Joint α-Fair Allocation of RAN and Computing Resources to URLLC Users in 5G

Valentin Thomas Haider, *Graduate Student Member, IEEE*, Fidan Mehmeti, *Member, IEEE*, Ana Cantarero, and Wolfgang Kellerer, *Senior Member, IEEE*

***Abstract*—5G networks have emerged as the only viable solution to accomplish a satisfying performance level for various types of services, where each of them has very challenging traffic requirements. One of those services are ultra-reliable low-latency communications (URLLC), which are characterized by the stringent demand to deliver packets within a very short time with high reliability. Besides being successfully transmitted, the data must be processed as well. To maximize the number of users the network can serve, the interplay of the different resources must be understood and adequate resource allocation schemes must be devised. In this paper, we consider the joint allocation of uplink and downlink radio access network (RAN) and edge computing resources such that the traffic requirements of individual users are met and the network utility is maximized for different types of fairness. To this end, an optimization problem for the general case of α-fairness is formulated and its properties are explored. For the special cases of no fairness, proportional fairness, minimum potential delay fairness, and max-min fairness, polynomial-time allocation approximation algorithms are proposed. Using data from real traces, it is shown that the performance deviation of these approaches from the continuous optimum (upper bound) rarely exceeds 2%.**

***Index Terms*—5G, URLLC, α-fairness, resource allocation.**

I. Introduction

Currently, three main types of services are provided by 5G networks: massive machine-type communications (mMTC), enhanced mobile broadband (eMBB), and ultra-reliable low-latency communications (URLLC). For eMBB services, very high data rates with high spectral efficiency are needed [2]. Low energy consumption and the support of a high number of devices are required by mMTC services [3]. Finally, URLLC services necessitate extremely low latencies and the support for high mobility and availability [4]. The latter service type is our focus in this paper.

Examples of URLLC services are applications like autonomous driving, remote surgery, or remote monitoring and control [5]. The main requirement for these services is the ability to deliver packets with very high reliability within a short time (on the order of ms), which is quite challenging. Besides being transmitted, those data need to be processed and often a response needs to be returned to the initial sender within a given latency. This impedes handling the data even further. Lastly, given the constrained network resources in the radio access network (RAN) and the edge and the ever-increasing number of devices competing for those resources, the challenge becomes even more complicated. For the aforementioned URLLC services, it is not only essential to comply with these strict requirements, but, given their nature, any violation of the specifications may bring a severe risk to human lives. Thus, enabling the impeccable operation of URLLC services is paramount.

In cellular networks, where the channel characteristics of users show dynamic behavior over time due to mobility and processes like shadowing, facilitating this flawless operation is particularly challenging. To provide the needed data rates in the up- and downlink, as well as to provide a sufficient amount of processing resources, an appropriate resource allocation scheme is needed on the transmission side, i.e., the RAN, and on the analyst side, i.e., the edge computing resources. Furthermore, to increase the revenue and utilization of the system, the operator needs to allocate these resources in an *efficient way* such that as many users as possible can be served.

The joint allocation of two types of resources, i.e., RAN and edge computing resources, where the uplink and downlink RAN resources are separated, renders the development of efficient allocation schemes quite challenging. The reason for this complexity of the problem is that one resource can compensate for the other, meaning that, e.g., the assignment of fewer uplink RAN resources (implying a higher uplink transmission delay) can be compensated by the allocation of more computing resources (implying a lower processing delay). This compensation is not possible if the allocation approach separates the various resources, as for example done in [6], where the demands of each user are specified separately in terms of RAN, computing, and storage resources.

Two important questions come up related to the joint allocation of uplink/downlink network and edge cloud resources that provides fairness among the users:

- Firstly, what is the policy that allows achieving different types of fairness with a joint allocation of uplink and downlink RAN as well as edge computing resources, while all the relevant traffic requirements are fulfilled?
- Secondly, how does the maximal tolerable latency that is set for the entire process of transmitting the data from and to a user, including the processing at the edge, influence the overall utility?

To answer the previous questions, we formulate an optimization problem, where the aim is to provide α-fairness while

Valentin Thomas Haider (valentin.haider@tum.de), Fidan Mehmeti (fidan.mehmeti@tum.de), and Wolfgang Kellerer (wolfgang.kellerer@tum.de) are with the Chair of Communication Networks, Technical University of Munich, 80333 München, Germany.

Ana Cantarero (ana.cantarero@bmw.de) is with the BMW Group, 80788 München, Germany.

Initial results of this paper have been presented at the IEEE International Conference on Communications and Networking (IEEE CCNC), 2023 [1].

This work was supported in part by the BMW Group and in part by the German Federal Ministry of Education and Research (BMBF) as part of the project “6G-ANNA” with project identification number 16KISK107.

TABLE I: List of Acronyms

BS	Base Station	CQI	Channel Quality Indicator
CRV	constrained rate variability	eMBB	enhanced Mobile Broadband
gNodeB	Next Generation Node B	MCS	Modulation and Coding Scheme
MEAR	Minimum Expected Achieved Rate	mMTC	massive Machine-Type Communication
NUM	Network Utility Maximization	PRB	Phyiscal Resource Block
QoS	Quality of Service	RAN	Radio Access Network
RR	Round-Robin	SB	subband
SCS	Subcarrier Spacing	SINR	Signal-to-Interference-Plus-Noise-Ratio
SLA	Service Level Agreement	SotA	State of the Art
URLLC	Ultra Reliable Low Latency Communication	WB	wideband

fulfilling the maximum delay requirements for every user and meeting the resource constraints on the uplink/downlink RAN and the edge computing side. After analyzing this optimization problem regarding its solvability, we specifically contemplate the four main types of fairness, i.e., throughput maximization (no fairness), proportional fairness, minimum potential delay fairness, and max-min fairness, and propose polynomial-time approximation algorithms. We work with very realistic assumptions in this paper, e.g., we assume that a user experiences different channel gains over different channel resources (blocks), irrespective on how close the resources are in the frequency dimension. As the results indicate how resources should be allocated to increase the total network utility while providing certain types of fairness, they are especially important for the network operator. Additionally, the results are valuable in order to get an idea of the interplay of the assignment of different separated resources while a delay constraint is active. There are two key messages of this paper: First, the integer nonlinear optimization problem for α-fair resource allocation that is generally unbounded in time can be solved close to optimality using the solution of the continuous relaxation embedded in appropriate algorithms. Second, only for $\alpha = 0$ the overall system throughput decreases with a tighter delay constraint. For all other types of fairness, no dependence of the system throughput on the delay constraint was observed. Specifically, our main contributions are:

- We formulate the optimization problem of jointly allocating uplink and downlink RAN and computing resources as a network utility maximization (NUM) problem for vehicular users within the same cell and solve the integer-relaxed version for general α.
- As the original integer problem is NP-hard, we propose polynomial-time approximation algorithms that provide near-optimal performance for the cases $\alpha = 0$, $\alpha = 1$, $\alpha = 2$, and $\alpha \to \infty$.
- Using data from real measurements, we evaluate our approach and provide some interesting engineering insights.

The remainder of this work is structured as follows: We discuss some related work in this field in Section II. These elaborations are succeeded by the introduction of the system model and the formulation of the optimization problem in Section III. In Section IV, we analyze the properties of the optimization problem, whereas approximation algorithms for the four fairness cases are proposed in Section V. The performance of the algorithms is evaluated in Section VI. Finally, Section VII concludes this paper. A summary of the acronyms used throughout this paper is given in Table I, while the symbols are listed in Table II.

II. Related Work

In [7], the authors consider a two-level network architecture comprising a lower-level RAN with edge computing resources as well as an upper-level transport network with central cloud computing resources. They investigate a network slicing process for the three types of services in 5G where they especially examine the partitioning ratios between the lower- and upper-level resources for the service types. While they constrain their optimization problem with a maximum delay requirement for the services, their objective is to minimize an over-provisioning ratio defined as the ratio of the required delay divided by the achieved delay. Moreover, since the authors consider slices as the unit of allocation, the granularity of the units is much larger than in the present work. A work that is concerned with uplink communication of URLLC traffic is [8]. However, the authors neither formulate an optimization problem nor do they consider the processing of the data; instead, two protocols for connection-less transmission of URLLC traffic are assessed.

Further, the work in [9] considers the optimal allocation of transmission attempts and communication channels for URLLC traffic in a cellular system. Two optimization problems for the resource allocation are formulated: in the first scenario, the number of transmission attempt assignments is fixed before starting the transmission, whereas it is adaptive in the second scenario. While [9] is also concerned with reducing the required resources, the setup and the objective are different from our work, and providing fairness is not one of the aims. To meet the latency and reliability requirements of URLLC traffic, the authors in [10] propose a periodic resource allocation scheme. While minimizing the needed network resources, i.e., choosing the best modulation and coding scheme (MCS) when considering retransmissions and the latency and reliability constraints, the scope of [10] is limited due to the assumption of a factory environment, which implies that channel conditions are not significantly changing over time. Furthermore, the objective does again not include providing any fairness.

Other related works are [11], [12]. In [11], three objectives similar to the present paper are considered: maximize the total throughput in the network, provide proportional fairness, and achieve max-min fairness. There are some important differences between this work and [11] though. In [11], the primary goal is to provide a given constant data rate to everyone and then reallocate the unused resources to the users according to the respective fairness policies. Besides, while the setup in our work is related to URLLC traffic, the target of [11]

TABLE II: List of Symbols

α	fairness value, $\alpha \in [0, \infty)$	β	slack variable, equal to $1 - \alpha$
$\Delta_{\{u,d\}}$	uplink/downlink packet size	$\mathcal{E}$	exponential cone, defined in (8)
$f_i^{\alpha}(\boldsymbol{I}_{u,i}, \boldsymbol{I}_{d,i}, m_i)$	utility function of user i, defined in (2)	g	slack variable, objective function for the optimization problem in epigraph form
$\Gamma_{\mathcal{P}}(\boldsymbol{x})$	generalized logarithm for the n-dimensional power cone $\mathcal{P}_{\zeta}^{n}$, defined in (22)	$\Gamma_{\mathcal{Q}}(\boldsymbol{x})$	generalized logarithm for the n-dimensional quadratic cone $\mathcal{Q}^n$, defined in (21)
$\gamma_{\{u,d\},i}$	uplink/downlink RAN data rate of user i, defined in (3)	$\boldsymbol{I}_{\{u,d\}}$, $\boldsymbol{I}_{\{u,d\},i}$, $I_{\{u,d\},ij}$	uplink/downlink PRB allocation matrix, vector of user i, indicator of user i for PRB j
$\boldsymbol{J}_{\{u,d\}}$, $\boldsymbol{J}_{\{u,d\},i}$, $J_{\{u,d\},ij}$	uplink/downlink integer PRB allocation matrix, vector of user i, indicator of user i for PRB j	$K_{\{u,d\}}$	number of available uplink/downlink PRBs
$\mathcal{K}$	set of all PRBs	L	number of available edge computing resources
λ_a	roots of the characteristic polynomial of a Hessian matrix	$\Lambda(\boldsymbol{w})$	logarithmic barrier function for the optimization problem defined in (19)
$\boldsymbol{m}$, m_i	edge computing resource allocation vector, indicator of user i	N	number of users in the system
$\boldsymbol{n}$, n_i	integer edge computing resource allocation vector, indicator of user i	p	processing rate of one edge computing unit
$\boldsymbol{\Phi}_{\{u,d\}}$, $\Phi_{\{u,d\},i}$, $\Phi_{\{u,d\},ij}$	uplink/downlink data rate matrix, vector of user i, indicator of user i for PRB j	$\mathcal{P}_{\zeta}^{n}$	n-dimensional power cone, defined in (7)
$\mathcal{Q}_r^n$	n-dimensional rotated quadratic cone, defined in (6)	$\boldsymbol{s}$, s_{1i}, s_{2i}, s_{3i}	slack variable vector with entries s_{ki} corresponding to uplink ($k = 1$), processing ($k = 2$), and downlink ($k = 3$) rate of users i, defined in (9f)-(9h)
$\mathcal{S}$	helper set, $\{1, 2, 3\}$	$t_i(\boldsymbol{I}_{u,i}, \boldsymbol{I}_{d,i}, m_i)$	delay of user i in up-/downlink scenario, see (5)
T_{max}	maximum allowed delay a packet can experience	$\boldsymbol{u}$, u_{ki}	slack variable vector with entries u_{ki} bounding an expression including the corresponding s_{ki}
$\mathcal{U}$	set of all users		

are users with eMBB traffic and satisfying the requirements of users with URLLC traffic is more challenging. Lastly, the authors of [11] only consider a one-dimensional allocation problem, as they assume that the channel conditions are equal across all PRBs.

There exist a lot of works focusing on the joint allocation of resources to eMBB and URLLC users [13]–[19]. In particular, the authors in [16] aim to provide long-term proportional fairness to eMBB users in the downlink, while simultaneously fulfilling the latency and reliability demands of URLLC users. Although they jointly consider eMBB and URLLC users, their resource allocation scheme is in fact a two-step process. First, downlink RAN resources are allocated with the objective of providing proportional fairness to the eMBB users. Thereafter, the demands of the URLLC users are considered and RAN resources are reallocated to fulfill the delay requirements of URLLC users. While the presented approach uses very strict assumptions regarding the latency, it lacks realistic assumptions regarding the channel conditions, i.e., varying channel quality indicator (CQI) values across PRBs in the frequency domain. Furthermore, only one type of fairness, i.e., proportional fairness, is considered and the processing of the data is not included in the system model. A similar approach of puncturing and providing proportional fairness is taken by the authors of [17]. Their objective function, however, benefits from a risk measure that counteracts the reallocation of PRBs belonging to users with sparse resources. In [19], the objective is to maximize the minimum expected achieved rate (MEAR) of eMBB users while instantaneously providing the resources to URLLC users' requests. The authors show that their approach outperforms other State of the Art (SotA) solutions in terms of MEAR and fairness. Similar to the work in [16], the authors in [18] aim to maximize the utility of eMBB users while fulfilling the latency requirements of URLLC users. They again consider puncturing, i.e., the reallocation of RAN resources, for fulfilling the demands of URLLC users and use three different loss models (linear, convex, and threshold) for modeling the influence of puncturing on the eMBB user utility maximization. Despite developing three scheduling policies for the different loss models, providing fairness to the users is not a goal. Furthermore, in [18], it is assumed that all URLLC users can be served.

Finally, different questions on URLLC RAN resource allocation are analyzed by the authors of [20]. An optimization problem where the sum over users satisfying their service level agreement (SLA) is maximized is defined, however, no solution to the problem but just an analysis of its NP-hardness is provided. Besides, the authors of [20] address the research question of deciding whether a given set of users can be scheduled such that their SLAs are fulfilled. A feasible resource allocation that is attained in polynomial time is provided by the authors. However, per-PRB rates are either zero or a fixed number, which is a simplifying assumption compared to the channel modeling in our work where we take real CQI measurements to determine per-PRB rates. Moreover, the given solution is not optimal.

III. Problem Formulation

In this section, first, the system model will be introduced in detail and important parameters are defined. Based on these elaborations, the optimization problem that is associated with the contemplated setup is mathematically stated.

A. System Model

With the possibility of network slicing in 5G [21], *dedicated* network resources can be allocated to users requiring the same service quality, e.g., users with URLLC type of traffic that have the same reliability and latency demands. In 5G, *PRBs* are used as the unit of allocation on a per-slot basis [22]. Over

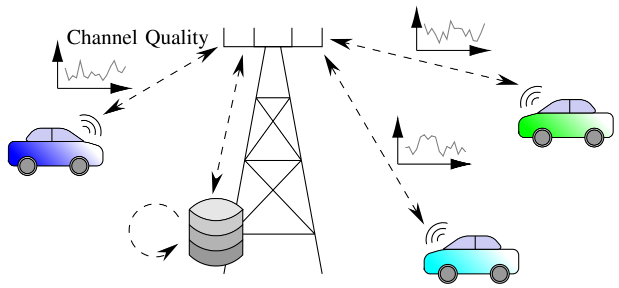

Fig. 1: Illustration of the system model.

the course of this work, we assume that the considered users are situated in the coverage area of a 5G macro base station (gNodeB) operating in the sub-6 GHz band. All users require the same service quality. The focus is set on the uplink and downlink communication and we assume that the processing of the data is executed at the edge, which is collocated with the base station (BS).

The system consists of a single BS, multiple users contained in the set $\mathcal{U}$, and edge computing resources (Fig. 1). There are N users simultaneously requesting a service by sending a packet to the BS, where the inquiry is handled. Afterward, each user is receiving a response that is sent from the BS. To enable the communication, K_u and K_d PRBs are available in the uplink and downlink RAN, respectively. The set of all PRBs is denoted as $\mathcal{K}$. Moreover, there are L edge computing resources accessible at the BS to process the information.

We assume that the channel conditions change over time, i.e., they vary from one radio frame to another. Furthermore, the channel conditions differ across the PRBs for a single user even within the same frame. Because of the time-varying nature of the channels and the mobility of the users, the per-PRB CQI (which is a function of the signal-to-interference-plus-noise-ratio (SINR)) changes from one radio frame to another. The CQI can take 15 different values [22]. Finally, the per-PRB CQI sets, depending on the used MCS, the per-PRB rate for a user. Thus, scheduling must be conducted across two dimensions, *time* and *frequency*.

The process of transmitting and processing the packets must be carried out within a maximum time of T_{max}, as we are considering URLLC traffic in this paper.[1] Hence, every user must be allocated at least one PRB as well as one edge computing resource, since otherwise the user cannot fulfill its delay constraint. Naturally, a PRB and also an edge computing resource can only be allocated to one user and the resource can either be fully allocated or left unassigned.

Finally, due to the consideration of services that are organized in small packets, we assume that the packet sizes $\Delta_{\{u,d\}}$ are fixed [23]. Furthermore, packets are generated for each user at the beginning of each radio frame, i.e., the number of transmitted packets is fixed.

B. Optimization Problem Formulation

The objective of this work is to maximize the overall network utility while guaranteeing that all users satisfy their traffic requirements and taking into account the constrained RAN and edge computing resources. The focus is set on providing α-fairness, in the same spirit as the NUM approach [24]. We thus can formulate the following optimization problem:

$$\max_{\boldsymbol{I}_u,\boldsymbol{I}_d,\boldsymbol{m}} \sum_{i=1}^{N} f_i^{\alpha}(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) \tag{1a}$$

$$\text{s.t.} \quad \frac{\Delta_u}{\gamma_{u,i}} + \frac{\Delta_u}{m_i p} + \frac{\Delta_d}{\gamma_{d,i}} \leq T_{max}, \quad \forall i \in \mathcal{U}, \tag{1b}$$

$$\sum_{i=1}^{N} m_i \leq L, \tag{1c}$$

$$\sum_{i=1}^{N} I_{\{u,d\},ij} \leq 1, \quad \forall j \in \mathcal{K}, \tag{1d}$$

$$\sum_{j=1}^{K_{\{u,d\}}} I_{\{u,d\},ij} \geq 1, \quad \forall i \in \mathcal{U}, \tag{1e}$$

$$I_{\{u,d\},ij} \in \{0,1\}, \quad \forall i \in \mathcal{U},\, j \in \mathcal{K}, \tag{1f}$$

$$m_i \in \mathbb{N} \setminus \{0\}, \quad \forall i \in \mathcal{U}, \tag{1g}$$

where

$$f_i^{\alpha}(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) = \\ = \begin{cases} \frac{1}{1-\alpha}\left(\gamma_{u,i}^{1-\alpha} + (m_i p)^{1-\alpha} + \gamma_{d,i}^{1-\alpha}\right), & \alpha \neq 1 \\ \log(\gamma_{u,i}) + \log(m_i p) + \log(\gamma_{d,i}), & \alpha = 1 \end{cases}, \tag{2}$$

and

$$\gamma_{\{u,d\},i} = \sum\nolimits_{j=1}^{K_{\{u,d\}}} I_{\{u,d\},ij} \Phi_{\{u,d\},ij} \tag{3}$$

describes the uplink/downlink RAN data rate of user i. The decision variable $\boldsymbol{I}_{\{u,d\}} = \{I_{\{u,d\},ij}\}$ denotes the $N \times K_{\{u,d\}}$ PRB allocation matrix in a given radio frame. This means that if $I_{\{u,d\},ij} = 1$, then PRB j is allocated to user i in that frame. The data rates user i would experience when being allocated PRB j in the uplink and downlink, respectively, are contained in the $N \times K_{\{u,d\}}$ matrix $\boldsymbol{\Phi}_{\{u,d\}} = \{\Phi_{\{u,d\},ij}\}$. These data rates are deduced from the CQI values given for each user. The number of allocated edge computing resources for user i is given by the $N \times 1$ decision variable $\boldsymbol{m} = \{m_i\}$ and the amount of information sent or received by each user is denoted by $\Delta_{\{u,d\}}$. Finally, the static variable p stands for the processing rate that one edge computing resource can provide.

The objective (1a) maximizes the overall utility for general $\alpha \in [0, \infty)$. Note that the special values $\alpha = 0$, $\alpha = 1$, $\alpha = 2$, and $\alpha \to \infty$ correspond to the cases of *no fairness* (throughput maximization), *proportional fairness*, *minimum potential delay fairness*, and *max-min fairness*. Clearly, since three resource parts are allocated, they all affect the overall gained utility. The first and third term in (2) (both for $\alpha \neq 1$ and $\alpha = 1$) corresponds to the utility from assigning uplink or downlink RAN resources to user i, whereas the second term denotes the utility obtained from allocating a fraction of the edge computing resources.

The maximum tolerable latency for every user is described by constraint (1b). Constraint (1c) captures the finite amount

[1] While the requirement for ultra-high reliability for URLLC traffic is to transmit the packets successfully within the maximum latency in more than 99% of the attempts, here we are even more conservative and require that the procedure must be executed within the deadline every time. In case a less strict reliability criterion should be modeled, the delay constraint changes to $\mathbb{P}\left(\frac{\Delta_u}{\gamma_{u,i}} + \frac{\Delta_u}{m_i p} + \frac{\Delta_d}{\gamma_{d,i}} \leq T_{Max}\right) \geq 1 - \epsilon$, where ϵ denotes the outage probability that one wants to allow. Therefore, $1 - \epsilon$ denotes the reliability.

of available computing resources. Constraint (1d) merely indicates that every block can be assigned to at most one user, whereas (1e) dictates that every user must receive at least one PRB both in the uplink and downlink. Lastly, the integer nature of the decision variables is described by (1f) and (1g), where the latter constraint includes the minimum number of one edge computing resource that needs to be assigned to every user.

IV. Analysis

The previously introduced optimization problem belongs to the class of Integer Nonlinear Programs, which are generally known to be NP-hard [25]. Hence, approximation algorithms are needed to obtain a solution to (1).

The procedure that we follow in this paper comprises two main steps. First, we relax the integer nature of the decision variables and allow them to be continuous. Next, the transformed optimization problem is shown to be convex and solvable in polynomial time under these conditions. As the second step, in Section V, we propose special approximation algorithms to obtain an integer solution to the aforementioned optimization problem.

We continue with the first step of showing the convexity of (1) when $I_{\{u,d\},ij} \in [0,1]$ and $m_i \in [1,\infty)$. Since the constraints (1c)-(1g) are linear inequalities, they are apparently convex. To prove the concavity of the objective function, the concavity of $f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ needs to be shown, as the sum of concave functions is a concave function itself. We have:

Lemma 1. *The function* $f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ *is concave.*

Proof. The gradient of $f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ for $\alpha \neq 1$ is

$$\nabla f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) = \big[\Phi_{u,i1}\gamma_{u,i}^{-\alpha} \quad \dots \quad \Phi_{u,iK_u}\gamma_{u,i}^{-\alpha}$$
$$p(m_ip)^{-\alpha} \quad \Phi_{d,i1}\gamma_{d,i}^{-\alpha} \quad \dots \quad \Phi_{d,iK_d}\gamma_{d,i}^{-\alpha}\big]^T .$$

Next, the Hessian matrix of $f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ for $\alpha \neq 1$ is calculated as

$$\nabla^2 f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) =$$
$$= -\alpha \left[\begin{matrix} \Phi_{u,i1}^2/\gamma_{u,i}^{\alpha+1} & \dots & \Phi_{u,i1}\Phi_{u,iK_u}/\gamma_{u,i}^{\alpha+1} \\ \vdots & \ddots & \vdots \\ \Phi_{u,iK_u}\Phi_{u,i1}/\gamma_{u,i}^{\alpha+1} & \dots & \Phi_{u,iK_u}^2/\gamma_{u,i}^{\alpha+1} \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}\right.$$
$$\left.\begin{matrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \\ p^2/(m_ip)^{\alpha+1} & 0 & \dots & 0 \\ 0 & \Phi_{d,i1}^2/\gamma_{d,i}^{\alpha+1} & \dots & \Phi_{d,i1}\Phi_{d,iK_d}/\gamma_{d,i}^{\alpha+1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \Phi_{d,iK_d}\Phi_{d,i1}/\gamma_{d,i}^{\alpha+1} & \dots & \Phi_{d,iK_d}^2/\gamma_{d,i}^{\alpha+1} \end{matrix}\right] .$$

Then, the characteristic polynomial of $\nabla^2 f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ for $\alpha \neq 1$ is computed as

$$\det(\nabla^2 f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) - \lambda\mathbb{I}) =$$
$$= (-1)^{K_u+K_d-1} * \lambda^{K_u+K_d-2}\left(\alpha p^2(m_ip)^{-\alpha-1} + \lambda\right)$$
$$* \Big(\lambda^2 + \alpha^2\gamma_{u,i}^{-\alpha-1}\gamma_{d,i}^{-\alpha-1}\left(\Phi_{u,i1}^2\Phi_{d,i1}^2 + \dots + \Phi_{u,iK_u}^2\Phi_{d,iK_d}^2\right)$$
$$+ \lambda\alpha\Big(\Phi_{u,i1}^2\gamma_{u,i}^{-\alpha-1} + \dots + \Phi_{u,iK_u}^2\gamma_{u,i}^{-\alpha-1}$$
$$+\Phi_{d,i1}^2\gamma_{d,i}^{-\alpha-1} + \dots + \Phi_{d,iK_d}^2\gamma_{d,i}^{-\alpha-1}\Big)\Big),$$

where $\mathbb{I}$ denotes the identity matrix in the corresponding dimension and λ denote eigenvalues. For $\alpha \neq 1$, the eigenvalues of the Hessian $\nabla^2 f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ are determined as

$$\lambda_1, \dots, \lambda_{K_u+K_d-2} = 0,$$
$$\lambda_{K_u+K_d-1} = -\alpha\gamma_{u,i}^{-\alpha-1}\left(\Phi_{u,i1}^2 + \dots + \Phi_{u,iK_u}^2\right),$$
$$\lambda_{K_u+K_d} = -\alpha\gamma_{d,i}^{-\alpha-1}\left(\Phi_{d,i1}^2 + \dots + \Phi_{d,iK_d}^2\right),$$
$$\lambda_{K_u+K_d+1} = -\alpha p^2(m_ip)^{-\alpha-1}.$$

The proof for $\alpha = 1$ is omitted here, as it follows the exact same proposition as for $\alpha \neq 1$. Conclusively, the Hessian $\nabla^2 f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ is negative semidefinite for any α as all eigenvalues of the Hessian are less than or equal to 0 and thus the function $f_i^\alpha(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ is concave for all α. □

Next, the characteristics of (1b) are explored. We have:

Lemma 2. *Constraint* (1b) *is convex.*

Proof. We denote the left-hand side of (1b) as

$$t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) = \frac{\Delta_u}{\sum\limits_{j=1}^{K_u} I_{u,ij}\Phi_{u,ij}} + \frac{\Delta_u}{m_ip} + \frac{\Delta_d}{\sum\limits_{j=1}^{K_d} I_{d,ij}\Phi_{d,ij}}$$
$$= \frac{\Delta_u}{\gamma_{u,i}} + \frac{\Delta_u}{m_ip} + \frac{\Delta_d}{\gamma_{d,i}}. \tag{5}$$

Calculating the gradient of $t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ leads to

$$\nabla t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) = \Big[\tfrac{-\Delta_u\Phi_{u,i1}}{\gamma_{u,i}^2} \quad \dots \quad \tfrac{-\Delta_u\Phi_{u,iK_u}}{\gamma_{u,i}^2}$$
$$\tfrac{-\Delta_u}{m_i^2p} \quad \tfrac{-\Delta_d\Phi_{d,i1}}{\gamma_{d,i}^2} \quad \dots \quad \tfrac{-\Delta_d\Phi_{d,iK_d}}{\gamma_{d,i}^2}\Big]^T .$$

For the Hessian of $t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$, we have

$$\nabla^2 t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) =$$
$$= \left[\begin{matrix} 2\Delta_u\Phi_{u,i1}^2/\gamma_{u,i}^3 & \dots & 2\Delta_u\Phi_{u,i1}\Phi_{u,iK_u}/\gamma_{u,i}^3 \\ \vdots & \ddots & \vdots \\ 2\Delta_u\Phi_{u,iK_u}\Phi_{u,i1}/\gamma_{u,i}^3 & \dots & 2\Delta_u\Phi_{u,iK_u}^2/\gamma_{u,i}^3 \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}\right.$$
$$\left.\begin{matrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \\ 2\Delta_u/m_i^3p & 0 & \dots & 0 \\ 0 & 2\Delta_d\Phi_{d,i1}^2/\gamma_{d,i}^3 & \dots & 2\Delta_d\Phi_{d,i1}\Phi_{d,iK_d}/\gamma_{d,i}^3 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 2\Delta_d\Phi_{d,iK_d}\Phi_{d,i1}/\gamma_{d,i}^3 & \dots & 2\Delta_d\Phi_{d,iK_d}^2/\gamma_{d,i}^3 \end{matrix}\right] .$$

For the determinant of $\nabla^2 t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) - \lambda\mathbb{I}$, we get

$$\det(\nabla^2 t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i) - \lambda\mathbb{I}) =$$
$$(-1)^{K_u+K_d-1}\lambda^{K_u+K_d-2}\left(2\Delta_u m_i^{-3}p^{-1} + \lambda\right)$$

$$* \Big(\lambda^2 + 4\Delta_u\Delta_d\gamma_{u,i}^{-2}\gamma_{d,i}^{-2}\left(\Phi_{u,i1}^2\Phi_{d,i1}^2 + \cdots + \Phi_{u,iK_u}^2\Phi_{d,iK_d}^2\right) + 2\lambda\Big(\Delta_u\gamma_{u,i}^{-2}\left(\Phi_{u,i1}^2 + \cdots + \Phi_{u,iK_u}^2\right) + \Delta_d\gamma_{d,i}^{-2}\left(\Phi_{d,i1}^2 + \cdots + \Phi_{d,iK_d}^2\right)\Big)\Big).$$

Lastly, the eigenvalues of the Hessian $\nabla^2 t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ can be found as

$$\begin{aligned}
&\lambda_1, ..., \lambda_{K_u+K_d-2} = 0,\\
&\lambda_{K_u+K_d-1} = 2\Delta_u\gamma_{u,i}^{-3}\left(\Phi_{u,i1}^2 + \cdots + \Phi_{u,iK_u}^2\right),\\
&\lambda_{K_u+K_d} = 2\Delta_d\gamma_{d,i}^{-3}\left(\Phi_{d,i1}^2 + \cdots + \Phi_{d,iK_d}^2\right),\\
&\lambda_{K_u+K_d+1} = 2\Delta_u m_i^{-3} p^{-1}.
\end{aligned}$$

Thus, the Hessian $\nabla^2 t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ is positive semidefinite and the function $t_i(\boldsymbol{I}_{u,i},\, \boldsymbol{I}_{d,i},\, m_i)$ is convex, as all the eigenvalues of the Hessian are greater than or equal to zero. □

Theorem 3. *The integer relaxed optimization problem* (1) *is a convex optimization problem.*

Proof. Given the linearity of (1c)-(1g) as well as Lemmas 1 and 2 proves that (1) is a convex optimization problem. □

For the next main step of proving the polynomial-time solvability of the integer-relaxed optimization problem, (1) is rewritten into a convex optimization problem with generalized inequality constraints. For the following derivations, we define the n-dimensional rotated quadratic cone as

$$\mathcal{Q}_r^n = \left\{\boldsymbol{x} \in \mathbb{R}^n \,|\, 2x_1x_2 \geq x_3^2 + \cdots + x_n^2,\, x_1,\, x_2 \geq 0\right\}. \quad (6)$$

Additionally, the n-dimensional power cone parameterized by a real number $\zeta \in [0, 1]$ is specified as

$$\mathcal{P}_\zeta^n = \left\{\boldsymbol{x} \in \mathbb{R}^n \,|\, x_1^\zeta x_2^{1-\zeta} \geq \sqrt{x_3^2 + \cdots + x_n^2},\, x_1,\, x_2 \geq 0\right\} \quad (7)$$

and the exponential cone as

$$\mathcal{E} = \left\{\boldsymbol{x} \in \mathbb{R}^3 \,|\, x_1 \geq x_2 e^{x_3/x_2},\, x_1,\, x_2 > 0\right\}. \quad (8)$$

As a first step, we introduce the slack variables s_{ki}, $k \in \mathcal{S} = \{1, 2, 3\}$, $i \in \mathcal{U}$, and write the relaxed optimization problem in epigraph form[2], such that it reads as

$$\begin{aligned}
\min_{g,\, \boldsymbol{I}_u,\, \boldsymbol{I}_d,\, \boldsymbol{m},\, \boldsymbol{s}} \quad & g && (9a)\\
\text{s.t.} \quad & -\sum_{i=1}^{N} h_i^\alpha(s_{1i},\, s_{2i},\, s_{3i},\, g) \leq 0, && (9b)\\
& \frac{\Delta_u}{s_{1i}} + \frac{\Delta_u}{s_{2i}} + \frac{\Delta_d}{s_{3i}} - T_{max} \leq 0, \quad \forall i \in \mathcal{U}, && (9c)\\
& \text{(1c), (1d), (1e),} &&\\
& 0 \leq I_{\{u,d\},ij} \leq 1, \quad \forall i \in \mathcal{U},\, j \in \mathcal{K}, && (9d)\\
& 1 - m_i \leq 0, \quad \forall i \in \mathcal{U}, && (9e)\\
& s_{1i} = \sum_{j=1}^{K_u} I_{u,ij}\Phi_{u,ij}, \quad \forall i \in \mathcal{U}, && (9f)\\
& s_{2i} = m_i p, \quad \forall i \in \mathcal{U}, && (9g)\\
& s_{3i} = \sum_{j=1}^{K_d} I_{d,ij}\Phi_{d,ij}, \quad \forall i \in \mathcal{U}, && (9h)
\end{aligned}$$

where

$$h_i^\alpha(s_{1i},\, s_{2i},\, s_{3i},\, g) = \begin{cases} \frac{1}{1-\alpha}\left(s_{1i}^{1-\alpha} + s_{2i}^{1-\alpha} + s_{3i}^{1-\alpha}\right) + g, & \alpha \neq 1\\ \log(s_{1i}) + \log(s_{2i}) + \log(s_{3i}) + g, & \alpha = 1 \end{cases}. \quad (10)$$

We continue with the introduction of conic inequalities[3] for the constraints (9b) and (9c).

Lemma 4. *The constraint* (9c) *can be written as*

$$\sum_{k=1}^{3} u_{ki} \leq T_{max}, \quad (11a)$$

$$(u_{ki},\, s_{ki};\, \Delta_k) \in \mathcal{Q}_r^3, \quad \forall k \in \mathcal{S}, \quad (11b)$$

where $\Delta_k = \Delta_u$ for $k = \{1, 2\}$ and $\Delta_k = \Delta_d$ for $k = 3$.

Proof. Constraint (11b) is by definition transformed to

$$u_{ki}s_{ki} \geq \sqrt{\Delta_k^2},\, u_{ki},\, s_{ki} \geq 0.$$

Dividing this term by s_{ki} and extracting the root leads to

$$u_{ki} \geq \frac{\Delta_k}{s_{ki}},\, u_{ki} \geq 0,\, s_{ki} \geq 0,$$

from where it can be observed that

$$\sum_{k=1}^{3} \frac{\Delta_k}{s_{ki}} \leq T_{max}.$$

Due to the positiveness of Δ_k and s_{ki}, which follows from the constraints (9d) and (9e), the constraints $u_{ki}, s_{ki} \geq 0$ that are introduced with this reformulation are always met. □

By setting $\beta = 1 - \alpha$ and using the slack variable u_{ki}, we transform the constraint (9b) to the constraints (13a)-(13d) for the cases $\alpha \in (0, 1)$ and $\alpha \in (1, \infty)$. Bringing the sum over s_{ki}^β to the right-hand side of the inequality results in

$$-g \leq \frac{1}{\beta}\sum_{k=1}^{3}\sum_{i=1}^{N} s_{ki}^\beta, \quad (12)$$

which can be converted into

$$(12) = \begin{cases} -g\beta \leq \displaystyle\sum_{k=1}^{3}\sum_{i=1}^{N} u_{ki}, & (13a)\\ u_{ki} \leq s_{ki}^\beta, \quad \forall k \in \mathcal{S},\, i \in \mathcal{U};\, \alpha \in (0, 1) & (13b)\\ g\,|\beta| \geq \displaystyle\sum_{k=1}^{3}\sum_{i=1}^{N} u_{ki}, & (13c)\\ u_{ki} \geq s_{ki}^\beta, \quad \forall k \in \mathcal{S},\, i \in \mathcal{U};\, \alpha \in (1, \infty) & (13d) \end{cases}.$$

We start with the case $\alpha \in (0, 1)$, i.e., $\beta \in (0, 1)$.

Lemma 5. *The constraint* (13b) *can be written as*

$$(s_{ki},\, 1;\, u_{ki}) \in \mathcal{P}_\beta^3. \quad (14)$$

Proof. By definition, the expression (14) is equivalent to

$$s_{ki}^\beta 1^{1-\beta} \geq \sqrt{u_{ki}^2},\, s_{ki} \geq 0,$$

which simplifies to

$$s_{ki}^\beta \geq u_{ki},\, s_{ki} \geq 0,$$

by extracting the root and dropping the factor 1. Due to the constraints (9d) and (9e), the additional constraint $s_{ki} \geq 0$ is again fulfilled. □

[2] The epigraph form is an equivalent form of a standard form optimization problem which has a linear objective function that can be introduced at the cost of an additional constraint and an additional scalar decision variable [26].

[3] A conic inequality has the general form $x \in K$, where K is a pointed and closed convex cone with non-empty interior in $\mathbb{R}^n$ [27].

Now, we proceed with the case $\alpha \in (1, \infty)$, which implies that $\beta \in (-\infty, 0)$.

Lemma 6. *The constraint* (13d) *can be written as*

$$(u_{ki}, s_{ki}; 1) \in \mathcal{P}^3_{1/(1-\beta)}. \tag{15}$$

Proof. The expression (15) is by definition converted into

$$u_{ki}^{1/(1-\beta)} s_{ki}^{-\beta/(1-\beta)} \geq \sqrt{1^2}, u_{ki} \geq 0, s_{ki} \geq 0,$$

which simplifies to

$$u_{ki} \geq s_{ki}^{\beta}, u_{ki} \geq 0, s_{ki} \geq 0,$$

when taking the entire expression to the power of $(1-\beta)$ and multiplying both sides of the inequality by s_{ki}^{β}. Due to the positiveness of s_{ki}, implied by (9d) and (9e), the additional constraints $u_{ki} \geq 0$ and $s_{ki} \geq 0$ are always met. □

As a last step, we contemplate the case $\alpha = 1$. Then, (9b) must be reformulated to

$$-g \leq \sum_{k=1}^{3} \sum_{i=1}^{N} \log s_{ki}, \tag{16}$$

by bringing the sum over the logarithms to the right-hand side of the inequality. Inequality (16) can be written as

$$-g \leq \sum_{k=1}^{3} \sum_{i=1}^{N} u_{ki}, \tag{17a}$$

$$u_{ki} \leq \log s_{ki}, \quad \forall k \in \mathcal{S},\, i \in \mathcal{U}, \tag{17b}$$

by introducing the slack variable u_{ki} again.

Lemma 7. *Constraint* (17b) *can be reformulated as*

$$(s_{ki}, 1, u_{ki}) \in \mathcal{E}. \tag{18}$$

Proof. By definition, (18) is equivalent to

$$s_{ki} \geq 1 * e^{u_{ki}/1}, s_{ki} > 0,$$

where we take the logarithm on both sides and write it as

$$\log s_{ki} \geq u_{ki},\, s_{ki} > 0.$$

Due to the constraints defined in (1e), (9d), and (9e), the additional constraint $s_{ki} > 0$ that is introduced with this reformulation is fulfilled. □

Theorem 8. *The integer-relaxed version of the optimization problem* (1) *can be written as a convex optimization problem with generalized inequality constraints.*

Proof. Given the fact that (9b) is linear for $\alpha = 0$ as well as Lemmas 4, 5, 6, and 7 concludes the proof. □

With the preceding derivations, the optimization problem (1) reads for any $\alpha \in [0, \infty)$, written in its integer-relaxed version as a convex optimization problem with generalized inequality constraints, as:

$$\min_{g, \boldsymbol{I}_u, \boldsymbol{I}_d, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{u}} \quad g \tag{19a}$$

$$\text{s.t.} \qquad -\sum_{i=1}^{N} e_i^{\alpha}(s_{1i}, s_{2i}, s_{3i}, u_{1i}, u_{2i}, u_{3i}, g) \leq 0, \tag{19b}$$

$$(1c), (1d), (1e), (9d), (9e),$$
$$(9f), (9g), (9h), (11),$$

where

$$(19b) \equiv \begin{cases} (9b), & \alpha = 0 \\ (13a), (14), \quad \forall k \in \mathcal{S},\, i \in \mathcal{U}, & 0 < \alpha < 1 \\ (17a), (18), \quad \forall k \in \mathcal{S},\, i \in \mathcal{U}, & \alpha = 1 \\ (13c), (15), \quad \forall k \in \mathcal{S},\, i \in \mathcal{U}, & \alpha > 1 \end{cases}. \tag{20}$$

We define the following generalized logarithms and note their degrees for the final verification of the polynomial-time solvability of the optimization problem stated in (19). Further information on the generalized logarithm can be found in Section 11.6 of [26]. For the n-dimensional rotated quadratic cone $\mathcal{Q}_r^n$, a generalized logarithm can be designed as

$$\Gamma_{\mathcal{Q}}(\boldsymbol{x}) = \log\left(x_1^2 - \sum_{i=2}^{n} x_i^2\right), \tag{21}$$

which is the generalized logarithm for an ordinary quadratic cone, cf. [26]. Since the rotated n-dimensional quadratic cone can be written as an ordinary quadratic cone by a rotation of coordinates, $\Gamma_{\mathcal{Q}}(\boldsymbol{x})$ is also valid for the rotated n-dimensional quadratic cone $\mathcal{Q}_r^n$. The degree of a generalized logarithm can be calculated as $\theta_\Gamma = \nabla\Gamma(\boldsymbol{x})^T\boldsymbol{x}$, cf. [26]. Therefore, the degree of the function $\Gamma_{\mathcal{Q}}(\boldsymbol{x})$ is 2. Furthermore, for the n-dimensional power cone $\mathcal{P}_\zeta^n$, we define a generalized logarithm as

$$\Gamma_{\mathcal{P}}(\boldsymbol{x}) = \log\left(x_1^{2\zeta} x_2^{(2-2\zeta)} - \sum_{i=3}^{n} x_i^2\right) + (1-\zeta)\log(x_1) + \zeta\log(x_2), \tag{22}$$

as introduced in [28]. Its degree is 3. For the exponential cone $\mathcal{E}$, we define the generalized logarithm as [28]

$$\Gamma_{\mathcal{P}}(\boldsymbol{x}) = \log\left(x_2 \log\left(\frac{x_1}{x_2}\right) - x_3\right) + \log x_1 + \log x_2, \tag{23}$$

and note its degree as 3. Lastly, note that a slack variable that is attached to the system of equality constraints can be inserted for every linear inequality constraint of the optimization. The corresponding generalized logarithm for these slack variables has degree 1, as the slack variable needs to be in $\mathbb{R}_+$. With the preceding definitions of the generalized logarithms, a logarithmic barrier function[4] $\Lambda(\boldsymbol{w})$ can be given as

$$\Lambda(\boldsymbol{w}) = -\sum_{c=1}^{Z} \Gamma_c(\boldsymbol{w}),$$
$$\textbf{dom}\,\Lambda = \{\boldsymbol{w} \mid f_c(\boldsymbol{w}) \prec_{K_c} 0,\, c = 1, ..., Z\},$$

where $Z = (7+2K_u+2K_d)N+2+K_u+K_d$ for $\alpha = 0$ and $Z = (10+2K_u+2K_d)N+2+K_u+K_d$ for $\alpha \neq 0$. The vector $\boldsymbol{w}$ is composed of the vectorized matrices $\boldsymbol{I}_u$ and $\boldsymbol{I}_d$ as well as the vectors $\boldsymbol{m}$, $\boldsymbol{s} = \{s_{ki}\}$, and $\boldsymbol{u} = \{u_{ki}\}$. The function $\Gamma_c(\boldsymbol{w})$ denotes the generalized logarithms defined above for each generalized inequality constraint $f_c(\boldsymbol{w})$ in the convex optimization problem with generalized inequalities given in (19). As a logarithmic barrier function can be defined for the

[4] A barrier function is a function that represents the feasible set of an optimization problem. The domain of the barrier function is this feasible set. The barrier function is greater or equal to zero in the interior of the feasible set while it approaches infinity as the input approaches the boundary of the feasible set. It is used to incorporate inequality constraints of an optimization problem in the objective function such that the barrier method can be applied to solve the problem. [26]

optimization problem, the barrier method can be applied to solve the problem.

In the following, a complexity analysis that is based on the property of self-concordance is given[5].

Lemma 9. *The logarithmic barrier function $\Lambda(\boldsymbol{w})$ is self-concordant.*

Proof. First, note that the sum of self-concordant functions is again self-concordant [26]. Hence, the logarithmic barrier for the positive orthant defined by all slack variables corresponding to linear inequalities is a self-concordant function, because $-\log x$ is self-concordant. The logarithmic barriers established using the generalized logarithms defined in (21)-(23) are self-concordant as well; see Section 11.6 in [26] and Sections 2.4 and 3.1 in [28]. This concludes the proof. □

Lemma 10. *The number of total Newton steps*[6] *excluding the initial centering step for solving* (19) *using the Barrier method can be bounded by [26]*

$$T_{Barrier} = \left\lceil \frac{\log(\bar{\theta}/(t^{(0)}\epsilon))}{\log \mu} \right\rceil * \left(\frac{\bar{\theta}(\mu - 1 - \log \mu)}{\chi} + \log_2 \log_2(1/\epsilon) \right). \quad (24)$$

Proof. Given the fact that (19a) is linear and using Lemma 9, the objective of the Barrier method, i.e., the function $tg + \Lambda(\boldsymbol{w})$, is self-concordant. Given the additional properties that this function is closed and the sublevel sets of the optimization problem (19) are bounded leads to (24). □

In (24), $t^{(0)} > 0$ is the initial value of the algorithm parameter t of the barrier method, the parameter $\mu > 1$ is an algorithm parameter of the barrier method, and $\epsilon > 0$ is the specified tolerance of the barrier method, see Algorithm 11.1 in [26]. The parameter χ is a constant that depends on the backtracking parameters κ and τ, Alg. 9.2 in [26], which is used for line search in Newton's method. It is given as

$$\frac{1}{\chi} = \frac{20 - 8\kappa}{\kappa\tau(1 - 2\kappa)^2}.$$

Finally, $\bar{\theta}$ stands for the sum of the degrees of the generalized logarithms Γ_c, which for the contemplated problem is calculated as

$$\bar{\theta} = \begin{cases} (10 + 2K_u + 2K_d)N + 2 + K_u + K_d, & \alpha = 0 \\ (19 + 2K_u + 2K_d)N + 2 + K_u + K_d, & \alpha \neq 0 \end{cases}. \quad (25)$$

Theorem 11. *The complexity of solving the optimization problem* (19) *in terms of Newton steps is*

$$T_{Barrier} = \mathcal{O}\left(\log\left((K_u + K_d)N/\epsilon\right) * \left((K_u + K_d)N + \log_2 \log_2 (1/\epsilon)\right)\right). \quad (26)$$

Proof. Plugging (25) into (24) and simplifying this term leads to the bound given in (26). □

The order of growth of (26) is a function of $n \log n$ and of $\log(1/\epsilon)$, which is a desirable complexity for these types of problems.

V. Conversion Algorithms

In the previous section, it was shown that the integer-relaxed optimization problem can be solved optimally in polynomial time. However, the obtained allocation allows the assignment of arbitrary fractions of resources, which breaks the natural limitation that only integer parts of RAN and edge computing resources can be allocated. Hence, we developed specific approximation algorithms for obtaining an integer solution to the optimization problem for the particular values of $\alpha = 0$, $\alpha = 1$, $\alpha = 2$, and $\alpha \to \infty$. These algorithms rely on converting the continuous solution to an integer resource allocation. The reason for investigating these special α-values is described as follows: On the one hand, the guaranteed delay (constraint (1b)) is of interest to the individual users to fulfill their quality of service (QoS) requirements. On the other hand, the α-fairness is of interest to the network operator, as the allocation problem is a NUM. The specific α-values thereby offer the possibility to advertise different network properties. In detail, for $\alpha = 0$, the users are guaranteed a maximum delay, but if they experience good channel conditions, they might experience much better performance. For $\alpha = 1$, no user is punished for its channel conditions, i.e., other users' channel conditions have no influence on the experienced performance if there are no constraints [29], while for $\alpha \to \infty$, users with bad channel conditions are favored such that every user experiences roughly the same QoS. Lastly, for $\alpha = 2$, the overall delay is minimized, which implies that a minimum amount of resources is used.

Subsequently, first, the conversion algorithm for the edge computing resource allocation is presented. Afterward, the algorithms for the specific fairness cases are introduced. Throughout the following subsections, $\boldsymbol{J}_{\{u,d\}}$ indicates the $N \times K_{\{u,d\}}$ RAN allocation matrix with entries $J_{\{u,d\},ij} \in \{0,1\}$ and $\boldsymbol{n}$ denotes the $N \times 1$ edge computing resource allocation vector with entries $n_i \in \mathbb{N} \setminus \{0\}$. The variables $\boldsymbol{I}_{\{u,d\}}$ and $\boldsymbol{m}$ are their continuous equivalents.

A. Conversion Algorithm for Edge Computing Resources

Simple mathematical rounding is conducted to convert the continuous edge computing resource allocation to an integer assignment. To prevent an allocation of more or less than L computing resources, a limit check is performed after rounding. The user with a continuous allocation value closest above $\star.5$, where $\star$ denotes an arbitrary integer, is assigned one computing resource less than it would have received by mathematical rounding if more than L edge computing resources were allocated. This procedure is executed until L edge computing resources are assigned. Likewise, the users closest below $\star.5$ will receive one more resource until L resources are assigned if less than L resources are allocated after the rounding. Algorithm 1 summarizes the outlined approach. Its complexity is $\mathcal{O}(N)$, i.e., it is linear.

[5] A convex function $f : \mathbb{R} \to \mathbb{R}$ is called self-concordant if it fulfills the inequality $|f'''(x)| \leq 2f''(x)^{3/2}$ for all $x \in \textbf{dom} f$ [26]. For a more thorough discussion on self-concordance, see Section 9.6 in [26].

[6] The vector $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$ denotes the Newton step for f at x, for $x \in \textbf{dom}\ f$ [26]. The Newton step is used in Newton's method, which is an iterative multidimensional search method used in optimizations.

Algorithm 1 Integer Edge Computing Resource Allocation

```
Input: $N$, $L$, $\boldsymbol{m}$
Output: $\boldsymbol{n}$
function ECRAlloc($N$, $L$, $\boldsymbol{m}$)
    for all $m_i$ do
        $n_i = \lfloor m_i + 0.5 \rfloor$
    end for
    if $\sum_{i=1}^{N} n_i > L$ (Case 1 (C1)) then
        $l = 1$, $k = 0$, create empty list $w$.
    else if $\sum_{i=1}^{N} n_i < L$ (Case 2 (C2)) then
        $l = -1$, $k = 0$, create empty list $w$.
    end if
    while $\sum_{i=1}^{N} n_i > L$ (C1) or $\sum_{i=1}^{N} n_i < L$ (C2) do
        for $i = 1$ to $N$ do
            if $i \notin w$ then
                $r_i = m_i \mod \lfloor m_i \rfloor - 0.5$
                if $r_i \in ]0, l[$ (C1) or $r_i \in ]l, 0[$ (C2) then
                    $l = r_i$, $k = i$
                end if
            end if
        end for
        $n_k = \begin{cases} \lfloor m_i \rfloor, & \text{(C1)} \\ \lceil m_i \rceil, & \text{(C2)} \end{cases}$, attach $k$ to list $w$.
    end while
    return $\boldsymbol{n}$
end function
```

Algorithm 2 Integer Resource Allocation for $\alpha = 0$

```
Input: $N$, $K_u$, $K_d$, $L$, $\boldsymbol{m}$, $\boldsymbol{I}_u$, $\boldsymbol{I}_d$, $\boldsymbol{\Phi}_u$, $\boldsymbol{\Phi}_d$
Output: $\boldsymbol{n}$, $\boldsymbol{J}_u$, $\boldsymbol{J}_d$
1:  function AllocA0($N$, $K_u$, $K_d$, $L$, $\boldsymbol{m}$, $\boldsymbol{I}_u$, $\boldsymbol{I}_d$, $\boldsymbol{\Phi}_u$, $\boldsymbol{\Phi}_d$)
2:      $\boldsymbol{n} = \text{ECRAlloc}(N, L - N, \boldsymbol{m}) + \mathbf{1}$
3:      $\boldsymbol{J}_u = 0$, $\boldsymbol{J}_d = 0$
4:      for $i = 1$ to $N$ do
5:          Calculate $w_{\{u,d\},i} = \sum_{j=1}^{K_{\{u,d\}}} I_{\{u,d\},ij} \Phi_{\{u,d\},ij}$.
6:      end for
7:      Create list $z$ with users $i$ ordered
8:      s.t. $\Delta_u / w_{u,i} + \Delta_d / w_{d,i}$ is decreasing.
9:      while list $z$ is non-empty do
10:         for user $i$ in list $z$ do
11:             for uplink $u$ and downlink $d$ do
12:                 Find $\arg\max_j I_{\{u,d\},ij} \Phi_{\{u,d\},ij}$.
13:                 if $\exists$ more than one $j$ then
14:                     Choose randomly between those $j$.
15:                 end if
16:                 Allocate PRB $j$ to user $i$,
17:                 update $\boldsymbol{J}_{\{u,d\},j}$ and set $\boldsymbol{I}_{\{u,d\},j} = \mathbf{0}$.
18:             end for
19:             Calculate delay $\delta_i$ using $n_i$ and $\boldsymbol{J}_{u,i}$, $\boldsymbol{J}_{d,i}$.
20:             if $\delta_i \leq T_{max}$ then
21:                 Remove user $i$ from list $z$.
22:             end if
23:         end for
24:     end while
25:     for all non-allocated uplink/downlink PRBs $k$ do
26:         Find $\arg\max_i \Phi_{\{u,d\},ik}$.
27:         Allocate PRB $k$ to user $i$ and update $\boldsymbol{J}_{\{u,d\},k}$.
28:     end for
29:     return $\boldsymbol{n}$, $\boldsymbol{J}_u$, $\boldsymbol{J}_d$
30: end function
```

B. No Fairness

In the case $\alpha = 0$ (throughput maximization), when neglecting all constraints, every PRB would be assigned to the user with the highest CQI value across that PRB. Moreover, as each edge computing resource offers the same processing rate and hence contributes in the same way to the objective no matter to which user the resource is assigned, the allocation of the edge computing resources could be done randomly. However, each user must fulfill the delay constraint (1b), meaning that its packet must be sent and processed and a response must be received within the maximum time T_{max}. Thus, users experiencing worse channel conditions are allocated more computing resources such that the number of necessary PRB allocations for that user is minimized because allocations of PRBs to users with low CQI values negatively impact the maximization of the overall objective.

The approximation algorithm for $\alpha = 0$ can be explained as follows: First, all users are allocated enough edge computing and RAN resources such that they can fulfill their delay constraints, see lines 2 to 24 from Algorithm 2. This is done using the continuous allocations $\boldsymbol{I}_{\{u,d\}}$ and $\boldsymbol{m}$. Note that when finding the optimal solution to the continuous problem only $L - N$ edge computing resources are allocated. Afterward, one "extra" resource is assigned to each user during the conversion process, such that the integer edge computing resource allocation per user is at least as high as the continuous allocation. This ensures the feasibility of the integer solution.

For the RAN resource allocation, users are ordered such that those who were assigned the scarcest amount of continuous resources get their fixed integer allocation for complying with the delay constraint first.[7] Afterwards, the remaining PRBs are assigned to the users experiencing the best channel conditions. The procedure is recapitulated in Algorithm 2. The complexity of this algorithm is $\mathcal{O}(N + K_u + K_d)$.

C. Proportional Fairness

Unconstrained proportional fairness is characterized as the assignment where every user gets the same amount of resources, independent of the channel conditions it is experiencing (assuming the same CQI for all PRBs of a user). Mathematically, this can be explained as follows: for $\alpha = 1$, the pure objective is to maximize the sum of the natural logarithms of the RAN data and edge processing rates. The natural logarithm is characterized by the fact that its output value increases by a constant number whenever the argument of the logarithm

[7] The described procedure for the allocation of the edge computing resources as well as the assignment of the RAN resources needed to fulfill the delay constraint applies to the approximation algorithms for all four types of fairness considered in this work.

Algorithm 3 Integer Resource Allocation for $\alpha = 1$

```
Input: $N$, $K_u$, $K_d$, $L$, $\boldsymbol{m}$, $\boldsymbol{I}_u$, $\boldsymbol{I}_d$, $\boldsymbol{\Phi}_u$, $\boldsymbol{\Phi}_d$
Output: $\boldsymbol{n}$, $\boldsymbol{J}_u$, $\boldsymbol{J}_d$
 1: function ALLOCA1($N$, $K_u$, $K_d$, $L$, $\boldsymbol{m}$, $\boldsymbol{I}_u$, $\boldsymbol{I}_d$, $\boldsymbol{\Phi}_u$, $\boldsymbol{\Phi}_d$)
 2:     Follow lines 2 to 24 from Algorithm 2.
 3:     for $i = 1$ to $N$ do
 4:         Calculate $w_{\{u,d\},i} = \sum_{j=1}^{K_{\{u,d\}}} J_{\{u,d\},ij}$.
 5:     end for
 6:     Create lists $z_{\{u,d\}}$ with users $i$ ordered
 7:     s.t. $w_{\{u,d\},i}$ is increasing.
 8:     for all non-allocated uplink/downlink PRBs $k$ do
 9:         Take $z_{\{u,d\}}(1)$, find
10:         $\arg\min_k \left( \max_i \left( \Phi_{\{u,d\},ik} \right) - \Phi_{\{u,d\},z_{\{u,d\}}(1)k} \right)$.
11:         Allocate PRB $k$ to user $z_{\{u,d\}}(1)$
12:         and update $\boldsymbol{J}_{\{u,d\},k}$.
13:         Set $w_{\{u,d\},z_{\{u,d\}}(1)} = \sum_{j=1}^{K_{\{u,d\}}} J_{\{u,d\},z_{\{u,d\}}(1)j}$.
14:         Reorder list $z$ with users $i$
15:         s.t. $w_{\{u,d\},i}$ is increasing.
16:     end for
17:     return $\boldsymbol{n}$, $\boldsymbol{J}_u$, $\boldsymbol{J}_d$
18: end function
```

doubles. This implies that the objective value increases by the same amount irrespective of which user can double its resources. Conclusively, this provokes that every user should receive the same amount of resources since it is more costly (in terms of resources) to double the resources of a user who already has 6 PRBs than to double the resources of a user who only got assigned 2 PRBs.

In case the assumption of equal CQI values for all PRBs of a user does not hold, i.e., a user experiences very different channel conditions over the frequency range, the overall objective is still sensitive to these channel conditions. This means that a user should be allocated the PRBs for which he experiences the best channel conditions, even if the amount of PRBs should still be roughly the same for all users. These insights lead to the design of Algorithm 3. Thereby, again, first, the required resources for fulfilling the delay constraint are allocated and then the proportional fairness aim is followed when assigning the remaining PRBs. The allocation of the remaining RAN resources is done one after the other, where users are considered following an ordered list which is created according to a weight metric representing proportional fairness (see line 4 of Algorithm 3). Note that the remaining uplink and downlink resources can be handled independently, as their contribution to the overall objective value is combined with the sum of the uplink and downlink data rate a user is experiencing. This means that an allocation in the uplink does not influence the downlink objective values after fulfilling the delay constraint.[8] The complexity of Algorithm 3 is $\mathcal{O}(N + K_u + K_d)$.

D. Minimum Potential Delay Fairness

For the $\alpha = 2$ scenario, the prefactor in the objective (2) turns into -1, transforming the maximization problem into a minimization problem. Furthermore, the exponent of the RAN data rate and the edge processing rate of each user turns into -1 as well, leading to the minimization of the reciprocals of these rates. Comparing this objective function with the left-hand side of the delay constraint (1b), it is observable that the two functions are the same, with the only difference being the missing packet size $\Delta_{\{u,d\}}$ in the objective function. Since the packet sizes are equal for all users, they are, however, just a constant not influencing the optimization. To minimize the overall system delay, the knowledge of all RAN assignment combinations is needed. Since this knowledge is nonexistent when applying the approximation algorithm, the focus of the developed algorithm is to minimize the maximum experienced delay by any user. The algorithm for $\alpha = 2$ works in the exact same way as Algorithm 3, with the only two differences being the weight formulas representing the fairness, i.e., the equation in line 4 reads as

$$w_{\{u,d\},i} = \sum_{j=1}^{K_{\{u,d\}}} J_{\{u,d\},ij} \Phi_{\{u,d\},ij},$$

and the equation in line 13 is

$$w_{\{u,d\},z_{\{u,d\}}(1)} = \sum_{j=1}^{K_{\{u,d\}}} J_{\{u,d\},z_{\{u,d\}}(1)j} \Phi_{\{u,d\},z_{\{u,d\}}(1)j}.$$

E. Max-Min Fairness

Unconstrained max-min fairness, i.e., $\alpha \to \infty$, corresponds to the minimization of the sum of the reciprocals of the data and processing rates raised to the power of a large positive number. Once each user's data and processing rates are equal to each other, this minimization is achieved. This means that the edge computing resources are split equally among the users and the PRBs are allocated such that the difference between the users' data rates is minimized while the minimum data rate any user is experiencing is maximized. When a delay constraint is introduced, resources are allocated such that every user fulfills its constraint, which implies that the minimum data rate achieved by any user might decrease and the differences between the users' data rates might increase. The redistribution might also lead to a larger imbalance between the users' processing rates. The algorithm adjusted for $\alpha \to \infty$ follows the same concept as the approximation algorithms for the other special fairness cases. Again, the weight formulas in Algorithm 3 are the only factors that need to be adjusted to get the algorithm for max-min fairness. In that sense, the equation in line 4 is given as

$$w_{\{u,d\},i} = \left(\sum_{j=1}^{K_{\{u,d\}}} J_{\{u,d\},ij} \Phi_{\{u,d\},ij} \right)^{|1-\alpha|},$$

and the equation in line 13 reads as

$$w_{\{u,d\},z_{\{u,d\}}(1)} = \\ = \left(\sum_{j=1}^{K_{\{u,d\}}} J_{\{u,d\},z_{\{u,d\}}(1)j} \Phi_{\{u,d\},z_{\{u,d\}}(1),j} \right)^{|1-\alpha|}.$$

[8] This fact also applies to the approximation algorithms for the cases $\alpha = 2$ and $\alpha \to \infty$.

VI. Performance Evaluation

In the penultimate section of this work, we first describe our simulation setup and introduce the benchmarks. This is followed by an evaluation of the different fairness cases and analyses on the effect of fairness on the system throughput as well as the fairness scores of individual users.[9]

A. Simulation Setup

A 5G trace with data measured in the Republic of Ireland was used as input to the simulations. A detailed description of the traces can be found in [30], and a statistical analysis is given in [31]. The CQI with 15 levels is the parameter of interest from the trace, which specifies a user's experienced rate in a radio frame. The corresponding data rates per CQI are given in Table III. The measurements were conducted for a single user, but on various days, for different applications, and when the user was static or moving around. Only measurements where the user was moving were picked for the simulations in order to mimic the dynamic nature of the users.

Since there is only one CQI value given per time step in each measurement, the per-PRB CQI values were derived from the measured CQI value, denoted $\overline{\text{CQI}}$, by generating a population of CQI values in $\{\overline{\text{CQI}} - 1, ..., \overline{\text{CQI}} + 1\}$ or $\{\overline{\text{CQI}} - 3, ..., \overline{\text{CQI}} + \ 3\}$, respectively. Thereby, the mean value of the population is equal to the measured indicator $\overline{\text{CQI}}$ and the amounts of values per CQI were uniformly distributed when possible. In case the range of CQI values would exceed the technically possible range of CQI values, a biased distribution (larger amount of CQI values close to the boundary of the possible range) with the mean being $\overline{\text{CQI}}$ was used in order to still adequately represent the measurement. In case the CQI value was 1 or 15, no population was generated since CQI values of 0 and 16 are impossible and hence the population's mean would not have been equal to $\overline{\text{CQI}}$. The values from the population were then randomly assigned to the PRBs in the frequency range. Various measurements were taken to mimic various users.

The simulation parameters used in this evaluation are summarized in Table IV. Note that the presented mathematical analysis and subsequent evaluations are oblivious to the chosen subcarrier spacing (SCS). Only the achievable data rates dependent on the SCS do influence the achievable latencies, however, both for the optimal allocation and the assignment attained with the approximation algorithms. For all types of fairness, simulation data were gathered for different combinations of maximum latencies and user numbers as well as for the CQI populations with small and large variance. The simulations were conducted in MATLAB R2022b. To solve the optimization problems, CVX [32] together with Mosek [33] was used. In all cases, the solutions from the approximation algorithms are compared to the continuous (cont.) optimum obtained by solving the relaxed optimization problem, i.e., an upper bound. This procedure is followed since an integer optimum satisfying our accuracy requirements could not be found due to the NP-hardness of the integer optimization problem. Although the continuous allocation (allowing for arbitrary splitting of PRBs) is infeasible in reality, it gives a good indication of the performance of our algorithms.

B. Benchmarks

In total, besides the upper bound, our heuristics are compared to four different benchmarks. The first one is the Round-Robin (RR) principle [34]. This means all users are allocated one computing resource and one uplink and downlink PRB in each iteration. The PRBs are allocated one after another from the lowest to the highest frequency and independent of the users' channel conditions. Once a user fulfills its delay constraint, it will not be assigned any more resources until every user complies with its latency target. Thereafter, the remaining computing and RAN resources are allocated one by one to all users, until no resources are available anymore. The second benchmark policy provides a constrained rate variability (CRV) [12], where every user's PRB share corresponds to the fraction of the reciprocal of its experienced data rate divided by the sum of all reciprocals of the users' experienced data rates. Since in [12] it is assumed that the CQI value reported for one user is valid for the entire frequency range, the average CQI over a user's CQI values $\Phi_{\{u,d\},i}$ is taken as input to the allocation algorithm for all the user's PRBs. After the allocation of the PRBs according to the presented algorithm, the performance is measured based on the actual experienced CQI distributions $\mathbf{\Phi}_{\{u,d\}}$. Lastly, the assumption that one CQI value is reported per PRB reflects possible developments within the scope of 6G. The current 5G standards support either a wideband (WB) or a subband (SB) CQI reporting, where the width of the SB depends on the used bandwidth part size [22]. Thus, we compare the results from our heuristics for three different CQI input granularities (per-PRB CQI, SB CQI (8 PRBs), and WB CQI). Note that the resource allocation is done based on the various input types, the performance is however evaluated given the exact experienced per-PRB channel conditions. With this comparison, the potential of a more granular CQI reporting is shown.

C. Results for No Fairness (Throughput Maximization)

In Fig. 2a), the average objective value is shown for every possible combination of the number of users N and the delay constraint T_{max} for the CQI input with large variance. Thereby, and also in all following average plots, the averages are taken over 100 measurement points. Moreover, the average system througput, i.e., the data rates gained by RAN resource allocation, is depicted in Fig. 3a) for selected scenarios of both CQI distributions. For the bechmarks where only limited channel state information is available during the allocation, it can happen that the delay constraint is not fulfilled. In this case, the average is taken only over the valid measurement points. When analyzing the objective values (and also the system throughputs) for a specific number of users, it is observable that the average value decreases when tightening

[9] Corresponding to footnote 1, it is noted here that insignificant changes were observed when relaxing the reliability requirement from 100% to 99.999%. Due to the nature of the considered vehicular applications in this work, any further reduction in the required reliability would pose serious risks on human lives, which is the reason why no further reductions are considered.

TABLE III: Per-PRB rates for different CQIs [11]

CQI	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R (kbps)	48	73.6	121.9	192.2	282	378	474.2	712	772.2	874.8	1063.6	1249.6	1448.4	1640.6	1778.4

TABLE IV: Simulation Parameters

subcarrier spacing (SCS)	30 kHz
slot duration	0.5 ms
number of uplink PRBs (K_u)	80
number of downlink PRBs (K_d)	100
number of edge computing resources (L)	120
processing rate per edge computing resources (p)	500 kbps
uplink packet size (Δ_u)	6 kbit
downlink packet size (Δ_d)	4 kbit
number of users (N)	$\{5, 8, 10\}$
maximum latency (T_{max})	$\{3, 5, 10\}$ ms

the delay constraint. This was especially detectable for the CQI population with a small variance. The reason for this reduction can be explained by the allocation of more PRBs to users that are experiencing worse channel conditions, which is needed to fulfill their latency requirements. Additionally, it is observable that the algorithm outperforms the RR as well as the CRV scheme by far and is very close to the average continuous optimum. A more granular CQI reporting leads to a better objective value and system throughput, as the PRBs can be utilized more efficiently. Moreover, regarding the number of invalid measurement points, three things were seen. For the CRV scheme, all delay constraints could be fulfilled. Although the algorithm uses a WB CQI value, the allocation rule ensures a kind of fairness among the users such that all of them receive enough PRBs. In case of WB and SB CQI reporting, on average, 45.8 or 33.8 delay constraint violations out of 100 simulations could be observed due to the mismatch between the experienced and the reported CQI values. Especially for the CQI distribution with large variance, a lot of constraint violations could be detected. When comparing the average values for the CQI populations with small and large variance for a specific scenario, it was noticeable that the average objective value (and also the system throughput) is higher for the case where users experience higher variance in their CQI values. This can be explained by the fact that users who experience a bad average CQI value sometimes still have comparably good channel conditions if they experience larger variances in their CQI distribution.

The deviation of the objective value from the algorithm to the continuous optimum (upper bound) is shown for different CQI inputs for four exemplary scenarios in Fig. 4. Two conclusions can be drawn from this plot. Firstly, the deviation from the continuous optimum decreases with an increasing delay constraint T_{max}, as fewer PRBs need to be allocated to users with bad channel conditions with increasing T_{max}, which also leads to fewer allocations of *split* PRBs in the continuous assignment. Secondly, in most cases, the deviation for a scenario with a large variance of the CQI population is less than for a scenario with a small variance. In case there is a large variance in the per-user CQI populations, also the different CQI values per PRB have a larger variance, which almost leads to an integer solution when solving the relaxed optimization problem with continuous decision variables. The second observation is also reflected when comparing the maximum and average deviations among 100 data points among all scenarios for $\alpha = 0$; the maximum deviation for the small variance input is 1.79% (average: 0.47%), while it is 0.99% (average: 0.17%) for the large variance input. These values prove the very good performance of our algorithm in the no fairness case.

D. Results for Proportional Fairness

When looking at the average objective value and at the average system throughput in Fig. 2b) and Fig. 3b) for various scenarios, it is observable that all benchmarks are again outperformed by the approximation algorithm. Due to the similarity of the objectives of the RR principle and the proportional fairness, the results from the RR scheme are much closer to the algorithm objective values and to the optimal objectives than for $\alpha = 0$. The reason for this behavior is the allocation of the same amount of resources to all users in the proportional fair case. This is very similar to the RR principle, where all the users are assigned one resource after another until no resources are available anymore. Because users get assigned the resources where the difference between the data rate they experience and the maximum experienced data rate of any user for a PRB is the smallest, the approximation algorithm for proportional fairness still outperforms the benchmark algorithm. Due to the introduced fairness, only very few delay constraint violations could be observed for WB and SB CQI reporting. Again, a more precise CQI report leads to a better utility and system throughput.

Another observation from Fig. 3b) is that the system throughput does not depend on the delay constraint. Furthermore, also no influence of the number of present users in the network on the system throughput could be detected. The throughput depends, however, on the variance of the CQI values, i.e., the average system throughput is larger for CQI values with a larger variance due to the same reasons as in the no fairness case.

Finally, from the deviation plot in Fig. 5, it is recognizable that the approximation algorithm performs worse for CQI inputs with large variance. The reason for this deterioration is characteristic to proportional fairness, which is achieved easiest in case all CQI values of a single user are the same. Moreover, in our simulations, the algorithm performs worse if fewer users are present. For the case of 5 users, a lot of PRBs are split among these users in the continuous solution in order to achieve the same data rate for every user. In case there are more users, the amount of splitted PRBs decreases, as the variance of the CQI values of one PRB increases, which leads to a more integer-like solution of the continuous optimization problem. Still, our algorithm for $\alpha = 1$ exhibits an excellent performance. The maximum deviation to the continuous optimum among 100 data points and all scenarios is 0.14% (CQIs with small variance) or 0.32% (CQIs with large variance), respectively, whereas the average deviation is only 0.04% (CQIs with small variance) or 0.13% (CQIs with large variance).

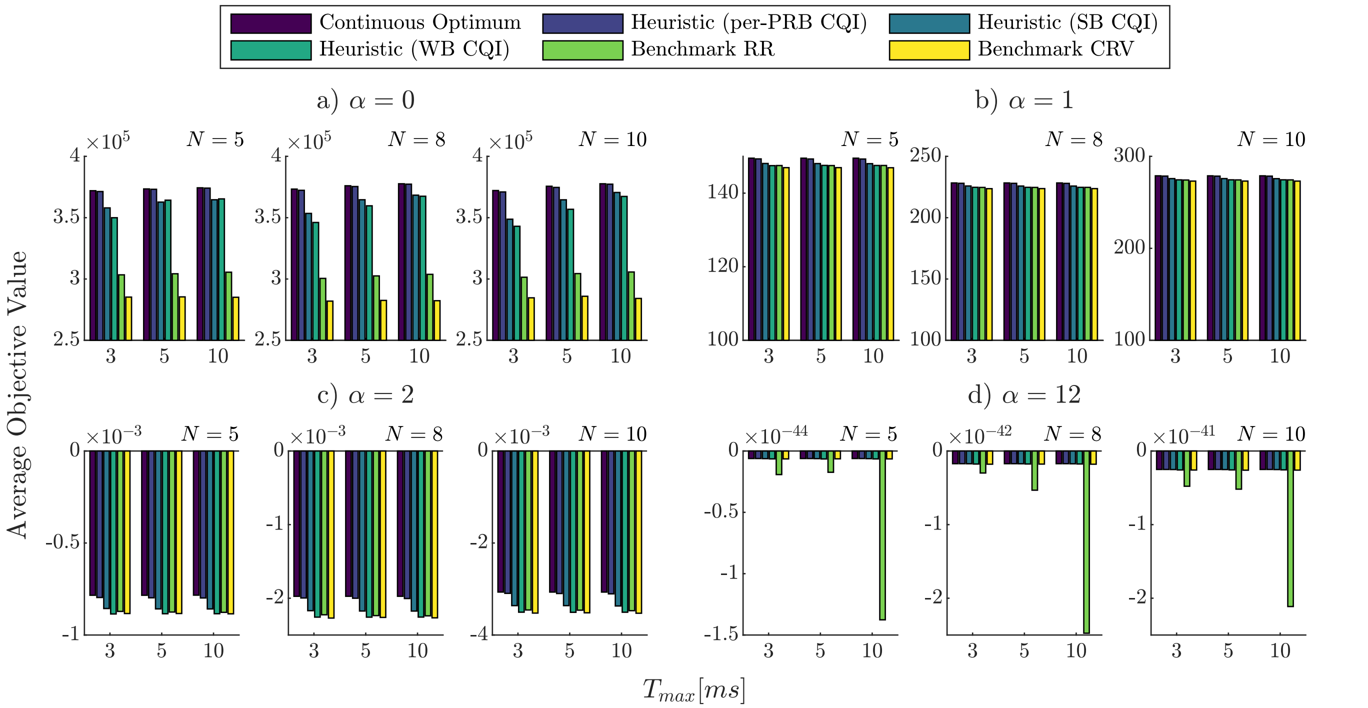

Fig. 2: Average objective values for no fairness ($\alpha = 0$), proportional fairness ($\alpha = 1$), minimum potential delay fairness ($\alpha = 2$), and max-min fairness ($\alpha = 12$) for CQI $\in \{\overline{\text{CQI}} - 3, \ldots, \overline{\text{CQI}} + 3\}$.

E. Results for Minimum Potential Delay Fairness

From the average objective values depicted in Fig. 2c), it is discernible that the special approximation algorithm for $\alpha = 2$ surpasses all benchmarks, as was the case for $\alpha = 0$ and $\alpha = 1$. Since a larger α-value corresponds to a more regular distribution of the resources, no delay constraint violations were observed in any of the results. It is observable that the RR principle outperforms both the result achieved with the heuristic and WB CQI input as well as the CRV scheme. To minimize the maximum delay, it is beneficial to allocate PRBs to the users which benefit most, i.e., the users which will experience the best data rates. However, Fig. 2c) shows that this is very hard given the limited channel state information, especially in case of strongly varying channel conditions, as then many transmission errors can occur or the potentially good channel cannot be fully utilized. When evaluating the average system throughputs, no dependence on the maximum acceptable delay was observable. They depend, however, on the number of users that are present in the system, which can be seen in Fig. 3c). The reason for this dependence is the objective of the approximation algorithm, i.e., minimizing the maximum encountered delay. In case more users are present, a user with bad channel conditions gets less PRBs, as the resources are shared among more users, which in the end leads to higher system throughput. Once more, for the same reasons as in the previous two fairness cases, the system throughput is larger in case the experienced channel conditions show a larger variance.

In most cases, the objective value obtained with the approximation algorithm is very close to the upper bound, see Fig. 6. However, for the measurements 15 to 18, some outliers are detectable in the deviation plot. The reason for these outliers is the presence of a user who is experiencing very bad channel conditions compared to the other users. Due to the slightly changed objective of the approximation algorithm, there are certainly resource allocations possible where the objective value can be maximized, i.e., the overall system delay is minimized, compared to the minimization of the maximum encountered delay by any user. This applies especially in case there are users whose data rates are a lot worse than all other users' data rates. The cost for this objective maximization is an increased delay experienced by the user with bad channel conditions. It is noticeable that the impact of this user with bad channel conditions gets smaller the higher the number of users in the network is. Furthermore, it can be seen that the variance of the CQI inputs does not have an influence on the deviation from the optimum for the delay minimization algorithm. Despite the aforementioned drawback, it can be concluded that the performance of the approximation algorithm for minimum potential delay fairness is still very good, as the maximum observed deviation from the continuous optimum is 11.62% ($\overline{\text{CQI}} \pm 1$) or 11.41% ($\overline{\text{CQI}} \pm 3$), while the average deviation among 100 data points and across all scenarios is only 1.59% ($\overline{\text{CQI}} \pm 1$) or 1.92% ($\overline{\text{CQI}} \pm 3$).

F. Results for Max-Min Fairness

Finally, also for the max-min fairness, satisfying evaluation results were obtained. The highest possible α that allowed for acceptable simulation outcomes was $\alpha = 12$. Higher values lead to numerical issues during the optimization process of the solver. The channel-agnostic RR scheme performs very poor in the presence of a user with bad channel conditions, which highly influences the average objective values depicted in Fig. 2d). This is not the case for the CRV algorithm, as it is

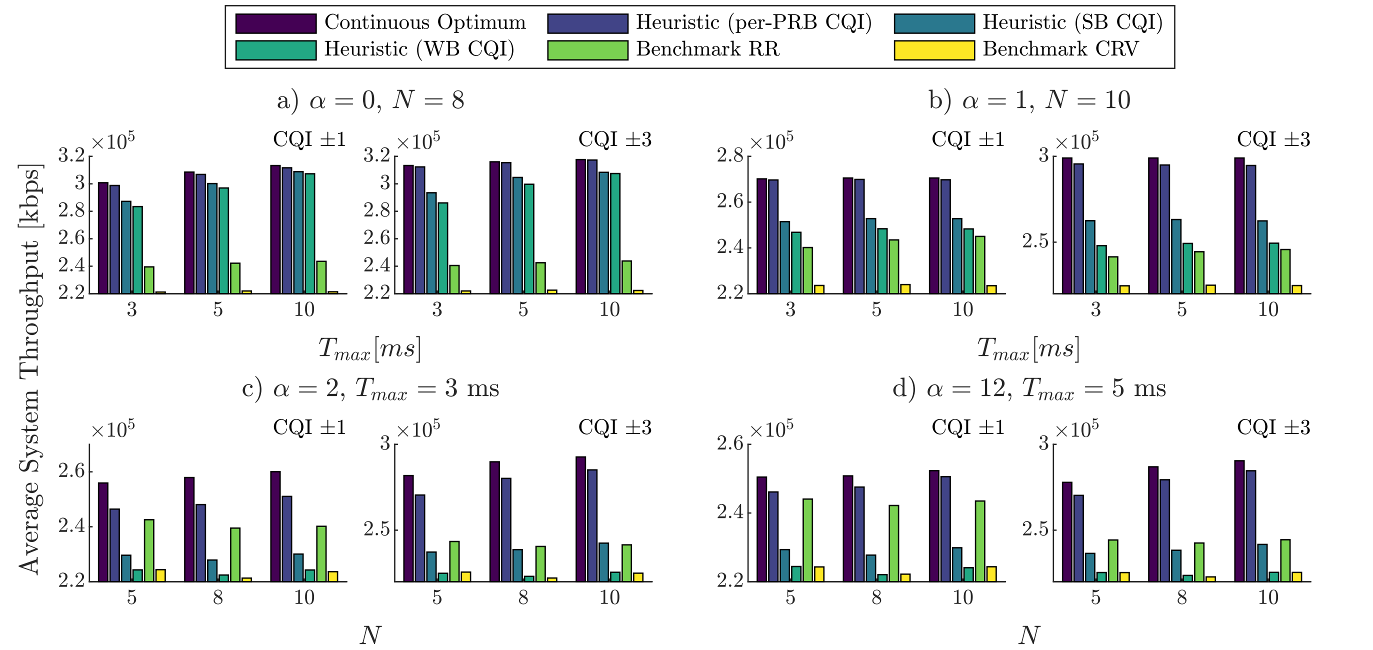

Fig. 3: Average system throughputs for no fairness ($\alpha = 0$), proportional fairness ($\alpha = 1$), minimum potential delay fairness ($\alpha = 2$), and max-min fairness ($\alpha = 12$) for selected scenarios.

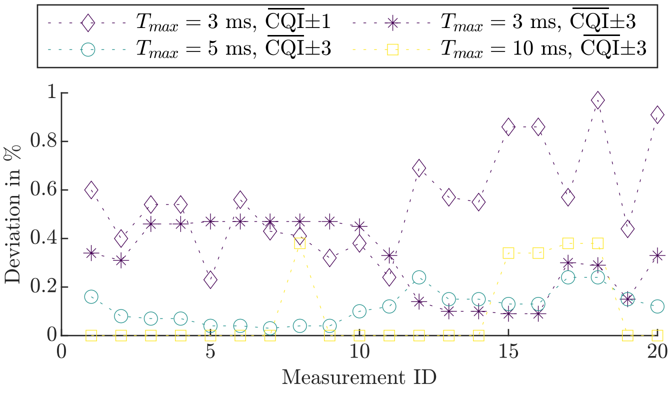

Fig. 4: Deviation of the objective value from the cont. optimum for different CQI inputs for $\alpha = 0$ and $N = 8$.

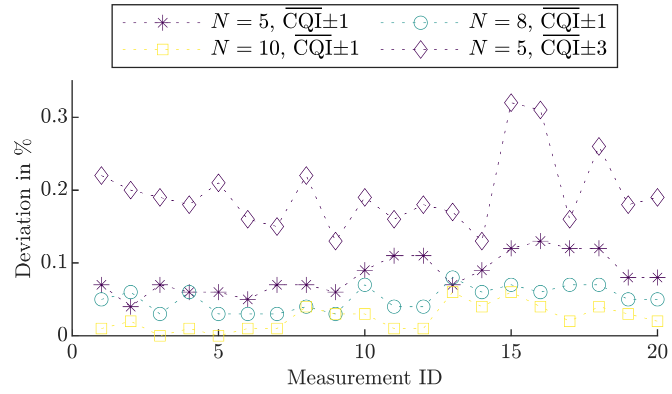

Fig. 5: Deviation of the objective value from the cont. optimum for different CQI inputs for $\alpha = 1$ and $T_{max} = 10$ ms.

based on reported CQI values. Nevertheless, both benchmark algorithms are outperformed by the proposed heuristic solution. In the deviation plot in Fig. 7, it is observable that also the approximation algorithm shows a much larger deviation in case of the presence of a user with very bad channel conditions (measurements 15 and 16) than for other inputs. The reason for this observation is that $\alpha = 12$ is only an approximation of $\alpha \to \infty$. Therefore, the presented results are only an approximation of the max-min fairness. When comparing the minimum data rate that any user is encountering in the optimal continuous solution and in the solution from the algorithm, it was perceivable that this data rate is larger in the solution from the algorithm. Thus, the algorithm actually provides a better solution in the sense of max-min fairness. However, since $\alpha = 12$ is not perfectly equal to max-min fairness, there are allocation scenarios where the minimum data rate a user is experiencing is worse than observed in the algorithm solution but the overall objective value is still better. Concluding, this implies that the degraded performance of the approximation algorithm is only detectable due to the approximation of $\alpha \to \infty$ and will diminish the greater α gets.

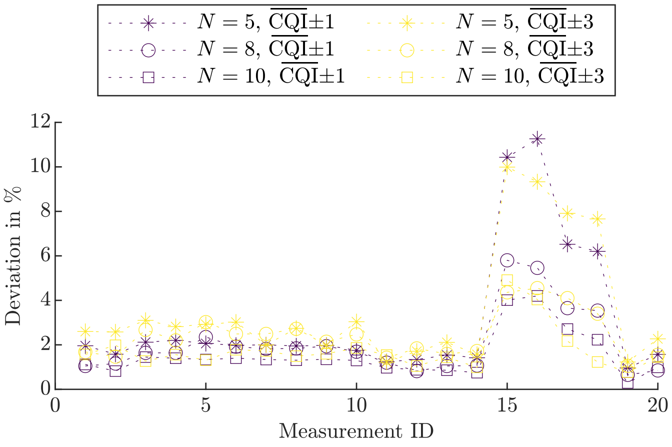

Fig. 6: Deviation of the objective value from the cont. optimum for different CQI inputs for $\alpha = 2$ and $T_{max} = 3$ ms.

TABLE VI: Per-user utilities (fairness scores) per α-value (fairness) of a selected simulation run for $N = 5$, $T_{max} = 3$ ms, and $\overline{\text{CQI}} \pm 3$ for the continuous optimum, the heuristic, and the RR scheme

Per-User Utility / α-value	Continuous Optimum					Heuristic					RR scheme				
	User 1	User 2	User 3	User 4	User 5	User 1	User 2	User 3	User 4	User 5	User 1	User 2	User 3	User 4	User 5
0 $[10^4]$	5.99	11.36	11.32	5.70	2.04	6.17	15.89	6.94	5.61	1.71	5.54	6.68	6.58	4.06	4.78
1	29.93	30.21	30.21	28.83	29.77	29.94	30.08	30.11	28.79	29.68	29.35	29.79	29.75	28.53	28.96
2 $[10^{-4}]$	-1.57	-1.52	-1.52	-1.82	-1.61	-1.63	-1.62	-1.63	-1.65	-1.65	-1.77	-1.61	-1.62	-2.23	-1.97
12 $[10^{-46}]$	-1.22	-1.22	-1.22	-1.23	-1.22	-1.23	-1.23	-1.23	-1.23	-1.23	-1.23	-1.96	-1.96	-1.32	-1.28

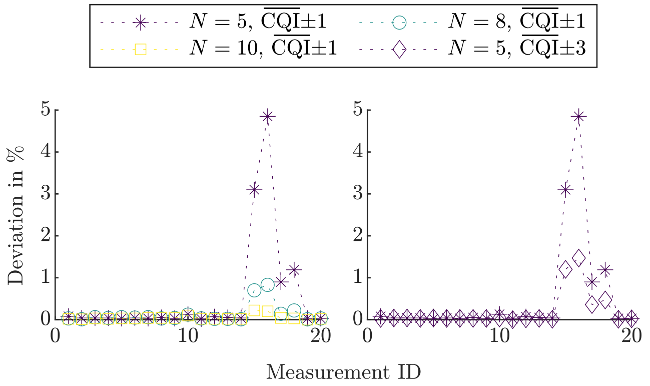

Fig. 7: Deviation of the objective value from the cont. optimum for different CQI inputs for $\alpha = 12$ and $T_{max} = 5$ ms.

Note that the influence of this approximation also decreases with an increasing number of users (less influence of a user's bad channel conditions), as could be observed for the case of minimum potential delay fairness (see Fig. 7). Furthermore, it is detectable that a CQI population with large variance leads to a lower deviation from the optimum, as a larger CQI variance allows for an integer-like solution of the continuous optimization problem.

For the case of max-min fairness, the system throughput again increases with the number of users that are present in the system. The rationale behind this observation is the same as for the minimum potential delay fairness. Moreover, as before ($\alpha = 12$), no connection between the delay constraint and the system throughput could be observed and the system throughput is again higher for CQI inputs with larger variance.

Even though the maximum deviation from the algorithm to the upper bound among 100 data points and across all scenarios is 7.20% (CQIs with small variance) or 1.47% (CQIs with large variance), the average deviation is only 0.1% (CQIs with small variance) or 0.03% (CQIs with large variance), which certifies the excellent performance of the approximation algorithm.

G. Effect of Fairness on the System Throughput

In the penultimate subsection of the performance evaluation, the influence of the fairness metric on the overall system throughput is shortly evaluated. To this end, the plots in Fig. 3 can be compared. It is observable that the system throughput decreases with an increasing amount of fairness that is introduced, i.e., with an increasing α. This is also reflected by the average system throughputs per fairness and CQI input distribution, which are given in Table V. Additionally, it is perceivable that the algorithms perform well compared to the continuous optimum also in terms of the overall system throughput, which again certifies the very good performance of all algorithms.

TABLE V: Average system throughput per α-value (fairness) for the continuous optimum and the heuristic

Sys. TP [10^5 bps] / α-value		0	1	2	12
$\overline{\text{CQI}} \pm 1$	Continuous Optimum	3.06	2.69	2.58	2.51
	Heuristic	3.04	2.68	2.48	2.48
$\overline{\text{CQI}} \pm 3$	Continuous Optimum	3.15	2.97	2.88	2.85
	Heuristic	3.14	2.91	2.78	2.78

H. Fairness among Different Users

In the last evaluation subsection, we provide insights on the per-user utility distributions, i.e., their achieved fairness scores. To this end, for one simulation run of a specific scenario configuration, all users' utilites are summmarized in Table VI. As expected, for $\alpha = 0$, the utilities are very diverse and reflect the experienced channel conditions of the users. Since the RR scheme is agnostic to these channel conditions, our approach outperforms this scheme by far, which can for example be seen when comparing the utilities of user 2 for both solution approaches. For higher α-values, the single utilities get closer to each other, as the resources are distributed such that all users experience the same data rate. Overall, it is observable that the users' utilities achieved with the heuristic are close to the continuous optimum users' utilities and larger than the ones from the RR scheme.

VII. Conclusion

In this paper, we addressed the problem of jointly allocating uplink and downlink RAN as well as edge computing resources to URLLC users so that their latency requirement is met, while simultaneously providing α-fairness. For the special cases $\alpha = 0$ (no fairness, i.e., throughput maximization), $\alpha = 1$ (proportional fairness), $\alpha = 2$ (minimum potential delay fairness), and $\alpha \to \infty$ (max-min fairness) we developed approximation algorithms with polynomial-time complexity. We have shown that their performance is very close to the optimum and that they considerably outperform the well-known RR as well as another SotA allocation scheme. Simulation results were obtained with input parameters taken from real datasets. For future research questions, we are going to implement the developed algorithms on a 5G testbed to evaluate the performance in a practical setup and we also plan to consider even more enhanced scenarios, e.g., including data storage and backhaul communication links.

References

[1] V. T. Haider, F. Mehmeti, A. Cantarero, and W. Kellerer, "Joint α-fair allocation of RAN and computing resources to vehicular users with URLLC traffic," in *Proc. of IEEE CCNC*, 2023.

[2] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, 2018.

[3] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. De Carvalho, Y. Ji, Č. Stefanović, *et al.*, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, 2018.

[4] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. of IEEE ISWCS*, 2018.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE Inst. Electr. Electron. Eng.*, vol. 106, no. 10, 2018.

[6] H. Halabian, "Distributed resource allocation optimization in 5G virtualized networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, 2019.

[7] 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 Trans. Vehi. Technol.*, vol. 69, no. 2, 2019.

[8] M. Centenaro, L. Vangelista, and S. Saur, "Analysis of 5G radio access protocols for uplink URLLC in a connection-less mode," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 5, 2020.

[9] H. Shariatmadari, S. Iraji, Z. Li, M. A. Uusitalo, and R. Jäntti, "Optimized transmission and resource allocation strategies for ultra-reliable communications," in *Proc. of IEEE PIMRC*, 2016.

[10] 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 (Spring)*, 2018.

[11] F. Mehmeti and T. F. La Porta, "Reducing the cost of consistency: Performance improvements in next generation cellular networks with optimal resource reallocation," *IEEE Trans. Mob. Comput.*, no. 7, 2022.

[12] 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.

[13] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," in *Proc. of IEEE VTC (Spring)*, 2019.

[14] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 7, 2021.

[15] J. Li and X. Zhang, "Deep reinforcement learning-based joint scheduling of eMBB and URLLC in 5G networks," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 9, 2020.

[16] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, vol. 69, no. 2, 2020.

[17] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, "eMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019.

[18] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, 2020.

[19] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and URLLC in 5G wireless networks," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1736–1749, 2021.

[20] A. Destounis and G. S. Paschos, "Complexity of URLLC scheduling and efficient approximation schemes," *arXiv preprint arXiv:1904.11278*, 2019.

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

[22] ETSI, "5G NR physical layer procedures for data: 3GPP TS 38.214 version 17.1.0 release 17." www.etsi.org, 2022. Technical Specification.

[23] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 2, 2020.

[24] R. Srikant, *The mathematics of Internet congestion control*. Springer, 2004.

[25] J. Lee and S. Leyffer, *Mixed integer nonlinear programming*, vol. 154. Springer Science & Business Media, 2011.

[26] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[27] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[28] R. Chares, *Cones and interior-point algorithms for structured convex optimization involving powers and exponentials*. PhD thesis, Université Catholique de Louvain Louvain-la-Neuve, Louvain, Belgium, 2009.

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

[30] 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.

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

[32] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1." http://cvxr.com/cvx, Mar. 2014.

[33] MOSEK ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 10.0.20.*, 2022.

[34] G. Miao, J. Zander, K. W. Sung, and S. B. Slimane, *Fundamentals of mobile data networks*. Cambridge University Press, 2016.

Valentin Thomas Haider received his bachelor's degree in electrical engineering and information technology from the Deggendorf Institute of Technology in 2020. He obtained his master's degree with high distinction from the Technical University of Munich in 2022. During his bachelor's and his master's studies, he received the SpeedUp and Fastlane scholarship from the BMW Group, respectively. Since November 2022, he is a research and teaching associate at the Chair of Communication Networks at the Technical University of Munich.

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

Ana Cantarero received her MSc degree in Communications Technology from the University of Ulm, Germany in 2012. At BMW Group she has done research work on connected vehicles across several national and European projects. Her emphasis has been V2X communications and advanced automotive use cases. From 2008 to 2010, she was a Radio Access Network Planning Engineer at Millicom International Cellular SA. From 2012 to 2018, she was a Software Developer for 4G and 5G Test Systems at Rohde & Schwarz GmbH.

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 with an ERC Consolidator Grant from the European Commission for his research on flexibility in communication networks. He currently serves as an associate editor for IEEE Transactions on Network and Service Management and as the area editor for network virtualization for IEEE Communications Surveys and Tutorials.