Minimizing energy consumption by joint radio and computing resource allocation in Cloud-RAN

Mahdi Sharara, Francesca Fossati, Sahar Hoteit, Véronique Vèque, Francesca Bassi

Minimizing energy consumption by joint radio and computing resource allocation in Cloud-RAN

Mahdi Sharara, Francesca Fossati, Sahar Hoteit, Véronique Vèque, Francesca Bassi

Minimizing Energy Consumption by Joint Radio and Computing Resource Allocation in Cloud-RAN

Mahdi Sharara[a], Francesca Fossati[b], Sahar Hoteit[a], Véronique Vèque[a], Francesca Bassi[c]

[a] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, , Gif-sur-Yvette, 91190, , France, (emails: {mahdi.sharara,sahar.hoteit,veronique.veque}@university-paris-saclay.fr)*
[b] *Sorbonne Université - LIP6, , Paris, 75252, , France (email: Francesca.Fossati@lip6.fr)*
[c] *IRT SystemX, , Palaiseau, 91120, , France (email: francesca.bassi@irt-systemx.fr)*

Abstract

Cloud-RAN is a key 5G enabler; it centralizes the baseband processing of several base stations by executing the baseband functions in a centralized, virtualized, and shared entity known as the Base Band Unit (BBU)-Pool. Cloud-RAN paves the way for joint management of the radio and computing resources of multiple base stations. In fact, centralization and virtualization allow for decreasing energy consumption which decreases Capital Expenditure (CAPEX) and Operational Expenditure (OPEX). Cloud-RAN architecture permits jointly allocating the radio and computing resources of multiple base stations. The radio resources include the Resource Blocks (RBs), the transmission power, and the Modulation Coding Scheme (MCS), whereas the computing resources include the CPUs resources. This paper investigates the potential benefits that could be scored thanks to the joint allocation of these two types of resources, with respect to energy consumption and overall throughput, when radio resources are finite and computing resources are not. The latter is an effect of the C-RAN architecture, which allows scalability and fast computing resource provisioning. Due to the unconstrained availability of computing resources, the joint allocation of radio and computing resources has a negligible impact when the objective is throughput maximization. However, it is highly beneficial when the target is energy consumption minimization in comparison to the sequential allocation that consists of allocating radio resources first, and then computing resources are allocated. For that, we formulate a Mixed Integer Linear Programming (MILP) problem

having the objective of minimizing energy consumption. When the goal is to minimize energy consumption, the joint allocation of radio and computing resources reduces the total energy consumption by up to 21.3% when compared to the case where radio and computing resources in the BBU pool are allocated sequentially. Furthermore, given the NP-hardness of solving a MILP problem, we propose a two-step low-complexity matching game-based algorithm with a transmission power adjustment mechanism that aims at performing close to the MILP solver. The results show that our proposed matching game algorithm is a good alternative for solving the joint-allocation MILP problem, producing results that are very close to the MILP optimal solutions.

Keywords: Cloud-RAN, Joint Radio and Computing Resource Allocation, Energy consumption minimization, Mixed Integer Linear Programming (MILP), Matching Game.

---

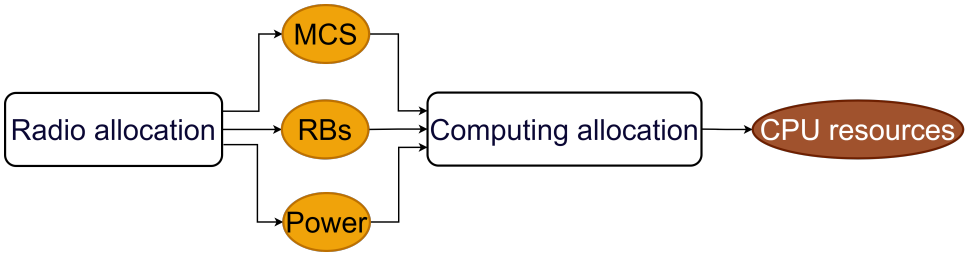
1. Introduction

The demand for mobile data is increasing at an extraordinary pace. The fifth generation of mobile networks (5G) attempts to address this issue through various technologies such as Cloud-Radio Access Network (Cloud-RAN), among others [1, 2]. A base station traditionally comprises a Remote Radio Head (RRH) and a Base Band Unit (BBU). The RRH is in charge of executing radio frequency functions, while the BBU is in charge of executing baseband functions. In Cloud-RAN, the BBU is decoupled from the RRH such that the BBUs are hosted virtually in a cloud known as the BBU Pool [3]. Cloud-RAN carries many advantages, including the ability to create a scalable and flexible network. Furthermore, centralization and cloudification enable more efficient resource utilization while lowering CAPEX and OPEX [1].

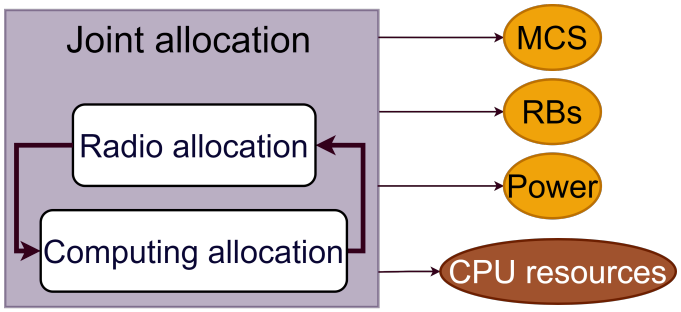
Strategies to reduce energy consumption are critical in 5G [1, 2] since a massive number of base stations must be deployed to satisfy users' demands. In this paper, we consider the joint allocation of radio and computing resources in Cloud-RAN. Precisely, we are interested in understanding how the joint allocation of radio and computing resources, compared to a sequential allocation of these resources, may help minimize the energy consumed for transmission and baseband processing while still meeting the users' quality of service requirements. On the one hand, radio resource allocation con-

sists of assigning to each user a number of Resource Blocks (RBs) as well as a Modulation and Coding Scheme (MCS) index and transmission power. These resources are necessary to transmit users' data over the air interface. The RBs provide the carrier frequencies to transmit data symbols, while the MCS index specifies the modulation and code rate. The power has to be allocated such that it achieves the minimum Signal to Interference plus Noise Ratio (SINR) required to use the selected MCS. On the other hand, computing resource allocation consists of assigning users' radio frames to CPUs in the BBU pool and ensuring that their processing deadlines are met. It is worth mentioning that the amount of time required to process users' data by the CPUs in the BBU pool is heavily dependent on radio parameters (i.e., MCS, number of RBs, transmission power): the transmission power allocation affects the user's signal to interference plus noise ratio (SINR), hence limiting the maximum MCS index that can be adopted. The MCS index, along with the number of allocated RBs for each user, affects its processing demand [4] and hence the amount of required processing resources in the BBU pool [5, 6]. Besides, if users' quality of service requirements have to be satisfied, a minimum throughput target per user has to be guaranteed. This throughput increases with the MCS and the number of allocated RBs [7, 5].

We formulate the joint allocation of radio and computing resources in Cloud-RAN as a Mixed Integer Linear Programming (MILP) problem that jointly allocates the transmission power, the resource blocks, the MCS indexes, and the CPU time to process the data of each user. We consider the goal of minimizing energy consumption and compare the results to those obtained when the goal is to maximize the system's throughput. To quantify the impact of *joint* radio and computing resources allocation, we compare it to a sequential scheme that performs radio resources allocation followed by allocating computing resources. Figure 1 shows the difference between sequential and joint allocation of radio and computing resources. In a setting where radio resources (transmission power, RB number, MCS) and computing resources (CPU time) are allocated sequentially, as in Figure 1a, minimizing a user's transmission power while still meeting its throughput target may result in MCS and RBs assignments that require more computing resources, consuming more energy. A joint radio and computing resource allocation scheme that controls the radio and computing resource parameters at the same time would be more capable of minimizing total energy consumption (i.e., the sum of transmission *and* processing energy). Even though joint allocation may be more computationally complex than sequential allocation,



(a) Sequential allocation



(b) Joint allocation

Figure 1: Sequential allocation vs. Joint allocation of radio and computing resources

it would be advantageous if it exhibits significant energy consumption reductions.

Recently, the matching theory has gained a lot of attention because of the possibility of modeling many problems in wireless networks as matching games and solving such problems using various matching algorithms [8]. Players should try to match with other players in such a way that their utility and benefits are maximized. Given that solving an MILP problem is highly complex and is NP-hard [9], we design a low-complexity two-step algorithm that allocates RBs, transmission power, and MCS indexes to users using a matching game and a transmission power adjustment mechanism.

We analyze the convergence and the complexity of the matching-based algorithm and compare its results to the optimal performance of the MILP problem. The main contributions of this paper are summarized as follows:

- We model the joint radio and computing resource allocation problem as an MILP problem that allocates RBs, power, MCS indexes, and CPU resources to users.
- We consider two objectives; energy consumption minimization and

throughput maximization. From the operator's perspective, these two objectives are critical in Cloud-RAN and 5G. We show that joint radio and computing resource allocation may greatly benefit energy consumption minimization.

- We compare the performance of the joint scheme of radio and computing resource allocation to that of a sequential method in which radio and computing resource allocation is done sequentially.
- We propose a two-step algorithm based on a matching game and a transmission power adjustment mechanism to achieve solutions close to the optimal solution of the joint-allocation MILP problem.
- We provide the convergence and complexity analysis of the matching-based algorithm. We also compare the run-time of the MILP solver with the matching-based algorithms.
- We compare the performance of the proposed matching-based algorithm to the optimal MILP problem considering the metrics related to energy consumption, throughput, and fairness.

The rest of the paper is organized as follows: Section 2 surveys the related work. The MILP problem is formulated in section 3. In section 4, the matching-based algorithm with transmission power adjustment is presented. The simulation settings are described in section 5, and the results of the simulations are discussed in section 6. Finally, our work is concluded in section 7.

2. Related Work

2.1. Radio and Computing Resource Allocation

Works in the literature have considered radio or computing resource allocation independently [10, 11, 12, 13]. Authors in [10] formulate a radio allocation MILP problem. It considers RBs and MCS assignments in addition to power allocation. Their model is quite limited as it considers only one base station but none of the interference caused by other base stations. Additionally, the problem only considers radio allocation without considering computing resource allocation. Moreover, it does not aim at minimizing the total energy consumption. To optimize system throughput and energy

efficiency, the authors of [11] formulate a Mixed Integer Non-Linear Programming (MINLP) problem, then relax it into a lower-complexity two-step approach for both radio and compute resources. The computing resource allocation occurs first by mapping users to Virtual Machines (VM). Secondly, the radio resource allocation is done by controlling the beam-forming vectors. Nevertheless, the scope of this allocation is limited as the algorithm does not consider the existence of multiple RBs or sub-carriers nor the selection of MCS indexes based on the SINR. The authors in [12] consider joint beamforming vector design and BBU computational resources allocation. They aim to minimize the total system power consumption while considering the constraints of users' Quality of Service (QoS), fronthaul capacity, transmit power per RRH, and per Antenna. Compared to our work, the paper in [12] does not consider the RBs or MCS assignments nor the effect on the required processing time. Similarly, [14] considers joint radio and computing resource allocation where they control the beamforming vectors and use a bin-packing algorithm to allocate virtual machines to process users' data, aiming to maximize the weighted sum-rate. In [13], the authors investigate the joint communication (i.e., radio) and computing resource allocation. They consider power allocation and RB assignment in addition to mapping RRH to BBUs running as virtual machines. The problem is formulated using queuing theory to minimize the mean response time. Then, an auction-theory-based algorithm is proposed. Unlike our work, this paper does not tackle the joint allocation problem with the goal of minimizing the total energy consumption.

A coordination scheme between radio and computing resources allocation has recently been considered in [5], where computing resources cannot satisfy all users' demands. The authors propose, in particular, two coordination schemes between radio and computing resources that maximize throughput and users' satisfaction. The proposed schemes permit feedback from the computing scheduler to the radio scheduler to update radio resource parameters based on computing resource availability. Authors demonstrate a significant ability to decrease the amount of wasted transmission power. In the same context, and to reduce the complexity of the Integer Linear Programming (ILP)-based coordination algorithms, lower complexity Recurrent Neural Network (RNN)-based algorithms are developed in [15]. They are trained to perform close to the ILP solver. Results show a significant reduction in the execution time compared to the ILP problems. Unlike our work, these coordination schemes are not based on the joint allocation between radio and computing resources and do not seek to optimize overall energy

consumption.

To jointly optimize the throughput and the functional split, the authors of [16] formulate the problem as an ILP problem that allocates resource blocks, fronthaul bandwidth, and computing resources. Considering service-aware resource allocation in an Open-RAN context and aiming to maximize the sum-rate, [17] formulates a Mixed Integer Non-Linear Programming (MINLP), and allocate RBs, power, and VNF processing resources subject to limited fronthaul capacity and end-to-end delay constraints. Due to the NP-hardness of the problem, the authors devise a low-complexity sub-optimal algorithm to solve the problem.

Considering latency and limited resource constraints, [18] tackles the problem of radio and computing resource allocation for edge computing. The authors allocate uplink, downlink, and computing resources. To solve the highly complex problem, they formulate the problem as a generalized Nash equilibrium problem and devise an algorithm to find this equilibrium. Considering edge computing and aiming to minimize the delay, [19] consider joint time allocation and offloading using Non-Orthogonal Multiple Access (NOMA) and demonstrate its improvement in comparison with Orthogonal Multiple Access schemes.

2.2. Matching Games

Due to the NP-Hardness of the ILP and the MILP problems, many papers in the literature propose low-complexity tools as an alternative. In this context, the matching-game theory appears at the forefront as one of the promising tools. The fundamental properties of matching games in wireless networks, along with several applications, have been presented in [8]. Different matching models have been used, including one-to-one, one-to-many, and many-to-many matching models, and algorithms have been proposed to solve each of these problems. The stable marriage problem is an example that can help us understand the basics of matching theory [20]. Consider a set of men $\mathcal{M} = A, B, C$ and a set of women $\mathcal{W} = a, b, c$. Each man and each woman has a preference list for the opposite gender; the ones they mostly prefer to be married to. The fundamental solution concept for a matching game is the notion of stable matching. A stable matching is a matching where no blocking pair exists. Suppose that A is matched with a, B is matched with b, and C is matched with c. However, A prefers b over its current match a, and b prefers A over its current match B. Hence the pair (A,b) is a blocking pair as each has the incentive to leave its current partner and switch to another.

Hence, a matching algorithm should converge to a stable matching with no blocking pair.

To solve the Stable-Marriage problem, Gale-Shapely proposed the Deferred Acceptance algorithm. In this algorithm, each player from each set creates its preference list. Players from one set make proposals, and the players from the other set accept or reject the proposals. For example, the men can be the ones who propose, and the women accept or reject. The other option is to reverse the roles. Supposing that men are the ones who propose, each man proposed to his most preferred woman. Then, each woman examines the proposals she received, accepts the most preferred man among those who proposed and rejects the other proposals. In the next iteration, men who are already matched will not propose, but those who are unmatched will propose to their second preferred woman. The women who receive the proposals accept the proposals if they are better than their current matches and reject them if they are not. Men who become unmatched will join the next round and propose to the next preferred women on their list. As this algorithm is guaranteed to converge to a stable matching, this iterative process will be repeated until a stable matching is attained.

Similarly, matching theory has applications in mobile communication. More precisely, we have users, base stations, operators, etc., competing with each other and forming preferences over other sets (i.e., a user prefers to match with RBs that help satisfy his requirements, and an RB, on behalf of the operator, prefers to match with a user who results in the least interference). A bipartite one-to-one matching game is used to model the assignment of LTE-U users to Unlicensed bands in [21]. The famous Gale-Shapley algorithm is used to realize a stable matching; then, an algorithm is proposed to improve the total throughput by allowing users to switch their assignments if this increases their utilities. In [22], the issue of caching in small base stations is modeled as a many-to-many matching game between small base stations and video content from service providers. In [23], a many-to-many matching game is formulated to model the assignment in relay networks between source nodes and sub-channels where Non-Orthogonal Multiple Access (NOMA) is used for channel access.

In [24], the authors consider the issue of resource allocation for a full-duplex OFDMA Network. A full duplex base station communicates with half duplex uplink and downlink users, and the objective is to maximize the network sum-rate by joint user pairing, sub-channel assignment, and power allocation. Because of the non-convexity of the optimization problem, it is

solved using a cyclic three-sided matching between the sets of transmitting users (uplink), sub-channels, and receiving users (downlink). In contrast to this paper, our work aims to minimize the total energy consumption, and we use our proposed matching game with a transmission power adjustment mechanism to achieve this objective.

The matching-coalition approach is used also in [25], where a NOMA-based MEC model that aims at improving energy efficiency is considered. A joint optimization problem is formulated where the problems of user association, power control, and computational resource allocation are combined to optimize energy consumption. Due to the NP-hardness of the model, the authors propose to use a matching and coalition framework. First, the generalized (resident-oriented) Gale-Shapely algorithm is used to formulate a matching between users and access points while neglecting co-channel interference. Then, the coalition approach is used to allow users to enhance their results by considering externalities. Externalities mean that assignments/matchings lead to changes in players' preferences. The preference lists are assumed to be fixed to assure the problem theoretically converges to a stable matching. However, the obtained solution might not be optimal because considering fixed preference lists ignores the externalities. Accordingly, a coalitional game is proposed allowing the modification of choices, but this time accounting for co-channel interference. However, unlike our work, the authors do not consider OFDM or assign MCS indexes to users. In [26], the problem of resource allocation of sub-channels in heterogeneous Cloud-RAN is addressed. The authors consider a downlink model that consists of one eNB transmitting to mobile users, RRHs transmitting to RRH users, and D2D transmitters transmitting to D2D receivers. However, not more than one radio sub-channel is used by an RRH or D2D user. An optimization problem aiming to maximize the system throughput is modeled as a mixed integer non-linear programming. Then a two-step algorithm based on matching theory and coalition games is used to find sub-optimal solutions. In contrast, our paper does not restrict RB allocation to one per user, and it also allocates radio power, MCS, and CPU resources.

In contrast to these research papers, we consider in this paper the joint radio and computing resource allocation. Our model considers transmission power allocation, MCS assignment, RBs allocation, CPU assignment, and computing resource allocation. To the best of our knowledge, In the current literature, there is no explicit comparison of joint vs. sequential resource allocation. Furthermore, the advantages, limitations, and influence of the

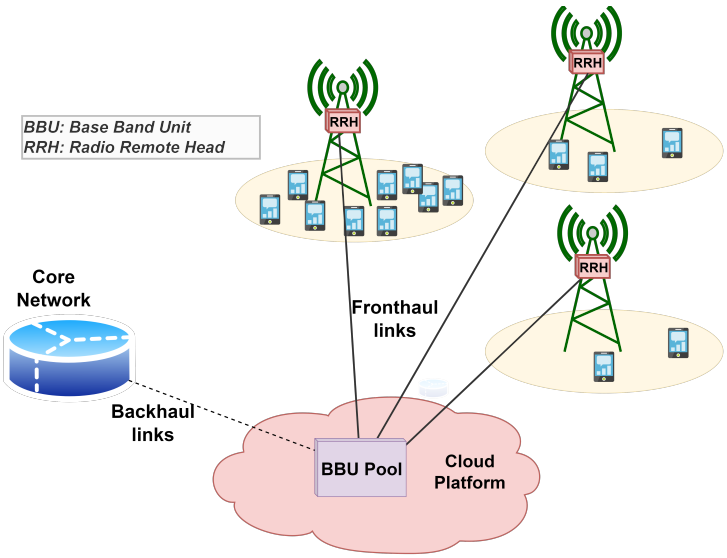


Figure 2: Cloud-RAN basic architecture

chosen objective function have not yet been investigated. Moreover, our paper proposes a matching-based algorithm that aims at providing solutions close to those of the optimal MILP problem.

This paper is an extension of our work in [27]. While [27] only models the joint MILP problem, which is NP-hard, and shows the benefits of having a joint radio and computing resources allocation, this extension provides a low-complexity alternative to the MILP problem. It is based on matching theory, and it yields solutions that are close to the optimal MILP problem solutions. Also, this extended version also includes convergence and complexity analyses of the proposed matching-theory-based algorithm.

3. Problem Formulation

In this section, we first present the problem of joint allocation of radio and computing resources allocation in Cloud-RAN, and we model it as a Mixed Integer Linear Programming (MILP) problem.

3.1. Cloud-RAN

In traditional RAN, the base station consists of a Radio Remote Head (RRH) responsible for Radio Frequency functions and a Base Band Unit

Table 1: Summary of the general notations

Parameters	***Definition***
$\mathcal{B}$	Set of base stations
$\mathcal{U}_b$	Set of users for each BS $b \in B$
$\overline{\mathcal{U}}_b$	Set of users in BS $b \in B$ who have achieved the demanded throughput
N	The total number of users in the system
$\mathcal{R}$	Set of RBs that can be used in the system
$\mathcal{R}^{b,u}$	Set of RBs assigned to user $u \in \mathcal{U}_b$
$\mathcal{I}$	Set of MCS indexes that can be used in the system
$\mathcal{C}$	Set of CPU cores in the shared BBU pool (multi-core data center).
$MCS_{b,u}$	the assigned MCS index to user $u \in \mathcal{U}_b$.
$th_{r,u}$	The achieved throughput by user $u \in \mathcal{U}_r$
$th_{r,u}(b)$	The possible achieved throughput by user $u \in \mathcal{U}_r$ if it uses RB b
$t_{s,i,c}$	Data processing time when a frame is transmitted over s number of resource blocks using MCS index i, and processed over CPU c
t_{TTI}	The duration of one Transmission Time Interval (TTI)
$R_{s,i}$	Data length (in bits) when a frame is transmitted over s number of resource blocks using MCS index i
$R_{min}^{b,u}$	minimum required rate for a user $u \in \mathcal{U}_b$
d	Processing time deadline
$x_{r,i}^{b,u}$	Binary variable that assigns the RB r and MCS index i to user $u \in \mathcal{U}_b$
$y_{s,i,c}^{b,u}$	A binary variable that indicates a user $u \in \mathcal{U}_b$ uses s total number of RBs and MCS index i and that its frame is processed on CPU c
$\beta_i^{b,u}$	Binary variable that indicates that user $u \in \mathcal{U}_b$ uses MCS index i
$z_{r,i}^{b,u}$	An auxiliary binary variable used to enforce the condition of not using an MCS index unless the SINR exceeds a specific threshold
$P_r^{b,u}$	A continuous variable that indicates the transmission power of user $u \in \mathcal{U}_b$ over RB r.
$P_{T_x}^{max}$	Maximum total radio transmission power of user $u \in \mathcal{U}_b$.
P_{comp}^c	Computing Power of a CPU $c \in \mathcal{C}$
γ_i^{th}	The minimum SINR threshold to use MCS index i
$g_r^{b,u,b'}$	The channel gain between user $u \in \mathcal{U}_b$ and the base station $b' \in B$ on RB $r \in \mathcal{R}$,
$\sigma^{b,u}$	The channel noise for user $u \in \mathcal{U}_b$
PL_r	The preference list of an RB r
PL_u	The preference list of a user u
PL_u^1	The first element in the preference list of a user u
prop(u)	The RB the that a user u proposes to be matched with
prop(r)	The users that have proposed to the RB r in the current iteration
$ACCEPT^b(u)$	A Boolean to indicate RB b accepts user u
$REJECT^b(u)$	A Boolean to indicate RB b rejects user u
$SINR^{r,u}(b)$	The possibly achieved SINR of user $u \in \mathcal{U}_r$ if it uses RB b in the next iteration

(BBU) responsible for all other functions. To account for peak hours, operators usually allocate excess resources to base stations that are rarely fully used. This increases capital and operational expenditure. To minimize the latter, Cloud-RAN emerges as an alternative based on centralization and virtualization. Cloud-RAN centralizes the baseband processing of multiple

base stations running as virtual machines in a powerful data center called the BBU-pool. This adds scalability and flexibility to the network as resources can be supplied on-demand. Additionally, it allows cutting down expenses as it is possible to release and shut down resources that are not needed. Thanks to virtualization and centralization, reducing the overprovisioning of resources will be possible. Moreover, centralization allows Cloud-RAN to better manage and optimize the network due to the globalized control of resources. The simplified architecture of Cloud-RAN is shown in Figure 2.

3.2. Model input and parameters

To study the performance of joint radio and computing resources allocation, we consider the following scenario: a set of base stations $\mathcal{B}$, a set of users $\mathcal{U}_b$ of each BS b, a set of Resource Blocks $\mathcal{R}$ for each base station, a set of MCS indexes $\mathcal{I}$ that can be used in the system, and a set of CPU cores $\mathcal{C}$ in the shared BBU pool (multi-core data center). As in [5], we focus on the uplink direction in which the BBU pool needs to execute the complex and energy-consuming decoding function. We assume that each user has a maximum transmission power equal to P_{Tx}^{max} and that the CPU power consumption is equal to P_{comp}^{c}. We define the following parameters: $g_r^{b,u,b'}$ is the channel gain between user $u \in \mathcal{U}_b$, who belongs to base station $b \in \mathcal{B}$, and the base station $b' \in \mathcal{B}$ on RB $r \in \mathcal{R}$, γ_i^{th} is the minimal SINR threshold that allows using a given MCS index $i \in \mathcal{I}$. We denote by $\sigma^{b,u}$ the channel noise for user $u \in \mathcal{U}_b$, $R_{s,i}$ the throughput of transmission when the data are transmitted over s number of RBs using an MCS index $i \in \mathcal{I}$, and $t_{s,i,c}$ the required time to process these data on CPU core $c \in \mathcal{C}$. Each CPU should process the assigned data before the deadline d. This deadline is imposed by the Hybrid Automatic Repeat Request (HARQ) mechanism and equals $2ms$ [5]. Not respecting this deadline would lead to retransmitting the radio frame, thus, wasting the initial transmission. We also suppose that users have different QoS requirements; each user requests a minimum throughput $R_{min}^{b,u}$ that must be satisfied. We note that the duration t_{TTI} of the Transmission Time Interval (TTI) over which a user transmits its frame is 1 ms. The system model is shown in Fig. 3.

To determine the time required to process each user's frame, we use the model in [4] built using the Open Air Interface (OAI) RAN simulator. This model provides the required processing time, $t_{s,i,c}$, of a user's frame as a function of the total number of used RBs, the MCS index, and the CPU

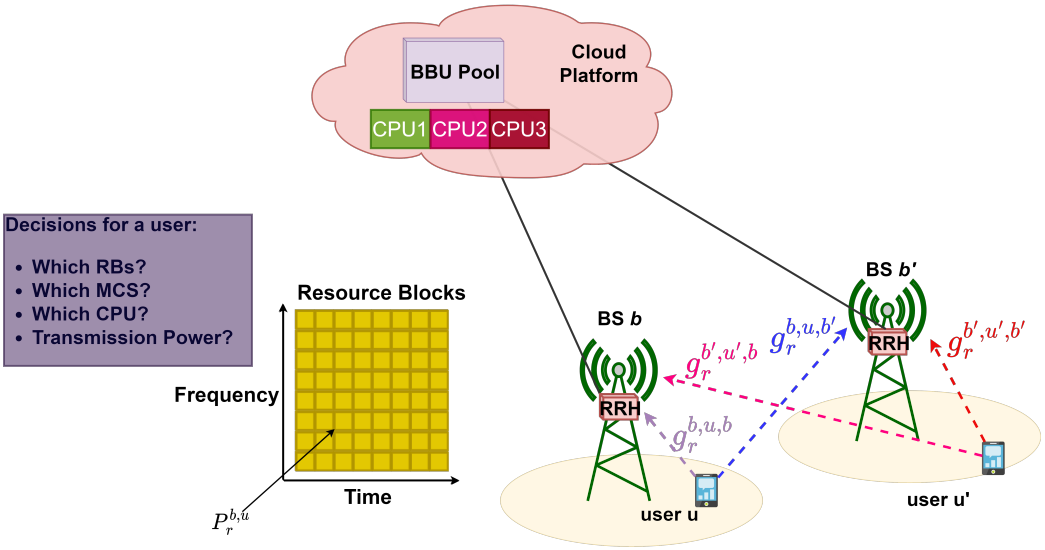


Figure 3: The system model where uplink transmission is considered. Resource Blocks, radio power, MCS indexes, and CPUs are the decision variables.

frequency. The formula is given by:

$$t_{s,i,c}[\mu s] = \frac{s}{f_c^2[\text{GHz}]} \sum_{j=0}^{2} \alpha_j i^j \tag{1}$$

In Eq. (1), s represents the total number of RBs , f_c is the working frequency of the CPU, and i is the used MCS index. Additionally, j defines the exponent, given the model is a second-order polynomial.

Based on experimental studies, [4] provides the values of alpha corresponding to the overall uplink processing: $\alpha_0 = 35.545$, $\alpha_1 = 1.623$, and $\alpha_2 = 0.086$.

On the other hand, we use the 3GPP standard [7] to determine the Transport Block Size (TBS), which is the number of bits transmitted by a transport block in 1 ms, as a function of the number of RBs and the MCS index. Then we get the throughput by dividing the TBS by the transmission duration. We note that using the MCS index to calculate the throughput is more realistic than using Shannon's capacity formula. The latter only gives the upper bound of the channel's throughput and does not distinguish useful bits from redundancy and physical layer overhead bits, as the TBS does. The summary

of general notations used throughout this chapter is shown in Table 1.

3.3. MILP problem

We formulate a Mixed Integer Linear Programming Model (MILP) for joint radio and computing resource allocation, which minimizes the total energy consumption. This MILP problem should be optimized by assigning RBs and MCS indexes to users, the power of their signals, and the CPUs that will process their data. The MILP problem contains the following decision variables:

- $x_{r,i}^{b,u}$ is a binary decision variable equal to 1 if user $u \in \mathcal{U}_b$ uses an MCS index $i \in I$ on RB $r \in \mathcal{R}$; otherwise, it is zero.
- $y_{s,i,c}^{b,u}$ is a binary decision variable that is equal to 1 if and only if a user $u \in \mathcal{U}_b$ transmits data using an MCS index $i \in I$ over a total of s resource blocks, and the user's data are processed on CPU $c \in \mathcal{C}$; and zero otherwise.
- The binary decision variable $\beta_i^{b,u}$ is equal to 1 if and only if a user $u \in \mathcal{U}_b$ uses MCS $i \in I$ on any of its RBs; and zero otherwise.
- Finally, $p_r^{b,u}$ is a continuous variable that indicates the transmission power of user $u \in \mathcal{U}_b$ on RB $r \in \mathcal{R}$.

We note that M is the big-M notation used to enforce the conditions explained below. The formulated MILP optimization problem is defined as follows:

$$\min \quad \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_b}\sum_{r\in\mathcal{R}} p_r^{b,u}\cdot t_{TTI} + \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_b}\sum_{s\in\mathbb{N}\cap[1,|\mathcal{R}|]}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} t_{s,i,c}\cdot y_{s,i,c}^{b,u}\cdot P_{comp}^{c} \tag{2}$$

$$\text{s.t.} \quad x_{r,i}^{b,u} \in \{0,1\},\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{3}$$

$$y_{s,i,c}^{b,u} \in \{0,1\},\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, s\in\mathbb{N}\cap[1,|\mathcal{R}|],; i\in\mathcal{I}, c\in\mathcal{C} \tag{4}$$

$$z_{r,i}^{b,u} \in \{0,1\},\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{5}$$

$$\beta_i^{b,u} \in \{0,1\},\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, i\in\mathcal{I} \tag{6}$$

$$p_r^{b,u} \geq 0,\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R} \tag{7}$$

$$\sum_{u\in\mathcal{U}_b}\sum_{i\in\mathcal{I}} x_{r,i}^{b,u} \leq 1,\ \forall b\in\mathcal{B}, r\in\mathcal{R} \tag{8}$$

$$\sum_{s\in\mathbb{N}\cap[1,|\mathcal{R}|]}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} y^{b,u}_{s,i,c}R_{s,i} \geq R^{b,u}_{min}, \forall b\in\mathcal{B}, u\in\mathcal{U}_b \tag{9}$$

$$\sum_{r\in\mathcal{R}} p^{b,u}_r \leq P^{max}_{Tx}, \forall b\in\mathcal{B}, u\in\mathcal{U}_b \tag{10}$$

$$p^{b,u}_r \leq M\sum_{i\in\mathcal{I}} x^{b,u}_{r,i}, \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R} \tag{11}$$

$$x^{b,u}_{r,i} \leq \beta^{b,u}_i,\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{12}$$

$$\sum_{i\in\mathcal{I}} \beta^{b,u}_i \leq 1, \forall b\in\mathcal{B}, u\in\mathcal{U}_b, \tag{13}$$

$$g^{b,u,b}_r p^{b,u}_r \geq \gamma^{th}_i(\sigma^{b,u} + \sum_{b'\in\mathcal{B}-\{b\}}\sum_{u'\in\mathcal{U}_{b'}} g^{b',u',b}_r p^{b',u'}_r) - Mz^{b,u}_{r,i} \quad \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{14}$$

$$M(1-z^{b,u}_{r,i}) \geq x^{b,u}_{r,i} \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{15}$$

$$\sum_{r\in\mathcal{R}}\sum_{i\in\mathcal{I}} x^{b,u}_{r,i} = \sum_{s\in\mathbb{N}\cap[1,|\mathcal{R}|]}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} s\times y^{b,u}_{s,i,c} \forall b\in\mathcal{B}, u\in\mathcal{U}_b, \tag{16}$$

$$\sum_{s\in\mathbb{N}\cap[1,|\mathcal{R}|]}\sum_{c\in\mathcal{C}} y^{b,u}_{s,i,c} \leq \beta^{b,u}_i,\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, i\in\mathcal{I}, \tag{17}$$

$$\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_b}\sum_{s\in\mathbb{N}\cap[1,|\mathcal{R}|]}\sum_{i\in\mathcal{I}} t_{s,i,c}\times y^{b,u}_{s,i,c} \leq d,\ \forall c\in\mathcal{C} \tag{18}$$

The objective function in (2) minimizes the total energy consumption of radio transmission and BBU processing. Equations (3), (4), (5), and (6) ensure that the decision variables are binary, while (7) ensures that the power variable is continuous and non-negative. Equation (8) ensures that users belonging to one base station cannot use the same RB and that no more than one MCS can be used on this RB. The minimum throughput requirement of a user is ensured by (9), while the limit on the total transmission power of a user is imposed by (10). Equation (11) ensures that the signal power of a user on an RB is zero if this RB is not used. Equations (12) and (13) together ensure that a user transmits using the same MCS index over all its assigned RBs. Knowing that using an MCS index requires the SINR to be above a threshold, equations (14) and (15) together make sure that if the SINR is lower than the threshold of an MCS index, then the user cannot use this MCS index. This condition is enforced by using an auxiliary binary decision variable $z^{b,u}_{r,i}$. To find the processing time and throughput for a user, it is

necessary to know the total number of used resource blocks by a user [7, 4]; this is done by (16) and (17). Finally, (18) ensures that each CPU can process the data assigned to it without violating the deadline constraint.

On the other hand, to understand how different objectives can affect the benefit of joint allocation, we consider a modified optimization problem that maximizes the total throughput while maintaining the same constraints as before. While Eq. (2) minimizes the total energy consumption, the modified objective in Eq. (19) maximizes the total throughput of users in the system. The objective function becomes as follows:

$$\max \quad \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_b}\sum_{s\in\mathbb{N}\cap[1,|\mathcal{R}|]}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} R_{s,i}\times y_{s,i,c}^{b,u} \tag{19}$$

4. Matching Game-Based Solution

The optimization problem proposed in the previous section is an NP-hard problem [9]. For this reason, we are interested in proposing a lower-complexity algorithm that can perform as close to the MILP problem solver's optimal solution as possible, which jointly allocates radio and computing resources to reduce overall energy consumption. In other words, such an algorithm seeks to output the optimal MILP problem solutions regarding the MCS index, the number of RBs, the power, and the CPU resources. We note that we are considering the case where computing resources and radio resources are. Hence, our algorithm neglects computing resource allocation, as all frames can get CPU resources within the deadline. However, the required computing resources depend on the MCS and the number of RBs, which are radio parameters. To optimize the total energy consumption, our algorithm takes the interest of the computing schedulers (i.e., minimizing the total energy consumption) when allocating RBs and MCS indexes. To that end, we propose a two-phase algorithm described as follows:

- Step 1: In the first step of the algorithm, we apply matching theory to associate RBs and MCS index to each user. The details of this step are shown in Section 4.1.

- Step 2: The second step aims at adjusting the radio transmission power to minimize the total energy consumption. This step is detailed in Section 4.2.

Before delving into the details of these two steps, we first go over the definitions of the matching games and stable matching.

Definition 1. *Given the set of users $\mathcal{U}_b$ associated to a base station b and the set of resource blocks $\mathcal{R}$, a many-to-one matching*[1] *μ is a function on the set $\mathcal{U}_b \cup \mathcal{R}$ such that:*

- *$\forall r \in \mathcal{R}$ either $\mu(r) \in U$ or $\mu(r) = \emptyset$, i.e., each RB is either matched to one user or unmatched;*
- *$\forall u \in \mathcal{U}$, $\mu(u) \in \mathcal{P}(R)$ where $\mathcal{P}(R) = \{R' \subseteq R\}$, i.e., each user is matched to a feasible set of resource blocks;*
- *$\forall r \in \mathcal{R}$ and $\forall u \in \mathcal{U}$, $\mu(r) = u$ if and only if $r \in \mu(u)$, i.e., an RB r is matched to a user u if and only if the user u is matched to the RB r.*

In a matching game, the two sets of agents rank their preferences and attempt to find a stable match [28]. We define $\succ_u$ as the "prefer" relation of a user u such that $r \succ_u r'$ when user u prefers r to r'. Similarly, $\succ_r$ is the "prefer" relation of RB r such that $u \succ_r u'$ when RB r prefers u to u'.

The definition of a stable matching is given as follows:

Definition 2. *A matching is said to be stable if there are no blocking pairs and it is individually rational, where*

- *a blocking pair is a couple of agents (r,u), such that $u \succ_r \mu(r)$ and (a) $r \succ_u \emptyset$ or (b) there exist $t \in \mu(u)$ such that $r \succ_u t$;*
- *an individually rational matching is when no agents of the two sets would be better off by breaking the current matching, i.e., if $\mu(r) = u$ then $u \succ_r \emptyset$ and $\mu(u) \succ_u (\mu(u) \setminus r)$.*

The problem of allocating RBs to users from different base stations can be seen as a many-to-many matching that maps the users of all base stations to the available RBs. Each user can be associated to many RBs and each RB to many users, under the constraint that no two users from the same base station can use the same RB.

[1]This model is also known as the *college admission problem*. When each element of the set $\mathcal{U}_b$ can be matched with at most one element of $\mathcal{R}$ and vice-versa, each element of the set $\mathcal{R}$ can be matched with at most one element of $\mathcal{U}_b$, the matching μ is said *one-to-one*. This model is also known as the *marriage problem*. When each element of $\mathcal{U}$ can be matched with many elements of $\mathcal{R}$ and vice-versa the matching μ is said *many-to-many*.

4.1. Step 1: The matching algorithm for associating RBs and MCS index to users

We describe and analyze in this section the Step 1 of our proposed matching game-based solution.

4.1.1. Algorithm Description

In this step, users assume that they fully use the total available radio transmission power P_{Tx}^{max}. The algorithm operates iteratively. In each iteration, each user forms a preference list (PL) by ranking the set of available RBs in decreasing order of achievable SINR. To find the achievable SINR, each user will equally divide P_{Tx}^{max} on all the RBs already assigned to the user plus the RB it is measuring the SINR on. The user proposes to only one RB in each iteration. In contrast, RBs rank users based on the decreasing channel gain. After receiving all users' proposals, the RBs accept or reject users. We note that an essential condition must be met: two users from the same base station cannot be assigned to the same RB.

In the next iteration, a user only proposes to a new RB if its minimal throughput target is not yet satisfied and if this new RB will not worsen the throughput. We note that if the interference on this new RB is high, the MCS index may become so low that the throughput worsens.

We should point out that a user only proposes to a given RB once and does not propose again if it is rejected. Algorithm 0 describes this procedure. At the end of this algorithm, all users who transmit operate using their maximum transmission power P_{Tx}^{max}.

It is worth mentioning that in each iteration, the preference of users on RBs could change due to dynamics caused by other users matching with other RBs. However, only semi-dynamic Preference Lists (PLs) are considered to ensure the algorithm converges to a stable matching; this practically means that a user can change its preference for the RBs to which it has not yet proposed, but it does not change the preference over RBs to which it has proposed earlier. The stability proof is demonstrated in the following subsection. At the end of the algorithm, the SINR for each user is calculated on each RB assigned to this user. The minimum SINR of these RBs should be higher than the threshold of the assigned MCS. The assigned MCS index is chosen such that it is the highest one satisfying this condition.

Algorithm 1 Matching Game for RBs and MCS index allocation

Input: $\mathcal{B}$,$\mathcal{U}_{b\in\mathcal{B}}$, $\mathcal{R}$, $\mathcal{I}$...
Output: RB and MCS assignment.

Forming RBs' preference lists:
for $\forall r \in \mathcal{R}$ **do**
  $PL_r = \{u_1, u_2...., u_i...|u_i \in \mathcal{U}_b, \forall b \in \mathcal{B}$ and $g_r^{b',u_i,b'} >= g_r^{b'',u_{i+1},b''}, \forall b', b'' \in \mathcal{B}\}$
end for
Forming users' preference lists:
for $b \in \mathcal{B}$ **do**
  for $b \in \mathcal{U}_{b\in\mathcal{B}}$ **do**
    $PL_u = \{r_1, r_2, ..., r_i, ...|r_i \in \mathcal{R}$ *and* $SINR^{b,u}(r_i) \geq SINR^{b,u}(r_{i+1})\}$
  end for
end for
REPEAT=1
while REPEAT **do**
  REPEAT=0
  for $b \in \mathcal{B}$ **do**
    for $u \in \mathcal{U}_b$ **do**
Semi-dynamic adjustment of users' preference lists:
      $PL_u = \{r_1, r_2, ..., r_i, ...|r_i \in PL_u$ *and* $SINR^{b,u}(r_i) \geq SINR^{b,u}(r_{i+1})\}$
      if $th_{b,u} < R_{min}^{b,u}$ and $th_{b,u} < th_{b,u}(PL_u^1)$ **then**
Users propose to RBs:
        $prop(u) = \{PL_u^1\}$
        $prop(r) = prop(r) \bigcup \{u\}$
        $PL_u = PL_u \setminus prop(u)$
      else
        prop(u) = ϕ
      end if
    end for
  end for
  for $\forall r \in \mathcal{R}$ **do**
    for $\forall u \in prop(r)$ **do**
A RB accept or reject a user:
      $ACCEPT^r(u)$ or $REJECT^r(u)$
      if $ACCEPT^r(u)$ **then**
        REPEAT = 1
      end if
    end for
  end for
end while

4.1.2. Proof of Matching-Stability

In many-to-one matching games, a stable matching may not exist. We introduce a restriction on users' preferences to guarantee the solution's stabil-

ity. In particular, we assume that users can modify their preferences for RBs to which they have not proposed yet, but they can not change their preferences for RBs to which they have already proposed. Under this assumption, we can state the following stability theorem along with its proof.

Theorem 1. *Step 1 of the matching-based algorithm for associating RBs and MCS index to users converges to a stable matching.*

Proof. Using Definition 2, in order to show that our algorithm converges to a stable matching, we have to prove that (i) there are no blocking pairs and (ii) the matching is individually rational.

Given that the algorithm has converged such that RB r is matched with user u' and RB r' is matched with user u, suppose there is a blocking pair (u, r) which consists of a user u and an RB r, such that r prefers u to its current matching u' and u prefers r to one of its associated RB r'. However, if user u prefers r to r', it should have proposed to it following the preference list, and either it did so and got rejected by r, which means that r prefers its current matching to u and this is a contradiction of the assumption above, or u prefers r' to r which also contradicts the assumption above. Consequently, our algorithm converges to a stable matching without any blocking pair.

The matching is individually rational because a user u will have proposed to an RB r only because its utility (i.e., throughput) increases such that $\mu(u) \succ_u (\mu(u) \setminus r)$, and on the other hand an RB does not accept a user that is not in its preference list, i.e., a user that is ranked inferior to the "no match" option. □

4.2. Step 2: Transmission power adjustment for energy consumption reduction

For each user, the transmission radio power is equal. This could lead to having a high SINR on an RB but a lower one on another. The MCS index that can be used is limited by the lowest SINR on the assigned RBs. Hence in this second step of our proposed solution, users who are already satisfied decrease their MCS indexes as much as possible under the condition that they remain satisfied. Then all the satisfied users minimize their radio power by solving the following convex problem. To reduce power consumption, users need to set their SINR to be precisely equal to the threshold of the MCS they are using. This problem is solved again if at least one previously non-satisfied user becomes satisfied due to the reduction of interference and the

ability to use a higher MCS index; this problem will be repeated, including the newly satisfied users. We reuse the notations in Table 1. We define $\overline{\mathcal{U}}_b$ as the set of users who attained their requirement, $\mathcal{R}^{b,u}$ as the set of RBs assigned to user u, and $MCS_{b,u}$ as the lowest MCS index a user can use while remaining satisfied. The following optimization problem minimizes the total radio power consumption; it only has $p_r^{b,u}$ as a decision variable, given that the RBs and MCS will already be determined before solving the problem.

$$\min \quad \sum_{b\in\mathcal{B}} \sum_{u\in\overline{\mathcal{U}}_b} \sum_{r\in\mathcal{R}} p_r^{b,u} \tag{20}$$

$$\text{s.t.} \quad \sum_{r\in\mathcal{R}^{b,u}} p_r^{b,u} \leq P_{Tx}^{max}, \forall b \in \mathcal{B}, u \in \overline{\mathcal{U}}_b \tag{21}$$

$$\sum_{r\notin\mathcal{R}^{b,u}} p_r^{b,u} = 0, \forall b \in \mathcal{B}, u \in \overline{\mathcal{U}}_b \tag{22}$$

$$p_r^{b,u} \geq \frac{\gamma_{MCS_{b,u}}^{th}}{g_r^{b,u,b}} (\sigma^{b,u} + \sum_{b'\in\mathcal{B}-\{b\}} \sum_{u'\in\mathcal{U}_{b'}} g_r^{b',u',b} p_r^{b',u'}) \quad \forall b \in \mathcal{B}, u \in \overline{\mathcal{U}}_b, r \in \mathcal{R}^{b,u}, i \in \mathcal{I} \tag{23}$$

We note that, as presented in Eq. (1), the computing resources required by users depend on the number of allocated RBs and the MCS index. Hence, after the power-adjustment phase, the chosen radio parameters (i.e., MCS, RBs) will determine the computing resources required by users. We recall that our study considers there is no bottleneck for computing resources at the BBU pool. Hence the scheduling is done in a First-In-First-Out manner.

4.3. Algorithmic complexity

The worst-case scenario of the matching game algorithm (i.e., Step 1 of our proposed solution) refers to the case where all users propose to all RBs. The maximum number of iterations, in this case, is equal to: $|\mathcal{B}|\cdot|\max\mathcal{U}_b|\cdot|\mathcal{R}|$.

Regarding the power adjustment phase (i.e., Step 2 of our proposed solution), it is a convex optimization problem; hence, it can be solved efficiently with simplex/interior point methods. Given that the power adjustment would allow some users to become satisfied because the interference would decrease, allowing a higher MCS to be used, the algorithm will be repeated every time a user becomes satisfied. The worst scenario for this step refers to the case where at each iteration, only one more user gets satisfied.

If all users will eventually be satisfied, the number of iterations will not be repeated more than $|\mathcal{B}| \cdot |\max \mathcal{U}_b|$ times.

5. Simulation Settings

In this section, the simulation environment is presented along with the performance metrics used for evaluating our proposed solution.

5.1. Simulation environment

To code and run the simulation, we use MATLAB. The MATLAB code calls GUROBI optimizer to solve the MILP problem. We acknowledge that solving an MILP problem is an NP-hard problem, and it is impossible to use it in a real setting where allocation decisions have to be made every 1 ms; however, the MILP solver allows us to measure the potential gains of the optimal joint allocation of radio and computing resources. Our study considers an area with a variable number of base stations ranging from 1 to 8. Each base station is separated from its neighboring base stations by a minimal distance that follows a uniform distribution between 0.4km and 0.6km. Each base station has 24 RBs for transmission, where the frequency reuse equals 1. Each base station serves two users. One would argue that having only two users per base station is not realistic. However, the reason behind this simple choice is two-fold: On the one hand, as the MILP problem in NP-hard, adding more users and RBs could prevent our MILP problem from producing an optimal solution. On the other hand, we recall that our goal is to compare the performance of the joint allocation of radio and computing resources vs. the sequential allocation and to study the potential benefits of the joint allocation when the goal is energy consumption minimization. As we are targeting the case where radio and computing resources are sufficient to satisfy the QoS of users, having only two users is adequate to model this scenario.

The position of each user follows a Poisson Point Process (PPP) such that each user is located in a disk of a radius of 200m and centered at the base station. Each user demands a throughput that follows a uniform distribution between 0.25 and 8 Mbps, and the total demand of the users from the same base station does not exceed 8 Mbps. To find the SINR threshold of the MCS indexes γ_i^{th}, we use the tables in [29], which map the SINR threshold to a Channel Quality Indicator (CQI) with specified modulation order and code rate. Then, we use the MCS table in [7] to map each CQI to its corresponding

MCS index. Hence, we end up with a set of possible MCS indexes:{0, 2, 4, 6, 8, 11, 13, 15, 18, 20, 22, 24, 26, 28}. The maximum user transmission power respects the 3GPP specifications in [30]. Based on it, P_{Tx}^{max} should be equal to 23 dBm with a tolerance of +/- 2 dBm. Hence we fix $P_{Tx}^{max} = 250$ mW ≈ 24 dBm. We suppose the noise spectral density is -174 dBm/Hz, and the Noise Figure is 8dB. For the channel gain, we use the ABG model [31] that models path loss and shadowing at a carrier frequency equal to 2GHz. Moreover, we consider the effect of Rayleigh fading such that it follows an exponential distribution with a unit mean. Considering a Cloud-RAN architecture, the baseband processing of these base stations is hosted in a shared BBU pool. For simplicity, we consider just one CPU core with a power consumption of $P_{comp}^{c} = 30$W and a clock frequency equal to 2.4GHz. We also assume that when the CPU core executes BBU functions for users, it consumes the max CPU power P_{comp}^{c}. In contrast, we suppose the power consumption is zero when the CPU is idle. Our simulation setting focuses on the case where the sum of users' throughput demand is smaller than the system capacity, and the computing resources are sufficient to process the data of all users.

5.2. Performance metrics

To analyze the performance of our model, we consider the following performance metrics:

- *Radio transmission energy consumption;* the total radio transmission energy of all users in the system.
- *Computing energy consumption;* the total computing energy consumption of all users in the system.
- *Total energy consumption;* the sum of the total transmission and computing energy consumption in the system.
- *Throughput;* the sum of throughput of all users.
- *CPU Idle time*; The percentage of time for which the CPU is idle
- *Satisfaction ratio*; The ratio of achieved throughput of a user divided by its minimum requirement. For a user $u \in \mathcal{U}_b$, the satisfaction ratio is defined as

$$SAT(b,u) = \frac{th(u)}{R_{min}^{b,u}}$$

where $th(u)$ is the achieved throughput by user u, and $R_{min}^{b,u}$ is its requested throughput.

- *Fairness;* using Jain's fairness index [32] defined as:

$$J_I = \frac{\left(\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}_b} SAT(b,u)\right)^2}{(N \times \sum_{r\in\mathcal{R}}\sum_{u\in\mathcal{U}_b}(SAT(b,u))^2)}$$

- *Non-satisfied users;* the number of users that fail to achieve their requested throughputs.

- *RB utilization;* The percentage of utilization of the total RBs from all base stations.

6. Results

In this section, we plot and analyze the performance of the joint radio and computing resource allocation vs. the sequential resource allocation, considering the two objectives of minimizing the total energy consumption and maximizing the total throughput. As we mentioned before, the sequential allocation separates the allocation of radio resources from that of computing resources by solving them in order. So, when considering the objective of energy consumption minimization in the sequential allocation, the radio allocation aims at minimizing the radio transmission energy consumption, while the computing allocation minimizes the computing energy consumption. The radio allocation is done by modifying the joint MILP formulation such that the parameters, variables, and constraints related to computing resource allocation are removed. Then, after choosing the radio parameters, the CPU resources are allocated. Moreover, we compare the performance of the MILP-based solutions to that of the proposed low-complexity algorithms:

- the matching with fixed preference lists; this is step-1 of our solution, except that the preference lists are initialized in the beginning and are no more modified in each iteration;

- the matching with semi-dynamic preference lists; this is step-1 of our algorithm.

- the matching with semi-dynamic preference lists followed by radio power adjustment; this is our proposed two-step matching algorithm.

We have opted to compare the proposed matching-based solutions to the optimal values output by the solver. To the best of our knowledge, there is no algorithm that jointly allocates RBs, MCS, and power while considering the effect on the computing requirements. Hence, the best way to benchmark the performance of the proposed matching-based algorithms is to compare them to the optimal, which is the MILP. The performance is measured using the metrics defined in Section 5.2 as a function of the number of base stations managed by the same BBU pool. In the simulation, we only consider the allocation for one Transmission Time Interval (TTI) that is equal to 1 ms. The simulation is repeated 25 times for the MILP problems and 100 times for the matching algorithms, and the 95% confidence intervals are plotted.

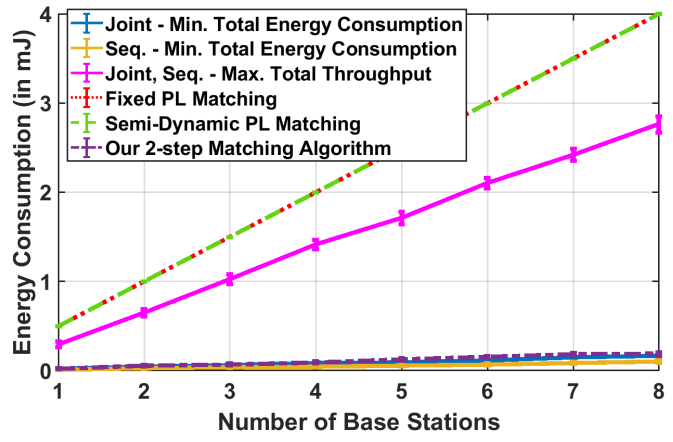
6.1. Radio transmission and computing energy consumption

Figures 4a, 4b, and 4c show the performance of the joint allocation problems vs. the sequential concerning the radio transmission energy consumption, computing energy consumption, and total energy consumption, respectively. The energy consumption is measured during one TTI. We clearly notice that :

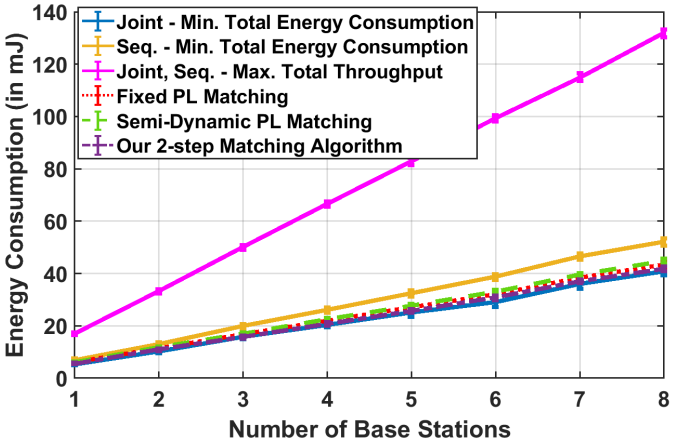
- When considering the objective of minimizing total energy consumption, the joint radio and computing resources allocation, compared to the sequential one, consumes more energy for radio transmission (as shown in Figure 4a), less energy for computing (Figure 4b) and less total energy (Figure 4c). For instance, when the number of BSs is equal to 8, the joint radio and computing allocation reduces the total energy consumption by 21.3% compared to the sequential counterpart.

- When considering the objective of throughput maximization, both joint and sequential allocation use all the available radio and computing resources to maximize the throughput. Hence, they both have similar performance for throughput maximization. This will be further explained later on in the paper. On the other hand, maximizing throughput objective would consume up to 325% more total energy than the joint allocation with the objective of minimizing energy consumption would do.

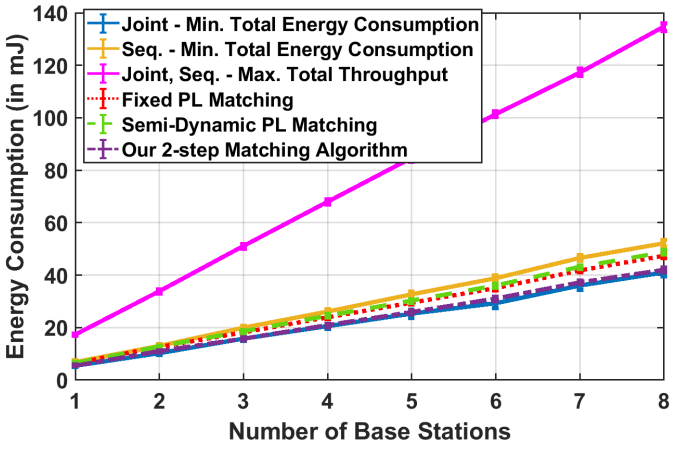
- The proposed matching algorithms without power adjustment fully use the available transmission power. However, as our two-step matching



(a)



(b)



(c)

Figure 4: Energy consumption as a function of the number of base stations managed by the BBU pool: (a) Tx energy consumption (in mJ), (b) Computing energy consumption (in mJ), (c) Total energy consumption (in mJ)

algorithm combines the power adjustment algorithm with the matching-based algorithm, the radio transmission energy consumption is very close to that of the joint variation of the energy consumption minimization algorithm.

- Regarding computing energy, the fixed PL and the two-step algorithm performs very close to the joint allocation with the objective of minimizing energy consumption. The Semi-Dynamic-PL Matching consumes a little more. The reason behind these results is that the Semi-Dynamic-PL Matching algorithm allows for using even higher MCS indexes, which require more computing resources. On the other hand, combining it with power adjustment in the two-step algorithm allows for decreasing the MCS to the least one that can satisfy the required throughput of a user. This permits to reduce the excess allocated resources and helps satisfy more users. Hence the computing energy consumed by the latter algorithm is close to the MILP-based joint allocation of the energy consumption minimization. Overall, the total energy consumption for our two-step algorithm is the closest to the joint energy consumption minimization algorithm and is less than the other two matching algorithms.

6.2. Throughput

The performance concerning the throughput metric as a function of the number of base stations in the BBU pool is plotted in Fig. 5. Since the objective of minimizing the total energy consumption must guarantee the requested throughput for every user, both the joint and sequential allocation achieve similar results with this objective. The slight differences result from the different decisions on the MCS indexes and number of RBs; together, they control the TBS size, which indicates the throughput. Our two-step algorithm achieves the lowest throughput while it remains close to the optimal solution of the MILP problem. The other matching algorithms achieve slightly higher throughput The reason behind this tendency is that some users can use high MCS indices when employing high radio power, allowing them to do far more than what they had expected.

6.3. CPU idle time

Fig. 6 shows the percentage of CPU idle time as a function of the number of base stations connected to the BBU pool. Using the computing model

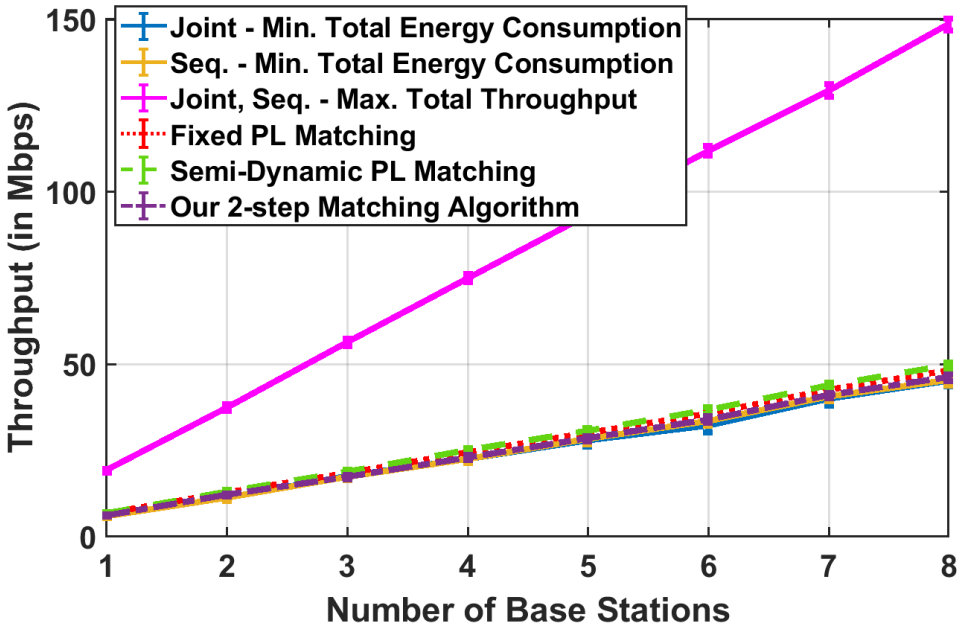


Figure 5: Throughput as a function of the number of base stations in the BBU pool

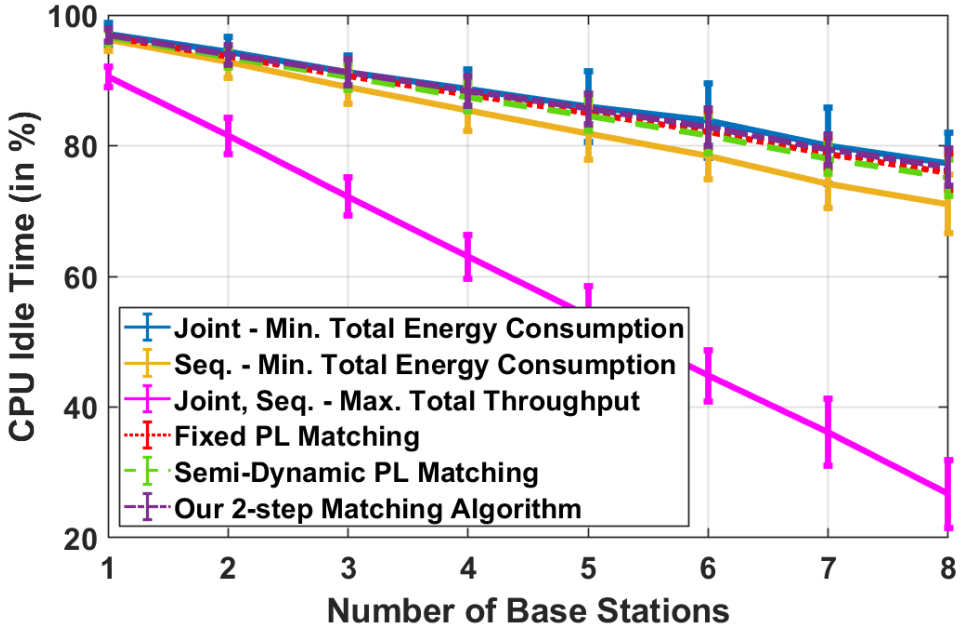


Figure 6: CPU Idle time as a function of the number of base stations in the BBU pool

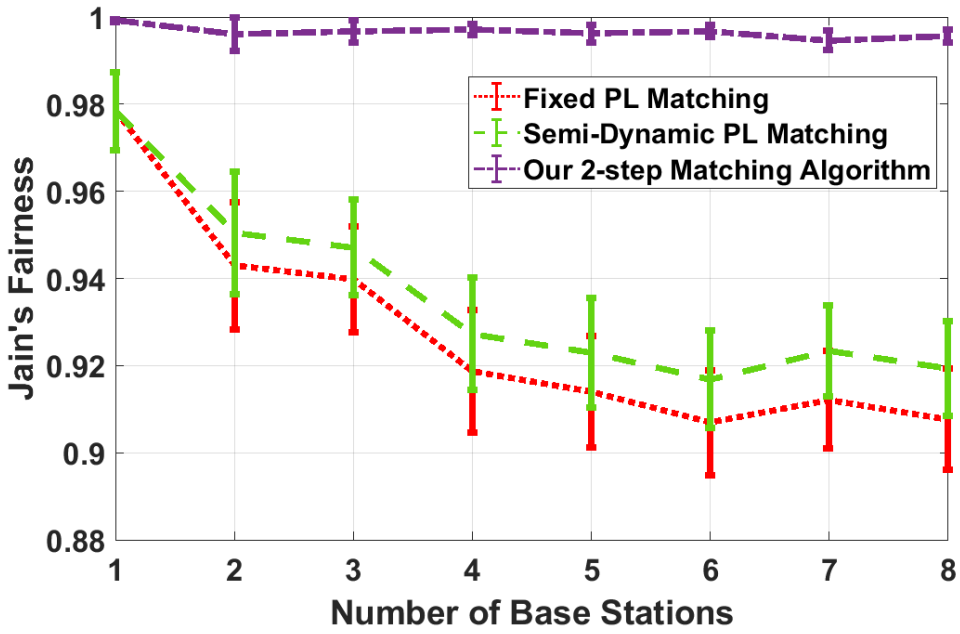


Figure 7: Fairness as a function of the number of base stations in the BBU pool

described in section 3.2, minimizing the computing energy consumption is consistent with reducing the CPU utilization, or in other words, increasing the CPU idle time. This explains why the joint allocation with the objective of minimizing energy consumption, which is the best algorithm that reduces compute energy consumption, achieves higher CPU idle times than the sequential allocation. Again, both the joint and sequential allocation with the objective of maximizing throughput have much lower CPU idle time than energy consumption minimization objectives counterparts. This is interpreted by the fact that maximizing throughput algorithms make the most use of all available computing resources. On the other hand, the matching-based algorithms, and especially our two-step algorithm, preserve outcomes that are similar to the joint MILP-based problem with energy consumption minimization.

6.4. Fairness and users' satisfaction

Given that the proposed matching algorithms are not guaranteed to satisfy the requirements of all users as the MILP problems do, we study and plot the graphs of fairness and the percentage of non-satisfied users for the matching-based algorithms as a function of the number of base stations in

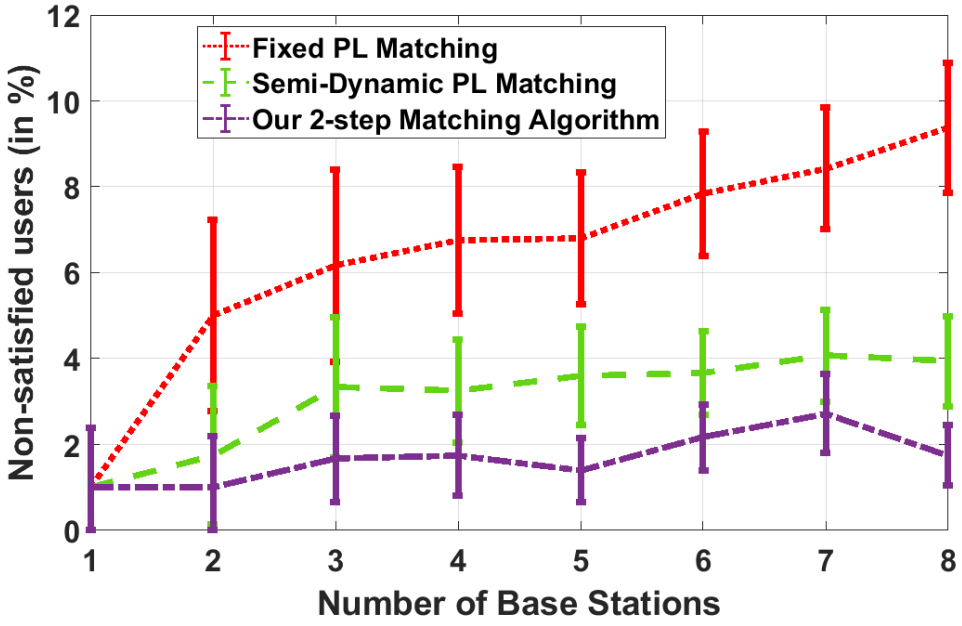


Figure 8: Percentage of Non-satisfied users as a function of the number of base stations in the BBU pool

the BBU pool in figures 7 and 8. Compared to the fixed PL matching, the results show that allowing the semi-dynamic preference lists increases the fairness and the number of satisfied users. Additionally, the performance in terms of fairness and the number of satisfied users is further enhanced when enabling power adjustment in our two-step algorithm. Furthermore, we plot the box-plots of the satisfaction ratio of non-satisfied users in figure 9. We define the satisfaction ratio as the achieved throughput divided by the required throughput.The results show that combining matching and power adjustment in the two-step algorithm is better than the other two algorithms, recalling that the number of non-satisfied users dropped when combining the algorithms. This combination not only satisfies more users but also improves the satisfaction ratio among non-satisfied users.

6.5. MCS and RB selection

To understand how the various algorithms behave, we observe the algorithms' decisions based on the MCS indexes assignment and the number of RBs assigned to users. Fig. 10 shows the percentage of utilized RBs in each base station as a function of the total number of base stations connected

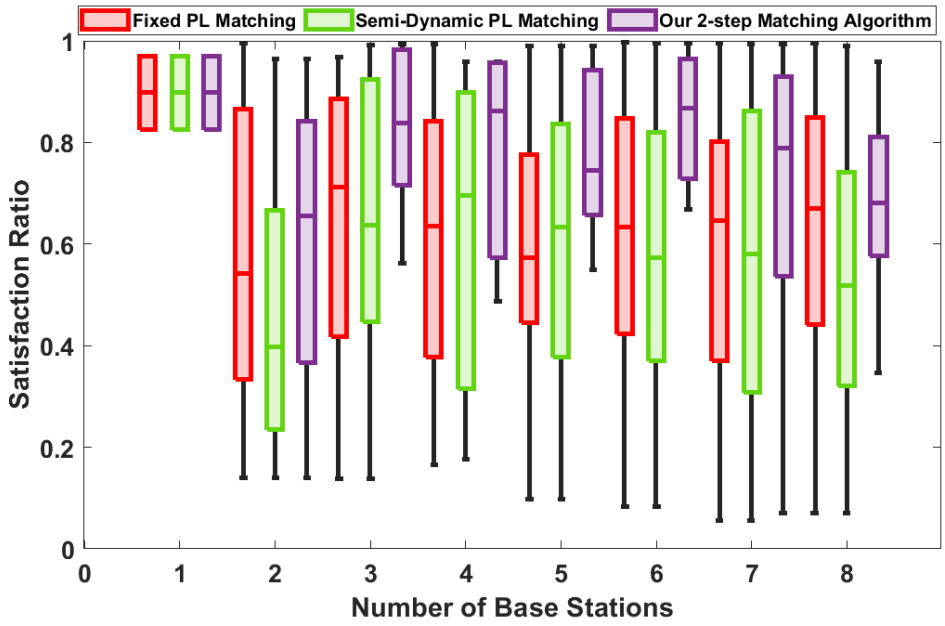


Figure 9: Box plot of the satisfaction ratio of non-satisfied users

to the BBU pool, and Fig. 11 shows the cumulative distribution probability of the selected MCS indexes for each algorithm. Considering the energy consumption minimization objective and analyzing both figures, the joint algorithm tends to favor allocating users with a low number of RBs but high MCS indexes. In fact, achieving the same throughput for a given user could be done by either using a lower number of RBs and a high MCS index if the SINR level permits so, or using a higher number of RBs with a lower MCS index [7]. The joint allocation with the objective of minimizing energy consumption would go for the first alternative which is increasing the MCS index but transmitting over a low number of RBs. This requires increasing the transmission power to increase the MCS over these RBs but would decrease the required computing resources. As a result, the computing energy consumption decreases, so the total combined energy consumption decreases. In contrast, the sequential allocation firstly solves the radio allocation that minimizes transmission power independently of the computing resources allocation.The results show that the sequential algorithm minimizes the transmission power and spreads the data over a higher number of RBs but with a lower MCS index. On the other hand, the maximizing throughput algorithm

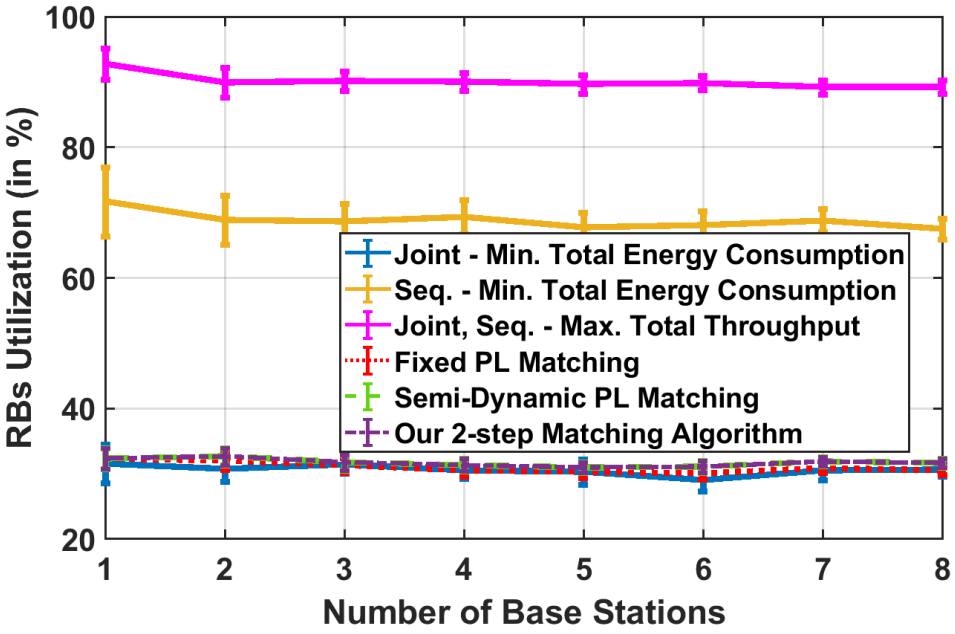


Figure 10: RBs Utilization as a function of the number of base stations in the BBU pool

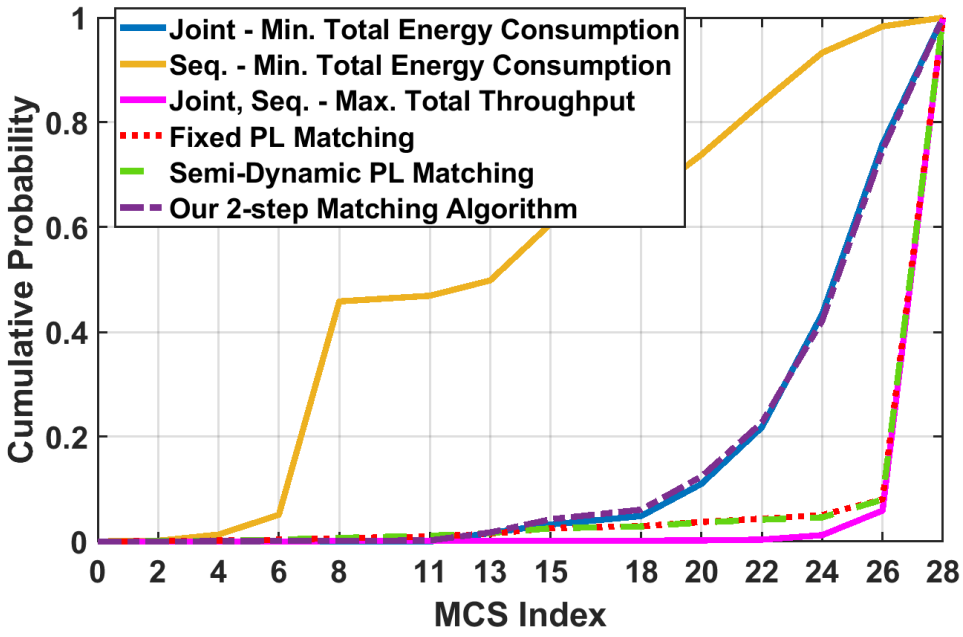


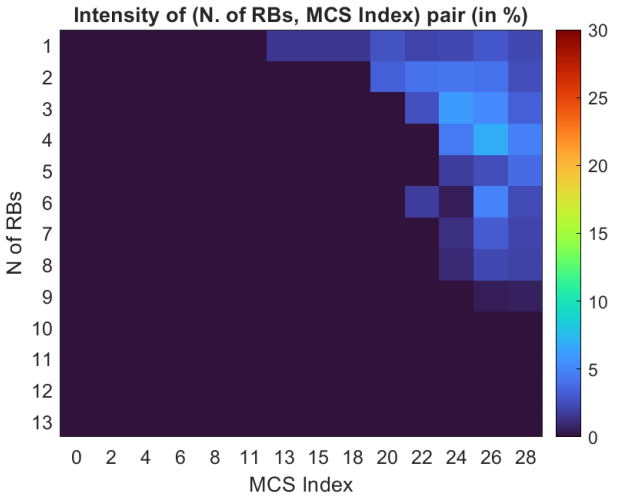
Figure 11: The Cumulative Distribution Function of MCS assignment

uses all the RBs in each base station and the maximum transmission power for every user. This justifies the very high RB utilization and the high MCS indexes, as Fig. 10 and Fig. 11 show. However, the maximizing throughput algorithm should ensure its selections do not increase the interference, worsening performance. On the other hand, the fixed PL matching algorithm has an RB utilization that is very close to that of the joint energy consumption minimization. When the preference lists are allowed to be partially modified, some users will take advantage of that in an attempt to improve their throughput by using more resource blocks. That is why the semi-dynamic PL matching with and without power adjustment uses a slightly higher number of RBs. Regarding the MCS, Fig. 11 shows that both matching algorithms without power adjustment use high MCS indexes. When power adjustment is applied in our two-step algorithm, and the power gets decreased so that a user does not receive more than its request, it would be possible to use lower MCS indexes.

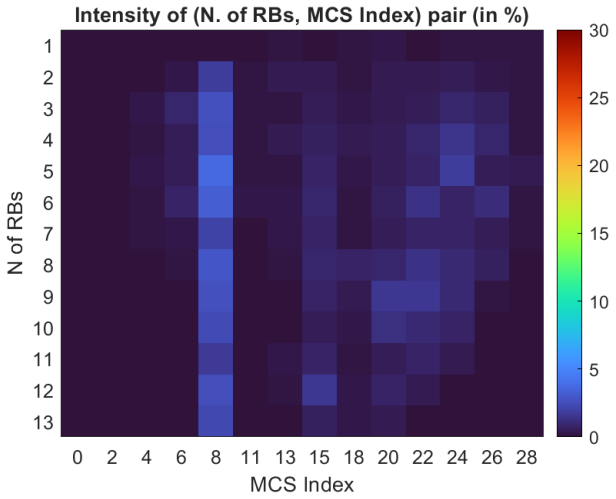
Fig. 12 further supports this previous explanation. In fact, these heat maps show the intensity of assigning the pair composed of 1) the number of resource blocks and 2) MCS indexes to users. While the joint energy consumption minimization algorithm intensely allocates a lower number of RBs to users and a very high MCS index, the sequential favors assigning a higher number of RBs but lower MCS indexes. On the other hand, the maximizing throughput algorithms assign more RBs and high MCS indexes to users. Moreover, the matching algorithms without power adjustment tend to use fewer RBs and excessively high MCS indexes. In contrast, the two-step algorithm decreases the MCS index for some users, which automatically improves the MCS for others. Hence, while it uses a lower number of RBs, this algorithm uses high, but not excessively high, MCS indexes to satisfy more users.

6.6. Execution Time

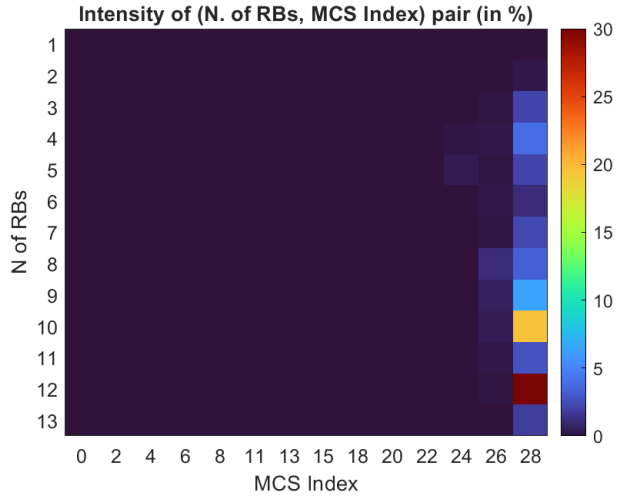
Fig. 13 shows the reduction of the execution time of the matching algorithms relative to the Joint-MILP-Energy Consumption Minimization. Our algorithm significantly reduces the execution time. However, we are measuring the performance using MATLAB on a Core-i9-9880H CPU. This adds overhead. Hence, our results achieved times that are higher than 1 ms. However, since our algorithm has a tractable complexity, it would be possible to implement our algorithms using low-level code, which would permit to respect the 1 ms requirement.



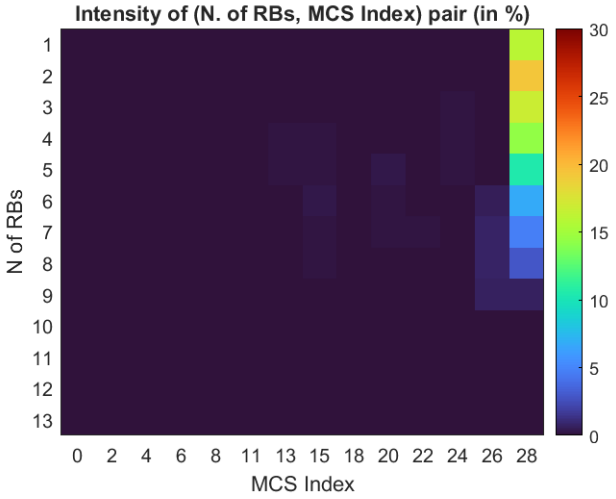
(a) Joint - Minimize energy consumption



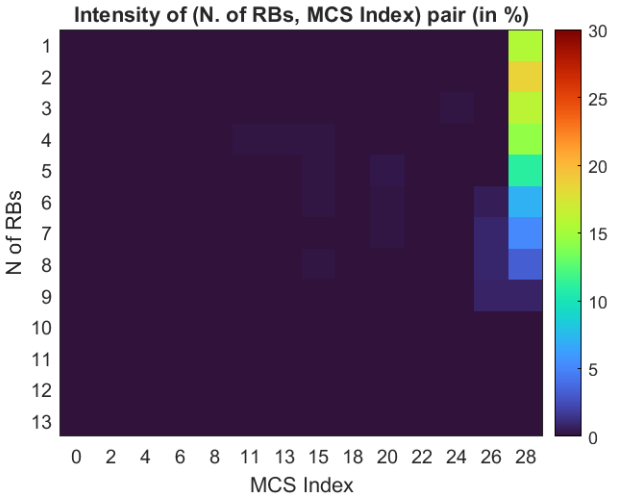
(b) Sequential - Minimize energy consumption



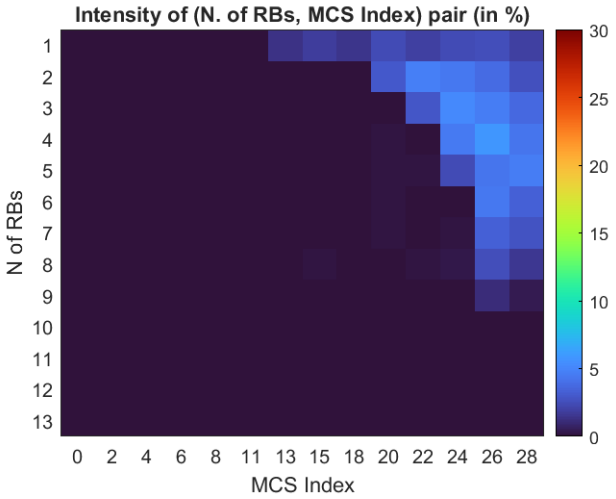
(c) Joint, Sequential - Maximize throughput



(d) Matching - Fixed PL



(e) Matching - Semi-Dynamic PL



(f) Our two-step Matching Algorithm

Figure 12: The intensity of assigning a pair of (number of Resource Blocks, MCS index) to users

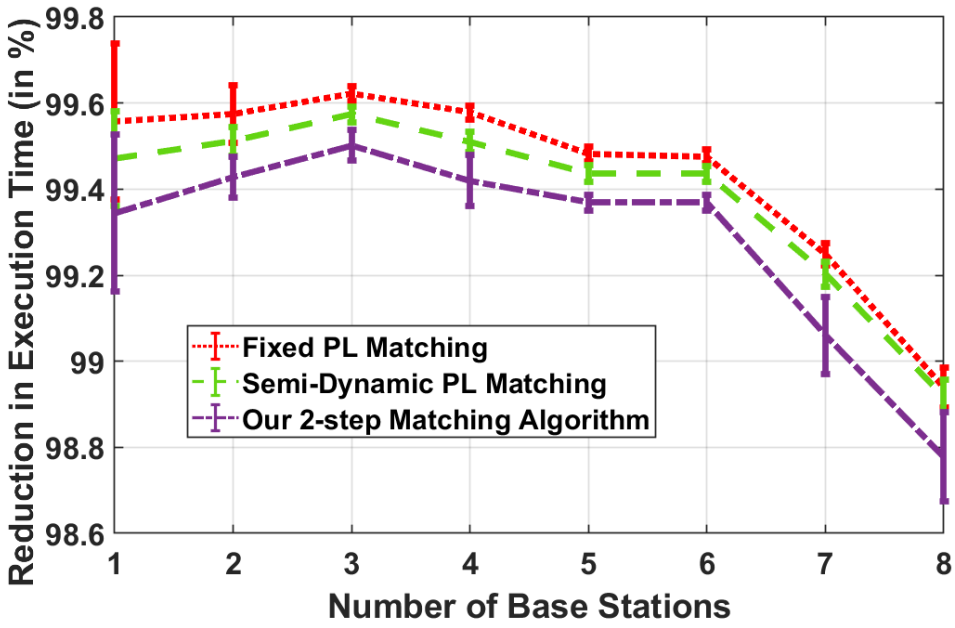


Figure 13: Redution of execution Time relative to Joint - Min. Total Energy Consumption as a function of the number of base stations in the BBU pool

As final notes, we have reduced the size of the MILP problem and made the problem tractable by using a small number of resource blocks, a small number of users, and a small number of base stations. The reason is that an MILP problem is known to be NP-hard and is unable to output results for a larger network. However, the tendency of the three variants of the matching-based solutions will persist for a larger network. The matching with semi-dynamic PL is at least as good as the one with fixed PL. On the other hand, including power adjustment in the 2-step solution will produce results that are at least as good as the matching without power-adjustment. Moreover, the results of the maximizing throughput objective help us understand the performance of the energy minimization objective in case all the radio resources are required to satisfy the demands of users. Suppose that the QoS requirements (i.e., minimum throughput) of users are increased so that the throughput maximization objective can fulfill the needs without being able to improve the assigned data rates. This means, more or less, all the radio resources (i.e., RBs and transmission power) are needed to satisfy the minimum throughput requirement (i.e., Equation (9)). On the other hand, since the energy consumption minimization objective should fulfill the mini-

mum requirement constraint, it will give the same results as the throughput maximization objective. In other terms, all the radio resources are needed, and it is not possible to use less to save energy. In such a case, the joint and sequential allocation will perform the same, even if the goal is energy consumption minimization, as long as the bottleneck happens at the level of the radio resources.

Finally, the proposed 2-step algorithm can perform close to the joint energy consumption minimization MILP algorithm, with good fairness among users and much lower complexity. Not only does our algorithm achieve near-optimal solutions, but also it reduces the execution time significantly relative to the MILP solver. Hence, it can serve as a suitable alternative to the MILP problem for efficient implementation by operators.

7. Conclusion

In this paper, we have studied the performance of joint radio and computing resources allocation in Cloud-RAN. We have formulated a Mixed Integer Linear Programming model and compared the performance of the joint allocation with respect to a sequential allocation, considering the two objectives of minimizing energy consumption and maximizing throughput. The results demonstrate that when the computing resources are sufficient, the joint allocation with the objective of minimizing energy consumption is beneficial and achieves performance gains by reducing the total energy consumption. Given that we used a high-complexity problem solver to analyze the benefits of joint allocation and that it is impractical to use such a solver in an actual implementation, we proposed a lower-complexity two-step matching algorithm with a power adjustment mechanism to perform close to the MILP optimization problem. Our results showed that the proposed algorithm could perform close to the joint allocation for the energy consumption minimization problem but with much lower complexity. As the RAN architecture is moving towards Open RAN which encourages multi-vendor support and openness, we plan to test the benefits of joint allocation of radio and computing resources in the context of RAN sharing between multiple operators considering different objectives including profit maximization and cost minimization.

Acknowledgments

This work has been partially carried out in the framework of ANR HEIDIS (https://heidis.roc.cnam.fr/ ; ANR-21-CE25-0019) project.

References

[1] M. A. Habibi, M. Nasimi, B. Han, H. D. Schotten, A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System, IEEE Access 7 (2019) 70371–70421. doi:10.1109/ACCESS.2019.2919657.

[2] A. Gupta, R. K. Jha, A survey of 5g network: Architecture and emerging technologies, IEEE Access 3 (2015) 1206–1232. doi:10.1109/ACCESS.2015.2461602.

[3] C. Mobile, C-RAN: the road towards green RAN, White Paper, ver 2 (2011) 1–10.

[4] S. Khatibi, K. Shah, M. Roshdi, Modelling of Computational Resources for 5G RAN, in: 2018 European Conference on Networks and Communications (EuCNC).

[5] M. Sharara, S. Hoteit, P. Brown, V. Vèque, Coordination between Radio and Computing Schedulers in Cloud-RAN, in: 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM).

[6] M. Sharara, S. Hoteit, P. Brown, V. Vèque, On Coordinated Scheduling of Radio and Computing Resources in Cloud-RAN, IEEE Transactions on Network and Service Management (2022). doi:10.1109/TNSM.2022.3222068.

[7] 5G; NR; Physical layer procedures for data, ETSI TS 138 214 V15.3.0 (October 2018).

[8] Z. Han, Y. Gu, W. Saad, Matching Theory for Wireless Networks, Springer, 2017. doi:10.1007/978-3-319-56252-0.
URL `http://link.springer.com/10.1007/978-3-319-56252-0`

[9] B. Korte, J. Vygen, Combinatorial Optimization: Theory and Algorithms, 5th Edition, Springer Publishing Company, Incorporated, 2012.

[10] S. Bian, J. Song, M. Sheng, Z. Shao, J. He, Y. Zhang, Y. Li, I. Chih-Lin, Sum-rate maximization in OFDMA downlink systems: A joint subchannels, power, and MCS allocation approach, in: 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC).

[11] Y. Li, H. Xia, J. Shi, S. Wu, Joint optimization of computing and radio resource for cooperative transmission in C-RAN, in: 2017 IEEE/CIC International Conference on Communications in China (ICCC).

[12] M. M. Abdelhakam, M. M. Elmesalawy, Joint Beamforming Design and BBU Computational Resources Allocation in Heterogeneous C-RAN with QoS Guarantee, in: 2018 International Symposium on Networks, Computers and Communications (ISNCC).

[13] L. Ferdouse, A. Anpalagan, S. Erkucuk, Joint Communication and Computing Resource Allocation in 5G Cloud Radio Access Networks, IEEE Transactions on Vehicular Technology 68 (9) (2019).

[14] F. Shirzad, M. Ghaderi, Joint computing and radio resource allocation in cloud radio access networks, in: 2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS), 2021, pp. 518–526. doi:10.1109/MASS52906.2021.00070.

[15] M. Sharara, S. Hoteit, V. Vèque, A recurrent neural network based approach for coordinating radio and computing resources allocation in cloud-ran, in: 2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR), 2021, pp. 1–7. doi:10.1109/HPSR52026.2021.9481812.

[16] S. Matoussi, I. Fajjari, N. Aitsaadi, R. Langar, User Slicing Scheme with Functional Split Selection in 5G Cloud-RAN, in: 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020, pp. 1–8. doi:10.1109/WCNC45663.2020.9120828.

[17] M. K. Motalleb, V. Shah-Mansouri, S. Parsaeefard, O. L. A. López, Resource allocation in an open ran system using network slicing, IEEE Transactions on Network and Service Management (2022) 1–1doi:10.1109/TNSM.2022.3205415.

[18] C. W. Zaw, N. H. Tran, Z. Han, C. S. Hong, Radio and computing resource allocation in co-located edge computing: A generalized nash equilibrium model, IEEE Transactions on Mobile Computing 22 (4) (2023) 2340–2352. doi:10.1109/TMC.2021.3120520.

[19] Y. Wu, K. Ni, C. Zhang, L. P. Qian, D. H. K. Tsang, Noma-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation, IEEE Transactions on Vehicular Technology 67 (12) (2018) 12244–12258. doi:10.1109/TVT.2018.2875337.

[20] Z. Han, D. Niyato, W. Saad, T. Başar, Game Theory for Next Generation Wireless and Communication Networks, Cambridge University Press, 2019. doi:10.1017/9781108277402.
URL https://doi.org/10.1017/9781108277402

[21] Y. Gu, C. Jiang, L. X. Cai, M. Pan, L. Song, Z. Han, Dynamic Path to Stability in LTE-Unlicensed with User Mobility: A Matching Framework, IEEE Transactions on Wireless Communications 16 (7) (2017) 4547–4561. doi:10.1109/TWC.2017.2699966.

[22] K. Hamidouche, W. Saad, M. Debbah, Many-to-many matching games for proactive social-caching in wireless small cell networks, in: 2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WiOpt 2014, IEEE Computer Society, 2014, pp. 569–574. doi:10.1109/WIOPT.2014.6850348.

[23] S. Zhang, B. Di, L. Song, Y. Li, Radio resource allocation for non-orthogonal multiple access (NOMA) relay network using matching game, in: 2016 IEEE International Conference on Communications, ICC 2016, Institute of Electrical and Electronics Engineers Inc., 2016. doi:10.1109/ICC.2016.7510918.

[24] B. Di, S. Bayat, L. Song, Y. Li, Z. Han, Joint user pairing, subchannel, and power allocation in full-duplex multi-user ofdma networks, IEEE Transactions on Wireless Communications 15 (12) (2016) 8260–8272.

[25] C. Xu, G. Zheng, L. Tang, Energy-aware user association for NOMA-based mobile edge computing using matching-coalition game, IEEE Access 8 (2020) 61943–61955. doi:10.1109/ACCESS.2020.2984798.

[26] B. Zhang, X. Mao, J. Yu, Z. Han, Resource allocation for 5g heterogeneous cloud radio access networks with d2d communication: A matching and coalition approach, IEEE Transactions on Vehicular Technology 67 (7) (2018) 5883–5894.

[27] M. Sharara, S. Hoteit, V. Vèque, F. Bassi, Minimizing power consumption by joint radio and computing resource allocation in cloud-ran, in: 2022 IEEE Symposium on Computers and Communications (ISCC), 2022, pp. 1–6. doi:10.1109/ISCC55528.2022.9912943.

[28] K. Bando, Many-to-one matching markets with externalities among firms, Journal of Mathematical Economics 48 (1) (2012) 14–20.

[29] J. Fan, Q. Yin, G. Y. Li, B. Peng, X. Zhu, MCS Selection for Throughput Improvement in Downlink LTE Systems, in: 2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN).

[30] 5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (July 2018).

[31] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partyka, J. Jarvelainen, Propagation Path Loss Models for 5G Urban Micro- and Macro-Cellular Scenarios, in: 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring).

[32] R. Jain, D. Chiu, W. Hawe, A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems, DEC Research Report TR-301, 1984.