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

Mahdi Sharara, Sahar Hoteit, Véronique Vèque, Francesca Bassi

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

Mahdi Sharara*, Sahar Hoteit*, Véronique Vèque*, and Francesca Bassi†
*Laboratoire des Signaux et Systèmes, Université Paris Saclay-CNRS-CentraleSupélec, France
†Institut de Recherche Technologique SystemX, Palaiseau, France
Emails: {mahdi.sharara, sahar.hoteit, veronique.veque}@universite-paris-saclay.fr,
francesca.bassi@irt-systemx.fr

***Abstract*—Cloud-RAN is a key 5G-enabler; it consists in centralizing the baseband processing of 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 resources of multiple base stations. This paper aims to analyze the potential reduction in power consumption brought by the joint allocation of the radio and computing resources. We formulate a Mixed Integer Linear Programming (MILP) problem, considering the objective of power consumption minimization. For comparison, we consider the objective of throughput maximization. When the goal is power minimization, the joint allocation can minimize the total power consumption by up to 21.2%, with respect to the case where radio and computing resources in the BBU pool are allocated sequentially.**

***Index Terms*—Cloud-RAN, 5G, Joint Resource Allocation, Power Minimization**

I. Introduction

The demand for mobile data is experiencing unprecedented, massive growth. The 5th generation of mobile networks (5G) attempts to tackle this issue using different technologies, including Cloud Radio Access Network (Cloud-RAN), among many [1], [2]. Traditionally, a base station is composed of a Remote Radio Head (RRH) and a Base Band Unit (BBU). The RRH is responsible for the radio frequency functions, while the BBU executes the 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; it allows for a scalable and flexible network. In addition, centralization and cloudification allow for more efficient use of resources and reduce CAPEX and OPEX [1].

Strategies to reduce power consumption are very important in 5G [1], [2], since, to satisfy users' demand, a massive number of base stations must be deployed. We consider the joint allocation of radio and computing resources in Cloud-RAN, and we investigate how this may help minimize the power consumed for transmission and baseband processing. In this work, we assume Single-Input-Single-Output (SISO) transmission. Radio resources allocation consists in assigning to each user some Resource Blocks (RBs) and a Modulation and Coding Scheme (MCS), along with the transmission power. On the other hand, computing resources allocation consists in assigning users' radio frames to CPUs in the BBU pool, such that the processing deadline is met.

The throughput of users depends on their MCSs and on the number of RBs over which they transmit [4], [5]. The transmission power assigned to users affect the Signal-to-Noise-Ratio (SINR) of each user, which in turn controls the maximum MCS that the user is allowed to adopt. Moreover, the execution time of the base band processing of a user's radio frame increases with the number of RBs and especially with the MCS [6]. Hence, these two radio parameters affect the required computing resources and the computing power consumption. In a system where the radio and the computing resources allocations are done sequentially, the radio resources allocation minimizes only the transmission power. Thus, the computing resources allocation will not control the radio parameters that affect the computing power consumption. In contrast, a joint radio and computing resource allocation scheme would be able to simultaneously control all the radio and computing parameters to minimize the total transmission and processing power. It is thus interesting to investigate the benefits of having joint radio and computing resources allocation. While joint allocation would have increased computational complexity with respect to sequential allocation, it would be worthwhile if it were to exhibit significant power consumption reductions.

In this paper, 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 objective of minimizing the power consumption, and we compare it with the objective of throughput maximization. To quantify the impact of *joint* allocation, we compare it to a sequential scheme that performs radio resources allocation (via MILP) followed by computing resources allocation.

The rest of the paper is organized as follows: Section II surveys the related work. The MILP problem is formulated in section III. The simulation settings are described in section IV, and the results of the simulations are discussed in section V. Finally, our work is concluded in section VI.

II. Related Work

Works in the literature have considered radio or computing resource allocation, independently [7]–[9]. To optimize system throughput and energy efficiency, the authors of [7] have formulated a Mixed Integer Non-Linear Programming (MINLP) problem, then relaxed it into a lower-complexity two-step approach for each resource type. The computing

resource allocation occurs first by mapping users to Virtual Machines (VM). Secondly the radio resource allocation is done by controlling the beamforming vectors. Nevertheless, the scope of this allocation is limited as the algorithm did not consider the existence of multiple RBs or sub-carriers nor the selection of MCS indexes based on the SINR. The authors in [8] considered joint Beamforming vector design and BBU computational resources allocation. They aimed at minimizing the total system power consumption while considering the constraints of users' Quality of Service (QoS), fronthaul capacity, transmit power per RRH, and per Antenna. However this paper did not consider the RBs or MCS assignments and the effect on the required processing time. In [9], the authors investigated the joint communication and computing resource allocation. They considered power 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. Joint management of radio and computing resources has recently been considered in [5] in the case where the computing resources are insufficient to satisfy all users' demands. The authors propose two coordination schemes between radio and computing resources that aim at maximizing throughput and users' satisfaction. The considered schemes demonstrated a significant ability to decrease the amount of wasted transmission power. To reduce the complexity of the Integer Linear Programming (ILP)-based coordination algorithms, lower complexity Recurrent Neural Network (RNN)-based algorithms were developed in [10]. They were trained to perform close to the ILP solver and were shown to significantly reduce the execution time with respect to the ILP problems. Aiming to maximize the sum-rate under limited computing resources, the authors in [11] proposed an algorithm that schedules users who contribute less to computational outages and permits downgrading MCS indexes if the computing resources are insufficient. In another context, [12] formulated a radio allocation MILP problem. It considers RBs and MCS assignment, in addition to power allocation. However, the model is quite limited as it takes into account only one base station but none of the interference caused by other base stations. Additionally, the problem only considers radio allocation without considering computing resources allocation. Different from these research papers, we consider joint radio and computing resource allocation. Our model considers power allocation, MCS assignment, RBs allocation, CPU assignment, and computing resources allocation. To the best of our knowledge, the current literature neither explicitly compares joint vs. non-joint allocation, nor examines its advantages, limitations, or the influence of the chosen objective function.

III. Problem Formulation

A. Model Input and Parameters

To study the performance of joint radio and computing resources allocation, we consider the following: a set of Base Stations (RRHs) $\mathcal{B}$, 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). We focus on the uplink direction in which the BBU pool should execute the complex and energy consuming decoding function [5]. 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 $u' \in \mathcal{U}_{b'}$ and the base station $b \in B$ on RB $r \in \mathcal{R}$, γ_i^{th} is the SINR threshold to use an MCS $i \in I$. If the SINR is lower than the threshold, then using MCS i will increase the decoding error. $\sigma^{b,u}$ is the channel noise for user $u \in \mathcal{U}_b$, $R_{s,i}$ is the throughput of a transmission when the data are transmitted over s number of RBs using an MCS index $i \in \mathcal{I}$, and $t_{s,i,c}$ is 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 is equal to 2ms [5]. Not respecting this deadline will lead to retransmission of data; thus, wasting the initial transmission. Additionally, we suppose that users have different QoS requirements; each user requests a minimum throughput $R_{min}^{b,u}$ that has to be satisfied.

We use the model from [6] modeled using Open Air Interface (OAI) RAN simulator to determine how much time is needed to process each user's data. This model provides the required processing time, $t_{s,i,c}$, of a user's data as a function of the total number of used RBs, the MCS index, and the CPU frequency. The formula is given by:

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

The parameters of this model are the total number of RBs s, the used MCS index i, and the working frequency of CPU, f_c. Based on experimental studies, [6] 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 [4] to determine the Transport Block Size (TBS), which is the amount 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, as the latter just 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.

B. MILP Problem

We formulate a Mixed Integer Linear Programming Model (MILP), which minimizes the total power 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 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 this user's data are processed on CPU $c \in \mathcal{C}$. 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. 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 and is used to enforce the conditions explained below. The formulated MILP optimization problem is:

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

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

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

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

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

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

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

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

$$\sum_{r\in\mathcal{R}} p_r^{b,u} \leq P_{Tx}^{max}, \forall b\in\mathcal{B}, u\in\mathcal{U}_b \tag{10}$$

$$p_r^{b,u} \leq M\sum_{i\in\mathcal{I}} x_{r,i}^{b,u}, \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R} \tag{11}$$

$$x_{r,i}^{b,u} \leq \beta_i^{b,u},\ \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{12}$$

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

$$g_r^{b,u,b} p_r^{b,u} \geq \gamma_i^{th}(\sigma^{b,u} + \sum_{b'\in\mathcal{B}-\{b\}}\sum_{u'\in\mathcal{U}_{b'}} g_r^{b',u',b} p_r^{b',u'}) - M z_{r,i}^{b,u}, \quad \forall b\in\mathcal{B}, u\in\mathcal{U}_b, r\in\mathcal{R}, i\in\mathcal{I} \tag{14}$$

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

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

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

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

The objective function in (2) minimizes the total power 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 users belonging to one base station cannot use the same RB and ensures 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 a 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 $z_{r,i}^{b,u}$. To find the processing time and throughput for a user, it is necessary to know the total number of used resources blocks by a user [4], [6]; this is done by (16) and (17). Finally, (18) makes sure that each CPU can process the data assigned to it without violating the deadline constraint.

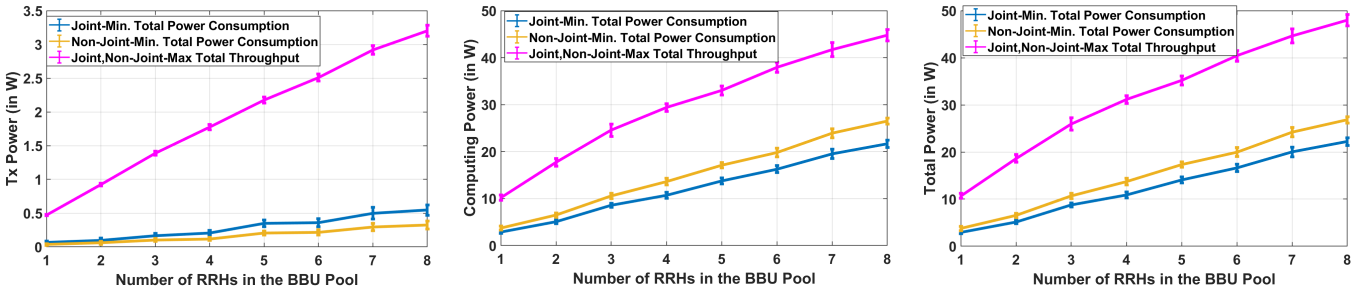
On the other hand, to understand how different objectives can affect the benefit of joint allocation, we consider a modified optimization problem that maximizes the total throughput but with the same constraints as before. The objective function becomes as follows:

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

IV. Simulation Settings

A. Simulation Environment

Our study considers an area with a variable number of Base Stations ranging from 1 to 8. Each base station is separated from neighboring base stations with a minimal distance. This distance follows a uniform distribution between 0.75km and 1.5km. Each base station serves two users. The position of each user follows a Poisson Point Process (PPP) such that each user is located in a disk of radius 300m and centered at the base station. Each user demands a throughput that follows a uniform distribution between 0.25 and 4 Mbps, and the total demand of the users from the same base station does not exceed 4 Mbps. Each base station has 24 RBs for transmission, where the frequency reuse is equal to 1. To find the SINR threshold of the MCS indexes γ_i^{th}, we used the tables in [13], 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 [4] 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 [14]. Based on it, P_{Tx}^{max} should be equal to 23 dBm with tolerance of +/- 2dBm. Hence we fix P_{Tx}^{max} = 250mW $\approx$ 24dBm. We suppose that the noise spectral density is 110 dBm/Hz, and the Noise Figure is 8dB. To model the channel gain, we consider the ABG model in [15] 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 unit mean. Considering a Cloud-RAN architecture, the baseband processing of these base stations is hosted in a shared BBU pool. We consider just one CPU core with power consumption



(a) Tx Power (in W) as a function of N of RRHs in the BBU pool (b) Computing Power (in W) as a function of N of RRHs in the BBU pool (c) Total Power (in W) as a function of N of RRHs in the BBU pool

Fig. 1. Power Consumption

P^c_{comp} = 30W 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^c_{comp}. In contrast, we suppose that 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 that the computing resources are sufficient to process the data of all users.

B. Performance Metrics

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

- *Transmission Power Consumption;* the total transmission power of all users in the system.
- *Computing Power Consumption;* the total computing power consumption of all users in the system.
- *Total Power Consumption;* the sum of the total transmission and computing power consumption in the system.
- *Throughput;* the sum of throughput of all users.
- *CPU Idle time*; the ratio for which the CPU is idle

In addition, to analyze the behavior of the joint resource allocation model vs. the non-joint, we monitor the percentage of utilized RBs in each base station and the selection of the MCS indexes.

C. Simulation Tools and Procedures

To code and run the simulation, we use MATLAB. The Matlab code calls GUROBI Optimizer to solve the MILP problem. We acknowledge the fact that solving an MILP problem is an NP-Hard problem, and it is not possible to use it in a real setting where allocation decisions have to be made every 1 ms; however, we recall that our goal is to probe the potential gains of optimal joint allocation. Finding efficient, implementable, real-time, and low-complexity joint-allocation algorithms will be left for future work. In the next section, we plot and analyze the performance of the joint allocation vs. the non-joint, considering the respective objectives of minimizing the total power consumption and maximizing the total throughput. The performance is measured as function of the number of RRHs connected and managed by the same BBU pool (i.e., number of base stations managed by the same BBU pool). The non-joint model separates the allocation of radio resources from the allocation of computing resources to solve the two problems sequentially. In the case of power minimization, the radio allocation should minimize the radio transmission power. Then the computing resources allocation should minimize the computing power. The simulation is repeated 25 times, and the 95% confidence intervals are plotted.

V. Results

We plot the graphs of the joint radio and computing resources allocation performance with respect to each of the following metrics.

A. Transmission and Computing Power Consumption

Figures 1(a), 1(b), and 1(c) show the performance of the joint allocation problems vs. the non-joint concerning the transmission power consumption, computing power consumption, and the total power consumption, respectively. Considering the objective of minimizing total power consumption, the joint radio and computing resources allocation consumes more radio transmission power but less computing power than the non-joint allocation. Hence, the joint algorithm consumes less total power. When the BBU pool has just one active base station, the joint algorithm reduces the total power consumption by 21.2% compared to the non-joint counterpart. When the number of base stations connected to the BBU pool increases to 8, the improvement falls to 17.15%. This results from increasing the number of users and the total demand in the system; this will be explained later in this section. On the other hand, adopting maximizing-throughput objective leads to the exact behavior for joint and non-joint variants. As long as the computational resources are sufficient, as [5] shows, joint and non-joint allocation of radio and computing resources produce the same results when the objective is total throughput maximization. Therefore, the algorithm will use all the available radio and computing power to maximize the throughput. Hence, maximizing-throughput would consume up to 260% more total power than the joint allocation that aims to minimize power consumption would do.

B. Throughput

The performance concerning the throughput metric as a function of the number of base stations (RRHs) in the BBU pool is plotted in Fig. 2. Since total power minimization must guarantee the demanded throughput for every user, both the joint and non-joint variants achieve similar results. 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.

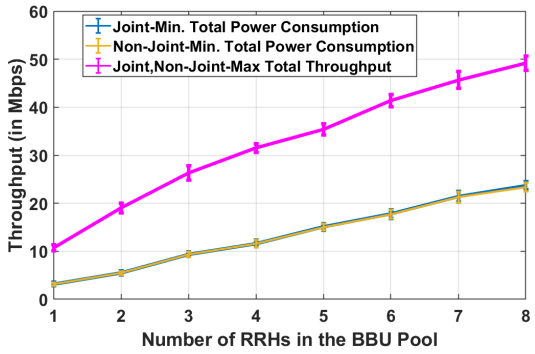


Fig. 2. Throughput as a function of N of RRHs in the BBU pool

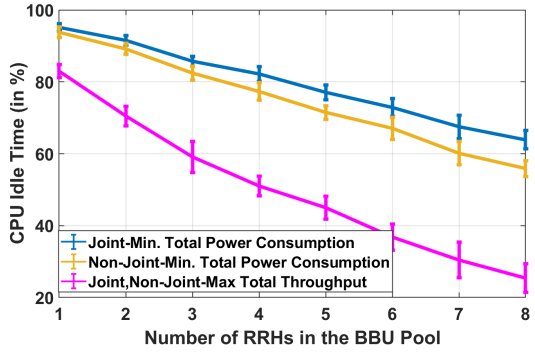


Fig. 3. CPU Idle time as a function of N of RRHs in the BBU pool

C. CPU Idle Time

Fig. 3 shows the percentage of CPU idle time as a function of the number of RRHs connected to the BBU pool. Using the computing model described in section III-A, minimizing the power consumption is consistent with reducing the CPU utilization, or in other words, increasing the CPU idle time. This explains why power minimization with joint allocation, which can best minimize the computing power consumption, achieves higher CPU idle time than the non-joint variant. Again, the joint and non-joint variants of maximizing-throughput objectives have much lower CPU idle time than power minimization algorithms. This is interpreted by the fact that maximizing-throughput algorithms aim at exploiting all the computing resources as much as possible to maximize the throughput.

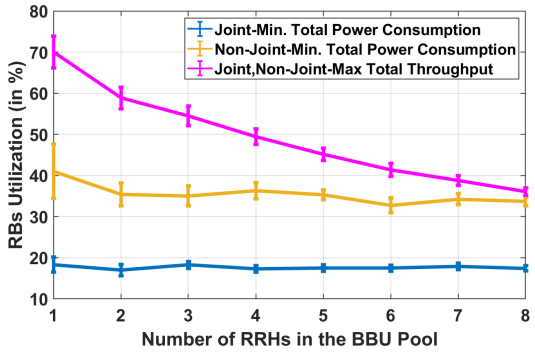


Fig. 4. RBs Utilization as a function of N of RRHs in the BBU pool

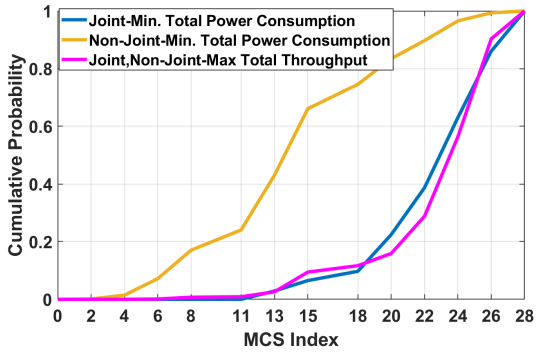
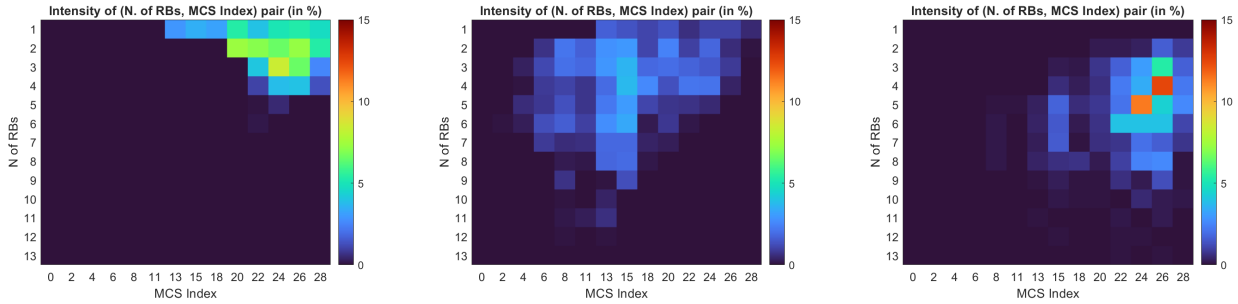


Fig. 5. Cumulative distribution function of MCS assignment

D. MCS and RB Selection

To understand the behavior of the different algorithms, we observe the decisions of the algorithms on MCS indexes assignment and on the number of RBs assigned to users. Fig. 4 shows the percentage of utilized RBs in each base station as a function of the total number of RRHs (Base Stations) connected to the BBU pool, and Fig. 5 shows the cumulative distribution probability of selecting an MCS index for each algorithm. Considering the power minimization objective and analyzing both figures, the joint algorithm tends to assign fewer number of RBs but higher MCS indexes to users as much as possible. Achieving the same throughput could be done by either using a lower number of RBs and a high MCS index, if the SINR is good, or using a higher number of RBs with a lower MCS index. The joint allocation variant of power minimization would go for increasing the MCS index. This requires increasing the transmission power, but overall would decrease the required computing resources, as the computing model in section III-A shows. As a result, the computing power consumption decreases, and thus the total combined power consumption decreases. In contrast, the non-joint sequential allocation firstly solves the radio allocation that minimizes transmission power independently of the computing resources allocation. While the results of the radio allocation (i.e., the MCS index and the number of resource blocks) affect the computing resources requirements, the computing resources allocation has to satisfy these requirements, without modifying any of the radio decisions such as the MCS index or the number of RBs. In general, this leaves a tighter room to minimize computing power consumption. It could only be possible to adjust the CPU frequency and select which CPU to process the data in case of non-homogeneous CPU power consumption. In short, the non-joint sequential algorithm minimizes the transmission power and tends to spread the data over a higher number of RBs but with a lower MCS index. On the other hand, maximizing-throughput algorithm tends to use all the RBs in each base station as well as the maximum transmission power for every user. This justifies the very high RB utilization and the usage of high MCS indexes, as Fig. 4 and Fig. 5 show. However, maximizing-throughput algorithm should make sure its selections do not increase the interference worsening performance.

Fig. 6 further supports this previous explanation. The heat maps show the intensity of assigning the pair composed of 1) the number of resources blocks, 2) MCS indexes to



(a) Joint - Min. Power (b) Non-Joint - Min. Power (c) Joint,Non-Joint - Max. Throughput

Fig. 6. Intensity of assigning a pair of (number of Resource Blocks, MCS index) to users

users. While the Joint-Power Minimization algorithm intensely allocates a lower number of RBs to users and a very high MCS index, the Non-Joint-Power Minimization favors assigning a higher number of RBs but lower MCS indexes. On the other hand, maximizing-throughput algorithm assigns more RBs and high MCS indexes to users.

As a final note, we have reduced the size of the 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. However, we can generalize our conclusions to a higher-dimensional problem. Moreover, the results of maximizing-throughput objective help us understand the performance, in case the minimum throughput requirement for users is high to the degree that all the radio resources are required to satisfy these demands. Suppose that the QoS requirements (i.e., minimum throughput) of users are increased, so that the throughput maximization objective can satisfy 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. (9)). On the other hand, since power-minimization objective should satisfy the minimum requirement constraint, it will give the same results as throughput maximization objective; all the radio resources are needed and it is not possible to use less to save power. In such a case, the joint and non-joint allocation will perform the same, even if the goal is power-minimization, as long as the bottleneck happens at the level of the radio resources. In case the computing resources are scarce and the radio resources are sufficient, (9) must be relaxed. In such case, the joint allocation should perform better than the non-joint as [5] shows.

VI. 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 non-joint sequential model, considering the objectives of minimizing power consumption and throughput, respectively. The results demonstrate that when the computing resources are sufficient, the joint allocation is beneficial and achieves performance gains by reducing the total power consumption when the objective is power minimization. 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 a real implementation, we aim to improve our study in the future by proposing low-complexity sub-optimal algorithms that can achieve the benefits of the joint allocation, while outputting results in real-time.

References

[1] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," *IEEE Access*, vol. 7, pp. 70 371–70 421, 2019.

[2] A. Gupta and R. K. Jha, "A survey of 5g network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[3] C. Mobile, "C-RAN: the road towards green RAN," *White Paper, ver*, vol. 2, pp. 1–10, 2011.

[4] (2018, October) 5G; NR; Physical layer procedures for data, ETSI TS 138 214 V15.3.0.

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

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

[7] Y. Li, H. Xia, J. Shi, and 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)*.

[8] M. M. Abdelhakam and 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)*.

[9] L. Ferdouse, A. Anpalagan, and S. Erkucuk, "Joint Communication and Computing Resource Allocation in 5G Cloud Radio Access Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, 2019.

[10] M. Sharara, S. Hoteit, and 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)*.

[11] D. Bega, A. Banchs, M. Gramaglia, X. Costa-Pérez, and P. Rost, "CARES: Computation-Aware Scheduling in Virtualized Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, 2018.

[12] S. Bian, J. Song, M. Sheng, Z. Shao, J. He, Y. Zhang, Y. Li, and 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)*.

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

[14] (2018, July) 5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone.

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