Ref. Ares(2017)2423262 - 12/05/2017

D2.2 WP2 Research Report I

Project Name: Anticipatory Networking Techniques in 5G and Beyond

Acronym: ACT5G

Project no.: 643002

Start date of project: 01/05/2015 **Duration:** 48 Months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions.

Document Properties

Document ID	EU-H2020-MSCA-ITN-2014-643002-ACT5G-D1.2
Document Title	D2.2 WP2 Research Report I
Contractual date of delivery to REA	Month 24
Lead Beneficiary	Politecnico di Milano (PoliMI)
Editor(s)	A. Capone, M. Cesana – PoliMI
Work Package No.	2
Work Package Title	Reaction Techniques
Nature	Report
Number of Pages	27
Dissemination Level	**PUBLIC**
Contributors	LiU: V. Angelakis, D. Yuan POLIMI: A. Capone, M. Cesana Bell labs: I. Malanchini
Version Number	1

Contents

0 Executive Summary

This report details the work progress within work package two Reaction Techniques of the ACT5G project. More specifically, the document provides information of the conducted and expected work of early-stage researcher (ESR) three and four. The document first gives an overview of the focus area and research topics. Technical details of the work are then presented by means of research paper to be published in 2017.

1 Work Plan and Progress of ESR 3

The research task of ESR 3 is to investigate resource optimization techniques for 5G networks. More specifically, the line of research consists of the design of schemes for resource allocation among the users, in particular for scenarios with mixed types of traffic requirements, and the optimization in software-defined networks (SDNs). In the study, 5G-specific assumptions and new transmission techniques are to be addressed in the system models.

In November 2015, Wenjian Li was recruited as ESR 3 of the ACT5G project. Together with the academic and industrial supervisors, a list of candidate topics was developed, including resource allocation for massive MIMO, millimeter Wave systems, cooperative non-orthogonal multiple access (NOMA), and flexible frame structure for 5G. Wenjian Li came up with a survey of state-of-the-art of these research fields. Then the ESR has narrowed down the topic of scheduling with scalable transmission time interval (TTI). While working in this direction, Wenjian Li requested to be detached from the project in June 2016, but continue as a regular (faculty-supported) PhD student, by personal preference, and the request was accepted by the project consortium.

The recruitment of the new ESR 3, Emmanouil Fountoulakis, was made in August 2016. Emmanouil Fountoulakis initiated his PhD study by pursuing the topic under study by the project: resource allocation with scalable TTI. The basic idea of scalable TTI is to dynamically adjust the TTI length to the service requirements. A short TTI reduces delay but is less resource efficient for providing high throughput. Hence, for scenarios where both mission critical communications (MCC) and conventional mobile broadband (MBB) services are both present (and arrive dynamically), optimizing the TTI and channel allocation for each scheduling instance is a highly relevant research problem.

Within a few months' time, Emmanouil Fountoulakis has developed system model, optimization formulation, and solution algorithm for channel allocation with scalable TTI. In the system model, a utility function is designed to address the drop of service and system throughput. The simulation results show that optimizing scalable TTI can address better this trade-off than fixed TTI. The work has resulted in a conference paper, accepted for presentation and publication by the Resource Allocation, Cooperation and Competition in Wireless Networks (RAWNET) workshop of the 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT) that will take place in May, 2017. A copy of the paper is reported in the Appendix.

As the next step, the research will extend the concept of scalable TTI in the direction of both the time and frequency domains, such that resource allocation accounts for the dependency between TTI length and subcarrier spacing (or bandwidth). This direction is very well in line of the emerging technique of flexible transmission frame structure for 5G.

2 Work Plan and Progress of ESR 4

A multitude of applications, driven in part by the Industry 4.0 initiative, are envisioned for future networks (5G and beyond). Most of these applications require not only high data rates, but also low latencies. One of the potential solutions to this problem is considered to be denser and more heterogeneous network deployments. This, however, places an enormous strain on the already decreasing profitability of mobile operators, and thereby, necessitates a change in their current business modus operandi. One of the solutions proposed to cope with increasing operational costs and decreasing profitability is Infrastructure Sharing. As the name suggests, this idea proposes that mobile network operators (MNOs) share a common infrastructure in order to reduce their capital and operational expenditure as well as to offer their customers better prices, a larger number of services, and a better quality of service.

The research activities of the Early Stage Researcher 4, Özgür Umut Akgül, cover the definition and analysis of a context-aware resource market for short term infrastructure sharing with trading and pricing framework. Flexible resource sharing at short time scales in multi-tenant shared radio access networks has proven to be quite a challenge.

In the early stages of the research, the ESR 4 developed a techno-economic model that enables dynamic short-term resource sharing as well as resource pricing, while simultaneously collecting revenue for network expansion. The proposed framework allows operators to meet their individual utility targets while optimizing their expenditures based on their respective budgets. The work generated two conference publications which are included in this deliverable in the Appendix.

The planned activities for the following reporting period include the model refinement by including the end users in the techno-economic framework, and the design of practical network slicing techniques which are driven by the proposed optimization framework.

3 Appendix

This appendix includes the published/submitted research papers which were generated from the research of the ESR 3 and the ESR 4. Namely, the following manuscripts are included:

- **E. Fountoulakis**, N. Pappas, Q. Liao, V. Suryaprakash, D. Yuan, *An Examination of the Benefits of Scalable TTI for Heterogeneous Traffic Management in 5G Networks*, accepted at the Resource Allocation, Cooperation and Competition in Wireless Networks (RAWNET) workshop, 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT), May 2017
- **O. U. Akgul**, I. Malanchini, V. Suryaprakash, A. Capone, *Dynamic Resource Allocation and Pricing for Shared Radio Access Infrastructure*, accepted at the IEEE International Conference Communications, May 2017
- **O. U. Akgul**, I. Malanchini, V. Suryaprakash, A. Capone, *Service-aware Network Slice Trading in a Shared Multi-tenant Infrastructure*, submitted to the IEEE Global Communications Conference (GLOBECOM) 2017

An Examination of the Benefits of Scalable TTI for Heterogeneous Traffic Management in 5G Networks

Emmanouil Fountoulakis[1], Nikolaos Pappas[1], Qi Liao[2], Vinay Suryaprakash[2], Di Yuan[1]
[1] Department of Science and Technology, Linköping University, Sweden
[2] Nokia Bell Labs, Stuttgart, Germany
E-mails: {emmanouil.fountoulakis, nikolaos.pappas, di.yuan}@liu.se
{qi.liao, vinay.suryaprakash}@nokia-bell-labs.com

***Abstract*—The rapid growth in the number and variety of connected devices requires 5G wireless systems to cope with a very heterogeneous traffic mix. As a consequence, the use of a fixed transmission time interval (TTI) during transmission is not necessarily the most efficacious method when heterogeneous traffic types need to be simultaneously serviced. This work analyzes the benefits of scheduling based on exploiting scalable TTI, where the channel assignment and the TTI duration are adapted to the deadlines and requirements of different services. We formulate an optimization problem by taking individual service requirements into consideration. We then prove that the optimization problem is NP-hard and provide a heuristic algorithm, which provides an effective solution to the problem. Numerical results show that our proposed algorithm is capable of finding near-optimal solutions to meet the latency requirements of mission critical communication services, while providing a good throughput performance for mobile broadband services.**

***Index Terms*—5G, scalable TTI, deadline-constrained traffic, low latency, channel allocation, service-centric scheduler**

I. Introduction

The statement, "Future wireless access will extend beyond people, to support connectivity for anything that may benefit from being connected.", by the authors of [1] has far reaching implications. This entails that a variety of new autonomous devices, such as drones, sensors, etc., will communicate using the same network that simultaneously has to serve conventional mobile broadband (MBB) services. Thus, next generation wireless communications systems will be characterized by their service requirement heterogeneity [2]. A characteristic example of services, which have requirements vastly different from MBB services, are those that fall under the category of machine type communications (MTC) [3]. Two subcategories of MTC services are the mission critical communications (MCC) and the massive machine type communications (MMC). MCC services are characterized by small packets and require ultra low latency ($\leq$ 1ms, [1]) and high reliability [4]. On the other hand, MMC envisions tens of billions of connected devices [1]. Therefore, it is not farfetched to assume that the use of a fixed TTI length for catering to such a diverse set of services could be suboptimal. For traffic types in which the ratio between the size of signaling and data is greater than or equal to 1, fixed TTI leads to a significant wastage of resources and – as a result – inefficient communications. The promise of scalable TTI as a potential solution was demonstrated in [5], where the TTI length could be scaled according to the traffic type.

To support a mix of services with heterogeneous requirements, in [3] and [6] the authors propose a flexible frame structure in frequency division duplex (FDD) networks. In these works, the delay constraints are reverse engineered based on the channel state information and the delay budgets. Along similar lines, the authors in [7] apply the variable frame structure in the context of millimeter wave communications. However, these works aim to prioritize active services with strict latency requirements, while sacrificing the throughput of mobile broadband users. In a recent work [5], scalable TTI lengths are introduced in dynamic time division duplex (TDD) mode in order to consider the requirements of each individual service and provide a good trade-off between heterogeneous performance metrics (with respect to their corresponding traffic demands and latency requirements). Moreover, the dynamic TDD scheme offers greater flexibility than the FDD scheme, in terms of adaptability to an asymmetry in uplink (UL) and downlink (DL) traffic. However, none of the works mentioned above jointly considers dynamic TTI length adaptation and channel allocation. In addition to scheduling flexibility in the time domain, jointly considering scalable TTI and channel allocation provides a more flexible frame structure, which is better at exploiting channel diversity and improving spectral efficiency.

In this paper, we aim to develop a scheduling approach that strives to fulfill the (service) deadlines and requirements of different types of services by scaling the length of the TTI to be used. To this end, we formulate an optimization problem whose solution provides the appropriate TTI length and the channel allocation for each service. We then prove that the optimization problem formulated is NP-hard. Therefore, in order to have a scheduler that works in polynomial time, we propose a greedy algorithm that finds an approximate solution to the optimization problem. Numerical results show that the formulated optimization problem tries to cater to all MCC services within their latency requirements, while providing a higher throughput for MBB services in comparison to the other methods commonly considered. They also indicate that the improvement in performance provided by our formulation over the shortest deadline first scheduler (SDFS) increases as the number of active MCC services increases.

II. System Model

We consider a single cell of an FDD network in downlink mode [1]. We also consider services, each with a deadline within which all their requirements must be met. Henceforth, we will use the term services rather than users in recognition of the fact that a user can request more than one service. In this paper, we assume discretized time and '*one time unit*' refers to the minimum amount of time during which a transmission can occur. Let the TTIs be indexed in the time domain by $n \in \mathbb{N}$. The length of each TTI $\Delta(n), \forall n \in \mathbb{N}$ is scalable and can be selected from a finite set $\Delta(n) \in \{1, 2, \ldots, L\}$, where $L \in \mathbb{N}$ is the largest number of time units that can be assigned to a particular TTI. The active set of services at the beginning of the n-th TTI is denoted by $\mathcal{S}_n$ with cardinality $|\mathcal{S}_n|$.

Let $\mathcal{K} \triangleq \{1, 2, \ldots, K\} \subset \mathbb{N}$ be the set of available channels with cardinality $|\mathcal{K}|$, and assume that the same TTI size is retained for all the channels. Each service $s \in \mathcal{S}_n$ can be allocated to a number of channels. We use the vector $\boldsymbol{a}_s(n) \in \{0,1\}^{|\mathcal{K}|}$ to denote the allocation of channels to a service s. The i-th element of $\boldsymbol{a}_s(n)$, $a_{i,s}(n)$, takes the value one if the i-th channel is assigned to the service s during the n-th TTI, and takes the value zero otherwise. Let $\mathcal{NZ}_s(n)$ denote the set of non-zero elements of vector $\boldsymbol{a}_s(n)$. Let the channel allocation for all services be collected in a binary matrix $\boldsymbol{A}(n) \in \{0,1\}^{|\mathcal{K}| \times |\mathcal{S}_n|}$, where the s-th column is $\boldsymbol{a}_s(n)$. Each channel can be assigned up to one service within a TTI and thus, we have the following constraint

$$\sum_{s \in \mathcal{S}_n} a_{i,s}(n) \leq 1, \forall i \in \mathcal{K}, \forall n \in \mathbb{N}. \tag{1}$$

Each channel i has a known channel state information (CSI) for every service s. The CSI in the i-th channel for the s-th service in the n-th TTI is a tuple defined as

$$\mathrm{CSI}_{i,s}(n) = (R_{i,s}(n), T_{i,s}(n)).$$

In this tuple, $R_{i,s}$ denotes the transmission rate of the s-th service over the i-th channel (in bits/one time unit) that can be sustained without errors for $T_{i,s}$ time units, if the i-th channel is assigned to s. Note that the CSI of a channel still changes from one TTI to another.

At the beginning of the n-th TTI, each service s has a known data requirement denoted by $Q_s(n-1)$. Then, we denote $Q_s(n)$ as the amount of data (in bits) that still needs to be served at the end of the n-th TTI. The evolution of the backlog can be described by

$$Q_s(n) \triangleq \left[Q_s(n-1) - (\Delta(n) - \delta) \sum_{i \in \mathcal{K}} a_{i,s}(n) R_{i,s}(n) \right]^+, \tag{2}$$

where $[\cdot]^+ \triangleq \max\{0, \cdot\}$ and δ is the fraction of a time unit required for the transmission of the signaling overhead. We assume that δ is less than or equal to one time unit. Moreover, each service has a specific deadline before which the data has to be delivered. If a service is not completely served before the deadline, the system fails to meet its requirements and the service is dropped. This deadline is denoted by $D_s(n)$, and defined as

$$D_s(n) \triangleq [D_s(n-1) - \Delta(n)]^+. \tag{3}$$

If $Q_s(n) \neq 0$ and $D_s(n) = 0$, the service s is dropped from the system, whereas if $Q_s(n) = 0$ and $D_s(n) \geq 0$, the service s is completely served and exits the system. Additionally, we define the "emptying rate", $E_s(n)$, of a service s at the end of the n-th TTI by

$$E_s(n) \triangleq \frac{Q_s(n-1) - Q_s(n)}{Q_s(n-1)}, \tag{4}$$

where $E_s(n) \in [0, 1]$, represents the ratio between the data served within the n-th TTI and the amount of data remaining at the end of the $(n-1)$-th TTI. This implies: the larger the emptying rate, the faster the data is served with respect to what was remaining at the end of the previous TTI. For example, if service s is completely served at the end of the third TTI, then $Q_s(3) = 0$ and $E_s(3) = 1$; on the other hand, if s is not served at all during the third TTI, then $Q_s(2) = Q_s(3)$ and thus, $E_s(3) = 0$.

III. Problem Formulation

At the n-th TTI, the optimization variables for the TTI length and the channel allocation are $\{\Delta(n), \mathbf{A}(n)\}$, respectively. Our objective is to address the trade-off between the throughput performance and number of dropped services. To this end, we develop a scheduling scheme that will be able to either prioritize services with short deadlines, or(/and) services that can be completely served during the current round of scheduling.

A. Utility function

We define our utility function as

$$U(n) \triangleq \sum_{s \in \mathcal{S}_n} W_s(n) E_s(n), \tag{5}$$

where $E_s(n)$ is the emptying rate, and the weight $W_s \triangleq \frac{1}{D_s(n-1)}$. Note that W_s increases when the $D_s(n-1)$ decreases, i.e., its value increases if the deadline is soon to expire. Since we consider discrete time, the smallest value $D_s(n-1)$ can attain is one time unit. Therefore, the maximum value of W_s is one and as a result, the maximum value of function $U(n)$ is equal to $|\mathcal{S}_n|$. Hence, the function provides a higher reward when the following types of services are served: i) those having urgent deadlines; and, ii) those that can be served with higher emptying rates.

B. Optimization Problem

Although the utility $U(n)$ in (5) is designed to prioritize services with urgent deadlines, $U(n)$ alone cannot guarantee that services, which can be completely served during the current round of scheduling are chosen. Therefore, we formulate

[1] In this work, we assume that the downlink resources are always available since we consider an FDD system. However, the same formulation can also be applied to a TDD system, depending on whether the carriers are configured in uplink or downlink mode during a given time period.

the optimization problem by augmenting the utility function and by introducing additional constraints, as given below.

$$\max_{\Delta(n),\boldsymbol{A}(n)} \quad U(n)+\theta(n) \tag{6a}$$

$$\text{s. t.} \quad \Delta(n) \leq \min_{s\in\mathcal{S}_n} \min_{i\in\mathcal{NZ}_s(n)} T_{i,s}(n), \tag{6b}$$

$$\sum_{s\in\mathcal{S}_n} a_{i,s}(n) \leq 1, \forall i \in \mathcal{K}, \tag{6c}$$

$$\Delta\left(n\right) \in \{1,\ldots,L\}, \tag{6d}$$

$$\boldsymbol{A}(n) \in \{0,1\}^{|\mathcal{K}|\times|\mathcal{S}_n|}, \tag{6e}$$

$$\theta(n) = M \sum_{s\in\mathcal{S}_n} \mathbb{1}_{\{Q_s(n)=0\}}, \tag{6f}$$

where $M = (|\mathcal{S}_n| - 1)$. Moreover, $\mathbb{1}_{\{B\}}$ is the indicator function which takes the value one if the event B occurs, and the value zero otherwise. For the rest of this paper, we refer to the problem above as scalable-TTI enabled channel allocation (STCA). The objective function (6a) is the sum of the utility function (5) and an additional reward $\theta(n)$. The function $\theta(n)$, defined in (6f), is equal to the product of a constant M and the number of completely satisfied services at the end of the current TTI. This, therefore, ensures that the number of completely served services is included in the objective function (6a). Furthermore, $\theta(n)$ also ensures that if at least one service is completely served, the value it takes in the corresponding term of the objective function (6a) is greater than the sum of the other $(|\mathcal{S}_n| - 1)$ terms of the objective function. As a result, we prioritize services that can be completely served after the current scheduling instance.

Additionally, constraint (6b) ensures that the selected TTI size does not violate the minimum TTI size for a given channel and service. Constraint (6c) ensures that a channel can be assigned to up to one service.

IV. Complexity

This section addresses the complexity of the optimization problem. Specifically, we prove that the optimization problem, as defined in Section III, is NP-hard. However, as shown later on in Theorem 2, the problem admits a polynomial-time algorithm guaranteeing optimality, if flat channels are assumed. By flat channels, we mean that for each service, the channel gains are the same for all channels within a given TTI.

Theorem 1. *STCA is NP-hard.*

Proof. We prove that the decision version of the STCA problem is NP-complete by a polynomial-time reduction to and from the Partition Problem (PP) in three steps, [8]. The decision version of the STCA problem can be stated as:

Given a set of services $\mathcal{S}_n$, the backlogs $Q_s(n-1)$, the deadlines $D_s(n-1)$, a set of channels $\mathcal{K}$, and the achievable rates $R_{i,s}(n)$, $\forall i \in \mathcal{K}$ and $\forall s \in \mathcal{S}_n$, is there a solution of the given STCA instance such that the value of the objective function is at least f, where f is a given positive number?

Step 1: We prove that the STCA problem belongs to the NP class of problems, i.e. given an STCA instance, a positive answer and its associated solution, it takes polynomial time to verify whether the answer to the question posed is indeed YES. It is a plain to see that, given a solution, computing $U(n) + \theta(n)$ takes polynomial time. Therefore, STCA is in the NP class of problems.

Step 2: We now show that there is a polynomial-time reduction from the PP to the STCA problem. In the PP, for a set of positive integers $\{p_1,\ldots,p_m\}$, the task is to determine whether or not this set can be partitioned into two subsets of equal sums, i.e. $\sum_{i\in\mathcal{A}'} p_i = \sum_{i\in\mathcal{A}\setminus\mathcal{A}'} p_i$, where $\mathcal{A} = \{1,\ldots,m\}$ and $\mathcal{A}' \subset \mathcal{A}$. Without loss of generality, we can assume that $\sum_{i\in\mathcal{A}} p_i$ is even. Then, given an instance of the PP, we can define an instance of the STCA problem as follows:

- $\mathcal{S}_n = \{1,2\}, \Longrightarrow |\mathcal{S}_n| = 2$. $|\mathcal{K}| = |\mathcal{A}|$.
- $D_s(n-1) = 1$ time unit, $\forall s \in \mathcal{S}_n$.
- $\Delta(n) = 1$ time unit.
- $\delta = 0$. $R_{i,s}(n) = p_i$, $\forall s \in \mathcal{S}_n, \forall i \in \mathcal{A}$.
- $Q_s(n) = \frac{1}{2}\sum_{i\in\mathcal{A}} p_i$, $\forall s \in \mathcal{S}_n$.

Based on the instance defined above, the value of f in the decision version of this STCA instance is set to 4, i.e., $f = 4$. From the assignments above, there is a one-to-one mapping between the elements in the PP and the channels in the STCA problem. In particular, we associate the i-th element in $\mathcal{A}$ with the i-th element in $\mathcal{K}$. Therefore, the above definition clearly represents a polynomial-time reduction.

Step 3: We now prove that the PP instance has the answer YES if and only if the answer to the defined STCA decision instance is YES. If the answer to the PP instance is YES, there are two sets $\mathcal{A}'$ and $\mathcal{A}\setminus\mathcal{A}'$, such that $\sum_{i\in\mathcal{A}'} p_i = \sum_{i\in\mathcal{A}\setminus\mathcal{A}'} p_i = \frac{1}{2}\sum_{i\in\mathcal{A}} p_i$. We assign the channels corresponding to the set $\mathcal{A}'$ to one service, and the channels corresponding to the set $\mathcal{A}\setminus\mathcal{A}'$ to the other. Hence, for the STCA instance, we have $\sum_{i\in\mathcal{A}'} R_{i,1} = \sum_{i\in\mathcal{A}\setminus\mathcal{A}'} R_{i,2} = \frac{1}{2}\sum_{i\in\mathcal{A}} p_i$. Since $Q_s(n) = \frac{1}{2}\sum_{i\in\mathcal{A}} p_i$, $\forall s \in \mathcal{S}_n$, both services are completely served and therefore, $f = 4$. Hence, the instance above is a YES instance of the defined STCA decision problem.

Conversely, if the answer to the defined STCA decision instance is YES, there are two sets $\mathcal{K}'$ and $\mathcal{K}\setminus\mathcal{K}'$ which correspond to the channel assignments for the services one and two, respectively. Since the answer is YES, there is a solution such that the value of the objective function is equal to 4. Note that this value can be reached if and only if both services are completely served. Hence, we have

$$\sum_{i\in\mathcal{K}'} R_{i,1}(n) \geq \frac{1}{2}\sum_{i\in\mathcal{A}} p_i, \tag{7}$$

$$\sum_{i\in\mathcal{K}\setminus\mathcal{K}'} R_{i,2}(n) \geq \frac{1}{2}\sum_{i\in\mathcal{A}} p_i. \tag{8}$$

We also have, by definition, that $\sum_{i\in\mathcal{K}} R_{i,s}(n) = \sum_{i\in\mathcal{A}} p_i$, for $s \in \{1,2\}$, and $R_{i,1}(n) = R_{i,2}(n) = p_i$, $\forall i$. Therefore, the conditions (7) and (8) hold if and only if they are equal. Hence,

$\sum_{i\in\mathcal{K}'} p_i = \sum_{i\in\mathcal{K}\setminus\mathcal{K}'} p_i = \frac{1}{2}\sum_{i\in\mathcal{K}} p_i$, and $\{\mathcal{K}, \mathcal{K}\setminus\mathcal{K}'\}$ is a feasible partition. This establishes the NP-completeness of the decision version of the STCA problem. Therefore, the STCA problem is NP-hard. □

This leads us to the proof that the global optimum of *STCA* can be computed in polynomial time for the special case of flat channels.

Theorem 2. *The global optimum of STCA can be computed in polynomial time for flat channels.*

Proof. If we have K flat channels, then $\mathrm{CSI}_{k,s_i} = \mathrm{CSI}_{l,s_j}$, for all channels k and l, and for all services s_i and s_j. Let g_k^s denote the value of the objective function when k channels are allocated to service s, i.e.

$$g_k^s = \begin{cases} W_s(n) + M, & \text{if } Q_s(n) = 0 \equiv E_s(n) = 1, \\ W_s(n)E_s(n), & \text{otherwise.} \end{cases} \tag{9}$$

Moreover, if there is no channel assigned to the service s, then $g_0^s = 0$. Let $h_s(i)$ denote the objective function value of optimally allocating i channels to services $\{1, \ldots, s\}$. The optimal objective value can be computed by the recursive function

$$h_s(k) = \max_{k=0,1,\ldots,K} \left\{ g_k^s + h_{s-1}(K-k) \right\}. \tag{10}$$

We then construct a $|\mathcal{S}_n| \times K$ matrix whose elements are computed using (10). The (s,k)-th element of the matrix includes the optimal value of the objective function for services $\{1, \ldots, s\}$ using k channels. Hence, the $(|\mathcal{S}_n|, K)$-th element gives the value of the optimum solution of the entire optimization problem.

For the first row of the matrix, computing the entries $h_1(1), \ldots, h_1(k)$ in the given order are straightforward, and each entry requires a computational complexity of $\mathcal{O}(1)$. Each element of the s-th row requires $\sum_{i=1}^{K} i = K(K+1)/2$ computations. Hence, the computational complexity that is required for each row is $\mathcal{O}\left(K^2\right)$ and thus, the total computational complexity is $\mathcal{O}\left(|\mathcal{S}_n|K^2\right)$. Therefore, the optimum solution of the STCA problem, in the case of flat channels, can be computed using dynamic programming in polynomial time. □

V. Integer Linear Programing Formulation

In this section, we develop an Integer Linear Program (ILP) in order to compute the optimal solution of the STCA problem, which enables a more detailed study of the performance of scalable TTI. First, we solve the problem in (6a) with a fixed TTI length as an input. Note that the problem is solved for each viable TTI length separately. Then, we compare the value of the objective function for all the TTI lengths considered, and subsequently select the TTI length and the channel assignment for which the objective function is maximized. The pair $\{\Delta(n), \boldsymbol{A}(n)\}$ for which the objective function in (6a) is maximized is the optimal solution. It should be noted that, for each possible TTI length, if the TTI length is greater than a given service's deadline, we remove the corresponding service from the optimization problem; thereby, considering the service dropped. In other words, the services whose deadlines will expire despite choosing the optimal Δ (denoted by Δ') have a utility equal to zero. Thus, for each fixed Δ', we consider the set of services $\{s \in \mathcal{S}_n : D_s(n-1) \geq \Delta'\}$.

In this section, we omit the index n for notational brevity and redefine some of the parameters as follows:

- Q'_s – the data backlog of s during the current TTI.
- $W'_s = \frac{W_s}{Q'_s}$.
- β_s – amount of data served to the service s at the end of the current TTI.
- $R'_{i,s} = (\Delta - \delta)R_{i,s}$ is the amount of data that could be transmitted to service s, if the channel i is assigned to it.
- $Y_s = \begin{cases} 1, \text{ if the service } s \text{ is completely served,} \\ 0, \text{ otherwise.} \end{cases}$
- D'_s– the deadline of service s after the $(n-1)$-th TTI.
- $\mathcal{S}_{\Delta'} = \{s \in \mathcal{S}_n : D'_s \geq \Delta'\}$.

The rest of the notations remain unchanged. The optimization problem can then be formulated as the following ILP for a given $W_s \in \mathbb{R}^+$ and Δ'.

$$\max_{\mathbf{A}} \sum_{s\in\mathcal{S}_{\Delta'}} W'_s\beta_s + M \sum_{s\in\mathcal{S}_{\Delta'}} Y_s \tag{11a}$$

$$\text{s. t.} \quad \Delta' - T_{i,s} \leq J_1(1 - a_{i,s}), \forall i \in \mathcal{K}, \forall s \in \mathcal{S}_{\Delta'}, \tag{11b}$$

$$\sum_{s\in\mathcal{S}_{\Delta'}} a_{i,s} \leq 1, \forall i \in \mathcal{K}, \tag{11c}$$

$$\beta_s \leq \sum_{i\in\mathcal{K}} R'_{i,s}a_{i,s}, \forall s \in \mathcal{S}_{\Delta'}, \tag{11d}$$

$$Y_s \leq \frac{\beta_s}{Q'_s} \leq 1, \forall s \in \mathcal{S}_{\Delta'}, \tag{11e}$$

where the constant $J_1 \gg L$ in (11b) guarantees that $a_{i,s} = 0$ if $T_{i,s} < \Delta'$. The constraint (11c) ensures that each channel is assigned up to one service and (11d) makes sure that the maximum value β_s can attain is the amount of data remaining for service s. Therefore, if the service s is completely served, the corresponding term in (11a) takes the maximum value, which is equal to W_s. Note that the ratio $\frac{\beta_s}{Q'_s}$ in (11e) represents the emptying rate in (4). Additionally, if s is completely served, constraint (11e) ensures that Y_s is assigned a value equal to one.

VI. Algorithm

In order to have a polynomial time scheduling algorithm, we propose a heuristic called channel allocation with scalable TTI (CAST) algorithm. For each channel $i \in \mathcal{K}$, the CAST algorithm finds the service $s \in \mathcal{S}_n$, which has the maximum corresponding value of the objective function (6a) – should the channel i be assigned to service s. The algorithm calculates the objective function for each possible TTI length, and selects the channel assignment and the TTI length for which the objective function is maximized.

The CAST algorithm decides the channel assignment for each TTI length in two steps. During the first step, the

Algorithm 1: CAST algorithm

1 $G_{\max} \leftarrow -\infty$, $W_s = \frac{1}{D_s(n-1)}, \forall s \in \mathcal{S}$
2 **for** $\Delta' = 1 : L$ **do**
3 $\quad \mathbf{A}' \leftarrow \mathbf{0}_{K \times |\mathcal{S}|}$, $\mathcal{S}' \leftarrow \mathcal{S}$, $Q'_s \leftarrow Q_s$
4 $\quad$ **if** $D_s(n-1) - \Delta' < 0$ **then**
5 $\quad\quad \mathcal{S}' \leftarrow \mathcal{S} \setminus \{s\}$
6 $\quad$ **for** $i \in \mathcal{K}$ **do**
7 $\quad\quad g_{\max} \leftarrow -\infty$
8 $\quad\quad$ **for** $s \in \mathcal{S}'$ **do**
9 $\quad\quad\quad$ **if** $\Delta' \leq T_{i,s}$ **then**
10 $\quad\quad\quad\quad Q_{\text{temp}} \leftarrow \left[Q'_s - (\Delta' - \delta) R_{i,s}\right]^+$
11 $\quad\quad\quad\quad E'_s \leftarrow \frac{Q_s(n-1) - Q_{\text{temp}}}{Q_s(n-1)}$
12 $\quad\quad\quad\quad g \leftarrow W_s E'_s + M \mathbb{1}_{\{Q_{\text{temp}}=0\}}$
13 $\quad\quad\quad\quad$ **if** $g > g_{max}$ **then**
14 $\quad\quad\quad\quad\quad s_{\max} \leftarrow s$, $g_{\max} \leftarrow g$
15 $\quad\quad\quad\quad\quad Q_{s_{\max}} \leftarrow Q_{\text{temp}}$
16 $\quad\quad\quad\quad$ **if** $Q_{s_{max}} = 0$ **then**
17 $\quad\quad\quad\quad\quad \mathcal{S}' \leftarrow \mathcal{S} \setminus \{s_{\max}\}$
18 $\quad\quad\quad$ **else**
19 $\quad\quad\quad\quad A'_{i,s} \leftarrow 0$
20 $\quad\quad G \leftarrow G + g_{\max}$, $A'_{i,s_{\max}} \leftarrow 1$
21 $\quad$ **if** $G > G_{max}$ **then**
22 $\quad\quad \mathbf{A}_{\max} \leftarrow \mathbf{A}'$
23 $\quad\quad \Delta_{\max} \leftarrow \Delta'$
24 $\mathbf{A}(n) \leftarrow \mathbf{A}_{\max}, \Delta(n) \leftarrow \Delta_{\max}$

algorithm excludes the services whose deadlines cannot be met (lines 4 – 5). The variable g, whose value is calculated in lines 9 – 12, is the objective function value, if the channel i is assigned to the service s. Note that a channel i can be assigned to service s only if the TTI length Δ' is less than the duration $T_{i,s}$ within which an error-free computation of the rate is possible (cf. line 9). During the second step, the algorithm allocates each channel to a corresponding service with the maximum value of the objective function (cf. lines 14 – 15) and removes the service if it is completely served (lines 16 – 17). The algorithm then compares the value of the objective function for each possible TTI length and selects the channel assignment as well as the TTI length maximizing the value of the objective function (lines 21 – 24). Based on the description of ILP above, the complexity of the CAST algorithm is found to be $\mathcal{O}\left(|\mathcal{K}||\mathcal{S}_n|L\right)$.

VII. Numerical Results

In this section, we compare the performance of the CAST algorithm with the optimal solution (OS) for the STCA problem. Additionally, we also compare our approach with a simpler version of the shortest deadline first scheduler (SDFS) proposed by the authors in [6]. The above mentioned comparisons are undertaken using the simulations based on the parameters that follow.

We consider one time unit to be equal to 0.1ms, and the TTI length can be selected from a finite set $\Delta(n) \in \{0.2\text{ms}, 0.3\text{ms}, \ldots, 1\text{ms}\}$ in a single cell scenario where the FDD is in downlink mode [2]. We also assume that the transmission of control signaling requires $\delta = 0.05$ms per TTI (regardless of the length of the TTI chosen). We consider a system with an 8 MHz bandwidth that works on a frequency selective channel with a coherence bandwidth of 0.5 MHz. The achievable rate for a service s in the i-th channel during the n-th TTI is computed using the Shannon formula and is given by $R_{i,s}(n) = B \log_2\left(1 + |h_{i,s}(n)|^2 \frac{S}{N}\right)$, where the channel gains $h_{i,s}(n)$ are distributed as a zero-mean complex Gaussian with variance σ^2, i.e., $h_{i,s}(n) \sim \mathcal{CN}\left(0, \sigma^2\right)$, S is the transmit power, N is the noise power, and B is the bandwidth of each channel, i.e., $B = 0.5$ MHz. The average value of the signal-to-noise ratio (SNR) is equal to 5 dB. Moreover, we consider that the base station caters to services generated by three MCC sources and one MBB source. Each source generates services per time unit (0.1ms) according to a Bernoulli distribution with probability r_{MCC} and r_{MBB} for MCC sources and the MBB source, respectively. Lastly, each MCC service has a demand of 125 bytes and deadline of 1ms, and each MBB service has a demand of 1125 bytes and a deadline of 10ms. In the following paragraphs, we study the behavior of the algorithms proposed for various values of r_{MCC}, while the probability of MBB service arrivals is constant and equal to 0.2, i.e., $r_{\text{MBB}} = 0.2$.

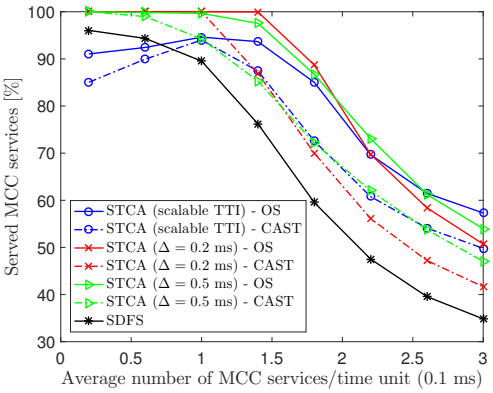


Fig. 1: Variations in MCC services.

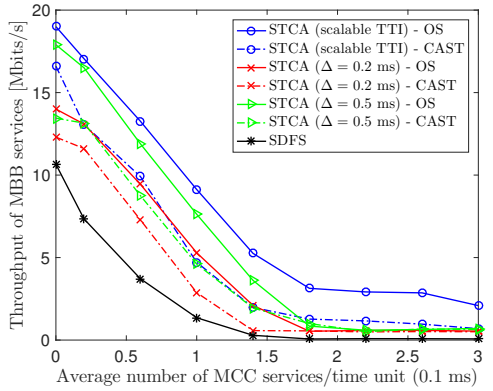


Fig. 2: Variations in the throughputs of MBB services.

Fig. 1 depicts the variations in the percentage of MCC services dealt with as the average number of MCC service requests per time unit (0.1ms) increases. It documents the aforementioned variations for both the optimal solution and the heuristic of the STCA in scenarios where the TTI lengths are scalable and fixed, as well as the variations seen in the

[2]Note that $\Delta(n)$ here is presented with the units 'milliseconds' for improved readability. The value of $\Delta(n)$ in milliseconds is obtained by multiplying the original $\Delta(n)$ with the duration of one time unit (0.1ms).

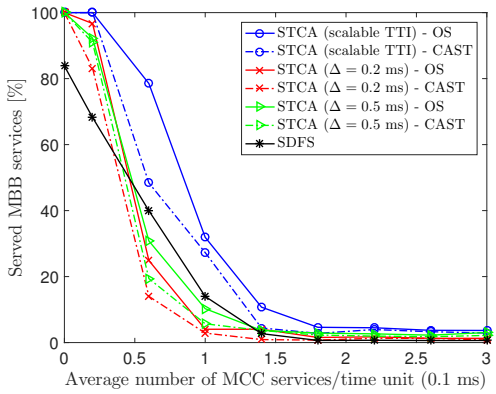


Fig. 3: Variations in MBB services.

behavior of the SDFS. This figure indicates that a scheduler using the STCA with short but fixed TTI lengths outperforms the one using the STCA with scalable TTI as well as the SDFS. The reason why the STCA with short, fixed TTI outperforms the STCA with scalable TTI is because the latter tends to select longer TTI lengths in order to be able to *completely* serve as many services as possible during each scheduling period. This sort of selection implies that a greater portion of the MCC services end up being dropped. However, as the arrival rate of MCC services continues to increase, the STCA with scalable TTI starts to select shorter TTI lengths; thereby, resulting in the increase in the percentage of MCC services catered to between 0.2 and 1 MCC arrivals$/0.1$ms before eventually decreasing beyond 1.5 MCC services$/0.1$ms. It is noteworthy that the STCA with scalable TTI eventually outperforms the STCA with fixed TTI, i.e., beyond 2.5 MCC services$/0.1$ms.

As commonly known, the amount of signaling overhead increases quite substantially when shorter TTI lengths are selected. The cost of an increase in the signaling overhead is born a decrease in the throughput delivered to the MBB services. Fig. 2 demonstrates the variations in the throughput of the MBB services as the average number of MCC service requests$/0.1$ms increases. Clearly, of the methods considered, the SDFS is the one that is most significantly affected. This figure also indicates that, though the MBB services see an inevitable drop in their throughput, the STCA with scalable TTI is able to cope much better than the STCA with short, fixed TTI – especially when the average number of MCC service requests$/0.1$ms is greater than 1.5. A reason why the STCA with scalable TTI outperforms the STCA with short, fixed TTI is because of its ability to contain (and regulate) the amount time spent in transmitting the control signaling more effectively.

Lastly, Fig. 3 – as in Fig. 2 – depicts the unavoidable decrease in the percentage of MBB services satisfied when the average number of MCC service requests$/0.1$ms increases. It does, however, highlight the fact that the STCA with scalable TTI is able to serve a far greater percentage of MBB services when compared to the others in the face of increasing MCC service requests$/0.1$ms. This behavior can, once again, be attributed to the fact the STCA with scalable TTI can control the fraction of time spent transmitting the control signaling by periodically choosing larger TTI lengths and thereby, ensuring that MBB services are also furnished with the resources they need. Also, the results illustrate that there is a visible gap between the performance of the CAST algorithm and the OS, though the CAST algorithm significantly outperforms the SDFS. This gap is expected because of the low complexity of the CAST algorithm.

Overall, when one considers all the results collectively, it can be said that a scheduler which jointly considers scalable TTI and channel allocation into account is better at being able to handle traffic heterogeneity and has the ability to improve the spectral efficiency of individual service types.

VIII. Conclusions

In this paper, at each scheduling time, we propose a joint optimization of the TTI lengths and the channel allocation depending on the traffic type. The joint optimization problem formulated is then proven to be NP-hard due to which we provide a heuristic akin to a greedy algorithm. However, for flat channels, we also demonstrate that the problem admits a polynomial-time solution that guarantees optimality. The optimization problem and its heuristic are then compared not only with one another for the cases of fixed and scalable TTI lengths, but also with the shortest deadline first scheduler. These evaluations illustrate that our proposal of a joint optimization of TTI lengths and channel allocation is better equipped to handle traffic heterogeneity and provide improved spectral efficiency, due to its ability to regulate the amount of time spent on control signal transmissions and maximize the number of services satisfied.

IX. Acknowledgment

The authors would like to thank Dr. Ilaria Malanchini for numerous fruitful discussions and her valuable suggestions. This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 643002.

References

[1] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, and Y. Selén, "5G radio access," *Ericsson review*, vol. 6, pp. 2–7, 2014.

[2] N. Alliance, "NGMN 5G white paper," *Next generation mobile Networks, white paper*, 2015.

[3] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, March 2016.

[4] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept 2016.

[5] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," in *2016 IEEE Globecom Workshops*, Dec 2016, pp. 1–7.

[6] K. Pedersen, F. Frederiksen, G. Berardinelli, and P. Mogensen, "A flexible frame structure for 5G wide area," in *2015 IEEE 82nd Vehicular Technology Conference*, Sept 2015, pp. 1–5.

[7] T. Levanen, J. Pirskanen, and M. Valkama, "Radio interface design for ultra-low latency millimeter-wave communications in 5G era," in *2014 IEEE Globecom Workshops*, Dec 2014, pp. 1420–1426.

[8] M. R. Garey and D. S. Johnson, *A guide to the theory of NP-Completeness*. John Wiley & Sons, 1979, vol. 70.

Dynamic Resource Allocation and Pricing for Shared Radio Access 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*—Flexible resource sharing at short time scales in multi-tenant shared radio access networks has proven to be quite a challenge. In this study, we develop a techno-economic model that enables dynamic short-term resource sharing as well as resource pricing, while simultaneously collecting revenue for network expansion. In order to regulate the resource costs and to prevent monopolization of resources, we define a unit cost of resources which can be scaled dynamically. The proposed framework allows operators to meet their individual utility targets while optimizing their expenditures based on their respective budgets. This work demonstrates that dynamic short timescale resource sharing can help network operators achieve their utility targets while minimizing their total expenditure.**

I. Introduction

A multitude of applications, driven in part by the Industry 4.0 initiative, are envisioned for future networks (5G and beyond), [1]. Most of these applications require not only high data rates, but also low latencies. One of the potential solutions to this problem is considered to be denser and more heterogeneous network deployments, [2]. This, however, places an enormous strain on the already decreasing profitability of mobile operators, [3] and thereby, necessitates a change in their current business modus operandi. One of the solutions proposed to cope with increasing operational costs and decreasing profitability is *Infrastructure Sharing*, [4]. As the name suggests, this idea proposes that mobile network operators (MNOs) share a common infrastructure in order to reduce their capital and operational expenditure as well as to offer their customers better prices, a larger number of services, and a better quality of service.

As detailed in the Organization for Economic Co-operation and Development (OECD) report, [4], infrastructure sharing can be undertaken at various levels. One of the most comprehensive methods of sharing is where there are multiple mobile virtual network operators (MVNOs) who lease or rent the infrastructure from an infrastructure provider and such a type of sharing is the focus of this paper. In general, sharing takes place based on service level agreements (SLAs) between the parties who intend to share the infrastructure and it usually takes the form of contracts which are enforced over a long period of time. However, the OECD report, [4, Pg. 65], also states: "savings from active sharing are greater than for passive sharing as a higher proportion of costs are shared". Active sharing implies sharing radio access network resources including the spectrum. This type of sharing, however, quickly becomes infeasible if today's (i.e. long term) SLAs are used. This is because the MVNOs will not have the ability to accommodate fluctuations in their traffic and might often find themselves in scenarios where they risk being unable to cater to their customers. Active sharing, therefore, requires a more dynamic sharing methodology which allows MVNOs to share and trade resources at much shorter timescales, i.e., in a few seconds or minutes. In order for such a system to work, viz. for it to be profitable for all the parties involved, each of them should have a good understanding of their own budgets, the implications of short-term fluctuations on them, and an accurate estimate of their traffic load. It, therefore, becomes imperative that each of the parties involved, be it MVNOs or infrastructure providers, have sound techno-economic models that are able to estimate aspects like resource allocation, the required network expansion, and their implications on resource pricing. As detailed in Subsection I-B, this is precisely where today's models fall short and this is an aspect this paper tries to address.

A. Contributions

In the interest of facilitating the active sharing promoted by the OECD, we propose a techno-economic model that allows dynamic short term resource sharing as well as short term price negotiations between the MVNOs and the infrastructure provider. The contributions of this paper are as follows:

- It provides *a first step* towards a more comprehensive *techno-economic market model* for radio access networks.
- It proposes *a short time scale* dynamic trading model wherein: i) the cost of resources is market driven, and ii) the MVNOs trade resources based on their ability to satisfy customer demands as well as meet their respective budget constraints.

B. Related Work

Relatively speaking, technological models have garnered more attention only in the recent past. Works such as [5]–[9] estimate the performance and provide a comparison of networks wherein both physical and virtual sharing of capacity or spectrum occurs. These works, however, tend to be system or technology dependent (e.g., focusing solely on LTE, etc.). Lately, there have been attempts in papers such as [10] or

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

[11] to provide more generic resource sharing algorithms. The economic aspects salient to MVNOs are, however, not considered in the aforementioned works.

Economic models, on the other hand, have been used quite extensively to motivate the need for network sharing. There are numerous works such as [4] and the references therein, which focus on various aspects related to the costs of sharing specific network components. More specifically, papers such as [12] and [13] argue in favor of site sharing, radio access network sharing, and core network sharing as ways towards a sustainable business platform for the future. Other works such as [14] and [15] also explore the relationship between network costs and the extent to which networks are shared. Admittedly, there is an implicit link between the technological and economic aspects when varying degrees of network sharing are explored. These works, however, do not shed sufficient light on the technological implications (i.e., on the ability to satisfy customer demands) of economic decisions made.

Another aspect - overlooked in most works - is the fact that the models proposed still focus on long term SLAs, which do not provide the flexibility required to enable active network sharing. An added degree of flexibility in the SLAs is provided in [11], where a sharing model, which allows the parties to deviate from the constraints agreed upon in the SLA to a certain extent while abiding by the SLA's constraints on average. This idea forms the basis of our work in this paper.

II. System Model

In our model, we considered two different types of stakeholders, i.e. a single infrastructure provider and multiple MVNOs. Let M represent the set of MVNOs and let $|M|$ be its cardinality. Then, let K represent the set of active users that are distributed between the MVNOs and let the set of active users of MVNO m be represented by K_m. We assume that the decisions taken by a base station's scheduler are not directly effected by schedulers in the neighbouring base stations and thereby, focus on the downlink of a single base station. Like [11], we assume that there exists an initial agreement between the infrastructure provider and the MVNOs, which sets the initial values of the network resources to be shared. However, unlike [11], we do not consider statically shared network resources; instead, MVNOs update their share of the network resources based on their respective traffic and utility targets.

A. Notations and the Model

To ensure consistency and clarity, this work uses the same notations as those used in [11]. In this framework, $S_m \in [0, 1)$ represents the sharing ratio, i.e., the percentage of resources, for operator m based on predefined SLAs and $\Delta_m \geq 0$ denotes the maximum deviation from S_m (when averaged over a certain time window). The average resources that a particular MVNO gets cannot deviate from S_m by more than Δ_m within W. Now, recall that our goal is to further a more dynamic resource trading environment in which the MVNOs are free to pursue their individual interests. With this objective in mind, in the proposed framework, S_m and Δ_m are MVNO specific variables that can be re-negotiated periodically, where the period specified by a time window W is determined by the infrastructure provider. Note that, in this work, we consider the existence of just a single infrastructure provider, who is not subject to conventional market pressures.

Similar to most works having to do with schedulers, time is discretized and partitioned into time slots. As in [11], $x_k[n]$ denotes the fraction of resources assigned to the user k at time slot n. Depending on the resources assigned, the deviation of operator m from S_m at time slot n is denoted by $\epsilon_m[n]$. Additionally, $r_k[n]$ represents the achievable rate of user k during the time slot n. Apart from the notation used in [11], we also define new parameters relevant to a techno-economic model. In this model, each operator has a budget, B_m, that can be spent at any time instance n. We define three types of cost, namely: capital expenditure (CapEx), operational expenditure (OpEx), and pressure cost denoted by C_{ca}, C_{op}, and C^m_{pre} (for MVNO m), respectively. The MVNO has a CapEx proportional to its S_m, whereas the OpEx is based on the actual resources obtained. This definition incentivizes the MVNOs to utilize the added flexibility to deviate from the original resource sharing limits agreed upon by coupling each MVNO's expenses with their needs and budget constraints.

The pressure cost, C^m_{pre}, ensures that the costs of network resources scale according to their demand and, from an infrastructure provider's point of view, provides a steady revenue stream for expenses like (longterm) capacity expansion. In the market model considered, when the number of available resources decreases, the cost of purchasing a given unit of resource increases. In order to create this inversely proportional dependence between the scarcity of resources and their cost, the term $\xi_m[n]$ – reflecting the difference between an operator's utility target ($U_{\text{th},m}$) and the actual utility they obtained – is used. Throughout the paper, we refer to resource scarcity as the case where $\xi_m > 0$ and $\sum_{k\in K} x_k = 1$ and to resource surplus as the case where $\xi_m = 0,\ \forall m \in M$ and $\sum_{k\in K} x_k[i] < 1$. The product of $\xi_m[n]$ and C^m_{pre}, therefore, provides the surcharge for the resources requested at time slot n. Since the pressure cost is proportional to a given operator's utility target, a long term aggregate of this cost results in the amount (proportional to the sum of the utility requirements of all MVNOs) which should be invested towards network or capacity expansion.

B. Assumptions

The salient assumptions are as follows:

1) *Operator's gap, $\xi_m[n]$,* gives complete information about *the additional resources* required to satisfy an MVNO's target.
2) All the traffic is *elastic*, i.e., the traffic is not sensitive to delays.

III. Formulation and Analysis of the Model

A. Problem Formulation

Based on the notation defined in Section II, the generic optimization problem solved at the base station's scheduler

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

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

$$\xi_m[n] \geq \max(0, U_{\text{th},m} - \frac{1}{(a+1)|K_m|} \sum_{i=n-a}^{n} \sum_{k \in K_m} U_k(x_k[i],\ r_k[i])), \quad \forall m \in M,\ a \equiv (n-1 \mod W), \tag{1c}$$

$$\epsilon_m[n] = \left(\frac{1}{(a+1)} \sum_{i=n-a}^{n} \sum_{k \in K_m} x_k[i] \right) - S_m, \quad \forall m \in M, \tag{1d}$$

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

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

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

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

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

to dynamically optimize the resource allocation and pricing is given by Equations (1a)-(1i). Ideally, the scheduler has the achievable rates of users for each MVNO and schedules using these rates. However, since the algorithm is proposed for a real time scheduling problem, the optimizer does not know the rates in the future. Therefore, the selection of efficient S_m and Δ_m is not trivial as the MVNOs have to predict their future needs extremely accurately. In order to solve this challenge, the optimization problem (1a)-(1i) is split into a two stage optimization problem denoted by P1 and P2. The individual objective functions used and their respective constraints are explained in Subsection III-B.

Our optimization problem considers a continuous objective function (1a), which depends on two parts. The first part minimizes the total gap between the individual MVNO's desired utility and their actual utility values at time slot n. In the second part, in order to guarantee a fair distribution of the CapEx among MVNOs, we minimize $S_{\max}$, which denotes the maximum between the SLA based resources available to an MVNO (S_m) and the remaining resources ($1 - S_m$) as defined in constraint (1b). The minimum of the right-hand side (RHS) of (1b) can be achieved if S_m is selected equal to $1 - S_m$. Based on this logic, the optimizer assigns $S_m = \frac{1}{|M|}$, $\forall m \in M$, if the budgets of all MVNOs permit it. Therefore, this constraint, i.e. (1b), ensures that fairness is achieved in terms of the initial sharing of resources. The MVNO can obtain additional resources (if available) by selecting a higher Δ_m value. The gap of operator m, $\xi_m[n]$, is constrained by (1c). The first term within the maximization function in (1c) prevents the gap from being lower than zero, and reflects the fact that the network (provided by the infrastructure provider) is able to handle traffic effectively enough that no expansion is necessary even in the long run. The second term in the maximization function, on the other hand, computes the difference between the desired utility – denoted by $U_{\text{th},m}$ – of a given MVNO m and the utility function $U_k\left(x_k[n], r_k[n]\right)$ measured at time n, where $x_k[n]$ and $r_k[n]$ are the percentage of resources and the rate assigned to a user k during a particular slot n, respectively. Constraint (1d) sets the value of $\epsilon_m[n]$, which is the instantaneous deviation from the agreed sharing ratio. The first term on the RHS of the equation is the average resources that MVNO m obtained from the beginning of the current time window to the current time slot n, whereas the second term is the SLA based sharing ratio. When $\xi_m[n]$ and $\epsilon_m[n]$ are calculated, each W is considered to be independent of the other.

Constraint (1e), in which $\epsilon_m[n]$ is computed using (1d), limits the maximum deviation between the agreed sharing ratio and the obtained resources from exceeding Δ_m. Constraint (1f) is the budget constraint, which ensures that the overall expenditure in a time window cannot exceed an MVNO's budget for that time window. The operator pays both CapEx and OpEx for the fixed resource shares agreed upon in the SLA, S_m, which is accounted for by the first term in the summation on the left-hand side (LHS) of the inequality. However, by choosing a higher deviation Δ_m, the operator has the ability to increase or decrease their expenditure in relation to the costs computed using the SLA. This aspect is taken into account by incorporating $\epsilon_m[n]$ in the second term on the LHS of the inequality (1f). If the MVNO receives fewer resources than S_m, as can be observed from (1d), the second term of (1f) becomes negative and decreases the total cost. The third term on the LHS of (1f) is the pressure cost that is designed in order to regulate the demands of individual MVNOs and to introduce a notion of 'supply and demand' economics to these short term resource acquisitions. From an infrastructure provider's perspective, it also acts as a means to collect the necessary revenue for network expansion. Since

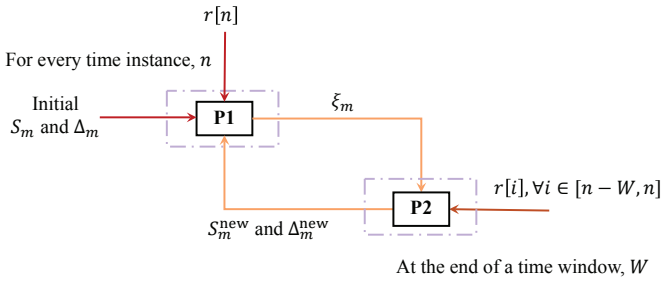


Fig. 1. Two stage optimization solved at the scheduler.

the MVNOs are not obligated to spend their entire budget during a given time slot, the unspent revenue from previous time slot can be used during the successive time slot. This effect is represented by the scaling factor $(1+a)$ on the RHS of (1f), where $a \equiv (n-1) \bmod W$. Then, constraint (1g) introduces the necessary coupling between Δ_m and S_m in order to prevent the MVNOs from trading resources that they do not have. Constraint (1h) ensures that the total number of resources consumed is always less than or equal to the total number of resources available and the network resources allocated to any user k cannot be lower than zero. Finally, constraint (1i) ensures non-negativity of the resources initially agreed upon in the SLA and it also prevents the sum of S_m over all the operators from being greater than one.

Note that the problem is presented in a non-linear form to improve readability. It can, nevertheless, be lineralized with standard techniques.

B. Applied Algorithm

Owing to the many fluctuations encountered during wireless communications, the optimization problem to determine the most cost effective resource allocation for a given MVNO is solved in two steps denoted by P1 and P2 (detailed below) and as illustrated in Fig. 1. During the first step indicated by P1, the optimizer accepts S_m and Δ_m as input parameters and finds the optimum resource allocations that minimizes the total gap between each MVNO's target utility and the utility they achieved. During each time slot within a given time window W, the optimizer runs this resource allocation optimization (i.e., P1) using the respective rate estimates of the active users. At the end of W, the optimizer switches to the second step, i.e. P2.

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

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

During P2, the optimizer determines the optimal resource allocation for the previous time window (i.e., the window that just ended) using the knowledge of all the rates actually achieved. Then, based on these 'optimum' resource allocations, the optimizer determines the optimal S_m and Δ_m, and updates their values for the upcoming time window. The update process is performed according to

$$S_m^{\text{new}} = (1-\alpha_m) S_m^{\text{old}} + \alpha_m S_m^{\text{opt}}, \tag{2}$$

$$\Delta_m^{\text{new}} = (1-\alpha_m) \Delta_m^{\text{old}} + \alpha_m \Delta_m^{\text{opt}}, \tag{3}$$

where α_m is the feature scaling coefficient and S_m^{opt}, Δ_m^{opt} are the optimum S_m, Δ_m values for the previous time window.

Both for P1 and P2, the optimizer uses the same objective function. However, since the variables of the two problems are different, the constraints that are applicable for each of the problems will also be different.

C. Effects of Feature Scaling

The input parameters for P1 during the upcoming time window are selected based on their initial values and the optimum values during the previous time window. However, the determination of the scaling coefficient is a challenging task as large values of α_m effectively leads to a memoryless network resource optimization and very small values result in a static network resource optimization. A comparison between different scaling coefficients is presented in Fig. 2. As presented in (4), the relative distance to the optimal (RDO) gives an understanding of how close the selected parameters are to their optimum values, ξ_m^{opt}. Note that due to (1c), the gap $\xi_m[n]$ cannot be negative for any time slot and $\xi_m[n] \geq \xi_m^{\text{opt}}[n]$, $\forall n \in N$. For the special case of $\xi_m[n] = \xi_m^{\text{opt}}[n] = 0$, the RDO is assumed to be 0; therefore, RDO $\in [0,\ 1]$.

$$\text{RDO} = \frac{1}{|M|} \sum_{m \in M} \frac{\sum\limits_{i=1}^{N} \xi_m[i] - \xi_m^{\text{opt}}[i]}{\sum\limits_{i=1}^{N} \xi_m[i]}. \tag{4}$$

The dynamic scaling coefficient is presented in (5), where $\xi_m^{\text{opt}}[n]$ is the optimum gap calculated by the optimizer during P2. Since it is determined by the actual rates achieved, it gives an idea of the minimum achievable gap if the scheduler has a-priori knowledge of the rates.

$$\alpha_m = \frac{\left| \sum\limits_{i=n-a}^{n} \xi_m[i] - \sum\limits_{i=n-a}^{n} \xi_m^{\text{opt}}[i] \right|}{\sum\limits_{i=n-a}^{n} \xi_m[i] + \sum\limits_{i=n-a}^{n} \xi_m^{\text{opt}}[i]}, a \equiv (n-1 \bmod W). \tag{5}$$

The scaling coefficient (α_m) provides information about the difference between the observed gap and the minimum gap achievable per time window. Therefore, α_m measures how close the real-time scheduler performs to the optimum. Since the network parameters are selected according to a given MVNO's targets, each of them has a different α_m parameter that reflects the optimality of their decision.

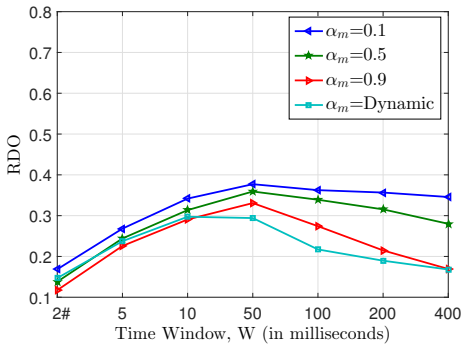


Fig. 2. Comparison of different scaling coefficients.

TABLE I
THE APPLIED PARAMETERS AND THEIR VALUES.

Parameter	Definition	Value
C_{ca}	CapEx Cost	35.97
C_{op}	OpEx Cost	25.69
U_{th}^{m}	Rate target of operator m	2 bps/Hz
$\lvert K_m \rvert$	Cardinality of the set of active users	1
W	Time window	100 ms
$\lvert M \rvert$	Number of MVNOs	3
N	Duration of simulation	5000 ms
B_1	Budget of MVNO 1	88.45
B_2	Budget of MVNO 2	100
B_3	Budget of MVNO 3	56.17
$C_{\text{pre}}^{\text{unit}}$	Unit pressure cost per resource	35.97

Fig. 2 presents the variation in the achievable gap for time windows of different lengths for various scaling coefficients. Since we would like to evaluate the merits of dynamic short time-scale resource sharing, we focus on time windows of length $50 - 200$ ms. For this range, dynamic scaling is better than static scaling because it is better at coping with changes in window size. Therefore, the dynamic scaling coefficient is utilized in the simulations.

D. Effects of Pressure Cost

As previously mentioned, the motivation behind the introduction of the pressure cost is twofold. First, it helps regulate the price of resources (in scenarios of both resource surplus as well as resource scarcity), while ensuring that the price a given MVNO pays is proportional to their respective budget. Scaled pressure costs ensure that MVNOs will have the same chance to obtain resources and will be charged in proportion to their budgets. More specifically, the operators will not face scenarios where neither the purchase of resources nor the pressure costs are unaffordable. Therefore, the pressure cost is defined as

$$C_{\text{pre}}^{m} = \frac{B_m}{\sum\limits_{m \in M} B_m} \times C_{\text{pre}}^{\text{unit}}, \tag{6}$$

where $C_{\text{pre}}^{\text{unit}}$ is the unit pressure cost of a given resource.

Second, in cases where the gap between the desired utility of MVNOs and their achieved utility is non-zero, since the pressure cost is proportional to the difference between the actual and desired utility values for each MVNO, the infrastructure provider – by means of the aggregated pressure costs collected – has the added advantage of knowing exactly how much has to be invested in capacity expansion.

IV. Simulation Results

In this section, the simulation results of the two-step optimization problem are detailed.

A. Parameters and the Scenarios Studied

In order to analyze the applicability of our mathematical model, we considered the downlink of a base station that is shared by three MVNOs, i.e. $|M| = 3$. All the users are uniformly distributed throughout the coverage area of the base station and, at each time slot, only one user from each operator becomes active. The simulation is run on a standard commercially available laptop for $N = 5000$ time slots, where each time slot is assumed to be 1 ms long, and the total run time of the algorithm (including both P1 and P2) is 0.998 sec. All the costs as well as the budgets are normalized to take values between 0 and 100 $\left(C_{\text{ca}},\ C_{\text{op}},\ C_{\text{pre}}^{m},\ B_m \in [0,\ 100],\ \forall m \in M\right)$ such that they can be considered as a generic value which can be spent during each time slot n. It is important to mention that the values of the budgets and costs are purely illustrative, whose purpose is to help understand the characteristic behavior of the model. Since the actual values that MVNOs use will merely be affine functions of the values used here, the behavior observed remains unchanged.

We model the channel between the user and the base station using a frequency-flat block fading channel with i.i.d. Rayleigh coefficients – resulting in exponentially distributed random channel gains $|h_k[n]|^2$. The Signal to Interference-plus-Noise Ratio (SINR) at any time instance is calculated as

$$\gamma_k[n] = |h_k[n]|^2 \text{SINR}_k, \tag{7}$$

where SINR_k is the average SINR of user k. This is calculated according to the Okumura-Hata propagation model as $\text{SINR}_k = P d_k^{-\alpha}/(\sigma^2 + I_0)$, where P is the transmit power (in Watts [W]), d_k is the user's distance to the base station (in meters [m]), α is the path-loss exponent, σ^2 is the thermal noise, and I_0 is the average interference power. Based on this, the spectral efficiency of a user (in bits/s/Hz) at time n is

$$r_k[n] = \log_2(1 + \gamma_k[n]). \tag{8}$$

Although the utility function can be something more intricate, the utility of the operator is measured in terms of the actual rate that a given MVNO's user achieves. Therefore, in a user-centric manner, the utility of each user is measured as

$$U_k(x_k[n],\ r_k[n]) = x_k[n] r_k[n]. \tag{9}$$

B. Performance Results

In Fig. 3, we begin by comparing the ability of various scaling coefficients to cope with variations in ξ_m due to the updates in the values of S_m and Δ_m caused by using (2) and

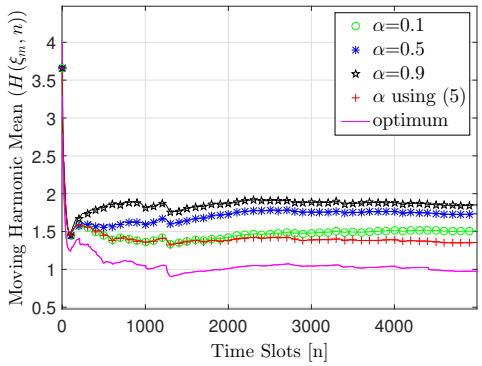
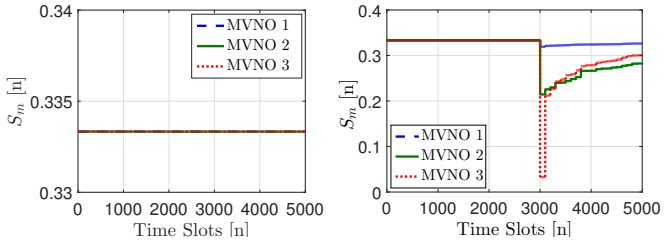


Fig. 3. Moving harmonic mean of the 'total' gap of the MVNOs computed over the all the previous time slots up to the current time slot n, for a time window $W = 100$ ms.

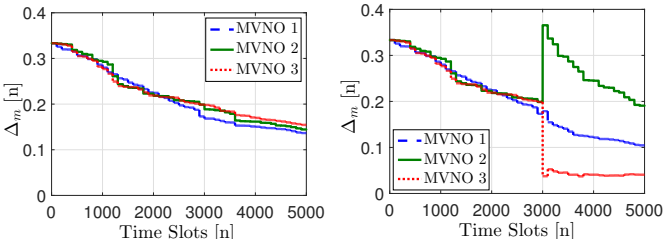
(3). The results of using various scaling coefficients are also compared with the case where the achievable rates for the upcoming time slots are known and the optimization problem can be solved for the entire time window as a whole. Fig. 3 plots the variations in the moving harmonic mean of ξ_m, $H(\xi_m, n)$, over all the time slots up to the time slot n in order to ensure that its 'peak' variations are more accurately captured than can be done when the arithmetic mean is used. We observe that the dynamic scaling coefficient computed using (5) outperforms the fixed scaling coefficients and is, as a result, closest to the 'optimal' case, i.e., the case where the rates for the subsequent time slots are known.

Fig. 4 and Fig. 5 document the variations in S_m and Δ_m, under two different cost scenarios. The objective function minimizing the maximum S_m in (1a), results in the same value of S_m for all the MVNOs as long as the MVNOs have the necessary budgets. Therefore, for the sufficient budget scenario (listed in Table I), since all the MVNOs never face a budget shortfall, $S_m = 0.33$ for all the MVNOs (Fig. 4(a)). Since each MVNO has the same S_m and incurs the same CapEx, this can be considered as the cost incurred to enter the coalition. In contrast to this initial sufficient budget scenario, for the second scenario (Fig. 4(b)) all the costs are doubled and obtaining network resources becomes more expensive. For this case, due to the infeasibility of the MVNOs' budgets, S_m takes a smaller value than 0.33. By decreasing S_m, MVNOs decrease their overall CapEx and also try to achieve their objective in (1a). However, this CapEx adjustment is required only during budget shortfall in order to avoid underutilized resources. Fig. 5(a) shows the changes in Δ_m when the MVNOs have sufficient budgets. Since the window W is large, MVNOs have the ability to balance their utility targets and resource consumption based solely on S_m and their willingness to trade short-term resources, i.e. Δ_m, decreases over time. However, during a budget shortfall, the MVNOs are not able to buy enough resources due to budget infeasibilty. Therefore, they have a higher incentive to share unused resources. In other words, as observed in Fig. 5(b), the MVNOs compensate for fluctuations in their resource requirements using Δ_m.



(a) Sufficient budget scenario (b) Insufficient budget scenario

Fig. 4. Variation in S_m over time ($W = 100$ ms).



(a) Sufficient budget scenario (b) Insufficient budget scenario

Fig. 5. Variation in Δ_m over time ($W = 100$ ms).

Fig. 6 and Fig. 7 provide additional insights into fairness in terms of resource distribution when there is no budget shortage. Fig. 6 is the cumulative distribution function (CDF) of the achieved rates for each MVNO. For all the MVNOs, we see that the probability of having a rate equal to 0 Mbps is around 0.6, which is a direct result of insufficient resources and maximum-rate scheduling. More specifically, the scheduler assigns all the resources to the user with the best channel conditions. Therefore, in a crowded network with similar user distributions, each MVNO can have the channel for $1/|M|$ of the time. Fig. 6 also shows that the MVNOs' achievable rate distributions are very close to each other. This similarity shows that, despite the initial differences in resource distributions, MVNOs achieve similar rates on a relatively long-term. Fig. 7 plots the fluctuations in the moving arithmetic mean of ξ_m, $A(\xi_m, n)$, over all the time slots up to the time slot n. Despite the large deviations due to the channel quality and the initial S_m and Δ_m values, it is seen that $A(\xi_m, n)$ attains a stable characteristic around 2000 ms. The fluctuations observed till 2000 ms can be attributed to the transient state of the model and the non-optimal selection of S_m and Δ_m. However, after this point, they reach a steady state which suggests that no further improvements can be achieved just by changing S_m and Δ_m for given channel conditions.

Finally, the costs for sharing over various time windows and for not sharing are given in Fig. 8 for each MVNO. For the no-sharing scenario, the MVNOs are assumed to have their own infrastructure; whereas, for the static sharing case, the MVNOs share a fixed portion of resources. For static sharing case, since the MVNOs share a fixed portion of the resources, we assume that they will also share the overall expansion cost equally. The MVNOs' costs are calculated using (1f) and averaged over the simulation duration N. As observed in

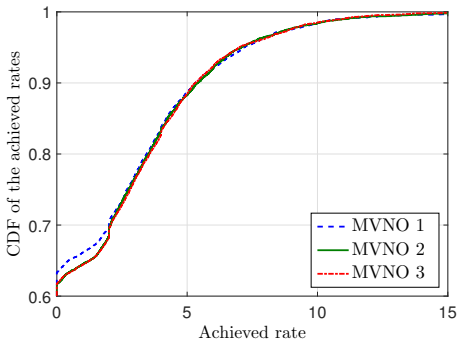


Fig. 6. Empirical CDF of achieved rates.

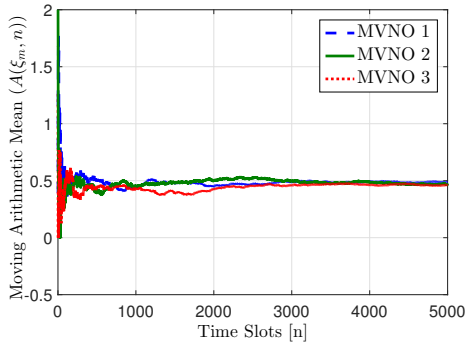


Fig. 7. Moving arithmetic mean of the 'total' gap of the MVNOs computed over all the previous time slots up to the current time slot n, for $W = 100$ ms.

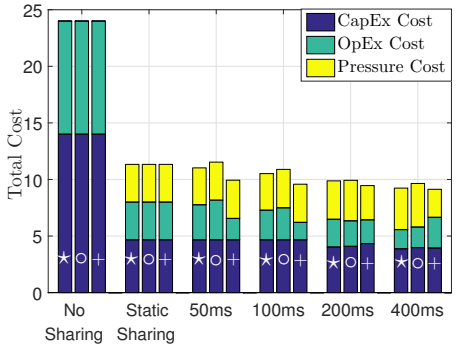


Fig. 8. Distribution of the costs (i.e. CapEx, OpEx and pressure) over MVNOs for different sizes of time windows and sharing options. (In the figure, ⋆ represents MVNO 1, ○ represents MVNO 2 and, + represents MVNO 3.)

Fig. 8, our framework provides an expenditure scaling based on the MVNOs' utility targets and the resources they utilize. In conclusion, by using a more flexible model, we obtain a higher spectral efficiency than in static/no sharing scenarios (as shown in [11]) at comparable costs while ensuring that the MVNOs pay solely for what they use.

V. Conclusion

In this paper, we proposed a novel dynamic pricing and resource sharing algorithm for multi-tenant networks. The framework proposes a real-time wireless resource market and adjusts resource prices based on their scarcity and the need for possible expansion in the future. This models also enables MVNOs to adjust their total expenditure based on their utility targets and the flexibility that can be tolerated while achieving them. It also imposes fairness in terms of the MVNO's SLA based sharing ratio, which is considered as the cost of entering the coalition. This model affords the MVNOs the ability to adjust their total expenditure based on their individual budgets and (user-dependent) utility targets. Finally, by proposing pressure costs proportional to the MVNOs' budgets, the threat of monopolization is reduced by ensuring that MVNOs with large budgets are penalized if they try to hoard resources in order to artificially inflate the unit cost of resources.

Acknowledgement

The authors would like to thank Dr. -Ing. Thomas Haustein for his valuable suggestions which gave rise to this work.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 643002.

References

[1] NGMN, "5G White Paper," 2015. [Online]. Available: http://www.ngmn.de/5gwhite-paper.html

[2] Alcatel-Lucent, "5G is coming: Are you prepared?" 2015. [Online]. Available: {http://www2.alcatel-lucent.com/landing/5g/}

[3] Alcatel-Lucent Bell Labs, "The declining profitability trend of mobile data: What can be done?" 2011. [Online]. Available: http://www3.alcatel-lucent.com/belllabs/advisory-services/documents/Declining_Profitability_Trend_of_Mobile_Data_EN_Market_Analysis.pdf

[4] OECD, "Wireless market structures and network sharing," 2014. [Online]. Available: http://dx.doi.org/10.1787/5jxt46dzl9r2-en

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

[6] V. Heinonen, P. Pirinen, and J. Iinatti, "Capacity gains through inter-operator resource sharing in a cellular network," in *11th International Symposium on Wireless Personal and Multimedia Communications (WPMC)*, Sept 2008.

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

[8] 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.

[9] G. Salami and R. Tafazolli, "Interoperator dynamic spectrum sharing (analysis, costs and implications)," *International Journal of Computer Networks (IJCN)*, vol. 2, pp. 47–61, 2010.

[10] J. Luo, J. Eichinger, Z. Zhao, and E. Schulz, "Multi-carrier waveform based flexible inter-operator spectrum sharing for 5G systems," in *Dynamic Spectrum Access Networks (DYSPAN), 2014 IEEE International Symposium on*, April 2014, pp. 449–457.

[11] 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.

[12] F. Berkers, G. Hendrix, I. Chatzicharistou, T. De Haas, and D. Hamera, "To share or not to share?" in *Intelligence in Next Generation Networks (ICIN), 2010 14th International Conference on*, Oct 2010.

[13] D.-E. Meddour, T. Rasheed, and Y. Gourhant, "On the role of infrastructure sharing for mobile network operators in emerging markets," *Computer Networks*, vol. 55, no. 7, pp. 1576–1591, 2011.

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

[15] Y.-K. Song, H. Zo, and S. Lee, "Analyzing the economic effect of mobile network sharing in korea," *ETRI Journal*, vol. 34, no. 3, pp. 308–318, 2012.

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.

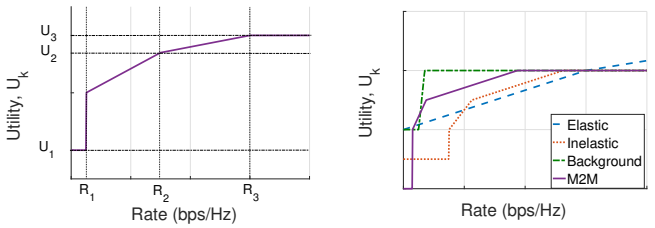
B. Utility Functions

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

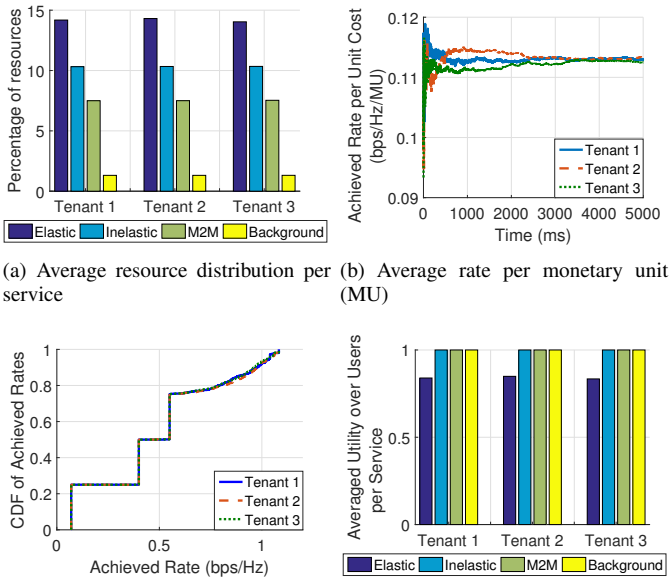
We consider the downlink of a single base station shared by 3 tenants, i.e., $M = 3$. The total number of active users is $|K| = 24$ and they are distributed equally among the 3 tenants, i.e., $|K_m| = 8,\ \forall m \in M$. Users are uniformly distributed within the coverage area of the base station 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

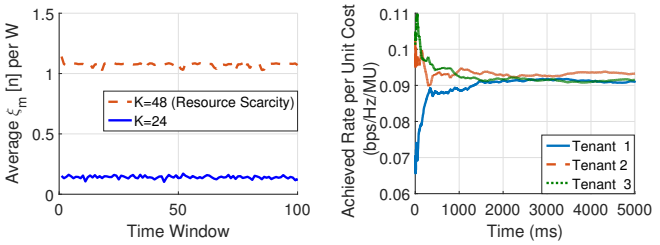
(d) Average utility over users per service per tenant

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

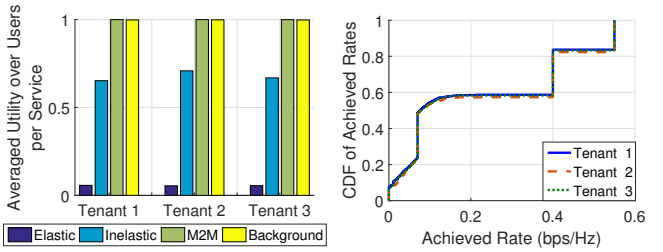
The findings in [10] also showed that the size of the time window plays a significant role in the ability of the framework to adapt to network fluctuations. Using a metric called the "Relative Distance to Optimum" (RDO), which described how close the selected parameters are to their optimum values, [10] 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

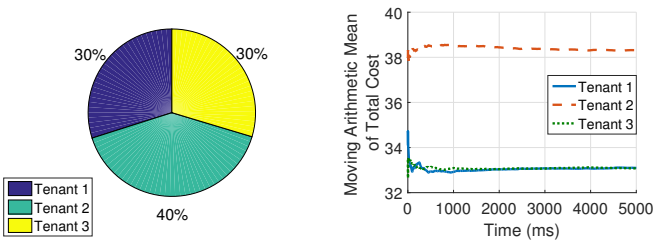


(a) Average gap of tenants over W (b) Average achieved rates per monetary unit (MU)

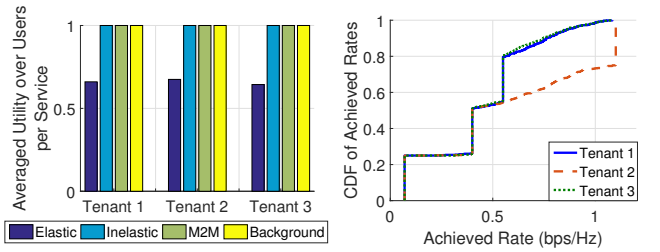


(c) Average utility over users per service per tenant (d) CDF of achieved rates per tenant

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

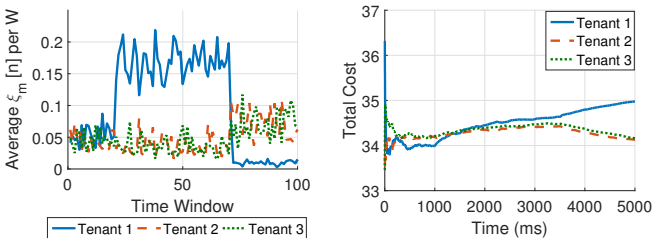
The effects of resource scarcity, documented in Fig. 3, are studied by increasing the number of active users. Fig. 3(a) shows the increase in the average difference between the utility target of the tenants and the utility they actually achieved over a time window, when the number of active users are 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.