Enhanced Mobility Management with SD-RAN in 5G Networks

Anna Prado, Merve Ciki, Fidan Mehmeti, Wolfgang Kellerer
Chair of Communication Networks, Technical University of Munich, Germany.
{anna.prado, merve.ck, fidan.mehmeti, wolfgang.kellerer}@tum.de

***Abstract*—Mobility Management in 5G is challenging due to the usage of high frequencies and dense cell deployments. This often results in frequent handovers for users, causing disruptions in transmission and reception and adversely affecting network capacity. The crucial task is to integrate handover decisions with resource allocation, ensuring the target base station guarantees the minimum required user rate while optimizing metrics that are essential for the operator, such as network sum throughput. The dynamic allocation of resources to BSs, facilitated by Software-Defined Radio Access Network (SD-RAN), emerges as a solution for efficient resource utilization. This paper aims to maximize network sum throughput, ensure a minimum user rate, and minimize handovers. We adopt a two-level approach, integrating resource allocation and mobility management using SD-RAN. This is modeled as an integer nonlinear program, and by relaxing it, we obtain an upper bound. Given the NP-hard nature of the problem, we introduce two heuristics (deterministic and probabilistic) which yield near-optimal user-to-BS assignments and efficiently allocate resources to serving BSs and end users. Our proposed algorithms outperform state of the art, significantly reducing the handover rate while remaining within 2% of the optimum, with user rate satisfaction reaching 99%.**

***Index Terms*—Handover, mobility management, SD-RAN, 5G.**

I. Introduction

Software-Defined Radio Access Networks (SD-RANs) were proposed for the first time in 5G networks to enhance their flexibility and performance [1]. In this architecture, the control plane functions are decoupled from the data plane and are shifted to centralized entities known as SD-RAN controllers. This separation allows for more flexibility, efficiency, and dynamic management of resources in the Radio Access Network (RAN) [2].

Mobility management is one of the functions in 5G that can benefit from SD-RAN. In Ultra-Dense Networks (UDNs) [3], [4], mobility management becomes more challenging, and the traditional LTE/5G handover approach performs poorly due to the handling of numerous cells and users. Consequently, users may encounter issues such as handover failures, unnecessary back-and-forth handovers (referred to as *ping-pong* handovers), and an increase in signaling related to mobility. These challenges contribute to increased mobility-related latency and service interruptions. To address these issues and fully leverage the advantages of UDNs, the separation of control and data planes in Software-Defined Networking (SDN) appears to be an effective solution [3]. This becomes even more pronounced in later releases of 5G-advanced, where novel mobility solutions are expected to optimize multiple metrics such as interruption time, mobility robustness, and throughput [5].

In a traditional RAN, each Base Station (BS) is configured with a predetermined set of resources (static assignment) that can be allocated to the users within its coverage area. Differently, in SD-RAN, the controller has a broader view of the network and can assign the resources to BSs based on the current distribution of users in the network, as well as their channel conditions. Additionally, SD-RAN allows the controller to re-assign the resources among BSs instead of performing handovers, thus, facilitating mobility management. Handovers are often necessary not because the user leaves the coverage of the serving BS, but because there is a better BS (in terms of signal strength in case of the baseline handover or in terms of the user rate in our approach presented later). So, instead of a handover that causes a Handover Interruption Time (HIT), the user might stay connected to the serving BS when their channel conditions degrade, but are still acceptable. This is especially beneficial when a user experiences a short channel degradation, e.g., due to intermittent blockage of Line of Sight (LoS).

Further, the possibility of global knowledge in SD-RAN empowers the controller to take into account the channel conditions and requirements of all users. In the conventional LTE/5G handover algorithm [6], the serving BS must initiate a request before a handover to the target BS to verify if it possesses enough resources for accommodating the user. If the target BS has the necessary resources, it responds with an acknowledgment. However, there is no mechanism to switch a user to another satisfactory BS to accommodate a newly arrived user, either to maximize certain utility or meet the user's rate, in the existing handover approach.

Therefore, optimizing handover decisions and minimizing their occurrence is of utmost importance. This goal is achieved by strategically assigning each user to the most suitable BS, as well as dynamically allocating Physical Resource Blocks (PRBs) to BSs according to the distribution of users and their

This work was supported in part by the Bavarian Ministry of Economic Affairs, Regional Development and Energy under the project "6G Future Lab Bavaria", and in part by the Federal Ministry of Education and Research of Germany (BMBF) under the project "6G-Life", with project identification number 16KISK002.
ISBN 978-3-903176-63-8 © 2024 IFIP

channel conditions. Different approaches have already been presented to reduce handover rate [4], [7], [8]. But, they suffer from certain shortcomings, either as being too challenging to implement or not considering all the causes for user handovers. For instance, in [7], the main triggering-event thresholds are based on the speed of the users, without taking into account the LoS blockages. Also, the user assignment should not be based solely on the signal-related metrics, because this may lead to overcrowded cells. In prior works, only resource allocation from BSs to the users is considered [4], [9], while the resources at every BS remain fixed. Modern networks should be able to adapt to user demands dynamically and have ideally the exact amount of required resources available without over- or under-provisioning. This can be achieved by utilizing SD-RAN to perform joint mobility management and two-level resource allocation.

Providing a smooth operation of cellular networks, with satisfied users who do not often experience service interruptions, is challenging, mainly due to the limited network resources and the dynamic nature of channel characteristics [10]. Also, the cellular operator aims at assigning users and allocating resources to the users so that the resources across different BSs are fully utilized to maximize the utility. Hence, of particular importance is to determine the number of PRBs a BS should be assigned at every time slot, as well as the number of PRBs every user should be assigned and with which BS should the user be associated.

Several important questions related to highly efficient mobility management in 5G networks arise:

- What is the policy that performs joint mobility management and two-level resource allocation with the goal of maximizing the sum throughput in the network, while satisfying user's rates and reducing handover rate?
- How does this approach perform compared to the conventional baseline and state-of-the-art algorithms?

To answer these questions, we model our problem aimed at reliably capturing system behavior, with the objective of optimizing network sum throughput (operator's Key Performance Indicator (KPI)) and guaranteeing a minimum rate to every user (end user's KPI). The main novelty lies in the utilization of a centralized controller, such as SD-RAN, for mobility management in 5G. We allocate the resources to the BSs dynamically using SD-RAN, and our approach is 3GPP-compliant with a minimal change in the signaling flow. Due to the complex nature of the problem, besides using a solver (Gurobi), we find an exact transformation of the original Integer Nonlinear Program (INLP) into an Integer Linear Program (ILP). Then, we relax the problem into a linear one and provide an upper bound in polynomial time. Finally, we propose two approaches on how to convert a Linear Program (LP) solution into a solution for the original problem, which can be obtained in polynomial time. Specifically, our main contributions are:

- We provide a solver-based solution to the optimization problem of jointly performing two-level resource allocation and mobility management.
- We provide the exact transformation of the nonlinear integer problem into an integer linear one, and then relax the integrality constraints. We provide a formulation of a linear problem that provides an upper bound to the original NP-hard problem.
- We propose two approaches on how to round the values of relaxed decision variables back to integer and obtain a solution for INLP/ILP from a solution for LP. Our approaches using SD-RAN perform close to the optimum, reduce significantly the handover and ping-pong handover rates while simultaneously achieving a high user rate satisfaction compared to the baseline solutions.

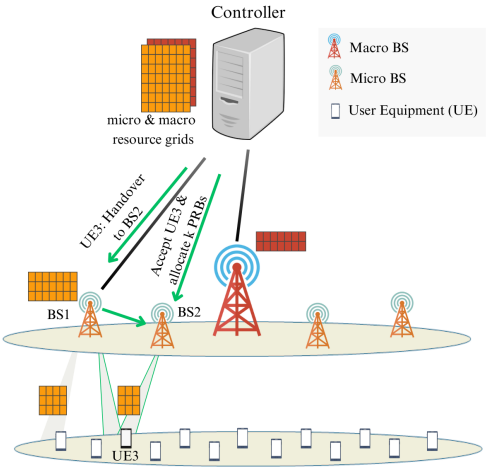


Fig. 1: Illustration of the system model.

II. Problem Formulation

We present the system model first and then provide the optimization formulation.

A. System Model

The network consists of multiple BSs, which are controlled by a central controller. The sets $\mathcal{U}$ and $\mathcal{B}$ denote the sets of all users and BSs in the network, respectively. Furthermore, we consider a two-tier scenario with two types of BSs: macro and micro: $\mathcal{B} = \mathcal{B}_{macro} \cup \mathcal{B}_{micro}$. The system model is depicted in Fig. 1. We assume a similar network architecture as in [3], where the controller performs two tasks: *mobility management* and *admission control*.

The controller allocates PRBs to BSs and their connected users from two pools, one for each BS type (macro and micro). One *PRB* (the unit of resource allocation in 5G) is defined as 12 consecutive subcarriers in the frequency domain and one slot in the time domain. Macro BSs have $\mu = 1$ with a subcarrier spacing of 30 kHz, while micro BSs have $\mu = 2$ with 60 kHz. The bandwidth per PRB is 360 kHZ for macro and 720 kHz for micro BSs.

We consider the problem over a time horizon of a duration T time slots. We denote by $R_{u,b}$ the rate per PRB of user $u \in \mathcal{U}$ from BS $b \in \mathcal{B}$. As such, $R_{u,b}$ depends on the Signal to Interference plus Noise Ratio (SINR) of the user, the type

of BS b and its bandwidth. We calculate $R_{u,b}$ using Shannon's formula, as in [4], [9]. Let b' denote the BS to which the user u was connected at time step $t-1$. If there was a handover at time t, $b \neq b'$. Otherwise, $b = b'$.

The focus is on an application that is throughput-sensitive, such as 4K Ultra HD content [11]. Therefore, the aim is to ensure that each user receives a minimum data rate, denoted as r_u that satisfies the traffic requirements. Latency-sensitive applications are deferred to future work.

The decision variable $x_{u,b}$ is a binary variable that states whether user *u* is connected to BS *b* or not. If user *u* is served by BS *b*, $x_{u,b} = 1$; otherwise, $x_{u,b} = 0$. The other decision variable $k_{u,b}$ denotes the number of PRBs the user u receives from BS b, thus, $k_{u,b}R_{u,b}$ is the rate of user *u* from BS *b* (when $x_{u,b} = 1$).

B. Optimization Formulation

In line with the objective of this paper, the formulation of the optimization problem is as follows:

$$\max_{x_{u,b},k_{u,b}} \quad \sum_{u=1}^{|\mathcal{U}|}\sum_{b=1}^{|\mathcal{B}|} x_{u,b}k_{u,b}R_{u,b}\cdot(1-\eta_{u,b',b}) \tag{1}$$

$$\text{s.t.} \quad \sum_{b=1}^{|\mathcal{B}|} x_{u,b} = 1, \qquad \forall u \in \mathcal{U}, \tag{2}$$

$$\sum_{b=1}^{|\mathcal{B}|} x_{u,b}\cdot k_{u,b}\cdot R_{u,b}\cdot(1-\eta_{u,b',b}) \geq r_u, \quad \forall u \in \mathcal{U}, \tag{3}$$

$$\sum_{u=1}^{|\mathcal{U}|}\sum_{b=1}^{|\mathcal{B}_{macro}|} x_{u,b}\cdot k_{u,b} \leq K_{macro}, \tag{4}$$

$$\sum_{u=1}^{|\mathcal{U}|}\sum_{b=1}^{|\mathcal{B}_{micro}|} x_{u,b}\cdot k_{u,b} \leq K_{micro}, \tag{5}$$

$$x_{u,b} \in \{0,1\}, \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}, \tag{6}$$

$$k_{u,b} \in \{0,1,...K_{macro}\}, \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}_{macro}, \tag{7}$$

$$k_{u,b} \in \{0,1,...K_{micro}\}, \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}_{micro}. \tag{8}$$

The objective (1) is to maximize network sum throughput considering the overhead to account for rate reduction during handover. Constraint (2) ensures that every user is served by exactly one BS in every slot. Constraint (3) guarantees a minimum rate to the user accounting for HIT. Constraints (4)-(5) state that the total number of PRBs in the network for macro BSs is limited by K_{macro}, whereas for micro it is K_{micro}. Solving this problem is of significant importance as it addresses the needs of both the operator (maximizing sum throughput) and the users, who seek a smooth service with a guaranteed rate and minimal mobility-related interruptions.

To account for the data rate reduction during a handover due to HIT, we consider the handover overhead $\eta_{u,b',b}$ in the objective (1) and constraint (3), which can be expressed in terms of the current allocation $x_{u,b}$ and the previous allocation $x'_{u,b}$ as $\eta_{u,b',b} = (1 - x'_{u,b}) \cdot \frac{T_{HIT}}{T_{slot}}$ [4], where T_{HIT} is the interruption time when there is a handover, and T_{slot} is the slot duration.

Note that $x'_{u,b}$ is fixed at the current time slot and is not a decision variable. The problem is formulated considering the previous user-to-BS assignment to be able to introduce a penalty for a handover $\eta_{u,b',b}$ in the objective. Then, the optimization problem is solved repeatedly at every time slot.

The problem (1)-(8) is an Integer Non-linear Program (INLP). Therefore, it is NP-hard [12]. Because of that, we can obtain the optimal user-to-BS assignment and the number of PRBs allocated to every user only using a solver, like Gurobi, assuming that the information about all users in the network is available at the beginning of the slot. We refer to this solution as *INLP* and evaluate its performance in Section VI.

In the next section, we transform the problem into an ILP and then the latter into an LP to derive the upper bound of the objective function in polynomial time.

III. Problem Transformation/Relaxation

Here, we first transform the INLP into an ILP by performing a variable replacement and adding some extra constraints. Then, we relax the integrality constraint on decision variables.

A. Transforming INLP into ILP

Let us examine the possible values of the product $x_{u,b}k_{u,b}$. If $x_{u,b} = 0$, the value of $k_{u,b}$ does not play a role and can be assumed to be 0. The product equals 0 as well. Alternatively, if $x_{u,b} = 1$, the product reduces to $k_{u,b}$. We replace the product $x_{u,b}k_{u,b}$ with a new variable $y_{u,b}$ such that

$$y_{u,b} = \begin{cases} k_{u,b}, & \text{if} \quad x_{u,b} = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The following constraints need to be added to the problem

$$0 \leq y_{u,b} \leq K_{macro}\cdot x_{u,b},\ \forall u \in \mathcal{U}, \forall b \in \mathcal{B}_{macro}, \tag{10}$$

$$0 \leq y_{u,b} \leq K_{micro}\cdot x_{u,b},\ \forall u \in \mathcal{U}, \forall b \in \mathcal{B}_{micro}. \tag{11}$$

The objective function (1) and constraints (3)-(5) become linear in terms of the newly introduced decision variable $y_{u,b}$:

$$\max_{x_{u,b},y_{u,b}} \quad \sum_{u=1}^{|\mathcal{U}|}\sum_{b=1}^{|\mathcal{B}|} y_{u,b}R_{u,b}\cdot(1-\eta_{u,b',b}) \tag{12}$$

$$\sum_{b=1}^{|\mathcal{B}|} y_{u,b}\cdot R_{u,b}\cdot(1-\eta_{u,b',b}) \geq r_u, \quad \forall u \in \mathcal{U}, \tag{13}$$

$$\sum_{u=1}^{|\mathcal{U}|}\sum_{b=1}^{|\mathcal{B}_{macro}|} y_{u,b} \leq K_{macro}, \tag{14}$$

$$\sum_{u=1}^{|\mathcal{U}|}\sum_{b=1}^{|\mathcal{B}_{micro}|} y_{u,b} \leq K_{micro}. \tag{15}$$

Since $y_{u,b}$ is the product of the binary variable $x_{u,b}$ and integer variable $k_{u,b}$, it takes the same values as $k_{u,b}$, specifically,

$$y_{u,b} \in \{0,1,...K_{macro}\}, \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}_{macro}, \tag{16}$$

$$y_{u,b} \in \{0, 1, ...K_{micro}\}, \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}_{micro}. \tag{17}$$

The ILP can be then expressed as (12), (2), (6), (10)-(11), (13)-(17). This problem formulation is equivalent to the one in Section II, and we show in Section VI that both problem formulations lead to exactly the same result. So, the problem formulation is simplified by replacing the non-linear terms with their transformed linear expressions.

B. Relaxing the integer variables

To obtain an upper bound in polynomial time, we proceed by relaxing $x_{u,b}$ from (6) and $y_{u,b}$ from (16)-(17). The decision variables $x_{u,b}$ and $y_{u,b}$ then become

$$x_{u,b} \in [0, 1], \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}, \tag{18}$$

$$y_{u,b} \in [0, K_{macro}], \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}_{macro}, \tag{19}$$

$$y_{u,b} \in [0, K_{micro}], \quad \forall u \in \mathcal{U},\ \forall b \in \mathcal{B}_{micro}. \tag{20}$$

This transformation and relaxation provide an upper bound to the original problem, denoted as LP. The LP formulation is expressed as (12), (2), (10)-(11), (13)-(15), (18)-(20).

This is a linear optimization, which can be solved with an off-the-shelf solver for linear programs, such as SciPy [13]. Utilizing LP allows us to efficiently establish an upper bound, which can be used to assess the performance of the algorithms introduced in the subsequent section. A reliable upper bound provides a rapid assessment of how closely our solution's objective value approaches the optimum in polynomial time.

IV. Proposed approaches

Our algorithms are developed by building upon LP discussed in Section III, after solving which the decision variables $x_{u,b}$ and $y_{u,b}$ must be converted back to an integer value. We propose two policies for this. We also introduce an additional constraint to handle issues arising from allowing continuous values for decision variables. Finally, we apply post-processing to ensure that the constraints of the optimization problem are not violated.

A. Avoiding mismatches in LP

Our simulations indicate that there are often "mismatches" between $x_{u,b}$ and $y_{u,b}$ when decision variables are continuous. For example, for user u, $x_{u,b}$ is close to 1 and the corresponding $y_{u,b} = 0$, while another $x_{u,b}$ that is almost 0 has its corresponding $y_{u,b}$ set very high. These values of $x_{u,b}$ and $y_{u,b}$ satisfy the constraints in LP, but the constraints become violated after rounding the decision variables to integer. Specifically, the minimum required rate constraint (13) and the resource constraints (14), (15). Hence, to restrict the values of $x_{u,b}$ and $y_{u,b}$, we add the following constraint

$$0 \leq x_{u,b} \leq y_{u,b},\ \forall u \in \mathcal{U}, \forall b \in \mathcal{B}, \tag{21}$$

which states that resource amount $y_{u,b}$ cannot be set to a non-zero value without setting the assignment variable $x_{u,b}$ to 1.

After adding constraint (21) to LP from Section III-B, the values of $y_{u,b}$ and $x_{u,b}$ "match", meaning that if BS b allocates resources to user u ($y_{u,b} > 0$), the corresponding $x_{u,b} \approx 1$. This facilitates the conversion back to an integer.

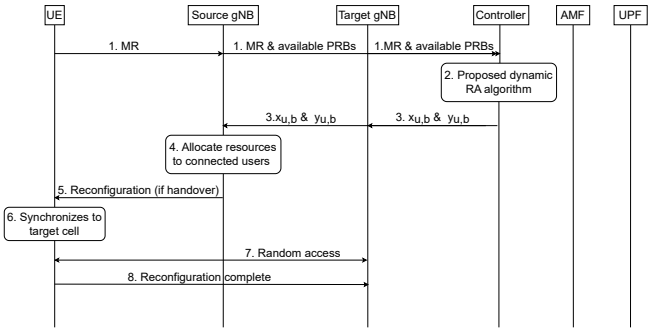


Fig. 2: Signaling diagram of the proposed approaches.

B. Rounding policies

1) Hard policy: Both decision variables, $x_{u,b}$ and $y_{u,b}$, are rounded to the nearest integer. This is a deterministic policy.

2) Soft policy: This approach adopts a probabilistic strategy. In the first step, to round $x_{u,b}$, a random value between 0 and 1 is sampled for each user at every time slot from a uniform distribution. We sort the array of $x_{u,b}$ values for every user from high to low. The likelihood of rounding an element to 1 is determined by its value in the array; in other words, the higher the value, the greater the probability of setting this $x_{u,b} = 1$. If the first $x_{u,b}$ is larger than the sampled value, we round it to 1 and stop. Otherwise, we set the value of the next $x_{u,b}$ to the sum of the first and the second $x_{u,b}$ values (hence, the cumulative probability). The process is repeated until an $x_{u,b}$ that is set to 1 is found.

Similarly, to convert $y_{u,b}$, we select probabilitstically a value from a distribution based on the weights calculated from the proximity of the current decision variable value to its floor and ceiling. For instance, if $y_{u,b} = 2.6$, we round it to 3 with a probability of 0.6 and to 2 with a probability of 0.4. Then, we again need to apply some post-processing to ensure that resources are not exceeded and user rates are satisfied.

C. Post-processing

Next, we implement post-processing on $y_{u,b}$ to prevent the violation of the constraints, which consists of 5 steps.

In the first step, we solve LP expressed in Eqs. (12), (2), (10)-(11), (13)-(15), (18)-(21) and round $x_{u,b}$ and $y_{u,b}$ using either soft or hard policies, denoted as LP-Hard Decision (LP-HD) and LP-Soft Decision (LP-SD), repeatedly. Then, if constraints (14)-(15) are violated, we reduce accordingly the number of PRBs allocated to the user (or users) with the largest $y_{u,b}$ (corresponds to the users with the best channels since our objective is to maximize the sum throughput) in macro and micro pools. Next, if some resources are still available, we allocate them to the user (or users) with the best channel, which aligns with our objective. There might be some resources available due to rounding down $y_{u,b}$ with any of the policies.

In the third step, we compute the minimum number of PRBs that is required to satisfy the user demand r_u, given per PRB rate $R_{u,b}$ that is computed based on the channel between user u and b at every time slot. Note that this is an integer value.

Then, we initialize multiple hash tables to store the data and to be able to look it up easily when performing the next operations, specifically, the mapping of BS pool to the available extra PRBs that are not necessary to satisfy the user rate in *bs_pool_extra_prbs* (dynamically updated) and *bs_pool_extra_prbs_const* (constant and used for reference), as well as the mapping of the user ID to the extra number of PRBs that it was allocated in *user_id_y_extra*.

In the forth step, we loop over the users with unsatisfied rate and try to satisfy it if there are available resources in the pool of their serving BS. We also update the hash table *bs_pool_extra_prbs* that tracks in every pool.

In the last step, we need to subtract the resources that we reallocated in the previous step from the users who received more PRBs than necessary to satisfy the user rate of unsatisfied users. If the hash tables are not the same, specifically, *bs_pool_extra_prbs_const* $\neq$ *bs_pool_extra_prbs*, which means that some resources were reallocated in step 3. We iterate over the users and BS pools and balance the resources to avoid the violation of resource constraints by subsequently deducting the number of reallocated PRBs from the *user_id_y_extra* hash table. The reallocated PRBs are subtracted from the extra PRB count of each user *user_id_y_extra*, which was sorted from high to low based on the number of extra allocated PRBs. Since the objective is to maximize the sum throughput, these are the users with the best channel conditions in the macro and micro pools of BSs.

D. Complexity analysis

The complexity of the first step from Section IV-C is $O(|\mathcal{U}| \cdot |\mathcal{B}|)$ because it requires rounding of every $x_{u,b}$ and $y_{u,b}$. Subsequent steps have a worst-case complexity of $O(|\mathcal{U}|)$ except for the last step that requires sorting, thus, has a complexity of $O(|\mathcal{U}| \cdot \log(|\mathcal{U}|))$. The total complexity of LP-HD and LP-SD algorithms becomes then $O(|\mathcal{U}| \cdot |\mathcal{B}| + |\mathcal{U}| \cdot \log(|\mathcal{U}|))$.

E. SD-RAN controller

Our approach reuses most of the signaling implementation proposed by 3GPP in [6] and the signaling flow is presented in Fig. 2. Users send their Measurement Reports (MRs) to their serving BSs, which forward them to the SD-RAN controller, where LP-HD and LP-SD are executed. Then, the controller communicates the assignment and resource allocation decisions to the BSs. To summarize, the controller needs to know user SINR values to the neighboring BSs and user required rates to make a decision at every time slot. We assume that the controller "remembers" the previous user-to-BS assignment as well as the total available resources at every BS. If a handover decision is made, the BSs send a handover reconfiguration message to the user. The serving BS does not need to send an admission request to the target BS in case of a handover because the controller has global knowledge about network resources and informs the target BS if it should accept a new user. We focus on the scenario with a single SD-RAN controller, but it can be extended to a distributed control plane scenario. Note that our approach is agnostic to mobility patterns, as the user trajectory is transformed into SINR.

V. Baseline Models

In this section, we describe the two baseline models against which we are going to compare the performance of our approaches later in Section VI.

A. SINR-based (Baseline) Handover

The user periodically measures the channel and sends the measurement report to its serving BS, which contains the signal strength of the serving and neighboring BSs. 3GPP allows for periodic and event-triggered MRs [14]. We assume that users signal their measured channels to their serving BSs periodically. In the baseline handover algorithm, the network makes handover decisions and signals them to the users [6]. Before reporting the measurements, the user applies Layer-3 filtering and averages Reference Signal Received Power (RSRP) or SINR values over 200 ms [6]. Based on these measurements, the serving BS selects the target BS that should be prepared for handover. The network initiates a handover when a neighboring BS becomes better than the serving BS by a certain margin (e.g., 3 dB) during a certain period of time (e.g., during 320 ms), similarly to [9], [15]. The handover and Radio Link Failure (RLF) rates greatly depend on these handover parameters, and to achieve optimal performance, they can be adjusted per user and per cell, considering the user speed and other system- and user-related parameters. We use SINR measurements to make handover decisions, hence, we refer to this baseline algorithm as the SINR-based handover.

B. Adaptive handover parameter baseline

The authors in [8] adjust handover parameters, like the handover margin (in dB) and Time-to-Trigger (TTT) (in ms), based on user velocity. They propose to use two thresholds to split users in different groups based on their speed. When the calculated user velocity is up to 10 km/h, then they set handover margin and TTT to 6 dB and 512 ms, respectively. If the user velocity is in the range between 10 and 45 km/h, then they set the handover parameters to 4 dB and 128 ms. Finally, in case user velocity is above the second threshold of 45 km/h, they select small handover parameters (2 dB and 32 ms) to speed up the handover process and avoid delayed handovers that might result in an RLF.

VI. Performance Evaluation

First, we describe the simulation setup. Then, we present results for a smaller network evaluated with 5G traces, which is followed by results for a network with a larger number of entities with simulated channels.

A. Simulation Setup

We use two types of data for the evaluation. First, we evaluated the algorithms with 5G traces [16] for a *small scenario* with 4 BSs (one macro and three micro) and 15 users over 100 s. To the best of our knowledge, there are no traces available for multiple BSs in the area for a *large network*,

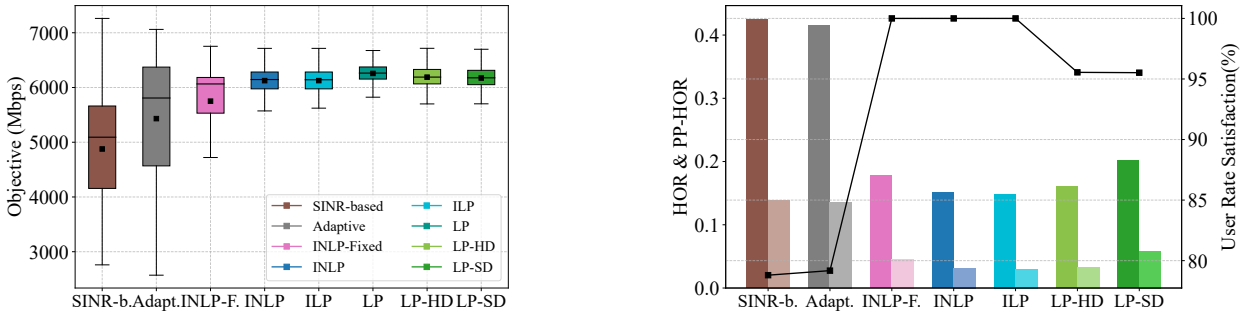


Fig. 3: The performance with different algorithms for the small scenario. In subfigure (b), the bars show the Handover and Ping-Pong Handover Rates denoted as HOR and PP-HOR, while the black line presents the user rate satisfaction.

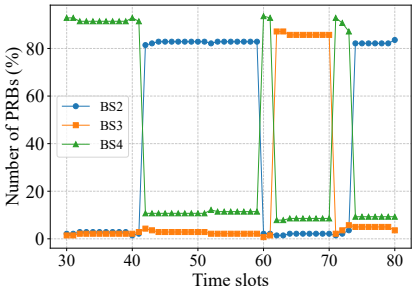


Fig. 4: PRB allocation to micro BSs over time with ILP.

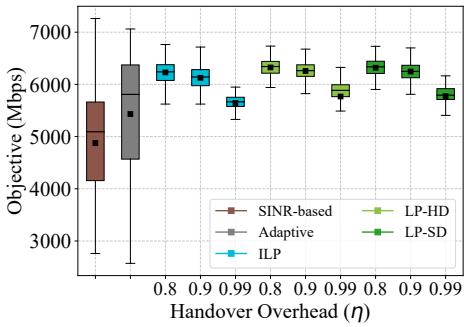


Fig. 5: Impact of η on the objective function.

which are suited for mobility-related evaluations. Thus, in the second case (for a larger network scenario) we consider a two-tier network with urban channel models for macro and micro cells, respectively, from 3GPP 5G Release 14 [17] and simulate the channel for a higher number of BSs. In the latter, we model the path loss and shadowing for LoS and no LoS as in [17]. The large network consists of 10 BSs (3 macro and 7 micro BS) and 50 users. We run the simulation over 1000 s. We refer to the first case as the small network scenario, whereas to the second as the large[1] network scenario. Random Waypoint mobility model is used to generate the mobility-related data [18] for pedestrian users, bikes, and cars for the large scenario. The other simulation parameters are provided in Table I. An RLF occurs when the user's SINR falls below an RLF T_{out} threshold during *T310* timer. For the large scenario, the frequency reuse factor is 3 for the baseline algorithms, while in our algorithms (INLP, ILP, LP-HD, LP-SD) PRBs are distributed among BSs within a common pool, meaning that both macro and micro BSs draw upon resources from their respective shared pool.

To illustrate the influence of the SD-RAN controller on performance, we contrast the optimal approach with fixed and dynamic allocation of PRBs to BSs. The method, employing a fixed number of PRBs per BS, is referred to as INLP-fixed, while the dynamic resource allocation strategy is denoted as INLP. The proposed transformed problem is denoted as ILP. The overall count of PRBs remains constant across

[1]We refer to this scenario as *large* to distinguish it from the small scenario with no intention of implying how many users comprise a large network.

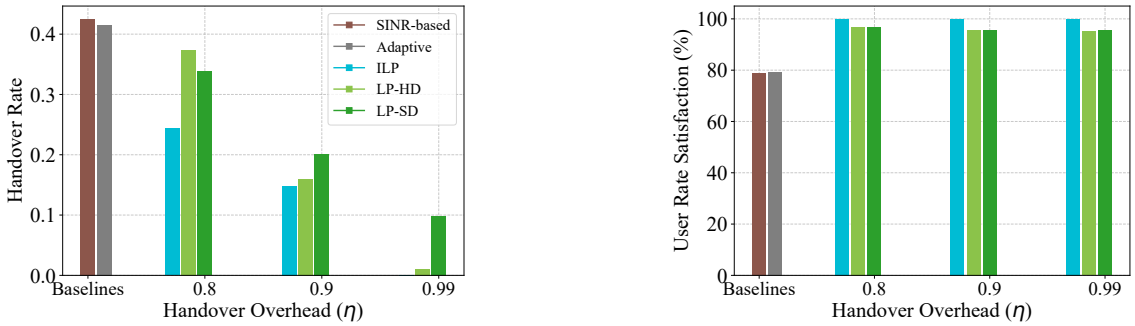
TABLE I: Simulation Parameters [6], [17]

Parameter	Value
Carrier frequency (macro)	2.5 GHz
Carrier frequency (micro)	28 GHz
Channel measurement periodicity	10 ms
L3 filtering time constant	200 ms
Ping-pong window	3000 ms
HIT	80 ms
Handover preparation time	28.5 ms
Handover execution margin and TTT	3 dB, 320 ms

all algorithms. However, in INLP-fixed, PRBs are set at a fixed value for each BS. In the dynamic approach, PRBs are allocated to each BS every 100 ms, with both macro and micro BSs having their respective pools of available PRBs. Then, we compare our approaches (LP-hard and LP-soft), proposed in Section IV, against INLP-Fixed and two baselines that were explained in Section V. Finally, we provide a relaxed solution (denoted LP), which is an upper bound to INLP.

B. Small network scenario with 5G traces

Fig. 3a shows the computed objective function from Eq. (1) for the evaluated handover algorithms. These are box plots, where the median is denoted as a horizontal line, the mean as a black square, error bars show the minimum and maximum values and empty circles show the outliers. INLP and ILP achieve exactly the same objective value, which demonstrates that we achieve the exact transformation from a non-linear into a linear optimization problem as described in Section III-A. One can notice a slight difference in some metrics between



(a) (b)

Fig. 6: Impact of the handover overhead η (the legend bar is provided in subfigure (a)).

INLP and ILP, e.g., handover rate in Fig. 3b is slightly different, namely, 0.151 vs. 0.149. This happens because there can be multiple optimal user assignments and resource allocation decisions with INLP that lead to the same objective value, but other KPIs such as handover rate can be different. Moreover, the decision at the current slot impacts the decisions in the next time slots because the previous assignment is considered in the handover overhead $\eta_{u,b',b}$.

The proposed algorithms LP-HD and LP-SD are within 1% of the ILP obtained with Gurobi (see Fig. 3a); they even achieve a larger objective because they trade off user rate satisfaction for the objective. They achieve a user rate satisfaction of 96% (while ILP reaches 100%), as shown in Fig. 3b. LP provides an upper bound to ILP because the integrality constraints are relaxed. The relaxed and transformed solution, LP, proposed in Section III-B, obtains the optimal solution with an optimality gap of 2%. So, LP provides a good insight on what the optimal value can be in polynomial time. The adaptive handover baseline outperforms the SINR-based baseline by 11%, while the proposed LP-HD and LP-SD by 27%. Baseline approaches achieve a low user rate satisfaction (SINR-based 78% and adaptive handover baseline 79%) because they do not consider available resources at the target BS before a handover and do not consider the handover overhead to compensate for HIT by allocating more resources before/after the handover. LP-HD and LP-SD enhance the user rate satisfaction by $\approx 18\%$, surpassing the baselines and achieving a user rate satisfaction of $\approx 96\%$.

Finally, ILP with dynamic resource allocation to BSs using SD-RAN elevates the objective function by 7% compared to INLP-fixed, as depicted in Fig. 3a. While this percentage might seem relatively small, it becomes crucial in resource-constrained scenarios. As demonstrated later in this section, it is challenging for INLP-fixed to find a feasible solution, whereas ILP, with dynamic resource allocation, successfully identifies a viable solution.

Dynamic resource allocation brings many other advantages such as significantly lower handover and ping-pong handover rates as shown in Figs. 3b (handover rates are shown per user per second). The baseline algorithms have a very high handover rate of ≈ 0.43, which implies that the user would experience on average an outage of more than 2 seconds every minute, while the proposed LP-HD algorithm achieves a total HIT of 0.75 seconds. ILP reduces the handover rate by $\approx 65\%$ compared to the baselines, while LP-HD by 62% and LP-SD by 53%. Even though randomized rounding usually performs well [19], due to the probabilistic nature, LP-SD has a higher handover rate than LP-HD. Furthermore, the ping-pong handover rate reduces by 75%, from 0.13 to 0.03 with LP-HD compared to the baselines, so ping-pong handovers are almost completely avoided with our approach. Reducing handover and ping-pong handover rates reduces the signaling in the network and increases user satisfaction because HIT is reduced.

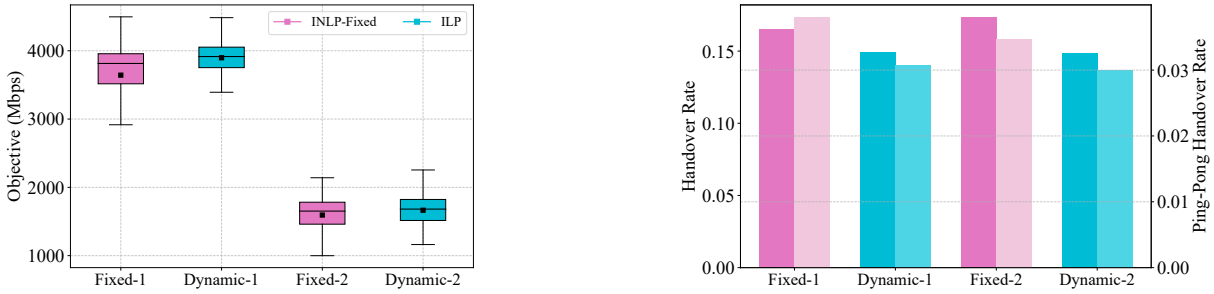
The optimal distribution of resources among micro BSs varies significantly from one time slot to another, as illustrated in Fig. 4 (for 80 time slots or 8 seconds). One reason for this is that instead of performing a handover, the resources might be allocated to the currently serving BS to satisfy the user's required rate as the handover and ping-pong handover rates reduce with ILP by 17% and 32%, accordingly, compared to INLP-fixed. Moreover, due to a poor channel or an unavoidable handover the user might require more resources to satisfy its user rate, so more PRBs are allocated to the serving BS.

C. Impact of the handover overhead

Figs. 5-6 show the impact of the handover overhead on the objective function, handover rate, and user rate satisfaction. As expected, the objective value reduces with a larger overhead, while the handover rate decreases, especially with the very large handover overhead of 0.99. User rate satisfaction is not impacted by the handover overhead. A larger overhead implies a larger penalty in constraint (13), which decreases the objective and reduces the handover rate. The reason is that more resources are allocated to the user that is performing a handover to satisfy its required rate instead of allocating them to the BS with the best channel, or a handover is avoided at the cost of less efficient resource utilization.

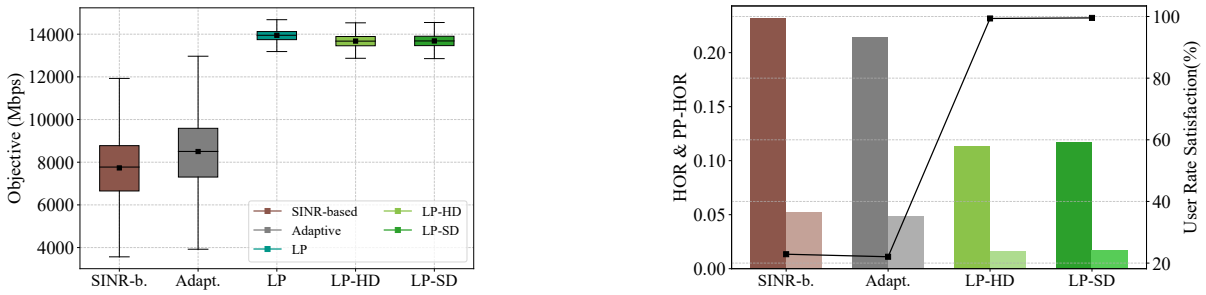
D. Resource-constrained scenarios

Fig. 7 displays the results for two resource-constrained scenarios. In the first, (denoted as Fixed-1 and Dynamic-1,



(a) (b)

Fig. 7: Comparison between the Fixed and the Dynamic resource allocation with SD-RAN among BSs is conducted in two scenarios, with the second scenario having only half of the resources of the first one. Note that with Fixed-2, the problem is infeasible in 15% of the time slots.



(a) (b)

Fig. 8: The performance results for the large scenario.

which correspond to INLP-Fixed and ILP), it is challenging for INLP-Fixed to find a feasible solution in all time slots. In the second scenario, the number of resources in the micro-pool is halved, and Fixed-2 cannot find a feasible solution in 15% of the time slots. For Fixed-2, we exclude those time slots and depict the results only for the slots with feasible solutions. In contrast, the Dynamic approach utilizes all traces. The proposed dynamic approach enhances the objective by 7%, reduces the handover rate by $\approx$ 10%, and diminishes the ping-pong handover rate by $\approx$ 20%. Crucially, SD-RAN-enabled ILP identifies a feasible solution, meeting all user requirements, a task in which INLP-Fixed falls short. This underscores the rationale for deploying SD-RAN to dynamically allocate resources to BSs.

E. Larger network scenario

Finally, we evaluate the proposed LP-HD and LP-SD algorithms along with LP and the baseline approaches for a large scenario. The observed trend aligns with the small scenario; notably, the objective function sees an increase of 77% with LP-HD and LP-SD compared to SINR-based baseline, while the handover rate experiences a decrease of 52% with LP-HD and LP-SD compared to the baselines. The current ping-pong handover rate with SINR-based baseline is notably low, however, LP-HD and LP-SD further diminish it by 70%. LP-HD and LP-SD closely approach the upper bound LP, with a marginal difference of less than 2%. The proposed approaches enhance the user satisfaction, progressing from 23% with the SINR-based baseline to a remarkable 99% with LP-HD and LP-SD approaches. These substantial improvements underscore the significant advantage our approach offers for effective mobility management and resource allocation in wireless networks.

VII. Related Work

Numerous works on mobility management have focused on adapting handover parameters [8], performing joint resource allocation and handover management [4], [9], reducing mobility-related signaling [20], [21] and using lower level signaling for mobility [15]. In [21], the authors analyze in depth mobility-related issues and propose a handover prediction system to improve the quality of experience of mobile users. Furthermore, their measurements show that frequent handovers in 5G diminish throughput and deplete user batteries, causing in the worst case a complete service outage.

Several approaches aimed at reducing handover rate by setting the handover parameters based on some criteria [7], [8]. In [7], the suggestion is to adjust the handover margin and TTT based on user speed and measured channel conditions. Similarly, the authors in [8] propose dynamic adjustments to handover parameters based on user speed, cell load, and load balancing between neighboring cells. The authors in [22] also propose a dynamic adjustment of TTT and handover margin aimed at minimizing latency and enhancing overall throughput. An intelligent approach using Deep RL (DRL) to

provide proportional fairness among users and perform user-to-BS assignment with equal resource split among connected users was proposed in [4]. The authors in [3] employ SDN for the management of handovers, and their approach leads to reductions in both delay and handover failures. However, to our best knowledge, the gains of utilizing SD-RAN for mobility management in cellular networks have not been studied so far.

The concept of SD-RAN has gathered significant attention in recent years. Early studies, such as [2] and [23], advocate for transferring control decisions from the BS to a centralized controller. These works highlight the increased flexibility achieved through the adoption of SD-RAN.

The gains of enhanced overall throughput through dynamic resource allocation to BSs under resource constraints using SD-RAN were explored in [24]. Additionally, the investigation into the advantages of SD-RAN flexibility in resource allocation, aiming to achieve proportionally fair resource distribution among users, is presented in [25].

The works most closely related to ours in the realms of mobility management and the utilization of SD-RAN in the context of 5G are [4] and [24]. However, the authors in [4] focus on a one-level resource allocation scenario, specifically without SD-RAN, where the resources at every BS remain fixed. On the other hand, we optimize the resource proportion of every user and ensure a minimum rate guarantee for every user to satisfy their Quality of Service (QoS) requirement. In [24], the authors propose a policy for achieving proportionally-fair resource allocation among the BSs and users, employing a two-level resource allocation - from the controller to the BSs and then to the users. However, their study does not consider mobility or handover management, crucial components that are inherently connected with the process of resource allocation.

VIII. Conclusion

In this paper, we considered the problem of joint mobility management and two-level resource allocation using SD-RAN. To that end, we formulated an optimization problem and transformed it into a linear one. Because of the complexity of the optimization problem, we proposed two near-optimal solutions (hard and soft policies) and showed that our approaches significantly outperform other state-of-the-art baselines while being very close to the optimal solution. Moreover, we provided an upper bound that the objective function of the optimization problem can achieve, which can be found in polynomial time, and showed that our approaches are not far from it. In the future, we plan to consider the allocation of computational resources on top of the RAN PRBs and distributed control plane.

References

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

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

[3] T. Bilen, B. Canberk, and K. R. Chowdhury, "Handover management in software-defined ultra-dense 5G networks," *IEEE Network*, vol. 31, no. 4, 2017.

[4] A. Prado, F. Stoeckeler, F. Mehmeti, K. Patrick, and W. Kellerer, "Enabling proportionally-fair mobility management with reinforcement learning in 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, 2023.

[5] Nokia, "Rock solid mobility innovations from 5G to 5G-advanced," June 2022, accessed on Jan, 2024. [Online]. Available: https://onestore.nokia.com/asset/212564

[6] 3GPP, "NR; NR and NG-RAN Overall description; Stage-2," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.300, 3 2021, version 16.5.0. [Online]. Available: http://www.3gpp.org/DynaReport/38300.htm

[7] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alraih, and K. S. Mohamed, "Auto tuning self-optimization algorithm for mobility management in LTE-A and 5G HetNets," *IEEE Access*, vol. 8, 2019.

[8] A. Hatipoğlu, M. Başaran, M. A. Yazici, and L. Durak-Ata, "Handover-based load balancing algorithm for 5G and beyond heterogeneous networks," in *Proc. of ICUMT*, 2020.

[9] A. Prado, D. Gölitz, F. Mehmeti, and W. Kellerer, "Proportionally Fair Resource Allocation Considering Geometric Blockage Modeling for Improved Mobility Management in 5G," in *Proc. of ACM Q2SWinet*, 2022.

[10] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[11] 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 Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[12] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2013.

[13] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, 2020.

[14] 3GPP, "NR; User Equipment (UE) conformance specification; Radio Resource Management (RRM)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.533. [Online]. Available: http://www.3gpp.org/DynaReport/38533.htm

[15] A. Gündogan, A. Badalıoğlu, P. Spapis, and A. Awada, "On the modelling and performance analysis of lower layer mobility in 5G-advanced," in *Proc. of IEEE WCNC*, 2023.

[16] 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.

[17] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.901, 1 2020, version 16.1.0. [Online]. Available: http://www.3gpp.org/DynaReport/38901.htm

[18] X. Lin, R. K. Ganti, P. J. Fleming, and J. G. Andrews, "Towards understanding the fundamentals of mobility in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, 2013.

[19] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

[20] A. Prado, F. Mehmeti, and W. Kellerer, "Cost-efficient mobility management in 5G," in *Proc. of IEEE WoWMoM*, 2023.

[21] A. Hassan, A. Narayanan, A. Zhang, W. Ye, R. Zhu, S. Jin, J. Carpenter, Z. M. Mao, F. Qian, and Z.-L. Zhang, "Vivisecting mobility management in 5G cellular networks," in *Proc. of ACM SIGCOMM*, 2022.

[22] R. Karmakar, G. Kaddoum, and S. Chattopadhyay, "Mobility management in 5G and beyond: a novel smart handover with adaptive time-to-trigger and hysteresis margin," *IEEE Transactions on Mobile Computing*, vol. 22, no. 10, 2023.

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

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

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