Minimizing Rate Variability with Effective Resource Utilization in Cellular Networks

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

Abstract—While one of the main features of 5G networks is provisioning very high rates with low (or no) variability to cellular users, it has been shown that this turns out to be very ineffective for operators because it leads to an abundance of unused network resources. Yet, reallocating the unused resources to the same users, after providing them with the same constant rate, increases back the variability in data rates. A more efficient way would be to provide different low-variability data rates to the users depending on their channel conditions while trying to bring the wasted resources to the lowest possible extent. To that end, in this paper, two approaches are considered; one with reserved resources for every user and the other where the amount of resources is decided on the fly, depending on their current channel conditions. Then, for each approach, we look at different allocation policies and derive the corresponding maximum achievable constant rate for every user jointly with the level of resource utilization, showing which policy is more beneficial. Further, the performance is evaluated on a real 5G trace using both extensive simulations and real measurements conducted on OpenAirInterface. Results show that no-resource reservation policies increase the utilization of resources and data rates at the expense of increased rate variability across all the users. Moreover, all the policies proposed in this paper outperform state-of-the-art approaches by at least $2\times$, bringing the waste of resources down to 15%.

Index Terms—5G, Resource allocation, Optimal performance, Rate variability, QoE.

1 Introduction

5G networks have emerged as the solution that renders the best performance for applications and services that require constant or very low-variability data rates [2], such as real-time video streaming, online gaming, augmented reality, and virtual reality, among other features like providing ultra-low latency with extremely high reliability [3] or providing service to a large number of devices within a given area [4].

Providing stable data rates in current wireless communications systems is very challenging, mainly because of the fact that cellular networks are characterized by very dynamic channel conditions due to the mobility of the users and effects like shadowing [5]. Consequently, a different amount of resources needs to be assigned to users at different times to satisfy the rate stability requirement. Therefore, it is very important to determine the appropriate resource allocation policy that maintains the communication quality of the users irrespective of the application/service they are running while utilizing network resources efficiently.

While providing the *same* constant data rate has been considered before [6], it has also been shown that this approach leads to very inefficient usage of network resources, leaving an abundance of resources unused [7]. Furthermore, reallocating those unused resources to the same users increases back the variability of the experienced data rates, which harms the applications with stable rate requirements. On the other hand, allocating all the available resources without any guarantees on providing a stable throughput [8] leads to cellular users running applications with low-variability rate requirements to experience severe Quality of Experience (QoE) deterioration.

To reconcile the need for stable throughput (low-variability data rates) with effective utilization of network resources (not waste large portions of the resources), in this paper, we propose an approach in which users still experience constant rates for most of the time, but these rates differ among different users, depending on their channel conditions.

Several important research questions arise with regard to the provisioning of low-variability data rates with effective utilization of resources in cellular networks:

- How to achieve the optimal trade-off between providing as high constant data rates as possible for all the users and not wasting network resources?
- What is the parameter that describes jointly the throughput stability over all the users and the effective utilization of network resources?
- Is it better to reserve the resources for every user or to assign them dynamically over time?
- In an application that requires low-variability data rates, such as real-time video streaming, what is the best policy that provides the highest QoE to cellular users while simultaneously keeping it beneficial for the operator?

In this paper, we address the problem of *effectively allocating network resources* within the cell in 5G networks that results in the highest possible data rates that are characterized by low variability, i.e., they are almost always constant for a user. To that end, two different approaches are considered; one with reserving resources for each user and the other with no resource reservation. Within each approach, different allocation policies are proposed.

F. Mehmeti (fidan.mehmeti@tum.de), A. Papa (arled.papa@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 MobiWac 2021 [1].

This work was supported in part by the Federal Ministry of Education and Research of Germany (BMBF) under the project "6G-Life", with project identification number 16KISK002.

Common to all these resource allocation policies is that they provide different constant rates (for almost 100% of the time) to different users. We derive the achievable data rates for every policy and the level of resource utilization when using each of them. In terms of performance, the focus is first on minimizing rate variability and maximizing resource utilization separately. Then, the joint problem of minimizing the rate variability and efficiently utilizing the network resources is considered. To that end, we introduce a new variable, coined *joint satisfaction efficiency*, which is defined as the ratio of the average utilization of resources and the sum of the coefficient of variation of data rates over all users in the cell across time. Then, it is shown which policy under which conditions and for which objective improves performance. The results provided here are helpful for cellular network operators in appropriately allocating resources so that the QoE of mobile users related to the aforementioned services/applications is maximized, and at the same time will prevent from wasting the valuable network resources. The main message of this paper is that providing different constant rates to users with different channel conditions can result in lower wastage of network resources and in minimizing rate variability compared to other resource allocation policies, irrespective of whether they provide constant rates or not. Further, reserving resources for the users is shown to decrease the rate variability but at the expense of lower utilization of network resources and lower data rates compared to the no-resource reservation policies. While the focus is on 5G when evaluating performance, this work could equally apply to 4G as well, where the latter shares some similarities with 5G in terms of resource structure, but with considerably lower rates.

Specifically, our main contributions are:

- We first propose an approach in which the resources for each user are determined *a priori* based on their channel statistics, and do not change over time.
- We also consider another approach in which resource allocation is flexible and there are no reserved resources, but they are allocated on the fly, depending on the channel conditions of all the users. In both approaches, the achievable data rates for several pertinent policies are derived.
- The performance of these policies are compared analytically in two cases. In the first, we compare separately the variability of data rates across all the users and the total resource utilization in the cell. In the second case, there is a joint minimization of rate variability and the utilization of network resources through the new parameter introduced in this paper, the joint satisfaction efficiency.
- Using both realistic simulations that are run on a 5G trace as well as measurements conducted on our own built system, based on OpenAirInterface [9], it is shown that our policies minimize the coefficient of variation of data rates across all the users compared to other state-of-the-art policies while enabling effective utilization of network resources.

The remainder of this paper is organized as follows. Section 2 presents some related work. The system model and the metrics of interest are introduced in Section 3. Then, in Section 4 the detailed analyses are presented for four resource-reservation policies, one of which is derived as a solution to two optimization problems. This is followed by the analyses for two no-resource reservation policies in Section 5. Some performance evaluation results, including both outcomes from simulations and measurements, are provided in Section 6, together with additional engineering insights. Finally, Section 7 concludes the paper.

2 Related Work

Reference [10] provides a detailed overview of various service requirements in 5G. The authors in [11] focus on providing a latency (a type of delay consistency) as low as possible. However, that work is not concerned with throughput stability nor with the usage of network resources. Additionally, in [12], the goal is to minimize the end-to-end delay. The authors propose an architecture, coined 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. The work in [12] is consistent in the delay sense but does not consider resource allocation to provide stable rates. A work similar to [12] is [13], where the goal is to provide a *consistent delay* for Machine-to-machine (M2M) communications. The analysis in [13] relies on large deviation theory. To meet the latency and reliability constraints in 5G, a periodic radio resource allocation is proposed in [14], 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 [13], which is the case with our work.

In [15], the authors propose a use case of 5G deployment and derive the achievable throughput by combining four traffic types and find the rate distributions that are required for each service type. However, the analysis in [15] is constrained by considering only one user with four applications. On the other hand, the approach followed in this work is valid for any number of users.

The highly variable nature of data rates in cellular networks has already been documented in [16], with a coefficient of variation going as high as 3. Similar conclusions were obtained in [17], where even for static users the data rates were exhibiting highly non-stable properties, with a coefficient of variation around 2. While quantifying the rate variability is certainly useful, neither [16] nor [17] provide insights on how to reduce rate variability, which is what we do in this work. In [18], where the focus is 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 [19] it is shown that constant playout rates outperform by a significant margin the adaptive streaming approaches. Similarly, the lack of throughput stability has been shown in [20] as well. But, there are no allocation policies that prevent this from happening. On the other hand, in our work, several policies are proposed that pertain to various scenarios that provide throughput with low variability and at the same time prevent the network from leaving its resources non-utilized.

The works in which there is a strict requirement on the constant data rate at almost all times for all users (consistent rate) in cellular networks, with the focus on 5G, are [21], [7], [6], [22]. In [21], the focus is on determining the maximum number of consistent users that can be admitted in the cell without violating the QoS. In [22], the problem of providing a consistent backhaul rate to public urban transportation systems is analyzed. The analysis captures the scenario with two bus lines having a different number of vehicles, where within the same line all vehicles experience identical per-PRB rate distributions. The most important outcome from [22] is that on average *about* $2/3$ *of the resources remain unused*. However, in all these works the focus is only on providing consistent rates and not on increasing the utilization of network resources. Yet, in our work, both the minimization of rate variability and the reduction of unused resources are considered.

The problem of determining the maximum consistent rate that can be offered to a group of users is analyzed in [6]. One of the main outcomes from [6], similarly to [22], is that providing a consistent rate to everyone leads to highly inefficient utilization of network resources. As a way of alleviating this problem, in [6] it is proposed 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 because the amount of unused resources changes rapidly from one time step to another. This plummets the satisfaction of the users when running applications with a stable throughput requirement. A similar approach is followed in [7], where after providing the maximum achievable constant rate (the same rate to everyone), the authors propose to reallocate the unused resources to satisfy two different objectives. The first is to maximize the total cell throughput after reallocation, whereas the second is to provide fairness. In addition to these two objectives, in [23] the goal is to allocate the unused resources (after guaranteeing a constant rate) in order to provide max-min fairness. While in these scenarios the resources are fully utilized, there is high variability in the data rate over all the users, and those policies are not suitable for applications and that require rates with low variability. More importantly, in Section 6 it is shown that in a practical scenario, the approaches proposed in this work outperform considerably state of the art.

One of the standard approaches of resource allocation mentioned in the 5G standard is round-robin [24]. Following on that, in [25] the authors propose an equal share of resources not only in the RAN but also for edge resources in computing tasks. However, simply sharing resources does not provide any consistency or guarantee on the data rate, due to the varying channel conditions of mobile users. In Section 6 the advantages of different consistent rates against round-robin are documented.

There are several other works related to resource allocation in 5G [26], [27], [28], [29], [30]. In [26], the authors consider resource allocation in a multi-tier mobile edge computing system. The considered resources are the computational units in the cloud, but not the spectrum resources as is the case with 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 data rate guarantees in [26]. Slice dimensioning is a resource allocation problem too, where the goal is 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 [27], where the three main 5G service types are considered (eMBB, URLLC, and mMTC) in a multi-tenant 5G system. There are considerable differences between our work and [27]. 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 different consistent rates improves network resource utilization. As such, it could be used by the approach in [27] to improve the network efficiency.

5G network optimization with massive MIMO has been considered in [28]. The authors formulate a multi-objective optimization problem, where one of the objectives is to maximize the user's average data rate. However, while maximizing the rate is important, its variability can lead to severe performance degradation, especially for services that require stable throughput. As we show in the case of live streaming in our work, a no-rate-guarantee policy leads to lower quality of experience among users. In [29], the authors consider another aspect of optimal resource allocation. Specifically, the goal is to jointly decide 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 rate stability, which can hurt the performance of applications like live video streaming. In contrast, in our work the focus is on rate-sensitive applications.

In [30], the authors focus on allocating resources for coordinated multipoint in 5G networks. URLLC is the traffic of interest. The objective is to minimize the required bandwidth subject to limited network resources and a maximum latency. But, there are no rate stability guarantees. Conversely, the consistent rate that can be provided to each user is determined with current approach, and these rates differ among users. This provides both rate consistency and effective allocation. More importantly, using our approach and knowing the achievable rate, one can predict quite accurately the transmission delay, and if low enough, it can provide the reliability guarantee for delay-sensitive traffic (for low/moderate traffic).

In [31], the authors consider the joint problem of adaptation of MCS and resource allocation in a cellular network. The objective is to minimize the usage of resource blocks while satisfying the minimum rate requirement for each user. The optimization problem is solved using deep reinforcement learning. As opposed to [31], in our work the focus is to effectively utilize network resources, dedicated to a network slice. Furthermore, with our approach one can determine the minimum rate itself (which is constant for the vast majority of the time). The problem of cross layer allocation in heterogeneous cloud RANs was considered in [32]. The focus is to enhance the spectral and energy efficiency. While the objective in the current work is providing consistent rate, the authors in [32] maintain a maximum-delay requirement. Of different nature is the problem in [33], where the device-to-device communications and their interference with cellular networks are considered. The objective in [33] is to maximize the total sum throughput while guaranteeing a minimum probability of coverage of every device. In our work, there is a different setup, with the focus on cellular networks. Furthermore, the requirement of providing a constant rate at almost all times captures the requirement for coverage probability. In similar spirit to [33] is [34], where the goal is again to maximize the total network throughput, while deciding jointly on the resource allocation and user assignment processes. However, in [34] there is no guarantee on the consistency of data rate. In [35], the optimum resource management concerning two objectives (separately) was the problem of interest. The first goal, similarly to [33] and [34], was to maximize the network sum throughput. Optimizing the 5*th percentile rate* was the second objective. The current work provides closed-form optimal resource allocation policies for the two metrics of interest not considered elsewhere.

An alternative policy, considered in [36], proposed to allocate all the resources to the user with the best channel conditions in a frame. While this could lead to maximizing the network throughput, it will penalize heavily the users with bad channel conditions. In Section 6, it is shown that this policy performs worse than different consistent-rate approaches.

In [37], a different approach is followed. The goal is not to guarantee a constant data rate at all times, but rather a data rate that is within some bounds most of the time. Also, each user in [37] is

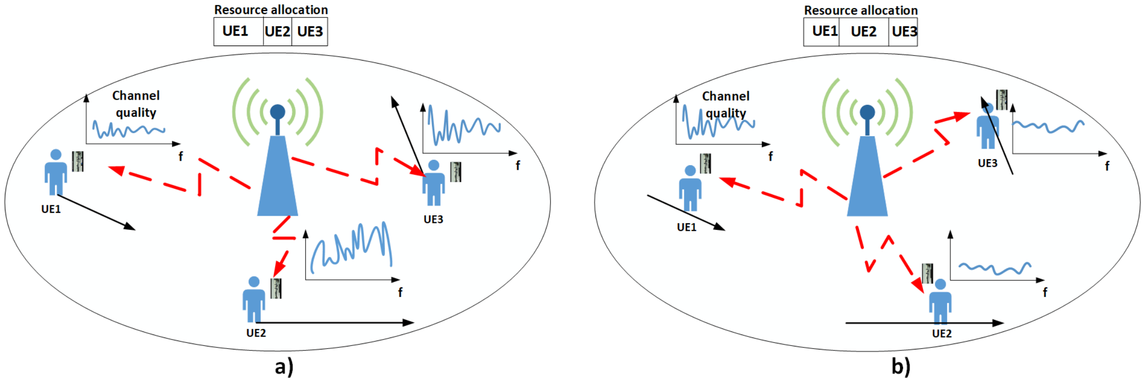


Fig. 1: Users and their channel qualities in different positions at: a) time t_1, b) time $t_2 > t_1$.

expected to receive the same data rate in a frame (which is within the prescribed bounds), as opposed to our approach here in which the users among themselves in general experience different data rates, but each user sees a non-varying data rate almost at all times.

Finally, this work is a considerable extension of [1]. In the current paper, the performance is optimized while considering together the variability of data rates of all the users and the utilization of network resources. Furthermore, we build a system, based on OpenAirInterface [9], which provides various relevant measurement outcomes, resulting further in a very good match with theoretical and simulation results.

3 Performance Modeling

In this section, the system model is introduced first. This is followed by the problem formulation.

3.1 System Model

Cellular users within the coverage area of a 5G macro base station (gNodeB) in the sub-6 GHz band (see Fig. 1) are considered. The focus in this work is the downlink.

As in 4G, the block resource allocation scheme is used in 5G as well, with *physical resource blocks (PRB)* being the allocation unit [8], 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*. The total number of available PRBs within the cell covered by a BS is assumed to be K.

All users send periodically to the BS the data on their channel conditions [24]. This parameter is known as Channel Quality Indicator (CQI). This value ranges from 1 (very poor channel conditions) to 15 (excellent channel conditions).

In general, a user experiences different channel conditions (different CQIs) at different frequencies (PRBs) even within the same radio frame, and hence every user has a different per-PRB CQI, which is a function of Signal-to-Interference-plus-Noise Ratio (SINR). The latter is a function of the transmission power of the BS from which the user received service, the transmission power of neighboring cells gNodeB's transmitting on the same frequencies (inter-cell interference), Additive White Gaussian Noise (AWGN), and the corresponding channel gains [5], [6]. Due to the user's mobility and time-varying channel characteristics, per-PRB SINR changes from one frame to another (according to some distribution) even for the same PRB. This value of per-PRB CQI, depending on the MCS used, sets the per-PRB rate. In our system, the MCS is with 15 possible values, which is the typical value encountered in practice [24], and is the same as the number of possible CQIs, mentioned above. For example, if at time t the per-PRB SINR lies within the interval $[\gamma_j, \gamma_{j+1}]$, with γ_j and γ_{j+1} being the thresholds of the MCS ($j = 1, \ldots, 15$), the per-PRB rate at that time would be $r_j(t)$ [38].

Further, for every user,"flat PRBs" are assumed only during a frame (flat fading), i.e., the per-PRB rate (of any PRB) does not change during the frame (for 10 ms). However, the per-PRB rate changes from one frame to another randomly for all users. Moreover, the per-PRB rates of different users are assumed to be mutually independent.

Although in practice different PRBs "bring" different rates, in order to preserve the analytical tractability, we make a simplifying assumption. Specifically, the assumption is that the gNodeB transmission power and channel characteristics of a user remain unchanged across all K PRBs in a frame (identical CQI over all PRBs for a given user). Thereby, the problem reduces to one-dimensional scheduling, *in time*, in which instead of deciding which PRBs to assign to a user over frames, another parameter is defined as:

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

PRBs are assigned orthogonally through the frame duration so that no two users receive the PRBs simultaneously. This is a reasonable simplification as the frame ratio can be translated into the corresponding number of PRBs assigned to a user per frame. In this paper, the frame ratio will be the quantitative measure of interest related to resource allocation in the cell.

Having in mind the previous assumptions, it follows that in every frame user's i per-PRB rate can be modeled as a discrete random variable, R_i, with values in the set $\{r_1, r_2, \ldots, r_{15}\}$, such that $r_1 < r_2 < \ldots < r_{15}$, with a probability mass function (PMF) $p_i(x)$. The latter is a function of user's i SINR over time.[1]

Number of users: There are n users in the cell with different per-PRB rate distributions. The set of users is denoted by $\mathcal{N}$.

Data rate: As the focus of this work are applications which require stable throughput, it is reasonable to assume that users most

[1] For the matter of notational convenience, we omit the reference to time in the remainder of this paper.

of the time will experience constant rates. In order to efficiently utilize network resources, the assumption is that users will have different data rates among themselves, but they are constant most of the time. These are denoted as $U_i, i \in \mathcal{N}$. As shown in [6], relaxing the requirement on providing the constant rate from 100% of the time to $(1-\epsilon)\cdot 100\%$ of the time increases considerably the achievable rate. Therefore, the data rate is allowed to be different from $U_i, i \in \mathcal{N}$ for $\epsilon \cdot 100\%$ of the time, where ϵ is known as the *outage probability*. Its value is usually very small. These *almost-always constant data rates* are denoted as $U_i(\epsilon), \forall i \in \mathcal{N}$. These rates are also known as *consistent rates* [6].[2]

Each user sends information about her CQI to the base station she is receiving service from. Then, based on the resource allocation policy used by the BS (which is usually for an optimization objective), the latter allocates the resources to the users (in our case the BS determines the value of Y_i and acts accordingly in sending the corresponding resources).

The relation between the frame ratio and data rate (in a frame) for user i is

$$U_i = KY_iR_i, \tag{1}$$

which implies that to provide the constant rate U_i, user i needs all the PRBs for the frame ratio of $Y_i = \frac{U_i}{KR_i}$. This leads to:

Definition 2. *The utilization of resources after providing the constant rates $U_i(\epsilon), \forall i \in \mathcal{N}$, in a frame is $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \frac{U_i(\epsilon)}{KR_i}$, where $R_i, \forall i \in \mathcal{N}$, are the per-PRB rates of the users in that frame.*

3.2 Problem Formulation

The first objective of this paper is to minimize the variation of data rates across all users in the cell. The parameter that captures the variability of a random variable is the coefficient of variation [39], which for the random variable X is defined as[3]

$$c_V(X) = \frac{\sqrt{\text{Var}(X)}}{\mathbb{E}[X]}. \tag{2}$$

Therefore, the first optimization objective in this paper is to minimize the sum of the coefficient of variations of all the users in the cell, i.e.,

$$\min_{Y_1,\ldots,Y_n} \sum_{i=1}^{n} c_{V,i} = \sum_{i=1}^{n} \frac{\sqrt{\text{Var}(U_i)}}{\mathbb{E}[U_i]}. \tag{3}$$

To realize this, the data rate should be almost always constant for every user. As data rates U_i are constant for all the users for $1-\epsilon$ of the time, the variances of data rates are small already. Therefore, our second goal is to effectively utilize network resources or equivalently, to reduce the amount of wasted resources, i.e., to make the total frame ratio as high as possible. Ideally, $\sum_{i=1}^{n} Y_i \to 1$.

While considering *separately* the minimization of rate variability and the maximization of resource utilization (first and second goal mentioned above) is used as a first step, it does not guarantee an overall joint optimization. Therefore, we go a step further and combine these two requirements into a single objective. To that end, a new metric is introduced:

Definition 3. *The ratio of the average level of utilization of resources and the sum of the coefficients of variation of data rates of all the users is called* **joint satisfaction of rate variability and utilization efficiency**, *or referred to as shortly* **joint satisfaction efficiency**, *and is defined as*

$$JSE = \frac{\sum_{i=1}^{n} \mathbb{E}[Y_i]}{\sum_{i=1}^{n} c_{V,i}}. \tag{4}$$

TABLE 1: Notation

$\mathcal{N}$	Set of all the users
$n = \lvert\mathcal{N}\rvert$	Number of users in the cell
$R_i(t)$	Per-PRB rate of user i in frame t
$p_i(x)$	PMF of user's i per-PRB rate
K	Total number of PRBs
ϵ	Outage probability
$U_i(\epsilon)$	Data rate with outage ϵ of user i
Y_i	Frame ratio resources are allocated to user i
$c_V(X)$	Coefficient of variation of X
f_i	Resource effectiveness for user i
JSE	Joint satisfaction efficiency

As the goal is to jointly maximize the level of resource utilization (the numerator of (4)), or equivalently, to minimize the amount of wasted resources, and to minimize the data rate variability across all the users in the cell (hence maximizing its inverse $\frac{1}{\sum_{i=1}^{n} c_{V,i}}$), this leads to the equivalent objective of maximizing the joint satisfaction efficiency (4). Note that in this objective, i.e., maximizing JSE, there are the first- and second-order statistics (the mean and the standard deviation) for one of the metrics (c_V). This makes the problem with a time horizon dimension. Therefore, the decision on the amount of resources to be allocated to every user has to be made by the base station in every frame, and will in general differ from one frame to another for all the users.

In this paper, we consider two approaches. In the first, described in detail in Section 4, the resources (PRBs) are reserved for the users. Then, depending on a user's channel conditions in a frame and the targeted data rate for that user, the BS decides to what extent the reserved resources for that user are utilized. In the second approach, resources are not reserved; they are allocated on the fly for all the users. The comprehensive analysis for the latter is presented in Section 5. For each of these two approaches, in the corresponding sections, the analyses for several policies are developed.

Before proceeding any further, Table 1 summarizes the notation used throughout this paper.

4 Performance Analysis for Reserved Resources

In this section, we present the approach in which resources are reserved for all the users from the beginning, and then in the next section the second, more flexible approach is presented in which there is no reservation of resources. Rather, they are allocated on the fly on a per-frame basis depending on the channel qualities of all the users in the frame. Common to both techniques is that they allocate resources in such a way that any user experiences a constant data rate at almost all times. These rates differ among different users and depend on the channel characteristics (CQI distribution) of all the users.

[2] The terms *constant* and *consistent* rate are used interchangeably in this paper.

[3] Note that $\mathbb{E}[X]$ denotes the expectation of the random variable X, whereas Var(X) represents the variance of the random variable X.

4.1 The Notion of Resource Effectiveness

With this approach, PRBs are reserved from the beginning. User i will receive K_i PRBs in each frame, where $\sum_{i=1}^{n} K_i = K$. As the requirement is to guarantee the data rate U_i to user i for $1-\epsilon$ of the time, the rate constraint can be expressed as

$$\mathbb{P}\left(\frac{U_i}{K_i R_i} \leq 1\right) \geq 1-\epsilon, \quad \forall i \in \mathcal{N}. \tag{5}$$

Inequality (5) is equivalent to

$$\mathbb{P}\left(\frac{1}{R_i} \leq \frac{K_i}{U_i}\right) \geq 1-\epsilon, \quad \forall i \in \mathcal{N}. \tag{6}$$

In (6), there are two unknowns, K_i and U_i. To determine the ratio $\frac{U_i}{K_i}$, (6) should be closely examined. Since its left-hand side (LHS) is the Cumulative Distribution Function (CDF) of $\frac{1}{R_i}$, it is an increasing function in $\frac{K_i}{U_i}$ (i.e., decreasing in $\frac{U_i}{K_i}$), and it should not be smaller than $1-\epsilon$. Therefore, the point at which the strict equality in (6) holds is of interest. At that point, the maximum allowed value of $\frac{U_i}{K_i}$ is achieved. As a result, for every user i, based on its channel characteristics, the maximum value of U_i/K_i can be determined, so that the constraint (5) is not violated. It follows:[4]:

Result 1. *The maximum value of $\frac{U_i}{K_i}$ for user i is*

$$\left(\frac{U_i}{K_i}\right)_{\max} = \frac{1}{F^{-1}_{\frac{1}{R_i}}(1-\epsilon)}, \quad \forall i \in \mathcal{N}, \tag{7}$$

where $F^{-1}_{\frac{1}{R_i}}(1-\epsilon)$ denotes the inverse of the CDF of $\frac{1}{R_i}$ at $1-\epsilon$.

Definition 4. *The parameter $f_i(\epsilon) = \frac{U_i(\epsilon)}{K_i}, \forall i \in \mathcal{N}$, is called the* ***resource effectiveness*** *of user i, and represents the data rate user i experiences per unit of allocated resource (PRB).*

Essentially, the meaning of this parameter is that *the user with higher resource effectiveness will have a higher data rate for the same amount of allocated PRBs compared to the user with lower resource effectiveness.* As the goal is to improve the performance of the users, in this paper it is assumed that all the users operate with their maximum resource effectiveness.

Intuitively, users with good channel conditions should have higher resource effectiveness. However, higher resource effectiveness is not necessarily related to the first moment of per-PRB rates. In fact, it depends on the entire distribution of the per-PRB rate, and not solely on the first moment. This will be shown in Section 6.

Fig. 2 illustrates the way in which $\frac{1}{f_i(\epsilon)}$, i.e., $F^{-1}_{1/R_i}(1-\epsilon)$, is determined. Point A in Fig. 2 is the crossing-point of the CDF of $\frac{1}{R_i}$ with $1-\epsilon$. The value that corresponds to that point on the x-axis is $\frac{1}{r}$ (r can take its value from the discrete set $\{r_1, \ldots, r_{15}\}$). Therefore, the resource effectiveness is $\frac{1}{\frac{1}{r}} = r$, which means that it can take a value only from the discrete set of possible per-PRB rates. Also, relaxing the requirement on the time to provide the guaranteed data rate, i.e., increasing the value ϵ, leads to higher values of the resource effectiveness. The last claim comes from the fact that increasing ϵ decreases $1-\epsilon$, hence the crossing point of $1-\epsilon$ and $F^{-1}_{1/R_i}(1-\epsilon)$ shifts down to the lower values on the x-axis, i.e., $\frac{1}{r}$ declines, increasing thus the value of $f = r$.

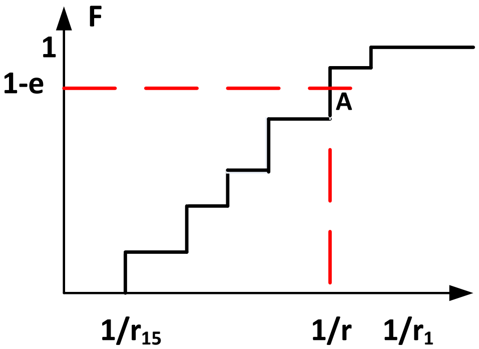


Fig. 2: Illustration of how f is determined.

Note: We remind the reader that the corresponding constant data rate is guaranteed to every user for $1-\epsilon$ of the time. For the remaining ϵ, when the data rate guarantee is not met, resources are split equally among all the users.

4.2 Maximizing Network Utilization

The approach we follow in this paper is a consequence of the fact that by allowing for a small ratio of time a deviation from an at-all-times constant rate, the data rates that a user experiences increase significantly [23], but the ratio of time this deviation happens is quite small, not increasing thus the coefficient of variation of data rates. Therefore, it is a reasonable approach to obtain the "best of both worlds." In that case, assuming that irrespective of the actual allocation policy with reserved resources the coefficients of variation of all the users would be small (as an argument in that direction are Fig. 7 and Fig. 10 later in Section 6), we look only at the level of resource allocation and derive the policy which maximizes it. To that end, the following optimization problem needs to be solved:

$$\max_{K_1,\ldots,K_n} \mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right] \tag{8}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} K_i \leq K, \tag{9}$$

$$K_i \geq 1, \quad \forall i \in \mathcal{N}. \tag{10}$$

The objective is a function of ϵ because for different outages different allocation policies will need to be enforced. This will be more clear after solving the aforementioned optimization problem. In terms of the constraints, every user will need at least one PRB reserved (10). Otherwise, she would experience a 0 data rate, which does not make sense.

Let us pay a closer attention to the objective function $\mathbb{E}\left[Y_i(\epsilon)\right]$, which yields

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right] = \frac{\sum_{i=1}^{n}\left(\mathbb{P}(R_i < f_i) + \mathbb{P}(R_i \geq f_i) f_i \mathbb{E}\left[\frac{1}{R_i}\middle| R_i \geq f_i\right]\right) K_i}{K}, \tag{11}$$

where

$$\mathbb{E}\left[\frac{1}{R_i}\middle| R_i \geq f_i\right] = \frac{\sum_{r_l=f_i}^{r_{15}} \frac{1}{r_l} p_i(r_l)}{\mathbb{P}(R_i \geq f_i)}. \tag{12}$$

[4] The distribution of $\frac{1}{R_i}$ is used instead of that of R_i for practical purposes, in order to be able to express the maximum value of $\frac{U_i}{K_i}$ in closed form.

The expression under the sum on the right-hand side (RHS) of (11) denotes the average frame ratio all K_i PRBs are utilized by user i. In cases when the consistent rate is not provided, resources are fully utilized, i.e., the entire frame. This occurs with probability $\mathbb{P}\left(\frac{U_i}{K_iR_i} > 1\right) = \mathbb{P}\left(\frac{f_i}{R_i} > 1\right) = \mathbb{P}(R_i < f_i)$. On the other hand, in frames in which the consistent rate is provided, which occurs with probability $\mathbb{P}\left(\frac{U_i}{K_iR_i} \leq 1\right) = \mathbb{P}(R_i \geq f_i)$, on average the dedicated PRBs are allocated to user i for the frame ratio of $\mathbb{E}\left[\frac{U_i}{K_iR_i}\middle| R_i \geq f_i\right] = f_i\mathbb{E}\left[\frac{1}{R_i}\middle| R_i \geq f_i\right]$.

Substituting (12) into (11) yields

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{K}\sum_{i=1}^{n}\left(\mathbb{P}(R_i < f_i) + f_i\sum_{r_l=f_i}^{r_{15}}\frac{1}{r_l}p_i(r_l)\right)K_i. \tag{13}$$

The general form of the resource utilization level can be rewritten as

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{K}\sum_{i=1}^{n} A_iK_i, \tag{14}$$

where

$$A_i = \mathbb{P}(R_i < f_i) + f_i\sum_{r_l=f_i}^{r_{15}}\frac{1}{r_l}p_i(r_l), \tag{15}$$

and A_i does not depend on the resource allocation policy, but only on the channel characteristics of the user.

As, in line with constraint (10), at least one PRB must be assigned to every user, objective (8) reduces to

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{K}\left(\sum_{i=1}^{n} A_i(K_i-1) + \sum_{i=1}^{n} A_i\right). \tag{16}$$

In (16), the only term that depends on the PRB allocation is $\sum_{i=1}^{n} A_i(K_i-1)$. Therefore, as equivalent objective arises:

$$\max_{K_1,\ldots,K_n}\sum_{i=1}^{n} A_i(K_i-1). \tag{17}$$

The maximum of (17) is achieved if first all the users receive one PRB, and the remaining PRBs are reserved for the user with the highest A, i.e., $j = \underset{i\in\mathcal{N}}{\mathrm{argmax}} A_i$ is the user that receives the remaining PRBs. Hence, it follows:

Result 2. *With resource reservation, the maximum average utilization of network resources where different constant data rates are guaranteed to all the users with outage ϵ is achieved if the resource allocation is done according to the policy:*

$$K_i = 1, \quad \forall i \in \mathcal{N}\setminus j, \textit{ and } \quad K_j = K - n + 1, \tag{18}$$

where $j = \underset{i\in\mathcal{N}}{\mathrm{argmax}} A_i$. The maximum average utilization that can be achieved is

$$\underset{K_1,\ldots,K_n}{\mathit{max}}\ \mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right] = \frac{A_j(K-n+1)}{K} + \sum_{i=1,i\neq j}^{n}\frac{A_i}{K}, \tag{19}$$

where A_i's are given by (15).

This policy will be referred to as **RR-OPT**. As will be seen in Section 6, **RR-OPT** outperforms the other policies with resource reservation in terms of resource utilization.

4.3 Maximizing JSE

Next, the goal is to optimize the performance by jointly considering the variability of data rates (through their coefficient of variation) and the level of resource utilization. To that end, the following optimization problem has to be solved:

$$\max_{K_1,\ldots,K_n}\frac{\sum_{i=1}^{n}\mathbb{E}[Y_i]}{\sum_{i=1}^{n}c_{V,i}} \tag{20}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} K_i \leq K, \tag{21}$$

$$K_i \geq 1, \quad \forall i \in \mathcal{N}. \tag{22}$$

The same explanation of the constraints holds for this optimization formulation as for (8)-(10). This is a stochastic optimization problem [40]. Due to its structure, it is possible to fins a solution in a "non-conventional way." Let us pay closer attention to the objective. Substituting (14) for $\mathbb{E}[Y_i]$ and (3) for $\sum_{i=1}^{n} c_{V,i}$ into (20) yields

$$\frac{\sum_{i=1}^{n}\mathbb{E}[Y_i]}{\sum_{i=1}^{n}c_{V,i}} = \frac{\frac{1}{K}\sum_{i=1}^{n}A_iK_i}{\sum_{i=1}^{n}\frac{\sqrt{\mathbb{E}[\tilde{U}_i^2]-(\mathbb{E}[\tilde{U}_i])^2}}{\mathbb{E}[\tilde{U}_i]}}. \tag{23}$$

We have made a slight change in notation in (23) and now $\tilde{U}_i$ denotes the data rate instead of U_i. This is done as for up to ϵ of the time, the rate is different from U_i. The data rate of user i when the advertised rate cannot be provided is K_iR_i. This happens in instances in which $R_i < f_i$. As the quasi-constant rate U_i is provided whenever the per-PRB rate of the user is at least equal to the resource effectiveness of user i (see the discussion succeeding (12)), then

$$\mathbb{E}[\tilde{U}_i] = \mathbb{P}(R_i \geq f_i)U_i + \mathbb{P}(R_i < f_i)\mathbb{E}[K_iR_i|R_i < f_i], \tag{24}$$

which after substituting $U_i = K_if_i$, results in

$$\mathbb{E}[\tilde{U}_i] = K_if_i\mathbb{P}(R_i \geq f_i) + K_i\mathbb{E}[R_i|R_i < f_i]\mathbb{P}(R_i < f_i). \tag{25}$$

Note that K_i does not depend on the current R_i in the frame (reserved resources), hence, it goes outside of the expectation. Similarly, for the second moment of the data rate it holds

$$\mathbb{E}[\tilde{U}_i^2] = K_i^2f_i^2\mathbb{P}(R_i \geq f_i) + K_i^2\mathbb{E}[R_i^2|R_i < f_i]\mathbb{P}(R_i < f_i). \tag{26}$$

Combining (25) and (26), for the coefficient of variation of the data rate of user i, it holds

$$c_{V,i} = \sqrt{\frac{f_i^2\mathbb{P}(R_i \geq f_i) + \mathbb{E}[R_i^2|R_i < f_i]\mathbb{P}(R_i < f_i)}{\left(f_i\mathbb{P}(R_i \geq f_i) + \mathbb{E}[R_i|R_i < f_i]\mathbb{P}(R_i < f_i)\right)^2} - 1}, \tag{27}$$

which leads to the conclusion that

$$\sum_{i=1}^{n} c_{V,i} \neq f(K_i). \tag{28}$$

Having this in mind, the optimization problem (20)-(22) attains the same solution as the optimization problem (8)-(10). Namely, as already shown, there are no decision variables in the denominator of (23) - see (28). So, the objective (23) is equivalent to the objective function

$$\max_{K_1,\ldots,K_n}\sum_{i=1}^{n} A_iK_i, \tag{29}$$

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.2024.3389589

where the latter, as already shown, given that at least one PRB has to be allocated to every user, is equivalent to (17). Therefore, the same solution for the resource allocation is obtained as in Result 2. The value of the objective is different, however, because of the sum of the coefficients of variation of data rates in the denominator. Nevertheless, the actual value of JSE is not as descriptive as the average level of resource utilization. Hence, it is omitted here. To summarize, the following result is obtained:

Result 3. *With resource reservation, the maximum JSE is achieved if the resource allocation is done according to the policy:*

$$K_i = 1, \quad \forall i \in \mathcal{N} \setminus j, \text{ and } \quad K_j = K - n + 1, \tag{30}$$

where $j = \underset{i \in \mathcal{N}}{\operatorname{argmax}} A_i$.

The importance of this result is highlighted by the fact that irrespective of whether one is interested in maximizing only the resource utilization (and hence minimizing the amount of wasted resources) or by considering jointly the utilization level and the variability of data rates (maximizing JSE), the same optimal allocation policy is obtained, i.e., **RR-OPT**, which further implies that it suffices to look only for maximizing the average level of network utilization[5], as long as the same rate is provided to a user for most of the time. Needless to say, for different values of ϵ, different resource allocation policies will be in place. This is reminiscent of the fact that the resource effectiveness depends on ϵ, and for different values of the latter, the resource effectiveness of one user can be larger than the resource effectiveness of another user, but this can change for another value of ϵ. More on this in Section 6.

4.4 Other Resource-Reservation Policies

In the previous two subsections, we proved that the policy **RR-OPT** is optimal in terms of both maximizing the resource utilization and JSE. However, there is a drawback associated with **RR-OPT**. Namely, its main feature is that only the user with the highest A_i will receive many PRBs, and hence will experience a very high data rate, while all the other users will receive only one PRB each. This leads to extremely low rates for most users.[6] Hence, the performance of three other resource-reservation policies is considered here. These are: 1) equal-share of resources (**RR-ES**), 2) resources are allocated directly proportionally to the resource effectiveness of the users (**RR-P**), and 3) resources are allocated inversely proportionally to the resource effectiveness of the users (**RR-IP**).

4.4.1 Equal-share of resources (*RR-ES*)

With this policy, each user will in total have $K_i = \frac{K}{n}$ PRBs (reserved) in every frame. Therefore, from (7), it follows:

Result 4. *Using the **RR-ES** policy, user i will have a constant data rate of*

$$U_i(\epsilon) = \frac{K f_i}{n} = \frac{K}{n F^{-1}_{\frac{1}{R_i}}(1-\epsilon)}, \quad \forall i \in \mathcal{N}. \tag{31}$$

With this policy, users with higher resource effectiveness (higher f_i) will experience higher data rates.

4.4.2 Users with higher resource effectiveness receive proportionally more resources (*RR-P*)

In this case, the relation between the allocated resources for two users is expressed as

$$\frac{K_i}{K_j} = \frac{f_i}{f_j}, \tag{32}$$

i.e., each user will receive the number of PRBs proportionally to its resource effectiveness. Further, (32) yields

$$K_j = K_i \frac{f_j}{f_i}. \tag{33}$$

Substituting (33) into $\sum_{j=1}^{n} K_j = K$, and solving for K_i, leads to

$$K_i = K \frac{f_i}{\sum_{j=1}^{n} f_j}, \quad \forall i \in \mathcal{N}. \tag{34}$$

Finally, from $\frac{U_i(\epsilon)}{K_i} = f_i(\epsilon)$ and (34), the following result is obtained:

Result 5. *When allocating the PRBs according to the **RR-P** policy, user i will receive the constant data rate of*

$$U_i(\epsilon) = \frac{K f_i(\epsilon)^2}{\sum_{j=1}^{n} f_j(\epsilon)}, \quad \forall i \in \mathcal{N}. \tag{35}$$

With this policy, users with good resource effectiveness (higher f_i) are rewarded, whereas there is a penalization for users with low resource effectiveness (lower f_i).

4.4.3 Users with worse channels receive proportionally more resources (*RR-IP*)

In this case, for two users i and j, the number of PRBs allocated to them satisfies the condition

$$\frac{K_i}{K_j} = \frac{f_j}{f_i}, \tag{36}$$

i.e., each user will receive the number of PRBs which is inversely proportional to its resource effectiveness. This yields

$$K_j = K_i \frac{f_i}{f_j}. \tag{37}$$

Next, substituting (37) into $\sum_{j=1}^{n} K_j = K$, and solving for K_i yields

$$K_i = K \frac{\frac{1}{f_i}}{\sum_{j=1}^{n} \frac{1}{f_j}}, \quad \forall i \in \mathcal{N}. \tag{38}$$

Finally, substituting (38) into $\frac{U_i(\epsilon)}{K_i} = f_i(\epsilon)$ leads to:

Result 6. *When allocating the PRBs according to the **RR-IP** policy, user i will receive the constant data rate of*

$$U_i(\epsilon) = \frac{K}{\sum_{j=1}^{n} \frac{1}{f_j(\epsilon)}}, \quad \forall i \in \mathcal{N}. \tag{39}$$

Result 6 shows that every user will receive the same constant data rate under **RR-IP**. This policy penalizes users with good resource effectiveness because they receive fewer network resources than users with low resource effectiveness so that everyone has the same rate for the vast majority of the time.

Note: The levels of resource utilization achieved by **RR-ES**, **RR-P**, and **RR-IP** can be compared among themselves, after substituting the corresponding K_i policy into (13), and based on that one can decide which policy to use for a given scenario. We omit this analysis here. It can be found in our technical report [41].

[5] The allocation policy in Result 2 is identical to that of Result 3.

[6] Nevertheless, **RR-OPT** can serve as an indicator of the upper bound of the resource utilization and JSE.

5 Performance Analysis for No-Reservation of Resources

When there is no reservation of resources for the users, the decision on what amount to allocate to every user is made dynamically over time, depending on the channel conditions of all the users in a given frame. The constraint on service outage with this approach is

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{U_i}{KR_i}\leq 1\right)\geq 1-\epsilon, \tag{40}$$

or equivalently,

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{U_i}{R_i}\leq K\right)\geq 1-\epsilon. \tag{41}$$

Determining a single set of maximum (constant) values $U_i(\epsilon), \forall i\in\mathcal{N}$, from the previous inequality is impossible, as it is an under-determined inequality, i.e., one needs to determine n unknowns from a single inequality. Nevertheless, we will approach this problem in a slightly different way in order to obtain the maximum constant rates for every user. It is worth mentioning that now since the amount of allocated resources is not fixed (no resource reservation), using the resource effectiveness does not make sense. Therefore, in the policies that follow, a different strategy is used to determine the maximum U_i for a given ϵ.

The main question that arises is: *Knowing that frame ratio (resource utilization) changes across frames, how does one decide on the constant rate to guarantee to a user with probability* $1-\epsilon$*?*

With this dynamic approach, solving any of the optimization problems (8)-(10) or (20)-(22) is not feasible. Hence, regarding this approach, the performance of two policies is considered: 1) All users receive the same frame ratio on average (**NR-EY**), and 2) users will have data rates proportionally to their average per-PRB rates (**NR-P**). As will be seen in Section 6, both these dynamic allocation policies outperform the optimal resource-reservation policy in the level of resource utilization.

5.1 All Users Receive the Same Frame Ratio on Average (NR-EY)

The amount of resources (frame ratio) user i receives in frame t is

$$Y_i(t)=\begin{cases}\frac{U_i}{KR_i(t)}, & \sum_{i=1}^{n}\frac{U_i}{KR_i(t)}\leq 1\\ \frac{1}{n}, & \sum_{i=1}^{n}\frac{U_i}{KR_i(t)}>1\end{cases}$$

In frames in which resources are not sufficient to provide the rate U_i, the frame ratio is $\frac{1}{n}$, i.e., every user will receive all the PRBs for $\frac{1}{n}$ of the frame. In the frames in which the (different) constant rates of all the users can be provided, if there is a requirement for the average utilization ratio to be the same across all the users, i.e., $\mathbb{E}[Y_1]=\ldots=\mathbb{E}[Y_n]$, for those frames it must hold (note that in frames in which resources are not sufficient to provide the constant rates to all the users, the resources are anyway split among the users)

$$U_i\mathbb{E}\left[\frac{1}{R_i}\right]=const, \quad \forall i\in\mathcal{N}. \tag{42}$$

Expressing (42) in terms of user 1, it follows that

$$U_i=U_1\cdot\frac{\mathbb{E}\left[\frac{1}{R_1}\right]}{\mathbb{E}\left[\frac{1}{R_i}\right]}, \quad \forall i\in\mathcal{N}. \tag{43}$$

Substituting (43) for all the users in (41), and rearranging, the following holds:

$$\mathbb{P}\left(\sum_{i}^{n}\frac{d_i}{R_i}\leq\frac{K}{U_1}\right)\geq 1-\epsilon, \tag{44}$$

where $d_i=\frac{\mathbb{E}\left[\frac{1}{R_1}\right]}{\mathbb{E}\left[\frac{1}{R_i}\right]}$.

The LHS of (44) is the CDF of the sum of independent random variables $\frac{d_i}{R_i}$. As is well known [39], the CDF of the sum of independent random variables is the convolution of the CDF of the first variable with the probability mass functions (PMF) of the other variables. In this case, it would correspond to

$$F_{\sum_i\frac{d_i}{R_i}}(x)=F_{\frac{d_1}{R_1}}(x)*p_{\frac{d_2}{R_2}}(x)*\ldots*p_{\frac{d_n}{R_n}}(x). \tag{45}$$

The first RHS term of (45) transforms into

$$\mathbb{P}\left(\frac{d_1}{R_1}\leq x\right)=\mathbb{P}\left(\frac{1}{R_1}\leq\frac{x}{d_1}\right)=\sum_{k_1=1}^{15}p_1(r_{k_1})u\left(x-\frac{d_1}{r_{k_1}}\right), \tag{46}$$

where $u(x)$ represents the Heaviside (unit) step function [42], whose value is 1 for $x\geq 0$. Otherwise, its value is 0. Further, for the other RHS terms ($i\neq 1$), it holds:

$$p_{\frac{d_i}{R_i}}(x)=\mathbb{P}\left(\frac{1}{R_i}=\frac{x}{d_i}\right)=\sum_{k_i=1}^{15}p_i(r_{k_i})\delta\left(x-\frac{d_i}{r_{k_i}}\right), \tag{47}$$

where $\delta(x)$ represents the delta function [42], whose value is 1 only at $x=0$. Otherwise, it is 0.

Before proceeding with (45), it should be pointed out that the delta function is the unit element with respect to the operation of convolution. Also, it holds that $x(t)*\delta(t-t_0)=x(t-t_0)$ [42], i.e., it shifts the function.

With the previous facts in mind, combining (46) and (47), $\forall i\in\mathcal{N}\backslash\{1\}$, into (45), and rearranging, the following is obtained:

$$F_{\sum_i\frac{d_i}{R_i}}\left(\frac{K}{U_1}\right)=\sum_{k_1=1}^{15}\cdots\sum_{k_n=1}^{15}\prod_{i=1}^{n}p_i(r_{k_i})u\left(\frac{K}{U_1}-\sum_{i=1}^{n}\frac{d_i}{r_{k_i}}\right). \tag{48}$$

Obtaining U_1 from (48) is not tractable. Nevertheless, there exists a simpler way. As the CDF is an increasing function in $\frac{K}{U_1}$, it follows that it is a decreasing function in U_1. Therefore, as maximum value of U_1 should be taken the one for which (44) reduces to a strict equality, i.e.,

$$F_{\sum_i\frac{d_i}{R_i}}\left(\frac{K}{U_1}\right)=1-\epsilon, \tag{49}$$

which yields

$$U_1(\epsilon)=\frac{K}{F^{-1}_{\sum_i\frac{d_i}{R_i}}(1-\epsilon)}. \tag{50}$$

Finally, substituting (50) into (43) yields:

Result 7. *The maximum constant data rate that can be guaranteed to user i in the cell for* $1-\epsilon$ *of the time with the* ***NR-EY*** *policy is*

$$U_i(\epsilon)=\frac{K}{F^{-1}_{\sum_j\frac{d_j}{R_j}}(1-\epsilon)}\cdot\frac{\mathbb{E}\left[\frac{1}{R_1}\right]}{\mathbb{E}\left[\frac{1}{R_i}\right]}, \quad \forall i\in\mathcal{N}. \tag{51}$$

Observing Result 7, one can infer that better channel conditions (lower $\mathbb{E}\left[\frac{1}{R_i}\right]$) imply a higher rate for user i. This is coherent with the fact that, on average, all the users receive the same amount of resources. Consequently, the user with better channel conditions experiences a higher data rate.

Given that for ϵ of the time the data rates U_i cannot be provided, and then all the resources are split among the users (utilization is 100%), the average resource utilization with this policy is

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right]_{\textbf{NR-EY}} = \epsilon + (1-\epsilon)\sum_{i=1}^{n}\mathbb{E}\left[\frac{U_i(\epsilon)}{KR_i}\middle|\sum_{i=1}^{n}\frac{U_i(\epsilon)}{KR_i}\leq 1\right]. \tag{52}$$

The second RHS term in (52) corresponds to the frames in which resources are sufficient to provide the constant rates U_i when the utilization is less than or equal to 100%. As average utilization is the same for all the users, (52) transforms into

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right]_{\textbf{NR-EY}} = \epsilon + \frac{(1-\epsilon)n}{K}\mathbb{E}\left[\frac{U_i(\epsilon)}{R_i}\middle|\sum_{i=1}^{n}\frac{U_i(\epsilon)}{R_i}\leq K\right], \tag{53}$$

where $U_i(\epsilon)$ are obtained using (51). The previous expectation part leads to

$$\mathbb{E}\left[\frac{U_i}{R_i}\middle|\sum_{i=1}^{n}\frac{U_i}{R_i}\leq K\right] = \frac{\sum_{r_{k_i}=r_1}^{r_{15}}\cdots\sum_{r_{k_n}\geq y_n}^{r_{15}}\frac{U_i}{r_{k_i}}\prod_{i=1}^{n}p_i(r_{k_i})}{\mathbb{P}\left(\sum_{i=1}^{n}\frac{U_i}{R_i}\leq K\right)}, \tag{54}$$

where $y_n = \frac{U_n}{K-\sum_{i=1}^{n-1}\frac{U_i}{r_{k_i}}}$. Note that the denominator of the RHS of (54) is simply (assuming the network is pushed to operate to the maximum values of U_i until the rate constraint holds)

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{U_i}{R_i}\leq K\right) = 1-\epsilon. \tag{55}$$

In terms of y_n, that value was obtained in the following way: For a given realization of $(R_1, \dots R_{n-1})$, say $(r_{k_1}, \dots, r_{k_{n-1}})$, for the realization of $R_n = r_{k_n}$, given that the expectation is conditioned upon $\sum_{i=1}^{n}\frac{U_i}{R_i}\leq K$, it follows

$$\frac{U_n}{r_{k_n}} \leq K - \sum_{i=1}^{n-1}\frac{U_i}{r_{k_i}}, \tag{56}$$

which leads to

$$r_{k_n} \geq \frac{U_n}{K-\sum_{i=1}^{n-1}\frac{U_i}{r_{k_i}}} = y_n. \tag{57}$$

A similar expression is obtained for other values in the lower bound of the corresponding summation term.

Substituting (54) into (53), for the average utilization it holds

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right]_{\textbf{NR-EY}} = \epsilon + \frac{n}{K}\sum_{r_{k_i}=r_1}^{r_{15}}\cdots\sum_{r_{k_n}\geq y_n}^{r_{15}}\frac{U_i}{r_{k_i}}\prod_{i=1}^{n}p_i(r_{k_i}). \tag{58}$$

Finally, substituting (51) into (58), the following result is obtained:

Result 8. *The average utilization of network resources with the **NR-EY** policy is*

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right]_{\textbf{\textit{NR-EY}}} = \epsilon + \frac{n\mathbb{E}\left[\frac{1}{R_1}\right]\sum_{r_{k_i}=r_1}^{r_{15}}\cdots\sum_{r_{k_n}\geq y_n}^{r_{15}}\frac{\prod_{i=1}^{n}p_i(r_{k_i})}{r_{k_i}}}{\mathbb{E}\left[\frac{1}{R_i}\right]F^{-1}_{\sum_i\frac{d_i}{R_i}}(1-\epsilon)}. \tag{59}$$

Result 8 suggests that the lower the value of $F^{-1}_{\sum_i\frac{d_i}{R_i}}(1-\epsilon)$, the higher the utilization of network resources is.

5.2 Users with Better Channel Conditions Receive Proportionally Higher Data Rates (NR-P)

With this policy, the relation between the data rates of two users is expressed as

$$\frac{U_i}{U_j} = \frac{\mathbb{E}[R_i]}{\mathbb{E}[R_j]}, \tag{60}$$

i.e., each user will receive a data rate that is proportional to its first moment of per-PRB rate. Adapting (60) with respect to user $j=1$, the following holds:

$$U_i = U_1 e_i, \quad \forall i \in \mathcal{N}, \tag{61}$$

where $e_i = \frac{\mathbb{E}[R_i]}{\mathbb{E}[R_1]}$. Combining (61) into (41) leads to

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{e_i}{R_i}\leq\frac{K}{U_1}\right)\geq 1-\epsilon. \tag{62}$$

The remainder of the procedure for deriving the maximum value of $U_1(\epsilon)$ is the same as when obtaining (50), yielding

$$U_1(\epsilon) = \frac{K}{F^{-1}_{\sum_i\frac{e_i}{R_i}}(1-\epsilon)}. \tag{63}$$

Finally, substituting (63) into (61) results in:

Result 9. *The maximum constant rate that can be guaranteed to user i in the cell for $1-\epsilon$ of the time with **NR-P** is*

$$U_i(\epsilon) = \frac{K}{F^{-1}_{\sum_i\frac{e_i}{R_i}}(1-\epsilon)}\cdot\frac{\mathbb{E}[R_i]}{\mathbb{E}[R_1]}, \quad \forall i \in \mathcal{N}. \tag{64}$$

In a similar vein with **NR-EY**, the average utilization of network resources when using **NR-P** is

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right]_{\textbf{NR-P}} = \epsilon + (1-\epsilon)\sum_{i=1}^{n}\mathbb{E}\left[\frac{U_i(\epsilon)}{KR_i}\middle|\sum_{i=1}^{n}\frac{U_i(\epsilon)}{KR_i}\leq 1\right], \tag{65}$$

or equivalently,

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i(\epsilon)\right]_{\textbf{NR-P}} = \epsilon + \frac{1-\epsilon}{K}\sum_{i=1}^{n}\mathbb{E}\left[\frac{U_i(\epsilon)}{R_i}\middle|\sum_{i=1}^{n}\frac{U_i(\epsilon)}{R_i}\leq K\right]. \tag{66}$$

Substituting (54) into (66) leads to

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right]_{\textbf{NR-P}} = \epsilon + \frac{\sum_{i=1}^{n}\sum_{r_{k_i}=r_1}^{r_{15}}\cdots\sum_{r_{k_n}\geq y_n}^{r_{15}}\frac{U_i}{r_{k_i}}\prod_{i=1}^{n}p_i(r_{k_i})}{K} \tag{67}$$

Finally, substituting (64) into (67) yields:

Result 10. *The average utilization of network resources with the **NR-P** policy is*

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right]_{\textbf{\textit{NR-P}}} = \epsilon + \frac{\sum_{i=1}^{n}\mathbb{E}[R_i]\sum_{r_{k_i}=r_1}^{r_{15}}\cdots\sum_{r_{k_n}\geq y_n}^{r_{15}}\frac{\prod_{i=1}^{n}p_i(r_{k_i})}{r_{k_i}}}{\mathbb{E}[R_1]F^{-1}_{\sum_i\frac{e_i}{R_i}}(1-\epsilon)} \tag{68}$$

Because of the complex interplay of $\mathbb{E}[R_i]$ and $\mathbb{E}\left[\frac{1}{R_i}\right]$ on the one hand, as well as $F^{-1}_{\sum_i\frac{d_i}{R_i}}(1-\epsilon)$ and $F^{-1}_{\sum_i\frac{e_i}{R_i}}(1-\epsilon)$ on the other, it is difficult to analytically predict under what channel conditions the utilization will be higher with **NR-P** than with **NR-EY** and vice versa.

Substituting (59) and the corresponding data rates achieved with **NR-EY** into (4), JSE for **NR-EY** is obtained. Similarly, JSE for **NR-P** is obtained by substituting (68) and the corresponding data rates into (4).

TABLE 2: Per-PRB rates and the corresponding probabilities for every user from the Republic of Ireland trace [43]

R (kbps)	48	73.6	121.8	192.2	282	378	474.2	612	772.2	874.8	1063.8	1249.6	1448.4	1640.6	1778.4
$p_1(r_k)$	0	0	0	0	0	0	0.01	0.05	0.11	0.13	0.14	0.18	0.06	0.11	0.21
$p_2(r_k)$	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
$p_3(r_k)$	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
$p_4(r_k)$	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
$p_5(r_k)$	0	0	0	0	0	0	0.04	0.07	0.13	0.17	0.22	0.2	0.05	0.06	0.06
$p_6(r_k)$	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
$p_7(r_k)$	0	0	0	0	0	0	0.04	0.07	0.13	0.17	0.22	0.2	0.05	0.06	0.06
$p_8(r_k)$	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

In Section 6, results for the variability of data rates and resource utilization for these policies will be shown, and also their performance will be compared against other approaches.

Note that it is completely inefficient to provide higher data rates to users with worse channel conditions, and also unfair to users with good channel conditions. Therefore, we do not pursue that policy here.

The analyses derived in the paper assume there is no link outage. If the connectivity drops when a UE is supposed to send its CQI, the BS could assume the value of the last received CQI from that UE in making the resource allocation decision, or it can exploit the knowledge of the channel statistics. It is also worth pointing out that the computational complexity for both approaches can be shown to be $O(n^2)$, hence, scalability is guaranteed.

6 Performance evaluation

First, the simulation setup and the measurement setup are described. Then, our results are validated using our own built system relying on OpenAirInterface. This is followed by results related to both approaches (resource reservation and no reservation) for different policies, and comparisons with other state-of-the-art techniques. Finally, the outcomes from a practical use-case scenario are presented.

6.1 Simulation Setup

For input parameters, we used a 5G trace with data measured in the Republic of Ireland, as this is the best data trace available to date. These traces can be found in [44], with a detailed description in [43], and a statistical analysis in [17]. The parameter of interest from the trace is CQI with 15 levels (with data showing a certain degree of correlation), which serves to determine the per-PRB rate of a user in a frame. These measurements were conducted for one user, but on different days, for different applications, both for the cases when the user was static and moving around. To mimic the dynamic nature of these users, we picked data for eight different days when the user was moving around and represented them in our analysis as eight different users in the same cell. It should be mentioned that we could not find any trace with CQI data in a 5G network for multiple users simultaneously, despite our comprehensive search. Based on the frequency of occurrence of a per-PRB rate for every user, the corresponding per-PRB rate probabilities were obtained (Table 2). [7]

The frame duration is $10\,\text{ms}$. The subcarrier spacings of 15 KHz and 30 KHz are considered, with 12 subcarriers per PRB. Hence, in scenarios corresponding to the first case the PRB width is $180\,\text{KHz}$, whereas in the others it is $360\,\text{KHz}$. In $30\,\text{KHz}$ scenarios, the number of PRBs is 273[8] [24]. The simulations are conducted in MATLAB R2022b.

6.2 Measurement Setup

In order to validate the theoretical results presented in the previous sections, here details are provided with respect to the measurement setup. Our setup is based on the OpenAirInterface architecture [9], which resembles the cellular network infrastructure. More specifically, our setup is split into three parts, namely core network, BS, and UEs. For our measurements, the 4G version of OpenAirInteface core [45] is utilized due to its stability compared to the 5G counterpart as well as the fact that the choice of the core network does not alter the decision of the scheduling policies in RAN. Furthermore, in this work, for the BS and UEs, the emulation mode of OpenAirInterface based on the `mosaic5g-oai-sim` branch [46], [47] is used. The rationale for this choice is twofold. Firstly, the `mosaic5g-oai-sim` possesses a realistic wireless channel model based on 3GPP standardization and secondly, it renders an easier deployment and assessment of results with realistic UE traces but also enables reproducibility. Within the BS, several UEs can be created and depending on the scheduling policy running at the BS, 25 PRBs are shared among eight UEs. Similar to the OpenAirInterface tutorials (see [48] for details), we measure the throughput of each UE within a BS using iperf [49] every $1\,\text{s}$, whereas measurements are performed for $1200\,\text{s}$. The subcarrier spacing supported by our system is 15 KHz. To match the OpenAirInterface deployment, the subcarrier spacing in the validation part (Section 6.3) is 15 KHz. In the other scenarios, the subcarrier spacing (in the simulations) is assumed to be $30\,\text{KHz}$, to distinguish the 5G from the 4G setup.

In our setup, an Intel(R) Core(TM) i7-7700T CPU @ 2.9 GHz is used for both OpenAirInterface BSs and the core network operation. The PC that operates the BS contains four physical CPU cores and 16 GB of RAM, whereas the core network contains one CPU and 4 GB of RAM. The operating system is Ubuntu 16.04.04 LTS with 4.4.0-116-generic kernel.

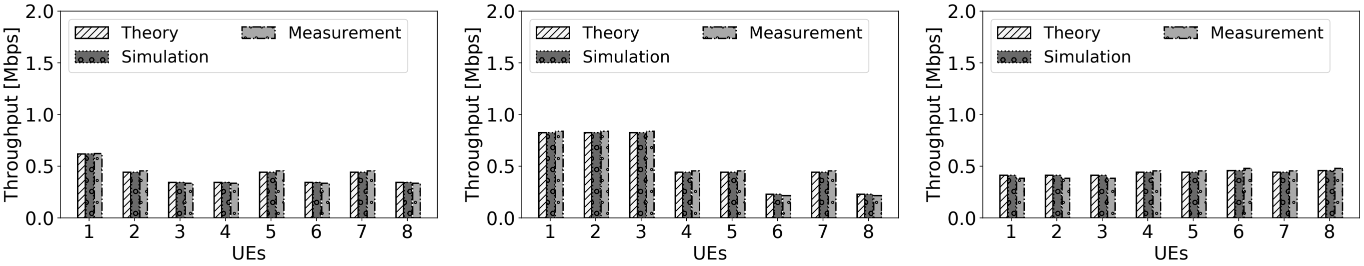
6.3 Validations

Some of the analytical results obtained in this paper for throughput and resource utilization are validated in this subsection.

Apart from realistic measurements provided with our OpenAirInterface setup, we further verify our theoretical findings with simulations. In order to do that, the input simulation parameters need to match with those of the measurements.

[7] Evaluations with different configurations have been conducted with conclusions remaining unchanged compared to the presented results. Due to space limitations, we omit showing other results.

[8] Higher subcarrier frequencies pertain to the mmWave communications, which are not considered in this work.

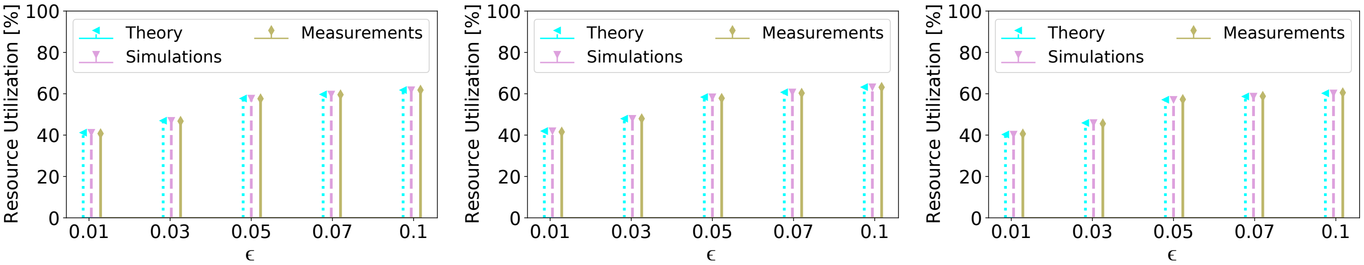


(a) ϵ = 0.01

(b) ϵ = 0.03

(c) ϵ = 0.05

Fig. 3: Validating the results for the throughput with **RR-ES** (a), **RR-P** (b), and **RR-IP** (c) for different values of ϵ.



(a) **RR-ES**

(b) **RR-P**

(c) **RR-IP**

Fig. 4: Validating the results for the average resource utilization with **RR-ES** (a), **RR-P** (b), and **RR-IP** (c) for different ϵ.

TABLE 3: Sum rate (Mbps) for the scenarios of Fig. 3

Policy	Theory	Simulations	Measurements
RR-ES ($\epsilon = 0.01$)	3.32	3.32	3.33
RR-P ($\epsilon = 0.03$)	4.26	4.26	4.31
RR-IP ($\epsilon = 0.05$)	3.48	3.48	3.47

The simulation environment consists of a single BS and eight UEs, whose CQI distributions are given in Table 2, from which also the conversion between the UE CQI and per-PRB rate is obtained. In total, there are 25 PRBs shared among UEs. For every UE within the BS, a list that contains a CQI value for the UE according to its CQI distribution is created. Since the BS possesses the information about all CQIs at each point and depending on the scheduling policy, the resources are distributed among UEs. The final throughput is calculated based on the CQI in the list entry and the applied policy.

Once the measurements and simulations have been performed for each ϵ, the corresponding throughput of each UE is assessed and then compared with the corresponding theoretical values. The list of the measured and simulated throughput for each UE is sorted in descending order. Afterward, for each ϵ, the respective point within the list is captured for each policy.

Fig. 3 shows the results for the achievable rate with **RR-ES** (for $\epsilon = 0.01$), **RR-P** ($\epsilon = 0.03$), and **RR-IP** ($\epsilon = 0.05$) for all eight users. The analytical results are obtained using (31), (35), and (39) accordingly for the corresponding policy. As can be observed, there is a good match among analytical, simulation, and measurement results in all cases. Table 3 summarizes the sum throughput for these scenarios.

Having demonstrated the effectiveness of our theoretical model with measurements and simulations for the resource-reservation policies in terms of throughput, we further evaluate each policy's resource utilization in Fig. 4. Similarly to the previous scenario, we provide not only simulation results but also measurements to verify the theoretical results, obtained using (13) accordingly, for each of the three policies with different outage probabilities. As can be observed, our theoretical values match closely the simulation and measurement results. Another observation is that increasing the outage leads to an increase in the utilization level. Due to space limitations, other validation results are not shown. In all cases, there is a match between the three types of results (theory, simulations, and measurements).

6.4 Resource-Reservation Policies

First, we look at the maximum values that can be obtained for the resource effectiveness of all eight users. From now on, the subcarrier frequency is 30 KHz. Fig. 5 shows the results for an outage of $\epsilon = 0.03$. Users 1-3 have the highest resource effectiveness, whereas users 6 and 8 have the lowest. The reason is that users 1-3 for at least 97% of the time have at least a per-PRB rate of r_7, while users 6 and 8 have for at least 97% of the time a per-PRB rate of at least r_5. If **RR-ES** is used, each user would have a data rate of $K/8$ times its resource effectiveness from Fig. 5. Fig. 6 shows the values of resource effectiveness when ϵ increases to 0.05. Now, user 3 alone has the highest resource effectiveness. Apparently, increasing ϵ can never decrease the resource effectiveness for a user.

Another important outcome is that for a given ϵ the user with the highest average per-PRB rate is not necessarily the user with the highest resource effectiveness. Namely, from Table 2 one can find that the average per-PRB rate for user 6 is $\mathbb{E}[R_6] = 1.07$ Mbps, whereas for user 7 it is $\mathbb{E}[R_7] = 0.92$ Mbps, i.e., $\mathbb{E}[R_6] > \mathbb{E}[R_7]$. On the other hand, from Fig. 5 and Fig. 6 one can see that $f_6 < f_7$. This implies that resource effectiveness is a function of the entire distribution of the per-PRB rate and ϵ, and not only of the first-order statistics of channel conditions.

Next, the impact of outage probability on the four resource-reservation policies - **RR-ES**, **RR-P**, **RR-IP**, and **RR-OPT** - is explored. Fig. 7 illustrates the results for the sum of the

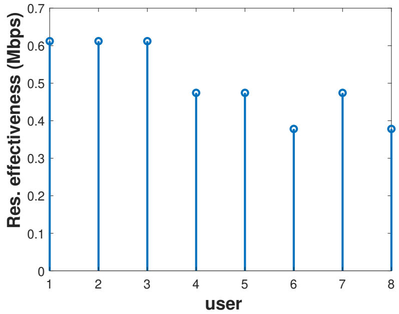


Fig. 5: Resource effectiveness f_i for $\epsilon = 0.03$.

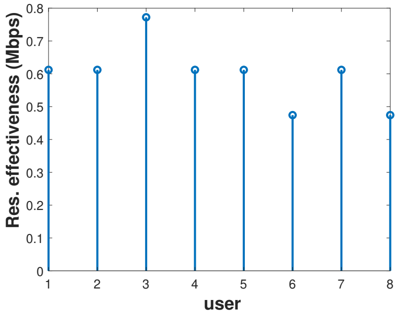


Fig. 6: Resource effectiveness f_i for $\epsilon = 0.05$.

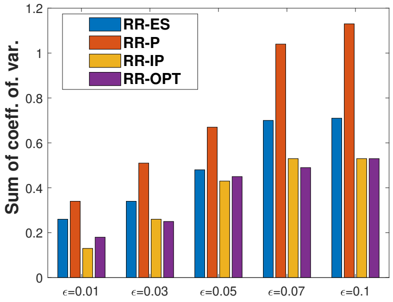


Fig. 7: The sum of the coefficients of variation with the resource-reservation policies.

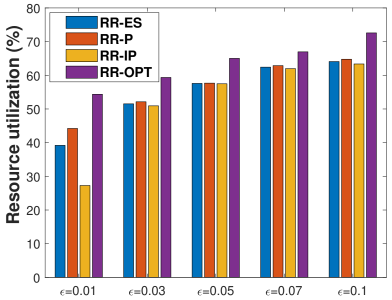


Fig. 8: Average resource utilization with the resource-reservation policies.

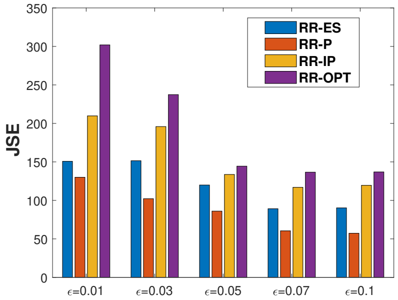


Fig. 9: Joint satisfaction efficiency with the resource-reservation policies.

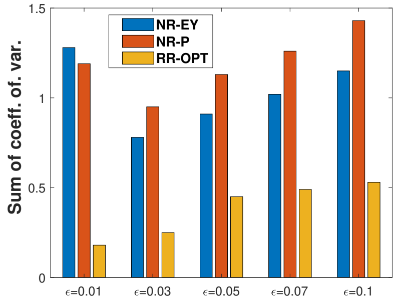


Fig. 10: The sum of the coefficients of variation with no-resource-reservation policies and **RR-OPT**.

coefficients of variation of data rates of each user. While a higher ϵ provides a higher U_i, it also increases the coefficient of variation, as more often there will be deviations from the constant value. This conclusion propagates across the four policies. These policies provide a very low sum of coefficients of variation. Note that there are eight users, so for instance when $\epsilon = 0.03$, each user will roughly have a c_V lower than 0.08 with **RR-P**, and even lower with other policies. This is a rate that is characterized by very low variance.

The resource utilization for the same policies and the same values of ϵ is investigated next. Fig. 8 shows the results (expressed in %). As can be observed from Fig. 8, increasing ϵ provides higher resource utilization. The reason is the higher U which requires more resources, whereas ϵ does not increase significantly. Regarding the best policy, the highest resource utilization, up to 72%, can be achieved using **RR-OPT**, which is expected as it is the solution to the optimization problem. However, the penalty when using one of the other policies is not large, especially for higher values of ϵ.

To complete the picture of resource efficiency for this approach from both the operator's and user's perspective, the attainable JSE values are investigated too. Fig. 9 portrays the results for the same policies in terms of JSE, corroborating that **RR-OPT** is indeed the solution to the optimization problem (20)-(22). As for the other three policies, from Fig. 8, the resource utilization is roughly the same (except for $\epsilon = 0.01$), and given the lower variability for **RR-IP** (see Fig. 7), the latter is the policy with the second highest JSE.

6.5 No-Resource-Reservation Policies

While the rate variability is very low with the reservation policies, the resource utilization is not satisfactorily high even with the optimal policy. Therefore, the performance of the policies with no reservation of network resources is looked at next. Fig. 10 illustrates the sum of the coefficients of variation vs. ϵ (the same eight users) for the two no-reservation policies, **NR-EY** and **NR-P**, as well as for the resource-reservation policy **RR-OPT**.

Note that, same as previously, increasing ϵ from 0.03 leads to a higher U_i and higher sum of coefficient of variation. In those frames in which the constant data rate cannot be provided, the resources are split equally among the users. An interesting outcome from Fig. 10 is that only for $\epsilon = 0.01$, **NR-P** provides a lower sum of c_V's. For higher values of ϵ, **NR-EY** performs better than the other no-reservation policy. The resource-reservation **RR-OPT** outperforms heavily both no-reservation policies in terms of lower variability. However, there is a price to pay for this outcome. Namely, regarding resource utilization, whose results are depicted in Fig. 11, it can be observed that both no-reservation policies perform almost identically, with insignificant differences across all considered values of ϵ but outperform **RR-OPT** on average by more than 10%. The level of utilization increases considerably with these policies, reaching a value as high as 85%, which is much higher than with the optimal resource-reservation policies (72%).

When it comes to JSE, **RR-OPT** outperforms both no-reservation policies considerably. This is shown in Fig. 12. The reason for such an outcome lies in the fact that the variability in data rates is much lower with **RR-OPT** than with no-reservation policies, and the advantage in terms of resource utilization of the latter is lower than the variability penalization (from the definition of JSE).

So far, we have compared the sum of the coefficients of variations, the level of utilization of network resources, and JSE for the six policies (four resource-reservation and two no-resource-

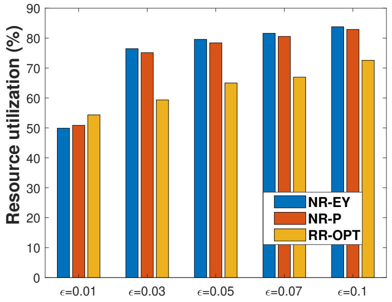


Fig. 11: Average resource utilization with **NR-EY**, **NR-P**, and **RR-OPT**.

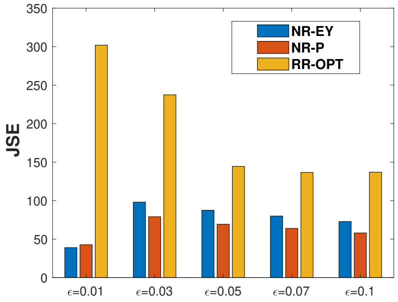


Fig. 12: Joint satisfaction efficiency with **NR-EY**, **NR-P**, and **RR-OPT**.

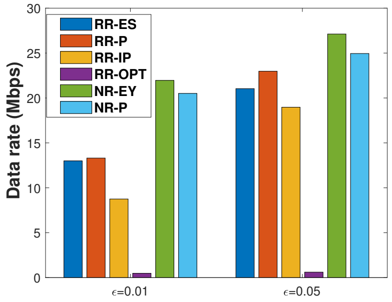


Fig. 13: The maximum rates for user 5 with reservation- and no-reservation policies.

reservation policies). But, what are the data rates provided by these policies? First, we of the users is picked (user 5), and for two of the outage probabilities, $\epsilon = 0.01$ and $\epsilon = 0.05$, the corresponding data rates with all the policies proposed in this paper are shown. Fig. 13 illustrates those results. Obviously, the two no-resource-reservation policies outperform all the resource-reservation policies. E.g., for $\epsilon = 0.01$, the data rate provided by **NR-EY** is at least 65% higher than with resource-reservation policies. For $\epsilon = 0.05$, the difference is at least 20%. It is noteworthy mentioning that **RR-OPT** performs the worst for user 5. The reason is that this user does not have the highest A_i among its peers (according to Result 2). Therefore, user 5 always receives only one PRB.

Motivated by the previous outcome, we further investigate the impact of different policies on the rate across all the users. This is done for $\epsilon = 0.05$. Table 4 shows the results. As can be observed, with **RR-OPT** user 3 experiences an extremely high rate (higher than 200 Mbps) because it has the highest A_i, receiving almost all PRBs, whereas the other users receive only one PRB each. Therefore, despite providing the highest utilization and JSE among the policies with (almost always) constant rates, **RR-OPT** is beneficial only for the user with the highest A_i. The performance of other users suffers severely under this policy. As expected, with **RR-IP** all the users experience the same throughput. The third outcome is that no-resource-reservation policies provide higher rates for almost all users compared to the resource-reservation policies.

6.6 Comparisons with Other Approaches

One of the policies from each approach is chosen for comparison with other state-of-the-art resource allocation schemes. From resource-reservation policies, **RR-ES** is selected. The reason for not choosing one of the three other policies, including **RR-OPT**, is that **RR-ES** provides the best trade-off between data rates across all the users and resource effectiveness, expressed through the sum of c_V's, resource utilization, and JSE. From no-resource-reservation policies, **NR-EY** is chosen as it provides lower variability (in the majority of the cases) and higher rates than **NR-P** for most of the users.

The performance of **RR-ES** and **NR-EY** is compared against some benchmark models. For the latter, these state-of-the-art policies are chosen:

- Non-consistent round-robin policy [24], [25]: Every user will receive $\frac{1}{n}$ of network resources in every frame with no consistent rates guaranteed but with full utilization. Obviously, the resource utilization, in this case, is 100%.
- The same consistent rate to everyone [6].

TABLE 4: Data rates with different policies for $\epsilon = 0.05$

User	RR-ES	RR-P	RR-IP	RR-OPT	NR-EY	NR-P
1	21.04	21.55	20.1	0.61	31.57	21.36
2	21.04	21.55	20.1	0.61	28.47	23.78
3	26.55	34.3	20.1	206.96	27.25	21.08
4	21.04	21.55	20.1	0.61	27.25	26.25
5	21.04	21.55	20.1	0.61	27.38	24.87
6	16.3	12.94	20.1	0.47	27.25	29.03
7	21.04	21.55	20.1	0.61	27.38	24.87
8	16.3	12.94	20.1	0.47	27.25	29.03

- Reallocating unused resources to the same users after providing the *same* consistent rate to everyone, such that unused resources are split equally among the users [7], [23].
- **Best-CQI** [36]: The user with the highest CQI in a frame receives all the resources (or they are split equally if two or more users have the same highest CQI).
- 5**th percentile rate** [35]: It guarantees that the data rate across all users and all frames is in 95% of the realizations higher than the given value.

Similar to the first benchmark, in the third and fourth benchmark models the resource utilization is 100% as well. So, for these policies the focus is only on the coefficient of variation.

Fig. 14 shows the results for the sum of the coefficients of variation of all the users vs. ϵ for the aforementioned policies, **RR-ES**, and **NR-EY**. As can be observed from Fig. 14, **RR-ES** provides the lowest values, outperforming the other policies by several times. An order of magnitude higher values were achieved with **Best-CQI** [36]; specifically, the sum of the coefficients of variation is 21.79. We do not show this result in Fig. 14. As far as 5**th percentile rate** [35] policy is concerned, the sum of the coefficients of variation for this scenario was found to be 0.51, which is still higher than for low values of ϵ with **RR-ES**.

On the other hand, the value of JSE with **Best-CQI** obtained via simulation was 4.59, which is almost two orders of magnitude lower than with **RR-OPT** (see Fig. 12). Higher values of JSE are achieved with 5**th percentile rate** policy; specifically, 155.93. The reason lies within the low values of coefficient of variation achieved with the latter. However, this value, as can be observed from Fig. 12 is lower than the highest value of **RR-OPT**.

Finally, Fig. 15 shows the utilization of network resources for **RR-ES**, **NR-EY**, and the same-consistent-rate-to-everyone policy [6]. For the other policies, the results are not shown here because all the resources are fully utilized. The results from Fig. 15 show that **NR-EY** outperforms the other two policies for all the values of ϵ, reaching a utilization level of up to almost

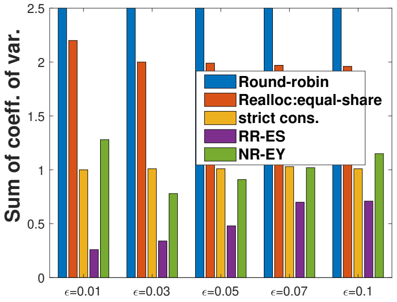


Fig. 14: The sum of the coefficients of variation with different approaches.

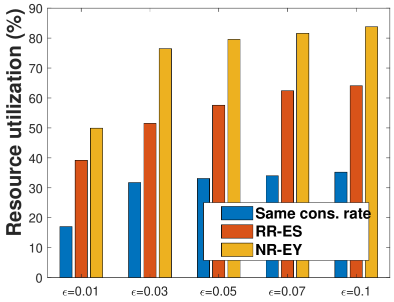


Fig. 15: Average resource utilization with different approaches.

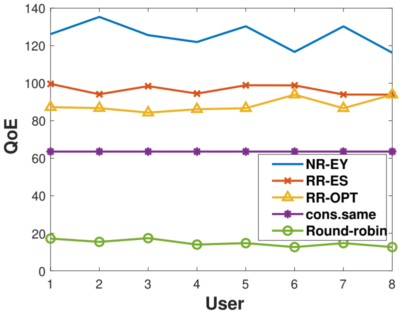


Fig. 16: QoE for mobile users with real-time video streaming.

85%. While the reallocation policy utilizes 100% of the resources, it leads to poor performance in terms of data rate variability (as shown in Fig. 14), being close to the policy with no consistency in rate provisioning (round-robin). Also, it is worth mentioning that the maximum achievable constant rate with the policy in which every user receives the same consistent rate is 6.54 Mbps for $\epsilon = 0.01$ and 12.71 Mbps for $\epsilon = 0.05$. Compared to the rates from Fig. 13 and Table 4, these values are much lower than any of the rates the policies proposed in this paper would provide, except for **RR-OPT** that rewards the user with the highest A_i, whose data rate outperforms that of [6] by almost two orders of magnitude. This corroborates the prediction that providing different constant rates, which we do in this paper, is more efficient than providing the same constant rate to everyone.[9]

The reason why **NR-EY** performs much better than **RR-ES** in terms of resource utilization and data rates is the dynamic adjustment feature of the number of allocated PRBs depending on the channel conditions in every frame, whereas **RR-ES** does not have that flexibility in allocating resources.

To summarize, it is up to the operator to decide whether the choice would be a resource-reservation or no-resource-reservation policy. If the goal is to increase the level of resource utilization (and provide higher rates to users), **NR-EY** should be the right choice. If, on the other hand, users are very conservative in terms of the rate variability, **RR-ES** is to be chosen as the resource allocation policy.[10]

6.7 Practical Scenario

Finally, the advantages offered by our approach in a practical setup are shown. The focus is on the use case of real-time video streaming. The performance of the eight users from Table 2 is considered. A high video resolution, or equivalently a high and stable playout rate, with very few video stalls (due to rebuffering events) and little loss of information (due to the buffer being full and discarding new packets/chunks) is one of the main indicators of high QoE for mobile users when streaming live video. The other important indicator for a good QoE is the small buffer size (resulting in low latency of playing the video from the occurrence of the event) needed to support that high playout rate. The latter is achieved if the variation in data rates is kept small (low coefficient of variation). Therefore, as the QoE metric in this setup is considered the ratio of the playout rate that can be guaranteed and the size of the buffer such that no more than 5% of the packets are lost and no more than 5% of the time the video stalls, i.e.,

$$QoE_i = \frac{U_{p,i}}{B_i}, \tag{69}$$

where $U_{p,i}$ denotes the (constant) playout rate of user i, whereas B_i is the required buffer size.

Performance should be compared when using our **RR-ES**, **NR-EY**, and **RR-OPT** against two of the already mentioned benchmarks: round-robin and the policy that provides the same consistent rate to everyone [6]. Fig. 16 portrays the results for the QoE of the eight users.

As can be observed from Fig. 16, the highest QoE is provided by our **NR-EY** policy. The reason is that **NR-EY** provides data rates with low variability, resulting in small buffer sizes, and high resource utilization, leading to high data rates and subsequently, to high playout rates. While **RR-ES**, as already shown, provides very low rate-variability, the data rates achieved by this policy, due to lower resource utilization levels, are considerably lower than with **NR-EY** (see Fig. 13 and Table 4). **RR-OPT** performs worse than **NR-EY** and **RR-ES** because all the users, other than the one with the highest value of A_i (see (15)), experience very low data rates, whereas user 3 (with the highest A_i) does not experience considerably lower rate variability than with **RR-ES** and **NR-EY**. The QoE with the policy in which everyone experiences the same consistent rate [6] is lower, mostly because users experience low data rates (12.71 Mbps in this scenario), whereas their buffer sizes are not lower than with **RR-ES** and **NR-EY**, due to the higher coefficient of variation of their data rates (see Fig. 14 for $\epsilon = 0.05$). On the other hand, while the round-robin policy [24] indeed provides higher data rates, especially for users with better channel conditions (the average can go up to 43 Mbps in this scenario), the variance of data rates is much higher with this policy, resulting in considerably larger buffers needed to support those high playout rates. Consequently, the QoE values with round-robin are lower than with the other four policies considered in this scenario.

7 Conclusion

This paper considered the problem of minimizing the rate variability across all mobile users in a cell while effectively utilizing network resources. To that end, two approaches were considered. In the first one, every user is allocated fixed resources (resource reservation). The second approach is dynamic, and resources are allocated differently across frames (no resource reservation). For each of the approaches, different allocation policies were proposed, and the corresponding (almost at all times) constant data

[9] The average resource utilization with 5**the percentile rate** is lower than 70%.

[10] When running the simulations, the results were obtained on the order of μs.

rates and the level of resource utilization were derived. For the resource-reservation approach, four policies were considered in total, one of which (**RR-OPT**) was derived as the solution to two optimization problems, with the objective to maximize the level of resource utilization and to maximize joint satisfaction efficiency. The other three resource reservation policies allocate resources proportionally to the resource effectiveness of users (**RR-P**), assign resources equally to all the users (**RR-ES**), and allocate resources inversely proportionally to user's resource effectiveness (**RR-IP**). For the no-resource reservation approach, two policies were considered. In the first policy, all the users receive the same frame ratio on average (**NR-EY**), whereas in the second, users will have data rates proportionally to their average per-PRB rates (**NR-P**). The performance of these policies was compared in terms of the sum of the coefficients of variation of data rates of all the users, the average utilization of network resources, and the joint satisfaction accuracy (which combines the previous two aspects). The results were validated using measurements conducted on OpenAirInterface and simulations run on a 5G dataset. Results show that our policies outperform other state-of-the-art approaches considerably by minimizing the rate variability and reducing the level of wasted resources. Also, we corroborate that no-reservation policies provide higher data rates and higher resource utilization than resource-reservation policies but at the expense of higher variability in data rates.

As part of our future work, we plan to consider the problem of minimizing the coefficient of variation of data rates in mmWave networks and to consider the general problem of providing α-fairness across all the users in the cell.

References

[1] F. Mehmeti and T. F. La Porta, "Minimizing rate variability with effective resource utilization in 5G networks," in *Proc. of ACM MobiWac*, 2021.

[2] "5G vision and requirements," 2014. white paper.

[3] 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.

[4] "5G radio access." www.ericsson.com/res/docs/whitepapers/wp-5g.pdf, 2016. Ericsson white paper, Uen 284 23-3204 C.

[5] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[6] 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.

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

[8] 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.

[9] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, 2014.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] F. Mehmeti and T. F. La Porta, "Resource allocation for improved user experience with live video streaming in 5G," in *Proc. of ACM Q2SWinet*, 2021.

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

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

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

[23] F. Mehmeti and T. L. Porta, "Reducing the cost of consistency: Performance improvements in next generation cellular networks with optimal resource reallocation," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, 2022.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] J. Khan and L. Jacob, "Resource allocation for CoMP enabled URLLC in 5G C-RAN architecture," *IEEE Systems Journal*, vol. 15, no. 4, 2021.

[31] X. Ye and L. Fu, "Joint MCS adaptation and RB allocation in cellular networks based on deep reinforcement learning with stable matching," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, 2024.

[32] N. Moosavi, M. Sinaie, P. Azmi, P.-H. Lin, and E. Jorswieck, "Cross layer resource allocation in H-CRAN with spectrum and energy cooperation," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, 2023.

[33] S. Shamaei, S. Bayat, and A. M. A. Hemmatyar, "Interference-aware resource allocation algorithm for D2D-enabled cellular networks using matching theory," *IEEE Transactions on Network and Service Management*, 2023.

[34] Y. Qiao, Y. Niu, Z. Han, S. Mao, R. He, N. Wang, Z. Zhong, and B. Ai, "Joint optimization of resource allocation and user association in multi-frequency cellular networks assisted by RIS," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 1, 2024.

[35] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, 2021.

[36] A. Mamane, M. Fattah, M. E. Ghazi, M. E. Bekkali, Y. Balboul, and S. Mazer, "Scheduling algorithms for 5G networks and beyond: Classification and survey," *IEEE Access*, vol. 10, 2022.

[37] F. Mehmeti, T. F. La Porta, and W. Kellerer, "Efficient resource allocation with provisioning constrained rate variability in cellular networks," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2023.

[38] D. Kim, B. C. Jung, H. Lee, D. K. Sung, and H. Yoon, "Optimal modulation and coding scheme selection in cellular networks with hybrid-ARQ error control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, 2008.

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

[40] J. F. Bonnans, *Convex and Stochastic Optimization*. Springer, 2019.

[41] F. Mehmeti, A. Papa, W. Kellerer, and T. L. Porta, "Minimizing rate variability with effective resource utilization in cellular networks." https://sites.google.com/view/fidanmehmeti, 2023. Tech report.

[42] A. Oppenheim and A. Willsky, *Signals and Systems*. Prentice Hall, 1996.

[43] 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.

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

[45] “oai cn.” https://gitlab.eurecom.fr/mosaic5g/mosaic5g/-/wikis/tutorials/oai-cn.

[46] “mosaic5g-oai-sim.” https://gitlab.eurecom.fr/oai/openairinterface5g/-/tree/mosaic5g-oai-sim.

[47] A. Papa, P. Kutsevol, F. Mehmeti, and W. Kellerer, “Delphi: Computing the maximum achievable throughput in SD-RAN environments,” *IEEE Transactions on Network and Service Management*, vol. 20, no. 4, 2023.

[48] “l2 sim.” https://gitlab.eurecom.fr/mosaic5g/mosaic5g/-/wikis/tutorials/l2-sim.

[49] “iperf.” https://iperf.fr/.

Fidan Mehmeti received his graduate degree in Electrical and Computer Engineering from the University of Prishtina, Kosovo, in 2009. He obtained his Ph.D. 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.

Arled Papa completed his Bachelor of Science in Electronics Engineering at the Polytechnic University of Tirana, Albania in 2015. He received his Master of Science in Communications Engineering at the Technical University of Munich (TUM) in November 2017 with high distinction. In February 2018 he joined the Chair of Communication Networks at TUM as a research and teaching associate. His research focuses on the design and analysis of QoS, Network Slicing and Programmability of Software-Defined Radio Access Networks.

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 career, 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 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.

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.