

A dynamic resource allocation framework in LTE downlink for Cloud-Radio Access Network

Mohammed Yazid Lyazidi, Nadjib Aitsaadi, Rami Langar

▶ **To cite this version:**

Mohammed Yazid Lyazidi, Nadjib Aitsaadi, Rami Langar. A dynamic resource allocation framework in LTE downlink for Cloud-Radio Access Network. Computer Networks, 2018, 140, pp.101-111. 10.1016/j.comnet.2018.05.008 . hal-01842110

HAL Id: hal-01842110

https://hal.sorbonne-universite.fr/hal-01842110

Submitted on 15 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Dynamic Resource Allocation Framework in LTE Downlink for Cloud-Radio Access Network

Mohammed Yazid Lyazidi, Nadjib Aitsaadi, and Rami Langar

***Abstract*—One main asset of Cloud-Radio Access Network (C-RAN) lies in its centralized architecture that allows network operators to serve dynamic flows of mobile traffic with efficient utilization of baseband resources and lesser operation costs than the distributed RAN architecture. For this very reason, the implementation of online resource allocation algorithms in the BaseBand Unit (BBU) pool for handling loads of multiple Remote Radio Heads (RRHs) is one of the most motivating challenges in C-RAN. Those centralized algorithms must be able to handle efficiently interference between users, as well as to dynamically select RRHs that can be turned on/off based on traffic variation. By doing so, the total RRHs transmission power can be minimized and the number of active BBUs within the cloud can also be reduced. In this paper, the issues of dynamic wireless resource allocation, transmission power minimization and BBU-RRH assignment in downlink C-RAN are addressed in one framework. We have previously attempted to address these problems by proposing a approach based on the branch-and-cut algorithm to solve small instances of the problem to optimality. However, due to the combinatorial complexity of the problem, finding optimal solutions for a large-scale network may take a fair amount of time and will not be suitable for online optimization. Towards this end, we propose a novel two-stage approach to address these issues for a large-scale problem. The first stage is a new proposal that addresses the problems of dynamic resource allocation and power minimization in C-RAN using a simulated annealing approach with a specific neighborhood search program. The BBU-RRH assignment is handled in the second stage using a multiple knapsack formulation. Through extensive event-based simulations, our proposal achieves significant reduction in time complexity and yields near optimal performance compared to state-of-the-art methods.**

***Index Terms*—Cloud-RAN; LTE; Resource allocation; Power minimization; BBU-RRH assignment; Simulated annealing.**

I. Introduction

Cloud Radio Access Network (C-RAN) has been recently introduced by China Mobile Research Institute as a novel cloud architecture for Long Term Evolution (LTE) and upcoming cellular standards (5G) [2]. It is a new RAN paradigm that can address the challenges mobile network operators are faced with and meet their requirements in terms of capital and operational expenditure costs reduction. The C-RAN architecture is illustrated in Fig. 1. It is based on a central cloud pool composed of BaseBand Units (BBUs) that perform Physical (PHY) and Medium Access (MAC) functions processing. The BBUs are connected to the Remote Radio Heads (RRHs) in the cell sites by means of a low-latency and high bandwidth fronthaul network. A cloud controller is situated in the BBU pool and performs resource and load balancing between BBUs that are interconnected through a high-speed backhauling network [3]. By replacing "hard" wireless network equipments by "soft" BBUs, the C-RAN capabilities can be dynamically adjusted based on the traffic load variations [4]. This not only fosters efficient resource utilization, but also allows the C-RAN to handle more areas than standalone clusters of base stations and facilitates service deployment on the edge [5].

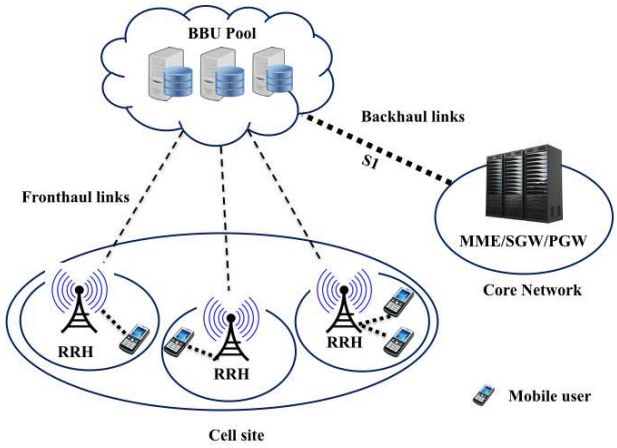


Fig. 1: Cloud Radio Access Network (C-RAN) Architecture

However, the design of dynamic schemes for C-RAN's radio resource management constitutes a major challenge that hinders its commercial expansion. In fact, the optimization of C-RAN baseband resource allocation needs methods to cater to time-varying traffic demands at different RRHs [6]. A centralized algorithm can help optimize the resource demands of mobile users located in different cells and with different bandwidth requests. Besides, such centralized approach will help network operators select the RRHs that can be dynamically turned on/off, based on their traffic loads patterns during the day. Consequently, the total RRHs transmission power can be minimized and the baseband resources can be efficiently utilized for handling traffic demands during the day. Moreover, lessening the number of active RRHs would help reduce the number of instantiated BBUs associated to them in the cloud and realize more power and cost savings. Therefore, for all these reasons, a careful C-RAN resource

M. Y. Lyazidi is with the LIP6, University of Pierre and Marie Curie (UPMC), Paris 75005, France (E-mail: yazid.lyazidi@lip6.fr).

N. Aitsaadi is with University of Paris-Est, LIGM-CNRS UMR 8049, ESIEE Paris, 93162 Noisy-le-Grand, France (E-mail: nadjib.aitsaadi@esiee.fr)

R. Langar is with University of Paris-Est, LIGM-CNRS UMR 8049, University Paris Est Marne-la-Vallée (UPEM), 77454 Marne-la-Vallée, France (E-mail: Rami.Langar@u-pem.fr).

A preliminary version of this paper appeared in the proceedings of the 2016 IEEE International Conference on Communications (ICC 2016) [1].

allocation strategy must be planned regarding users traffic demands, RRHs transmission power minimization and BBU pool capacity in terms of handled RRHs.

In [1], we presented two optimization models for the i) resource allocation and power minimization problem and ii) the BBU-RRH assignment problem in C-RAN. The proposed scheme based on the branch-and-cut algorithm [7] has permitted to achieve reasonable gain in throughput satisfaction rate and transmission power minimization over state-of-the-art algorithms and for small instances of the problem. However, due to the combinatorial nature of the first problem (NP-hard), the computational complexity is exponential if an exact optimal solution is to be calculated for a large-scale system. In this paper, a meta-heuristic algorithm, known as Simulated Annealing (SA), is used in providing fast and close-to-optimal solutions to the first-stage problem at a much reduced complexity. The near-optimality gap will be emphasized by comparison to solving the problem to optimality by the offline branch-and-cut algorithm used in [1].

In summary, our key contributions are the following:

- We express in the first stage the Centralized Resource Allocation and Power Minimization (C-RAPM) problem, which is formulated as an Mixed Integer Linear Programming (MILP) problem. A reformulation is proposed using the framework of the well-known big-*M* method [8]. A novelty in this paper compared to our previous approach is we consider here a power allocation model based on static transmission instead of continuous.
- We formulate in the second stage the BBU-RRH assignment problem as a Multiple Knapsack Problem (MKP). The latter can efficiently be solved by standard solvers such as IBM CPLEX [9].
- We present our new Dynamic Resource Allocation in C-RAN framework based on SA (DRAC-SA) to solve the C-RAPM problem with dynamic constraints.
- We compare our approach's results from event-based simulations to our previous approach DRAC in [1] and to different literature schemes. We also discuss the associated performance gains.

The remainder of the paper is organized as follows: Section II presents a review of related works regarding resource allocation, power minimization and BBU-RRH management in C-RAN. In Section III, we describe the two-stage system model for the C-RAPM and MKP problems, which is followed by Section IV that details our proposed SA approach. Discussion and analysis of simulation results are exposed in Section V. Finally, Section VI concludes the paper.

II. Related Work

C-RAN has received a considerable amount of research attention after its introduction by China Mobile Institue. Authors in [4] highlighted C-RAN's advantages for operators and vendors compared to distributed RAN. In fact, traditional base stations are often under-utilized during certain hours of the day, which results in wasteful use of radio resources and baseband capacity. The authors showcased C-RAN's ability to handle this issue by dynamically instantiating BBUs and allocating the baseband resources to RRHs depending on traffic volumes [10]. Furthermore, authors in [11] introduced the concept of coupling C-RAN with mobile cloud computing systems to enhance end-to-end cloud services for future 5G networks. In their work, the authors proposed a novel topology framework and rate-allocation configuration in C-RAN to improve end-to-end traffic performance of mobile cloud computing users.

Regarding the transmission power minimization issue, authors in [12] described a Group Sparse-based Beamforming approach (GSB), that can minimize the C-RAN RRHs transmission and fronthaul links power consumption in downlink. The authors outlined the problem as a joint RRH selection and transmit plus fronthaul links power minimization problem, with a Signal-to-Interference-plus-Noise Ratio (SINR) constraint at each user. Their proposed GSB algorithm solves the problem by starting to sort all RRHs following their transmitting power gains. The algorithm then iteratively turns off RRHs with minimum power gain, until the power minimization problem becomes infeasible. However, the GSB approach was not a C-RAN-specific solution for power minimization, since it can also be applied to traditional base station networks, with an extension of fronthaul links. Furthermore, the GSB scheme could not measure the number of necessary BBUs in the cloud that can handle the system.

Our paper, to the best of our knowledge, is one of the precursory attempts to present a high-level centralized approach combining dynamic resource allocation, transmission power minimization and BBU-RRH assignment in one framework. Other attempts regarding centralized resource allocation have been previously tackled under rate constraint such as [13] [14][15]. Authors in [13] presented a QoS-based Power Control and Resource Allocation in LTE Femtocell network (QP-FCRA). Although their approach is mainly within the context of femtocell networks, it can be applied to C-RAN thanks to its centralization nature. In their proposal, a joint resource allocation and power minimization algorithm is implemented at a central level of each clustering cells. The QP-FCRA algorithm then exploits cooperation between neighboring RRHs to periodically optimize the throughput satisfaction rates of users. However, their optimization scheme was run in offline mode and the algorithm's computational time was fairly big. In [14], we have addressed the problem of admission control considering individual UEs Quality of Service (QoS) requirement for guaranteed-service users but the transmission power aspect was however not considered. In this paper we encompass jointly maximizing the throughput of best-effort users while minimizing the total transmission power.

Although solutions for resource BBU-RRH assignment procedures in C-RAN have received some notable attention, the number of contributions for this problematic remains nonetheless very limited. Authors in [16] described a Colony RAN design that can lessen the number of BBUs by roughly 75% compared to distributed RAN. In [17], the same authors carried out their Colony RAN framework by proposing two mapping schemes for BBU-RRH assignment: Semi-Static (SS) and adaptive. The SS approach fixes the dichotomies of BBU-RRH subject to traffic peak hours of all network's RRHs in a

large time window (one day). On the other hand, the Adaptive scheme dynamically maps BBUs with RRHs based on BBUs resource capacity and neighboring RRHs loads within a short time interval (one hour). For a given business office area traffic distribution, the authors demonstrated that the SS and Adaptive schemes can help reduce the number of BBUs by 26% and 47%, respectively.

III. Problem Formulation

We detail in this section our two optimization models: C-RAPM and MKP. We consider a C-RAN system composed by a number of S RRHs within the set $\mathcal{S} = \{i | 1 \leqslant i \leqslant S\}$. The BBU pool jointly assigns to each RRH in $\mathcal{S}$ a number of K Physical Resource Blocks (PRBs) from the set $\mathcal{K} = \{k | 1 \leq k \leq K\}$. We assume that the fronthaul network has sufficient links capacity.

A. Centralized Resource Allocation and Power Minimization (C-RAPM) Problem Formulation

In our first optimization model, we consider N ($N \geq 1$) number of User Equipments (UEs) entering the system at a given epoch and connecting to a certain RRH from $\mathcal{S}$. Each UE $u \in \{1, .., N\}$ requests from its serving RRH a number of PRBs N_u to run its applications [18]. We suppose that each RRH i handles one cell in a delimited area, and that a UE u can only be served by the RRH covering the area it is positioned within. We consider a static transmission power from RRH i to UE u on each allocated PRB k. We suppose that the transmission power is quantized into $L \geq 2$ discrete power levels: $p_{min} = p_1 < p_2 < ... < p_L = p_{max}$, where p_{min} is the minimum power that can be transmitted to a UE u and p_{max} is the maximum transmitted power for each RRH. An increase in the number of power levels L pushes the discrete domain to be closer to a continuous one, but undoubtedly increases the problem's computational complexity [19]. Each resolution can lead to different transmission powers. We define our UE-RRH attachment, PRB allocation and transmit power variables:

$$x_i^u = \begin{cases} 1, & \text{if UE } u \text{ is attached to RRH } i, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$y_{ik}^u = \begin{cases} 1, & \text{if PRB } k \text{ is allocated to UE } u \text{ on RRH } i, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

$$p_{ik}^u = \begin{cases} p \in \{p_1, ..., p_L\}, & \text{if } y_{ik}^u = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The SINR achieved by UE u, attached to RRH i and on a given PRB k can be formulated as:

$$\gamma_{ik}^u = \frac{p_{ik}^u g_{ik}^u}{\sum_{j \neq i} \sum_{v \neq u} p_{jk}^v g_{jk}^u + \sigma^2} \quad (4)$$

Where g_{ik}^u is the path gain between RRH i and UE u, and σ^2 is the noise power. The SINR is expressed per PRB, as both channel/fading and interference vary over PRBs due to multi-path, frequency selectivity and domain scheduling [20]. Our objective in this first stage is to find the best PRBs allocation to serve in a best effort way all existing UEs, while minimizing the total downlink RRHs transmission power. The C-RAPM optimization problem can be expressed as follows:

$$\underset{x^u, y^u, p^u}{\text{minimize}} \quad \sum_{u=1}^{N} \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} (\epsilon \frac{p_{ik}^u}{P_{max}} - (1 - \epsilon) \frac{x_i^u y_{ik}^u}{K}) \quad (5)$$

$$\text{subject to} \quad \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} x_i^u y_{ik}^u \leqslant N_u, \ \forall u \quad (6)$$

$$\sum_{i \in \mathcal{S}} x_i^u \leqslant 1, \ \forall u \quad (7)$$

$$\sum_{u=1}^{N} \sum_{k \in \mathcal{K}} p_{ik}^u \leqslant p_{max}, \ i \in \mathcal{S} \quad (8)$$

$$\gamma_{ik}^u \geqslant y_{ik}^u \Gamma_k^u, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \quad (9)$$

$$p_{ik}^u \geqslant y_{ik}^u p_{min}, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \quad (10)$$

$$\sum_{u=1}^{N} y_{ik}^u \leqslant 1, \ i \in \mathcal{S}, k \in \mathcal{K} \quad (11)$$

$$y_{ik}^u \leqslant x_i^u, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \quad (12)$$

$$x_i^u, y_{ik}^u \in \{0, 1\}, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \quad (13)$$

We outline in the objective function (5) that we target to minimize the total transmission power while maximizing all possible UEs-PRBs assignments. The objective function is standardized so as to return values in the same order of magnitude. ϵ is a constant optimization weight between 0 and 1. Constraint (6) imposes that the total number of allocated resources for UE u cannot surpass its original demand N_u. Constraint (7) denotes that a UE can only be served by at most one RRH. This is already guaranteed by UEs' cell position, however, a decision should be made for edge UEs positioned in a coverage area overlapped with other RRHs. If so, the optimization should assign this UE to an RRH that satisfies the other problem constraints[1]. Conditions (8) and (10) are the power constraints on RRH and UE, respectively. Condition (9) ensures that the received SINR is equal to the required one Γ_k^u when the PRB k is in use (i.e., $y_{ik}^u = 1$) [13]. Constraint (11) stresses the fact that two users linked to the same RRH cannot be served with the same PRB. Constraint (12) enforces all $y_{ik}^u = 0$ if attachement variable x_i^u is equal to 0 (i.e., UE u is not receiving any PRBs from RRH i). Finally, constraint (13) refers that both y_{ik}^u and x_i^u are binary variables.

The optimization problem formulated in $(5) - (13)$ is a Mixed Integer NonLinear Programming (MINLP), which is NP-hard [21] due to the presence of the quadratic product in the objective function (5) and the non-linear SINR constraint (9). We propose next, to reformulate the problem in a Mixed Integer Linear Integer version (MILP) via the big-M method [8]. In fact, the product of two binary variables x_i^u and y_{ik}^u can be replaced with a single binary one z_{ik}^u that is defined in the following constraints:

$$z_{ik}^u \leqslant y_{ik}^u, \quad (14)$$

$$z_{ik}^u \leqslant x_i^u, \quad (15)$$

$$z_{ik}^u \geqslant x_i^u + y_{ik}^u - 1. \quad (16)$$

[1]The study of Cooperative Multipont Processing (CoMP) in C-RAN is out of the scope of this paper.

regarding constraint (9), we find it convenient to reformulate it as follows:

$$(1+\frac{1}{\Gamma_k^u})p_{ik}^u g_{ik}^u \quad \geqslant \quad y_{ik}^u \Upsilon_k^u + y_{ik}^u \sigma^2 \tag{17}$$

where Υ_k^u is equal to $\sum_j \sum_v p_{jk}^v g_{jk}^u$. Besides, the product between y_{ik}^u and Υ_k^u can be linearized using the big-M reformulation too; provided that Υ_k^u has explicit lower and upper bounds. From (3) and (10), we can deduce Lwr and $Uppr$, the respective lower and upper bounds of Υ_k^u. Hence, the binary-continuous product $y_{ik}^u \Upsilon_k^u$ can be substituted by a continuous variable w_{ik}^u and by including the following constraints:

$$y_{ik}^u Lwr \quad \leqslant w_{ik}^u \quad \leqslant y_{ik}^u Uppr \tag{18}$$

$$(1-y_{ik}^u)Lwr \quad \leqslant \Upsilon_k^u - w_{ik}^u \quad \leqslant (1-y_{ik}^u)Uppr \tag{19}$$

Hence, the ILP formulation of our C-RAPM problem can be expressed as follows:

$$\underset{x^u, y^u, p^u}{\text{minimize}} \quad \sum_{u=1}^{N} \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} \epsilon \frac{p_{ik}^u}{P_{max}} - (1-\epsilon)\frac{z_{ik}^u}{K} \tag{20}$$

$$\text{subject to} \quad \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} z_{ik}^u \leqslant N_u, \ \forall u \tag{21}$$

$$(7), (8), (10), (11), (12), (13)$$

$$(14), (15), (16) \tag{22}$$

$$(1+\frac{1}{\Gamma_k^u})p_{ik}^u g_{ik}^u \geqslant w_{ik}^u + y_{ik}^u \sigma^2 \tag{23}$$

$$(18), (19) \tag{24}$$

B. Multiple Knapsack Problem (MKP) Formulation for BBU-RRH Assignment

In a distributed RAN system, one BBU is entirely assigned to a single RRH in order to handle its total traffic load. Thanks to C-RAN's centralization and flexibility, the resources of one BBU can be shared across different RRHs that have few traffic loads [10]. For instance, if a remote site is covered by 4 RRHs and each has 25% of traffic load, one BBU is enough to manage all four RRHs. In our study, we can compute the optimal number of needed BBUs B to manage the S loaded RRHs as follows:

$$B = \lceil \frac{\textit{Sum of all RRHs traffic charges}}{K} \rceil \tag{25}$$

where $\lceil . \rceil$ is the ceiling function and K is the number of PRBs. The total charge of active RRHs corresponds to the total number of assigned PRBs from transmitting RRHs to all users, that are returned after solving the C-RAPM problem. Our goal in this second stage consists of properly assigning RRHs to BBUs according to their traffic charges and the number of available B BBUs. Towards this end, we consider a MKP formulation of the BBU-RRH assignment problem [22], where the *objects* and the *knapsacks* are represented by the RRHs and the BBUs, respectively. We introduce a new binary variable r_{ij}, which is equal to 1 if RRH i is attached to BBU j and 0 otherwise. From the results of the C-RAPM problem, we can compute the weight c_i of each RRH i as follows:

$$c_i = \sum_{k \in \mathcal{K}} y_{ik}^\star / K \tag{26}$$

where $y^\star$ is the returned solution of y from the C-RAPM problem. The value of c_i represents the percentage of traffic load RRH i handles. We suppose that each BBU j can handle 100% of a fully loaded RRH (i.e., all K PRBs are used). We then formulate our BBU-RRH MKP as follows:

$$\underset{r}{\text{maximize}} \quad \sum_{j=1}^{B} \sum_{i=1}^{S} r_{ij} \tag{27}$$

$$\text{subject to} \quad \sum_{i=1}^{S} c_i r_{ij} \leqslant 1, \ j \in \{1, ..., B\}, \tag{28}$$

$$\sum_{j=1}^{B} r_{ij} \leqslant 1, \ i \in \{1, ..., S\}, \tag{29}$$

$$r_{ij} \in \{0,1\}, \ i \in \{1, ..., S\}, j \in \{1, ..., B\} \tag{30}$$

where constraint (29) denotes that one RRH cannot be managed by more than one BBU. This formulated problem is an Integer Linear Program (ILP), which can be efficiently solved by standard ILP solvers such as CPLEX.

IV. Proposal: DRAC-SA Algorithm

In this section, we present our Dynamic Resource Allocation in C-RAN based on Simulated Annealing (DRAC-SA) meta-heuristic with defined neighborhood search to solve the C-RAPM problem formalized in $(20) - (24)$.

A. Algorithm Overview

The SA meta-heuristic [23] is a powerful stochastic algorithm used to solve many combinatorial optimization problems in a fixed amount of time. The framework is based on exploring the different states of the cooling process of a solid from an initial hot temperature to a fixed frozen one. Each state of the process corresponds to a solution of the optimization problem. From a given state, a subsequent one can be generated by performing a small perturbation mechanism. This corresponds to generating neighbors of the initial solution via some particular neighborhood structures. The acceptance rule of a new solution (or new state) to the initial one is defined by the Metropolis rule [24], which imposes a probabilistic decision based on the varying temperature and the energy of both states. The energy refers to the cost function of the optimization problem. If the generated state has lesser energy, it is accepted as the current state. Otherwise, it is admitted with a probability $\exp(\frac{-\Delta E}{T})$, where ΔE is the energy difference of the two states and T is the time varying temperature. It is worth noting that at high temperature $\exp(\frac{-\Delta E}{T})$ is close to 1, therefore the majority of moves can be accepted. Whereas at low temperature, $\exp(\frac{-\Delta E}{T})$ is close to 0, which severely limits the search process to only solutions decreasing the energy. Hereafter, we will define each function if the SA meta-heuristic to resolve C-RAPM.

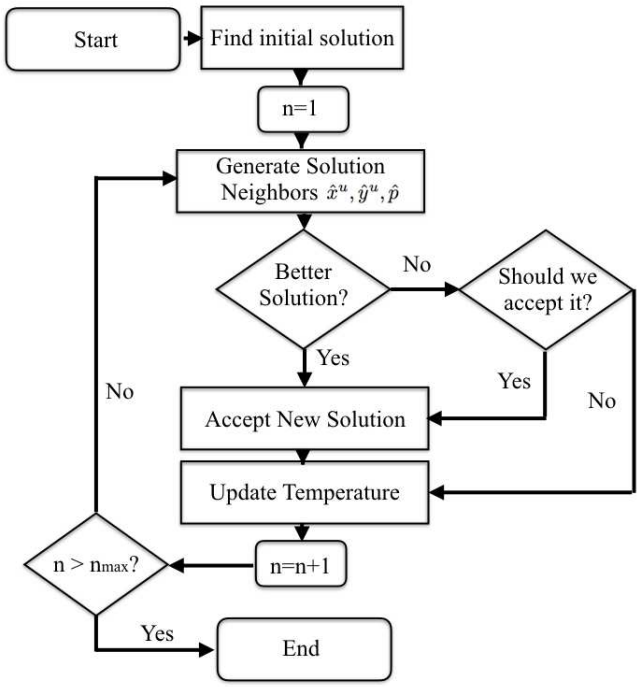


Fig. 2: DRAC-SA Flow Chart

B. Initial Solution

We first start by employing a greedy search method to generate the initial solution of the C-RAPM problem. It is based on performing linear relaxation of the integer variables and limiting the local search at the first nodes containing feasible integer solutions. Moreover, by focusing the resolution on a limited optimization space generated by fewer variables, the local search can be further reduced. In fact, we can consider z_{ik}^u as the "core" variable of our problem; since variables x_i^u and y_{ik}^u can be derived from the big-M constraints (14), (15), (16). On the other hand, p_{ik}^u comes as a "sub-core" optimization variable, which can be deduced from (8) and (10). We denote E_0 the cost function (or energy) of this initial solution and T_{max} the maximum annealing temperature. In what follows, we define TSR_u the throughput satisfaction rate of UE u, which is the ratio of its total allocated PRBs on its initial demand N_u.

C. Neighborhood Search Structure

Here, we define our specific neighborhood search stage to generate the states. We initiate the neighborhood generation by selecting a uniformly random UE u from the outputs of the initial solution and by computing its TSR_u. We define $\hat{x}^u$, $\hat{y}^u$ and $\hat{p}^u$, the solution neighbors of x^u, y^u and p^u for UE u, as follows:

- *Step 1:* UE u changes its RRH attachment following a discrete Bernoulli distribution with parameter $(1 - TSR_u)$. A new RRH attachment vector $\hat{x}^u$ is generated from this probability and by selecting the available RRHs to whom u can be linked to based on its geographical position.
- *Step 2:* We keep the existing PRB allocation in the new RRH $\hat{x}_i^u$ to other UEs untouched. For the available PRBs ($y_{ik}^u = 0$), we select the eligible ones that can be allocated to UE u based on the SINR constraint $\gamma_{ik}^u \geq \Gamma_k^u$, while determining for each one the minimal power that satisfies this constraint.
- *Step 3:* For the eligible PRBs that satisfy $\gamma_{ik}^u \geq \Gamma_k^u$, they are allocated to UE u following a Bernoulli distribution with parameter $TSR_u \times \frac{\gamma_{ik}^u}{SNR_{max}}$, where $SNR_{max} = p_{max} g_{ik}^u / \sigma^2$ represents the maximum Signal-to-Noise Ratio (SNR) achieved on UE u. This helps allocating PRBs to UE u, with respect to other users existing allocation and possible interference. After this, we set all allocated PRBs power levels to a unique one, corresponding to the highest level of the allocated PRBs (i.e., the maximum of all minimal powers that satisfy the SINR constraint or each PRB).

D. Equilibrium state

After generating the new solution neighbors, a new cost function E_n is calculated. We increase the neighborhood search structure to other UEs if and only if the current solution does not improve the objective function and satisfies the following equation:

$$\exp(-\frac{E_n - E_0}{T_n}) \geq \delta \tag{31}$$

where δ is a random number in $[0, 1]$, which refers to the random value of the equation to increase the neighborhood states in the SA meta-heuristic to see whether $\exp(\frac{-\Delta E}{T})$ in equation (31) is close to 0 or 1, and thus accept increasing the neighborhood tree. Additionally, in each iteration n we use the following cooling equation to decrease the temperature:

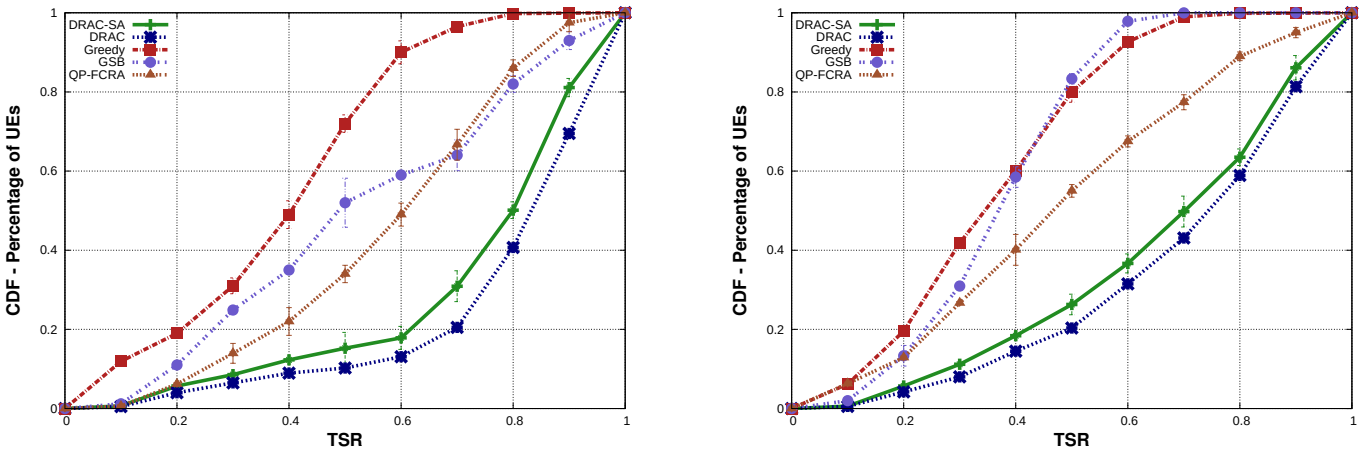
$$T_n \leftarrow \frac{T_n}{\ln(n)} \tag{32}$$

E. Stopping condition

Fig. 2 illustrates our DRAC-SA algorithm flow-chart. The algorithm converges as soon as the maximum number of iteration n_{max} is elapsed, which corresponds to the maximum CPU time. Therefore, its value should be scalable based on the processing machine so as to not exceed the delays of mobile users resources requests during their stay time in the system.

F. MKP Resolution

Once the C-RAPM problem is solved and the resources are allocated to UEs, the next step consists in calculating the number of needed BBUs B to handle the total traffic demand (25). Since the MKP problem (27)-(30) is a ILP, we make use of IBM's linear solver CPLEX to compute its solution using the solver's built-in algorithms. CPLEX's branch-and-bound is able to return optimal results within a computation time n_{MKP} very small compared to n_{max} ($n_{MKP} \ll n_{max}$). Hence, by summing the two computation times, solutions for the C-RAPM and BBU-RRH associations can be dynamically found while respecting mobile users requests delays.



(a) SINR = 10 dB

(b) SINR = 25 dB

Fig. 3: Throughput Cumulative Density Function

Table I: Simulation Parameters

Parameters	Values
Number of RRHs	100
Bandwidth	20 MHz
Total number of PRBs	100
Power levels L	6
$p_1(p_{min})/p_2/p_3/p_4/p_5/p_6(p_{max})$	0.1/1/5/10/15/20 mW
Constant ϵ	0.5
Path loss model[1]	148.1 + 37.6$\log_{10}(d)$, d in Km
Shadowing standard deviation[3]	5 dB
Fading model[3]	Normal distribution $\mathcal{N}(0, \mathbf{I})$
Thermal noise[3]	-174 dBm/Hz
Transmit antenna power gain[3]	8 dBi
Arrival rate of UEs	$\lambda \in [1, 10]$ (default 5)
Departure rate of UEs	$\mu = 0.1$
UE's PRB demand	Uniform distribution $\mathcal{U}(1, 25)$
BBU capacity[10] W	1 (100%)
Initial hot temperature	$T_{max} = 1000$
Max. number of iterations	1000

V. Performance Evaluation

In this section, we evaluate the benefits and performances of our proposed DRAC-SA algorithm, and compare the benefits of our solution with respect to state-of-the-art schemes: the QP-FCRA [13] and the Iterative GSB [12] algorithms for solving the C-RAPM problem. We also include comparisons to the greedy approach, which was used to generate the initial solution of the DRAC-SA algorithm, as well as to our previous DRAC approach in [1]. The latter was run in offline mode due to its high computation time for the chosen system parameters. On another hand, we also compare the SS and Adaptive switching algorithms in [17] to the returned solutions of our MKP regarding the BBU-RRH assignment problem. For our experimental environment, we simulated a wireless LTE environment consisting of 100 RRHs deployed in a $450\ m \times 450\ m$ square grid. Each RRH has a coverage radius of $35\ m$ and the distance between two nearest RRHs is $50\ m$. We considered the following channel model [1]: $h_i^u = 10^{-L(d_i^u)/20}\sqrt{\phi_i^u s_i^u} g_i^u$, where $L(d_i^u)$ is the path-loss at distance d_i^u between RRH i and UE u, ϕ_i^u is the antenna gain, s_i^u is the shadowing coefficient, and g_i^u is the fading coefficient. We generate a fixed poisson arrival rate of mobile users of $\lambda = 5$ arrivals per time, and vary at each simulation run the users' stay time and service demand following an exponential and uniform laws, respectively. Each UE's geographical position is randomly generated at each run and remains fixed during its stay-time in the system. The service demand of each user is expressed in terms of number of PRBs from a downlink LTE frame of 100 PRBs and follows a uniform distribution from 1 to 25 PRBs. We run 30 simulations for each scenario of SINR threshold Γ: 10 and 25 dB, to reach a confidence level of 97%. Table I reports the simulation parameters. In what follows, we present the corresponding simulation results in terms of Throughput Satisfaction Rate (TSR), computation time analysis, Spectrum Spatial Reuse (SSR), normalized throughput distribution, transmission power, and number of BBUs along with the active number of RRHs.

1) Throughput Satisfaction Rate (TSR): We present in Fig. 3 the Cumulative Distributed Function (CDF) of the TSR. The latter represents the ratio of the number of allocated PRBs to the total initial demands N_u. The CDFs of DRAC, GSB and QP-FCRA correspond to CDFs generated from offline resolutions, where we left the algorithms methods running until the end results. We emphasize the fact that they are not applicable in real-time context due to their high computational time, and we only added them for the sake of comparison. We can observe, by comparing the CDF of the offline methods and the online Greedy and DRAC-SA's ones, that for the latter, more than 50% of UEs have their TSR greater than 80% and 70% in SINR threshold equal to 10dB and 25dB, respectively. The TSR is lessened to 60% and 48% for QP-FCRA and GSB, respectively - as shown in Fig.3(a) - at low SINR threshold, and to 47% and 35%, respectively, in Fig.3(b), when the SINR threshold is high. Hence, our proposed DRAC-SA approach outperforms both QP-FCRA and GSB schemes, and approaches as well the highest throughput satisfaction rate given by DRAC, when the latter reaches the end of the resolution. However, we notice that the greedy online approach achieves better satisfaction rate at high SINR regime than the

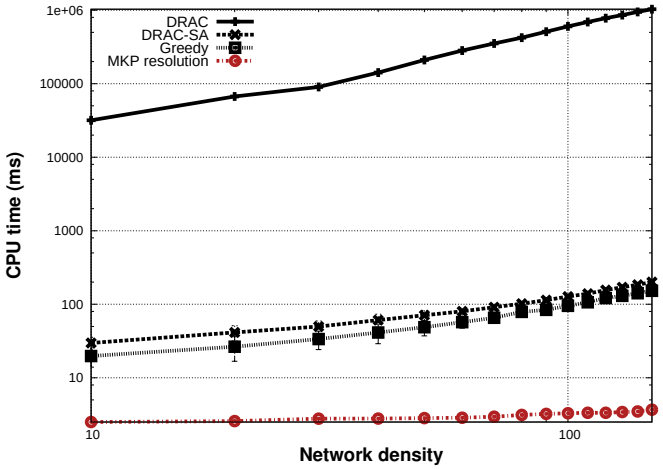


Fig. 4: CPU time vs number of UEs

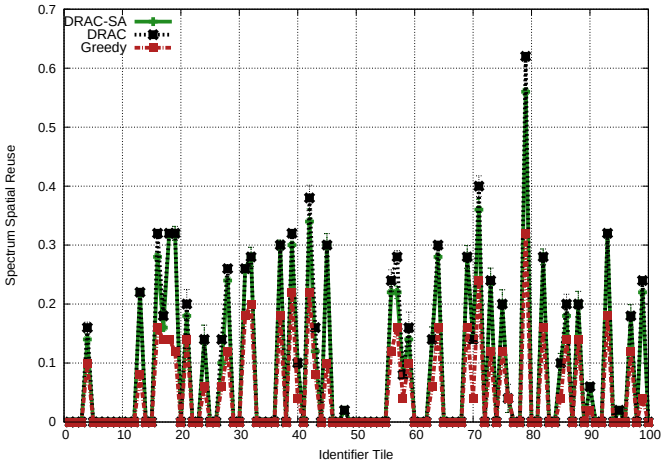


Fig. 5: SSR per PRB, SINR = 25 dB

Table II: Mean Spectrum Spatial Reuse

SINR	DRAC-SA	DRAC	Greedy	QP-FCRA	GSB
10 dB	4.29 $\pm$ 0.12	4.55	2.45 $\pm$ 0.5	4.05	2.25
25 dB	4.23 $\pm$ 0.15	4.55	2.45 $\pm$ 0.5	4.10	2.01

offline GSB scheme. In fact, the latter emphasizes on turning off as many as RRHs as possible to achieve maximum power savings, whereas the greedy approach turns a large number of RRHs on to find quick solutions for the C-RAPM problem.

2) CPU time vs network density: As shown in Fig. 4, the complexity evolution of DRAC-SA is polynomial in terms of network density, and is largely lower than that of DRAC. The time computation results indicate that the proposed DRAC-SA can solve the C-RAPM problem in less than 120 ms for a network with 100 mobile users. Besides, it can return solutions in a few ms when the number of active users is low (less than 20 mobile users). Overall, DRAC-SA achieves significantly high CPU time savings than DRAC, where the latter returns the optimum solutions after at least 1000 s for the highest dense network (150 users). This makes the DRAC approach unpractical for online optimization as it severely impacts the rate of served users and the global TSR, which will be described later on.

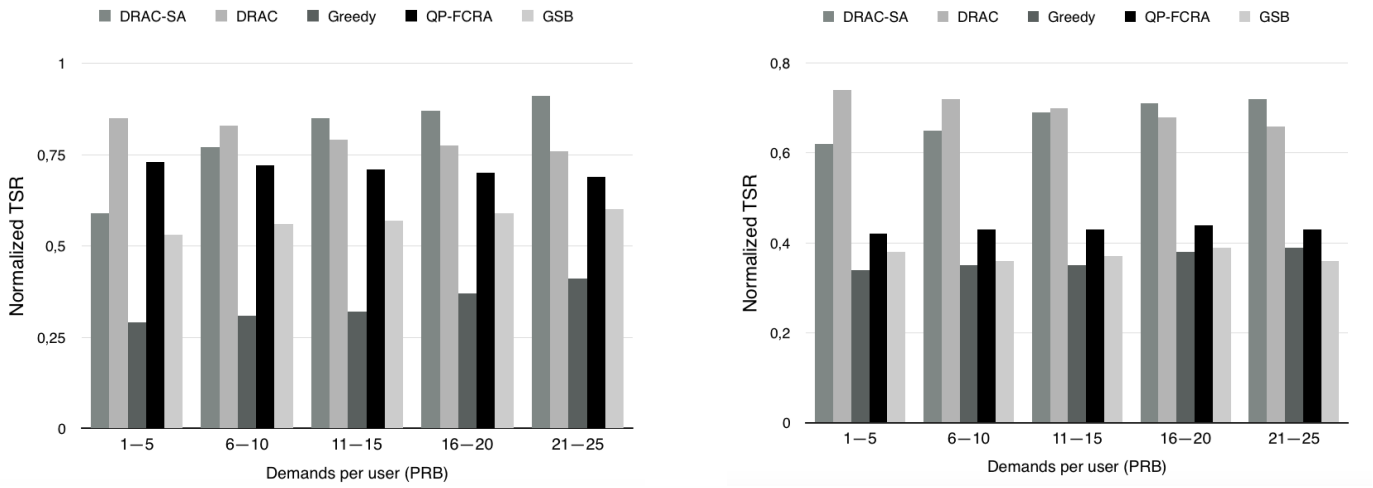
3) Spectrum Spatial Reuse (SSR): Table II reports the SSR of all aforementioned approaches. Note that SSR denotes the average portion of RRHs using identical PRB and can be expressed as follows:

$$SSR = \frac{1}{S \times K} \sum_{u=1}^{N} \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} y_{ik}^{u} \tag{33}$$

The more a PRB is reused, the better is the performance. Table II clearly shows that our proposal DRAC-SA increases PRBs reuse by a factor of 1.06 and 1.91 compared to QP-FCRA and GSB approaches, respectively, at low SINR threshold. When the SINR threshold is high, the reuse factor is enhanced by 1.03 and 2.13, respectively. We also notice that the gap between DRAC-SA's SSR and DRAC's is only of 5.71%, which exhibits the good performance of our algorithm. We further extend the analysis by investigating how each PRB is reused in the network compared to the greedy and the optimal solutions. Fig. 5 shows that DRAC-SA improves the reuse factor up to 32% for PRBs that are less re-used with the greedy method. In fact, DRAC-SA achieves globally 43.36% better performance in PRB reuse than the greedy approach.

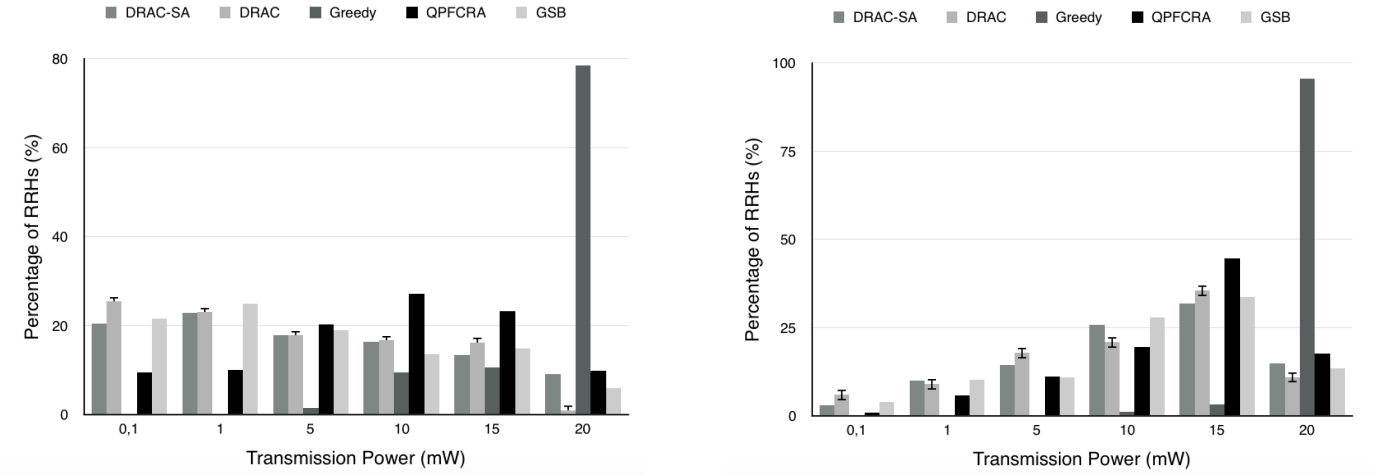
4) Normalized throughput vs UEs demand: In order to illustrate how the allocated resources are affected by UEs' demand volume, Fig. 6 presents the normalized throughput evolution as a function of UEs demands N_u, for both SINR regimes. Globally, QP-FCRA and GSB show a roughly constant behavior for SINR = 10 dB in Fig. 6(a), with an emphasis on low and high PRB demand, respectively. This implies that their resource allocation mechanisms are done independently of UEs' requested number of PRBs. On the other hand, DRAC-SA favors resource allocation of UEs with the highest demand N_u in order to increase their total satisfaction rate. This is more clearly shown in the high SINR regime (Fig. 6(b)), where DRAC-SA favors high demands significantly more than DRAC and the other schemes. This may be interpreted as unfair to users with low PRB demands. However, from a network management perspective, it is a positive behavior as DRAC-SA can dismiss resource allocation to low user demands that would cause interference to high-demanding users with greedy resource applications, and eventually increase their total transmitted power.

5) Transmitted Power per RRH: Fig. 7 illustrates the percentage of RRHs transmitting on each transmission power. We remark that at low SINR regime (see Fig. 7(a)), the majority of RRHs are transmitting on the lowest power levels: $p_{min} = 0.1\ mW$ and $p_2 = 1\ mW$, whereas for most RRHs the greedy method favors the highest power level $p_{max} = 20\ mW$, which results in a high total transmission power. What is more, DRAC-SA focuses mostly on the second power level $p_1 = 1\ mW$. At high SINR regime (see Fig. 7(b)), most approaches emphasize on higher transmission powers such as $p_4 = 15\ mW$ and $p_{max} = 20\ mW$. By scattering the transmission powers on the lowest levels, our approach can achieve less energy consumption compared to the greedy resolution method and QP-FCRA, as shown in Fig. 8, which presents the total C-RAN transmitted power. We can remark



(a) SINR = 10 dB (b) SINR = 25 dB

Fig. 6: Throughput distribution as a function of user demands



(a) SINR = 10 dB (b) SINR = 25 dB

Fig. 7: Percentage of RRHs vs transmission power levels

that the GSB scheme realizes minimum transmission power thanks to its successive RRH switching algorithm; however, this is negatively reflected on the TSR of mobile users as seen in Fig. 3, since they are less satisfied by their allocated PRBs. The QP-FCRA approach, on the other hand, supposes that all RRHs are turned on, which results in a higher power consumption but to a good TSR. As can be observed in Fig. 3 and Fig. 8, our proposed DRAC-SA scheme performs a good tradeoff balance between UEs satisfaction rates and overall transmission power minimization in both SINR threshold levels.

6) Number of BBUs and on RRHs: Fig. 9 shows the variations of the number of on RRHs and the number of instantiated BBUs B per time returned by DRAC-SA, when the SINR threshold is equal to 25 dB. As can be observed, the number of instantiated BBUs computed from the output of DRAC-SA can achieve important savings compared to a conventional RAN scenario. The latter follows a static variation due to the one-one mapping, which imposes as many BBUs as deployed RRHs. This consequently inflicts heavy investments from operators to manage their network and increase their total BBU equipment costs to handle the radio traffic loads of all cells. Our approach can realize up to 86% and 92% BBUs savings compared to a RRH-based RAN scenario, which will result in important OPEX savings for the operator.

Table III: Number of BBUs and RRHs

Scheme	Mean BBU	Mean on RRHs	Max RRHs/BBU
MKP	7.97 ± 0.06	55.4 ± 2.2	39
SS	15.1	59.7	27
Adaptive	13.5	62.4	28

For the BBU-RRH assignment problem, we solve the MKP in (26) using CPLEX, which is able to find optimal results

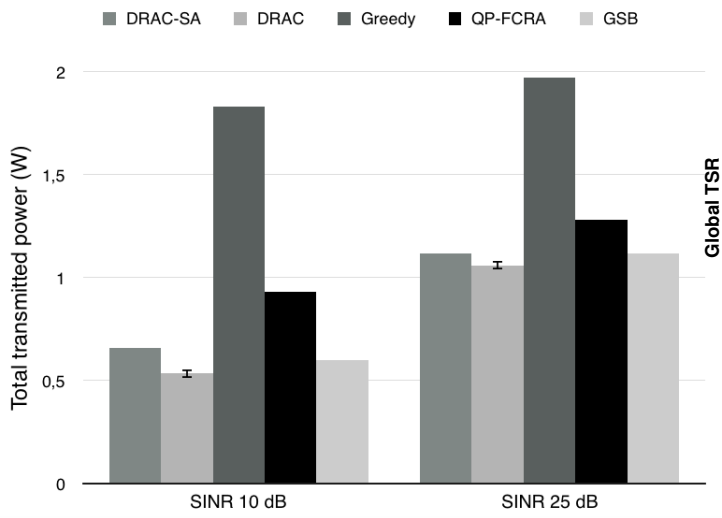


Fig. 8: Total RRHs transmitted power in C-RAN

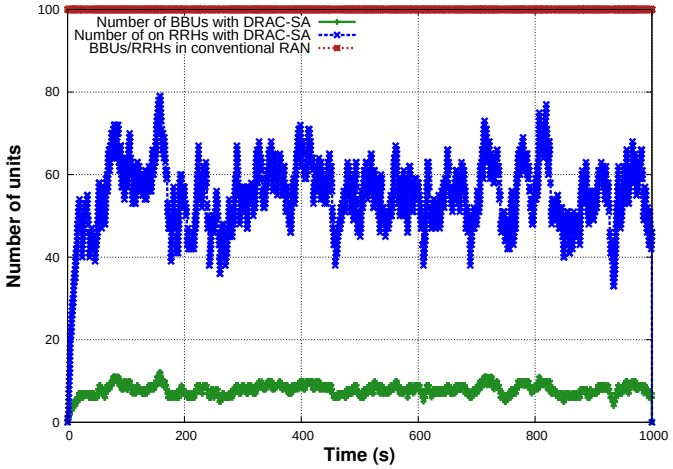


Fig. 9: Number of needed BBUs and on RRHs per time

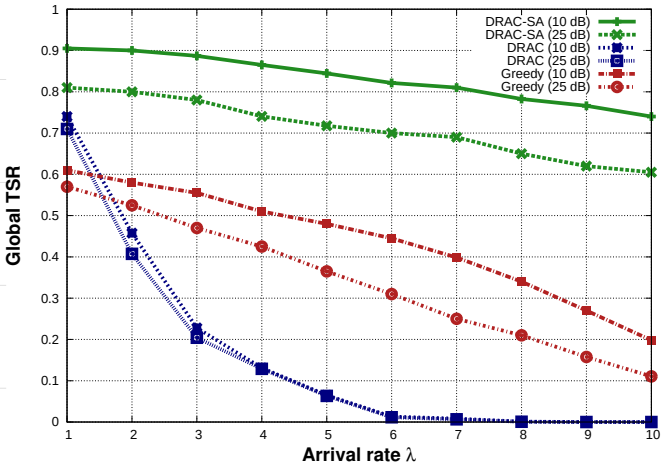


Fig. 10: Global TSR vs arrival rate λ

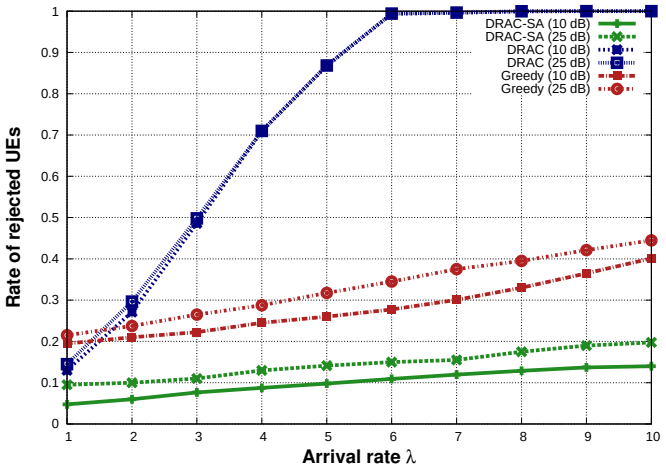


Fig. 11: Rejected users rate vs arrival rate λ

with very low computation time (at average 3 ms at each epoch). Table III presents the average number of BBUs and on RRHs as well as the minimum and maximum number of handled active RRHs per BBU. Clearly, DRAC-SA achieves more BBUs savings to cater to the same volume of traffic load with a reduced number of on RRHs. This not only improves the network capacity, since many RRHs can be handled by the same BBU, but also helps maximizing the efficiency of BBUs within the virtual pool.

7) Global TSR vs arrival rates: We vary next the arrival rates of mobile users in the system, where λ takes values in $[1, 10]$. Fig. 10 illustrates the evolution of the global TSR in both SINR regimes for each of DRAC-SA, DRAC and the greedy approaches. As λ increases, more users penetrate the system, which leads to less time intervals between each user arrival. As stated before, a large proportion of new arrived users are discarded by DRAC, since it is still solving the C-RAPM problem of the previous existing users. What is more, starting from $\lambda = 4$, the global TSR returned by DRAC is severely impacted and results in more than 70% of UEs not served by the system. This is depicted in Fig.11, which illustrates the evolution of rejected users' rate with different arrival rates. DRAC-SA, on the other hand, clearly outperforms DRAC thanks to its reduced complexity and possible online optimization, which provides a very global TSR at high arrival rate (74% and 61% for low and high SINR threshold, respectively) as well as a low rate of rejected UEs (14% and 19% for low and high SINR threshold, respectively).

8) Number of BBUs vs arrival rates: In the following, we present the variation of the number of instantiated BBUs as a function of the network density for each arrival rate. Fig. 12 presents the evolution of BBUs and the number of on RRHs in the system with the variation of the poisson arrival rate λ for SINR threshold equal to 25 dB. The plot illustrates the evolution for the DRAC-SA and Greedy methods as well as the number of BBUs required in case of the conventional RAN. The latter corresponds to the number of active RRHs, which has to be equal to the number of BBUs due to the one-one mapping in a distributed RAN deployment. As illustrated, the number of instantiated BBUs for the DRAC-SA solutions achieves important savings in BBUs compared to the conventional scenario. On another hand, we can remark

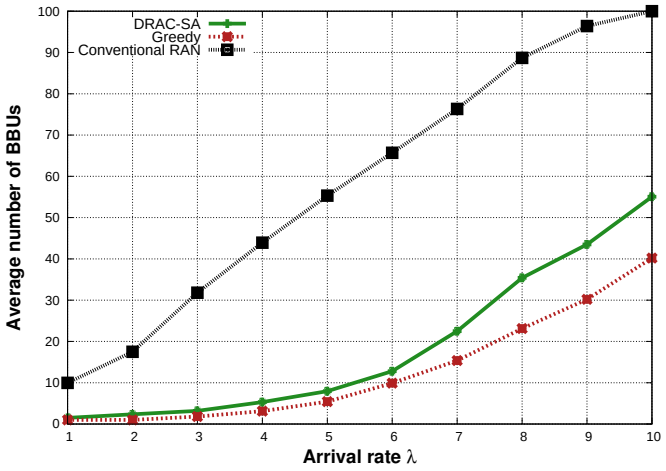


Fig. 12: Number of BBUs vs arrival rate λ

that for the highest arrival rate, $\lambda = 10$, the number of on BBUs is at its maximum capacity for the conventional case, whereas DRAC-SA and the greedy methods are still at 55% and 40% of the total system's capacity, respectively. Therefore, it is up to the operator to manage its C-RAN deployment: wether is increasing the number of RRHs to satisfy maximum users, or turning them off to achieve energy efficiency is the better choice.

VI. Conclusion

In this paper, we have portrayed a novel approach based on simulated annealing to address the problem of resource allocation and transmission power minimization in C-RAN for a dynamic flow of UEs traffic. Specifically, our newly improved DRAC-SA framework can quickly find the best PRB allocation and transmission power strategy to cater the traffic demand, while satisfying individual SINR constraints and maximum power limitations. Besides, our approach can dynamically determine the best (if not optimal) number of active RRHs and BBUs to instantiate in order handle the C-RAN traffic. Through our extensive event-based simulations, we have demonstrated that our method finds several good balances regarding, firstly, throughput satisfaction rate and total transmitted power and, secondly, resolution time and global user satisfaction. In fact, DRAC-SA achieves 43.36% better performance in PRB distribution than a greedy approach, and only 5.71% of difference is between the global optimum offline approach and DRAC-SA in terms of throughput satisfaction. Besides, the number of BBUs calculated from DRAC-SA can help increase the BBUs savings to 85.6% compared to distributed RAN scenarios. We hence believe that this approach represents a promising solution for centralized resource allocation in future C-RAN deployments.

Acknowledgment

This work is partially supported by the ANR ABCD project (Grant no. ANR-13-INFR-0005-01) and the FUI ELASTIC Network project (Grant no. C16/0287).

References

[1] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Resource Allocation in Cloud-RAN with Real-Time RRH-BBU Assignment," *IEEE International Conference on Communications (ICC)*, 2016.

[2] *C-RAN, The Road Towards Green RAN*, China Mobile Research Institute, White Paper, Version 3.0, 2013.

[3] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Schefczik, and M. Soellner, "Radio Base Stations in the Cloud," *Bell Labs Technical Journal 18(1), 129152 Alcatel-Lucent*, 2013.

[4] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[5] S. F. M. Paolini, "Benefits of C-RAN and adoption trends," *Senza Fili Consulting online webinar on C-RAN*, 2015.

[6] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 692–703, 2013.

[7] J. E. Mitchell, "Branch-and-cut algorithms for combinatorial optimization problems," *Handbook of applied optimization*, pp. 65–77, 2002.

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

[9] "IBM ILOG CPLEX Optimizer Version 1 v12.6. available: http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/."

[10] A. Checko, H. Holm, and H. Christiansen, "Optimizing small cell deployment by the use of C-RANs," *European Wireless Conference*, pp. 1–6, 2014.

[11] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (C-RAN) for mobile cloud computing systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1536–1548, 2016.

[12] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2809–2823, 2014.

[13] A. Hatoum, R. Langar, N. Aitsaadi, R. Boutaba, and G. Pujolle, "Qos-based power control and resource allocation in OFDMA femtocell networks," in *IEEE Global Communications Conference (GLOBECOM)*, 2012.

[14] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Resource allocation and admission control in OFDMA-based Cloud-RAN," *IEEE Global Communications Conference (GLOBECOM)*, 2016.

[15] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, 2015.

[16] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-RAN Architecture for Future Cellular Network," *Future Network Mobile Summit (FutureNetw)*, 2012.

[17] S. K. S. Namba, T. Warabino, "BBU-RRH Switching Schemes for Centralized Ran," *International ICST Conference on Communications and Networking in China (CHINACOM)*, 2012.

[18] F. Z. Kaddour, E. Vivier, M. Pischella, L. Mroueh, and P. Martins, "Power control in opportunistic and efficient resource block allocation algorithms for green lte uplink networks," in *IEEE Online Conference on Green Communications (OnlineGreenComm)*, 2013.

[19] H.-C. Jang and Y.-J. Lee, "Qos-constrained resource allocation scheduling for LTE network," in *International Symposium on Wireless and Pervasive Computing (ISWPC)*, 2013.

[20] R. Langar, S. Secci, R. Boutaba, and G. Pujolle, "An operations research game approach for resource and power allocation in cooperative femtocell networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 4, pp. 675–687, 2015.

[21] Y. Shi, Y. T. Hou, S. Kompella, and H. D. Sherali, "Maximizing capacity in multihop cognitive radio networks under the sinr model," *IEEE Transactions on Mobile Computing*, vol. 10, no. 7, pp. 954–967, 2011.

[22] D. Pisinger, "An exact algorithm for large multiple knapsack problems," *European Journal of Operational Research*, vol. 114, no. 3, pp. 528 – 541, 1999.

[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[24] P. J. Van Laarhoven and E. H. Aarts, "Simulated annealing," *Springer*, pp. 7–15, 1987.