This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3303097

Efficient Resource Allocation with Provisioning Constrained Rate Variability in Cellular Networks

Fidan Mehmeti, *Member, IEEE,* Thomas F. La Porta, *Fellow, IEEE*, and Wolfgang Kellerer, *Senior Member, IEEE*

Abstract—While LTE networks are known to provide relatively high data rates, reaching values as high as tens of Mbps, these rates exhibit considerable variability over time. The rate variability hurts especially the performance of applications and services that require stable data rates, such as real-time video streaming, online gaming, virtual reality, augmented reality, etc. 5G emerged as a solution to this as well as to many other problems. However, it has been shown that strict constant data rates come at the cost of underutilized network resources, resulting in inefficient operation of cellular networks. Therefore, a tradeoff between the data rate stability, important to cellular users, and the efficient utilization of resources, important to network operators, needs to be taken into account. To that end, in this paper, we consider the problem of allocating *all* the network resources to cellular users in such a way that it provides as high a data rate as possible to all users while limiting the rate variation within tight bounds. We do this for different scenarios in terms of the user activity, user type, and the nature of the policy. Firstly, we consider the case of static allocation policy, irrespective of channel conditions, for users that are always active. Then, for these same users, we look at the case when resources are allocated dynamically over time. Secondly, we consider static and dynamic policies for users that are only intermittently active. Thirdly, we consider the case with users having different Service Level Agreements (SLAs) with the cellular operator. Furthermore, we run extensive simulations with input parameters from real traces. Results show that allocating the resources dynamically improves performance in terms of data rates over static allocation mechanisms by an additional 10%, and that allowing a slightly higher outage in not complying with the guaranteed data rate further increases the user's throughput by at least 20%.

Index Terms—Resource allocation, 5G, QoE, Cellular networks, Network slicing.

1 Introduction

While LTE networks are capable of providing relatively high data rates (up to tens of Mbps) [2], these rates are characterized by a high variability over time [3], irrespective of which of the known resource allocation policy is being used. This renders the performance of applications and services that require *constant* or *very low-variability* data rates, such as live video streaming, online gaming, augmented reality, and virtual reality unacceptable, leading to severe deterioration in the Quality of Experience (QoE) of cellular users. As the most viable solution to alleviate or completely overcome this and several other challenges, like providing ultra low latency with high reliability [4] or providing service to a large number of devices within a given area [3], 5G networks were developed and have already been deployed for three years now.

Providing a strict constant data rate has been considered comprehensively before [5], both from the users' and operator's perspective. However, it has been shown [6] that it leads to very inefficient usage of network resources, leaving an abundance of resources unused (of interest to cellular operators) and not providing as high a data rate as expected (of interest to cellular users). On the other hand, allocating all the available resources without providing any guarantees on the stability of the data rate, as already mentioned, deteriorates the user experience when running applications with stable-throughput requirements.

The expected question that arises is *what is the tradeoff between these two seemingly contradictory approaches that satisfies everyone, i.e., cellular users and wireless operators?* To the best of our knowledge, the problem of stable data rate provisioning with efficient resource utilization has not been considered before. To reconcile the need for low-variability data rates with efficient utilization of network resources, in this paper we propose an approach in which we allow a *small deviation* of the data rate from a targeted value for all the users (i.e., a rate with *constrained variability*) while simultaneously exploiting all network resources. The magnitude of the allowed variation of the data rate can be controlled by the operator. The deviation that we allow represents only a small portion of the aimed data rate.

Provisioning a stable[1] data rate in current wireless communication systems is very challenging, mainly because of the fact that cellular networks are characterized by highly dynamic channel conditions because of the mobility of users and inherent effects like shadowing [7], together with the varying number of users receiving service simultaneously by the same base station. Things become even more intricate when the requirement of no-wastage of network resources is added.

F. Mehmeti (fidan.mehmeti@tum.de) and W. Kellerer (wolfgang.kellerer@tum.de) are with the Chair of Communication Networks, Technical University of Munich, Germany.
T. La Porta (tlp@psu.edu) is with the Department of Computer Science and Engineering, The Pennsylvania State University, USA.
Initial results of this paper have been presented at ACM MSWiM 2021 as a poster and have been published in Proceedings of ACM Q2SWinet 2021 [1].
This work was supported in part by the Bavarian Ministry of Economic Affairs, Regional Development and Energy under the project "6G Future Lab Bavaria", and in part by the Federal Ministry of Education and Research of Germany (BMBF) under the project "6G-Life", with project identification number 16KISK002.

1. Note that with *strict constant* data rate we refer to the case when the rate is constant over time, e.g., a user receives 10 Mbps (almost) at all times. On the other hand, *stable* rate (with constrained variability) refers to a value that exhibits a slight deviation from a value, e.g., providing to a user a data rate in the range $[9.5, 10.5]$ Mbps.

Consequently, in general, a different amount of network resources need to be assigned to the users at different times in order to satisfy the requirement of data rate stability. Hence, of particular importance is to determine the appropriate resource allocation policy that maintains the communication quality of the users and does not leave network resources non-utilized.

There are several important research questions that arise related to the joint efficient resource allocation and provisioning of *constrained-variability data rates* in cellular networks:

- Given the maximum allowed deviation of the rate from a target data rate[2] nearly *at all times*, what is the value of this targeted data rate? Which policy accomplishes that? Is it a *static* or a *dynamic* one?
- What is the gain in terms of increased targeted data rate if the outage probability for having the data rate within the prescribed region is relaxed?
- Are there any further gains if there is a differentiation between the users in terms of their Quality of Service (QoS)?

To answer these questions, in this paper we address the joint problem of *efficiently allocating all the resources* within the cell and *providing data rates which are characterized by low variability*, while achieving as high a data rate as possible. The results we provide here are helpful for cellular network operators in appropriately allocating the resources so that the QoE of mobile users related to the corresponding services/applications of interest, which can perform well even without a strictly constant rate *at all times*, is maximized. Also, we show that in a practical scenario (real-time video streaming) our approach can improve considerably the performance of mobile users. The main message of this paper is that by allowing a small deviation, and hence a very small variation in data rates, it is possible to provide high and stable data rates while simultaneously utilizing the entire spectrum of network resources. While we mainly focus on 5G, as the new cellular generation that is becoming ubiquitous, when evaluating the performance we also show results related to 4G, where the latter shares some similarities with 5G in terms of the structure of resources, with lower corresponding data rates only.

Specifically, our main contributions are:

- We determine the maximum achievable *center of region* data rate that can be guaranteed to all the users in the cell with a given outage probability for users that experience heterogeneous channel conditions. This is done under the assumption that all the network resources have to be fully allocated within the cell. The analysis is performed under a general setup, and can be suited to any kind of realistic system characterization.
- We then consider a dynamic policy where resources are allocated differently on different frames, depending on the channel conditions of every user in a given frame, and determine the maximum achievable targeted data rate with the given outage.
- We provide the analysis corresponding to the scenario when users are only intermittently active, showing that data rates are higher compared to the case with always-active users.
- We derive the maximum targeted data rate that can be achieved when users have different Service Level Agreements (SLAs) with the operator, showing also the benefits carried out by this approach.
- We validate our analysis with extensive realistic simulations, where the input parameters are taken from real-life traces. We also obtain some interesting engineering insights, such as an almost linear increase in the achievable data rates by only allowing a small widening of the allowed interval of the experienced data rates.
- We compare our approach and show its superiority against state-of-the-art benchmark models in terms of resource efficiency and in terms of QoE for a practical use case.

The remainder of this paper is organized as follows. Section 2 presents some related work. We introduce the system model and problem formulation in Section 3. This is followed by the analysis for the static policy in Section 4. The dynamic policy is presented in Section 5. The analysis for the users which are not always active is provided in Section 6. In Section 7, we analyze theoretically the best performance that can be achieved when users are split into different classes, according to their QoS determined by their SLA with the operator. Some performance evaluation results with additional engineering insights, including outcomes from a practical use case scenario, are provided in Section 8. Finally, Section 9 concludes the paper.

2 Related Work

To our best knowledge, there are only few works that address the problem of resource allocation which provides a constant or low-variability data rates to users in any cellular network in general, and in 5G in particular, especially from the analysis perspective.

Reference [8] provides a detailed overview of various service requirements in 5G together with a thorough description of physical layer characteristics, such as modulation, coding, and achievable capacities, spectrum sharing techniques and the architectures that enable offering these services. It also mentions the consistency requirement as one of the 5G features. Authors in [9] focus on providing the lowest possible latency (a type of delay consistency) and propose a reconfigurable architecture that enables it, together with new procedures for device attachment, connectivity and mobility management. However, that work is not concerned with throughput stability.

Additionally, in [10], the goal is to minimize the end-to-end delay. The authors propose an architecture, called SDUN, to accomplish that. A queueing network model (exhibiting memoryless properties mostly) is proposed and the presented theoretical analysis leads to finding the average waiting time in the system. This work is consistent in the delay sense, but does not consider resource allocation to provide stable rates. A work similar to [10] is [11], where the goal is to provide a *consistent delay* for Machine-to-machine (M2M) communications. The analysis in [11] relies on large deviation theory. To meet the latency and reliability constraints in 5G, a periodic radio resource allocation is proposed in [12] and the corresponding Modulation and Coding scheme (MCS) is selected to minimize resource consumption. However, providing a low-variability data rate is not considered in any of these works, and user mobility is not taken into account in [11], which is the case with our work.

In [13], the authors propose a use case for 5G deployment, and derive the average throughput by combining four different traffic types, which are ultra-high definition (UHD) video, content sharing, web browsing, and virtual reality (VR) experience. They also obtain the data rate distributions required for each of the four service types considered. However, while the analysis in [13] is

2. This data rate will be called *central value* or *center of region* data rate in Section 3.

quite detailed, it is constrained by considering only one user with four application types or four users with one application type, and it does not generalize to the case with any number of users, or applications/services.

The highly variable nature of data rates in cellular networks has been documented in [14], with a coefficient of variation going as high as 3. Similar conclusions were obtained in [15], where even for static users the data rates were exhibiting non-stable properties, with a coefficient of variation around 2. While quantifying the rate variability is certainly useful, neither [14] nor [15] provide insights on how to reduce the rate variability, which is what we do in this work. In [16], where the focus is on video streaming, the authors acknowledge the data rates with high variability, and propose a dynamic adaptation of the rate at which the video is rendered (video resolution) in order to avoid video stalling. However, in Section 8 we show that our approach of providing limited varying data rate outperforms by a significant margin the adaptive streaming approaches. Similarly, the lack of throughput stability has been shown in [17] as well. But, there are no allocation policies that prevent that to happen. On the other hand, in our work we propose several policies pertaining to various scenarios that provide throughput with low variability.

Some works related to resource allocation in 5G are [18], [19], [20], [21], and [22]. In [18], the authors consider the resource allocation problem in a multi-tier mobile edge computing setup. The considered resources are the computational units on the clouds, but not the spectrum resources as are in our work. They use the data rate as part of the calculations of the task processing delay, and more specifically, in the offloading part. However, there are no guarantees on the data rate in [18]. Slice dimensioning can be considered as a resource allocation problem too, where the goal is be to determine the number of PRBs that comprise a given slice, which would serve the same use-case users. A work in that direction is [19], in which three main types of 5G services are considered (eMBB, URLLC, and mMTC) in a multi-tenant 5G system. There are considerable differences between our work and [19]. Namely, the URLLC traffic is time-sensitive, whereas mMTC are characterized by massiveness. Our approach is more tailored towards eMBB services. We show that providing a rate within some (usually tight) limits leads to higher rates than when providing fixed data rates, which could improve the performance of eMBB users. As such, it could be used by the approach in [19] to improve the performance. Optimizing the performance of 5G networks with massive MIMO has been considered in [20]. The authors formulate a multi-objective optimization problem, where one of the objectives, related to our work, is to maximize the user's average data rate. However, while maximizing the data rate is important, having no guarantees on the range of the actual values can lead to severe performance degradation, especially for services that require stable throughput. As we show in the case of live video streaming in our work, no-rate-guarantee policy leads to lower quality of experience among users. Moreover, in [20], the allocation policy is not given in an explicit analytical form.

On a related note, in [21] the authors consider another aspect of optimal resource allocation. Specifically, the goal is to decide jointly on the allocation of the radio, optical, and mobile edge computing resources in a 5G network while minimizing the power consumption. However, there are no requirements on the data rate stability, which can hurt the performance of applications/services like real-time video streaming. In contrast, in our work we focus on these type of applications that are rate-sensitive and show that allowing a slight deviation from the strict-rate requirement can lead to considerable performance improvements for the users and the cellular operator. In [22], the authors focus on allocating resources for coordinated multipoint in 5G networks. The type of traffic of interest in [22] is URLLC. The objective is to minimize the required bandwidth subject to limited network resources and the latency that should be met. But, there are no guarantees on the rate stability. On the other hand, we try to provide a data rate that is as high as possible and within a given interval, where the width of the latter can be tuned by the operator to achieve a trade off between rate stability and magnitude. More importantly, using our approach and knowing the achievable stable data rate, we are able to predict quite accurately the transmission delay as well, and if low enough, it can provide the reliability guarantee to delay-sensitive traffic (if the traffic is not very intensive).

The works in which there is a strict requirement on the constant data rate for all users (also known as *consistent rate*) in cellular networks, with the focus on 5G[3], are [23], [24], [5], [25]. In [23], the objective is to determine the maximum number of consistent users, i.e., users with constant rate at almost all times, which can be admitted by the base station to receive service without violating the QoS. The significant impact of the channel variability on the number of consistent users to be admitted is also shown in [23]. In [25], the problem of providing a consistent backhaul rate to public urban transportation systems, like buses and trains, is analyzed. The analysis captures the scenario with two bus lines having a different number of vehicles. Within the same bus line, all the vehicles have identical statistics of channel conditions. The most important conclusion from [25] is that on average about 67% of the resources remain unused, i.e., they are wasted. In contrast, in our paper, the analysis captures the case with any number of users and utilizes network resources completely, while maintaining a low variability on the data rate, and improving considerably the user performance in terms of the achievable rates as well.

On a related note, the problem of determining the maximum consistent data rate that can be offered to a group of users in the cell is analyzed in [5]. One of the main outcomes from [5] is that providing a consistent rate to everyone leads to inefficient utilization of network resources. One of the ways, proposed in [5], to alleviate this problem is to reallocate the unused resources *equally* to the same users. However, assigning these unused resources to the same users leads to highly-variable data rates as channel conditions change rapidly over time, and hence the dynamic nature of the amount of used and consequently, unused resources. This leads to performance deterioration for applications and services which require a stable throughput.

A similar approach is followed in [24] and [6], where in [24] after providing the maximum achievable constant rate, the authors propose to reallocate the unused resources in order to satisfy two different objectives, separately. The first is to maximize the total cell throughput after reallocation, whereas the second is to provide proportional fairness. In addition to those two objectives, in [6] the goal is to allocate the unused resources (after guaranteeing a constant rate) in order to provide max-min fairness. While in all these scenarios the resources are fully utilized, there is still a high variability in the data rate over all users after reallocation, making those policies less suitable for applications and services

3. In the previous generations of cellular systems, reducing the rate variability was not of uppermost importance, as also that would have reduced even further the achievable data rates, which were already not that high.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3303097

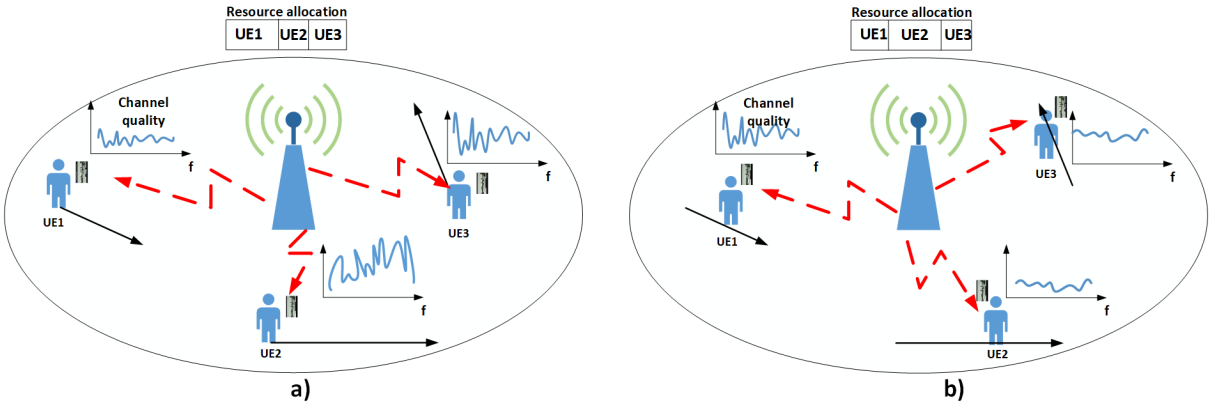

Fig. 1: Users and their channel qualities in different positions at: a) time t_1, b) time $t_2 > t_1$.

that require a stable throughput.

Finally, this work is a considerable extension of [1]. In the current paper, we provide the analysis for users that are not always active, which can also be pertinent to users coming in and going out of the cell. Additionally, we consider the scenario with two classes of users, contrary to [1] where all the users have the same SLA with the cellular operator.

To our best knowledge, no other works exist that consider the problem of resource allocation for providing stable rates in 5G and beyond, and neither those proposing or analyzing the impact on performance and operator efficiency when there is an aberration from the consistent-rate requirement. This is the first work that considers jointly the problem of minimizing rate variability and not wasting the (valuable) network resources. This fact emphasizes even further the importance of our results.

3 Performance Modeling

In this section, we first present the system model. Then, we describe the formulation of the problem considered in this paper.

3.1 System model

We consider mobile users (UEs) within the coverage area of a 5G macro base station (gNodeB) in the Frequency Range 1 (FR1), i.e., the sub-6 GHz band. The focus is on the downlink. An illustration of the system is depicted in Fig. 1.

Similar to 4G, the block resource allocation scheme is used in 5G as well, with *physical resource blocks (PRB)* being the allocation unit [26], but with higher flexibility in choosing the block bandwidth, and correspondingly, the duration of the unit of resource allocation. Within a frame, different blocks are assigned to different users. In general, the assignment will vary across frames. Consequently, scheduling is to be performed along two dimensions, *time* and *frequency*. To make the problem setup more general, we assume that the total number of available blocks for the users in a cell is K.

In general, due to the wide span of frequencies the blocks share, users will experience different channel conditions in different frequencies (different blocks) even within the same frame, and therefore a different *per-block* Signal-to-Interference-plus-Noise Ratio (SINR). The latter is a function of the base station transmission power of the cell in which the users are located, the transmission power of neighboring base stations transmitting on the same sets of frequencies (inter-cell interference), Additive White Gaussian Noise (AWGN), and the corresponding channel gains affected by free space loss and shadowing [5]. Because of user's mobility and time-varying channel characteristics, per-block SINR changes from one frame to another even for the same block. This changing *per-block* SINR translates into a varying *per-block rate*. The value of SINR in a frame determines the Channel Quality Indicator (CQI). It is a parameter sent by each UE to gNode, which depending on the Modulation and Coding scheme (MCS) sets the per-block rate. There are 15 possible values of CQI [27]. For instance, if at time t the per-block SINR lies in the interval $[\gamma_l, \gamma_{l+1}]$, with γ_l and γ_{l+1} being the thresholds of the CQI ($l = 1, \ldots, 15$), the per-block rate would be $r_l(t)$ [5].

Further, for every user, we assume flat blocks in a frame, i.e., the per-block rate (of any block) of a user does not change during the frame, but it changes from one frame to another randomly. In Section 8, we show that even when per-block rates are correlated, the analysis provides a close match to actual results.

Although in practice different blocks in general "bring" different rates, for the sake of analytical tractability, we make a simplifying assumption. Namely, we assume that gNodeB transmission power and channel characteristics of a user remain unchanged across all K blocks in a frame. In this way, our problem reduces to one-dimensional scheduling, in time. Hence, instead of deciding how many and which blocks to assign to every user in every frame, we use another (but related) parameter. It is defined as:

Definition 1. *The ratio of frame during which all the network resources (blocks) are allocated to user i is denoted by Y_i, and is called **frame ratio**. It can take any value in the interval* $[0, 1]$.

The blocks are assigned orthogonally across the frame duration, so that no two users receive the blocks simultaneously. This simplification is reasonable as frame ratio can be translated into the corresponding number of resource blocks per frame. Therefore, in this paper, the *frame ratio* will be the quantitative figure of merit related to resource allocation in a cell.

With the previous assumptions in mind, it follows that in every frame the per-block rate of user i can be modeled as a discrete random variable, R_i, with values in $\{r_1, r_2, \ldots, r_{15}\}$, such that $r_1 < r_2 < \ldots < r_{15}$, with a probability mass function (PMF) $p_{R_i}(x)$. The latter is a function of user's i SINR over time.[4]

Number of users: We consider two scenarios in terms of the number of users in the cell. In the first, there are n users in the cell with different per-block rate distributions, i.e., the users are

4. We omit the reference to time from now on for notational convenience.

heterogeneous, and all of these users are active *at all times*. We also consider the scenario in which the number of users changes over time, i.e., users enter and leave the cell. In that case, we use the random variable $N(t)$ to denote the number of users in the cell at frame t. We assume that the maximum number of users in the cell can be N_{max}.

User activity: Besides considering users that are always active, we also consider the case where users that are in the cell are active with a given probability $\pi_{i,active}$ in the frame. Essentially, this scenario is equivalent to the one where users come to the cell and leave it after some time. Therefore, we confine our analysis to the scenario where the number of users is fixed across time, with all the users being always active, and to the scenario in which the number of active users varies. We will refer to the former as *fixed user activity*, whereas to the latter as *dynamic user activity*. The set of active users in the cell is denoted by $\mathcal{N}$.

Types of users: There are two scenarios that we consider in terms of the types of users as well. In the first, all the users require the same level of service from the network, whereas in the second users have different SLAs with the operator. Based on that, for the latter, we split users in two groups - users that pay more for the service, and hence experience better performance, are referred to as *premium users*, while users that pay less are referred to as *regular users*.

3.2 Problem formulation

The possibility of network slicing in 5G [28], non-existing in the previous generations of cellular networks, enables assigning *dedicated* network resources to the same application type and to users requiring the same level of service, e.g., users watching live the same event or doing online gaming. Throughout this paper, we assume that the users belonging to the same class have the same SLA with the operator, and therefore will require the same service quality and will receive the resources from the same network slice.

Data rate: In order to efficiently utilize all network resources, which is not possible when providing a strict data rate [5], we allow a slight deviation from the strict rate. Nevertheless, this deviation is small enough to keep the variability of data rates limited and controlled. In this way, by trading off the requirement for strict data rate with efficient utilization of resources, the overall performance of cellular networks, from both the user's and operator's perspective, improves. To capture this effect, we need to introduce the following:

Definition 2. *The **targeted data rate** for all the users in the cell is denoted by U_c. This is also known as the **center of region (interval)** data rate or **central value**.*

To increase the efficiency, we assume that every user is allowed to have a data rate that is close to U_c, with a small deviation. To capture this effect, we introduce another variable.

Definition 3. *The allowed deviation ratio from U_c is denoted by θ, making the feasible data rate range to be $[(1-\theta)U_c, (1+\theta)U_c]$. The latter is known as the **feasible interval**.*

The deviation from the central value, which in percentage can be expressed as $100 \cdot \theta\%$, is usually small, and can be controlled by the operator. Allowing a higher θ increases the central value at the expense of an increased variability of experienced rates.

We specify another variable, which is used as a metric of interest when comparing different approaches in Section 8.

Definition 4. *The ratio between the maximum value and the minimum value of the feasible interval, i.e., $\frac{1+\theta}{1-\theta}$, is called the **span** of the data rate.*

Outage: According to our model, rates will need to fall within the required interval. However, imposing the strict requirement that rates need to be within the feasible interval 100% of the time is too restrictive. Instead, we relax this constraint to data rate being within the feasible interval with probability $1-\epsilon$, or equivalently for $(1-\epsilon)\cdot 100\%$ of the time, where ϵ is the *outage probability*.

The main question that arises is *what is the maximum achievable target data rate U_c for a given value of θ and ϵ, and what is the policy that achieves that?* We answer this question in the following sections, both for static and dynamic policies.

4 Static policy: Fixed user activity

In this section, we propose a static policy that provides the maximum achievable target data rate. It is static as the allocation is computed only once (at the beginning but taking into account the channel statistics of all the users) and then it is used throughout the entire process irrespective of the actual channel conditions in a frame. The number of users is fixed and they are all always active.

4.1 Analysis

As described in Section 3.1, the goal is to have the data rate in the range

$$[(1-\theta)U_c, (1+\theta)U_c],$$

with probability $1-\epsilon$. As a first step, we propose a *static policy*[5] that achieves this objective. First, we capture the constraint related to the feasible interval. Given that the data rate that user i receives in a frame is KY_iR_i, the rate constraint can be written as

$$\mathbb{P}\left((1-\theta)\,U_c \le KY_iR_i \le (1+\theta)\,U_c\right) \ge 1-\epsilon,\ \forall i \in \mathcal{N}, \quad (1)$$

which after some simple algebra transforms into

$$\mathbb{P}\left(R_i \le (1+\theta)M_i\right) - \mathbb{P}\left(R_i < (1-\theta)M_i\right) \ge 1-\epsilon, \quad (2)$$

where $M_i = \frac{U_c}{KY_i},\ \forall i \in \mathcal{N}$, which further yields $Y_i = \frac{U_c}{KM_i},\ \forall i \in \mathcal{N}$.

Since we assume full resource utilization in every frame, it must hold $\sum_{i=1}^{n} Y_i = 1$. From the last two conditions we obtain

$$U_c(\theta,\epsilon) = \frac{K}{\sum_{i=1}^{n}\frac{1}{M_i}}. \quad (3)$$

Summarizing, we have the following:

Result 1. *The maximum achievable center-of-region data rate with allowed deviation θ and outage ϵ in the cell with n users, whose per-block rate probability mass functions are $p_{R_i}(x)$ is given by Eq.(3), where M_i are the highest values that satisfy (2), $\forall i \in \mathcal{N}$. To accomplish this, user i needs to receive all the resources for a frame ratio of $Y_i = \frac{U_c(\theta,\epsilon)}{KM_i}$.*

In order to get the maximum achievable center-of-interval data rate U_c, we need to determine the unknown parameters $M_i,\ \forall i \in \mathcal{N}$. Based on the nature of (2), it can be concluded that the exact value of M_i can be obtained only numerically. But, there are some

5. We refer to this policy as the *static* policy in the sense that taking into account the statistics of the channel conditions the operator decides on the amount of network resources to allocate to each user and keeps that amount *fixed over time*.

insights that can be obtained related to the range of values of M_i's, which are described in the following. First, from (2) it is apparent that the following must hold:

$$\mathbb{P}\left(R_i \leq (1+\theta)M_i\right) \geq 1-\epsilon, \quad \forall i \in \mathcal{N}, \tag{4}$$

which is equivalent to

$$F_{R_i}\left((1+\theta)\,M_i\right) \geq 1-\epsilon, \quad \forall i \in \mathcal{N}, \tag{5}$$

where $F_{R_i}\left((1+\theta)\,M_i\right)$ denotes the Cumulative Distribution Function (CDF) of R_i at $(1+\theta)\,M_i$. Since CDF is a non-decreasing function, from the last inequality we have

$$(1+\theta)M_i \geq F_{R_i}^{-1}(1-\epsilon), \quad \forall i \in \mathcal{N}, \tag{6}$$

where $F_{R_i}^{-1}$ denotes the inverse function of CDF of R_i, and

$$M_i \geq \frac{F_{R_i}^{-1}(1-\epsilon)}{1+\theta}, \quad \forall i \in \mathcal{N}. \tag{7}$$

As far as the upper bound of M_i is concerned, we find it from the following inequality (in line with (2))

$$\mathbb{P}\left(R_i \leq (1+\theta)M_i\right) \leq 1, \quad \forall i \in \mathcal{N}, \tag{8}$$

which yields

$$(1+\theta)M_i \leq F_{R_i}^{-1}(1), \quad \forall i \in \mathcal{N}, \tag{9}$$

and the latter

$$M_i \leq \frac{r_{15}}{1+\theta}, \quad \forall i \in \mathcal{N}. \tag{10}$$

The parameter r_{15} is the highest possible value of per-block rate (corresponding to $CQI = 15$). Hence, $F_{R_i}(r_{15}) = 1$. From (7) and (10) we have

$$\frac{F_{R_i}^{-1}(1-\epsilon)}{1+\theta} \leq M_i \leq \frac{r_{15}}{1+\theta}, \quad \forall i \in \mathcal{N}. \tag{11}$$

Expression (11) provides the interval in which we should look for the corresponding value of M_i. The value of M_i itself, $\forall i \in \mathcal{N}$, is computed numerically.

4.2 Implementation and application

Note that, as already mentioned, this is a static policy, where the frame ratio the resources need to be allocated to every user is determined only once (at the beginning)[6] and does not change over time, irrespective of the *actual* channel conditions of the users in a given frame. This is very practical from operator's standpoint as it reduces the complexity and signaling overhead considerably.[7]

Users running applications that require stable throughput benefit most from this approach. The actual impact on real-time video streaming applications is illustrated in Section 8.4.

6. As already seen in Section 4.1, the operator considers the distribution of the per-block rates of all the users when making the decision on the frame ratio to be allocated to every user at the beginning. This amount (Y_i) remains fixed for the user over time.

7. The analyses in this work are derived under the assumption that we have exact knowledge of the distributions of channel conditions of all the users. Considering in detail the impact of inaccurate CQI distributions on the targeted data rate requires a further study and is deferred to future work. Nevertheless, it should be mentioned that for low deviations of the CQI, we observed that the targeted data rate does not show significant discrepancy from the one obtained with our analysis under an assumed distribution.

5 Dynamic policy: Fixed user activity

The previous result is very convenient as the operator needs to compute the frame ratio for every user only once, at the beginning of the process. In this section, we improve the previous result by exploiting the knowledge of the per-block rates in a frame, which can vary significantly from one frame to another. This implies the need for a dynamic policy.

5.1 Analysis

Based on the nature of the problem setup, according to which all the users need to receive the data rate within the same feasible interval, it follows that in a frame a user having a good channel will require a smaller frame ratio Y, and vice versa. As a result, user i in frame t will receive all the resources for the following frame ratio:

$$Y_i(t) = \frac{\frac{1}{R_i(t)}}{\sum_{j=1}^{n}\frac{1}{R_j(t)}}, \quad \forall i \in \mathcal{N}, \tag{12}$$

resulting in a total data rate of

$$KY_i(t)R_i(t) = \frac{K}{\sum_{j=1}^{n}\frac{1}{R_j(t)}}, \quad \forall i \in \mathcal{N}. \tag{13}$$

Hence, given that the data rate has to be in the interval $[(1-\theta)U_c, (1+\theta)U_c]$, we have the following inequality that must be satisfied:

$$(1-\theta)\frac{U_c}{K} \leq \frac{1}{\sum_{i=1}^{n}\frac{1}{R_i}} \leq (1+\theta)\frac{U_c}{K}. \tag{14}$$

So, for the rate constraint in this case, where the outage probability is ϵ, we have

$$\mathbb{P}\left(\frac{K}{(1+\theta)U_c} \leq \sum_{i=1}^{n}\frac{1}{R_i} \leq \frac{K}{(1-\theta)U_c}\right) \geq 1-\epsilon. \tag{15}$$

This further transforms into

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i} \leq \frac{K}{(1-\theta)U_c}\right) - \mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i} < \frac{K}{(1+\theta)U_c}\right) \geq 1-\epsilon. \tag{16}$$

The first left-hand side (LHS) term of (16) represents the CDF of $\sum_{i=1}^{n}\frac{1}{R_i}$ at point $\frac{K}{(1-\theta)U_c}$. Therefore, we will look at the CDF of $\sum_{i=1}^{n}\frac{1}{R_i}$ at any point x. It is known that the CDF of a sum of independent random variables is equal to the convolution of the CDF of the first term with the PMFs of the other terms [29]. So,

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i} \leq x\right) = \mathbb{P}\left(\frac{1}{R_1} \leq x\right) * \mathbb{P}\left(\frac{1}{R_2} = x\right) * \ldots * \mathbb{P}\left(\frac{1}{R_n} = x\right), \tag{17}$$

where * denotes the convolution operation. The first right-hand side (RHS) term of Eq.(17) is equivalent to

$$\mathbb{P}\left(\frac{1}{R_1} \leq x\right) = \mathbb{P}\left(R_1 \geq \frac{1}{x}\right) = \sum_{k_1=1}^{m} p_{1,k_1} \cdot u\left(x - \frac{1}{r_{k_1}}\right), \tag{18}$$

where $u(x)$ is the Heaviside step function [30]. The other terms of the RHS of Eq.(17) yield

$$\mathbb{P}\left(R_i = \frac{1}{x}\right) = \sum_{k_i=1}^{15} p_{i,k_i} \cdot \delta\left(x - \frac{1}{r_{k_i}}\right), \tag{19}$$

where $\delta(x)$ is the delta function [30].

Substituting Eq.(18) and Eq.(19) into Eq.(17), and after rearranging, we obtain

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i}\leq x\right)=\sum_{k_1=1}^{15}\cdots\sum_{k_n=1}^{15}p_{1,k_1}\ldots p_{n,k_n}\cdot u\left(x-\frac{1}{r_{k_1}}-\ldots-\frac{1}{r_{k_n}}\right). \quad (20)$$

Replacing $x=\frac{K}{(1-\theta)U_c}$ in Eq.(20), we obtain the first RHS term of (16).

Note that while deriving the previous expression we have used the fact that the convolution of a signal with a shifted Dirac delta function is just the shifted signal itself [30], i.e., $x(t)*\delta(t-t_0)=x(t-t_0)$.

The second LHS term of (16) is not an exact CDF. It is

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i}<\frac{K}{(1+\theta)U_c}\right)=\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i}\leq\frac{K}{(1+\theta)U_c}\right)-\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i}=\frac{K}{(1+\theta)U_c}\right). \quad (21)$$

The first RHS term in Eq.(21) is computed from (20) for $x=\frac{K}{(1+\theta)U_c}$. As far as the second RHS term in Eq.(21) is concerned, it represents the PMF of $\sum_{i=1}^{n}\frac{1}{R_i}$ at $K/((1+\theta)U_c)$. In a similar context, the PMF of a sum of independent random variables is the convolution of the PMFs of the random variables [30]. Hence, we have

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i}=x\right)=\mathbb{P}\left(\frac{1}{R_1}=x\right)*\mathbb{P}\left(\frac{1}{R_2}=x\right)*\ldots*\mathbb{P}\left(\frac{1}{R_n}=x\right). \quad (22)$$

Substituting the corresponding Eqs.(19) ($\forall i\in\mathcal{N}$) into Eq.(22) and rearranging, we obtain

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{1}{R_i}=x\right)=\sum_{k_1=1}^{15}\cdots\sum_{k_n=1}^{15}p_{1,k_1}\ldots p_{n,k_n}\cdot\delta\left(x-\frac{1}{r_{k_1}}-\ldots-\frac{1}{r_{k_n}}\right). \quad (23)$$

Next, replacing $x=\frac{K}{(1+\theta)U_c}$ into Eq.(23), we obtain the second RHS term of Eq.(21). Finally, substituting Eq.(20) and Eq.(21) into inequality (16) and rearranging the obtained expression, we have the following:

Result 2. *The maximum achievable center-of-region data rate with allowed deviation θ and outage probability ϵ, $U_c(\theta,\epsilon)$, in the cell with n users, whose per-block rate probability mass functions are $p_{R_i}(x)$, following the dynamic policy, is the maximum value of U_c that satisfies the inequality*

$$\sum_{k_1=1}^{15}\cdots\sum_{k_n=1}^{15}\prod_{i=1}^{n}p_{i,k_i}\cdot u\left(\frac{K}{(1-\theta)U_c}-\sum_{i=1}^{n}\frac{1}{r_{k_i}}\right)-\sum_{k_1=1}^{15}\cdots\sum_{k_n=1}^{15}\prod_{i=1}^{n}p_{i,k_i}\cdot u\left(\frac{K}{(1+\theta)U_c}-\sum_{i=1}^{n}\frac{1}{r_{k_i}}\right)+\sum_{k_1=1}^{15}\cdots\sum_{k_n=1}^{15}\prod_{i=1}^{n}p_{i,k_i}\cdot\delta\left(\frac{K}{(1+\theta)U_c}-\sum_{i=1}^{n}\frac{1}{r_{k_i}}\right)\geq 1-\epsilon. \quad (24)$$

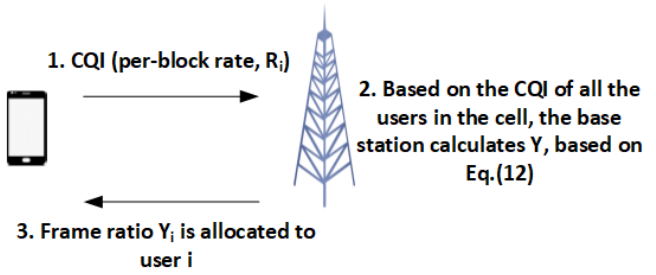

Fig. 2: Illustration of the implementation of the dynamic policy.

The resource allocation policy that enables this is given by Eq.(12), $\forall i\in\mathcal{N}$.

5.2 Implementation details

The central value U_c for the dynamic policy from (24) is computed numerically as well. To do that, the base station needs to know the distributions of per-block rates for every user. When it comes to the resource allocation policy, gNodeB has to compute the amount of frame ratio for every user at the beginning of the frame. This is done after getting the CQI in every frame, i.e., per-block rate R_i, from all the users at the beginning of the frame. Therefore, this needs to be done on a per-frame basis, which increases the implementation complexity compared to the static policy, where the latter is done only once - at the beginning. The process is illustrated in Fig. 2.

As will be seen in Section 8, the dynamic policy further improves the performance by exploiting the knowledge of channel conditions, providing a higher U_c compared to the static policy.

5.3 Solving alternative problems

In this paper, we assume that the deviation ratio θ is given, and we derive the maximum target data rate. The advantage of the approach we follow here is that we can solve other problems as well. For example, if the service requirement is that the span should not exceed a maximum value of c, then from $\frac{1+\theta}{1-\theta}\leq c$, we obtain the maximum deviation ratio the operator needs to provide:

$$\theta_{max}=\frac{c-1}{c+1}. \quad (25)$$

The procedure to determine the central value then follows as before.

Alternatively, we can solve the inverse problem - given the value of U_c, we can determine the maximum value of the deviation ratio for the data rate of all users. Or, given the maximum U_c and the maximum deviation ratio, we can answer the question of how many blocks (PRBs) are needed to accommodate a given number of users. The latter is useful for *resource planning* purposes.

6 Users with Dynamic Activity

In the analysis so far, we have assumed that the users are always active. However, in the general case, there are time periods when the users will be idle. In that case, there is no need to assign any resources to users that are not active in a frame. To capture this scenario, we introduce the indicator variable I which denotes whether a user is active ($I=1$) or idle ($I=0$). Further, we assume that when user i is active in a frame, the probability that she will keep being active in the next frame is α_i. So, the

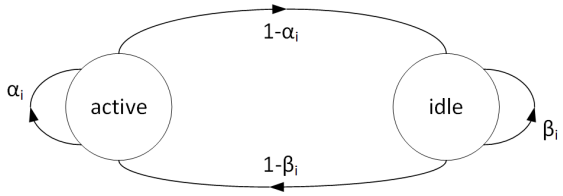

Fig. 3: Markov chain for the activity of users.

probability that the user is idle in the next frame is $1-\alpha_i$. Similarly, let β_i denote the probability that user i is idle in the next frame, given that she is idle in the current frame too. Hence, $1-\beta_i$ is the probability that the user will be active in the next frame, given that she is idle in the current one.

The Markov chain in Fig. 3 mimics the activity of user i. The local-balance equation for this chain for user i is

$$\pi_{i,active}(1-\alpha_i) = \pi_{i,idle}(1-\beta_i), \tag{26}$$

where $\pi_{i,active} = \mathbb{P}(I_i = 1)$ is the stationary probability of user i being active, whereas $\pi_{i,idle} = \mathbb{P}(I_i = 0)$ is her stationary probability of being idle. Combining $\pi_{i,active} + \pi_{i,idle} = 1$ with Eq.(26), we get the probabilities for user i being active and idle:

$$\pi_{i,active} = \frac{1-\beta_i}{2-\alpha_i-\beta_i}, \tag{27}$$

$$\pi_{i,idle} = \frac{1-\alpha_i}{2-\alpha_i-\beta_i}. \tag{28}$$

Eq.(27) is used in the analysis that follows.

6.1 Analysis

Reserving resources for users that are not always active in a given cell is very inefficient, especially when users are not highly active and when dealing with mobile users which come into and leave the cell very often. Therefore, in this scenario, we are considering only the dynamic case in terms of assigning the resource blocks. Similar to Section 5, we strive to provide the same rate to all the users that are active in a frame. No resources are allocated to the users that are idle in that frame.

The frame ratio during which user i needs all the resources is

$$Y_i(t) = \frac{\frac{I_i(t)}{R_i(t)}}{\sum_{j=1}^{N_{max}} \frac{I_j(t)}{R_j(t)}}, \tag{29}$$

which captures the case of both no resource provisioning when the user is not active, and assigning a given amount when she is active.[8] Note that we have already stated in Section 3.1 that the number of active users in the cell in this scenario is denoted by N, with a maximum possible value of N_{max}. The total rate user i experiences in frame t is

$$KY_i(t)R_i(t) = \frac{KI_i(t)}{\sum_{j=1}^{N_{max}} \frac{I_j(t)}{R_j(t)}}. \tag{30}$$

8. Note that since not all the users are active, the amount of resources allocated to a user is expected to be higher, resulting in a higher guaranteed center-of-interval data rate in this scenario.

As we need to guarantee that the data rate has to be within the feasible interval only in the periods of user activity, the rate constraint for user i is expressed as

$$\mathbb{P}\left((1-\theta)\frac{U_c}{K} \le \frac{1}{\sum_{j=1}^{N_{max}} \frac{I_j}{R_j}} \le (1+\theta)\frac{U_c}{K}\right) \ge (1-\epsilon)\pi_{i,active}, \tag{31}$$

Eq.(31) transforms into

$$\mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} \le \frac{K}{(1-\theta)U_c}\right) - \mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} < \frac{K}{(1+\theta)U_c}\right) \ge (1-\epsilon)\pi_{i,active}. \tag{32}$$

Eq.(32) is similar to Eq.(16) except for the indicator random variables capturing the activity or inactivity of the users. Looking for the general case of $x = \frac{K}{(1-\theta)U_c}$, the first LHS of Eq.(32) yields

$$\mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} \le x\right) = \mathbb{P}\left(\frac{I_1}{R_1} \le x\right) * \mathbb{P}\left(\frac{I_2}{R_2} = x\right) * \ldots \mathbb{P}\left(\frac{I_{N_{max}}}{R_{N_{max}}} = x\right). \tag{33}$$

Further, for the first RHS term of Eq.(33), we have

$$\mathbb{P}\left(\frac{I_1}{R_1} \le x\right) = \mathbb{P}(0 \le x)\mathbb{P}(I_1 = 0) + \mathbb{P}\left(\frac{1}{R_1} \le x\right)\mathbb{P}(I_1 = 1). \tag{34}$$

We got the last equation by conditioning on the possible values of the indicator random variable. Note that in the previous equation, $\mathbb{P}(0 \le x) = u(x)$, where $u(x)$ is the unit step function (the Heaviside function having value 1 for all the non-negative values of x, and 0 otherwise). Eq.(34) becomes

$$\mathbb{P}\left(\frac{I_1}{R_1} \le x\right) = \pi_{1,idle}u(x) + \pi_{1,active}\mathbb{P}\left(\frac{1}{R_1} \le x\right). \tag{35}$$

Observing Eq.(35), we can infer that $\mathbb{P}\left(\frac{I_1}{R_1} \le x\right)$ is a staircase function that is "elevated" by $\pi_{1,idle}$ from $\mathbb{P}\left(\frac{1}{R_1} \le x\right)$. Written in a more detailed form, Eq.(35) becomes

$$\mathbb{P}\left(\frac{I_1}{R_1} \le x\right) = \pi_{1,idle}u(x) + \pi_{1,active}\sum_{k_1=1}^{15} p_{1,k_1}u\left(x - \frac{1}{r_{k_1}}\right). \tag{36}$$

Similarly, for the other terms $i = 2, \ldots, N_{max}$ of Eq.(33), after some calculus we obtain

$$\mathbb{P}\left(\frac{I_i}{R_i} = x\right) = \pi_{i,idle}\delta(x) + \pi_{i,active}\sum_{k_j=1}^{15} p_{i,k_j}\delta\left(x - \frac{1}{r_{k_j}}\right). \tag{37}$$

The second LHS term of Eq.(32) yields

$$\mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} < \frac{K}{U_c(1+\theta)}\right) = \mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} \le \frac{K}{U_c(1+\theta)}\right) - \mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} = \frac{K}{U_c(1+\theta)}\right). \tag{38}$$

The first RHS term in Eq.(38) is computed from Eq.(33) for $x = \frac{K}{(1+\theta)U_c}$. As far as the second RHS term in Eq.(38) is concerned,

it represents the PMF of $\sum_{i=1}^{N_{max}} \frac{I_i}{R_i}$ at $K/((1+\theta)U_c)$. Similar to when obtaining Eq.(22), we have

$$\mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} = x\right) = \mathbb{P}\left(\frac{I_1}{R_1} = x\right) * \ldots * \mathbb{P}\left(\frac{I_{N_{max}}}{R_{N_{max}}} = x\right). \tag{39}$$

We have already derived the individual RHS terms of Eq.(39) in Eq.(37). We only need to substitute $x = \frac{K}{U_c(1+\theta)}$. Finally, we have the following:

Result 3. *The maximum achievable center-of-region data rate with allowed deviation θ and outage probability ϵ for user i, whose per-block rate PMF is $p_{R_i}(x)$, is the maximum value of $U_c(\theta, \epsilon)$ that satisfies the inequality*

$$\mathcal{L}_1(U_c) - \mathcal{L}_2(U_c) \geq (1-\epsilon)\pi_{i,active}, \tag{40}$$

where $\mathcal{L}_1(U_c) = \mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} \leq \frac{K}{(1-\theta)U_c}\right)$, and $\mathcal{L}_2(U_c) = \mathbb{P}\left(\sum_{j=1}^{N_{max}} \frac{I_j}{R_j} < \frac{K}{(1+\theta)U_c}\right)$.

For notational convenience, let us denote by $U_{i,max}$ the largest value of data rate for user i that satisfies (40). As the goal with our problem setup is to provide the same targeted data rate to all the users, we have the following:

Result 4. *The highest value of central data rate that can be provided to all the users in the cell is*

$$U_c(\theta, \epsilon) = \min\left(U_{1,max}, \ldots, U_{N,max}\right). \tag{41}$$

Note that the data rate for all users will be in the range $[(1-\theta)U_c, (1+\theta)U_c]$ for $1-\epsilon$ of the time they are active. It is also worth mentioning that the "bottleneck" in the system are the users with the highest activity (higher π_{active}) and those with bad channel conditions. Both these types tend to reduce the value of $U_c(\theta, \epsilon)$. Namely, users with bad channel conditions will require more resources for themselves, leaving the users with good channel conditions with fewer resources, resulting in a lower targeted data rate. On the other hand, highly-active users lead to higher RHS values in (40), which means lower corresponding central data rate.

6.2 Implementation insights

A similar procedure is followed by gNodeB in this case as in Section 5.2. Specifically, users that are active send their CQIs to gNodeB, which using Eq.(29) calculates the frame ratio for each active user in that frame and allocates the resources accordingly.

7 Different classes of users

In previous sections, we were assuming there was only one class of users, and all users are expected to be guaranteed the same maximum central rate with the same deviation ratio θ and outage probability ϵ. Supposedly, since all the users receive the same service, they are expected to pay the same (monthly) flat rate.

However, there may be users "that are happy with less", i.e., those that are not very interested in the highest possible quality, but instead are more interested in paying less. Moreover, some operators like AT&T in the USA are considering to introduce charging different flat rates based on the data rate and not the amount of data [31], [32]. When that is the case, in order to attract more customers, the operators may consider splitting users in groups according to the level of service they would be supposed to receive. This would be beneficial for both the mobile operators and the end users. For the former, they would be flexible in handling their resources, and for the latter, they would be paying for the service they are interested to receive. For those interested in the ultimate user experience, a higher cost would be imposed. Alternatively, users who are not very interested in a high QoS could be more interested in paying less, as is currently the case with media-service providers like Netflix, where users have the option to choose one of the packages and pay accordingly.

The development of network slicing in cellular networks [28] has fostered the operators to split the users in groups, having similar use cases or applications/services of users within the group (slice). Consequently, the operators can split the users of different classes (based on their QoS) in different slices. The network is split virtually and slices should operate in an isolated fashion, rendering this way a higher network efficiency.

In this section, we assume that users can choose one of the two classes.[9] The users with a better QoS are called *premium users*, whereas those with a worse QoS are called *regular users*. The former will be paying higher flat rates than the latter. The central value for premium users is $U_p(\theta_p, \epsilon_p) = kU_r(\theta_r, \epsilon_r)$, where $k > 1$, and $U_r(\theta_r, \epsilon_r)$ is the central value of regular users. Therefore, the interval at which the premium users' rate should be with a given outage is $[(1-\theta_p)U_p, (1+\theta_p)U_p]$, while for regular users it is $[(1-\theta_r)U_r, (1+\theta_r)U_r]$.

As far as the outage is concerned, there are two reasonable options. In the first, the operator can choose to guarantee the same outage to both premium and regular users, i.e., $\epsilon_p = \epsilon_r = \epsilon$. The second option is to provide the premium users with the data rate in the given interval with a higher probability (lower outage) than regular users. For instance, premium users could receive the data rate $[9, 11]$ Mbps for 99% of the time, while the operator can guarantee the regular users that they will receive the data rate in the interval $[4.5, 5.5]$ Mbps for 98% of the time. Therefore, in this second option, the relation between the outage probabilities of premium and regular users would be $\epsilon_p = \frac{\epsilon_r}{k}$, $k > 1$. In this paper, we consider the most general case, where ϵ_p and ϵ_r are different.

7.1 Static policy: Fixed user activity

We assume there are n_p premium users and n_r regular users. Their per-block rates are $R_{p,i}, \forall i \in \mathcal{N}_P$ for premium users, and $R_{r,j}, \forall j \in \mathcal{N}_R$ for regular users. We assume that $|\mathcal{N}_P| = n_p$, and $|\mathcal{N}_R| = n_r$.

As previously, there are in total K blocks dedicated to both classes of users. Network slicing in 5G enables splitting these resources between premium and regular users. Let K_p be the number of blocks dedicated to premium users and let K_r be the number of blocks dedicated to regular users. It holds that $K = K_p + K_r$.

There are two goals in front of a cellular operator. The first is to determine the maximum central value data rate with a given width of the feasible interval that can be guaranteed to premium and regular users with the corresponding outage probabilities. The second goal is to determine the optimal assignment of the number of blocks to both groups of users. In the following, we demystify both these riddles.

9. Considering more classes is straightforward with similar conclusions drawn. Therefore, to ease the presentation, we focus on two classes.

The rate constraint for the static policy for primary users is expressed as

$$\mathbb{P}\left((1-\theta_p)U_p \leq K_p Y_{p,i} R_{p,i} \leq (1+\theta_p)U_p\right) \geq 1-\epsilon_p, \forall i \in \mathcal{N}_P, \tag{42}$$

which yields

$$\mathbb{P}\left((1-\theta_p)\frac{U_p}{K_p Y_{p,i}} \leq R_{p,i} \leq (1+\theta_p)\frac{U_p}{K_p Y_{p,i}}\right) \geq 1-\epsilon_p, \forall i \in \mathcal{N}_P. \tag{43}$$

Replacing

$$M_{p,i} = \frac{U_p}{K_p Y_{p,i}}, \quad \forall i \in \mathcal{N}_P, \tag{44}$$

into Eq.(43) yields

$$\mathbb{P}\left(R_{p,i} \leq (1+\theta_p)M_{p,i}\right) - \mathbb{P}\left(R_{p,i} < (1-\theta_p)M_{p,i}\right) \geq 1-\epsilon_p. \tag{45}$$

For the frame ratio of premium user $i \in \mathcal{N}_P$, we have

$$Y_{p,i} = \frac{U_p}{K_p M_{p,i}}. \tag{46}$$

Then, in line with Result 1, we have

$$U_p(\theta_p, \epsilon_p) = \frac{K_p}{\sum_{i=1}^{n_p} \frac{1}{M_{p,i}}}. \tag{47}$$

In order to determine the maximum central value for premium users, we need to determine the values of $M_{p,i}$. We obtain the latter as the maximum values that satisfy the corresponding Eq.(45), similarly to the analysis in Section 4.

Similarly, for regular users, we obtain

$$U_r(\theta_r, \epsilon_r) = \frac{K_r}{\sum_{j=1}^{n_r} \frac{1}{M_{r,j}}}. \tag{48}$$

Replacing Eq.(47) and Eq.(48) into $U_p(\theta_p, \epsilon_p) = kU_r(\theta_r, \epsilon_r)$, we get

$$\frac{K_p}{M_p} = k\frac{K_r}{M_r}, \tag{49}$$

where

$$M_p = \sum_{i=1}^{n_p} \frac{1}{M_{p,i}}, \text{and} \tag{50}$$

$$M_r = \sum_{j=1}^{n_r} \frac{1}{M_{r,j}}. \tag{51}$$

This results in

$$K_p = kK_r\frac{M_p}{M_r}. \tag{52}$$

As $K = K_p + K_r$, we have the following:

Result 5. *The number of blocks that need to be allocated to the slices of regular and premium users are*

$$K_p = \frac{Kk\frac{M_p}{M_r}}{1+k\frac{M_p}{M_r}}, \tag{53}$$

$$K_r = \frac{K}{1+k\frac{M_p}{M_r}}. \tag{54}$$

Substituting Eqs.(53) and (54) into Eqs.(47) and (48), respectively, we obtain:

Result 6. *The maximum central values for premium and regular users are*

$$U_p(\theta_p, \epsilon_p) = \frac{Kk}{M_r + kM_p}, \tag{55}$$

$$U_r(\theta_r, \epsilon_r) = \frac{K}{M_r + kM_p}. \tag{56}$$

Result 5 and Result 6 enable the operator to determine the level of service it can offer to users of different slices, as well as to perform resource allocation.

Note: The analysis can be extended to any number of classes. Assume that there are in total $L = |\mathcal{L}|$ classes. In each class, there are $n_l = |\mathcal{N}_l|$ users, with per-block rates R_{l,i_l}, $l \in \mathcal{L}, i_l \in \mathcal{N}_l$. Following a similar procedure as for two classes, for the targeted data rate of users of class l (assuming that users of class 1 are most privileged, whereas those of class L receive the lowest level of service), we obtain

$$U_l = \frac{k_l K}{\sum_{l=1}^{L} k_l M_l}, \quad \forall l \in \mathcal{L}, \tag{57}$$

where k_l is a coefficient denoting how much the targeted rate of class l users is higher than that of the users with the lowest level of service. Similarly as before, $M_l = \sum_{i_l=1}^{|\mathcal{N}_l|} \frac{1}{M_{l,i_l}}$.

As a final note, the analysis we perform here is computationally scalable. Namely, the complexity is $O(Ln)$, where n is the highest number of users across all classes, and L is the number of classes.

7.2 Dynamic policy: Fixed user activity

Similar to Section 7.1, the sizes of the slices dedicated to premium and regular users are fixed. But, within the slice the number of blocks allocated to a user changes over time, depending on the channel conditions of all the users belonging to that slice.

We need to determine the slice sizes for premium and regular users, i.e., K_p and K_r, respectively, the resource allocation policies within the slices, and the central values for premium (U_p) and regular users (U_r), respectively. The setup is identical to the one in Section 7.1. The procedure, for each slice, is similar to the one presented in Section 5.1. The only difference is that in Eq.(24), K and U_c are replaced by K_p and U_p for the corresponding premium users (and their per-block rate probabilities), and by K_r and U_r for regular users. However, there is another difference compared to the single-class users. Namely, while K was known previously, K_p and K_r need to be determined.

The corresponding Eq.(24) would yield the minimum value of $x_p = \left(\frac{K_p}{U_p}\right)_{\min}$ that can be achieved. Similarly, we obtain the minimum value of $x_r = \left(\frac{K_r}{U_r}\right)_{\min}$ for regular users from the corresponding Eq.(24).

Combining previous expressions with $K = K_r + K_p$ and $U_p = kU_r$, we obtain the slice sizes for premium and regular users.

Result 7. *The number of blocks that need to be allocated to the slices of premium and regular users are*

$$K_p = \frac{x_p kK}{kx_p + x_r}, \tag{58}$$

$$K_r = \frac{Kx_r}{kx_p + x_r}. \tag{59}$$

Similarly, we have the other outcome.

Result 8. *The maximum targeted data rates for premium and regular users are*

$$U_p(\theta_p, \epsilon_p) = \frac{kK}{kx_p + x_r}, \tag{60}$$

$$U_r(\theta_r, \epsilon_r) = \frac{K}{kx_p + x_r}. \tag{61}$$

Note that while premium users always receive a proportionally higher target data rate, the same is not the case for the sizes of the slices (i.e., the number of PRBs). For the latter, it depends on the channel statistics of the users from the two slices as well as on the number of corresponding users which of the slices will be larger.

As far as the block allocation policy is concerned, within the slice it is performed in line with Eq.(12), with the adjustment for the slice size and the corresponding statistics of premium and regular users.

7.3 Dynamic user activity

For this scenario, due to the consistently changes in the number of active users, it makes sense to look only at dynamic allocation policies. Resource reservation, i.e., static policy, does not make sense for users that are only intermittently active. The goal here is to decide on splitting the resources, i.e., the size of the corresponding slices for both the premium and regular users. We follow a similar approach as in Section 6. Note that the maximum number of premium users that can be simultaneously active is N_p. For regular users, the maximum number is N_r.

The data rate premium user $i \in \mathcal{N}_P$ experiences in frame t, after receiving a frame ratio of $Y_{p,i}(t)$ from premium slices with K_p blocks, is given by

$$K_p Y_{p,i}(t) R_{p,i}(t) = \frac{K_p I_i(t)}{\sum_{l=1}^{N_p} \frac{I_l(t)}{R_l(t)}}, \tag{62}$$

where again I_i denotes the indicator variable for the activity of user i. With a probability $\pi_{i,active}$ of user i being active, it can be shown that, similar to the previous scenarios, the rate constraint should satisfy

$$\mathbb{P}\left(\sum_{l=1}^{N_p} \frac{I_l(t)}{R_l(t)} \leq \frac{K_p}{(1-\theta)U_p}\right) - \mathbb{P}\left(\sum_{l=1}^{N_p} \frac{I_l(t)}{R_l(t)} < \frac{K_p}{(1+\theta)U_p}\right) \geq (1-\epsilon_p)\pi_{i,active}. \tag{63}$$

As opposed to Eq.(32) where K was known, our unknown variable here is the ratio $x_p = \frac{K_p}{U_p}$. It denotes the inverse of the number of blocks needed to provide rate U_p to premium users. Inequality (63) can be rewritten as (in line with (40))

$$\mathcal{L}_1(x_p) - \mathcal{L}_2(x_p) \geq (1-\epsilon_p)\pi_{i,active}. \tag{64}$$

We are interested at the point where x_p achieves its minimum (or the highest data rate) for which (64) still holds. For user i, we have

$$X_{i,\min} = \min x_{p,i} = \min\left(\frac{K_p}{U_{p,i}}\right), \forall i \in \mathcal{N}_P. \tag{65}$$

However, as all the premium users need to have the same central data rate, we have the following:

Result 9. *The maximum center-of-region data rate that can be guaranteed to premium users is*

$$U_p(\theta_p, \epsilon_p) = \frac{K_p}{\max\{X_{1,\min}, \ldots, X_{N_p,\min}\}}. \tag{66}$$

Following a similar approach with the regular users, we obtain:

Result 10. *The maximum center-of-region data rate that can be guaranteed to regular users is*

$$U_r(\theta_r, \epsilon_r) = \frac{K_r}{\max\{Z_{1,\min}, \ldots, Z_{N_r,\min}\}}, \tag{67}$$

where

$$Z_{j,\min} = \min x_{r,j} = \min\left(\frac{K_r}{U_{r,j}}\right), \forall j \in \mathcal{N}_R. \tag{68}$$

The final step is to determine the slices that need to be allocated to premium and regular users, and from that the corresponding central data rates. Combining Eqs.(66) and (67) into $U_p = kU_r$ and $K_p + K_r = K$, and solving the corresponding system of equations, finally we obtain:

Result 11. *The maximum targeted data rates for premium and regular users are:*

$$U_p = \frac{kK}{\max\{Z_{1,\min}, \ldots, Z_{N_r,\min}\} + k\max\{X_{1,\min}, \ldots, X_{N_p,\min}\}}, \tag{69}$$

$$U_r = \frac{K}{\max\{Z_{1,\min}, \ldots, Z_{N_r,\min}\} + k\max\{X_{1,\min}, \ldots, X_{N_p,\min}\}}. \tag{70}$$

Slice dimensioning is performed at the beginning and is kept fixed over time, despite the fact that the number of both types of users changes over time. Changing the slice dimensions dynamically is more involved as it requires careful consideration of interference issues when reallocating the blocks between slices. We defer this to future work.

As far as the implementation is concerned, having determined the slice sizes for premium and regular users, the process continues in a very similar way with previous scenarios.

8 Performance Evaluation

We first describe the simulation setup. Then, we validate our theoretical results on two traces, one from a 4G and the other from a 5G network. Finally, we illustrate several scenarios that provide some interesting engineering insights, including the outcomes from a practical use case scenario of real-time video streaming. To emphasize the advantages of using our approaches, we compare the results obtained using them against several state-of-the-art schemes [5], [33], [34], [35], [36] from different aspects.

8.1 Simulation setup

In order to corroborate the validity of our theoretical approach for cellular networks in general, as input parameters we use data obtained from real-life traces of both 4G and 5G networks, explained next. It is worth mentioning that as 4G networks have been deployed for a long time, there were multiple traces with 4G-related data publicly available. We chose the one, which contains all the input parameters we need. On the contrary, as 5G is relatively new and still subject of intense deployment, not many traces are publicly available. Nevertheless, we were able to find one, which we use in this paper.

8.1.1 Scenario 1: 4G trace

As input parameters, we have used data from a trace of the received signal characteristics of mobile users in a number of cities across Western Europe and North America. These traces can be found in [37], and their detailed description is provided in [2]. The parameters of interest are the Received Signal Strength Indicator (RSSI) and users' positions, where the latter are expressed in terms of their longitude and latitude. We chose 6 users in Amsterdam, such that their positions are close enough to be served by the same gNodeB. Then, RSSI values of every user over time were mapped to the corresponding SINR values taken from [38]. The number of CQI levels chosen was 15. This means that all SINR values were translated into 15 discrete per-block rates (second row of Table 1), according to the threshold values γ [39] that are shown in the first row of Table 1. For example, if a user's SINR in the current frame is 9 dB (i.e., it is between 8.5 and 10.3 dB), its per-block rate will be 386.1 kbps.

Based on the occurrence frequency of a per-block rate for every user, we obtained the corresponding per-block rate probabilities in Table 1. From the trace, we observed a strong correlation between the received signals of a user in contiguous frames, meaning that the per-block rate of a user does not really change independently from one frame to another. Nevertheless, as will be seen in Section 8.2, our theoretical results closely match simulated (actual) results despite this correlation feature in the data.

There are some notable differences between the channel characteristics of these 6 users. Based on the per-block rate distributions, we can infer that users 2 and 4 are most probably static, as they obtain only three per-block rates (implying that the distance-based component of the signal is unchanged and the change occurs only due to the shadowing component). The other four users are moving around the cell.

8.1.2 Scenario 2: 5G trace

To corroborate the validity of our approach in different generations of cellular technologies, in the second part, for input parameters, we have used a 5G trace with data measured in the Republic of Ireland. These traces can be found in [40], with a detailed description in [41], whose analysis is performed in [15]. The parameter of interest from the trace is CQI with 15 levels, which serves to determine the per-block rate of a user in a frame. These measurements were conducted for one user, but at different days, for different applications, and when the user is static and moving around. To mimic the dynamic nature of these users, we have picked 8 users that were moving around, and assume they are all in the same cell. As before, based on the frequency of occurrence of a per-block rate for every user, we obtained the corresponding per-block rate probabilities, which are depicted in Table 2.

The frame duration is 10 ms.[10] The simulation is run over $100,000$ frames. Unless stated otherwise, for Scenario 1 (4G network), the subcarrier spacing is 15 KHz, with 12 subcarriers per block, making the block width 180 KHz. The total number of PRBs in this case is $K = 100$ [42].

For Scenario 2 (5G network), the subcarrier spacing is 30 KHz, with 12 subcarriers per block, making the block width 360 KHz. The total number of PRBs in this scenario is $K = 273$ [27]. Note that in Scenario 2 the values of the per-block rates are $2\times$ higher than the corresponding Scenario 1 values (see Table 1 and Table 2) because the PRBs have $2\times$ higher bandwidth.

The simulations are conducted in MATLAB R2020b and we take the average of the metrics of interest over 1000 runs.

To introduce diversity in validations and evaluations, but at the same time to avoid being repetitive, we use the trace data from the 4G network for some of the scenarios, and 5G-related data for some others.

8.2 Validations

While deriving the theoretical results we assumed that the per-block rate of a user is independent in two contiguous frames. As mentioned above, the trace results exhibit a strong correlation in the signal quality across frames. Hence, we are interested in looking at how our results fare under realistic circumstances.[11]

In the first case, we validate the results of the static policy for fixed user activity for single-class users (Result 1). There are 6 users (users 1-6) from Scenario 1 (Table 1). All the other parameters are as stated previously in Section 8.1. There are three scenarios in terms of the outage probability: $\epsilon = 0.15$, $\epsilon = 0.25$, and $\epsilon = 0.3$. Finding the simulated center of interval data rate is not straightforward. The procedure is illustrated in the following.

As the data rate for user i is $KY_iR_i(t)$ and for $1-\epsilon$ of the time it must lie in the interval $[(1-\theta)U_c, (1+\theta)U_c]$, this implies that the span is $\frac{1+\theta}{1-\theta}$. Given that K is a constant and Y_i remains unchanged over time for the static policy, we only need to look at all the possible ratios of two per-block rates. There are $\frac{15\cdot 14}{2} = 105$ possible combinations, as we need to consider only ratios that are greater than 1. After finding the corresponding values whose ratio falls within the interval $\left[1, \frac{1+\theta}{1-\theta}\right]$, we look at the probability mass function of every user separately. Namely, if e.g., $[r_k, r_l], 1 \le k \le l \le 15$, is one of the candidate solutions, we need to check the sum of PMF's of per-block rates for all the users in that interval, i.e., $p_i(r_k) + \ldots + p_i(r_l),\ \forall i \in \mathcal{N}$. If the latter is $\ge 1-\epsilon$, then it is a feasible interval for user i. If not, we proceed with the other candidate intervals until we find one which is feasible. This is done for all the users. The next step is determining the value of Y_i. As the values of these intervals are generally different for different users, whereas the value of U_c has to be the same for all users, we need to choose higher values of Y_i for users with smaller values of the center value of the interval $[r_k, r_l]$, where the former for user i is $z_i = \frac{r_k+r_l}{2}$. With the previous discussion in mind, for Y_i we have $Y_i = \frac{\frac{1}{z_i}}{\sum_{j=1}^{n}\frac{1}{z_j}}$, which results in a central value of $U_c = KY_iz_i = \frac{K}{\sum_{j=1}^{n}\frac{1}{z_j}}$.

Fig. 4 illustrates the theoretical vs. simulation results for U_c for the static policy for different values of deviation θ and three different values of outage $\epsilon = \{0.15, 0.25, 0.3\}$. As can be observed from Fig. 4, the simulation results match closely the theoretical results for all ϵ (the level of discrepancy never exceeds 10%), despite the fact that in the theoretical analysis we have the independence assumption of the per-block rate in contiguous frames for every user, whereas in the trace there is a correlation. This shows the practical importance of our model. Another observation to be made is that increasing θ, the value of U_c increases almost linearly, and this conclusion propagates

10. A frame is a concatenation of slots, where the latter represent the actual unit of resource allocation. Nevertheless, our theoretical approach is oblivious to the duration of a slot/frame.

11. As opposed to [1], in which we used the 5G parameters on a 4G trace (in order to mimic to the best possible extent the former), in this work, we assume a more realistic setup, in which for the 4G network we use exactly 4G-standard parameters. Hence, some of the results in this section look different than in [1].

TABLE 1: Per-block rates and the corresponding probabilities for every user from the sampled 4G Amsterdam trace (Scenario 1) [37]

SINR thr. (dB)	-9.5	-6.7	-4.1	-1.8	0.4	2.4	4.5	6.4	8.5	10.3	12.2	14.1	15.8	17.8	19.8
R (kbps)	24	36.8	60.9	96.1	141	189	237.1	356	386.1	437.4	531.9	624.8	724.2	820.3	889.2
$p_{1,k}$	0	0.1	0.72	0.04	0.05	0.09	0	0	0	0	0	0	0	0	0
$p_{2,k}$	0	0	0.2	0.7	0.1	0	0	0	0	0	0	0	0	0	0
$p_{3,k}$	0	0	0	0	0.02	0.12	0.51	0.32	0.01	0.01	0.01	0	0	0	0
$p_{4,k}$	0	0	0	0	0	0.01	0.98	0.01	0	0	0	0	0	0	0
$p_{5,k}$	0.22	0.04	0.07	0.04	0.04	0.06	0.17	0.15	0.01	0.01	0.06	0.06	0	0.03	0.04
$p_{6,k}$	0.17	0.11	0.1	0.07	0.05	0.1	0.17	0.11	0.02	0.04	0	0.03	0	0.02	0.01

TABLE 2: Per-block rates and the corresponding probabilities for every user from the 5G Republic of Ireland trace (Scenario 2) [41]

R (kbps)	48	73.6	121.8	192.2	282	378	474.2	712	772.2	874.8	1063.8	1249.6	1448.4	1640.6	1778.4
$p_{1,k}$	0	0	0	0	0	0	0.01	0.05	0.11	0.13	0.14	0.18	0.06	0.11	0.21
$p_{2,k}$	0	0	0	0	0	0.01	0.02	0.06	0.13	0.14	0.2	0.21	0.07	0.09	0.07
$p_{3,k}$	0.01	0	0	0	0	0.01	0.01	0.02	0.06	0.13	0.17	0.18	0.08	0.18	0.15
$p_{4,k}$	0	0	0	0	0	0.02	0.03	0.13	0.06	0.2	0.32	0.11	0.01	0.09	0.03
$p_{5,k}$	0	0	0	0	0	0	0.04	0.07	0.13	0.17	0.22	0.2	0.05	0.06	0.06
$p_{6,k}$	0	0	0	0	0.01	0.03	0.11	0.12	0.19	0.15	0.15	0.12	0.05	0.04	0.03
$p_{7,k}$	0	0	0	0	0	0	0.04	0.07	0.13	0.17	0.22	0.2	0.05	0.06	0.06
$p_{8,k}$	0	0	0	0	0.01	0.03	0.11	0.12	0.19	0.15	0.15	0.12	0.05	0.04	0.03

across all the considered values of ϵ (see Fig. 4). Finally, relaxing the requirement for the data rate to fall within the feasible interval (increasing ϵ) increases U_c considerably.[12]

After validating the results related to the static policy, we proceed with the validation of the dynamic policy results. To this end, all the parameters remain unchanged from the static-policy scenario. The simulation value for U_c is obtained as follows. As the rate is the same for all the users, we rank the values obtained from Eq.(13) in descending order over all frames. Then, starting from the highest values of data rate we check whether the interval with that (highest) value on one side and the lowest value, that is $\frac{1+\theta}{1-\theta}$ times smaller than the highest value, on the other, contains at least $(1-\epsilon)\cdot 100,000$ values in between. When this is achieved, the simulated value of the center-of-interval value U_c is taken as the mean of the highest and lowest values of that same interval.

Fig. 5 shows the theoretical (obtained from Result 2) vs. simulation results for the maximum achievable value of U_c for different values of θ and ϵ for the dynamic policy (Scenario 1, single-class users, fixed user activity). Similar conclusions hold as in the scenario of Fig. 4. Most importantly, our theoretical result fares pretty well in realistic scenarios (the level of mismatch with the simulation result is less than 11%). The second thing to observe, comparing Fig. 5 with Fig. 4, is that the dynamic policy outperforms the static policy in terms of U_c by about 10%. The rationale behind this is that the dynamic policy takes into account the channel characteristics of all the users in the frame and reacts accordingly in making the decision on allocating resources. There is a tradeoff involved between the complexity and the data rates when deciding between the static and dynamic policy. Implementing the resource allocation scheme for the static policy is far less complex but provides lower U_c. If the goal is to increase the target data rate U_c without considering the computational cost involved, then the dynamic policy is the proper choice.

Having validated our theoretical results for single-class permanently-active users in Figs. 4 and 5, we proceed with validating the results for users that are only intermittently active, and for users with different levels of service (i.e., two classes of users). To that end, we use the input parameters pertaining to Table 2 (Scenario 2). In the first case, where we have intermittent users, all the eight users belong to the same class. The corresponding α_i parameters (see Section 6) for these users are $[0.8\ \ 0.95\ \ 0.7\ \ 0.2\ \ 0.8\ \ 0.87\ \ 0.9\ \ 0.1]$, whereas the corresponding β_i parameters are $[0.87\ \ 0.95\ \ 0.3\ \ 0.8\ \ 0.2\ \ 0.8\ \ 0.9\ \ 0.9]$, resulting in the users' probabilities for being active ($\pi_{i,active}$) equal to $[0.4\ \ 0.5\ \ 0.7\ \ 0.2\ \ 0.8\ \ 0.6\ \ 0.5\ \ 0.1]$. Fig. 6 shows the actual and theoretical results for three different values of $\epsilon = \{0.1, 0.15, 0.2\}$. Similar conclusions follow as in the scenarios from Figs. 4 and 5. The difference is that the target rates here are considerably higher than previously. There are several reasons for that. The results in Fig. 6 are for a 5G network, with a $2\times$ higher per-block rates, more PRBs (273 vs. 100), users are not active at all times, which means that those users which are active will have more resources. Yet another reason is that, comparing the per-block rates between the two scenarios, it can be observed that the Republic of Ireland users have much better channel conditions (more often higher CQIs).

Next, we validate the results of maximum target rates for two classes of users. In these cases, the premium users should receive two times higher data rates than regular ones ($k = 2$). The first four users from Table 2 are premium, and the last four are regular. Fig. 7 illustrates the results for the fixed user activity with the static policy, whereas Fig. 8 does that for the dynamic policy. The outage is $\epsilon = 0.1$. In all the cases, the theoretical predictions fare pretty well (the level of mismatch at most around 10%) despite the fact that in our theoretical approach we were assuming that the per-block rate from one frame to another changes randomly, while in the traces we observed a strong correlation in the values of CQI between contiguous frames for a user. Observing Figs. 7 and 8, we can infer that the dynamic policy yields better performance by around 10% compared to the static policy.

8.3 Performance comparisons

Having validated the accuracy of our theoretical results, we proceed with comparing the performance of the static and dynamic

12. The extreme value of $\theta = 0.5$ implies a very tolerant data rate, and hence not that constrained rate-variability. The operator would constrain itself to choosing lower values of θ. Nevertheless, we consider even these high values here to show the actual dependency of certain parameters as we relax the deviation ratio θ.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3303097

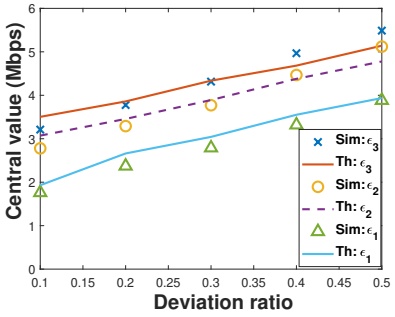

Fig. 4: Theoretical vs. simulated central values U_c with the static policy for $\epsilon = \{0.15, 0.25, 0.3\}$ in Sc.1 (single-class users).

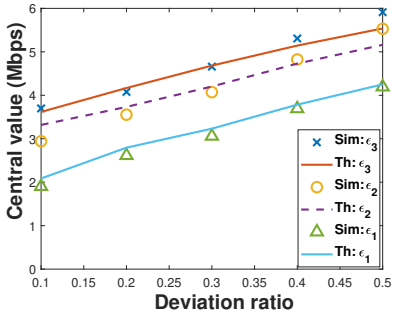

Fig. 5: Theoretical vs. simulated central values U_c with the dynamic policy for $\epsilon = \{0.15, 0.25, 0.3\}$ in Sc.1 (single-class users).

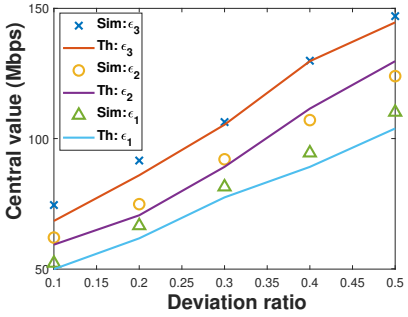

Fig. 6: Theoretical vs. simulated central values U_c for single-class users with dynamic activity and $\epsilon = \{0.15, 0.25, 0.3\}$.

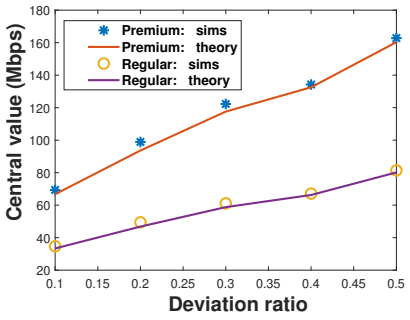

Fig. 7: Theoretical vs. sim. U_c for two classes of users for fixed user activity ($\epsilon = 0.1$), with static policies.

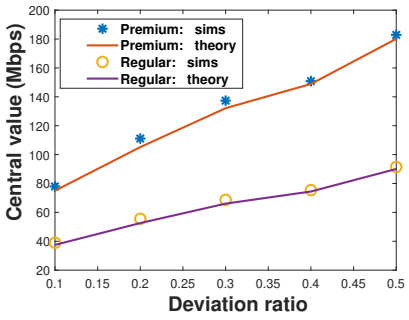

Fig. 8: Theoretical vs. sim. U_c for two classes of users for fixed user activity ($\epsilon = 0.1$), with dynamic policies.

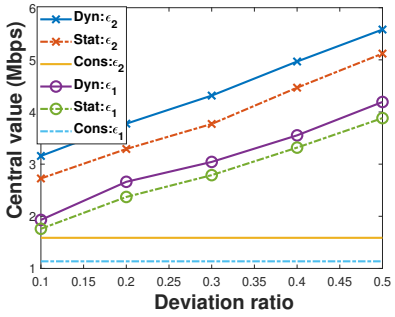

Fig. 9: Comparing static and dynamic (single-class, Sc.1) with consistent-rate policy [5], and $\epsilon = \{0.15, 0.25\}$.

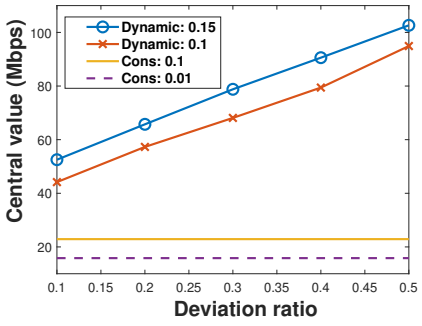

Fig. 10: Comparing the dynamic ($\epsilon = \{0.1, 0.15\}$) with consistent-rate policy [5] ($\epsilon = \{0.01, 0.1\}$) for dynamic user activity (Sc.2).

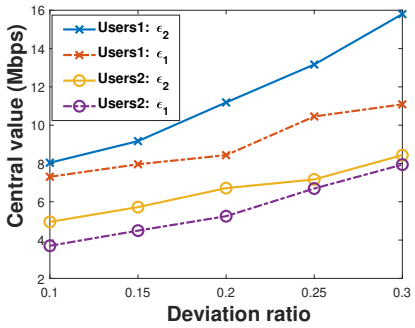

Fig. 11: Impact of SINR variability on U_c for type-1 and type-2 users (dynamic policy) in 5G for $\epsilon = \{0.1, 0.2\}$.

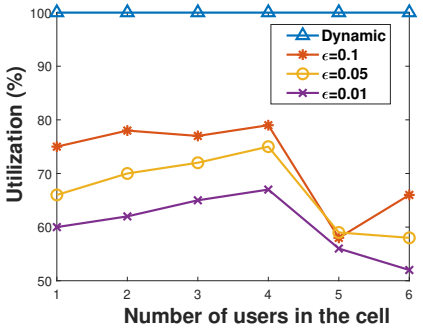

Fig. 12: Resource utilization for consistent policy [5] with $\epsilon = \{0.01, 0.05, 0.1\}$ vs. dynamic policy (single-class users, Sc.1).

policies with that of the strict consistent rate policy [5], in which the user has to receive the same data rate almost at all times. We do this for two different values of outage probability, $\epsilon = \{0.15, 0.25\}$, for single-class users with fixed activity from Scenario 1. Fig. 9 illustrates the maximum achievable values of the central rate vs. θ for the three aforementioned policies. As can be seen from Fig. 9, both the static and dynamic policy largely outperform the strict consistent-rate policy. The gain ranges from 60% to 250%, which is considerable. Furthermore, by allowing a wider feasible interval, the performance (in terms of data rate) improves linearly with the increase in deviation ratio.

Similar conclusions follow for the users that are not active at all times (single class of users), belonging to Scenario 2 (Fig. 10). Specifically, we can observe the linear increase with the deviation ratio and the improvements when relaxing the requirement for the rate to be within the feasible interval. Data rates are much higher than the consistent rates [5], where for the latter there are two scenarios in terms of the probability of outage, 0.01 and 0.1.

Next, we look at the impact of the SINR variability (expressed through the variability of the per-block rate) on the value of U_c. In this scenario, there are two types of users (in terms of their rate variability). The first, denoted as users of type 1, can take one of the two per-block rates: r_3 and r_{13} (whose values are shown in Table 2), with corresponding probabilities $p_{r_3} = 0.79$ and $p_{r_{13}} = 0.21$, respectively. The second type of users (denoted as users of type 2) have possible per-block rates of r_2 and r_{15}, with probabilities $p_{r_2} = 0.8$ and $p_{r_{15}} = 0.2$, respectively. We consider the 5G network setup and the users are always active (fixed user activity). The other parameters remain unchanged, including the number of users, which is 6 in both groups. The average per-block rate in both cases is the same (≈ 0.4 Mbps), but the variability is higher in the second case (channel conditions highly variable): $c_{v1} = 1.35$, $c_{v2} = 1.64$.[13] As the dynamic policy provides superior performance, due to space limitations, we only show results related to that policy.

13. The coefficient of variation of the random variable X is defined as $c_v = \frac{\sqrt{\text{Var}(X)}}{\mathbb{E}[X]}$.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3303097

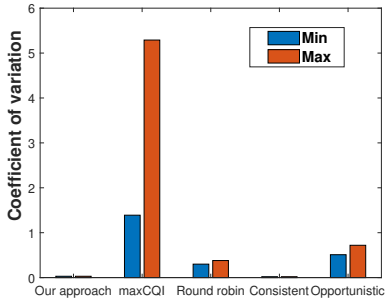

Fig. 13: Maximum and minimum values of coefficient of variation (across all users) with five different policies (single-class always-active users).

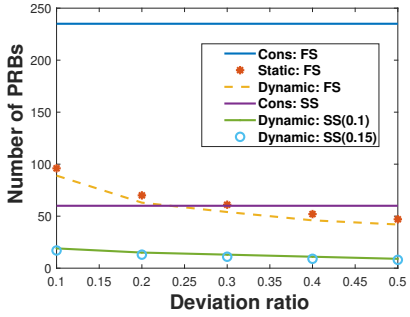

Fig. 14: Number of PRBs required to provide 5 Mbps with different policies for Sc.1 (FS) and Sc.2 (SS). For Static (FS) and Dynamic (FS), $\epsilon = 0.1$.

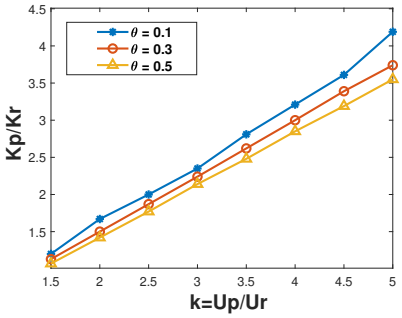

Fig. 15: The ratio between slice sizes for premium and regular users for different k, for Sc.2 and $\theta = \{0.1, 0.3, 0.5\}$.

Fig. 11 shows the maximum achievable U_c vs. θ for two outage probabilities $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.2$, for the two types of users. The SINR variability greatly affects the value of U_c. For instance, in the scenario with lower per-block rate variability (users of type 1), U_c is between 50 and 90% higher than in the second scenario, even though their average per-block rates are the same. The same conclusion holds for any value of ϵ and is almost insensitive to θ.

We proceed with comparing the efficiency in terms of resource utilization when using our approach against the efficiency when providing the consistent rate [5]. Regarding the latter, we consider three scenarios related to the outage of providing a consistent rate: $\epsilon_1 = 0.01$, $\epsilon_2 = 0.05$, and $\epsilon_3 = 0.1$. Note that in this approach ϵ defines the ratio of time during which the requirement for the *strict* constant (consistent) rate is not fulfilled. For each of these values of the outage, we run the simulation for different number of users. The input data are from Scenario 1 and we deal with fixed user activity.

Fig. 12 shows the corresponding values of the average utilization level of the network resources for the strict consistent-rate approach [5] for the three aforementioned probability of outage values and different number of users. The x-axis values have the following meaning: The value "1" means that only user 1 is in the cell, "2" corresponds to users 1 and 2 being considered, ..., "6" means that all the six users are taken into account. From Fig. 12 it can be observed that *the looser the consistency requirement, the higher the utilization is*. For example, for the same number of users, an outage probability of 0.1 leads to $10 - 15\%$ more utilization of the resources. Nevertheless, even for the relatively loose consistent requirement of $\epsilon = 0.1$ the utilization of the resources is much lower than in any of our two approaches (either static or dynamic), which is 100%. The other outcome from Fig. 12 is that for a given outage probability, the utilization does not necessarily increase with the number of users. This is especially emphasized when there are users 1-5 in the cell. This is because when there are more users their guaranteed consistent rate is lower. If one of the users has better channel conditions (like user 5), the amount of resources needed to provide the consistent rate is lower. Hence, the total utilization level can decrease, in certain cases, when the new user is added.

In the next scenario, we look at the span of data rates that can be achieved using our policies against an equal-share policy [34] of all the available resources between users (no guarantee on the data rate). We use the same users from Table 1, i.e., Scenario 1 single-class fixed user activity. Table 3 depicts the span of data rates for every user with the equal-share policy, and with the static and dynamic policies for $\theta = 0.1$. The equal-share policy for all

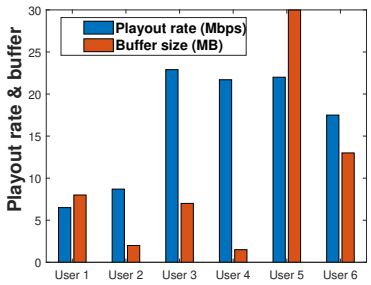

Fig. 16: The achievable playout rate and the required buffer size with the proportionally-fair policy [36] (single-class users).

TABLE 3: The span (ratio of maximum and minimum) of data rates for the static and dynamic policies with $\theta = 0.1$, and for the equal-share policy [34]

	Span	
	Equal share	Static & Dynamic
User 1	5.13	1.21
User 2	2.32	1.21
User 3	3.77	1.21
User 4	1.62	1.21
User 5	37	1.21
User 6	37	1.21

the users results in a higher span, which is considerably higher than $\frac{1+\theta}{1-\theta}$, which in this case is 1.21. We also observed that the total average data rate (over all users) with the static policy is 30% higher, whereas with the dynamic policy it is 40% higher than when using the equal-share policy. This represents a significant gain both in terms of throughput and its stability.

We proceed next with comparing the performance when using our approach with state of the art from two different viewpoints, in terms of the *coefficient of variation of data rate* and the *number of PRBs* required to guarantee a data rate. See the next paragraph for the latter. Users belong to Scenario 2, and are single-class always-active. First, we look at the coefficient of variation of the data rates of users. Fig. 13 shows the results for the following five policies: our dynamic policy, maxCQI [33], Round-robin [34] (known also as equal-share), consistent-rate policy [5], and opportunistic policy [35] (in which the amount of allocated resources is proportional to the channel conditions of the user). For each case, we show the lowest and the highest corresponding c_V among all the users. As can be observed from Fig. 13, using maxCQI, Round-robin, and opportunistic scheduling leads to higher variability in the data rates. For maxCQI, there is a huge difference in the c_v within the users as well. The reason is that a user with this policy

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3303097

will either have a data rate of 0 or will obtain all the resources (if she has the best channel conditions in that frame), leading to a very high data rate. As opposed to these three policies, our approach provides only slightly higher variability in the data rate compared to the consistent-rate policy (0.05 vs. 0.03), which given the gains in the magnitude of data rate our approach offers (see Fig. 9 and Fig. 10), can be neglected.

In the next scenario, we compare the number of blocks (PRBs) needed to guarantee the data rate of 5 Mbps with different approaches. Fig. 14 depicts the outcomes. The first three results pertain to users from Scenario 1 (marked with FS in the legend bar), where all of them are always active. For consistent users, whose outage probability in FS is 0.01, the number of blocks required is extremely high (243), whereas with the static policy, this number drops below 100, and even further with the dynamic policy. The other three results are related to Scenario 2 (marked with SS in the legend bar). The number of required blocks is then much lower due to the intermittent nature of user activity. With the consistent-rate approach [5], where the outage is 0.1, the number of required blocks is around 60, which plummets with the dynamic policies. This shows the significant advantages our approach offers in resource savings as well. These resources can then be used by users of other use cases.

Finally, we look into the sizes of the slices for premium and regular users as a function of the ratios of the corresponding rates, i.e., k. To that end, we consider the users of Scenario 2, where users 1-4 are premium, whereas users 5-8 are regular. Fig. 15 depicts the ratio of slice sizes for the two classes of users, when the total number of available PRBs is 273, as a function of the ratio of achieved targeted data rates for premium and regular users, for different values of deviation ratio. The first interesting thing to observe is the lower than 1 slope increase in the ratio of $\frac{K_p}{K_r}$ with k. Essentially, this means that as we want to provide better and better performance to premium users, we need fewer and fewer extra resources. The second interesting observation is that as we relax the constraint on the rate variability, this ratio increases slower. For instance, for $k = 2$ and $\theta = 0.5$, the premium slice should be $1.5\times$ larger, whereas for the same deviation ratio, but with $k = 5$, the premium slice should be $3.5\times$ larger. In the latter case, if the deviation ratio is 0.1, the premium slice has to be $4.2\times$ larger.

8.4 Practical scenario

To illustrate the practical applicability of our approach, we focus on the use case of real-time video streaming. We consider the performance of the six users described above (users from Table 1 single-class, but with 5G network parameters, i.e., 30 KHz sub-carrier spacing and $K = 273$) when streaming videos online. It is well known that for video streaming a high QoE for mobile users is achieved with a high video resolution (high playout rate (bitrate), i.e., the rate at which the video is being rendered on the smartphone) and a consistent resolution (stable playout rate). Therefore, the goal is to provide a high, fixed playout rate. To control the latency when streaming live video, the playout buffer size at the receiver should not be large. This buffer can be kept small if the variation in data rate (the rate at which packets are received from the network) is kept low. We look at the size of the buffer that is required for each user with three policies, such that both the packet drop rate (information loss) and frequency of rebuffering events (video stalling) do not exceed 5%. In the following, we compare the performance when using our dynamic policy, consistent-rate [5] and proportionally-fair [36] policies.

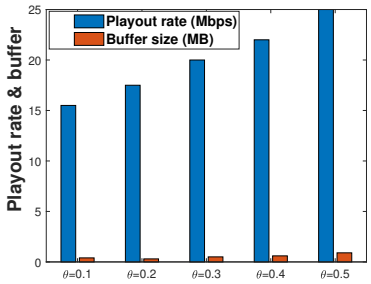

Fig. 17: The achievable playout rate and the required buffer size with the dynamic policy (single-class users).

1) Consistent Rate [5]: With the same input parameters as in the previous scenarios, the maximum consistent rate that can be provided 95% of the time is 5.14 Mbps. Since the network throughput is almost always constant, no buffering is needed, and the constant playout rate that can be provided to the users is the same as the data rate, i.e., 5.14 Mbps. The probability of rebuffering is the same as the probability of outage (0.05).

2) Proportionally-fair policy [36]: Using the proportionally-fair policy, which under special conditions in a single cell can be equivalent to equal-share [6], Fig. 16 shows the playout rates that can be guaranteed along with the required buffer size to provide a packet drop rate and frequency of rebuffering events lower than 0.05. As can be observed, while relatively high playout rates can be provided to users with good channel conditions, the required buffer sizes to support the performance are quite high. The size depends on the variability of the channel conditions. For example, for user 5, a buffer size as large as 30 MB is required, which would imply a high playout delay.

3) Dynamic Policy: As the dynamic policy outperforms the static, we consider only the former due to space limitations. Using our dynamic policy, all the users receive the same data rate. We consider five scenarios in terms of the allowed deviation ratio: $\theta = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. A deviation of 0.1 means that the data rate received is in the range $[0.9U_c, 1.1U_c]$, where U_c is the targeted data rate (central value). Following this policy, *the playout rates provided are equal to the corresponding central value*. The buffer is used to amortize the variability in the data rates. The playout rates obtained, and the corresponding sizes of the buffers are shown in Fig. 17. Obviously, the buffer sizes are very small (less than 1 MB), corresponding to an extra delay in playing out the video on the order of a second. This is very important as for live video streaming a small delay in playing out the video is required. The playout rates with the dynamic policy are much higher ($3 - 5\times$) than with the consistent-rate policy.

The take-away message from this simulation is that when using the consistent-rate policy the video must be played out at a lower resolution, hence is of lower quality. For instance, using the H.265 codec via the consistent-rate policy we can play the video in FHD resolution (1920×1080). On the other hand, when following the dynamic policy the same video can be played out with a considerably higher quality with very small buffer size (less than 1 MB). The dynamic policy supports playout in UHD resolution (3840×2160) [43], for which the required playout rate at H.265 is 12 Mbps, in 4K (4096×2160), or even in Netflix's UHD for which the required rate is 25 Mbps [44]. This shows the concrete advantage our approach offers in practice for real-time video streaming.

To summarize, relaxing the strict constant-rate requirement

results in a considerable improvement in the efficiency of resource allocation and in the achievable data rates, while providing much lower throughput variability than when network resources are split equally among all the users in the cell.

9 Conclusion

In this paper, we considered the problem of controlling the data rate variability of users in cellular networks within reasonable bounds while not wasting any network resources. We did this for three scenarios: (i) when users are always active, (ii) when users are not always active, and (iii) users belong to different service classes. Whenever feasible, we proposed two policies for resource allocation, one which is static and the other which depends on the channel conditions of all the users over time. We showed that allowing a slight increase in the width of the feasible interval leads to an almost linear increase in central data rates. The advantage of our approach is that the operator can decide between static and dynamic policies depending on whether it is more interested in reducing complexity or improving performance. If the former is the case, the static policy is the right choice, whereas if the goal is to improve the performance by all means, the dynamic policy should be chosen. We validated our theoretical results with extensive realistic simulations, which were run on data obtained from real traces, and compared our results against state of the art, showing the significant advantages our approach offers.

As part of our future work, we plan to consider the problem of providing α-fair resource allocation in terms of data rate with constrained rate variability in the cell, for users belonging to the same service level, as well as dimensioning slices so that users that have the same SLA, within the slice, obtain resources in such a way that there are fairness guarantees. We also plan to consider the discrepancy introduced when having inaccurate CQI distributions.

References

[1] F. Mehmeti and T. L. Porta, "Efficient resource allocation with constrained rate variability in cellular networks," in *Proc. of ACM Q2SWinet*, 2021.

[2] B. Meixner, J. W. Kleinrouweler, and P. Cesar, "4G/LTE channel quality reference signal trace data set," in *Proc. of ACM MM*, 2018.

[3] "5G radio access." www.ericsson.com/res/docs/whitepapers/wp-5g.pdf, 2016. Ericsson white paper, Uen 284 23-3204 C.

[4] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, 2018.

[5] F. Mehmeti and C. Rosenberg, "How expensive is consistency? Performance analysis of consistent rate provisioning to mobile users in cellular networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 5, 2019.

[6] F. Mehmeti and T. L. Porta, "Reducing the cost of consistency: Performance improvements in next generation cellular networks with optimal resource reallocation," *IEEE Tran. on Mobile Computing*, vol. To appear, 2021.

[7] A. Goldsmith, *Wireless communications*. Cambridge University Press, 2005.

[8] A. Gupta and J. R. Kumar, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, 2015.

[9] R. Trivisonno, R. Guerzoni, I. Vaishnavi, and D. Soldani, "Towards zero latency software defined 5G networks," in *Proc. of IEEE ICCW*, 2015.

[10] M. Erel-Özçevik and B. Canberk, "Road to 5G reduced-latency: A software defined handover model for eMBB services," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, 2019.

[11] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive massive access management for QoS guarantees in M2M communications," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, 2015.

[12] Y. Han, S. E. Elayoubi, A. Galindo-Serrano, V. S. Varma, and M. Messai, "Periodic radio resource allocation to meet latency and reliability requirements in 5G networks," in *Proc. of IEEE VTC*, 2018.

[13] Y. Qi, M. Hunukumbure, M. Nekovee, J. Lorca, and V. Sgardoni, "Quantifying data rate and bandwidth requirements for immersive 5G experience," in *Proc. of IEEE ICC Workshop on 5G RAN Design*, 2016.

[14] Y. Zhang, Arvidsson, M. Siekkinen, and G. Urvoy-Keller, "Understanding HTTP flow rates in cellular networks," in *Proc. of IFIP Networking Conference*, 2014.

[15] F. Mehmeti and T. L. Porta, "Analyzing a 5G dataset and modeling metrics of interest," in *Proc. of IEEE MSN*, 2021.

[16] H. Du, Q. Zheng, W. Zhang, and X. Gao, "A bandwidth variation pattern-differentiated rate adaptation for HTTP adaptive streaming over an LTE cellular network," *IEEE Access*, vol. 6, 2018.

[17] E. A. Walelgne, J. Manner, V. Bajpai, and J. Ott, "Analyzing throughput and stability in cellular networks," in *Proc. of IEEE/IFIP NOMS*, 2018.

[18] E. Šlapak, J. Gazda, W. Guo, T. Maksymyuk, and M. Dohler, "Cost-effective resource allocation for multitier mobile edge computing in 5g mobile networks," *IEEE Access*, vol. 9, 2021.

[19] H.-T. Chien, Y.-D. Lin, C.-L. Lai, and C.-T. Wang, "End-to-end slicing with optimized communication and computing resource allocation in multi-tenant 5g systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, 2020.

[20] S. K. Goudos, P. D. Diamantoulakis, and G. K. Karagiannidis, "Multi-objective optimization in 5g wireless networks with massive mimo," *IEEE Communications Letters*, vol. 22, no. 11, 2018.

[21] T. Lagkas, D. Klonidis, P. Sarigiannidis, and I. Tomkos, "Optimized joint allocation of radio, optical, and mec resources for the 5g and beyond fronthaul," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, 2021.

[22] J. Khan and L. Jacob, "Resource allocation for comp enabled urllc in 5g c-ran architecture," *IEEE Systems Journal*, vol. 15, no. 4, 2021.

[23] F. Mehmeti and T. L. Porta, "Admission control for consistent users in next generation cellular networks," in *Proc. of IEEE ICC*, 2019.

[24] F. Mehmeti and T. L. Porta, "Optimizing 5G performance by reallocating unused resources," in *Proc. of IEEE ICCCN*, 2019.

[25] F. Mehmeti and C. Rosenberg, "Providing consistent rates for backhauling of mobile base stations in public urban transportation," in *Proc. of IEEE ICC*, 2017.

[26] G. Ku and J. M. Walsh, "Resource allocation and link adaptation in LTE and LTE Advanced: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, 2015.

[27] ETSI, "5G NR overall description: 3GPP TS 38.300 version 15.3.1 release 15." www.etsi.org, 2018. Technical specification.

[28] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Communications Magazine*, vol. 57, no. 1, 2019.

[29] S. M. Ross, *Stochastic Processes*. John Wiley & Sons, 1996.

[30] A. Oppenheim and A. Willsky, *Signals and systems*. Prentice Hall, 1996.

[31] https://www.slashgear.com/att-5g-plans-speed-not-data-randall-stephenson-q119-24574643/.

[32] https://www.theverge.com/2019/4/24/18514518/att-ceo-5g-tiered-data-plans-randall-stephenson.

[33] N. Sharma, S. Zhang, S. R. Somayajula Venkata, F. Malandra, N. Mastronarde, and J. Chakareski, "Deep reinforcement learning for delay-sensitive LTE downlink scheduling," in *Proc. of IEEE PIMRC*, 2020.

[34] O. Grøndalen, A. Zanella, K. Mahmood, M. Carpin, J. Rasool, and O. N. Østerbø, "Scheduling policies in time and frequency domains for lte downlink channel: A performance comparison," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, 2017.

[35] R. Agrawal, A. Bedekar, R. J. La, and V. Subramanian, "Class and channel condition based weighted proportional fair scheduler," *Teletraffic Engineering in the Internet Era*, vol. 4, 2001.

[36] H. Gao, J. S. Bawa, and R. Paranjape, "An evaluation of the proportional fair scheduler in a physically deployed lte-a network," in *Proc. of IEEE ANTS*, 2019.

[37] https://zenodo.org/record/1220256#.XiZbo8hKhPY.

[38] B. Cells, "Baicells technical training." https://baicells.zendesk.com/hc/en-us/articles/115003137453-WISPAPALOOZA-2017-Technical-Training-Slides, 2017. Tech report.

[39] A. Samuylov, D. Moltchanov, R. Kovalchukov, R. Pirmagomedov, Y. Gaidamaka, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Characterizing resource allocation trade-offs in 5G NR serving multicast and unicast traffic," *IEEE Tran. on Wireless Comm.*, vol. 19, no. 5, 2020.

[40] https://github.com/uccmisl/5Gdataset.

[41] D. Raca, D. Leahy, C. J. Sreenan, and J. J. Quinlan, "Beyond throughput, the next generation: A 5G dataset with channel and context metrics," in *Proc. of ACM MMSys*, 2020.

[42] ETSI, "3GPP release 10." www.etsi.org, 2011. Technical specification.

[43] https://www.synopi.com/bandwidth-required-for-hd-fhd-4k-video/.

[44] https://help.netflix.com/en/node/306.

Fidan Mehmeti received the graduate degree in Electrical and Computer Engineering from the University of Prishtina, Kosovo, in 2009. He obtained his PhD degree in 2015 at Institute Eurecom/Telecom ParisTech, France. After that, he was a Post-doctoral Scholar at the University of Waterloo, Canada, North Carolina State University and Penn State University, USA. He is now working as a Senior Researcher and Lecturer at the Technical University of Munich, Germany. His research interests lie within the broad area of wireless networks, with an emphasis on performance modeling, analysis and optimization.

Thomas F. La Porta is the Director of the School of Electrical Engineering and Computer Science at Penn State University. He is an Evan Pugh Professor and the William E. Leonhard Chair Professor in the Computer Science and Engineering Department and the Electrical Engineering Department. He received his B.S.E.E. and M.S.E.E. degrees from The Cooper Union, New York, NY, and his Ph.D. degree in Electrical Engineering from Columbia University, New York, NY. He joined Penn State in 2002. He was the founding Director of the Institute of Networking and Security Research at Penn State. Prior to joining Penn State, Dr. La Porta was with Bell Laboratories for 17 years. He was the Director of the Mobile Networking Research Department in Bell Laboratories, Lucent Technologies where he led various projects in wireless and mobile networking. He is an IEEE Fellow, Bell Labs Fellow, and received the Bell Labs Distinguished Technical Staff Award. He also won two Thomas Alva Edison Patent Awards. Dr. La Porta was the founding Editor-in-Chief of the IEEE Transactions on Mobile Computing. He has published numerous papers and holds 39 patents.

Wolfgang Kellerer (M'96, SM'11) is a Full Professor with the Technical University of Munich (TUM), Germany, heading the Chair of Communication Networks at the School of Computation, Information and Technology. He received his Ph.D. degree in Electrical Engineering from the same university in 2002. He was a visiting researcher at the Information Systems Laboratory of Stanford University, CA, US, in 2001. Prior to joining TUM, Wolfgang Kellerer pursued an industrial careerwas being for over ten years with NTT DOCOMO's European Research Laboratories. He was the director of the infrastructure research department, where he led various projects for wireless communication and mobile networking contributing to research and standardization of LTE-A and 5G technologies. In 2015, he has been awarded with an ERC Consolidator Grant from the European Commission for his research on flexibility in communication networks. He currently serves as an associate editor for IEEE Transactions on Network and Service Management and as the area editor for network virtualization for IEEE Communications Surveys and Tutorials.