Delay Fairness in 5G Networks with SD-RAN

Fidan Mehmeti
Chair of Communication Networks
Technical University of Munich, Germany
Email: fidan.mehmeti@tum.de

Wolfgang Kellerer
Chair of Communication Networks
Technical University of Munich, Germany
Email: wolfgang.kellerer@tum.de

***Abstract*—The possibility of decoupling the operation of control plane from data plane in RANs, which became possible with the introduction of Software-Defined Networks in 5G, brought a paradigm shift in cellular network operation. The key element that enables this is a centralized controller, located away from base stations. This yields increased flexibility in the functioning of cellular networks, resulting in considerable enhancements compared to classical pre-5G resource allocation approaches. However, so far the range these improvements span is known only in terms of throughput. The advantages in terms of other metrics and objectives, like *delay fairness*, are not yet known. Therefore, in this paper, we derive analytically the resource allocation policies that lead to different delay fairness definitions among the entities in an SD-RAN-enabled network and show the advantages compared to the classical pre-5G approaches. We do this for different scenarios. First, we consider the minimum potential delay fairness in the network. Then, we consider the min-max delay fairness among base stations, and also the min-max delay fairness among users. We evaluate performance extensively with input data from a dataset. The results indicate that the introduction of SD-RAN improves the objective value up to $6\times$ compared to policies without SD-RAN.**

***Index Terms*—SD-RAN, 5G, Minimum potential delay fairness, Min-max fairness.**

I. Introduction

In the previous generations of cellular networks, including 4G, both data plane and control plane operations were performed jointly in Base Stations (BSs). With the advent of Software Defined Networks (SDNs) [1] and their adaptation in Radio Access Networks (RANs), widely known as SD-RAN [2], decoupling the control plane operation from the data plane became possible for the first time in 5G networks. The control is transferred to centralized entities, known as SD-RAN controllers, which are usually not co-located with BSs. This novelty introduced a paradigm shift in the resource allocation process in cellular networks in particular, and how the latter operate in general.

This change in operation brings considerable benefits to cellular networks [2]–[4], with increased *flexibility* being among the main ones. This increased degree of flexibility arises from having a broader view of the network topology, facilitated by the centralized SD-RAN approach. This way, depending on the current spread of users across BSs, and their channel conditions for which users periodically inform their serving BSs [5] that forward those data to the SD-RAN controller, the latter can allocate resources to BSs according to a given policy. In the second step then, BSs perform the resource assignment process across users within their region of operation. As a consequence, exploiting the wide network knowledge leads to performance improvements by allowing optimal assignment decisions, depending on the objective of interest. As opposed to SD-RAN, in a classical RAN, each BS is pre-assigned its own set of resources (frequencies) and allocates them to the users receiving service from it.

While the improvements that SD-RAN brings in terms of throughput and different forms of throughput fairness have already been documented [6], [7], [8], little is known on whether there are improvements on the delay at all compared to the traditional resource allocation approaches. Some open-source SD-RAN prototypes, like FlexRAN [2] and 5G-EmPOWER [3] already exist, but they do not provide an answer on the benefits the delegation of traditional RAN functions to centralized controllers brings in terms of delay. Moreover, delays can be associated with individual users, BSs, or the entire network. There may be other objectives related to delay, like providing delay fairness across different entities. To our best knowledge, these problems have not been tackled before in the context of SD-RAN.

Deriving and implementing resource policies in cellular networks that are delay-fair is quite strenuous mostly because of the varying nature of wireless channels, stemming from the mobility of the users and effects characteristic to wireless communications, like shadowing [9]. This dynamic channel behavior propels the need to vary the amount of assigned resources at the same granularity level at which the channel characteristics change, and also to consider the channel conditions of all users when making allocation decisions.

Some of the important research and practical questions that arise relating to provisioning delay fairness in SD-RAN-led 5G networks are:

- Which allocation policy minimizes the overall delay in the network, also known as *minimum potential delay fairness*, where the number of users, their channel conditions, and their association to BSs are known beforehand?
- What is the allocation policy that provides delay fairness among BSs, i.e., which minimizes the highest delay in a BS in the network?
- If the goal is to minimize the delay of the worst-performing user network-wide, which policy enables this?

This work was supported by the Federal Ministry of Education and Research of Germany (BMBF) under the projects "6G-Life" and "6G-ANNA" with project identification numbers 16KISK002 and 16KISK107.

To answer the aforementioned open questions, here we formulate and solve three optimization problems. In the first one, the goal is to minimize the total delay in the network when users send/receive packets. This turns out to be equivalent to minimizing the potential delay fairness, derived as a special case of the general Network Utility Maximization (NUM) problem. We show that minimum potential delay fairness is achieved if resources are allocated inversely proportionally to the square root of the channel conditions of the user in a slot.[1] The second problem we solve in this paper is to provide delay fairness among BSs. It is in fact a min-max problem in which the goal is to minimize the delay in the worst-performing BS, i.e., the BS where the sum of delays of users when transmitting data is the highest in the network. The solution to this problem results in an optimal allocation policy in which the amount of allocated resources should be again inversely proportional to the square root of the user's channel conditions, but as opposed to the first problem the impact of other users is described with a different function. Finally, in the third problem, we look at minimizing the delay of the worst-performing user in the network. The optimal solution in this case, as opposed to the previous two problems, yields an inverse proportionality of the number of assigned resources to the channel conditions (not the square root). The results we provide in this work are especially helpful for the cellular operator as they can provide an exact prediction of the delay a user can expect to experience given the network topology in a slot, without penalizing users with bad channel conditions. The main message of this paper is that the use of SD-RAN can improve performance considerably under any number of BSs and users.

Specifically, our main contributions are:

- We derive the allocation policies which provide minimum potential delay fairness among all users, min-max delay fairness among BSs, and min-max delay fairness among all users, given the channel characteristics of the users and their spread across BSs.
- We evaluate the performance using input data from real-life 5G traces.
- We show the concrete performance improvements when using SD-RAN compared to the traditional approach in terms of different delay-related fairness objectives.

The remainder of this paper is organized as follows. The system model and the problem formulations are presented in Section II. This is followed by the solutions to the optimization problems in Section III. Section IV introduces the benchmark models against which the performance of SD-RAN is compared. In Section V, we evaluate the performance and provide some additional insights. Some related work is discussed in Section VI. Finally, Section VII concludes the paper.

II. Performance Modeling

First, we introduce the system model and then define three optimization problems that we solve in this paper.

[1]As will be seen later, the quantity known as per-PRB rate is used to describe the rate of a user per unit of allocated resources.

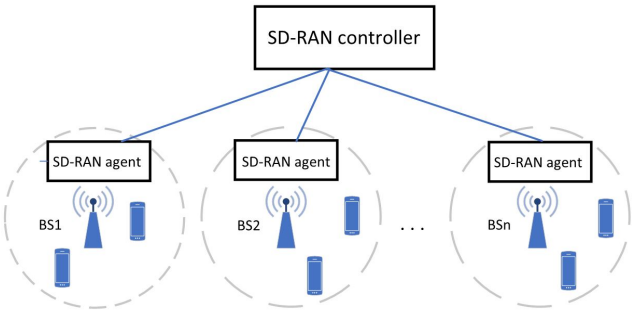

Fig. 1. Illustration of the SD-RAN environment.

A. System model

We consider an SD-RAN-led network (Fig. 1) with a single controller responsible for assigning resources (and the allocation decisions further to users) to BSs. For every BS there is an SD-RAN agent that communicates with the controller [10], using the Transport Control Protocol (TCP). We denote the set of all BSs by $\mathcal{N}$. There are in total $n = |\mathcal{N}|$ BSs in the coverage area of the controller. We denote by $\mathcal{M}_i$ the set of all users within the operational area of BS i, where $m_i = |\mathcal{M}_i|$ is the number of users in BS i. So, there are in total $\sum_{i=1}^{n} m_i$ users.

5G uses *Physical Resource Blocks (PRBs)* as the unit of allocation on a per-slot basis [11]. Each PRB consists of 12 subcarriers. The slot duration is a function of the subcarrier spacing. Specifically, if the subcarrier spacing is 15 KHz (PRB width of 180 KHz), the slot duration is 1 ms. If the subcarrier spacing is 30 KHz (PRB width of 360 KHz), the corresponding slot duration is 0.5 ms. The slot duration decreases further ($2\times$) when switching to subcarrier spacing of 60 KHz, and another $2\times$ when switching to 120 KHz [5]. Different PRBs are assigned to different users in a slot. The assignment varies across slots. Therefore, scheduling needs to be performed along two dimensions, *time* and *frequency*. In total, there are K available PRBs for n BSs.

Users experience different channel conditions (characterized by the Channel Quality Indicator (CQI) with discrete values from 1 (worst channel conditions) to 15 (excellent channel conditions)) across different PRBs even within the same slot. Because of the mobility and time-varying nature of the channels, per-PRB CQI (which is a function of Signal-to-Interference-Plus-Noise-Ratio (SINR)) changes from one slot to another, whose value depending on the Modulation and Coding Scheme (MCS) used sets the per-PRB rate. To keep the analysis tractable, we make a simplifying assumption. Namely, we assume that the BS splits the transmission power equally among all PRBs it transmits on and that the channel characteristics for a user remain static across all PRBs (identical CQI over all PRBs for a given user), but change randomly (according to some distribution) from one slot to another, and are mutually independent among users (users with heterogeneous channel conditions). These assumptions reduce the resource allocation problem to the number of allocated PRBs and not to which PRBs are assigned to a user.

The previous assumptions imply that in every slot user (i,j)[2], where $i \in \mathcal{N}$ and $j \in \mathcal{M}_i$, will have a per-PRB rate (i.e., the rate each assigned PRB brings to the user) that can be modeled with a discrete random variable, $R_{i,j}$, with values in $\{r_1, r_2, \ldots, r_{15}\}$, such that $r_1 < r_2 < \ldots < r_{15}$, with Probability Mass Function (PMF) $p_{R_{i,j}}(x)$, which is a function of user's (i,j) CQI over time.

B. Problem formulation

Every user sends periodically the information about its CQI to its serving BS. Then, every BS collects all the CQI information from the users in its area and forwards them to the SD-RAN controller (see Fig. 1). Based on the CQI values from all the BSs (and hence all users), the controller then, depending on the allocation policy used, decides on the number of PRBs to assign to each BS in a slot. Then, from the PRBs it receives, each BS further allocates those PRBs to the users in its area. Therefore, using SD-RAN, the resource allocation process is performed in two levels. First, among BSs, and then each BS allocates the PRBs it received from the controller to its users.

Let $K_{i,j}, \forall j \in \mathcal{M}_i$, denote the number of PRBs user j gets from BS i.[3] If $K_i, \forall i \in \mathcal{N}$, denotes the number of PRBs that BS i receives from the controller in a slot, then $K_i = \sum_{j=1}^{m_i} K_{i,j}$. The data rate of user (i,j) in a slot is $K_{i,j}R_{i,j}$.

The delay user (i,j) experiences when transmitting a packet of size Δ is $\frac{\Delta}{K_{i,j}R_{i,j}}$. In the analysis to follow, we assume that all users transmit packets of equal sizes. Therefore, w.l.o.g. we assume that packets are of unit sizes. Hence, the per-packet delay for user (i,j) is $\frac{1}{K_{i,j}R_{i,j}}$. In Section III-D, we provide a short discussion on the scenario when users transmit packets of different sizes.

1) Minimum potential delay fairness across users: In the first scenario, our goal is to minimize the total delay across the entire set of users in the SD-RAN-enabled network. This leads to the following optimization problem:

$$\mathcal{P}_1: \quad \min_{K_{i,j}} \quad \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{1}{K_{i,j}R_{i,j}} \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{i,j} \leq K, \tag{2}$$

$$K_{i,j} \geq 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{3}$$

Constraint (2) expresses the total number of PRBs that can be allocated to all users (which is K), whereas (3) captures the fact that the number of allocated PRBs to users should be non-negative. The decision variables are $K_{i,j}$.

Note that in the general Network Utility Maximization framework [12] the objective for $\alpha \neq 1$ is $\max \sum_i \frac{x_i^{1-\alpha}}{1-\alpha}$. When $\alpha = 2$, corresponding to minimum potential delay fairness, this objective reduces to $\max \sum_i -\frac{1}{x_i}$, which is equivalent to the objective $\min \sum_i \frac{1}{x_i}$. Hence, we refer to this problem as *minimizing potential delay fairness* across all users.

[2]We denote every user with the ordered pair (i,j), where i denotes the BS, and j indicates the user receiving service by that BS.

[3]Each user can receive resources only from one BS.

2) Delay (min-max) fairness among BSs: In the second scenario, the goal is to look at fairness among BSs. Namely, resources should be allocated in such a way that the total transfer delay among the users in a BS is not much higher than among the users in another BS. To capture this requirement, we formulate an optimization problem where the objective is to minimize the maximum BS delay (where the delay in a BS is the sum of delays across all users in that BS) in the entire network, i.e., to minimize the total delay of the worst-performing BS:

$$\mathcal{P}_2: \quad \min_{K_{i,j}} \max \quad \sum_{j=1}^{m_i} \frac{1}{K_{i,j}R_{i,j}} \tag{4}$$

$$\text{s.t.} \quad (2), (3).$$

The function in objective (4) denotes the total transfer delay in a BS. The decision variables are again $K_{i,j}$. So, here we are dealing with a min-max optimization problem.

3) Delay (min-max) fairness among users: Another interesting objective is to minimize the delay of the worst-performing user in the network. This would be of interest, for example, in a scenario in which all the packets have to be received within a time window and the latter to be as narrow as possible. This task translates into the following optimization problem:

$$\mathcal{P}_3: \quad \min_{K_{i,j}} \max \quad \frac{1}{K_{i,j}R_{i,j}} \tag{5}$$

$$\text{s.t.} \quad (2), (3).$$

We solve these optimization problems in the next section.

III. Performance optimization

In this section, first, we determine the optimal policy for minimum potential delay fairness by solving $\mathcal{P}_1$. We proceed then with the solution to $\mathcal{P}_2$ and $\mathcal{P}_3$. Finally, we provide a short discussion on packets of different sizes for different users.

A. Minimum potential delay fairness across users

We start the analysis by solving $\mathcal{P}_1$. Objective function (1) is concave. Namely, the main diagonal elements of the Hessian matrix $\mathbf{A}$ are $\frac{\partial^2}{\partial K_{i,j}^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{K_{i,j}R_{i,j}}\right) = \frac{2}{R_{i,j}K_{i,j}^3} > 0$, whereas the off-diagonal elements are all 0. This implies that for any non-zero vector $\mathbf{x}$, the following holds always

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0,$$

which is the condition satisfied by convex functions. Furthermore, as constraints (2) and (3) are linear, there exists a unique solution to $\mathcal{P}_1$. As a first step in solving this convex optimization problem, we define the Lagrangian function as

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{K_{i,j}R_{i,j}} - \lambda\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}K_{i,j} - K\right) + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\mu_{i,j}K_{i,j}, \tag{6}$$

where $\lambda \geq 0$ and $\mu_{i,j} \geq 0$, $\forall j \in \mathcal{M}_i$ are the slack variables. It can be shown in a straightforward fashion that $\mathcal{P}_1$ satisfies Slater's condition [12]. Hence, the strong duality holds. Therefore, Karush-Kuhn-Tucker (KKT) conditions [13] can be applied to the dual optimization problem, where the optimal solution needs to fulfill the following system of equations:

$$\frac{\partial \mathcal{L}}{\partial K_{i,j}} = 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i, \tag{7}$$

$$\lambda \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{i,j} - K \right) = 0, \tag{8}$$

$$\mu_{i,j} K_{i,j} = 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{9}$$

Substituting (6) into (7), we obtain

$$\frac{1}{R_{i,j} K_{i,j}^2} - \lambda + \mu_{i,j} = 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i, \tag{10}$$

or equivalently,

$$\lambda = \frac{1}{R_{i,j} K_{i,j}^2} + \mu_{i,j}, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{11}$$

From the objective (1), we can infer that $K_{i,j} > 0$, which together with (9) yields $\mu_{i,j} = 0$. Replacing the latter finding into (11) leads to

$$\lambda = \frac{1}{R_{i,j} K_{i,j}^2} > 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{12}$$

From (12) and (8), we obtain

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{i,j} = K. \tag{13}$$

This is an expected finding, as to improve the performance in terms of delay the resources need to be fully utilized. In order to derive the amount of allocated resources, we express the term for user (i, j) in (12) through that of user $(1, 1)$, to get

$$R_{i,j} K_{i,j}^2 = R_{1,1} K_{1,1}^2, \tag{14}$$

yielding

$$K_{i,j} = \sqrt{\frac{R_{1,1}}{R_{i,j}}} K_{1,1}. \tag{15}$$

Substituting (15) into (13) and performing some simple algebra, we obtain

$$K_{1,1} = \frac{K}{\sqrt{R_{1,1}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}}}. \tag{16}$$

Finally, substituting (16) into (15), we have:

Result 1. *A minimum potential delay fair allocation policy across all users in the network with SD-RAN is achieved if the number of assigned PRBs to user (i, j) follows the policy*

$$K_{i,j} = \frac{K}{\sqrt{R_{i,j}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}}}. \tag{17}$$

The most important observation to make from Result 1, besides that of the policy being dynamic, is the inverse proportionality between the square root of the channel conditions of the user and the amount of allocated resources; the worse the channel conditions of a user (lower $R_{i,j}$), the higher the $K_{i,j}$, and vice versa.

B. Min-max delay fairness among BSs

When it comes to solving $\mathcal{P}_2$, we need to transform it first to an equivalent problem:

$$\min_{K_{i,j}} \quad Z \tag{18}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{i,j} = K, \tag{19}$$

$$K_{i,j} \geq 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i, \tag{20}$$

$$Z \geq \sum_{j=1}^{m_i} \frac{1}{K_{i,j} R_{i,j}}, \quad \forall i \in \mathcal{N}. \tag{21}$$

Note that constraint (19) is strict equality now. We made this change because besides being interested in providing fairness, the users would also be interested to have satisfying performance (lower delay), which implies full utilization of network resources (also in line with the discussion of the solution to $\mathcal{P}_1$).

The Lagrangian for this optimization problem is

$$\begin{aligned} \mathcal{L} &= -Z - \lambda \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{i,j} - K \right) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mu_{i,j} K_{i,j} - \sum_{i=1}^{n} \theta_i \left(\sum_{j=1}^{m_i} \frac{1}{K_{i,j} R_{i,j}} - Z \right) \end{aligned} \tag{22}$$

where $\lambda \geq 0$, $\mu_{i,j} \geq 0$, $\forall i, j$, and $\theta_i \geq 0, \forall i$ are the slack variables. Applying KKT conditions results in

$$\frac{\partial \mathcal{L}}{\partial K_{i,j}} = -\lambda + \mu_{i,j} + \frac{\theta_i}{R_{i,j} K_{i,j}^2} = 0, \tag{23}$$

$$\frac{\partial \mathcal{L}}{\partial Z} = -1 + \sum_{i=1}^{n} \theta_i = 0, \tag{24}$$

$$\lambda \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{i,j} - K \right) = 0, \tag{25}$$

$$\mu_{i,j} K_{i,j} = 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i, \tag{26}$$

$$\theta_i \left(\sum_{j=1}^{m_i} \frac{1}{K_{i,j} R_{i,j}} - Z \right) = 0, \quad \forall i \in \mathcal{N}. \tag{27}$$

For the same reasons as when solving $\mathcal{P}_1$, $K_{i,j} > 0$, so from (26) we have $\mu_{i,j} = 0$. Replacing the latter into (23), we get

$$\lambda = \frac{\theta_i}{R_{i,j} K_{i,j}^2}. \tag{28}$$

As $\lambda > 0$, from (28) we obtain $\theta_i > 0$. With the previous finding, from (27) we get

$$\sum_{j=1}^{m_i} \frac{1}{K_{i,j} R_{i,j}} = Z, \quad \forall i \in \mathcal{N}. \tag{29}$$

From Eq.(29), minimizing the maximum total delay across BSs implies that *resources need to be assigned in the way that total delays are the same in all BSs.*

Next, observing (28) for a given BS i, we have that for all users within the serving area of that BS, it holds

$$R_{i,j} K_{i,j}^2 = \text{const.}$$

The previous expression is the same as the adjusted (14). If K_i denotes the total number of PRBs allocated to users in BS i, where $K_i = \sum_{j=1}^{m_i} K_{i,j}$, following similar reasoning as when obtaining (17), for the amount of resources that should be allocated to users of BS i, we have

$$K_{i,j} = \frac{K_i}{\sqrt{R_{i,j}} \sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}}}, \quad \forall j \in \mathcal{M}_i. \tag{30}$$

The next step is to establish the relationship between the allocated resources among BSs. To that end, substituting (30) into (29), after some simple algebra, we obtain

$$\frac{\left(\sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}} \right)^2}{K_i} = Z, \quad \forall i \in \mathcal{N}. \tag{31}$$

Next, expressing the general term K_i in terms of K_1 from (31), and replacing it into

$$\sum_{i=1}^{n} K_i = K, \tag{32}$$

solving the latter in K_1, and after a straightforward procedure for K_i we obtain

$$K_i = \frac{K \left(\sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}} \right)^2}{\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}} \right)^2}. \tag{33}$$

Finally, substituting (33) into (30), we have:

Result 2. *A min-max delay fair resource allocation policy across all BSs in the network with SD-RAN is achieved if the number of assigned PRBs to user (i,j) follows the policy*

$$K_{i,j} = \frac{K \sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}}}{\sqrt{R_{i,j}} \sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}} \right)^2}. \tag{34}$$

From Result 2 we can observe that the amount of allocated resources is inversely proportional to (the square root of) the user's channel conditions, and also inversely proportional to the channel conditions of the other users within the same BS (see the numerator of (34)).

C. Min-max delay fairness across users

The optimization problem, in this case, is $\mathcal{P}_3$. The objective and constraints are already well-known. Similarly to $\mathcal{P}_2$, we introduce the new variable Z, leading to the equivalent optimization problem:

$$\min_{K_{i,j}} \quad Z \tag{35}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{i,j} = K, \tag{36}$$

$$K_{i,j} \geq 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i, \tag{37}$$

$$Z \geq \frac{1}{K_{i,j} R_{i,j}}, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{38}$$

Note that here as well, for the same reasons as in $\mathcal{P}_2$, we impose the requirement that all PRBs must be allocated. Forming the Lagrangian, and using KKT conditions, in a similar vein as before, leads to

$$K_{i,j} R_{i,j} = \text{const}, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{39}$$

Using (39) and (36), after some simple calculus, we obtain:

Result 3. *A min-max delay fair resource allocation policy across all users in the network with SD-RAN is achieved if the number of assigned PRBs to user (i,j) follows the policy*

$$K_{i,j} = \frac{K}{R_{i,j} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{1}{R_{i,j}}}. \tag{40}$$

In this case, again, the better channel conditions the user experiences, the lower the number of PRBs needed. However, now the amount of PRBs is not inversely proportional to $\sqrt{R_{i,j}}$, but to $R_{i,j}$ instead. Interestingly, the same policy is obtained for max-min fairness in cellular networks with SD-RAN in terms of throughput [6].

D. Different packet sizes

In the analysis so far, we have assumed that all users have the same packet size. Hence, we normalized it to unity. In case users transmit packets of different sizes, then we can adapt each of the optimization problems by adding the corresponding weights to denote the packet sizes. Then, we would have *weighted delay fairness*. Nevertheless, the procedure for the solution of all the optimization problems would be the same, and the resource allocation results would be simply adjusted by the corresponding packet size for each user. Hence, due to space limitations, we omit further discussions on this.

IV. Benchmark models

In order to assess the performance of the SD-RAN-enabled network in terms of delay fairness, we need benchmark models (baselines). To that end, in this paper, we use two of them. The first suitable model is the one in which there is no SD-RAN, but where there are some delay fairness guarantees (in this case minimum potential delay fairness). Hence, we choose the baseline in which RAN operates in a classical way, where every BS is allocated its set of PRBs beforehand, and the allocation process undergoes minimum potential delay fairness within each BS separately. If K is the total number of PRBs

in the system, then w.l.o.g. we assume that each BS operates on $\frac{K}{n}$ PRBs, where n is the number of BSs.

In the no-SD-RAN setup, the optimization problem for BS i, whose solution guarantees minimum potential delay fair resource allocation to the users within its coverage area, can be formulated as:

$$\mathcal{P}_{0,1}(i): \quad \min_{K_{i,j}} \quad \sum_{j=1}^{m_i} \frac{1}{K_{i,j}R_{i,j}} \tag{41}$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} K_{i,j} \leq \frac{K}{n}, \tag{42}$$

$$K_{i,j} \geq 0, \quad \forall j \in \mathcal{M}_i. \tag{43}$$

Basically, for each BS we need to solve $\mathcal{P}_{0,1}(i)$ separately. The function in the objective is apparently concave. Namely, similar to $\mathcal{P}_1$, the main diagonal elements of its Hessian matrix are equal to $\frac{2}{R_{i,j}K_{i,j}^3} > 0$, whereas all the off-diagonal elements are 0, making the Hessian a positive definite matrix, resulting in a convex objective function [13]. Given also that the constraints are linear, there exists a solution to the problem, with the local optimizer being a global optimizer as well. The Lagrangian of this optimization problem is

$$\mathcal{L} = -\sum_{j=1}^{m_i} \frac{1}{K_{i,j}R_{i,j}} - \lambda \left(\sum_{j=1}^{m_i} K_{i,j} - \frac{K}{n} \right) + \sum_{j=1}^{m_i} \mu_{i,j} K_{i,j}, \tag{44}$$

where $\lambda \geq 0$ and $\mu_{i,j} \geq 0$, $\forall j \in \mathcal{M}_i$. It can be easily shown that $\mathcal{P}_{0,1}(i)$ satisfies Slater's condition [12]. Therefore, the strong duality holds in this case too. Therefore, KKT conditions can be applied to the dual optimization problem, and the optimal solution should satisfy the following conditions:

$$\frac{\partial \mathcal{L}}{\partial K_{i,j}} = 0, \quad \forall j \in \mathcal{M}_i, \tag{45}$$

$$\lambda \left(\sum_{j=1}^{m_i} K_{i,j} - \frac{K}{n} \right) = 0, \tag{46}$$

$$\mu_{i,j} K_{i,j} = 0, \quad \forall j \in \mathcal{M}_i. \tag{47}$$

Substituting Eq.(44) into Eq.(45), we obtain

$$\lambda = \frac{1}{R_{i,j}K_{i,j}^2} + \mu_{i,j}, \quad \forall j \in \mathcal{M}_i, \tag{48}$$

implying $\lambda > 0$, which in turn, from (46) results in the need for full utilization of network resources, i.e.,

$$\sum_{j=1}^{m_i} K_{i,j} = \frac{K}{n}, \tag{49}$$

as before. Further, since $K_{i,j} > 0$, from (47) we obtain $\mu_{i,j} = 0$. This reduces (48) to (12). Therefore, the remainder of the procedure is similar to the one which leads to the solution of $\mathcal{P}_1$. The difference is that we need to use (49) instead. Finally, for the optimal allocation policy, we have

$$K_{i,j} = \frac{K}{n\sqrt{R_{i,j}} \sum_{j=1}^{m_i} \frac{1}{\sqrt{R_{i,j}}}}, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{50}$$

The same conclusions as for Result 1 follow, except that now there is the factor $\frac{K}{n}$, instead of K, as the available resources are split equally among the BSs. Also, the summation has to be performed only across the users of the same BS (the denominator of (50)).

The previous benchmark model is not suitable for the optimization problems $\mathcal{P}_2$ and $\mathcal{P}_3$. Therefore, we use another baseline, in which the goal is to minimize the maximum delay in each cell, but without SD-RAN. In that case, each BS is pre-assigned its set of PRBs, which we assume is $\frac{K}{n}$. The optimization formulation for this baseline model is:

$$\mathcal{P}_{0,2}(i): \quad \min_{K_{i,j}} \max \quad \frac{1}{K_{i,j}R_{i,j}} \tag{51}$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} K_{i,j} \leq \frac{K}{n}, \tag{52}$$

$$K_{i,j} \geq 0, \quad \forall j \in \mathcal{M}_i. \tag{53}$$

Following a similar procedure as with the other optimization problems in this paper, we obtain the optimal solution for the resource allocation in this benchmark model as

$$K_{i,j} = \frac{K}{nR_{i,j} \sum_{j=1}^{m_i} \frac{1}{R_{i,j}}}, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}_i. \tag{54}$$

Having the benchmark models against which we can compare the results obtained with our approaches, we proceed next with assessing the performance under different policies.

V. Performance Evaluation

In this section, we describe the simulation setup first. Then, we compare the performance of our three approaches in SD-RAN-led networks; namely, the minimum potential delay fairness, min-max fairness across all BSs, and min-max fairness across all users, against the corresponding benchmark optimal policies without SD-RAN, and the equal-share policy again operating on a non-SD-RAN setup. We do this for different cases. This is followed by results on the impact of channel statistics on the resource allocation process, or more precisely, on the variability of the allocated resources over time.

A. Simulation setup

In this paper, we have used a 5G trace with data obtained in a measurement campaign conducted in the Republic of Ireland as input parameters. These datasets can be found in [14], with a detailed description in [15], and statistical analysis in [16]. From the trace data, the parameter of interest here is CQI with 15 levels, which is used to quantify the channel conditions of a user and determines the per-PRB rate of a user in a slot. These measurements were conducted for one user, but on different days (hence considering them as different users), for different services, and both in cases when the user is static and also when moving around in a vehicle. To mimic the dynamic nature of these users, we have picked 8 users that were moving around. Based on the frequency of occurrence of a per-PRB rate for every user, we obtained the corresponding per-PRB rate probabilities, which are shown in Table I.

TABLE I
PER-PRB RATES AND THE CORRESPONDING PROBABILITIES FOR EVERY USER FROM THE REPUBLIC OF IRELAND TRACE [15]

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

The subcarrier spacing is 30 KHz, with 12 subcarriers per PRB, making the PRB width 360 KHz. This incurs a slot duration of 0.5 ms. The total number of PRBs in the system is considered to be $K = 273$ [5]. The simulations were conducted in MATLAB R2022b.

In the simulator, every BS in each slot sends the information of the CQIs of its users to the SD-RAN controller in cases with SD-RAN. In the classical network setup without SD-RAN, users send the information of their CQIs to the associating BS, where the latter has a fixed set of PRBs. In the SD-RAN setup, with the full picture of all CQIs in the network, the controller according to the resource allocation policy used distributes the resources (PRBs) to BSs together with the information on how to further assign them to users in their coverage areas. Depending on the amount of resources assigned, and the user's per-PRB rate, we determine the data rate each user experiences in a slot, and consequently, the delay for transmitting a packet of a given size.

Unless stated otherwise, we show results for three cases:

- Case 1: 4 BSs; 2 users for BSs 1 and 2, 3 users for BSs 3 and 4.
- Case 2: 5 BSs; 2 users for BSs 1 and 2, 4 users for BSs 3 and 4, 6 users for BS 5.
- Case 3: 7 BSs; 2 users for BSs 1 and 2, 4 users for BSs 3 and 4, 6 users for BSs 5 and 6, 8 users for BSs 7.

Note that in all the cases, a user is chosen randomly from one of the eight types of Table I. Then, its CQI values across slots are taken from the trace of the corresponding user.

B. Performance comparisons

We start this section by comparing the performance obtained with one of our optimal resource allocation policies and the benchmark models. First, we compare our approach for minimum potential delay fairness across all users in an SD-RAN-led network (the solution to $\mathcal{P}_1$), to which from now on we refer to as SD-RAN in the plots, with the benchmark (the solution of $\mathcal{P}_{0,1}$) and another allocation policy (equal-share [17] of resources among the users in a BS in a non-SD-RAN setup).

We show results for the three aforementioned cases. Fig. 2 depicts the results for the sum of the delays across all users when transmitting unit-size packets (Eq.(1)) in the network with different policies for Case 1. As shown in Fig. 2, the solution to $\mathcal{P}_1$ always outperforms that of the benchmark $\mathcal{P}_{0,1}$ (marked as *no SD-RAN*), and equal-share allocation policy. The improvement is around 15%. As expected, the equal-share policy yields the worst results because is oblivious to the channel conditions of the users. On the other hand, as shown analytically, the optimum is achieved when users with worse channel conditions received more resources than those with good channel conditions. Note that we are showing results for only 30 slots to better emphasize visually the differences among the results with different policies.

Fig. 3 illustrates the results for Case 2, whereas Fig. 4 does it for Case 3. Similar to Fig. 2, on the y-axes we depict the total delay in the entire network for unit-size packets. In both scenarios, SD-RAN outperforms the other two approaches in terms of minimum potential delay fairness significantly. Note that as the number of BSs and users increases, the total delay is higher because users receive fewer resources, hence, introducing larger delays individually, and there are more of them. In Case 3, the total delay is more than an order of magnitude higher than in Case 1. It is worth mentioning that the values on the y-axes are always in seconds, while the packet size is 1 Mbit.

Next, we compare the results in terms of delay fairness among BSs (min-max fairness of total delay per BS). Fig. 5 shows the outcomes related to Case 1 for the value of the objective (4), i.e., the sum of delays per BS, with different policies, whereas Fig. 6 and Fig. 7 portray the results for Case 2 and Case 3, respectively. The parameters remain unchanged from the previous scenarios. Our approach, marked as SD-RAN in the plots, now uses the solution to $\mathcal{P}_2$. The benchmark, in this case, is the solution to $\mathcal{P}_{0,2}$, and marked as *no SD-RAN*, whereas the third policy is again the equal-share policy (without SD-RAN). In all three cases, SD-RAN outperforms the two no-SD-RAN approaches, up to $6\times$. In Figs. 5-7, the results for the equal-share policy are identical to the solution of benchmark $\mathcal{P}_{0,2}$, and hence cannot be discerned in the plot. To prove this, we substitute (54) into (4), and after rearranging we obtain

$$\sum_{j=1}^{m_i} \frac{1}{K_{i,j} R_{i,j}} = \frac{n m_i}{K} \sum_{j=1}^{m_i} \frac{1}{R_{i,j}}. \tag{55}$$

On the other hand, with the equal-share policy, user (i, j)

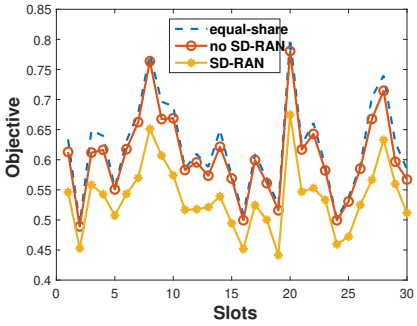

Fig. 2. The evolution of objective (1) for Case 1.

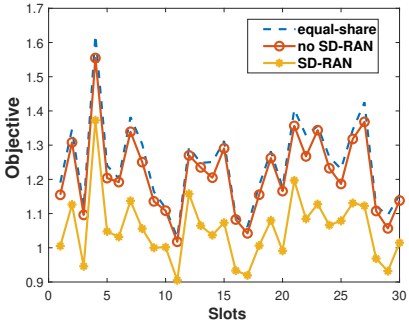

Fig. 3. The evolution of objective (1) for Case 2.

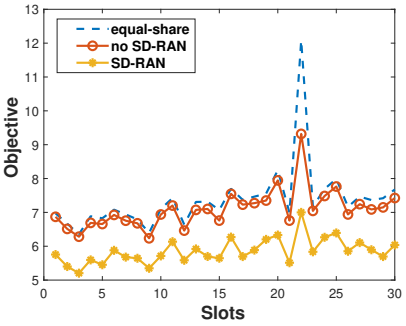

Fig. 4. The evolution of objective (1) for Case 3.

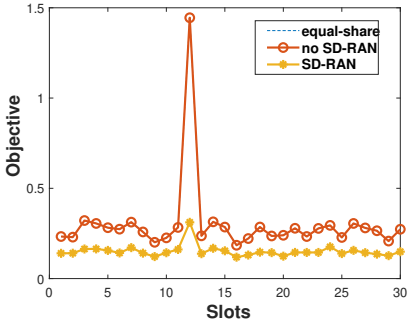

Fig. 5. The evolution of objective (4) for Case 1.

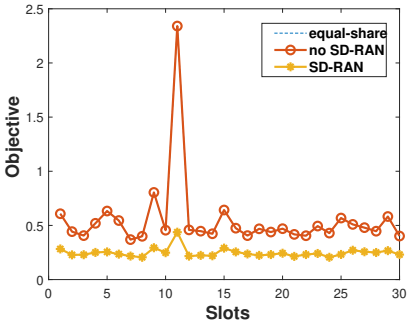

Fig. 6. The evolution of objective (4) for Case 2.

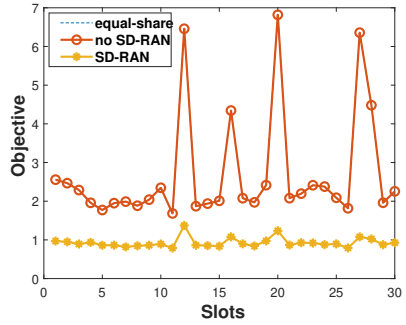

Fig. 7. The evolution of objective (4) for Case 3.

receives $K_{i,j} = \frac{K}{nm_i}$ PRBs. Substituting this into (4), we get

$$\sum_{j=1}^{m_i} \frac{1}{K_{i,j} R_{i,j}} = \frac{nm_i}{K} \sum_{j=1}^{m_i} \frac{1}{R_{i,j}}, \tag{56}$$

where the latter is identical to (55), proving that the equal-share policy is indeed identical to the optimal no-SD-RAN policy.

Again, as the number of users per BS increases, the total delay per BS increases (Case 3 has the highest delays).

Finally, we compare the performance in terms of the worst delay among all the users in the network, i.e., we are interested in providing min-max fairness among the users. Again, we compare three policies. The first is the solution to $\mathcal{P}_3$. The second policy is the no SD-RAN policy obtained by solving $\mathcal{P}_{0,2}$, whereas the third policy is, as before, the equal share. Needless to say, the latter two pertain to the classical cellular network operation (no SD-RAN). The other parameters remain unchanged compared to the previous scenarios. Fig. 8 depicts the results for Case 1, Fig. 9 shows the outcomes for Case 2, and Fig. 10 the results pertaining to Case 3. As can be observed, in all scenarios, our policy outperforms the other two policies, corroborating the advantages SD-RAN brings to the operation of cellular networks in this aspect as well.

The effects shown in the previous results can be observed in other scenarios too (different input parameters). Summarizing, common to all these is that SD-RAN is always more delay fair.

C. Impact of channel statistics

In the final scenario, we look at the impact of channel statistics (expressed through the first and second moments of the per-PRB rate) on the variability of the assigned number of PRBs to users. We do not look at the impact on the first moment (average) since from the analytical results it was clear that the number of allocated PRBs was inversely proportional to the (square root of) per-PRB rate. Therefore, users with a higher average per-PRB rate would be receiving fewer PRBs on average. We do the analysis for the three problems considered in this paper $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$. We assume there are 4 BSs, and in each BS there are 8 users (those from Table I). Of interest to us are user types 1, 3, and 6.

Our focus here is to look at how much varies the number of assigned PRBs to users with different channel statistics. To quantify the latter, we use the average per-PRB rate $\mathbb{E}[R]$ and the coefficient of variation (c_V), where the latter is defined as the ratio of the standard deviation to the mean of the per-PRB rate. Table II shows those two parameters for our users of interest in this scenario (row 2 and row 3). As can be seen, user types 1 and 3 have higher average per-PRB rates than user type 6. When it comes to the variability of channel conditions, again user types 1 and 3 are similar, whereas user type's 6 channel conditions are characterized by higher variability.

As was shown in Section III, the corresponding optimal allocation policies in this work react according to channel (CQI) changes at the users. Fig. 11 depicts the coefficient of variation of the number of assigned PRBs for these three user types over time. As can be observed from Fig. 11, when resources are allocated according to the optimal solutions of $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$, user types 1 and 6 have very similar variability in the number of allocated PRBs over time. On the other hand, user type 3 experiences the highest variability.

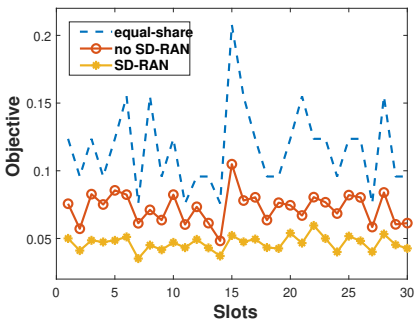

Fig. 8. The evolution of objective (5) for Case 1.

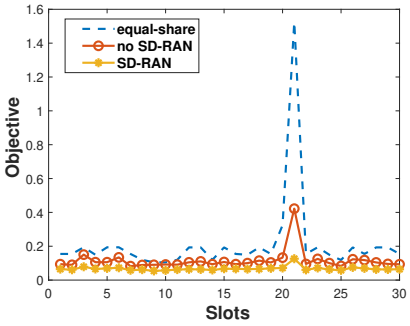

Fig. 9. The evolution of objective (5) for Case 2.

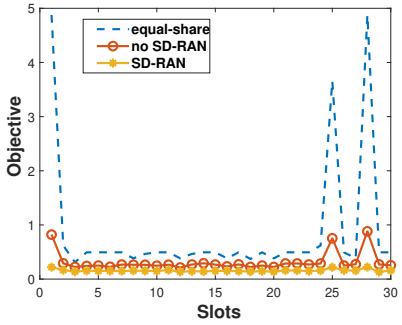

Fig. 10. The evolution of objective (5) for Case 3.

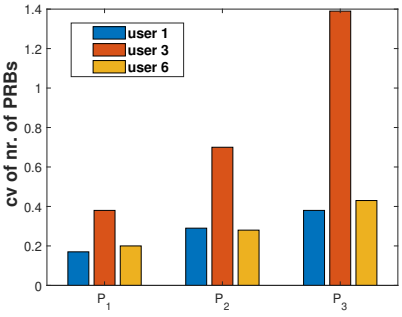

Fig. 11. The coefficient of variation of the number of assigned PRBs to users over time.

TABLE II
THE MEAN AND THE COEFFICIENT OF VARIATION OF PER-PRB RATES AND THE INVERSE OF PER-PRB RATES FOR USERS FROM TABLE I

	user type 1	user type 3	user type 6
$\mathbb{E}[R]$	1.25	1.27	0.92
$c_{V,R}$	0.31	0.3	0.38
$\mathbb{E}[\frac{1}{R}]$	0.9	1.07	1.27
$c_{V,\frac{1}{R}}$	0.36	1.88	0.43

Now, when comparing these findings with the first and second moments of per-PRB rates (rows 2 and 3 in Table II), we can observe that the results are contradictory. Namely, as already mentioned, user types 1 and 3 have similar per-PRB rate statistics but the coefficient of variation of the number of assigned PRBs over time is much higher for user type 3.

Let us see next how the statistics of the inverse of per-PRB rate affect the variability of the number of assigned PRBs over time. Row 4 and row 5 of Table II depict the values of the mean and the coefficient of variation of $\frac{1}{R}$ for the three user types considered in this simulation scenario. Comparing these results with those of Fig. 11, we can see that the first moment of $\frac{1}{R}$ does not exhibit an impact on the variability of the number of allocated PRBs. But, the coefficient of variation of $\frac{1}{R}$ does. As can be observed from Table II, user type 3 has the highest $c_{V,\frac{1}{R}}$ (considerably higher than the other two user types), and also (from Fig. 11) the highest variability (considerably higher) in the number of assigned PRBs over time. On the other hand, user type 1 has a slightly lower $c_{V,\frac{1}{R}}$ than user type 3, and this is also reflected in the variability in Fig. 11.

So, to summarize, the variability in the assigned number of PRBs for a user depends mostly on the coefficient of variation of the inverse of its per-PRB rate.

VI. RELATED WORK

SD-RAN has since recently attracted considerable attention both from research and industry players [18], [19]. Transferring control decisions to a centralized (SD-RAN) controller as a way to improve performance (via increased flexibility) has been suggested first in [20] and [21]. However, none of these works discuss the exact gains in terms of neither the throughput nor delay, or objectives like fair resource allocation.

The first known prototype implementations of SD-RAN are FlexRAN [2] and 5G-EmPOWER [3]. Both these implementations are constrained to serve only a limited number of users with a single server and also are not concerned with matters of delay, or its fairness. In [22], the problem of minimizing the number of assigned resources has been considered in an SD-RAN environment, by taking into account two types of slices, those for delay-sensitive traffic, and those for throughput-critical traffic. The other contribution of [22] is that slice isolation can be maintained. However, there is no discussion on the resource allocation policy that provides delay fairness.

On a related note, the authors in [23] consider the problem of allocating resources where network slices can be spread across multiple BSs. The objective in [23] is to allocate resources so that the overall throughput (across all users) is maximized, by guaranteeing a minimum data rate to everyone first. However, the solution in [23] is based on a non-closed form approximation approach, which does not allow to express explicitly the dependency of throughput on different input parameters. Furthermore, delay fairness is not considered in [23]. As opposed to [23], in this paper, we solve the problem over the entire network in its most general form for any number of users, BSs, and heterogeneous channel statistics while providing closed-form delay-fair resource allocation policies.

Deriving the maximum achievable throughput in an SD-RAN-enabled network has been the focus of [8], where it was shown that the maxCQI policy on both levels of resource allocation leads to maximum possible throughput. However, there is no fairness in resource allocation in [8]. Two works that consider different types of fair resource allocation, in

terms of throughput, are [7] and [6]. In [7], the resource allocation policies that provide proportional fairness are derived. This was done for two scenarios. In the first, the objective is proportional fairness among all users in the network, whereas in the second, the goal is proportional fairness among base stations. A similar approach has been followed in [6], but for max-min fairness. Both these works are concerned only with throughput fairness, while delay fairness is not considered, which we do in this work.

Some forms of delay fairness in cellular networks have been considered in [24] and [25], wherein the former deep reinforcement learning has been used to obtain the optimal allocation policies. However, SD-RAN is not considered and there are no closed-form expressions for the resource allocation policies, which is the case with the approach in our work. In [25], a trade-off between fairness and delay in wireless packet scheduling is considered. However, the approach there is not compliant with an SD-RAN setup.

When it comes to delay fairness in SD-RAN, the work closest in spirit to ours is [19], where the attention is turned to network slicing in SD-RAN. However, while there is an implementation of the approach in [19] in existing open-source SD-RAN systems, there is no closed-form analytical solution for resource allocation that provides delay fairness, but only simulation results. In contrast, in our paper, we derive explicit formulas for the resource allocation policies for minimum potential delay fairness across all users in the network and min-max delay fairness across BSs and users, while providing advantages compared to the traditional non-SD-RAN approaches.

VII. Conclusion

In this paper, we considered the problem of providing different types of fairness in terms of delay in cellular networks in an SD-RAN setup. First, we considered the problem of minimum potential delay fairness, as a special case of network utility maximization, across all users in the network. The optimal allocation policy in this case is achieved if the amount of allocated resources to a user is inversely proportional to the square root of the per-PRB rate of that user in a slot. Then, we turned our attention to the delay fairness across base stations. To that end, we looked at min-max fairness and showed that the optimum is achieved if the amount of assigned resources is inversely proportional to the square root of the per-PRB rate of a user in a slot, but also depends on a function of the inverses of square roots of the other users' per-PRB rates. Finally, we derived the policy which provides min-max fairness across all users in the network. In this case, the optimal allocation per user is inversely proportional to its per-PRB rate, as opposed to the previous two cases. We evaluated the performance for the three problems on realistic data from traces and compared the performance against the corresponding optimal policies without SD-RAN and another classical policy (equal share). While outperforming considerably the non-SD-RAN approach, the results demonstrate the obvious advantages SD-RAN brings into cellular networks in terms of delay fairness. In the future, we plan to consider the problem of resource allocation providing delay fairness for time-sensitive traffic.

References

[1] L. Cui, R. Yu, and Q. Yan, "When big data meets software-defined networking: SDN for big data and big data for SDN," *IEEE Network*, vol. 30, no. 1, 2016.

[2] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRan: A flexible and programmable platform for software-defined radio access networks," in *Proc. of ACM CoNEXT*, 2016.

[3] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A software-defined networking platform for 5G radio access networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, 2019.

[4] A. Papa, R. Durner, L. Goratti, T. Rasheed, and W. Kellerer, "Controlling Next-Generation Software-Defined RANs," *IEEE Communications Magazine*, vol. 58, no. 7, 2020.

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

[6] F. Mehmeti and W. Kellerer, "Max-min fair resource allocation in SD-RAN," in *Proc. of ACM Q2SWinet*, 2022.

[7] F. Mehmeti and W. Kellerer, "Proportionally fair resource allocation in SD-RAN," in *Proc. of IEEE CCNC*, 2023.

[8] F. Mehmeti, A. Papa, and W. Kellerer, "Maximizing network throughput using SD-RAN," in *Proc. of IEEE CCNC*, 2023.

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

[10] A. Papa, P. Kutsevol, F. Mehmeti, and W. Kellerer, "Effects of SD-RAN control plane design on user Quality of Service," in *Proc. of IEEE Netsoft*, 2022.

[11] 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.

[12] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhauser, 2004.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

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

[15] 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.

[16] F. Mehmeti and T. L. Porta, "Analyzing a 5G Dataset and Modeling Metrics of Interest," in *Proc. of IEEE MSN*, 2021.

[17] 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.

[18] Z. Zaidi, V. Friderikos, and M. A. Imran, "Future RAN architecture: SD-RAN through a general-purpose processing platform," *IEEE Vehicular Technology Magazine*, vol. 10, no. 1, 2015.

[19] Q. Qin, N. Choi, M. R. Rahman, M. Thottan, and L. Tassiulas, "Network slicing in heterogeneous Software-defined RANs," in *Proc. of IEEE INFOCOM*, 2020.

[20] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "OpenRAN: A software-defined RAN architecture via virtualization," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, 2013.

[21] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *Proc. of ACM SIGCOMM workshop on Hot topics in Software Defined Networking*, 2013.

[22] A. Papa, M. Klugel, L. Goratti, T. Rasheed, and W. Kellerer, "Optimizing dynamic RAN slicing in programmable 5G networks," in *Proc. of IEEE ICC*, 2019.

[23] A. Papa, A. Jano, S. Ayvaşık, O. Ayan, H. M. Gürsu, and W. Kellerer, "User-based Quality of Service aware multi-cell radio access network slicing," *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, 2022.

[24] M. López-Sánchez, A. Villena-Rodríguez, G. Gómez, F. J. Martín-Vega, and M. C. Aguayo-Torres, "Latency fairness optimization on wireless networks through deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, 2022.

[25] A. Dua and N. Bambos, "On the fairness delay trade-off in wireless packet scheduling," in *Proc. of IEEE GLOBECOM*, 2005.