Edge computing optimization for efficient RRH-BBU assignment in Cloud Radio Access Networks

Niezi Mharsi, Makhlouf Hadji

Version of Record: https://www.sciencedirect.com/science/article/pii/S1389128619302816
Manuscript_2db2bc53d8faecec44c0c7962c5ee1f8

Edge computing optimization for efficient RRH-BBU assignment in Cloud Radio Access Networks

Niezi Mharsi[a,b,*], Makhlouf Hadji[b]

[a] *Institut Mines Télécom, Télécom ParisTech, 46 Rue Barrault, 75013 Paris, France*
[b] *Technological Research Institute SystemX, 8 Avenue de la Vauve, 91120 Palaiseau, France*

Abstract

Cloud Radio Access Network (C-RAN) has been proposed as a promising architecture to overcome the challenges of next generation mobile networks (5G). The main concept of C-RAN is to decouple the BaseBand Units (BBU) and the Remote Radio Heads (RRH), and place the BBUs in common edge data centers (or BBU pools) for centralized processing. The optimal assignment of RRHs (or antennas) to edge data centers when jointly optimizing the fronthaul latency and resource consumption is one of the key issues in the deployment of C-RAN. This problem is NP-Hard and network operators need new assignment algorithms that can scale with large problem sizes and find good solutions in acceptable times. In this paper, we first model our constrained resource allocation problem by an exact approach based on Integer Linear Programming (ILP) formulation. Then, and for sake of scalability, we propose new heuristic algorithms with reduced complexity to rapidly achieve optimal (or near-optimal) solutions for the assignment of antennas demands to the available edge data centers. Simulation results highlight the efficiency and scalability of our proposed approximation algorithms and their ability to provide good solutions in negligible times.

Keywords: Edge computing, C-RAN, 5G, Optimization, Resource allocation

*Corresponding author
Email address: {niezi.mharsi, makhlouf.hadji}@irt-systemx.fr (Niezi Mharsi)

1. Introduction and motivation

To cope with dynamic and insatiable end-users demands in the telecommunications domain, Telecommunications Service Providers (TSPs) are investigating new solutions to reduce CAPEX (CAPital EXpenses) and OPEX (OPerating EXpenses) of their new infrastructures often based on Network Functions Virtualization (NFV) paradigm [1]. In fact, by sharing their NFV-Infrastructures (NFVI), these players can significantly reduce their operational costs and hence maximize their profit when satisfying larger number of end-users and subscribers of their new virtualized services.

The virtualization and cloudification of Radio Access Networks (RAN) have been identified as a good candidate to efficiently optimize the network deployment costs, e.g. CAPEX and OPEX, of these actors. In this context, C-RAN has been proposed as a promising network architecture for next generation mobile networks, that combines NFV and cloud computing concepts to enhance the network utilization efficiency and achieve cost savings. In fact, unlike conventional networks where the baseband functions reside on the cell sites along with the antennas, C-RAN decouples the traditional base station into RRHs and centralized BBUs that are pooled in common locations called BBU pools and used as shared resources between multiple cell sites. Figure 1 illustrates C-RAN architecture and focuses on three main components: (i) RRHs (antennas), (ii) BBU pools (edge data centers) and (iii) fronthaul network.

The centralization of computing resources in C-RAN enables to achieve costs savings and resource utilization gains (see [2], [3] and [4], for instance). However, such gains can be only reached when optimally assigning the heterogeneous antennas demands, with strict latency and processing expectations, to the available edge data centers. Hence, TSPs are investigating new resource allocation algorithms to efficiently allocate the limited processing resources of the edge data centers to the antennas demands when jointly meeting the latency and processing requirements.

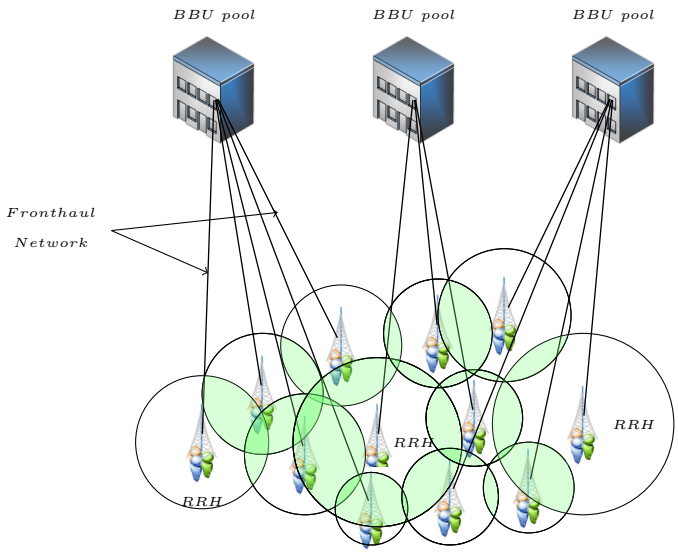


Figure 1: C-RAN architecture and components

1.1. Objective and contributions

This paper focuses on proposing new optimization algorithms to efficiently assign the antennas demands to the edge data centers in order to improve the efficiency of network resource utilization when meeting strong latency requirements on the fronthaul network. To reach these objectives, we propose an exact approach based on ILP formulation to derive an appropriate algorithm to find optimal solutions for the RRH-BBU assignment problem. This exact approach provides the best RRH-BBU assignment strategies by jointly reducing the fronthaul latency and the resource consumption. However, the ILP approach can only deal with small and medium problem instances. Thus, for larger problem instances, we propose three approximation algorithms, based on exact theories and approaches, that scale well and converge reasonably fast. Our proposed algorithms, exact and heuristics, are summarized as follows:

1. **ILP formulation**: is an exact approach based on the convex hull description of the constrained resource allocation problem to identify the most appropriate strategies for the assignment of antennas demands to

the available edge data centers. The proposed ILP formulation will jointly optimize the communication latency and the network resource consumption. This approach guarantees optimal (best) solutions for the RRH-BBU assignment problem.

2. **Matroid-based algorithm**: we propose a new approximation algorithm based on Matroid theory [5] to deal with RRH-BBU assignment problem. Matroid is an exact approach and we use it with minor modifications to propose a heuristic algorithm for the addressed problem. It is worth noting that this is the first time matroid theory will be used to address constrained resource allocation problems in the context of C-RAN.

3. **b-Matching-based algorithm**: we investigate new formulation based on b-matching approach [6] that aims to find the minimum weight matching between antennas and edge data centers, with limited capacity of processing, when satisfying the expected communication latency.

4. **Multiple knapsack-based algorithm**: we propose an approximation algorithm based on multiple knapsack formulation, which has been very used in the literature to solve many variants of resource allocation problems (for instance [7], [8], [9] and [10]). In this paper, we use the multiple knapsack formulation to address the RRH-BBU assignment problem in the context of C-RAN.

1.2. Paper organization

The rest of this paper is organized as follows: Section 2 is dedicated to deeply analyze the most relevant works in the literature addressing the RRH-BBU assignment problem. Section 3 describes our system model for the addressed problem and discusses its complexity. Section 4 introduces our proposed algorithms, exact and heuristics. Numerical results are presented in Section 5 to highlight the performance of our proposed algorithms using several scenarios. Conclusion and future research challenges are presented in Section 6.

2. Related work

The deployment of C-RAN architecture, where the infrastructure is shared across multiple cell sites, is expected to reduce network costs (CAPEX and OPEX) as well as to improve the resource utilization efficiency [11]. To achieve these goals, TSPs investigate new algorithms to determine the best strategies to assign RRHs to BBUs (known as RRH-BBU assignment problem) when jointly meeting the strong latency expectations and the processing requirements of antennas demands.

In this context, authors in [12] and [13] discussed new mathematical modeling to cope with RRH-BBU assignment problem. They proposed a mathematical formulation based on ILP approach in which only BBUs processing capacity constraints are considered. The proposed exact optimization model does not take into account the transmission delay on the fronthaul network and the latency requirements of antennas demands. To cope with scalability issues, both these references proposed approximation algorithms that do not guarantee the convergence to an optimal solution. In our paper, we address the RRH-BBU assignment problem when taking into account strong latency expectations and respecting the edge data centers' limited capacity constraints. Our joint optimization is represented by an exact formulation before investigating heuristic algorithms that converge to near-optimal solutions in acceptable times.

Authors of reference [14] proposed a load-aware dynamic mapping between RRHs and BBUs with the aim of minimizing the number of active BBUs required to process the computational resource demands. The authors introduced a heuristic DRA (Dynamic RRH Assignment) to dynamically optimize the BBU pooling gain. They claimed that their approach delivers an almost optimal performance in terms of computational resource gain and convergence time as compared to First-Fit Decreasing (FFD) algorithm. Similarly, another resource allocation algorithm was introduced in [15] to minimize the number of active BBUs, that are required to serve all users in the network, in order to save more energy. In our work, and in addition to the proposed ILP formulation used as

reference to benchmark other algorithms, we propose three heuristic approaches to guarantee the convergence of the constrained resource allocation problem to optimal solutions in negligible times.

Another work addressing the RRH-BBU assignment problem was proposed in [16]. Indeed, the authors of this paper proposed a greedy algorithm to assign the aggregated demands of each cell to the BBU pool in such a way that the power consumption of the physical resources is minimized. The authors did not consider the latency requirements of cells in their optimization model. Since the latency and transmission delay constraints are very strong in the context of C-RAN, we propose exact and heuristic algorithms based on a joint optimization of communication latency and computing resource allocation.

In [17], the authors introduced a mathematical formulation based on ILP to optimally assign antennas demands to different BBU pools. This work aims to minimize the length of fiber while maximizing the statistical multiplexing gain for each BBU pool hosting the baseband functions. The proposed approach shows that the optimal assignment of RRHs to the BBU pools depends on the length of fiber and BBU resources. In this paper, we proposed an exact formulation for the same problem and to scale, our contribution consists in investigating new and rapid approaches to guarantee the convergence to near optimal solutions when considering the same parameters than those used in [17].

Authors in [18] investigated new algorithms to determine the best strategies for RRH-BBU mapping by finding the optimal clustering of existing RRHs. They modeled this problem as bin packing problem when considering two main constraints : (i) the radio resources of each active BBU must be enough to meet the demands of its mapped RRHs and (ii) the set of antennas that will be assigned to a BBU should be geographically adjacent. Exact and heuristic algorithms are provided to reduce the network power consumption when guaranteeing good Quality of Service (QoS) for end-users. Nevertheless, the proposed formulation did not consider the communication latency on the fronthaul network joining RRHs to BBU pools. In our work, we address the same problem by proposing an exact approach based on ILP model and approximation algorithms

to find the best assignment of antennas to centralized data centers when jointly considering the limited processing capacity in BBU pools and the transmission delay on fronthaul links.

Authors in [19] proposed an exact approach based on ILP formulation to determine the optimal placement of BBU pools over a Wavelength Division Multiplexing (WDM) aggregation network. Their optimization proposal jointly minimizes the number of necessary BBU pools and the total number of optical fiber links when meeting the strong latency expectations on the fronthaul network. In our paper, and in addition to the ILP formulation, we propose three approximation algorithms to find good strategies to assign antennas demands to the centralized data centers when considering different transport requirements and limited capacity of processing in BBU pools. The aim of our algorithms is to jointly satisfy the latency requirements and achieve resource utilization gains in terms of number of necessary BBU pools. The obtained solutions by the ILP-based approach will be used to benchmark the performance of our heuristic algorithms in terms of resource utilization, convergence time and scalability.

Authors in [20] discussed the placement of the fog nodes in a Fog Computing/NFV environment (equivalent to BBU pools in the context of C-RAN) while meeting 5G mobile network requirements. They proposed a mathematical formulation based on Mixed-Integer Linear Programming (MILP) which consists in minimizing the number of fog nodes and their capacities under strict latency requirements and limited processing capacity constraints. Then, for sake of scalability, they proposed a heuristic algorithm called Hybrid Simulated Annealing (Hybrid-SA) that combines SA method and some local search techniques to reduce the necessary time to obtain solutions, especially for large problem instances. Simulation results highlight the efficiency of the Hybrid-SA algorithm and its ability in minimizing the number of fog nodes. However, the convergence time of this algorithm remains a bit high when considering large problem instances. In our paper, we investigate an exact approach based on ILP model and three heuristic algorithms with similar objective and constraints to deal with the RRH-BBU assignment problem in the context of C-RAN. Ac-

cording to different performance metrics, we deeply analyze the performance of our heuristic algorithms with the aim of providing optimal or near optimal solutions in a negligible convergence time (compared to the convergence time of the Hybrid-SA algorithm) even for large network sizes.

Some other existing works (for instance [21], [22] and [23]) addressed the resource allocation problem in the context of C-RAN by only focusing on minimizing the energy consumption in the BBU pool without taking into account the fronthaul latency constraints. In our work, we seek new algorithms to reduce the network costs by jointly optimizing the resource consumption and the communication latency in order to achieve optimal utilization of computing resources.

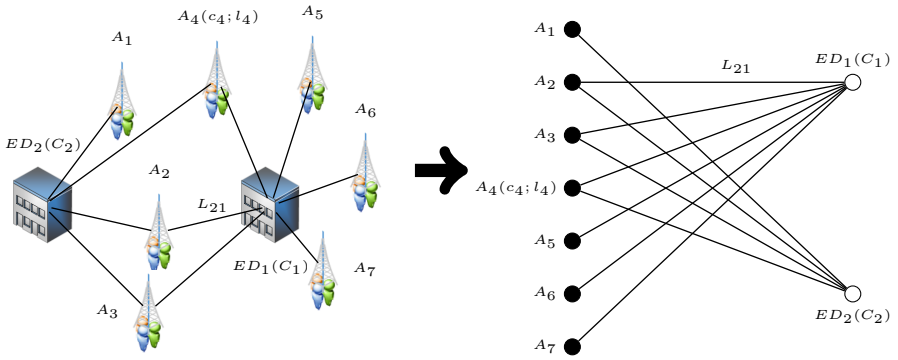
3. Problem statement

In this section, we describe the system model that we consider to address the RRH-BBU assignment problem and we introduce all variables and parameters used in the description of the problem. Then, and before providing our proposed algorithms, we discuss the complexity of the RRH-BBU assignment problem when considering all constraints that will be defined below.

3.1. System model

We consider the system model, as shown in Figure 2, to define the constrained resource allocation problem that aims to efficiently assign the antennas demands to the most appropriate edge data centers when strict latency and processing requirements are met. Our system model represents a C-RAN network where RRHs (antennas) and BBU pools (edge data centers) are deployed in a large area. As depicted in Figure 2a, our network architecture contains a set of antennas, denoted by I, each of which is defined by a position on the plane. These antennas $i \in I$ have variable expected latencies l_i and processing requirements in terms of CPU cores c_i, depending on aggregated end-users' demands. The RRHs are served by a finite set of available edge data centers denoted by

J. Each edge data center $j \in J$ has a limited computing processing capacity C_j expressed as number of CPU cores.



(a) Network topology (b) bipartite graph

C_1 (resp. C_2) : total number of available CPU cores in BBU pool ED_1 (resp. ED_2)
c_4 : number of CPU cores requested for processing the demands of antenna A_4
l_4 : expected latency for processing the demands of antenna A_4
L_{21} : communication latency on the fronthaul link between A_2 and ED_1

Figure 2: System model for constrained resource allocation problem

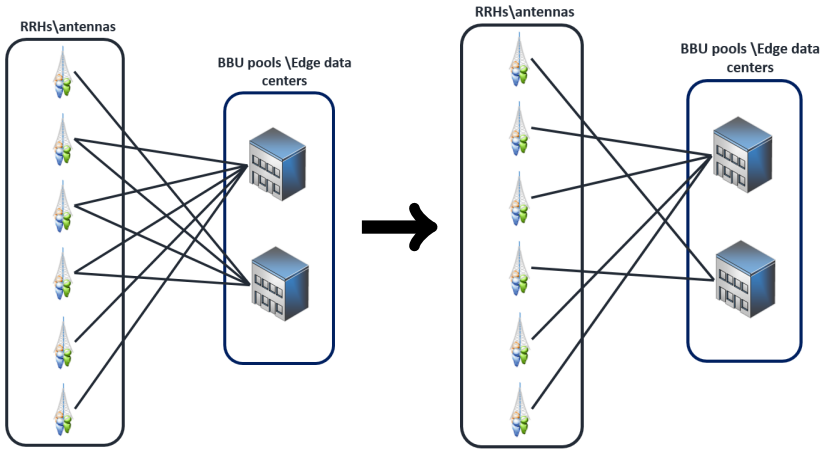
The antennas are connected to the edge data centers via fronthaul network, which is represented by a set of communication links. Each fronthaul link between an antenna $i \in I$ and an edge data center $j \in J$ has a transmission delay L_{ij} that should be kept below **1 millisecond** in order to meet HARQ[1] requirements (see [2], [24] and [25]). This requires that the maximum distance d_{ij} between RRH i and BBU pool j must not exceed **20** to **40 kilometers** ([2] and [26]). The data traffic on the fronthaul network can be transmitted using different protocols, most commonly CPRI [27], or in some cases OBSAI [28]. In our system model, and according to [2] and [29], the transmission delay on the fronthaul network is **5 microseconds per Kilometer** and thus the communication (fronthaul) latency between RRHs and BBU pools vary between **100**

[1]HARQ (Hybrid Automatic Repeat reQuest) is the process that poses the most stringent delay requirement for cellular networks

and **200 microseconds** at the most.

As depicted in Figure 2, our network topology (Figure 2a) can be modeled by a weighted bipartite graph $G = (I \cup J, E)$ containing a set of antennas I in one side, a set of edge data centers J in the other side and a set of fronthaul links represented by the set of edges E. The weight value, denoted by L_{ij}, on each edge in the graph G represents the communication latency between the antenna $i \in I$ and the edge data center $j \in J$. The bipartite graph $G = (I \cup J, E)$ will be used to efficiently assign each antenna **to exactly one edge data center** when meeting the processing and latency requirements.

For sake of clarity, we give in Figure 3 a simple example of C-RAN network which is composed by 6 RRHs (antennas), 2 edge data centers (BBU pools) and a fronthaul network represented by a set of communication links.



(a) Initial graph (b) Final graph

Figure 3: A solution example of the constrained resource allocation problem

The constrained resource allocation problem consists in determining the optimal strategies to assign the antennas demands to the available edge data centers under strict processing and latency requirements. Hence, we aim to select, in the bipartite graph of Figure 3a, the optimal matching of all considered antennas

with the available edge data centers. The optimal assignment of all considered antennas to the BBU pools is achieved when latency and resource consumption (number of used edge data centers) are minimized. The right graph (Figure 3b) represents a feasible solution of the RRH-BBU assignment problem.

For sake of clarity, we summarize in Table 1 all variables and parameters that will be used, in the following, to model the constrained resource allocation problem.

Table 1: Variables and parameters

$G = (I \cup J, E)$	:	weighted bipartite graph
I	:	set of antennas/RRHs
J	:	set of edge data centers/BBU pools
E	:	set of communication links between I and J
d_{ij}	:	distance between an antenna i (with coordinates (x_i, y_i)) and an edge data center j (with coordinates (x_j, y_j))
c_i	:	total number of CPU cores requested for processing the aggregated demands of antenna i
C_j	:	available computing resources (CPU cores) in each edge data center j
l_i	:	expected latency for processing the aggregated demands of antenna i
L_{ij}	:	transmission delay (latency) on the communication link between an antenna i and an edge data center j

3.2. Problem complexity

Before investigating new algorithms to solve the RRH-BBU assignment problem, we address in this section the problem's complexity. We provide a theorem and a proof confirming the problem's NP-Hardness.

Theorem 3.1. *Finding the optimal assignment of the antennas (RRHs) demands to available edge data centers (BBU pools) is an NP-Hard problem.*

Proof As it is described above, the constrained resource allocation problem consists in finding the optimal assignment of antennas demands to the available edge data centers with the aim of satisfying latency and processing requirements and minimizing network resource utilization. Our problem is close to the Generalized Assignment Problem (GAP) (see [30] for more details), which is a classical generalization of both multiple knapsack problem [31] and bin packing problem [32]. Indeed, GAP consists in finding a feasible packing of the items (each item is defined by a size and a profit) into the bins (each bin has a limited capacity) that maximizes the total profit.

Our constrained resource allocation problem is very similar to GAP in which the antennas can be considered as items and edge data centers are the bins. Furthermore, compared to GAP, our constrained resource allocation problem has additional constraints concerning the latency requirements on the communication links joining the antennas and edge data centers. Hence, the relaxation of these constraints give an instance of GAP which means that the **optimal** solution of GAP is a **feasible** (not necessarily optimal) solution for RRH-BBU assignment problem.

Authors in [31] and [33] have proven the NP-Hardness of GAP. Therefore, by using the previous linear reduction from our problem to GAP, we deduce that our RRH-BBU assignment problem is also NP-Hard which means that finding the optimal assignment of the antennas demands to the available edge data centers is an NP-Hard problem.

■

4. Proposed algorithms

In this section, we provide an exact approach based on ILP formulation to determine the optimal assignment of antennas demands to the edge data centers. Since the NP-hardness of the addressed problem (see the proof in Section 3.2), we propose three approximation algorithms to rapidly deal with the RRH-BBU assignment problem even for large problem instances.

4.1. Mathematical formulation based on ILP model

In this section, we investigate a new mathematical formulation based on ILP approach to optimally solve the RRH-BBU assignment problem. It is worth noting that this approach is proposed to provide optimal (best) solutions for small and medium network sizes and these solutions will be used then as **reference** to benchmark the performance of our proposed heuristic algorithms according to several performance metrics.

Decision variables

We start our problem's modeling by introducing two decision variables as follows:

- x_{ij} is a binary decision variable, the value of which is 1 if the antenna $i \in I$ is assigned to the edge data center $j \in J$, and 0 otherwise.
- y_j is a binary decision variable, the value of which is 1 if the edge data center j is used (activated) to host at least one RRH (antenna), and 0 otherwise.

Objective function

The objective of our RRH-BBU assignment problem is to efficiently allocate the computing resources of the most appropriate ("best") edge data centers to the antennas demands when jointly satisfying their processing and latency requirements. This objective will be reached by finding the best trade-off between transport requirements on the fronthaul network and the number of active edge data centers. In fact, similarly to [12], [19], [21] and [34], our objective function (1) contains two terms : the first denotes the total assignment cost in terms of communication latency on the fronthaul network and the second term represents the total network resource utilization in terms of used edge data centers. Using this objective function, we aim to find an optimal solution for the RRH-BBU assignment problem which is equivalent to select, in the final graph, the optimal matching of all antennas to the available edge data centers when jointly

optimizing the latency and the resource consumption (as shown in the example of Figure 2).

$$\min \quad \mathcal{F} \quad = \sum_{j \in J} \sum_{i \in I} L_{ij} \times x_{ij} \quad + \quad \sum_{j \in J} y_j \tag{1}$$

Constraints

Constraints (2) guarantee that each antenna i is connected **to exactly one** edge data center j. These constraints are considered in the graph solution of Figure 3 where each antenna is mapped on exactly one edge data center.

$$\sum_{j \in J} x_{ij} = 1, \quad \forall i \in I \tag{2}$$

Constraints (3) ensure that the assignment of antennas demands to the BBU pools does not violate the edge data centers' limited capacity constraints. In fact, as mentioned in Section 3.1, each edge data center j has a limited processing capacity C_j in terms of CPU cores and thus the total number of CPU cores requested for processing all antennas must not exceed the available computing resources of the selected edge data center.

$$\sum_{i \in I} c_i \times x_{ij} \leq C_j \times y_j, \quad \forall j \in J \tag{3}$$

Our optimization will select the most appropriate fronthaul links that satisfy the latency requirements of the antennas demands. In fact, constraints (4) impose that the transmission delay L_{ij} on the selected communication link between the antenna and the edge data center must not exceed the expected latency l_i. Thus, as shown in Figure 3, only expected latencies will be kept in the final solution. This is guaranteed by the following inequalities:

$$L_{ij} \times x_{ij} \leq l_i, \quad \forall i \in I, \forall j \in J \tag{4}$$

Constraints (5) ensure that if there exists at least one antenna assigned to the edge data center j (i.e. $\sum_{i \in I} x_{ij} \geq 1$), then this edge data center is activated (i.e. $y_j = 1$) and can be used to host other antennas as long as its processing capacity is not exceeded. We recall that the optimal assignment of antennas

demands to the edge data centers is reached when the number of used edge data centers is minimized. This will help network operators to reduce their network costs.

$$y_j \leq \sum_{i \in I} x_{ij}, \quad \forall j \in J \tag{5}$$

Our mathematical model is hence characterized by the above ILP formulation which is represented by the objective function (1) and the set of above constraints (2), (3), (4) and (5). Using a Branch-and-Bound methods [35], our proposed mathematical model explores all feasible solutions for the RRH-BBU assignment problem and selects the best one allowing to find the optimal strategies to assign the limited processing resources in the available edge data centers to the antennas demands. This allows to achieve resource utilization gains by using a small number of edge data centers when meeting latency requirements.

Nevertheless, our addressed RRH-BBU assignment problem is NP-Hard (see the proof in Section 3.2) and thus the necessary convergence time to obtain optimal solutions using this approach exponentially increases with the increase of number of antennas demands. Hence, we need to investigate new approximation algorithms that converge rapidly and provide optimal or near-optimal solutions for large problem instances.

In the following, we introduce three heuristic algorithms (i) matroid-based approach, (ii) b-matching formulation and (iii) multiple knapsack-based algorithm. We recall that the obtained solution by the exact approach based on ILP formulation is **optimum** ("best" solution) and will be used to evaluate the quality of solutions provided by the proposed heuristic algorithms.

4.2. Matroid-based algorithm

In addition to the above exact model based on ILP formulation, we investigate new polynomial time algorithm that can scale to larger number of antennas and edge data centers. Since the exact solution is efficiently optimizing the latency and the resource allocation jointly, we propose an approximation algorithm based on matroid theory with similar properties and criteria.

4.2.1. Matroids background and construction

In the following, we introduce the definition of a matroid using the theorem provided by [35].

Definition A matroid $M = (E, \mathcal{F})$ is a structure in which E is a finite set of elements and $\mathcal{F}$ is a family of subsets of E verifying the following principal properties:

1. $\emptyset \in \mathcal{F}$.
2. If $A \in \mathcal{F}$ and $B \subseteq A$, then $B \in \mathcal{F}$.
3. If $A, B \in \mathcal{F}$, and $\mid B \mid > \mid A \mid$ thus $\exists e \in B \setminus A$, such that $A \cup \{e\} \in \mathcal{F}$.

If $\mathcal{F}$ is only satisfying the properties (1) and (2), then we are invoking an independent system. A *basis* of E is a maximal set in E, and all basis of a matroid have the same cardinality. More details on matroid theory can be found in [5], [36] and [37].

Using the bipartite graph $G = (I \cup J, E)$ of Figure 3, the optimal solution of the RRH-BBU assignment problem consists in hosting each antenna demand in one edge data center. Similarly, in the bipartite graph G, each vertex $i \in I$ will be assigned to exactly one vertex $j \in J$, and each vertex $j \in J$ can be a neighbor of different vertices in I as each edge data center can host more than one antenna demand. This yields a solution as presented in Figure 3, showing a forest of trees optimally linking antennas (RRHs) and edge data centers (BBU pools). Thus, we propose the following theorem that defines our assignment algorithm using matroid theory.

Theorem 4.1. *Let $G = (I \cup J, E)$ be a weighted bipartite graph as shown in Figure 3. By relaxing data centers' limited capacities constraints, $M = (E, \mathcal{F})$ is a matroid, with $\mathcal{F} = \{I \subseteq E, \ \ I \ \ is \ \ a \ \ forest \ \ of \ \ trees\}$.*

For the best of our knowledge, our matroid-based algorithm is well known in the literature (see [5] for instance) and it is noted by the graphic matroid.

Proof The proof is given as follows:

- The first condition (1) of the definition 4.2.1 concerning matroids, is trivial.
- The second condition (2) of the definition 4.2.1: Suppose we have $A \in \mathcal{F}$, and according to the definition of $\mathcal{F}$, A is a forest of trees. Thus, if $B \subseteq A$, then the connected components of B are also trees even by deleting one or multiple edges in A. This leads to easily conclude that $B \in \mathcal{F}$.
- To prove the last condition (3) of the definition 4.2.1, we note by $A = \cup_{i=1}^{k} A_i$ which represents the connected components (trees) of A. Then, for all $i = 1, \ldots, k$, we suppose $G_i = (T_i, A_i)$, where G_i is a tree with $|T_i|$ vertices and $|A_i|$ edges. This leads to deduce the number of vertices of A given by

$$n_A = \sum_{i=1}^{k} \mid T_i \mid = \mid A \mid + k. \tag{6}$$

We also suppose $B = \cup_{j=1}^{t} B_j$, we note by $G_i^{'} = (T_i^{'}, B_i)$, where $G_i^{'}$ is a tree with $|T_i^{'}|$ vertices and $|B_i|$ edges. The number of nodes of B is then given by :

$$n_B = \sum_{j=1}^{t} \mid T_i^{'} \mid = \mid B \mid + t. \tag{7}$$

By using $\mid B \mid > \mid A \mid$, two cases are discussed:

1. If $n_B > n_A$ $(t > k)$: We suppose that B reaches more vertices than A, so there exists a vertex x covered by B and not by A. Suppose that $e \in B$ is an edge which contains x as one of its two extremities, we finally deduce that $A \cup \{e\} \in \mathcal{F}$.
2. If $n_B < n_A$: We suppose that the edges of B connects each couple of nodes in A in the same connected component (tree) A_i. Using the absurd reasoning, we suppose that there is no edge $e \in B \setminus A$, leading to get $A \cup \{e\} \in \mathcal{F}$. This means that:
   - The edge $e \in B$, relies two vertices in the same component (tree) A_i and forms a cycle.

In this case, the number of edges of B will verify $\mid B \mid \leq \mid V_1 \mid + \mid V_2 \mid + \ldots + \mid V_k \mid$, then $\mid B \mid \leq \mid A \mid$ which contradicts our hypothesis $\mid B \mid > \mid A \mid$.

The proposed matroid formulation, given by theorem 4.1, does not consider the hypothesis of edge data centers' limited capacity constraints, which are very important in our RRH-BBU assignment problem. In fact, these constraints influence the choice of the solicited edge data center to host antennas demands. To introduce these constraints in our solution, we propose a simple modification in the matroid-based algorithm as illustrated below.

Algorithm 1 Matroid-based algorithm for RRH-BBU assignment problem

Put $A = \emptyset$;
$l_{e_1} \leq l_{e_2} \leq \ldots \leq l_{e_m}$;
for $i = 1$ to m **do**
  if $A \cup \{e_i\} \in \mathcal{F}$ **then**
    if $c_{I(e_i)} \leq C_{T(e_i)}$ **then**
      $A := A \cup \{e_i\}$
      $C_{T(e_i)} - = c_{I(e_i)}$
    end if
  end if
end for

l_{e_i} is the communication latency on the edge e_i;
$I(e_i)$ (resp. $T(e_i)$) represents the initial (resp. terminal) extremity of the edge e_i;
$c_{I(e_i)}$ represents the number of CPU cores requested for processing the antenna demand $I(e_i)$;
$C_{T(e_i)}$ represents the available amount of CPU in an edge data center $T(e_i)$.

4.2.2. Matroid-based algorithm's complexity

It is important to evaluate the complexity of our proposed matroid-based algorithm (Algorithm 1). We note that the addressed problem is NP-Hard, and we need rapid and cost-efficient approaches to cope with this complexity.

Our proposed matroid-based algorithm, as described in Algorithm 1, has a global complexity (in the worst case) of $O(m\ln(m)+m)$, where $m\ln(m)$ is the complexity of sorting a set of m edges according to their weights (latency in our case), and the "**For**" loop indicated in Algorithm 1 iterates m times.

In addition to the matroid-based algorithm, we introduce in the following another heuristic algorithm based on b-matching approach. This proposal aims to find the optimal mapping between RRHs and BBUs, when satisfying all antennas demands. Using the b-matching algorithm, we seek to rapidly reach optimal or near-optimal solutions for large instances of RRH-BBU assignment problem. This may not be feasible with matroid-based approach, especially when the number of antennas demands becomes important (more than 100 antennas) and the computing resources in available edge data centers are limited.

4.3. b-Matching algorithm

To address larger problem instances, we propose a new heuristic approach based on b-matching theory to attend optimal or near optimal solution in negligible times. The proposed heuristic considers the bipartite graph described in Section 3.1 and consists in finding the minimum weight b-matching to rapidly assign the antennas demands to the available edge data centers. The definition of the b-matching problem is introduced in the following [35] :

Definition Let G be an undirected graph with integral edge capacities $u : E(G) \to \mathbb{N} \cup \{\infty\}$ and numbers $b : V(G) \to \mathbb{N}$. Then a b-matching in (G, u) is a function $f : E(G) \to \mathbb{Z}_+$ with $f(e) \leq u(e)$ for all $e \in E(G)$ and $\sum_{e\in\delta(v)} f(e) \leq b(v)$ for all $v \in V(G)$.

where $V(G)$ (resp. $E(G)$) denotes the set of vertices (resp. edges) in the graph G and $\delta(v)$ is a set of incident edges of v.

According to this definition, we introduce new algorithm that solves the constrained resource allocation problem by finding the minimum weight b-matching in the bipartite graph $G = (I \cup J, E)$. This algorithm will jointly consider the latency constraints and the edge data center capacity constraints.

Proposition 4.2. *Let $G = (I \cup J, E)$ be a weighted bipartite graph as shown in Figure 3. The RRH-BBU assignment problem can be solved by finding the minimum weight b-matching while considering the following parameters:*

- *The integral edge capacities :* $u = 1$.
- $b(v) = 1, \quad \forall v \in I$ *(I is a set of antennas).*
- $b(v) = min\{|I_v|, \lfloor \frac{C_v}{c(v)} \rfloor\}, \quad \forall v \in J$ *(J is a set of edge data centers).*

where :

- I_v is a subset of antennas that can be assigned to the edge data center $v \in J$ when satisfying the expected latency and CPU cores number requested for each antenna demand : $I_v = \{i \in I \quad | \quad l_i \geq L_{ij} \wedge c_i \leq C_j\}$.
- $\overline{c(v)}$ is the average number of CPU cores of antennas demands that can be assigned to the edge data center $v \in J$: $\overline{c(v)} = \frac{\sum_{i \in I_v} c_i}{|I_v|}$.

In addition and in order to help our optimization to find optimal solution with integer variables, we add the blossom inequalities given by the following formula :

$$\sum_{e \in E(G[X])} x_e + \sum_{e \in F} x_e \leq \lfloor \frac{1}{2} (\sum_{v \in X} b(v) + |F|) \rfloor, \quad \forall X \subseteq I \cup J, F \subseteq \delta(X) \qquad (8)$$

where $E(G(X))$ represents a subset of edges in the subgraph $G(X)$ generated by a subset of vertices X and $\delta(X)$ is a set of incident edges of X (for more details, see [35] and [38]).

Finally, we use the obtained result of Proposition 4.2 to provide a new minimum weighted b-matching formulation to polynomially solve the RRH-BBU assignment problem. The mathematical formulation is given by the following

model:

$$
\begin{aligned}
&\min \quad \mathscr{F} = \textstyle\sum_{e\in E} L_e \times x_e \\
&S.T.: \\
&\textstyle\sum_{e\in\delta(v)} x_e = 1, \quad \forall v \in I; \\
&\textstyle\sum_{e\in\delta(v)} x_e \leq min\{|I_v|, \lfloor \frac{C_v}{c(v)} \rfloor\}, \quad \forall v \in J; \\
&\textstyle\sum_{e\in E(G[X])} x_e + \sum_{e\in F} x_e \leq \lfloor \frac{1}{2}(\sum_{v\in X} b(v) + |F|) \rfloor, \quad \forall X \subseteq I \cup J, F \subseteq \delta(X); \\
&x_e \in \mathbb{R}^+, \quad \forall e \in E;
\end{aligned}
\tag{9}
$$

4.3.1. b-Matching algorithm's complexity

To assess the ability of the b-matching algorithm to find good solutions with large-scale graph instances in reasonable times, we analyze in this section the complexity of the proposed algorithm. We note that the objective of this algorithm is to assign antennas demands to available edge data centers under hard latency requirements and limited processing capacity constraints. The complexity of our proposed linear programming or b-matching solution is $O\left(|V||E|^2 \ln(\frac{|V|^2}{|E|})\right)$ where $V = I \cup J$ and E is the set of weighted links between I and J. This approach is a simple linear program with a negligible complexity. For interested readers, more details can be found in [39].

In the following, we introduce another heuristic algorithm using the multiple knapsack formulation. As mentioned before, the multiple knapsack approach has been very well used to address resource allocation problems in different contexts. In the context of C-RAN, we propose a modified algorithm based on multiple knapsack formulation to solve the RRH-BBU assignment problem. The solutions provided by this algorithm will be benchmarked with matroid, b-matching and ILP algorithms to better evaluate the performance of our algorithms under different simulation scenarios and performance metrics.

4.4. Multiple knapsack-based algorithm

In this section, we propose an approximation algorithm based on multiple knapsack formulation. In fact, the multiple knapsack formulation is a generalization of the classical knapsack problem (KP) from a single knapsack to m knapsacks with different capacities. The objective of multiple knapsack algorithm is to assign each item to at most one of the knapsacks such that none of the capacity constraints are violated and the total profit of the items put into knapsacks is maximized. The multiple knapsack algorithm is introduced in the following definition.

Definition Given a set of n items and a set of m knapsacks ($m < n$), with p_j = profit of item j, w_j = weight of item j, c_i = capacity of knapsack i, find m disjoint subsets of items with the total profit of the selected items is a maximum, and each subset can be assigned to different knapsacks whose capacity is less than the total weight of items in the subset.

According to this definition and by considering the bipartite graph $G = (I \cup J, E)$ described in Figure 3, we obtain the following equivalence between the constrained resource allocation problem addressed in this paper and the multiple knapsack formulation :

- The knapsacks are the edge data centers ($j \in J$).
- The antennas demands ($i \in I$) are the items to be inserted in the knapsacks (data centers).
- The weight w_j is the amount of CPU cores c_i requested for processing the antenna demand i.
- The profit p_j does not vary between different antennas demands and can be set to 1 ($p_j = 1$).

This formulation addresses our resource allocation problem by only focusing on the processing capacity of the edge data centers without considering the latency requirements of antennas demands. The relaxation of latency constraints

influences the choice of which edge data center will host the antennas demands. Therefore, in order to consider these constraints in our solution, we introduce a simple modification in the multiple knapsack algorithm which consists in checking if the expected latency is guaranteed before assigning the antenna demand to the edge data center. We illustrate our multiple knapsack formulation in Algorithm 2.

Algorithm 2 Modified multiple knapsack Algorithm

Input: $G = (I \cup J, E)$, Antenna demands, Edge data centers.

Output: A joint mapping (CPU, Latency) of all antennas demands on the available edge data centers.

This is summarized formally in steps:

Step 1: Sort the edge data centers ($j \in J$) in increasing order of their CPU capacities C_j;

Step 2: Select the antennas demands that can be assigned to the selected edge data center j by checking if :

- The expected latency of the antenna demand is provided by the communication link joining it to the selected edge data center j;
- The available computing resources in the selected edge data center j are greater than the number of CPU cores requested by the antenna demand;

Step 3: Pick as many antennas demands as possible to the selected edge data center using the dynamic programming approach (see [40], for instance);

Step 4: Update the total number of available CPU cores in the selected edge data center;

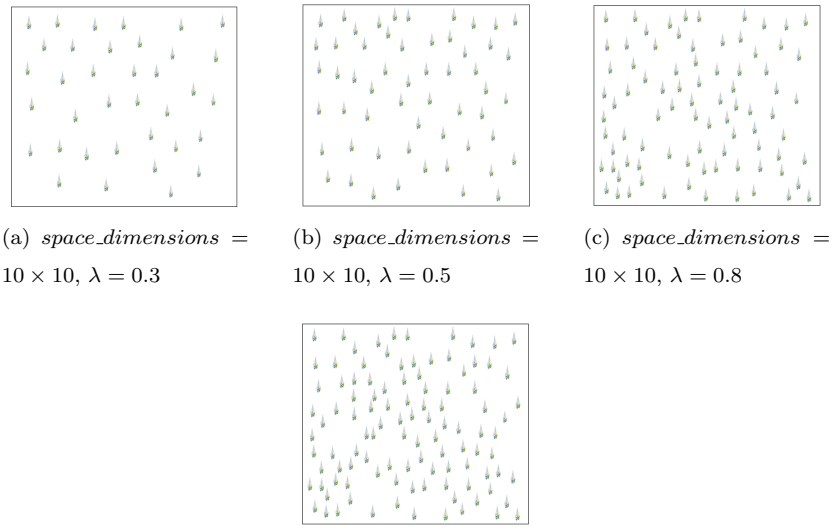
Step 5: Repeat Steps 2, 3 and 4 until all considered antennas demands are assigned to the edge data centers;

5. Performance evaluation

The simulation and experiments use the optimization solver Cplex [41] for the linear programming approaches, the exact approach based on ILP formulation (Section 4.1) and the b-matching formulation (given by formula (9) in Section 4.3). We first evaluate the performance of the exact algorithm and then we compare the obtained solutions (optimum) with those found by our heuristic algorithms in terms of convergence time, scalability and optimality. Each simulation scenario is run 100 times using different parameters.

5.1. Simulation settings and parameters

The performance evaluation of our algorithms is conducted using a 2.40 GHz PC with 8 GB RAM. The number of antennas is generated following a Poisson process with a parameter $\Lambda = \lambda \times space\_dimensions$, where λ is varying in the range $[0.1; 1]$, and $space\_dimensions$ in the range $[5; 20]$.



(a) $space\_dimensions = 10 \times 10$, $\lambda = 0.3$

(b) $space\_dimensions = 10 \times 10$, $\lambda = 0.5$

(c) $space\_dimensions = 10 \times 10$, $\lambda = 0.8$

(d) $space\_dimensions = 10 \times 10$, $\lambda = 1.0$

Figure 4: Example of simulation scenarios for RRH-BBU assignment problem

In Figure 4, we illustrate four examples of simulation scenarios when considering a cellular network in a region of space dimensions $space\_dimensions = 10 \times 10$ and varying the density of antennas $\lambda \in \{0.3; 0.5; 0.8; 1\}$.

Each antenna comprises a random number of demands (from end-users) presented in terms of an amount of CPU cores in the $[5; 10]$ interval (some papers such as [16] and [42] are considering allocation of Physical Resources Blocks PRBs, this is not changing our mathematical modeling and the convergence of our algorithms to good solutions). The number of edge data centers is set to 20, each of which has random computing resources (number of available CPU cores) drawn in the $[50; 200]$ CPU cores range. The workloads (i.e. aggregated amount of end-users demands in terms of equivalent CPU cores) of the antennas demands are expecting a latency to not exceed 1 millisecond and this is drawn randomly in the $[0.1; 1]$ milliseconds range. For sake of clarity, we summarize the simulation settings and parameters in Table 2.

Table 2: Simulation settings and parameters

Parameters	**Values**
Density of antennas	$\lambda \in [0.1; 1]$
Space dimensions	10×10; 20×20;...
Poisson parameter	$\Lambda = \lambda \times space\_dimensions$
Number of antennas	Poisson distribution: $\mathcal{P}(\Lambda)$
Antenna coordinates	Uniform distribution: $\mathcal{U}(0, space\_dimensions)$
Number of edge data centers	20
Latency between antenna i and edge data center j	5 µs/km
Expected latency of antenna i	$l_i \in [0.1ms;\ 1ms]$
Number of CPU cores required by each antenna i	$c_i \in [5; 10]$
Number of CPU cores in each edge data center j	$C_j \in [50; 200]$

5.2. Performance metrics

The metrics used for the performance assessment of our algorithms (exact and heuristics) are detailed in the following :

- **Convergence time**: is the time needed by the algorithms to converge to their best solutions.

- **Resource utilization rate**: is defined as the percentage of edge data centers that are used to host the aggregated antennas demands and it can be expressed as follows :

$$Resource\ utilization\ rate(\%) = \frac{\sum_{j \in J} y_j}{|J|} \times 100 \tag{10}$$

  where $|J|$ is the total number of available edge data centers.

- **Gap**: is used to benchmark the proposed heuristics with the exact ILP algorithm used as "reference and optimal solution". With no loss of generality, we focus on the comparison of CPU resource consumption (expressed by the percentage of edge data centers used to host all antennas demands). We note that the quality of the solution provided by the heuristic algorithms is better when the cost gap value is smaller (**optimum when the gap is equal to 0**). This metric is formally expressed as:

$$Gap(\%) = \mid Utilization\ rate(\textbf{ILP}) - Utilization\ rate(\textbf{Heuristic}) \mid \tag{11}$$

- **Rejection rate**: is the average of the percentage of antennas demands that cannot be assigned to each edge data center. This metric, can be expressed as a function of the decision variables (Section 4.1) and parameters described in Table 1 :

$$Rejection\ rate(\%) = \frac{|I| - \sum_{j \in J} \sum_{i \in I} x_{ij}}{|I|} \times 100 \tag{12}$$

  where $|I|$ is the total number of antennas.

- **SLA violations rate**: is the average of over-used edge data centers in terms of CPU cores. This metric will be mainly used to evaluate the ability

of the matroid-based approach in finding optimal solutions that **do not violate** the edge data centers' limited capacity constraints (which are defined, in the ILP formulation, by constraints (3)). We only focus on matroid-based algorithm (as defined by theorem 4.1) because there are no SLA violations with ILP, b-matching and multiple knapsack approaches. The average of SLA violations rate can be expressed as a function of decision variables (Section 4.1) and parameters (described in Table 1).

$$SLA\ violations\ rate(\%) = \frac{1}{|J|} \times \sum_{j \in J} \frac{\sum_{i \in I} c_i \times x_{ij} - C_j \times y_j}{C_j \times y_j} \times 100 \quad (13)$$

where $|J|$ is the total number of available edge data centers.

5.3. Performance analysis

5.3.1. Performance evaluation of ILP based approach

Table 3 depicts the performance results in terms of convergence time and rejection rate of the exact algorithm based on ILP formulation. This algorithm explores all feasible solutions before finding the optimum. This causes an exponential increase of the convergence time when increasing the number of antennas. Indeed, the ILP approach needs more than 4 minutes (4.39 minutes) to converge to optimal solutions for an instance of 400 antennas and 20 available edge data centers. This is expected since the addressed problem is NP-Hard. Thus, the ILP approach can be used for small or medium instances with a number of antennas not exceeding 100. Furthermore, the rejection rate is always equal to 0 which means that the exact approach based on ILP formulation is always able to assign all antennas demands to the available edge data centers.

Table 3: Performance of the exact approach based on ILP formulation

Space	λ	**#Antennas**	**Convergence time**	**Rejection rate**
10×10	0.3	30	**9.63s**	0
	0.5	50	**10.92s**	0
	0.8	80	**11.87s**	0
	1	100	**12.58s**	0
20×20	0.3	120	**62.09s**	0
	0.5	200	**86.56s**	0
	0.8	320	**2.87min**	0
	1	400	**4.39min**	0

5.3.2. Performance evaluation of heuristic algorithms

In Table 4, we consider different simulation scenarios by varying the dimensions of the considered space area as well as the density of deployed antennas (see the examples in Figure 4). Using these simulations, we would like to evaluate the performance of our proposed approximation algorithms: matroid-based algorithm (Algorithm 1), b-matching formulation given by (9) and the multiple knapsack-based approach (Algorithm 2).

As shown in Table 4, our heuristic algorithms are benchmarked with the ILP approach, that provides optimum solutions, using three performance metrics : the convergence time, the gap (11) to compare with optimal solutions provided by the exact approach and the rejection rate (12). We note that we calculate the gap only if the rejection rate is equal to 0, otherwise it is not really significant.

Table 4 highlights clearly the efficiency of the matroid-based algorithm in finding near optimal solutions faster than the exact approach based on ILP formulation. Indeed, the matroid approach provides good solutions with an average gap not exceeding 7% in worst cases and needs 2 **milliseconds** to converge when considering large graphs of 400 antennas and 20 available edge data centers. Thus, the matroid-based approach can be used to cope with large problem instances.

Table 4: Heuristic algorithms' performance assessment

Space	λ	**#Antennas**	**Heuristic algorithm**	**Convergence time**	**Gap(%)**	**Rejection rate(%)**
10×10	**0.3**	**30**	matroid	0.28**ms**	7	0
			b-matching	0.34s	7	0
			multiple knapsack	0.57**ms**	8	0
	0.5	**50**	matroid	0.38**ms**	5	0
			b-matching	0.36s	5	0
			multiple knapsack	1.01**ms**	11	0
	0.8	**80**	matroid	0.51**ms**	6	0
			b-matching	0.26s	5	0
			multiple knapsack	1.69**ms**	15	0
	1	**100**	matroid	0.88**ms**	6	0
			b-matching	0.39s	4	0
			multiple knapsack	3.35**ms**	15	0
20×20	**0.3**	**120**	matroid	0.94**ms**	-	1
			b-matching	0.4s	4	0
			multiple knapsack	4.35**ms**	-	1
	0.5	**200**	matroid	1.02**ms**	-	4
			b-matching	0.39s	6	0
			multiple knapsack	7.71**ms**	-	1
	0.8	**320**	matroid	1.75**ms**	-	17
			b-matching	0.34s	6	0
			multiple knapsack	25.44	-	3
	1	**400**	matroid	2**ms**	-	19
			b-matching	0.33s	5	0
			multiple knapsack	39.89**ms**	-	4

However, the matroid approach comes with some drawbacks such as it cannot assign all antennas demands for large problem instances. This is shown by the rejection rate metric of which its value can reach 19% for an instance of 400 antennas and 20 edge data centers.

To better evaluate the performance of our matroid-based algorithm, we calculate the rejection rate when increasing the number of considered edge data centers. For that, we consider two network instances of 320 and 400 antennas and we adjust the number of edge data centers from 20 to 60. The obtained results of these simulations are represented by Figure 5.

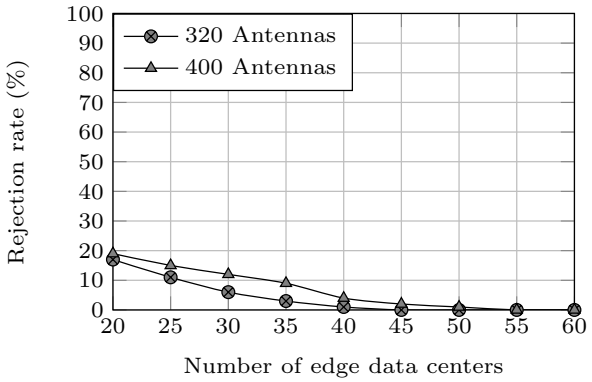


Figure 5: Matroid-based approach : rejection rate variation when increasing number of edge data centers

The simulation results in Figure 5 show that the rejection rate depends on the amount of available computing resources and thus decreases when the number of available edge data centers increases. In fact, for the first simulation scenario (320 antennas), matroid-based algorithm attends a rejection rate equal to 0 when there are at least 40 available edge data centers, while for the second simulation scenario (400 antennas), the rejection rate vanishes when there are at least 50 available edge data centers. This means that the matroid-based algorithm becomes more efficient when more resources (edge data centers) are considered.

In addition and in order to get a better grasp of the relative performance of the matroid-based approach, we illustrate in Figure 6 the SLA violations rate behavior according to different network sizes. In fact, we consider four simulation scenarios : $50, 100, 200, 320$ antennas to be efficiently assigned to a number of edge data centers ranging from 20 to 100. We recall that, for this simulation, we consider the matroid-based algorithm (as defined in theorem 4.1) when relaxing the edge data centers' limited capacity constraints and we calculate the SLA violations rate as defined by Formula (13).

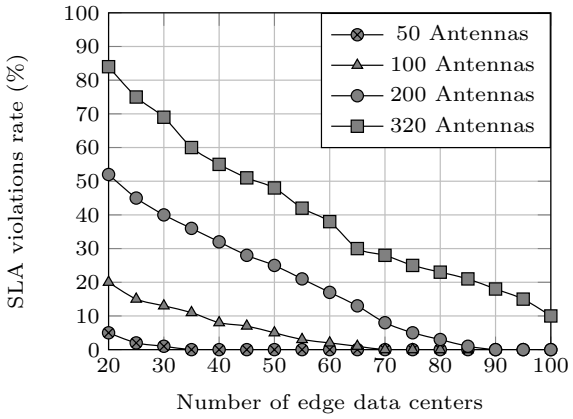


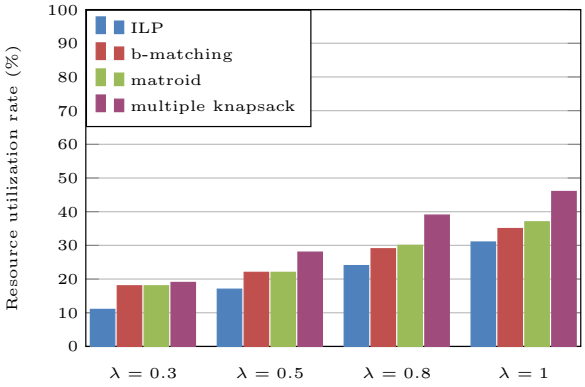
Figure 6: SLA violations rate behavior of the matroid-based approach

Simulation results in Figure 6 confirm that the SLA violations rate decreases when more processing resources (edge data centers) are considered. This confirms that the efficiency of the matroid-based algorithm depends on the amount of the available processing resources and attends good solutions when more resources (edge data centers) are used.

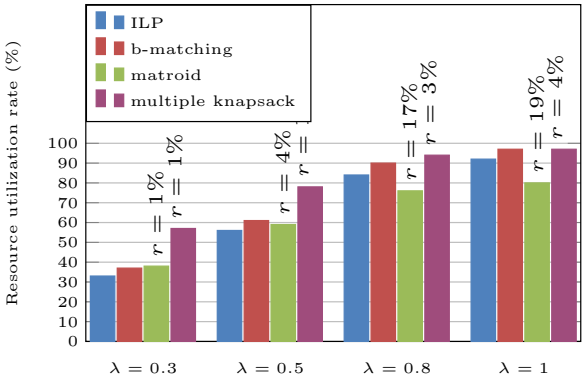
5.3.3. Resource utilization behavior

Figure 7 depicts the percentage of resource utilization (in terms of number of used edge data centers) obtained by the three approximation algorithms (matroid, b-matching and multiple knapsack) and the ILP approach. With a weak advantage of the ILP method which consists in investigating all the feasible solutions before keeping the optimal one, the matroid-based approach and b-matching algorithms can find an efficient assignment of antennas demands to the available edge data centers while the solution obtained by multiple knapsack algorithm consumes a larger number of edge data centers (as shown in Figure 7a).

It is important to mention that for larger problem instances (Figure 7b), b-matching algorithm always provides a near-optimal solution in terms of resource utilization compared to the ILP solution with **a rejection rate equal to 0**%. However, for matroid and multiple knapsack algorithms, the resource



(a) Space dimensions = 10×10



(b) Space dimensions = 20×20

r is the rejection rate calculated according to Formula (12)

Figure 7: Resource utilization in different space dimensions

utilization rate depends on the rejection rate (negligible but different from zero) in the case of large network size (see Figure 7b). Therefore, we deduce that **b-matching algorithm can easily scale when large problem instances are considered** and thus can be used by network operators to efficiently reduce their network costs (CAPEX and OPEX) and achieve network utilization gains.

5.3.4. Algorithms' performance evaluation using real traces

To better evaluate the performance of our proposed algorithms, we consider a real trace from a 4G-LTE cell map of the network operator Orange, in a small area in Paris [43]. As shown in Figure 8, this topology represents a cellular network containing 50 antennas with their given geographical positions (coordinates). Then, according to [2] and [24], we place 20 edge data centers on the cell map such that the distance separating the antennas and the edge data centers is between 20 and 40 Kilometers. Similarly to the simulation parameters described in Table 2, we consider that each edge data center has a limited capacity of processing in terms of CPU cores while the antennas demands have variable processing and latency requirements.

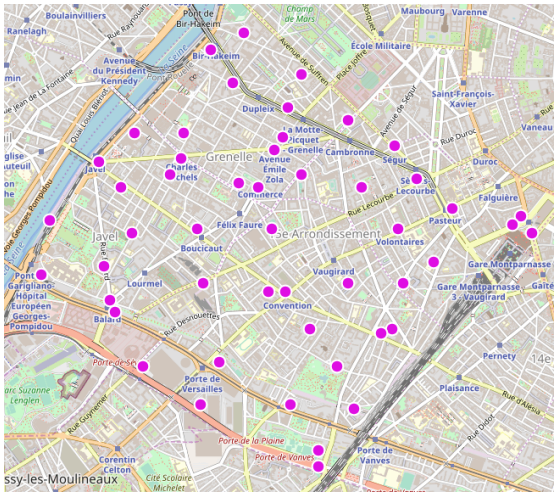


Figure 8: Real trace : Orange 4G-LTE cell map in Paris. Source: [43]

In this experimentation, we apply our exact approach based on ILP formulation (as described in Section 4.1) and the three proposed approximation algorithms, including matroid-based algorithm (Algorithm 1), b-matching formulation (9) and multiple knapsack-based approach (Algorithm 2), on the 4G-LTE cell map of Figure 8. The solutions provided by these algorithms are benchmarked according to three performance metrics : convergence time, resource utilization rate given by (10) and rejection rate defined by (12).

Table 5 shows that both matroid-based approach and b-matching formulation provide **optimal** solutions (the same solution provided by the ILP approach) in negligible times. In fact, with a weak advantage of the matroid-based approach which converges to the optimum in **0.58 ms**, the b-matching algorithm can also find an efficient assignment of antennas demands to the available edge data centers in **23.52 ms**. However, the solution obtained by multiple knapsack algorithm consumes a larger number of edge data centers, with a resource utilization rate equal to 25%. Regarding the rejection rate metric, all proposed algorithms can assign all considered antennas demands to the the available edge data centers and satisfy their latency and processing requirements without SLA violations.

Table 5: Performance evaluation using a real cellular network in Paris

Algorithm	**Convergence time (ms)**	**Resource utilization rate (%)**	**Rejection rate (%)**
ILP formulation	334.21	15	0
b-Matching algorithm	23.52	15	0
Matroid-based approach	0.58	15	0
Multiple knapsack algorithm	1.7	25	0

5.3.5. Scalability evaluation

The performance assessment would not be complete without addressing the scalability for very large problem instances. In fact, we propose a simulation scenario with an instance of **400** antennas and number of edge data centers in {**60**, **80**} which are both generated according to the parameters detailed in the Table 2. Simulation results in Table 6 show the efficiency of matroid-based approach and b-matching algorithm in finding good solutions in negligible times compared to ILP approach. Indeed, the matroid algorithm provides near optimal solutions (gap not exceeding 2%) in less than **28 milliseconds** and the b-matching algorithm can optimally solve the assignment problem in less than **4 seconds** (with gap value not exceeding 3%). However, the ILP approach is not converging in more than **1hour** due to the exploration of all feasible solutions.

Table 6: Algorithms' scalability assessment

#Antennas[1]	**#Edge data centers**[2]	**ILP**	**b-Matching**		**Matroid**		**Multiple knasapck**	
		Time	**Time**	**Gap**	**Time**	**Gap**	**Time**	**Gap**
400	60	**34.28min**	1.47s	3	6.42ms	2	82.84ms	18
	80	**1.02hour**	3.97s	2	27.16ms	2	107.4ms	19

This simulation is executed 100 times with different parameters.

[1] Antennas are generated as described in Table 2.

[2] Edge data centers are randomly distributed as mentioned in Section3.1.

5.3.6. Comparative analysis of proposed algorithms

In this section, we present a comprehensive comparison of the proposed algorithms for the joint constrained resource allocation and RRH-BBU assignment problem. A taxonomy of these approaches in terms of: i) computational complexity ii) cost savings(including OPEX and CAPEX), iii) scalability, iv) implementation difficulty are highlighted in Table 7. Thus, the matroid and b-matching algorithm are globally more efficient in finding good solutions in

negligible times and in scaling larger problem instances. However, we note that it is not easy to implement the b-matching algorithm (described in 9) due to the high difficulty in the implementation of the blossom inequalities (constraints 8).

Table 7: Algorithms' qualitative comparison

Algorithm	**Complexity**	**Cost savings**	**Scalability**	**Implementation difficulty**
ILP-based algorithm	Exponential	■■■■	■□□□	■■□□
b-Matching algorithm	Polynomial	■■■□	■■■□	■■■□
Matroid-based algorithm	Logarithmic	■■■□	■■■■	■□□□
Multiple knaspack algorithm	Linear	■□□□	■■■□	■■□□

6. Conclusion

In this paper, we addressed the RRH-BBU assignment problem with the objective of determining the best strategies to assign antennas demands to available edge data centers when jointly optimizing communication latency and resource consumption. For that, we proposed an exact algorithm based on ILP formulation to find optimal solutions for small and medium network sizes. The exact algorithm optimizes the resource consumption (in terms of used edge data centers) and communication latency associated for assigning antennas demands to the most appropriate edge data centers. However, this algorithm is known to not to scale for large problem instances. Therefore, we proposed three approximation algorithms : matroid-based approach, b-matching algorithm and multiple knapsack-based algorithm to meet larger number of antennas demands in negligible times.

The performance evaluation has been conducted using different simulation scenarios and a real 4G-LTE cellular network in a small region in Paris. According to several performance metrics, the simulation results have revealed the efficiency of the matroid-based approach and b-matching algorithm compared to multiple knaspack formulation (the most used approach the literature to address constrained resource allocation problems) and their ability in **rapidly finding optimal or near-optimal solutions even for large problem instances**. This also was confirmed by the numerical results when considering a real trace from a 4G-LTE cell map.

As a future work, we will consider the processing delay (compute latency) of antennas demands in edge data centers. In fact, for sake of simplicity, we only considered the communication latency (transmission delay) on the fronthaul network joining antennas and edge data centers to model our RRH-BBU assignment in the context of C-RAN. It would be very interesting to consider also the BBU processing time required to perform different BBU functions, colocated in the edge data centers. This can lead to nonlinear objective functions that should be efficiently optimized. The problem becomes more complex and requires depth studies relying on Lagrangian relaxations, for instance. Furthermore, the data traffic on the fronthaul network, which connects the antennas to the BBU pools, can be transmitted using different protocols including CPRI and OBSAI. The fronthaul network can be realized by different technologies, such as optical fiber communication, standard wireless communication, or mmWave communication [44]. The impact of these protocols and technologies can be investigated to better evaluate the performance of our proposed models and algorithms.

References

[1] M. Huang, X. Wang, K. Li, S. K. Das, A comprehensive survey of network function virtualization, Computer Networks 133 (2018) 212–262.

[2] K. Chen, R. Duan, C-RAN : The Road Towards Green RAN, Tech. Rep. V3.0 (December 2013).

[3] J. Wu, Z. Zhang, Y. Hong, Y. Wen, Cloud radio access network (C-RAN): a primer, IEEE Network 29 (1) (2015) 35–41. doi:10.1109/MNET.2015.7018201.

[4] M. Peng, Y. Sun, X. Li, Z. Mao, C. Wang, Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues, IEEE Communications Surveys Tutorials 18 (3) (2016) 2282–2308. doi:10.1109/COMST.2016.2548658.

[5] J. G. Oxley, Matroid Theory (Oxford Graduate Texts in Mathematics), Oxford University Press, Inc., New York, NY, USA, 2006.

[6] B. Korte, J. Vygen, b-Matchings and T-Joins, 6th Edition, Springer Publishing Company, Incorporated, 2018, pp. 305–324. doi:10.1007/978-3-662-56039-6_12.

[7] A. Li, Y. Sun, X. Xu, C. Yuan, An energy-effective network deployment scheme for 5G Cloud Radio Access Networks, 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (2016) 684–689doi:10.1109/INFCOMW.2016.7562164.

[8] X. Xu, J. Liu, W. Chen, Y. Hou, X. Tao, Storage and computing resource enabled joint virtual resource allocation with QoS guarantee in mobile networks, Science China Information Sciences 60 (4) (2017) 040304. doi:10.1007/s11432-016-9038-7.
URL https://doi.org/10.1007/s11432-016-9038-7

[9] L. Pu, L. Jiao, X. Chen, L. Wang, Q. Xie, J. Xu, Online Resource Allocation, Content Placement and Request Routing for Cost-Efficient Edge Caching in Cloud Radio Access Networks, IEEE Journal on Selected Areas in Communications 36 (8) (2018) 1751–1767. doi:10.1109/JSAC.2018.2844624.

[10] H. Kellerer, U. Pferschy, D. Pisinger, Knapsack Problems, Springer Publishing Company, 2004. doi:https://doi.org/10.1007/978-3-540-24777-7.

[11] M. Agiwal, A. Roy, N. Saxena, Next Generation 5G Wireless Networks: A Comprehensive Survey, IEEE Communications Surveys Tutorials 18 (3) (2016) 1617–1655. doi:10.1109/COMST.2016.2532458.

[12] R. Mijumbi, J. Serrat, J. Gorricho, J. Rubio-Loyola, S. Davy, Server placement and assignment in virtualized radio access networks, in: 2015 11th International Conference on Network and Service Management (CNSM), 2015, pp. 398–401. doi:10.1109/CNSM.2015.7367390.

[13] N. Yu, Z. Song, H. Du, H. Huang, X. Jia, Multi-resource allocation in cloud radio access networks, 2017 IEEE International Conference on Communications (ICC) (2017) 1–6doi:10.1109/ICC.2017.7997025.

[14] D. Mishra, P. C. Amogh, A. Ramamurthy, A. A. Franklin, B. R. Tamma, Load-aware dynamic RRH assignment in Cloud Radio Access Networks, 2016 IEEE Wireless Communications and Networking Conference.

[15] K. Wang, W. Zhou, S. Mao, On Joint BBU/RRH Resource Allocation in Heterogeneous Cloud-RANs, IEEE Internet of Things Journal 4 (3) (2017) 749–759. doi:10.1109/JIOT.2017.2665550.

[16] E. Aqeeli, A. Moubayed, A. Shami, Power-Aware Optimized RRH to BBU Allocation in C-RAN, IEEE Transactions on Wireless Communications 17 (2) (2018) 1311–1322.

[17] H. Holm, A. Checko, R. Al-obaidi, H. Christiansen, Optimal assignment of cells in C-RAN deployments with multiple BBU pools, 2015 European Conference on Networks and Communications (EuCNC) (2015) 205–209.

[18] K. Boulos, M. E. Helou, K. Khawam, M. Ibrahim, S. Martin, H. Sawaya, RRH clustering in cloud radio access networks with re-association consideration, in: 2018 IEEE Wireless Communications and Networking Conference (WCNC), 2018, pp. 1–6. doi:10.1109/WCNC.2018.8377287.

[19] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, S. Gosselin, Optimal BBU Placement for 5G C-RAN Deployment Over WDM Aggregation Networks, Journal of Lightwave Technology 34 (8) (2016) 1963–1970. doi:10.1109/JLT.2015.2513101.

[20] A. Santoyo-Gonzlez, C. Cervell-Pastor, Latency-aware cost optimization of the service infrastructure placement in 5g networks, Journal of Network and Computer Applications 114 (2018) 29 – 37. doi:https://doi.org/10.1016/j.jnca.2018.04.007.
URL http://www.sciencedirect.com/science/article/pii/S1084804518301334

[21] J. Tang, W. P. Tay, T. Q. S. Quek, Cross-Layer Resource Allocation With Elastic Service Scaling in Cloud Radio Access Network, IEEE Transactions on Wireless Communications 14 (9) (2015) 5068–5081. doi:10.1109/TWC.2015.2432023.

[22] M. Khan, R. S. Alhumaima, H. S. Al-Raweshidy, Reducing energy consumption by dynamic resource allocation in c-ran, 2015 European Conference on Networks and Communications (EuCNC) (2015) 169–174doi:10.1109/EuCNC.2015.7194062.

[23] Y. Zhong, T. Q. S. Quek, W. Zhang, Complementary Networking for C-RAN: Spectrum Efficiency, Delay and System Cost, IEEE Transactions on Wireless Communications 16 (7) (2017) 4639–4653. doi:10.1109/TWC.2017.2701359.

[24] NGMN, RAN evolution project backhaul and fronthaul evolution, NGMN Alliance.

[25] N. Nikaein, Processing Radio Access Network Functions in the Cloud: Critical Issues and Modeling, Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services (2015) 36–43doi:10.1145/2802130.2802136.
URL http://doi.acm.org/10.1145/2802130.2802136

[26] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, L. Dittmann, Cloud RAN for Mobile Networks - A Technology Overview, IEEE Communications Surveys Tutorials 17 (1) (2015) 405–426. doi:10.1109/COMST.2014.2355255.

[27] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, A. Azcorra, An overview of the CPRI specification and its application to C-RAN-based LTE scenarios, IEEE Communications Magazine 54 (2) (2016) 152–159. doi:10.1109/MCOM.2016.7402275.

[28] Open Base Station Architecture Initiative, BTS System Reference Document Version 2.0 (2006).

[29] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. A. Polakos, V. Srinivasan, T. Woo, CloudIQ: a framework for processing base stations in a data center, in: MobiCom, 2012.

[30] D. G. Cattrysse, L. N. V. Wassenhove, A survey of algorithms for the generalized assignment problem, European Journal of Operational Research 60 (3) (1992) 260 – 272. doi:https://doi.org/10.1016/0377-2217(92)90077-M.
URL http://www.sciencedirect.com/science/article/pii/037722179290077M

[31] S. Martello, P. Toth, Knapsack Problems: Algorithms and Computer Implementations, John Wiley and Sons 1 edition, 1990.

[32] R. E. Korf, A new algorithm for optimal bin packing, Eighteenth National Conference on Artificial Intelligence (2002) 731–736.
URL http://dl.acm.org/citation.cfm?id=777092.777205

[33] M. L. Fisher, R. Jaikumar, L. N. V. Wassenhove, A multiplier adjustment method for the generalized assignment problem, Management Science 32 (9) (1986) 1095–1103.
URL http://www.jstor.org/stable/2631537

[34] H. Holm, A. Checko, R. Al-obaidi, H. Christiansen, Optimal assignment of cells in C-RAN deployments with multiple BBU pools, in: 2015 European Conference on Networks and Communications (EuCNC), 2015, pp. 205–209. doi:10.1109/EuCNC.2015.7194069.

[35] B. Korte, J. Vygen, Combinatorial Optimization: Theory and Algorithms, 6th Edition, Springer Publishing Company, Incorporated, 2017.

[36] E. Lawler, Combinatorial Optimization: Networks and Matroids, Dover Books on Mathematics, Dover Publications, 2012.
URL https://books.google.fr/books?id=MTuoAAAAQBAJ

[37] L. Matthews, *Bicircular matroids*, Quart. J. Math. Oxford. 28 (1977) 213–228.

[38] J. Edmonds, E. L. Johnson, Matching: A Well-Solved Class of Integer Linear Programs, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 27–30. doi:10.1007/3-540-36478-1_3.
URL https://doi.org/10.1007/3-540-36478-1_3

[39] M. W. Padberg, M. R. Rao, Odd minimum cut-sets and b-matchings, Mathematics of Operations Research 7 (1) (1982) 67–80. arXiv:https://doi.org/10.1287/moor.7.1.67, doi:10.1287/moor.7.1.67.
URL https://doi.org/10.1287/moor.7.1.67

[40] M. A. Trick, A dynamic programming approach for consistency and propagation for knapsack constraints, Annals of Operations Research 118 (1)

(2003) 73–84. doi:10.1023/A:1021801522545.
URL https://doi.org/10.1023/A:1021801522545

[41] Ibm cplex optimizer, https://www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer (2019).

[42] Y. Li, H. Xia, S. Wu, C. Lu, Joint optimization of computing and radio resource under outage QoS constraint in C-RAN, 2017 International Symposium on Wireless Communication Systems (ISWCS) (2017) 107–111.

[43] Paris 4G LTE Map, https://www.anfr.fr/gestion-des-frequences-sites/lobservatoire-2g-3g-4g/lobservatoire-en-carte2/#menu2 (2018).

[44] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, F. Gutierrez, Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!, IEEE Access 1 (2013) 335–349. doi:10.1109/ACCESS.2013.2260813.