Profit-Based Radio Access Network Slicing for Multi-tenant 5G Networks

J. Pérez-Romero, O. Sallent, R. Ferrús, R. Agustí
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
[jorperez,sallent,ferrus,ramon]@tsc.upc.edu

***Abstract*—Network slicing is a key capability of 5G networks that facilitates the provision of multi-tenancy by allocating different slices to the tenants that share a certain infrastructure according to specific Service Level Agreements (SLAs). In this context, this paper focuses on radio admission control as the function that controls the amount of radio resources assigned to the different tenants in a 5G radio access network. Specifically, a novel approach is proposed that includes profit-related metrics in the decision-making process, accounting for the additional extra incomes that can be obtained by sporadically granting additional capacity beyond the SLA level and for the penalties incurred due to potential SLA breaches. The proposed approach is evaluated by means of simulations to assess its benefits in terms of achieved profit and throughput.**

Keywords—Multi-tenancy; RAN Slicing; Radio Admission Control; 5G New Radio.

I. INTRODUCTION

Network slicing is a key capability of 5G systems that allows for a single network to simultaneously support a wide range of application scenarios (e.g. automotive, utilities, smart cities, high-tech manufacturing) and business models that exhibit a high variety of requirements on network functionalities and expected performance [1]. It allows operators to compose and manage multiple dedicated logical networks with specific functionality running on top of a common infrastructure. Each one of these logical networks is referred to as *network slice*, and can be tailored to provide a particular system behaviour to best support specific service/applications domains. The realization of a network slice requires specific features and resources both in the 5G Core network part and in the New Generation Radio Access Network (NG-RAN) part, referred to as *RAN slice*. The realization of RAN slices is particularly challenging because it has to address how the pool of radio resources available to one gNB (i.e., a NG RAN node) can be configured and operated to simultaneously deliver multiple and diverse behaviours [2].

Network slicing capability facilitates the provision of *multi-tenant* networks in which a *tenant* (i.e. an organization or business entity that provides services to a group of subscribers) consumes the network slice services with the necessary resources and isolation (with regard to the operation of other concurrent slices) to meet the service performance requirements agreed between network operator and tenant.

This work has been supported by the EU funded H2020 5G-PPP project 5G ESSENCE under the grant agreement 761592 and by the Spanish Research Council and FEDER funds under SONAR 5G grant (ref. TEC2017-82651-R).

In a multi-tenant network, the Service Level Agreement (SLA) is the contract between the operator acting as infrastructure provider and each tenant that guarantees specific levels of performance and reliability at certain cost [3]. The SLA records a common understanding about the service and/or service behavior offered by the infrastructure provider, together with the measurable target values characterizing the level of the offered service (e.g. aggregated guaranteed bit rate that the provider commits to assure to each tenant) as well as related cost considerations (e.g. penalties for SLA breaching) [4]. SLAs are essential elements to drive the management of multi-tenant networks, as it has been considered in the literature in different contexts, including mobile networks [5], Software as a Service [6] or Database as a Service [7].

Consistently with the type of performance guarantees and cost considerations established in the SLAs, the network should include the appropriate resource management functions for configuring and operating the network slice associated with each tenant to fulfil the SLA. A particularly challenging scenario arises when tenants offer services that require a Guaranteed Bit Rate (GBR). In this respect, this paper focuses on multi-tenant NG-RAN scenarios where provisions for support of this type of services are part of the SLAs. Under such scenario, this paper proposes a novel Radio Admission Control (RAC) function to control the assignment of radio resources for GBR services to the different RAN slices that directly includes in the decision-making process profit-related metrics as derived from the SLA. In order to efficiently utilise the available radio resources and considering that the actual traffic demand required by a tenant in a certain area will be subject to variations with respect to the SLA, the main novelty of the proposed RAC is that it accounts for the additional extra incomes that can be obtained by the infrastructure provider when sporadically granting additional capacity to a tenant above its agreed SLA level and for the penalties incurred due to potential SLA breaches.

Different works in the literature have proposed solutions for managing the split of the available radio resources among different RAN slices in multi-tenant contexts. This split can be performed at different levels with the support of different Radio Resource Management (RRM) functions [8]. While most of the works have focused on the Packet Scheduling (PS) problem to determine the amount of resources available to each RAN slice [9]-[12], some works have also considered RAC for RAN slicing. In [13] a multi-tenant admission control for cellular networks was proposed while in [14] the Network Virtualisation Substrate concept of [11][12] was used in

978-1-7281-0546-8/19/$31.00 ©2019 IEEE

conjunction with a tenant-specific admission control. Similarly, [15] proposed a joint admission control and network slicing approach through a heuristic algorithm that integrates spectrum allocation, admission control and spatial multiplexing. In turn, the 5G Network Slice Broker of [5] was further developed in [16] incorporating traffic forecasting and admission control for slice instantiation requests. Nevertheless, none of these previous references incorporates business-related considerations in the RAN slicing process, which constitutes the main novelty of the approach presented here.

The rest of the paper is organized as follows. Section II presents the considered system model and defines the business model-related metrics. The proposed RAC algorithm is presented in Section III. Then, the proposed approach is evaluated through system-level simulations in Section IV. Finally, conclusions are outlined in Section V.

II. System and Business Model

Let us assume a 5G NR cell with a total of N Physical Resource Blocks (PRB), each one with bandwidth B that depends on the considered 5G NR numerology (see [17] for details). The cell is operated by an infrastructure provider that configures it to serve Q different slices. Each slice is associated with a different tenant and, therefore, the terms slice and tenant will be used interchangeably throughout this paper.

It is assumed that tenant q provides a total of S_q GBR services numbered as s=1,..,S_q through its associated slice. Therefore, following the Quality of Service (QoS) model of [1], the requirements of service s of tenant q are specified in terms of the Guaranteed Flow Bit Rate (GFBR) to be provided to the QoS flows of this service, denoted as $GFBR_{q,s}$.

The SLA between the infrastructure provider and each tenant is assumed to follow a mixed flat-rate and pay-per-use model in which the infrastructure provider gets a flat rate cost from tenant q for ensuring a contractual aggregate GFBR, denoted as SLA_q, to be guaranteed for the aggregate of all QoS flows of the tenant in the cell. In addition, the pay-per-use component of this model assumes a number of extra incomes or penalties for the infrastructure provider when there is traffic that exceeds the contractual aggregate or when SLA breaches occur, as specified in the following:

- The infrastructure provider gets extra incomes if, sporadically, it can grant excess capacity above the guaranteed SLA_q value to tenant q. Specifically, each time that a new QoS flow establishment request of a tenant is admitted and this involves exceeding the value of SLA_q for that tenant, the infrastructure provider gets an extra income of $\Delta I_{q,s}$ units (where s is the requested service of tenant q) if the QoS level of this flow is successfully provisioned.
- In case of SLA breaches, the infrastructure provider incurs a penalty with the corresponding tenant. This can happen in two different situations:
  - Each time that a new QoS flow establishment request is rejected for a tenant although the total requested capacity of this tenant is below its SLA_q, the infrastructure provider incurs a penalty of $\Delta C_{q,s}$ units (where s is the requested service of tenant q).
  - When congestion occurs (i.e. there are not sufficient PRBs to ensure the GFBR value of the admitted QoS flows), the infrastructure provider incurs a penalty. This penalty is $\Delta D_{q,s}$ units per each QoS flow of service s of tenant q that has not received its GFBR value during at least 1% of the QoS flow lifetime.

At the NG-RAN, each QoS flow is mapped to a Data Radio Bearer (DRB) that enables the data transfer through the radio interface between the User Equipment (UE) and the gNB according to the expected QoS. Without loss of generality, a one-to-one mapping between QoS flows and DRBs is assumed in this paper.

Whenever a new GBR QoS flow is established for a given service, the RAC function is triggered to determine if the DRB associated to this new QoS flow can be admitted or not. In general terms, the RAC decision should take into account the SLA of the different slices and the number of QoS flows that have been already admitted in the cell for each slice. In this context, the RAC algorithm proposed in this paper intends to maximize the profit of the infrastructure provider as per the pay-per-use component of the SLA model explained above. In particular, the algorithm considers the average profit P obtained by the infrastructure provider during a certain time window T as a result of the different extra incomes and penalties, which is given by:

$$P = \frac{1}{T}\sum_{q=1}^{Q}\sum_{s=1}^{S_q}\left(x_{q,s}(T)\Delta I_{q,s} - r_{q,s}(T)\Delta C_{q,s} - d_{q,s}(T)\Delta D_{q,s}\right) \quad (1)$$

This expression combines the following measurements performed during time window T: $x_{q,s}(T)$ is the number of QoS flows of service s and tenant q that have been admitted exceeding the SLA of this tenant and whose GFBR has been successfully provided. $r_{q,s}(T)$ is the number of QoS flows of service s and tenant q that have been rejected although the requested capacity of this tenant is below the SLA level. $d_{q,s}(T)$ is the number of admitted QoS flows of service s and tenant q that have experienced congestion during time window T. It is worth mentioning that the profit P in (1) does not consider explicitly the incomes from the fixed flat rate cost, since they can be regarded as just a constant additional term that will not be impacted by the decisions of the RAC.

III. Formulation of the Profit Based Radio Admission Control Algorithm

This section presents the proposed RAC algorithm that intends to maximize the profit obtained by the infrastructure provider. For this purpose, let us assume that, at a certain time, there is a new QoS flow establishment request for service s of tenant q. At this time, the number of already admitted QoS flows in the cell for each service and tenant is given by vector $\mathbf{x} = \left(n_{1,1},...,n_{1,S_1},...,n_{Q,1},...,n_{Q,S_Q}\right)$ where $n_{q,s}$ is the number of admitted QoS flows of service s of tenant q. In turn, the aggregate bit rate demand for all the services that have been already admitted for tenant q is given by:

$$R(q,\mathbf{x}) = \sum_{s=1}^{S_q} n_{q,s}\cdot GFBR_{q,s} \quad (2)$$

Let us also denote as $\mathbf{x}^+ = \left(n_{1,1},...,n_{q,s}+1,...,n_{Q,S_Q}\right)$ the

vector with the number of admitted QoS flows in case that the RAC admits the new QoS flow of service s of tenant q.

The RAC decision logic is based on comparing the estimated profit variation if the request is accepted, with the estimated profit variation if the request is rejected. To this end, the following possible situations are analysed:

Case 1: The new request is SLA compliant (i.e., $R(q,\mathbf{x}^+)\leq SLA_q$)

In this case, the acceptance of the request would be the expected behaviour, although it does not bring any extra income, since the total demand of slice q including the new request $R(q,\mathbf{x}^+)$ does not exceed the SLA level. Indeed, if the new request is rejected, an SLA breach occurs and the infrastructure provider incurs a penalty $\Delta C_{q,s}$ and, thus, a profit reduction of $\Delta R_r=\Delta C_{q,s}$.

However, there is also a risk of profit reduction in case of accepting the new request that has to be necessarily accounted. Specifically, in the eventuality that congestion appears after accepting the new request, there will be a penalty of $\Delta D_{q's'}$ for each admitted QoS flow of any service s' and tenant q'. Therefore, the estimated profit reduction if the request is accepted is given by:

$$\Delta R_a = \left(\Delta D_{q,s} + \sum_{q'=1}^{Q} \sum_{s'=1}^{S_{q'}} \Delta D_{q',s'} n_{q',s'} \right) P_{cong}\left(\mathbf{x}^+\right) \qquad (3)$$

where $P_{cong}(\mathbf{x}^+)$ is the estimation of the congestion probability when the number of admitted QoS flows is $\mathbf{x}^+$. The computation of the probability is detailed in the Appendix and depends on the propagation conditions existing in the cell that impact on the spectral efficiency achievable by the different UEs that transmit/receive the data of the QoS flows.

Based on these considerations, the new QoS flow can be admitted if the estimated profit reduction under acceptance ΔR_a is lower or equal than the estimated reduction under rejection ΔR_r, i.e. if the following condition holds:

$$\left(\Delta D_{q,s} + \sum_{q'=1}^{Q} \sum_{s'=1}^{S_{q'}} \Delta D_{q',s'} n_{q',s'} \right) P_{cong}\left(\mathbf{x}^+\right) \leq \Delta C_{q,s} \qquad (4)$$

Case 2: The new request is not SLA compliant (i.e., $R(q,\mathbf{x}^+)>SLA_q$)

In this case, if the request is accepted, the profit will be increased with the extra income $\Delta I_{q,s}$ as long as no congestion occurs. However, if congestion occurs, the profit will be reduced in $\Delta D_{q',s'}$ for each admitted QoS flow of any service s' of tenant q'. Therefore, the estimated profit increase if the request is accepted depends on the congestion probability $P_{cong}(\mathbf{x}^+)$ as:

$$\Delta P_a = \Delta I_{q,s}\left(1 - P_{cong}\left(\mathbf{x}^+\right)\right) - \left(\Delta D_{q,s} + \sum_{q'=1}^{Q} \sum_{s'=1}^{S_{q'}} \Delta D_{q',s'} n_{q',s'} \right) P_{cong}\left(\mathbf{x}^+\right) \qquad (5)$$

Instead, if the request is rejected, the profit will not be affected because the new request does not fulfil the SLA, so $\Delta P_r=0$. Correspondingly, the new QoS flow can be admitted if $\Delta P_a>\Delta P_r$, i.e. if the following condition holds:

$$\Delta I_{q,s}\left(1 - P_{cong}\left(\mathbf{x}^+\right)\right) - \left(\Delta D_{q,s} + \sum_{q'=1}^{Q} \sum_{s'=1}^{S_{q'}} \Delta D_{q',s'} n_{q',s'} \right) P_{cong}\left(\mathbf{x}^+\right) > 0 \qquad (6)$$

Based on the above considerations, Fig. 1 presents the flow diagram of the overall RAC process.

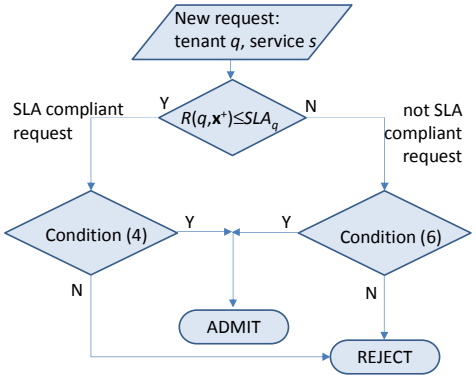


Fig. 1 Flow diagram of the Radio Admission Control process

IV. Performance Evaluation

A. Scenario description

The performance of the proposed approach has been evaluated in a scenario composed by one gNB with one omnidirectional cell that includes N=51 PRBs with bandwidth B=360 kHz, which corresponds to a subcarrier spacing Δf=30 kHz according to the numerologies defined in [17]. Table I summarizes the considered simulation parameters.

TABLE I. Simulation Parameters

Parameter	Value
Cell radius	115m
Path loss and shadowing model	Urban micro-cell model with hexagonal layout (see details in [18])
Shadowing standard deviation	3 dB in Line Of Sight (LOS) and 4 dB in Non Line Of Sight (NLOS) [18]
Base station antenna gain	5 dB
Frequency	3.6 GHz
Transmitted power per PRB	24 dBm
Number of PRBs (N)	51
UE noise figure	9 dB
Link-level model to map Signal to Interference and Noise Ratio and bit rate	Model in section A.1 of [19] with maximum spectral efficiency 8.8 b/s/Hz.
Session generation rate ($\lambda_{q,s}$)	$\lambda_{1,1}$: from $8.33\cdot10^{-4}$ to 0.1 sessions/s $\lambda_{2,1}$: from $4.17\cdot10^{-3}$ to 0.5 sessions/s
Average session duration ($T_{q,s}$)	120 s
Offered load per tenant ($\lambda_{q,s}\cdot GFBR_{q,s}\cdot T_{q,s}$)	From 1 Mb/s to 120 Mb/s
Simulation duration (T)	100000 s

The cell has been configured to support Q=2 tenants each one associated with a different slice. Tenant q=1, denoted as T1, provides a service of high definition video to its subscribers with $GFBR_{1,1}$= 10 Mb/s. In turn, tenant q=2, denoted as T2, provides a video service of standard quality with $GFBR_{2,1}$= 2 Mb/s. The SLAs are established in terms of an aggregate bit rate SLA_1=60 Mb/s for tenant 1 and SLA_2=40 Mb/s for tenant 2. These values are defined under the consideration that, based on the characteristics of the scenario, the planned average capacity of the cell is estimated to be around C=100 Mb/s. It is assumed that the sessions of both services are generated following a Poisson distribution process with exponentially distributed session duration and that each session involves one QoS flow. As indicated in Table I, the value of the session generation rate is varied in different simulations to consider different offered

load levels. Only traffic in the downlink direction is considered.

The profit is measured assuming the extra income of $\Delta I_{q,s}$ is proportional to the guaranteed bit rate $GFBR_{q,s}$ and the average session duration $T_{q,s}$ of the service with an income of 0.1 units/s per guaranteed Mb/s. Correspondingly, $\Delta I_{q,s}=0.1\cdot GFBR_{q,s}\cdot T_{q,s}$, which yields $\Delta I_{1,1}=120$ units and $\Delta I_{2,1}=24$ units. The values of $\Delta C_{q,s}$ and $\Delta D_{q,s}$ are varied to consider different types of SLAs.

B. Assessment in terms of profit and throughput

Fig. 2 plots the 3D representation of the obtained profit with the proposed RAC approach as a function of the offered load of each tenant. The profit result is averaged along the whole simulation time and measured in profit units per second. Fig. 2 includes three different configurations of the $\Delta C_{q,s}$ and $\Delta D_{q,s}$ values representing different parametrisations of the SLA. Fig. 2a represents the case $\Delta C_{q,s}=\Delta D_{q,s}=10$ units, in which the SLA establishes a lower value of the penalties than the extra incomes, $\Delta I_{q,s}$. In turn, Fig. 2b considers the case in which the penalties $\Delta C_{q,s}=\Delta D_{q,s}=100$ units are similar to the extra income. Finally, Fig. 2c assumes $\Delta C_{q,s}=\Delta D_{q,s}=200$ units, where the penalties are higher than the extra incomes.

It is observed in Fig. 2 that, when the offered load of both tenants is much lower than their contractual aggregate bit rates (SLA_1=60 Mb/s and SLA_2=40 Mb/s respectively), the profit is close to 0 in all the three configurations. Then, focusing on the configuration of Fig. 2a, it is observed that, as the load increases up to the contractual aggregates and beyond, there is an increasing trend of the profit. This increasing trend is more noticeable for the cases when the offered load of one tenant is much lower than its contractual agreement SLA_q value while at the same time the offered load of the other tenant is higher than SLA_q. For example, when the offered load of T1 is 20 Mb/s (i.e. one third of its SLA) and that of T2 is 80 Mb/s (i.e. twice its SLA) the obtained profit is about 7.5 units/s, and it increases further when the unbalance between offered load and SLA is higher, e.g. up to 10 units/s for an offered load of 1 Mb/s for T1 and 120 Mb/s for T2. This last effect is also observed in the configurations of Figs. 2b and 2c, where it can be clearly seen that the highest profits are achieved for a high offered load of one tenant and a low offered load of the other tenant. However, as a difference from Fig. 2a, in both figures Fig. 2b and 2c it is also noticed that, when the load of both tenants increases beyond their SLA_q values the profit starts to exhibit a decreasing trend due to the higher penalties in case of SLA breaches. These higher penalties make, on the one hand, that the admittance of requests above the SLA_q value becomes less beneficial from the infrastructure provider's perspective and, on the other hand, that the profit experiences higher reductions in case of SLA breaches. As a result of this decreasing trend, the profit in Fig. 2b ranges from -0.8 units/s when the offered load of both tenants is 120 Mb/s up to 10 units/s when the offered loads of T1 and T2 are 1 Mb/s and 120 Mb/s, respectively. In turn, due to the higher penalty of Fig. 2c the profit ranges in this case from -7.2 up to 9.8 units/s.

Fig. 3 depicts the total throughput in the cell as a function of the offered load of each tenant and considering the same three configurations as in Fig. 2. It is observed that the

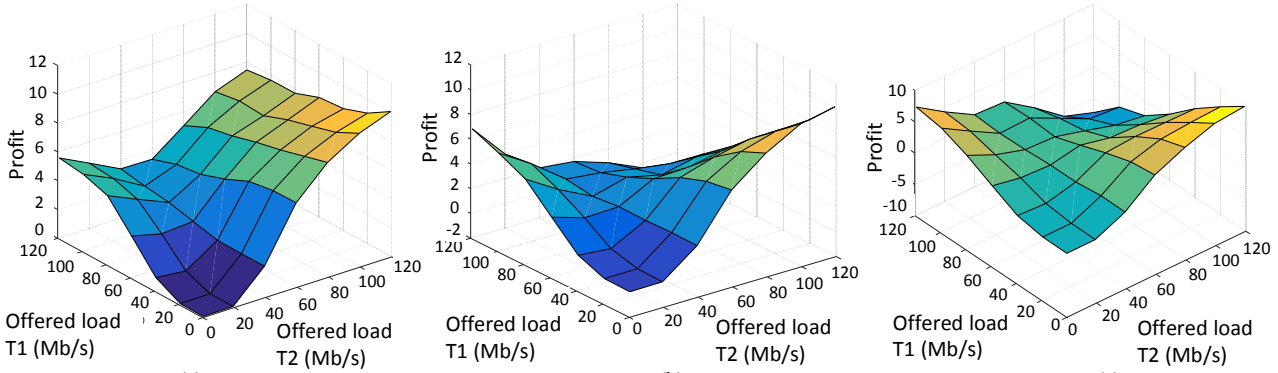


Fig. 2 Average profit for different values of the penalties in the SLA (a) $\Delta C_{q,s}=\Delta D_{q,s}=10$, (b) $\Delta C_{q,s}=\Delta D_{q,s}=100$, (c) $\Delta C_{q,s}=\Delta D_{q,s}=200$

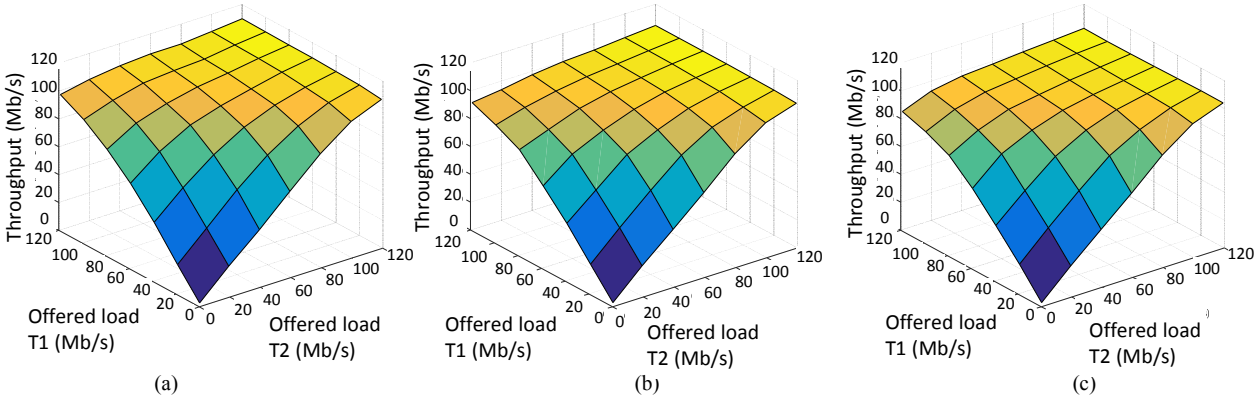


Fig. 3 Total throughput for different values of the penalties in the SLA (a) $\Delta C_{q,s}=\Delta D_{q,s}=10$, (b) $\Delta C_{q,s}=\Delta D_{q,s}=100$, (c) $\Delta C_{q,s}=\Delta D_{q,s}=200$

throughput increases with the offered load of the two tenants and it tends to stabilize at the value of about 100 Mb/s in accordance with the planned capacity of the cell. It is also remarkable that, when the offered load of one tenants is below its SLA_q value and the offered load of the other tenant is above its SLA_q the cell is also able to reach a throughput very close to the planned capacity of 100 Mb/s, reflecting that the proposed profit-based RAC allows giving to the high demanding tenant the unused capacity of the other tenant. This trend is more noticeable for the configurations of Fig. 3a and Fig. 3b, in which the penalty values are lower, than for the configuration in Fig. 3c. Overall, Figs. 2 and 3 reflect that the proposed RAC approach can achieve a good utilization of the available resources while at the same time providing some additional profit to the infrastructure provider.

C. Comparison against reference strategies

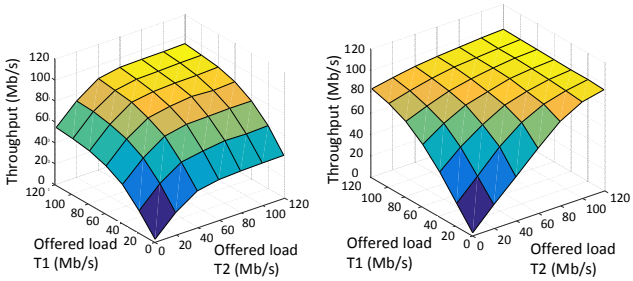
To further assess the behaviour of the proposed profit-based RAC, a comparison is performed against the two reference RAC strategies described in the following:

- Reference #1: This RAC strategy decides on the admission or rejection of a new QoS flow of the q-th tenant based on the agreed SLA_q of this tenant and considers a strict slicing mechanism that does not allow deviations exceeding this value. Therefore, a new QoS flow of the s-th service of the q-th tenant is admitted if condition $R(q,\mathbf{x})+GFBR_{q,s} \leq SLA_q$ is fulfilled. Otherwise, the QoS flow is rejected.
- Reference #2: This strategy considers a non slice-aware RAC that makes decisions based only on the planned capacity C of the cell. Then, a new QoS flow of the s-th service of the q-th tenant is admitted if the following condition is fulfilled:

$$\sum_{q'=1}^{Q} R(q',\mathbf{x}) + GFBR_{q,s} \leq C \qquad (7)$$

Fig. 4 depicts the total aggregate throughput in the cell as a function of the offered load of each tenant for the two reference strategies #1 (Fig. 4a) and #2 (Fig. 4b). For the reference strategy #1 it is observed in Fig. 4 that, when the load of one of the tenants is low, the total obtained throughput exhibits substantially lower values than the profit-based strategy (see Fig. 3). The reason is that reference strategy 1 does not allow the unused capacity left by one tenant to be exploited by the other tenant, thus resulting in worse resource utilization. This effect is somehow compensated by reference strategy #2 (see Fig. 4b), which allows increasing the throughput in the regions where the offered load of one tenant is low. However, by comparing Fig. 4b with Fig. 3 it is also observed that the profit-based approach allows achieving a slightly higher throughput than the reference strategy #2. The reason in this case is that the profit-based approach does not consider the planned capacity as a strict limit in the admission but, instead, it considers a more flexible approach based on estimating the actual congestion probability of the cell.

To better quantify the throughput improvements achieved by the proposed approach, Figs. 5 and 6 compare the obtained throughput with the reference strategies and with the profit-based approach configured with $\Delta C_{q,s}=\Delta D_{q,s}=100$ as a function of the offered load of T1. The offered load of T2 is 20 Mb/s in Fig. 5, i.e. below the SLA_2 value, and 80 Mb/s in Fig. 6, i.e. above the SLA_2 value. It is observed in Fig. 5 that, for low loads of T1, and given that the load of T2 is also low, the total throughput is similar in all the cases. However, when the offered load of T1 increases beyond the SLA_1 value of 60 Mb/s, the throughput obtained by the profit-based approach becomes higher than the throughput achieved by the other techniques. Specifically, there is an increase of about 28% with respect to reference 1 and about 8% with respect to reference 2.



(a) (b)

Fig. 4 Total throughput obtained with the reference RAC strategies (a) Reference 1, (b) Reference 2.

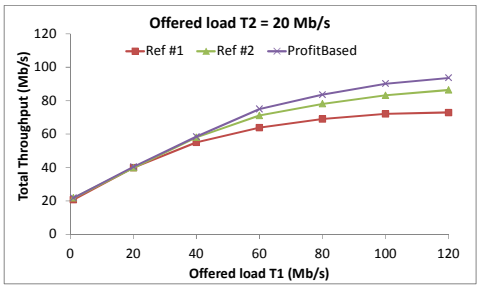


Fig. 5 Comparison in terms of total throughput achieved by the different strategies in the case that the offered load of T2 is 20 Mb/s.

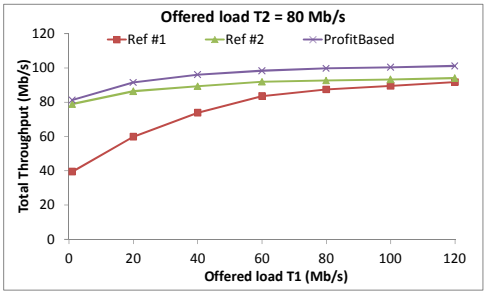


Fig. 6 Comparison in terms of total throughput achieved by the different strategies in the case that the offered load of T2 is 80 Mb/s.

The behaviour in Fig. 6 is similar than that of Fig. 5, with the exception that, for low loads of T1, the throughput of both the profit-based and the reference 2 technique exhibit almost twice the throughput of reference 1. This is due to the fact that reference 1 technique does not allow that T2 uses the capacity left by T1. In turn, for high loads of T1 the throughput obtained by the profit-based approach is about 10% and 7.5% higher than that of references 1 and 2, respectively.

V. Conclusions

This paper has proposed a flexible RAC algorithm to handle GBR traffic demands in multi-tenant radio access networks according to profit-related metrics directly derived from the SLA provisions. Remarkably, the decision-making accounts for the additional extra incomes that can be obtained

by the infrastructure provider when sporadically granting additional capacity to a tenant above its agreed SLA level and for the penalties incurred due to potential SLA breaches. Based on this, the admission control conditions have been formulated considering the estimated profit variations associated to the acceptance or rejection of a new QoS flow.

The algorithm has been evaluated by means of simulations, revealing the potential profit improvements that can be obtained through the proposed approach depending on the existing traffic mixes and the values of extra income and penalty terms that define the profit function. Specifically, results in a scenario with two tenants have shown that the profit is about 7.5 units/s for traffic mixes in which the offered load of one tenant is one third of its SLA while the offered load of the other tenant is twice its SLA, increasing this level to about 10 units/s when the unbalance between offered load and SLA becomes higher. In turn, when the penalty costs due to SLA breaches increase, the profit is reduced, being this particularly relevant for traffic mixes in which the offered load of both tenants is much higher than their SLAs. Besides, results have also revealed the throughput improvements that can be obtained with respect to two reference strategies.

Based on the considered developments in this paper, future work includes the optimisation of the decision making process exploiting analytical methods.

APPENDIX: ESTIMATION OF THE CONGESTION PROBABILITY

Congestion will occur whenever the number of required PRBs to satisfy the GFBR requirements of the all admitted GBR QoS flows, denoted as K_{TOT}, is higher than the number of available PRBs in the cell N. This can occur stochastically depending on the variations in the propagation conditions associated to the positions of the different UEs and the presence of shadowing losses. These effects will impact on the spectral efficiency achieved by each UE and therefore on the number of required PRBs.

Specifically, the number of required PRBs by one QoS flow with requirement GFBR is $K=GFBR/R$ where R is the bit rate per PRB, related with the actual spectral efficiency S as $R=S\cdot B$, and it is a random variable that depends on the propagation conditions experienced by the UE of the QoS flow. Then, the proposed methodology considers that, based on measurements collected from the different UEs, e.g. in terms of the wideband Channel Quality Indicator (CQI) distribution [20], it is possible to derive the probability density function (pdf) of the random variable $Y=1/R$, denoted as $f_Y(y)$. Correspondingly, the pdf of the number of required PRBs K by one QoS flow with GFBR is:

$$f_K(k) = f_Y\left(\frac{k}{GFBR}\right)\frac{1}{GFBR} \tag{8}$$

Then, assuming that, at a certain time, the number of admitted QoS flows of each service is vector $\mathbf{x} = \left(n_{1,1},...,n_{1,S_1},...,n_{Q,1},...,n_{Q,S_Q}\right)$ and that each QoS flow experiences independent propagation conditions, the aggregate number of required PRBs by all the QoS flows, K_{TOT}, is another random variable whose pdf is obtained by the convolution of (8) as many times as the number of QoS flows:

$$f_{K_{TOT}|\mathbf{x}}(k) = \left(\frac{1}{GFBR_{1,1}}\right)^{n_{1,1}} \cdots \left(\frac{1}{GFBR_{Q,S_Q}}\right)^{n_{Q,S_Q}} \cdot \left[\underbrace{f_Y\left(\frac{k}{GFBR_{1,1}}\right) * \ldots * f_Y\left(\frac{k}{GFBR_{1,1}}\right)}_{n_{1,1}}\right] * \ldots$$
$$\ldots * \left[\underbrace{f_Y\left(\frac{k}{GFBR_{Q,S_Q}}\right) * \ldots * f_Y\left(\frac{k}{GFBR_{Q,S_Q}}\right)}_{n_{Q,S_Q}}\right] \tag{9}$$

The congestion probability is thus given by:

$$P_{cong}(\mathbf{x}) = \Pr\left(K_{TOT} > N \middle| \mathbf{x}\right) = \int_N^\infty f_{K_{TOT}|\mathbf{x}}(k)\,dk \tag{10}$$

REFERENCES

[1] 3GPP TS 23.501 v15.2.0 "System Architecture for the 5G System; Stage 2 (Release 15)", June, 2018.

[2] R. Ferrús, O. Sallent, J. Pérez-Romero, R. Agustí, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration", IEEE Communications Magazine, May, 2018, pp.184-192.

[3] B-Y.Lee, G-H. Lee, "Service Oriented Architecture for SLA Management", 9th Int. Conf. on Adv. Comm. Technology, 2007

[4] O.Sallent, J. Pérez-Romero, R. Ferrús, R. Agustí, "Multi-tenant Mobility Control in Small Cells as a Service",Int. Symp. on Wireless Comm. Systems (ISWCS), Poznan, Poland, Sept. 2016.

[5] K. Samdanis, X. Costa-Perez, V. Sciancalepore, "From Network Slicing to Multi-Tenancy: The 5G Network Slice Broker", IEEE Communictions Magazine, July, 2016.

[6] R. Krebs, S. Spinner, N. Ahmed, S. Kounev, "Resource Usage Control in Multi-Tenant Applications", IEEE/ACM Int. Symp. on Cluster, Cloud and Grid Computing, 2014.

[7] S. Yi, A. Hameurlain, F. Morvan, "SLA Definition for Multi-tenant DBMS and its Impact on Query Optimization", IEEE Transactions on Knowledge and Data Engineering, 2018.

[8] 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.

[9] A. Aijaz, "Hap-SliceR: A Radio Resource Slicing Framework for 5G Networks with Haptic Communications", IEEE Systems Journal, 2017.

[10] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, "Network Slicing Games: Enabling Customization in Multi-Tenant Networks", IEEE INFOCOM, 2017.

[11] R. Kokku, R. Mahindra, H. Zhang, S. Rangarajan, "NVS: A substrate for Virtualizing Wireless Resources in Cellular Networks", IEEE/ACM Transactions on Networking, Vol. 20, No. 5, October, 2012.

[12] R. Mahindra, M. Khojastepour, H. Zhang, S. Rangarajan, "Radio Access Networks Sharing in Cellular Networks", 21st IEEE Int. Conference on Network Protocols (ICNP), Göttingen, Germany, October, 2013

[13] J. Pérez-Romero, O. Sallent, R. Ferrus, R. Agustí, "Admission Control for Multi-tenant Radio Access Networks", IEEE Int. Conference on Communicatins (ICC) - workshops, Paris, France, May, 2017.

[14] T. Guo, R. Arnott, "Active LTE RAN Sharing with Partial Resource Reservation", IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, USA, September, 2013.

[15] H. M. Soliman, A. Leon-Garcia, "QoS-Aware Frequency-Space Network Slicing and Admission Control for Virtual Wireless Networks", IEEE GLOBECOM, 2016.

[16] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, A. Banchs, "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization", IEEE INFOCOM 2017.

[17] 3GPP TS 38.300 v15.2.0, "NR and NG-RAN Overall Description; Stage 2 (Release 15)", June, 2018.

[18] 3GPP TR 36.814 v9.0.0, "E-UTRA; Further advancements for E-UTRA physical layer aspects (Release 9)", March, 2010.

[19] 3GPP TR 36.942 v12.0.0, "Radio Frequency (RF) system scenarios", September, 2014

[20] 3GPP TS 32.425 v15.0.0, "Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 15)", March, 2018.