Service-aware Network Slice Trading in a Shared Multi-tenant Infrastructure

Özgür Umut Akgül*†, Ilaria Malanchini†, Vinay Suryaprakash†, Antonio Capone*
*Dipartimento di Elettronica, Informazione e Bioingegneria Politecnico di Milano, Italy
Email: {oezguerumut.akguel, antonio.capone}@polimi.it
†Nokia Bell Labs, Stuttgart, Germany
Email: {ilaria.malanchini, vinay.suryaprakash}@nokia-bell-labs.com

***Abstract*—Maintaining service guarantees in a dynamic multi-tenant network, while ensuring an economically sustainable sharing platform, is a non-trivial problem. This paper, extending our previous work, develops a dynamic slicing and trading framework that can satisfy a variety of service guarantees. This framework not only determines the size of the network resource slices required for various active services, but it also adapts resource prices in accordance with the microeconomic laws of supply and demand. The proposed framework also ensures service continuity by learning the variations in the traffic mix as well as in the channel conditions, and by adjusting the slice assignments accordingly.**

I. Introduction

Stringent quality of service (QoS) requirements as well as lofty expectations of flexibility pose great challenges to 5G networks. One of the many technical solutions proposed – and widely accepted – is increasing network heterogeneity. However, in light of the steady decrease in network operator profits in last few years [1], this solution appears to pose a rather grave threat to the overall health of the mobile operator business. As shown in [2], increased heterogeneity and the demand for low service times decreases the profitability of operators and their impact is particularly severe on the smaller operators in the market. To alleviate this problem, the Organization for Economic Co-operation and Development (OECD) report [3] recommends various methods (and degrees) of infrastructure sharing among operators to increase operator profits as well as to ensure improved customer service.

The OECD report has admittedly lead to greater attention being paid to this topic. Works such as [4]–[6] focus on the comparisons between the technical aspects of sharing approaches like capacity or spectrum sharing. However, their technology specific focus (e.g., on LTE) makes it difficult to draw more generic conclusions from their findings. Malanchini et al. in [7] provide a generic (technology independent) resource sharing algorithm, but their algorithm is unable to cater to the flexibility guarantees that one expects in 5G networks. Although the OECD report, [3], and the references therein provide detailed economic analyses, only a handful of works deal with both the technical as well as the economic aspects. E.g., [8] and [9] investigate the relationship between the technical and the economic aspects, and provide an understanding of the tenants' (i.e., network operators') inclination to share as well as their related network costs. However, neither of these works provide a concrete techno-economic model. Another salient shortcoming is their strict adherence to state-of-the-art service level agreements (SLAs), which are intended to be fixed over a rather long time period (of months/years). This proves to be a major hurdle in allowing the network operators (or tenants) to adapt their resource consumption to the traffic traversing their network. As a result, operators in such a framework can often find themselves in situations of resource surplus, where they incur unnecessary expenditure by paying for unused resources, or resource scarcity, where they risk having dissatisfied customers. To address this issue, our previous work [10], while still relying on state-of-the-art SLAs and considering active sharing, provides a techno-economic model that permits short-term dynamic resource trading (i.e., on the order of seconds/minutes), wherein the mobile virtual network operators (MVNOs) can buy or sell resources based on their customers' needs and, as a consequence, deviate from the original SLA to a certain extent. While the idea proposed in [10] works quite well when the MVNOs happen to choose similar types of services, it struggles to accommodate scenarios wherein the service heterogeneity is large.

As detailed in [11] and [12], slicing the network and using dedicated resources for different services is deemed beneficial for achieving the service guarantees required by the heterogeneous applications of future networks (5G and beyond). However, as explained in [13], service scalability, adaptability to varying channel conditions and traffic types, and dynamic resource allocation are also of crucial importance within a particular network slice itself. While [14] provides an auction based pricing and dynamic slicing framework, it neither considers fluctuations in the channel quality nor variations in the traffic mix. Additionally, the applicability of the algorithm in a competitive shared infrastructure scenario is also unclear. [15] and [16] provide other dynamic slicing approaches, but they also ignore the fact that the algorithm needs to be able to adapt to varying channel conditions. The main reason for [14]–[16] not taking these aspects into consideration is because they are also reliant on traditional (long-term) SLAs for network slicing. In order to address the aforementioned issues while ensuring the profitability of stakeholders, in this work, we propose an automatic resource slicing algorithm, which works on short time scales and can provide the desired service guarantees, while exploiting the

economic benefits of infrastructure sharing. We assume that there are only two stakeholders in our scenario, namely: the *infrastructure provider* who owns the physical resources; and the *tenants* who do not own any physical resources, but trade resources they obtain from the infrastructure provider in order to provide for their designated services. The dynamic pricing structure proposed in this paper also allows the infrastructure provider to collect revenue, proportional to the performance expectations of the tenants, and use it for the infrastructure expansion necessary to satisfy the service guarantees.

The main contributions of this work can be summarized as follows:

- Automated network slice adjustment in order to guarantee a certain quality for each service type;
- Tenant centric resource provisioning – scaled according to the quality expectations, the channel conditions, and the mix of traffic;
- Short time scale (i.e. on the order of seconds) infrastructure sharing in a multi-tenant network.

The remainder of the paper is organized as follows: Section II contains the system model and the main assumptions. Following the system model, the optimization model is presented in Section III. In Section IV, the behavior and the validity of the optimization model are investigated through simulations, and Section V concludes the paper.

II. System Model

In this study, the downlink of a base station is shared by a set of tenants denoted by M. The base station is supplied (and operated) by an infrastructure provider and the tenants use the obtained resources to accommodate a set of active users, K, whose cardinality is given by $|K|$. In the scenario considered, the active users are distributed among tenants, and the subset of users belonging to a tenant $m \in M$ is given by $K_m \subseteq K$. As commonly practiced when dealing with resource allocation algorithms, time is discretized and separated into time slots (represented by n). The total number of slots contained in the entire time period of operation (during which the optimization is to be carried out) is denoted by N. For the sake of clarity and continuity, this work coopts the notations as used in [10]. Namely, the fraction of resources assigned to a user k at time slot n is represented by $x_k[n]$. The achievable rate for a user k during the time slot n is denoted by $r_k[n]$. The users are assumed to use a single service at each time instance.

To regulate the slicing of resources and the manner in which they are shared, we assume SLAs between the tenants and the infrastructure provider. The latter, i.e. the infrastructure provider, regulates the initial sharing values and prevents unfair scenarios, wherein a wealthier tenant tries to monopolize the market by artificially inflating the sharing parameters. However, the tenants are free to renegotiate their SLAs to fulfil their performance expectations and adapt to the fluctuations in their respective traffic.

In this paper, the SLA based sharing ratio for each tenant is represented by $S_m \in [0, 1)$ and indicates the fraction of resources assigned to tenant m. Notably, without introducing an added degree of flexibility, this would correspond to the static sharing scenario, where each tenant m obtains a resource share equal to S_m. The ability to trade resources is enabled by introducing Δ_m denoting a maximum deviation from the initial value S_m. It is through this parameter that the tenant has the opportunity to either trade unused resources or acquire additional resources from tenants who have a resource surplus. However, these trades are limited by the average deviation from S_m, represented by $\epsilon_m[n]$, which lies within the interval $[-\Delta_m, \Delta_m]$. Namely, the average deviation is calculated at every time slot n for a time window w (of length W), by considering the current and previous time slots from the beginning of the window. This implies that the time span over which the average is calculated varies at every n, and this time span is equal to $(a+1)$ time slots, where $a \equiv (n-1 \mod W)$. The sharing parameters (S_m, Δ_m) are negotiated at the end of each time window and are held constant for the window that follows. We assume that each tenant aims to fulfil its own utility target[1]. The difference between a tenant's utility target, denoted by $U_{\text{th},m}$, and the utility actually obtained during a given time slot is represented by $\xi_m[n]$.

To model the economic aspects of slicing, we introduce B_m, which denotes the budget per time slot for tenant m. Furthermore, we assume that each tenant pays a cost per assigned resource, which is composed of three parts, namely: capital expenditure (CapEx) represented by C_{ca}; operational expenditure (OpEx) denoted as C_{op}; and finally, the pressure cost given by C_{pre}. As discussed in [10], the pressure cost links the tenants' gaps between the desired utility and the utility achieved (i.e., $\xi_m[n]$) with the revenue necessary for expansion.

A. Assumptions

A couple of assumptions worth explicitly mentioning are as follows:

1) The tenants' gap, $\xi_m[n]$, provides a clear understanding of the capacity expansion required to reach their respective performance expectations.
2) All the resources are identical and services have no choice in terms of resource block assignment.

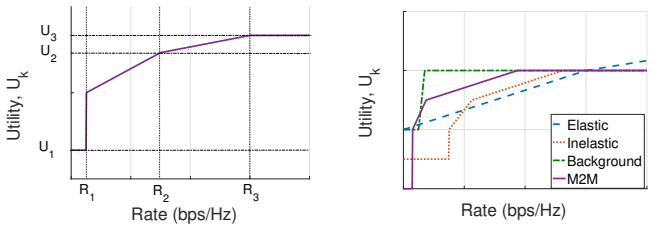
B. Utility Functions

We assume that the utility function of each tenant directly depends on the QoS of their respective users. Namely, it is a function of the average rate achieved within $[n-a, n]$ (i.e., the current time window), which is defined as

$$R_k[n] = \left(\frac{1}{(a+1)} \sum_{i=n-a}^{n} x_k[i] r_k[i] \right).$$

In order to incorporate the heterogeneity of services, we first define a generic function $U_k(R_k[n])$ (known henceforth as the "utility function") as illustrated in Fig. 1(a). This function – shaped by six parameters, namely, R_1, R_2, R_3, U_1, U_2,

[1]Here, utility is used as a generic synonym for the key performance indicators of a particular tenant and will be clarified subsequently.



(a) Generic utility function. (b) Exemplary utility functions.

Fig. 1. Generic utility function (left) and exemplary utility functions per service type (right).

and U_3 – can be used to describe a variety of services and their requirements as described in the paragraphs below. R_1 denotes the minimum rate required by a service if it has to be active. If the rate R_1 is achieved, the utility function takes the value zero. However, if a rate less that R_1 is achieved, the utility function takes the value U_1. R_2 is used to represent the rate necessary to achieve '*standard quality*' for which the utility function takes the value U_2. Note that the definition of standard quality depends on the service in use. We call R_3 the *saturation* point and use it to denote the rate that enables the utility function to attain its maximum value U_3. Note that although the utility function is only based on the achieved rate, the latency required by a service is implicitly taken into account by considering that the proposed utility is evaluated by considering the cumulative rate achieved within the current time window w, i.e. $R_k[n]$. Therefore, the latency is indirectly constrained by the length of the time window, W.

We then categorize the heterogeneous services envisioned in 5G networks into 4 broad categories, namely: elastic services, inelastic services, background services, and machine to machine (M2M) services. In what follows, we describe how a utility function for each of these categories can be obtained from the generic utility function in Fig. 1(a).

1) Elastic Services: By definition, elastic services do not have strict delay or rate constraints. Therefore, $R_1 = 0$ for this type of service. Moreover, since the service requirements are fairly lax, the slope of the utility function between R_1 and R_2 (cf. Fig. 1(a)) can be fairly gradual. Furthermore, since elastic users can usually 'take all they can get', the utility function does not really have a saturation point, i.e., theoretically $R_3 \longrightarrow \infty$ – albeit very slowly. This definition also provides tenants the possibility to increase their utility function's value by increasing the elastic rates. A visualization of the utility function for this service is given by the curve with the dashed blue line in Fig. 1(b).

2) Inelastic Services: A classic example for this type of service is video streaming. In particular, inelastic services need relatively large achieved rates even to guarantee service availability. Therefore, R_1 is assumed to be quite large. To reflect the fact that users are sensitive to variations in video quality, especially when it is low (e.g., the perceived difference between 144p and 720p videos), the slope of the utility function between R_1 and R_2 (cf. Fig. 1(a)) is assumed to be quite steep. However, since changes in the quality are less perceptible when quality is already high (e.g., the perceived difference between 720p and 1080p videos), the slope of the utility function between R_2 and R_3 (cf. Fig. 1(a)) is assumed to be gradual. In Fig. 1(b), the slope for this region (see the dotted red curve) is assumed to be same as that of the curve for elastic traffic. For such services, we assume the existence of a saturation region which corresponds to the fact that improving the achieved rates beyond what is required for the highest class of video transmission is unfruitful.

3) Background Services: This type of service is assumed to require considerably low rates and as soon as those rates are achieved, the utility function rapidly reaches the saturation point. As a result, the points R_2 and R_3 in Fig. 1(a) coincide, leading to the utility function looking like the curve with the dashed green line in Fig. 1(b). Notably, for such services, we assume the minimum value of the utility function U_1 to be zero, and thereby, indicating that the service is not critical and should not be prioritized over other services.

4) Machine to Machine (M2M) Services: M2M communications are the broadest group of services among the ones considered here. Thus, modeling their characteristics is quite a challenge. In this work, three major groups of M2M devices are considered and we assume that each M2M service request is a mix containing all three of them. Hence, the utility function shown in Fig. 1(b) (cf. the maroon curve) reflects this mix and resembles the generic utility function (see Fig. 1(a)) closely. The point R_1 in Fig. 1(a) corresponds to the minimum rate requirement for emergency services and the requirements of low rate and delay sensitive devices are modeled by the curve in the interval $[R_1, R_2]$. An example of devices requiring this type of service are sensors that send traffic periodically. For this region, we assume quite a steep curve within the interval $[R_1, R_2]$ (compare Fig. 1(a) and Fig. 1(b)) in order to prioritize the delivery of such messages. Additionally, the interval $[R_2, R_3]$ models rate sensitive devices, which are delay insensitive, and for whom the slope of the utility function can be gradual. An example of such a device is sensor aggregation node, wherein a large amount of sensor data is transmitted over a relatively large period. Lastly, as in the case of inelastic services, since providing a rate in excess of what is required does not bring any added benefits, the maroon curve in Fig. 1(b) also reaches a point of saturation (cf. R_3 in Fig. 1(a)).

III. Formulation and Analysis of the Model

A. Problem Formulation

Using the notations defined in Section II, the base station's scheduler solves the optimization problem described in (1a)-(1h) in order to perform real-time resource allocation, carry out sharing negotiations, and calibrate the dynamic pricing. Since the problem is intended to be solved in real-time, the achievable rates are not known to the scheduler. Thus, negotiating the sharing ratio for the upcoming time windows is quite a difficult hurdle to overcome. In order to realize this goal, the optimization is divided into two sub-problems P1 and

$$\min_{x_k} \quad f(\xi_m[n], S_{\max}) \tag{1a}$$

$$\text{s.t.} \quad S_{\max} \geq \max(S_m,\ 1 - S_m), \quad \forall m \in M, \tag{1b}$$

$$U_{\text{th},m} - \sum_{k \in K_m} U_k\left(R_k[n]\right) \leq \xi_m, \forall m \in M, \tag{1c}$$

$$|\epsilon_m[n]| \leq \Delta_m, \quad \forall m \in M, \tag{1d}$$

$$\sum_{i=n-a}^{n} \left(S_m(C_{\text{ca}} + C_{\text{op}}) + \epsilon_m[i]C_{\text{op}} + \xi_m C_{\text{pre}}\right) \leq B_m(a{+}1),$$
$$\forall m \in M, \quad a \equiv (n-1 \mod W), \tag{1e}$$

$$0 \leq \Delta_m \leq \max(S_m,\ 1 - S_m), \quad \forall m \in M, \tag{1f}$$

$$\sum_{k \in K} x_k[n] \leq 1, \quad x_k[n] \geq 0, \quad \forall k \in K, \tag{1g}$$

$$\sum_{m \in M} S_m \leq 1\ , \quad S_m \geq 0, \quad \forall m \in M, \tag{1h}$$

P2. The details of these problems are presented in Section III-B, while the remainder of this subsection describes the entire optimization problem (cf. (1a) - (1h)).

The continuous objective function depends on two factors, namely, ξ_m and $S_{\max}$. The first part minimizes the total gap of the tenants, ξ_m. By minimizing the total gap, instead of focusing on the tenants' individual gaps, a relaxation of the optimization problem is achieved. By using this approach, the optimizer can prioritize users with the best channel conditions and increase spectral efficiency. The second factor, $S_{\max}$, enables fairness among tenants in terms of their initial SLA based share of the resources, i.e., S_m. Constraint (1b) ensures that $S_{\max}$ is lower bounded by the larger of the two values between the amount of resources available to a tenant (S_m) and the remaining resources ($1 - S_m$). If one assumes the budgets of all tenants to be feasible, constraint (1b) ensures that resources are fairly (and equitably) distributed among all the tenants.

The primary constraint ensuring service-based resource slicing is presented in (1c). Namely, this constraint ensures that a given tenant's gap is the difference between the tenant's utility target (i.e., $U_{\text{th},m}$) and the achieved utility. Though visually similar to the formulation in [10], note that a tenant's achieved utility – in this formulation – is calculated as the sum of the utilities of all the tenant's services catered to[2]. The individual service utilities are computed using the utility functions illustrated in Fig. 1(b) and the average rate achieved by a particular service within the time window w, i.e. $R_k[n]$.

Constraint (1d) bounds the values taken by the maximum average deviation, $\epsilon_m[n]$, to those that lie within the interval $[-\Delta_m, \Delta_m]$. Constraint (1e) sets the budget constraint per tenant. In particular, for each time slot n, each tenant has a fixed budget. However, the right-hand side of (1e) allows tenants to use the unused budget from the previous time slots. The tenants have the flexibility to adjust their budget according to their users' channel conditions and their own long term fiscal strategies. On the left-hand side (LHS) of (1e), the total expenses incurred by a tenant is calculated. The first term represents the 'ownership' cost of the resources, i.e., each tenant incurs a CapEx and OpEx in proportion to their sharing ratio S_m. The second term of the LHS of (1e) is included to ensure that the tenants can adjust their resource use based on their own traffic estimates and QoS targets. If a particular tenant has surplus resources and wants to sell some, this term takes a negative value indicating that the total expenditure decreases in proportion to the OpEx. If, on the other hand, the tenant wants to buy resources due to a resource insufficiency, this term takes a positive value and the total expenditure increases. Finally, the last term on the LHS of (1e) is the pressure cost, which reflects the market driven price fluctuations as well as provides a means to collect the additional revenue required for future network capacity expansion.

Constraint (1f) sets an upper limit for the maximum deviation Δ_m that a given tenant can choose. This constraint ensures that a tenant cannot trade resources they do not own, and conversely, try to buy resources that the infrastructure provide does not yet have. Constraints (1g) and (1h) ensure that the total number of resources assigned cannot be larger than the system capacity and that the sum of the resources owned by individual tenants are not larger than the total number of resources available, respectively. Note that, for the sake of readability, all the constraints are given in their non-linear form. However, they can be linearized using standard techniques. The same applies to the proposed utility function, which has been expanded from the generic form presented in Fig. 1 during the solution of the optimization problem.

B. Applied Algorithm

As mentioned earlier, the optimization problem is divided into two parts, i.e., P1 and P2, to facilitate real-time applicability. The two sub-problems deal with slightly different optimization goals, while using each other's (previous) results as inputs. Formally, we have:

$$\text{P1} := \begin{cases} \text{(1a)} & \min\limits_{\xi_m,\ x_k,\ \epsilon_m} \ \sum\limits_{m \in M} \xi_m[n] \\ \text{s.t.} & \text{(1c)(1d)(1e)(1g)} \end{cases}$$

$$\text{P2} := \begin{cases} \text{(1a)} & \min\limits_{\substack{\xi_m,\ x_k,\ S_m,\\ \Delta_m,\ \epsilon_m}} \ \sum\limits_{m \in M} \xi_m[n] + S_{\max} \\ \text{s.t.} & \text{(1b) – (1h)} \end{cases}$$

P1, by taking S_m and Δ_m as input, finds the optimal resource allocation (i.e., $x_k[n]$) that minimizes the total gap between each tenant's target utility and the utility they achieved (i.e., $\xi_m[n]$). This optimization is run at each time slot within the time window w and the problem P2 is solved at the end of each time window w.

[2]For brevity and clarity, the utility function is presented in its aggregated form. The complete model can be found at https://tinyurl.com/akgul-model.

TABLE I
UTILITY PARAMETERS AND VALUES PER SERVICE TYPE.

	Elastic	Inelastic	Background	M2M
R_1	0 bps/Hz	0.1 bps/Hz	0.05 bps/Hz	0.01 bps/Hz
R_2	1.083 bps/Hz	0.225 bps/Hz	0.07 bps/Hz	0.075 bps/Hz
R_3	∞	0.55 bps/Hz	0.07 bps/Hz	0.4 bps/Hz
U_1	0	-0.5	0	-1
U_2	1	0.7	1	0.7
U_3	∞	1	1	1

The problem P2, then, uses the knowledge of all the rates actually achieved during the previous window (i.e., the window that just ended) to determine the optimal resource allocation for a given traffic mix and known channel states. The values of the optimal S_m and Δ_m determined are then used to update the input values for P1 in the upcoming time window.

IV. Simulation Results

The simulation setup and their results are discussed in the following subsections.

A. Parameters and investigated scenarios

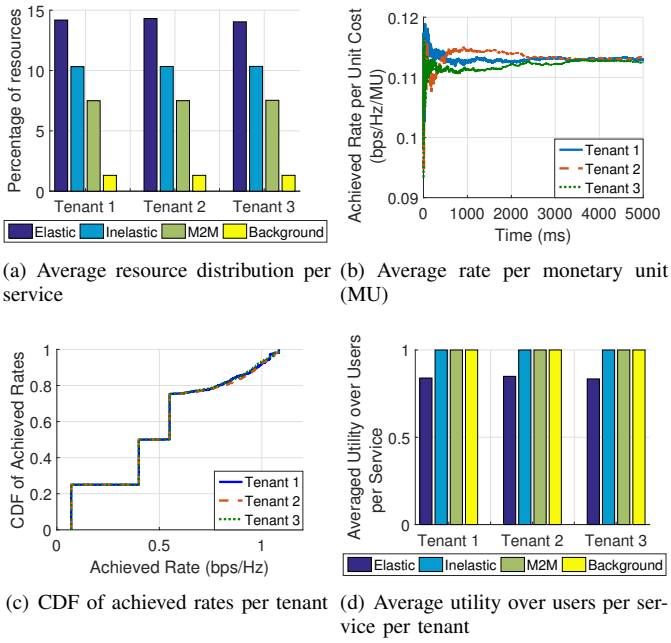
We consider the downlink of a single base station shared by 3 tenants, i.e., $M = 3$. The total number of active users is $|K| = 24$ and they are distributed equally among the 3 tenants, i.e., $|K_m| = 8,\ \forall m \in M$. Users are uniformly distributed within the coverage area of the base station and are active for the entire duration of the simulation. At each time window, w, a new set of active users, which replaces the set of active users in the previous window, is generated in the coverage area of the base station. The tenants provide the four service types described in Section II-B, where the parameters take values as reported in Table I. The number of users requesting each type of service is equal to $|K_m|/4$ for each tenant. Furthermore, each tenant has a utility target equal to $U_{\text{th},m} = |K_m|$. All the budgets and costs are normalized to take values between 0 and 100 (namely, $C_{\text{ca}} = 50, C_{\text{op}} = 50, C_{\text{pre}} = 16.66, B_m = 100,\ \forall m \in M$). Note that the values for costs and the budgets mentioned here are for purely illustrative purposes and are used with the sole intention of studying the characteristic behavior of the framework.

The channel between the user and the base station is modeled using a frequency-flat block fading channel with i.i.d. Rayleigh coefficients, which implies exponentially distributed channel gains, denoted by $|h_k[n]|^2$. Using the Okumura-Hata propagation model, the average signal-to-interference-plus-noise ratio (SINR) for user k, SINR_k, is computed as:

$$\text{SINR}_k = P d_k^{-\alpha}/(\sigma^2 + I_0),$$

where P is the transmit power (in Watts), d_k is the user's distance to the base station (in meters), α is the path-loss exponent, σ^2 is the thermal noise, and I_0 is the average interference power. From which, the instantaneous SINR of user k at a time slot n is calculated as $\gamma_k[n] = \text{SINR}_k |h_k[n]|^2$. The users' spectral efficiency at a time slot n is calculated as

$$r_k[n] = \log_2(1 + \gamma_k[n]).$$



(a) Average resource distribution per service

(b) Average rate per monetary unit (MU)

(c) CDF of achieved rates per tenant

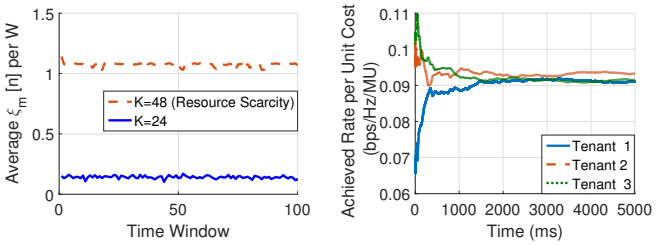
(d) Average utility over users per service per tenant

Fig. 2. Equitable distribution scenario with $K = 24$.

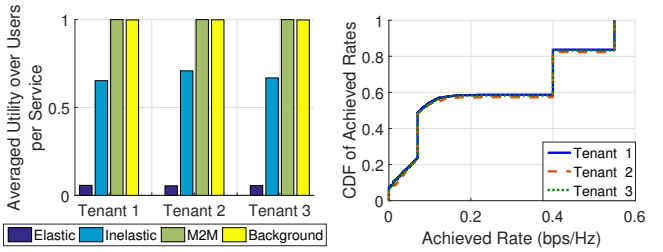
The findings in [10] also showed that the size of the time window plays a significant role in the ability of the framework to adapt to network fluctuations. Using a metric called the "Relative Distance to Optimum" (RDO), which described how close the selected parameters are to their optimum values, [10] showed that the best value was $W = 100$ ms. However, given that [10] considered an optimization framework wherein only a single service type existed, its complexity was significantly lower than the scenario considered here, where multiple service types need to be dealt with simultaneously. Since a comprehensive analysis of the impact of the window length W is still underway during the authorship of this work, we set $W = 50$ ms based on an empirical evaluation. The total duration of the simulation is 5000 time slots (i.e., $N = 5000$), where the length of each time slot is assumed to be 1 ms.

B. Equitable distribution scenario

Fig. 2 depicts the case where the set of active users are distributed equally among the tenants, who have the same initial sharing ratios. Fig. 2(a) shows the percentage of resources allocated to each of the service types per tenant, wherein one readily observes that there is an equitable share of resources. The instantaneous rates achieved per unit cost are given in Fig. 2(b). This figure along with Fig. 2(c), which depicts the cumulative distribution function (CDF) of the rates achieved per tenant, corroborates the fact that the tenants pay a similar price for obtaining a similar throughput; in essence, 'one gets what one pays for'. The variations seen early on during the simulation window are due to variations in the channel qualities of individual users. However, we observe that, as one starts to consider larger observation set, the three tenants

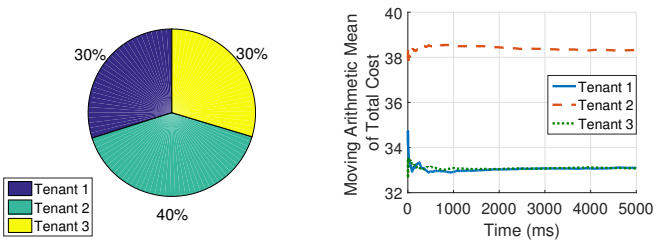


(a) Average gap of tenants over W (b) Average achieved rates per monetary unit (MU)

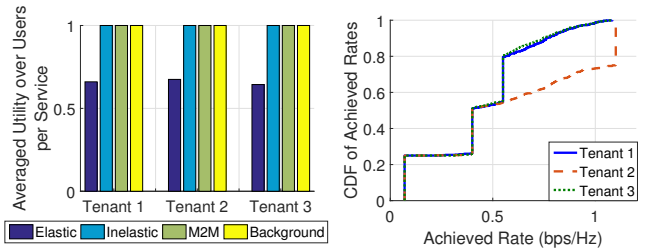


(c) Average utility over users per service per tenant (d) CDF of achieved rates per tenant

Fig. 3. Results for the resource scarcity scenario.



(a) Resource distribution per tenant (b) Moving arithmetic mean of total cost per tenant



(c) Average utility over users per service per tenant (d) CDF of achieved rates per tenant

Fig. 4. Results for the guaranteed services scenario.

obtain similar rates per monetary unit (MU) – as evidenced by the overlap of the curves beyond 3000ms in Fig. 2(b).

Finally, Fig. 2(d) plots the averaged sum of the utility achieved per service type for each of the tenants. The fact that the elastic services achieve the lowest average utility indicates that elastic services have the lowest priority and that they are assigned only when the other 3 service types no longer need resources, or have poor channel conditions.

C. Effects of resource scarcity

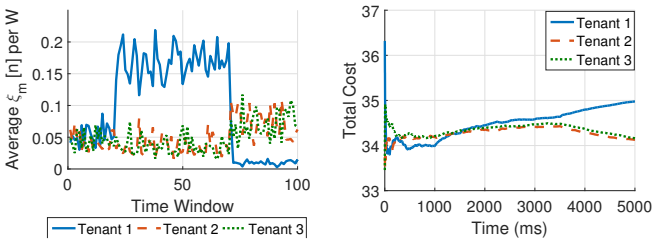
The effects of resource scarcity, documented in Fig. 3, are studied by increasing the number of active users. Fig. 3(a) shows the increase in the average difference between the utility target of the tenants and the utility they actually achieved over a time window, when the number of active users are doubled. Fig. 3(b), when compared with Fig. 2(b), illustrates a decrease in the average rate per unit cost. This behaviour can be understood as a decrease in the purchasing power of tenants due to an increase in the pressure cost, driven in turn by resource scarcity.

Fig. 3(c) shows the average sum of utility per tenant, demonstrating that the prioritization among service types still works efficiently and is unaffected by resource scarcity. We see that the framework continues to adhere to the priority set by the utility function design and tries to cater to all service types to the greatest extent possible. Finally, Fig. 3(d) plots the CDF of the rates achieved per tenant and shows that, despite being faced with situations of resource scarcity, the tenants pay a similar price for obtaining a similar throughput. The framework, therefore, ensures that all tenants are charged fairly for the resources they seek to purchase.

D. Guaranteed Services

An important use of network slicing is to ensure service guarantees. This also implies that service guarantees in one slice should have no perceptible effects on the service guarantees in other slices. This aspect is examined by doubling the rates required by the inelastic users of tenant 2. This increase also represents a case study, wherein one of the tenants promises a higher quality to their users than the others. These results are illustrated in Fig. 4. The average distribution of resources among tenants are given in Fig. 4(a), while Fig. 4(b) shows the moving arithmetic mean of total cost per tenant. As long as the tenants have sufficient budgets, the framework first satisfies the prioritized services (i.e., inelastic, M2M, and background services), regardless of the quality expectations of the tenants. Subsequently, the non-prioritized services (viz. elastic services) are satisfied in a fair manner. Consequently, the elevated quality expectations of second tenant do not effect the achieved quality of the critical services of other tenants. However, the tenant with a high quality target pays higher cost in comparison to the other tenants.

Fig. 4(c) shows that when tenants increase their quality expectations (i.e., increase the values of R_1, R_2, and R_3), there is no effect on the other services except for elastic traffic. However, this is reasonable since elastic traffic has the lowest priority. Fig. 4(c), also indicates that average utility obtained for a given tenant's users per service type continues to remain equitable even if one of the tenants increases their utility target for a specific service type. Finally, Fig. 4(d) shows the CDF of the rates achieved per tenant. This figure demonstrates that the second tenant is able to obtain the higher rates its users require. Note that tenant 2 is able to obtain higher rates only because



(a) Average gap of tenants over W (b) Tenants' total cost per n

Fig. 5. Framework's adaptability to the changes in the channel condition.

it can afford to pay for the additional resources required. Furthermore, we also observe that the CDFs of the other two tenants, whose requirements remained unaltered, have the same behavior. Therefore, this illustrates that the framework is able to cope with the increased demands of one of the tenants without affecting the equitable distribution of resources among the other tenants.

E. Adaptability to varying the channel conditions

Fig. 5 demonstrates our framework's ability to reshape the network slices according to variations in channel quality and the total expenses incurred by the tenants for the resources they obtain. In the scenario considered, all three tenants – at the beginning of the simulation – have the same statistical properties for the channel state distribution. During the 20^{th} time window (i.e, $w = 20$), path-loss exponent α of the users belonging to the first tenant is decreased and thereby, results in a corresponding decrease in the rates they achieve (i.e., $R_k[n]$). This decrease manifests itself as an increase in the average gap, $\xi_m[n]$, during a given time window as seen in Fig. 5(a). The change in the path-loss exponent mainly affects elastic services, since the other services are prioritized over elastic service by design. Fig. 5(b) illustrates the moving arithmetic mean of the tenants' costs over the simulation time. As long as the first tenant faces a larger gap due to poor channel quality, its total cost increases, while the costs of the other tenants remain fairly stable.

So far, the $U_{\text{th},m}$ values for all tenants are assumed to be equal – implying that their respective channel qualities play a central role in determining the inter-tenant resource distribution. In order to observe the behavior of the framework when tenants increase their utility targets to counteract the effects of bad channel quality, we assume that the first tenant increases its utility target $U_{\text{th},1}$ to $1.2|K_m|$ at $w = 70$ – denoted by a sharp dip in the blue curve in Fig. 5(a). This results in an increase in the total expenses of the first tenant as seen in Fig. 5(b). This leads us to conclude that, as long as a given tenant's budget is planned with a large enough[3] margin for 'contingencies', the tenant has the ability to satisfy its users by compensating for bad channel conditions by an overall increase in expenditure.

[3]Note that the budget per time slot is 100, while the expenses in Fig. 5(b) barely exceed 36.

V. Conclusion

This paper provides a framework that enables automatic network slice adjustment based on a tenant centric resource provisioning, which allows tenants to retain their autonomy in setting their quality targets. It provides a structure within which the slice sizes allocated to tenants can be adapted dynamically on short time scales based on the channel conditions faced by the tenant's users, the tenant's traffic mix, and their individual budget considerations. Dynamic network slice scaling in this framework is achieved by allowing tenants to trade unused resources and thereby, reduce expenditure. Simulations also show that this framework ensures that changes to service guarantees in one slice have no perceptible effects on the service guarantees in other slices.

Acknowledgment

This work is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 643002.

References

[1] CISCO, "Cisco Visual Networking Index: Forecast and Methodology, 2015-2020," 2015.

[2] H. Suomi, A. Basaure, and H. Hammainen, "Effects of capacity sharing on mobile access competition," in *21st IEEE International conference on Network Protocols (ICNP)*, Oct 2013, pp. 1–6.

[3] OECD, "Wireless Market Structures and Network Sharing," 2014. [Online]. Available: http://dx.doi.org/10.1787/5jxt46dzl9r2-en

[4] Y.-T. Lin, H. Tembine, and K.-C. Chen, "Inter-operator spectrum sharing in future cellular systems," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2012, pp. 2597–2602.

[5] A. P. Avramova and V. B. Iversen, "Radio access sharing strategies for multiple operators in cellular networks," in *IEEE International Conference on Communication Workshop*, June 2015, pp. 1113–1118.

[6] J. S. Panchal, R. Yates, and M. M. Buddhikot, "Mobile network resource sharing options: Performance comparisons," *IEEE Transactions on Wireless Communications,*, vol. 12, no. 9, pp. 4470–4482, 2013.

[7] I. Malanchini, S. Valentin, and O. Aydin, "Wireless resource sharing for multiple operators: Generalization, fairness, and the value of prediction," *Computer Networks*, vol. 100, pp. 110 – 123, 2016.

[8] I. Malanchini and M. Gruber, "How operators can differentiate through policies when sharing small cells," in *IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015.

[9] R. Berry, M. Honig, T. Nguyen, V. Subramanian, H. Zhou, and R. Vohra, "On the nature of revenue-sharing contracts to incentivize spectrum-sharing," in *IEEE INFOCOM*, 2013, pp. 845–853.

[10] O. U. Akgül, I. Malanchini, V. Suryaprakash, and A. Capone, "Dynamic resource allocation and pricing for shared radio access infrastructrue," in *2017 IEEE International Conference on Communications (ICC), to appear*, 2017. [Online]. Available: https://tinyurl.com/Akgul-icc2017

[11] China Mobile Communications Corporation, Huawei Technologies, Deutsche Telekom, and Volkswagen, "5G Service-Guaranteed network Slicing White Paper," 2017.

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

[13] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource Slicing in Virtual Wireless Networks: A Survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462 – 476, 2016.

[14] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing in 5G: an auction-based model," in *2017 IEEE International Conference on Communications (ICC)*.

[15] M. I. Kamel, L. B. Le, and A. Girard, "LTE wireless network virtualization: Dynamic slicing via flexible scheduling," in *IEEE Vehicular Technology Conference (VTC Fall)*, Sept 2014, pp. 1–5.

[16] X. Ting, P. Zhiwen, L. Nan, and Y. Xiaohu, "Inter-operator resource sharing based on network virtualization," in *International conference on wireless communication signal processing (WCSP)*, 2015, pp. 1–6.