of feasible states, which is formed by those states that satisfy the feasibility condition $f_{(i,j,k,l)}$ given in (2) that limits the maximum number of users of each service for the available capacity, $\omega_{max} \cdot N_{ava}$.

$$S = \{ S_{(i,j,k,l)} \mid f_{(i,j,k,l)} = 1 \ \} \tag{1}$$

$$f_{(i,j,k,l)} = \begin{cases} 1 & \text{if } (i < \left\lfloor \frac{\omega_{\max} N_{ava}}{N_{req,1,1}} \right\rfloor \text{ and } j < \left\lfloor \frac{\omega_{\max} N_{ava}}{N_{req,2,1}} \right\rfloor \\ & \text{and } k < \left\lfloor \frac{\omega_{\max} N_{ava}}{N_{req,1,2}} \right\rfloor \text{ and } l < \left\lfloor \frac{\omega_{\max} N_{ava}}{N_{req,2,2}} \right\rfloor), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

B. State transition rate matrix

Given the state space, the generic state transition diagram at a particular state $S_{(i,j,k,l)}$ is depicted in Fig. 1. It is assumed that transitions are only possible between neighbouring states, so only increases or decreases of a single user are allowed from a certain state. Besides, transitions are only possible between feasible states.

By analysing the presented state transition diagram, the Steady-State Balance Equation (SSBE) is given in (3), where $P_{(i,j,k,l)}$ corresponds to the steady state probability of being in $S_{(i,j,k,l)}$. Note that the feasibility condition is considered in the expression in order to generalise the expression to any state.

$$\begin{aligned} & P_{(i,j,k,l)}[i\mu_{1,1} f_{(i-1,j,k,l)} + j\mu_{2,1} f_{(i,j-1,k,l)} + \\ & + k\mu_{1,2} f_{(i,j,k-1,l)} + l\mu_{2,2} f_{(i,j,k,l-1)} + \\ & \lambda_{1,1} AC^{1,1}_{(i,j,k,l)} f_{(i+1,j,k,l)} + \lambda_{2,1} AC^{2,1}_{(i,j,k,l)} f_{(i,j+1,k,l)} + \\ & + \lambda_{1,2} AC^{1,2}_{(i,j,k,l)} f_{(i,j,k+1,l)} + \lambda_{2,2} AC^{2,2}_{(i,j,k,l)} f_{(i,j,k,l+1)}] = \\ & = P_{(i-1,j,k,l)} \lambda_{1,1} AC^{1,1}_{(i-1,j,k,l)} f_{(i-1,j,k,l)} + \\ & + P_{(i,j-1,k,l)} \lambda_{2,1} AC^{2,1}_{(i,j-1,k,l)} f_{(i,j-1,k,l)} + \\ & + P_{(i,j,k-1,l)} \lambda_{1,2} AC^{1,2}_{(i,j,k-1,l)} f_{(i,j,k-1,l)} + \\ & + P_{(i,j,k,l-1)} \lambda_{2,2} AC^{2,2}_{(i,j,k,l-1)} f_{(i,j,k,l-1)} + \\ & + P_{(i+1,j,k,l)} (i+1) \mu_{1,1} f_{(i+1,j,k,l)} + \\ & + P_{(i,j+1,k,l)} (j+1) \mu_{2,1} f_{(i,j+1,k,l)} + \\ & + P_{(i,j,k+1,l)} (k+1) \mu_{1,2} f_{(i,j,k+1,l)} + \\ & + P_{(i,j,k,l+1)} (l+1) \mu_{2,2} f_{(i,j,k,l+1)} \end{aligned} \tag{3}$$

When the SSBEs are obtained for all the feasible 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 normalisation constraint:

$$\sum_{S_{(i,j,k,l)} \in S} P_{(i,j,k,l)} = 1 \tag{4}$$

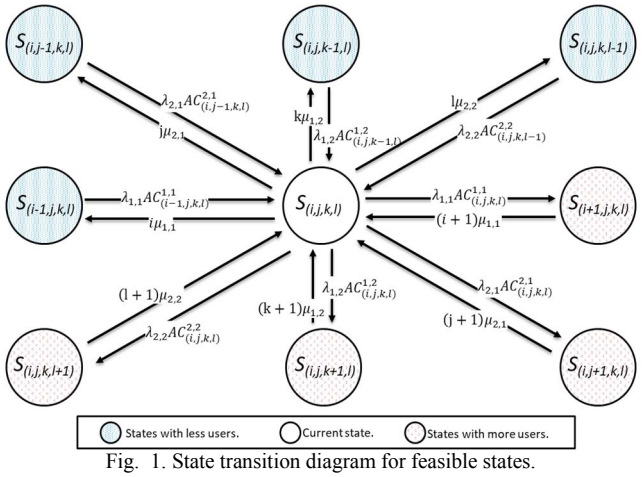

Fig. 1. State transition diagram for feasible states.

C. Admission Control

Diverse AC policies can be adopted in the proposed model to determine the acceptance of a user into the system according to its QoS parameters. In this sub-section, two specific AC policies are formulated.

1) Priority-based AC policy

This AC mechanism considers a common admission threshold ω_{max} established for both tenants and the priority $ARP_{s,n}$ indicator in order to resolve the admission of a user from the *s*-th service of the *n*-th tenant, according to:

$$AC^{s,n}_{(i,j,k,l)} = \begin{cases} 1 & \text{if } \omega_{\max} \geq \omega_{ARP,s,n} + \Delta\omega_{s,n}, \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $\omega_{ARP,s,n}$ measures the resource occupation of services with ARP lower or equal to $ARP_{s,n}$, and $\Delta\omega_{s,n}$ is the incremental proportion of needed resources to guarantee the $GBR_{s,n}$ of the user requesting admission. These parameters are given by:

$$\omega_{ARP_{s,n}} = \sum_{\substack{ARP_{s',n'} \leq ARP_{s,n} \\ s',n'=1,2}} \omega_{s',n'} \quad (6)$$

$$\Delta\omega_{s,n} = \frac{N_{req,s,n}}{N_{ava}} \quad (7)$$

where $\omega_{1,1} = i \cdot \Delta\omega_{1,1}$, $\omega_{2,1} = j \cdot \Delta\omega_{2,1}$, $\omega_{1,2} = k \cdot \Delta\omega_{1,2}$ and $\omega_{2,2} = l \cdot \Delta\omega_{2,2}$ correspond to the resource occupation of the different tenant's services.

2) Priority and Isolation-based AC policy

In order to guarantee that the admission of users from one tenant does not impact on the other tenant, the AC mechanism considers the admission threshold $\omega_{max,n}$ particularised to tenant *n*. In turn, $\omega_{ARP,s,n}$ only accounts for services with lower or equal $ARP_{s,n}$ belonging to tenant *n*. This is formulated as:

$$AC^{s,n}_{(i,j,k,l)} = \begin{cases} 1 & \text{if } \omega_{\max,n} \geq \omega_{ARP,s,n} + \Delta\omega_{s,n}, \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$\omega_{ARP_{s,n}} = \sum_{\substack{ARP_{s',n} \leq ARP_{s,n} \\ s'=1,2}} \omega_{s',n} \quad (9)$$

It is worth noting that, when considering this AC policy, the maximum number of users of a certain service in the state feasibility condition expressed by (2) should read $\omega_{max,n}$ instead of ω_{max}.

III. Performance Metrics

Based on the state probabilities, this section develops the different performance metrics of interest for the evaluation of the considered prioritisation and isolation mechanisms.

A. Blocking probability

There exists a subset of states inside the set of feasible states in which the acceptance of a new user would force the transition to an unfeasible state. Those states are known as blocking states. The set of blocking states for users of the *s*-th service of the *n*-th tenant is denoted as $S^b_{s,n}$. While extendable to other services and tenants, for the case *s*=1, *n*=1 it is defined as:

$$S^b_{1,1} = \{S_{(i,j,k,l)} \in S \mid S_{(i+1,j,k,l)} \notin S\} \quad (10)$$

The set of blocking states for the *n*-th tenant, S^b_n, are those states in which the arrival of one user from any of the services of this tenant forces the transition to an unfeasible state. Therefore, it is defined as the intersection of the sets of blocking states for the services of this tenant, i.e. $S^b_n = S^b_{1,n} \cap S^b_{2,n}$. 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 shown in (11). 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 in the summation, respectively.

$$P^b_{s,n} = \sum_{S(i,j,k,l) \in S^b_{s,n}} P_{(i,j,k,l)} \quad (11)$$

B. Degradation probability

Another subset of feasible states are the so-called degraded states, in which congestion is reached and some admitted users are not 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 resource allocation criteria adopted in the system. With the considered AC approaches, congestion may occur when the occupation of high ARP value (i.e. low priority) users is large and the system is close to its maximum capacity. Then, if a request of a user with low ARP value (i.e. high priority) arrives, the user will be admitted into the system. If this results in excess capacity, some performance degradation will be observed.

The set of degraded states for the *s*-th service of *n*-th tenant is expressed as:

$$S^{\deg}_{s,n} = \{S_{(i,j,k,l)} \in S \mid N_{ass,s,n} < N_{req,s,n}\} \quad (12)$$

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. $S^{\deg}_n = S^{\deg}_{1,n} \cup S^{\deg}_{2,n}$. 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 in (13). This 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.

$$P^{\deg}_{s,n} = \sum_{S(i,j,k,l) \in S^{\deg}_{s,n}} P_{(i,j,k,l)} \quad (13)$$

IV. Performance Evaluation

In this section, an illustrative scenario is described and the proposed analytical model is evaluated.

A. Considered scenario

The assumed scenario is comprised of a single cell that serves users from 2 Tenants providing 2 services each one. The smallest unit of radio resources that can be allocated to a user is a Physical Resource Block (PRB) of bandwidth *B*.

TABLE I. MODEL CONFIGURATION PARAMETERS

Parameter	Value
Number of available PRBs (N_{ava})	25 PRB
PRB Bandwidth (B)	180kHz
Spectral Efficiency (S_{eff})	5.6 b/s/Hz.
Data rate per PRB	1 Mbps/PRB
Guaranteed Bit Rate (GBR)	$GBR_{s,n}$=3Mbps for s,n=1,2
Average session generation rate	Tenant 1: varied from 0.001 to 0.06 sessions/s (corresponds to a variation from 0.36 Mb/s to 21.6 Mb/s) Tenant 2: - Low load: 0.02 session/s (corresponds to 7.2 Mbps) - High load: 0.035 sessions/s (corresponds to 12.6 Mb/s)
Average session duration	120 s
Tenant generation distribution	Tenant 1: 30% of generated traffic for service 1 and 70% for service 2. Tenant 2: 40% of generated traffic for service 1 and 60% for service 2

The required number of PRBs, $N_{req,s.n}$, is given by:

$$N_{req,s,n} = \frac{GBR_{s,n}}{B \cdot S_{eff}} \quad (14)$$

where S_{eff} is the spectral efficiency, which is assumed to be constant for the model's evaluation.

The configured parameters are summarised in Table I. For the specified scenario, the criteria followed to compute the assigned resources $N_{ass,s,n}$ given i, j ,k, and l admitted users in the system is done iteratively, 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 resources are reduced accordingly before moving to the next ARP. 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 distributing the available resources in proportion to the GBR required by each user.

The Markov model state probabilities have been computed through the Gauss-Seidel iterative method described in [23]. For implementation purposes of the model, it has been necessary to identify and remove those feasible states that are never reached because of an AC result being null, situation that depends on the priority awareness of the selected AC policy. The Markov model operation has been successfully validated by contrasting it with the performance results obtained from a system-level simulator. However, for the sake of brevity, the model validation results are not included in this paper.

TABLE II. ARP CONFIGURATIONS

ARP Configurations	$ARP_{1,1}$	$ARP_{2,1}$	$ARP_{1,2}$	$ARP_{2,2}$
Configuration 1	1	2	3	4
Configuration 2	3	4	1	2
Configuration 3	1	3	2	3

B. Performance results

This section includes the performance results analysed both from priority and isolation perspectives.

1) Priority analysis

For evaluating the effect of the ARP value on the performance, the ARP configurations in Table II have been tested for the priority-based AC policy with threshold ω_{max}=0.8. The offered load by Tenant 2 is set to the Low load level of Table I while Tenant 1 load is varied to observe different system load situations.

Fig. 2 shows the results obtained for the different services in the system considering the proposed ARP configuration in terms of blocking and degradation probability, both expressed in %. The comparison of the blocking probability of the different services reveals that higher blocking percentages are reached for those services with higher ARP value. This is the case of service 2 from Tenant 1 in configuration 1 or service 2 from Tenant 2 in configuration 2. In configuration 3, as services 2 from both Tenants share the same ARP value, which is the highest one in the system, the same performance in terms of blocking probability is found.

Focusing on the degradation probability, it is observed that the reached values are quite low (i.e. less than 3.5%) even for the highest considered load. This means that the guaranteed GBR of the admitted users is satisfactorily preserved. However, when the load is high, some differences can be perceived depending on the priority assigned to each of the services. Services with low ARP values are slightly degraded while services with high ARP values suffer from higher degradation. In addition, the effect of different loads can be noticed from Fig. 2f, where service 2 from Tenant 1 suffers from higher degradation than service 2 from Tenant 2, although both services have the same ARP value. The reason of this is that service 2 from Tenant 1 is more demanding because, as seen in Table I, 70% of the traffic from Tenant 1 belongs to service 2 while the traffic belonging to service 2 from Tenant 2 represents only 60%.

Based on these results, it can be concluded that the priority-based AC policy is capable of providing the required GBR to the admitted users with reduced degradation while differentiating between each service priority according to its ARP.

2) Isolation analysis

The impact of the priority and isolation-based AC policy on the achieved performance has been analysed by comparing different load conditions for two different AC tenant threshold configurations: configuration 1 sets $\omega_{max,1}$=0.4 for Tenant 1 and $\omega_{max,2}$=0.4 for Tenant 2 while configuration 2 sets $\omega_{max,1}$=0.6 and $\omega_{max,2}$=0.2. The selected ARP configuration corresponds to configuration 3 in Fig. 3. Both the Low and High load levels of Tenant 2 in Table I have been studied.

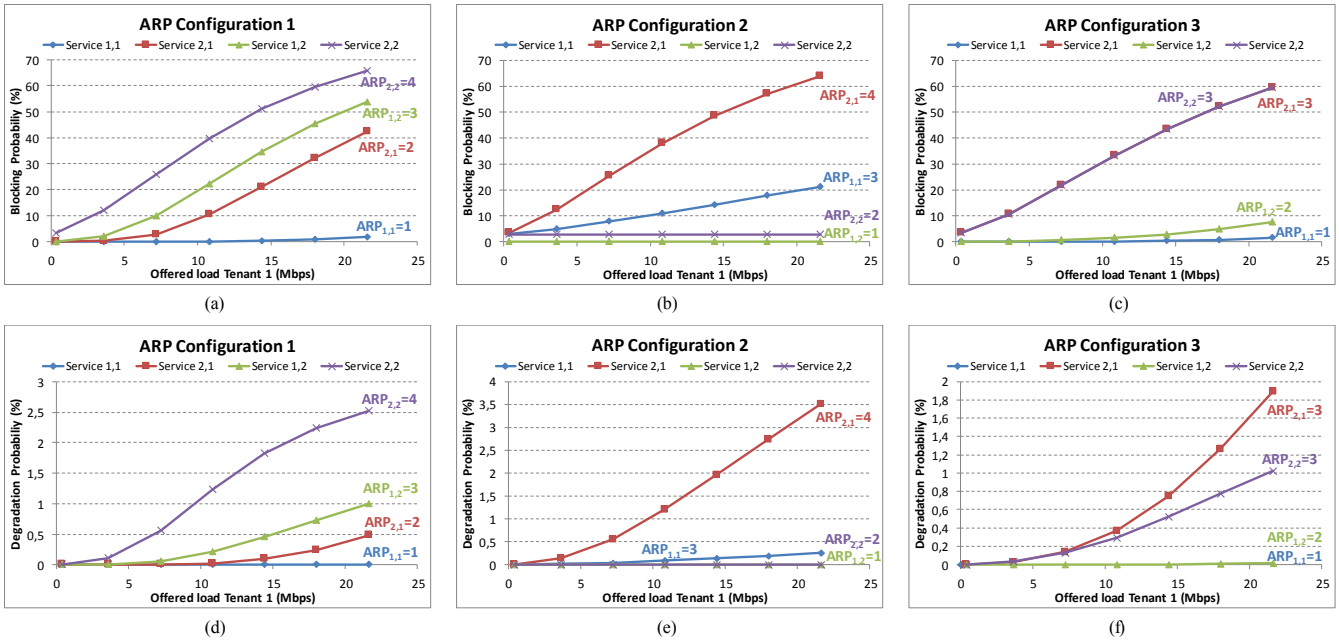

Fig. 2. Blocking probabilities for the different ARP configurations (a), (b), (c) and degradation probabilities for each configuration (d), (e), (f).

By comparing the results obtained for Tenant 1 and 2, it is clearly seen that, with this AC policy, and as a difference from the priority-based AC policy, the load variation of Tenant 1 has no impact on the blocking probability of Tenant 2, which remains constant for all the offered loads of Tenant 1 (see Fig. 3c, Fig. 3d). Moreover, no variation is found in the performance of Tenant 1 (see Fig. 3a, Fig. 3b) for the Low and High load values of Tenant 2. These results reflect the isolation achieved between tenants when this AC policy is applied.

Focusing on the effect of the AC threshold variation, it is observed that blocking probabilities of Tenant 1 are higher when configuration 1 is used, as the AC threshold $\omega_{max,1}$ is lower, so less Tenant 1 users can be admitted to the system. On the contrary, blocking probabilities for Tenant 2 are lower for configuration 1, since the value of $\omega_{max,2}$ is higher for this configuration.

As expected, the effect of ARP values is also observed in the provided results, as service 2 from both Tenants, which has the higher ARP value, also perceives higher blocking probabilities than service 1. When comparing the results for Low Tenant 2 load and configuration 1 in Fig. 3 and the results in Fig. 2c, it can be noticed that the blocking probabilities of Tenant 1 for the priority and isolation-based AC policy are in general higher than for the priority-based AC policy. This is because in the later approach, Tenant 1 improves its performance at the expense of Tenant 2 since no distinction in the resources used by each tenant is performed in the AC, while in the priority and isolation-based AC, admissions of Tenant 1 only consider the threshold defined for that tenant. Therefore, performance is dependent on how this threshold is configured. For example, with configuration 2 (see Fig. 3a, Fig. 3b), performance is improved as $\omega_{max,1}$ better fits with the actual Tenant 1 load.

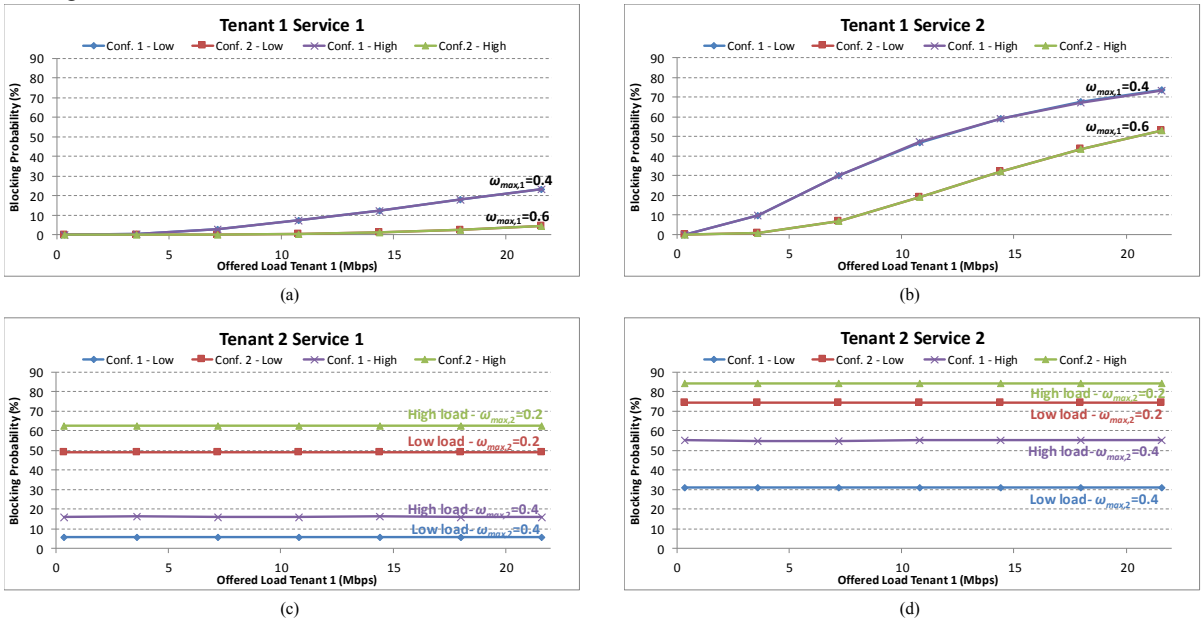

Fig. 3. Blocking probabilities for (a) service 1 and (b) service from Tenant 1 and (c) service 1 and (d) service 2 from Tenant 2 for configurations 1 and 2.

According to this, AC thresholds need to be configured according to the expected load for the tenant.

The above observations reflect that the priority and isolation-based AC policy effectively avoids that the overload in one Tenant may affect the other tenant while still respecting the priority established to each of the services of a Tenant. However, AC thresholds need to be properly defined in relation to the tenant traffic load.

V. CONCLUSIONS

This paper has proposed a Markov model for characterising the resource sharing in multi-tenant and multi-service scenarios. The model is able to capture different admission control policies by properly specifying the transition probabilities between states. In particular, two admission control mechanisms have been studied trough the analytical model: a priority-based policy and a priority and isolation based policy, which have been evaluated in terms of blocking and degradation probability for different configurations.

Results have revealed that: (i) For both admission control policies, better performance is obtained for those services with lower ARP (higher priority), (ii) The proposed admission control policies provide low degradation rates even for the highest loads considered in the evaluation, which implies that the required GBR is provided to the admitted users in the system, (iii) The priority and isolation-based policy is suitable to achieve isolation between tenants, avoiding that the traffic variations of one tenant negatively impacts on the performance of the other tenant, (iv) The proposed framework provides an appropriate platform for the further evaluation and characterisation of RAN slicing aspects.

ACKNOWLEDGEMENT

This work has been supported by the EU funded H2020 5G-PPP project 5G ESSENCE under the grant agreement 761592, by the Spanish Research Council and FEDER funds under SONAR 5G grant (ref. TEC2017-82651-R) and by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia under grant 2018FI_B_00412.

REFERENCES

[1] NGMN Alliance, “5G White Paper”, February 2015.

[2] P. Rost, et al. “Mobile network architecture evolution toward 5G”, IEEE Communication Magazine, May 2016.

[3] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso and C. Verikoukis, "Multi-tenant slicing for spectrum management on the road to 5G," IEEE Wireless Communications, vol. 24, no. 5, pp. 118-125, October 2017.

[4] P. Rost et al., "Network slicing to enable scalability and flexibility in 5G mobile networks," IEEE Communications Magazine, vol. 55, no. 5, pp. 72-79, May 2017.

[5] 3GPP TS 23.501 V15.0.0, “System architecture for the 5G system; stage 2 (Release 15)”, December 2017.

[6] 3GPP TS 38.300 V15.0.0, “NR; NR and NG-RAN overall description; Stage 2 (Release 15)”, December 2017.

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

[8] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: flexibility and resources abstraction," IEEE Communications Magazine, vol. 55, no. 6, pp. 102-108, June 2017.

[9] P. Caballero, A. Banchs, G. de Veciana and X. Costa-Pérez, "Multi-tenant radio access network slicing: statistical multiplexing of spatial loads," IEEE/ACM Transactions on Networking, vol. 25, no. 5, pp. 3044-3058, October 2017.

[10] O. Sallent, J. Perez-Romero, R. Ferrús, R. Agusti, "On radio access network slicing from a radio resource management perspective", IEEE Wireless Communications, October, 2017, pp. 166-174.

[11] R. Ferrús, O. Sallent, J. Pérez-Romero and R. Agustí, "On 5G radio access network slicing: radio interface protocol features and configuration," IEEE Communications Magazine, vol. 56, no. 5, pp. 184-193, May, 2018.

[12] J. Epperlein and J. Mareček, "Resource allocation with population dynamics," 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, October 2017, pp. 1293-1300.

[13] S. Jagannatha, N. S. Shravan and S. Kavya, "Cost performance analysis: Usage of resources in cloud using Markov-chain model," 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, January 2017, pp. 1-8.

[14] I. Lokshina and M. Bartolacci, "Effective assessment of mobile communication networks performance with clustering and neural modeling," 2008 Wireless Telecommunications Symposium, Pomona, CA, April 2008, pp. 9-16.

[15] A. Umbert and P. Diaz, "A generic radio channel emulator to evaluate higher layer protocols in a CDMA system," 11th IEEE International Symposium on Personal Indoor and Mobile Radio Communications. PIMRC 2000. Proceedings (Cat. No.00TH8525), London, September 2000, pp. 401-405 vol.1.

[16] H. Y. Ng, K. T. Ko and K. F. Tsang, "3G mobile network call admission control scheme using Markov chain," Proceedings of the Ninth International Symposium on Consumer Electronics, June 2005. (ISCE 2005), 2005, pp. 276-280.

[17] X. Gelabert, J. Pérez-Romero, O. Sallent and R. Agustí, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," in IEEE Transactions on Mobile Computing, vol. 7, no. 10, pp. 1257-1270, October 2008.

[18] 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," 2017 IEEE 42nd Conference on Local Computer Networks (LCN), Singapore, October 2017, pp. 539-542.

[19] S. Al-Rubaye, A. Al-Dulaimi, J. Cosmas and A. Anpalagan, "Call admission control for non-Standalone 5G ultra-dense networks," in IEEE Communications Letters, vol. 22, no. 5, pp. 1058-1061, May 2018.

[20] M. N. Patwary, R. Abozariba and M. Asaduzzaman, "Multi-Operator spectrum sharing models under different cooperation schemes for next generation cellular networks," 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, September 2017, pp. 1-7.

[21] S. Lin et al., "Advanced dynamic channel access strategy in Spectrum sharing 5G systems," in IEEE Wireless Communications, vol. 24, no. 5, pp. 74-80, October 2017.

[22] K. B. Ali, M. S. Obaidat, F. Zarai and L. Kamoun, "Markov model-based adaptive CAC scheme for 3GPP LTE femtocell networks," 2015 IEEE International Conference on Communications (ICC), London, June 2015, pp. 6924-6928.

[23] W.J. Stewart, Introduction to the Numerical Solution of Markov Chains. Princeton Univ. Press, 1994.