Dynamic resource allocation and network optimization in the Cloud Radio Access Network

Mohammed Yazid Lyazidi

Doctor of Philosophy
UPMC Sorbonne Universités

Specialization

COMPUTER SCIENCE
(École Doctorale Informatique, Télécommunication et Électronique "EDITE de Paris")

presented by

Mr. Mohammed Yazid Lyazidi

Submitted for the degree of

Doctor of Philosophy of UPMC Sorbonne Universités

Title:

Dynamic Resource Allocation and Network Optimization in the Cloud Radio Access Network

Defense: 27^{th} November 2017

Committee:

Mr André-luc Beylot	**Reviewer**	**Professor, IRIT/ENSEEIHT – Toulouse - France**
Mr Yacine Ghamri Doudane	**Reviewer**	**Professor, University of La Rochelle – La Rochelle - France**
Ms Ilhem Fajjari	**Examiner**	**Research engineer, Orange Labs – Châtillon - France**
Ms Anne Fladenmuller	**Examiner**	**Associate Professor, University Pierre et Marie Curie – Paris - France**
Mr Josep Mangues-Bafalluy	**Examiner**	**Senior Researcher, CTTC– Barcelona - Spain**
Mr Nadjib Ait Saadi	**Advisor**	**Professor, ESIEE Paris – Noisy-le-Grand - France**
Mr Rami Langar	**Supervisor**	**Professor, University of Paris Est – Marne-la-Vallée - France**

Acknowledgements

First and foremost I would like to express my sincere gratitude to my supervisor Pr. Rami Langar who has given me the opportunity to do my thesis and to be involved in the FUI project ELASTIC where I have learnt a lot.

A special thanks goes out to my advisor, Pr. Nadjib Ait Saadi, for his continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His excellent guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisor and mentor for my Ph.D study.

I would also like to thank all the jury members for reading my thesis, for offering their valuable time and for their constructive feedbacks.

I am grateful to Dr. Josep Mangues and Dr. Lorenza Giupponi, who provided me an opportunity to join their team in CTTC as a visiting Ph.D student, and who gave me access to the laboratory and research facilities. Without their precious support it would not have been possible to conduct this research.

My sincere thanks also goes to Pr. Paul Rubin from Michigan State University, for his collaboration, expertise and precious advices that helped me to learn more about the optimization research field.

I must acknowledge as well my friends and colleagues in the PHARE team who have been so supportive and helpful along the way of doing my thesis and for all the great times that we have shared in these three years.

And last but not least, I would like to thank my amazing family for their love, support, and constant encouragement I have gotten over the years. In particular, I am highly indebted to my wonderful mother, my sister and my brother. Without their faith in me, this thesis would never have been achieved.

Abstract

Cloud Radio Access Network (C-RAN) is a future direction in wireless communications for deploying cellular radio access subsystems in current 4G and next-generation 5G networks. In the C-RAN architecture, BaseBand Units (BBUs) are located in a pool of virtual base stations, which are connected via a high-bandwidth low latency fronthaul network to Radio Remote Heads (RRHs). In comparison to standalone clusters of distributed radio base stations, C-RAN architecture provides significant benefits in terms of centralized resource pooling, network flexibility and cost savings. In this thesis, we address the problem of dynamic resource allocation and power minimization in downlink communications for C-RAN. Our research aims to allocate baseband resources to dynamic flows of mobile users, while properly assigning RRHs to BBUs to accommodate the traffic and network demands. This is a non-linear NP-hard optimization problem, which encompasses many constraints such as mobile users' resources demands, interference management, BBU pool capacity, transmission power limitations and fronthaul links capacity. To overcome the high complexity involved in this problem, we will present several approaches for resource allocation strategies and will tackle this issue in three stages. In the first stage, a meta-heuristic algorithm using the simulated annealing will be considered in providing sub-optimal solutions to the resource allocation problem in C-RAN with an unconstrained fronthaul capacity. The goal is to provide near optimal online solutions at a much reduced complexity and in minimum time compared to offline optimization schemes. In the second stage, we will integrate different mobile users profiles and quality-of-service requirements, while considering a capacity-limited fronthaul network between BBUs and RRHs. A joint resource allocation and admission control approach is thus presented to handle this issue based on a two-stage algorithm and greedy fronthaul link selection scheme. Finally, in the third stage, we will consider a BBU-RRH assignment problem, while considering jointly several important objectives such as resiliency, operational costs, processing power and constraints on BBU processing and cost budgets. An algorithm based on the branch-and-price framework will be described to compute the optimal solution in minimum time period. Besides, our analysis will evaluate several policies and provide general guidelines that can be used by operators to decide the best optimization strategy according to their needs for their C-RAN infrastructure. Obtained results prove the efficiency of our proposed strategies in terms of throughput satisfaction rate, number of active RRHs, BBU pool processing power, resiliency, and virtualization cost.

Key Words

C-RAN, resource allocation, power minimization, admission control, quality-of-service, BBU-RRH assignment, BBU virtualization, optimization.

Table of contents

Chapter 1

Introduction

Contents

Mobile Network Operators (MNOs) are facing important growths in mobile data traffic on their networks due to the ever-increasing popularity of smartphones, tablets and new connected smart devices, which support a wide spectrum of resource-greedy applications and services. Recent reports [1] show that almost half a billion (429 million) mobile devices and connections were added in 2016. In fact, global mobile devices grew to 8.0 billion in 2016, up from 7.6 billion in 2015. Smartphones were accounted for a large part of that growth, which will carry on its volume's increase into the coming years due to the emergence of several new communication services including

Machine-to-Machine (M2M), Device-to-Device (D2D) communications and the Internet of Things (IoT). Future previsions have already been established where the number of mobile devices connected to a network, including M2M terminals, is expected to grow to 11.6 billions by 2021 and will exceed the world's population at that time (7.8 billion) [1] [2]. Besides, global mobile data traffic will reach 49.0 exabytes per month by 2021, realizing a compound annual growth rate of 47% in just five years.

On the other hand, MNOs are facing important issues with their current cellular systems to handle this traffic growth. In fact, macro cells are near their physical limitations and cause MNOs difficult and expensive plans for maintaining and upgrading them. Whereas small cells pose several challenges regarding strategy deployment, interference management and operating complexity, resulting in Capital (CAPEX) and Operating Expenditure (OPEX) costs inflation [3]. What is more, with the arrival of new fifth generation (5G) technology in multi-network environment, system designs and upgrading will become far more challenging and complex [4]. All of these issues are putting high pressures on MNOs to conceive and adopt a new cost-effective Radio Access Network (RAN).

Cloud-Radio Access Network, commonly known as Cloud-RAN or C-RAN, has been introduced by [5] as a new cloud architecture that can address the challenges MNOs are faced with and meet their requirements in terms of CAPEX and OPEX costs reduction. In the C-RAN philosophy, baseband processing is shifted away from the physical location of Base Station (BS) into a "virtual BS pool". This approach is adopted from the cloud computing concept [6], where resources are shared in a centralized data-center and allocated on demand. In the C-RAN application, baseband resources can be employed more efficiently based on the whole network's overall load instead of the maximum loads of individual BSs. Furthermore, this concept allows the processing power in the BS pool to be adapted to the network's instantaneous load.

In this thesis, we will exploit the flexibility of centralized C-RAN architecture regarding efficient baseband resource pooling and power scalability to propose dynamic resource allocation algorithms with the aim of minimizing the transmission and processing C-RAN power, and satisfying the Quality-of-Service (QoS) required by end-users with different profiles.

In what follows, we first introduce the C-RAN concept and the motivations behind its introduction. Secondly, we will describe its architecture compared to the conventional BS one. Thirdly, we will outline the different challenges C-RAN is currently facing and that hinder its commercial deployment. Afterwards, we will present the problematics of our thesis in Section 1.4. Then, we summarize our different contributions in Section 1.5, followed by a presentation of the thesis' organization in Section 1.6.

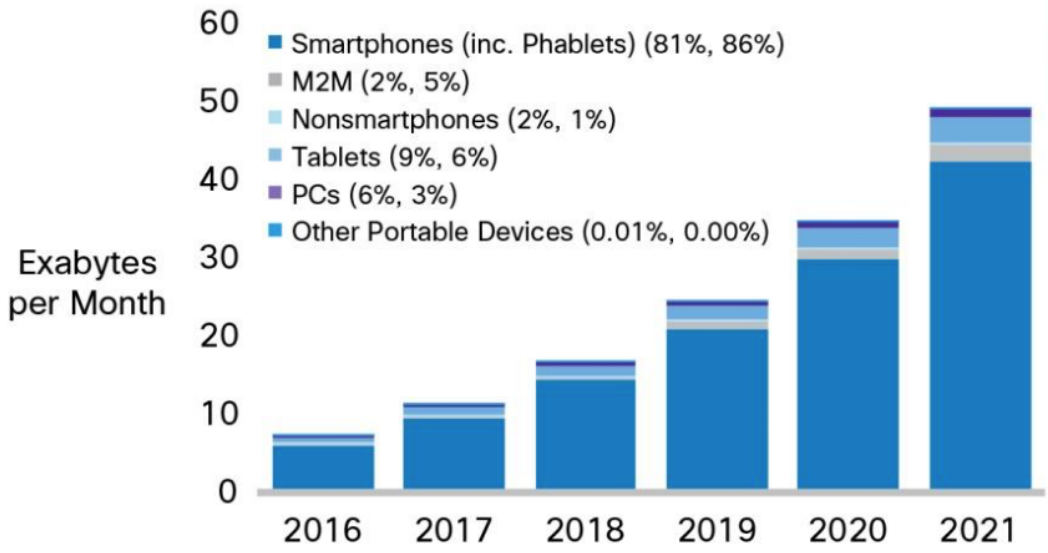


Figure 1.1 – Global Mobile Traffic Growth by Device Type [1] (numbers in parentheses refer to 2016 and 2021 traffic share)

1.1 C-RAN motivations

The premise behind the introduction of C-RAN was to propose a new cost-effective RAN solution that can cope with the current mobile traffic growth forecasted in Figure 1.1, while sustaining and alleviating the MNOs CAPEX and OPEX expenses. China Mobile Institue proposed first the idea of C-RAN in 2009 [5], that stands for Centralized, Collaborative, Cloud and Clean RAN, and that can address the aforementioned challenges. Further investigations led by Next Generation Mobile Networks (NGMN) in 2013 [7], highlighted the key technologies critical to C-RAN implementation, and promoted it as an essential direction in wireless communications for deploying current 4G cellular radio subsystems and future 5G networks.

C-RAN is based on the "Cloudification" or, in other terms, the migration of baseband processing from standalone BSs to a cloud data center composed of a pool of virtual BSs commonly known as BaseBand Units (BBUs). Thanks to this centralized network architecture, computation resources are virtualized in C-RAN; they are aggregated on a pool level and flexibly allocated on demand. This constitues the fundamental and basic feature of C-RAN. Within this cloud data center, BBUs "collaborate": they work together in a large BBU pool to share and exchange network information such as signalling, traffic information and Channel State Information (CSI) of mobile users in the system. Hence, by performing a proper load balancing between BBUs, the C-RAN can adapt itself to non-uniform traffic during the day, allowing efficient utilization of baseband resources and better

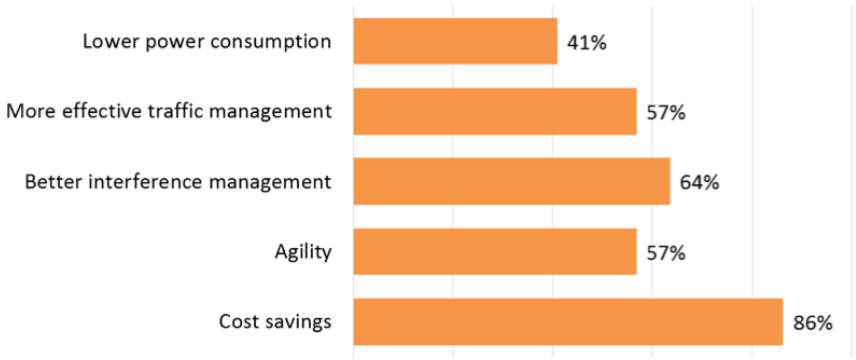


Figure 1.2 – MNOs motivations for C-RAN. Source: Monica Paolini, *Senza Fili Consulting* [9]

interference management across multiple cells [8].

By replacing "hard" wireless network equipments by "soft" BBUs, the C-RAN capabilities can be dynamically adjusted based on the traffic load. This not only fosters efficient resource utilization, but also allows the C-RAN to handle more areas than standalone clusters of BSs and facilitates service deployment on the edge. In fact, exploiting C-RAN advantages of reconfigurability and extensibility permits to move services or to directly deploy new ones on the RAN side with minimum reconfiguration and relieve the backhaul pressure. On another note, C-RAN appears as a successful way of speeding up the network construction by lowering down difficulties of site selection and civil work, since most equipments will be gathered in a central room. Site visits for maintenance and upgrades will also be reduced since most monitoring tasks can be done by software, which will contribute in more OPEX savings. What is striking, the CAPEX/OPEX cost savings brought by C-RAN constitutes the most alluring feature (Figure 1.2) that motivates MNOs for adopting C-RAN in their infrastructures [9].

C-RAN is not only applicable for existing wireless infrastructures but is also one of the key solutions for enabling future 5G systems [10]. In fact, thanks to its centralization, flexibility and cloud-based implementation, C-RAN can help MNOs foster their networks migration to LTE-Advanced Pro and meet the advent of 5G by enabling several 5G technologies such as: Large Scale Antenna Systems (LSAS), ultra-dense Multiple-Input Multiple-Output (MIMO) networks, full-duplex systems, and so on [11]. Additionally, C-RAN can help improve the resiliency of the wireless networks, paving the way towards ultra-reliable communications systems.

To discuss the different use cases for C-RAN implementation and deployment solutions, a number of C-RAN projects have been recently launched by NGMN and European Commission's Seventh Framework Programme (EU 7 FP). In 2012 started the "Interworking and JOINt Design of an Open Access and Backhaul Network Architecture for Small Cells based on Cloud Networks

(IJOIN) project [12] that introduces the novel concept of RAN as a Service (RANaaS), where RAN can be flexibly instantiated on demand from a centralized open IT platform based on a cloud infrastructure. In 2013, the "Mobile Cloud Networking" (MCN) project [13] was launched to assess the opportunities cloud computing can bring to mobile networks. In 2014, the "High capacity network Architecture with Remote Radio Heads & Parasitic antenna arrays" (HARP) project [14] was initiated for building an RRH-based C-RAN architecture with electronically steerable passive antenna radiators.

Initially steered by Japanese and South Korean MNOs, many Asian-Pacific operators have already been tempted by the C-RAN benefits and have started planning the deployment of their future networks. In fact, DoCoMO Japan and Korean SK Telecom both have announced early trials of 5G C-RAN in 2019 and for the upcoming 2020 Olympic Games [15]. Besdies, C-RAN technology has gained momentum worldwide with many operators and vendors, including Verizon Communications, AT&T, Vodafone, Orange, Intel, ZTE, Huawei and Nokia Bell Labs that are discussing the different investments opportunities of C-RAN and helping building its ecosystem.

1.2 C-RAN architecture

Traditionally, MNOs operate using the Distributed-Radio Access Network (D-RAN) architecture illustrated in Figure 1.3, which consists of two components: a BBU and a Remote Radio Head (RRH) collocated at the same macro site (or eNodeB). The interconnection between BBU and RRH is done by fiber optic cable using the Common Protocol Radio Interface (CPRI) [16]. The BBU is further connected to the Mobile Switching Center (MSC) via Carrier Ethernet backhaul where the traffic is processed.

In the C-RAN architecture illustrated in Figure 1.4, BBUs are relocated from individual cell sites to the centralized BBU hotel. The latter can be housed in a Central Office (CO) or a super macro site, regrouping many BBUs composed of high-performance programmable processors and real-time virtualization technology that perform baseband functions processing. Removing BBUs from cell sites means that operators can remove routers and other hardware too. This comes with reduced costs associated with space, heating, cooling, power and test access. What is more, the centralized topology simplifies networks management, deployment and scaling.

BBUs connection to RRHs requires a high-bandwidth low latency transport service. Dark fiber and Wavelength Division Multiplexing (WDM) are the most common used fronthaul solutions supporting CPRI to interconnect the many RRHs at each cell site to the central BBU at the BBU pool [17]. Transport reach, as limited by the strict Long Term Evolution (LTE) timing and latency requirements, is around 15 to 20 kilometers of dark fiber for the new centralized configuration [18]. This is essentially an extension from less than 60 meters at a typical eNodeB tower.

Since multiple BBUs are colocated in C-RAN, their enhanced X2 (eX2) interface can cost-

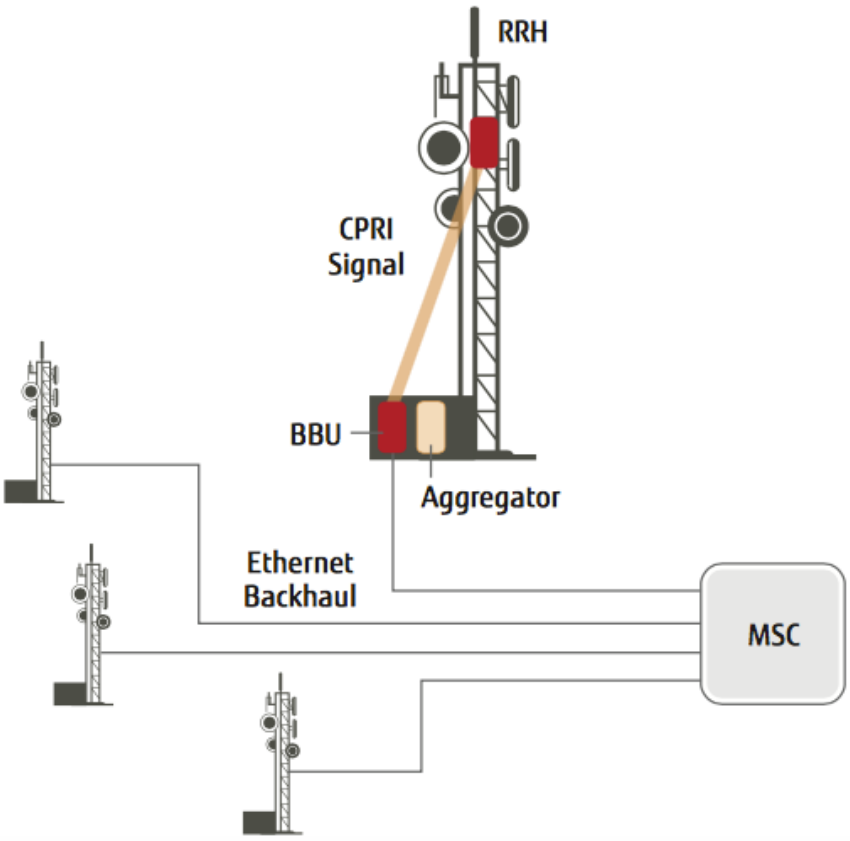


Figure 1.3 – Distributed RAN architecture

effectively interconnect resulting in improved performance of resource utilization. In fact, a cloud controller, located in the BBU pool, serves as a resource manager and performs load balancing between BBUs through eX2 to accommodate the traffic load of different RRHs [19]. Thanks to this centralized architecture, operators can deploy more RRHs in remote sites to increase their network coverage and allocate more baseband resources to RRHs in order to satisfy the traffic demand. This provides operators better flexibility in network upgrading, performed adaptability to non-uniform traffic and efficient utilization of baseband resources [20].

Furthermore, the S1 interface also comes from the BBUs to the router, providing aggregation and transport using the Carrier Ethernet backhaul connection to the MSC. A single intermediate or large-scale router can thus be used at the BBU pool instead of numerous small ones of individual cell sites as in D-RAN, further lowering equipment costs. Finally, the C-RAN topology positions the network for Software-Defined Networking (SDN) et Network Function virtualization (NFV) applications to enable dynamic resource allocation between BBUs. A server can ultimately replace

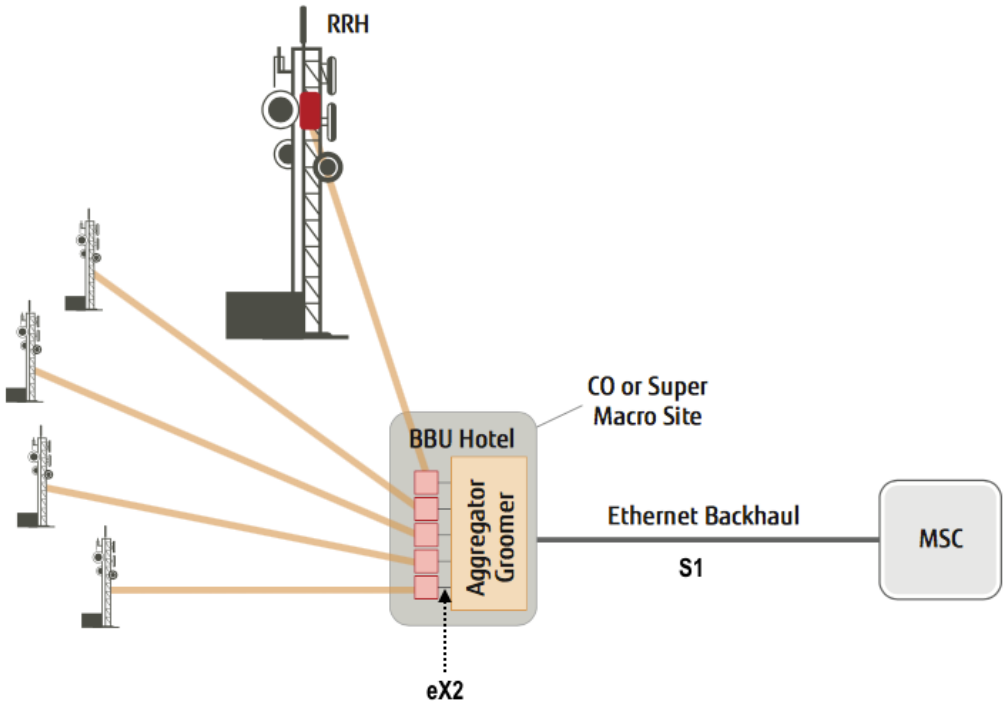


Figure 1.4 – Centralized RAN architecture

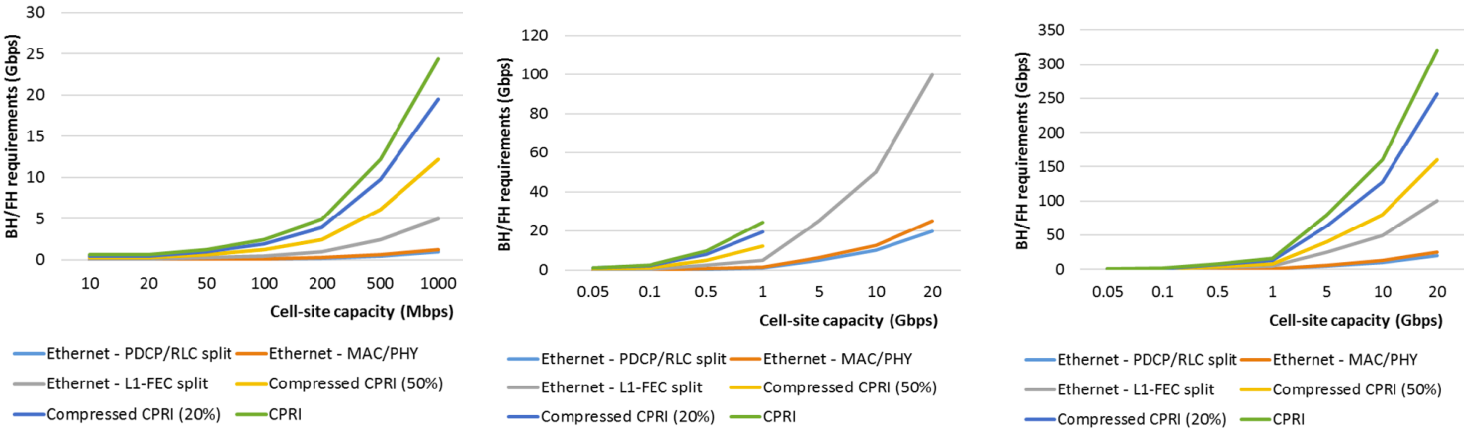
the purpose-built hardware and virtualize the BBU functionalities resulting in a virtual BBU. This means greater flexibility, even higher performance and even lower cost.

1.3 C-RAN challenges

C-RAN is designed to not only solve the cost and deployment issues of operators, but also to improve the network spectral efficiency via collaborative radio or joint processing techniques. However, while some C-RAN features are relatively easy to realize such as centralization, others require long-term planning and development. This section analyzes the major challenges for C-RAN.

1.3.1 Fronthaul limitations

In the C-RAN model, a transmission link with at least 10 Gbps and a maximum latency of 250 μs is necessary to convey the baseband data from the BBU pool to the cell site, with respect to LTE strict timing requirement [7]. As mentioned earlier, this also limits the maximum distance between BBUs and RRHs - up to 20 km, according to NGMN's specifications. Consequently, fiber connections from the cell sites are required at almost locations. In densified heterogeneous networks, the need of



(a) 4G, early 5G (b) path to 5G, current CPRI capacity (c) path to 5G, with CPRI expanded

Figure 1.5 – Backhaul and fronthaul requirements vs. cell-site capacity (Source: KDDI, Nokia, Viavi, Small Cell Forum 2016)

fiber-based links increases the CAPEX, especially for operators who do not own the fiber networks and limits the scope of C-RAN deployments.

BBUs can be kilometers away from RRHs, and fronthaul links have to serve multi-cell LTE sites that may use Carrier Aggregation (CA) or MIMO. For such complex network, CPRI can enable the transition to C-RAN topologies in 4G (Figure 1.5 (a)), however it will encounter fundamental limitations in 5G networks. In fact, as can be seen in Figure 1.5 (b), the fronthaul requirements will grow with the increase of the cell-site capacity, where next-generation cell sites will have a capacity of 10-20 Gbps and will require over 300 Gbps for CPRI. Currently, CPRI is limited to only 24 Gbps capacity, which does not make it a viable option for 5G. Expanding the CPRI capacity in the path to 5G can meet the network's requirements (Figure 1.5 (c)), however, since it heavily relies on fiber fronthaul links, its deployment will be overly expensive or unavailable.

Instead of relocating all baseband processing in a BBU pool, a functional split on the baseband processing chain can be defined to relax the most latency-sensitive functionalities in C-RAN [21] [22]. Recently, C-RAN functional splits have been introduced and are under extensive analysis to address the challenge of optimizing fronthaul bit rate and flexibility [23]. Figure 1.6 presents different use-cases of C-RAN functional splits. For instance, the PHY-MAC (also known as User-Cell) split is based on separating user and cell specific functionalities, occasioning the traffic between BBUs and RRHs to be traffic dependent. On the other hand, for the PDCP-RLC split, the majority of data processing will be distributed in the RRHs, and only a small portion will be done in the BBU pool. Such functional splits may help reduce the data bandwidth and latency requirements between the central point and remote sites. However, they may increase the CAPEX and OPEX costs since more equipment rooms will be needed to improve the system management.

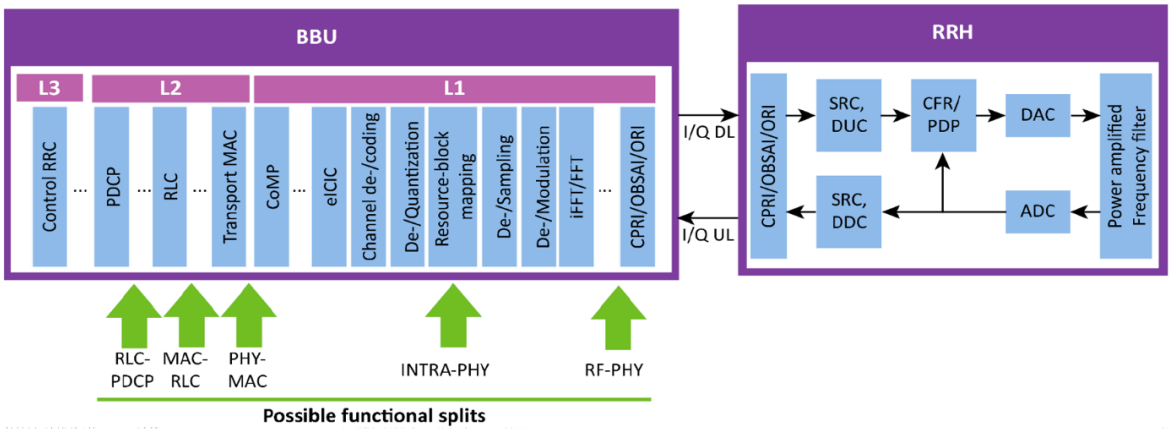


Figure 1.6 – BBU and RRH functionalities and possible functional splits

Generally, the choice of which split to activate is a multi-dimensional tradeoff between conflicting objectives [24], since it depends on the topology, the fronthaul network budget, the services MNOs want to support and their performance tradeoffs. Besides, choosing which functions to centralize and which to distribute must give MNOs the opportunity to foster the use of their network resources.

1.3.2 Complexity management

Collaborative Multipoint (CoMP) [25] and Enhanced Inter-Cell Interference Coordination (EICIC) [26] have been presented as efficient techniques for improving spectral efficiency and interference mitigation in C-RAN [20]. However, interference mitigation in centralized networks usually comes with the expense of higher complexity and system overhead due to the large number of RRHs and the different types of backhaul links associated with each cell type [27]. In fact, a challenge for C-RAN consists of implementing CoMP algorithms in BBUs for handling numerous cells with high data rates, while respecting the tight LTE requirement of 1 ms processing inside the BBU pool [12], and 3 ms of BBU-RRH transmission. These strict timing requirements are imposed by 3GPP LTE's Hybrid Automatic Retransmit reQuest (HARQ) protocol [7] [28]. To better clarify these timing requirements, we make a comparison between D-RAN and C-RAN processing delays:

- In a D-RAN environment, the eNodeB completes User Equipment (UE) data processing: (Uplink (UL) CPRI processing, frame decoding, ACK/NACK creation, Downlink (DL) frame creation, DL CPRI processing) within 3 ms [18].
- In C-RAN, additional fronthaul delays (like transmission delay via optical fiber, fronthaul equipment processing) are caused while the data is delivered from the RRH to the central

BBU pool. The sum of all these delays and baseband processing time at the BBU must be less than HARQ's 3 ms.

In order to maintain the timing requirement constraint, the additional delay caused in the fronthaul network must be compensated somehow, somewhere. Some vendors address the problem by expediting the BBU processing to finish the UE data processing and send the ACK/NACK in less than 2.75 ms instead of 3 ms, thus allowing an additional delay of 250 μs in the fronthaul links [18]. However, this imposes tighter C-RAN BBU time processing requirement than usual D-RAN to process mobile users data and monitor large-scale cellular networks. What is more, it severely impedes the implementation of centralized algorithms for multi-cell coordination, which are known for their higher complexity and processing time [29]. Besides, recent initiatives [30] [31] [32] have revealed that CoMP gains cannot be fully harnessed in C-RAN under practical constraints due to signalling and processing delays, signalling overhead, and limited backhaul capacities.

1.3.3 Resource management

Another challenge for C-RAN is the optimization of radio resources allocation for multiple cells. In fact, if the problem of Radio Resource Management (RRM) has already been immensely discussed in D-RAN for proposing diverse optimization schemes involving resource allocation (resource scheduling [33], power control [34], interference mitigation [35] and admission control [36]) under precise constraints, the C-RAN has other specific considerations; in particular, due to the fronthaul limitations and the computational complexity for finding solutions. Furthermore, the C-RAN RRM concerns also hybrid cellular networks (macro BS and small cells), which require resource sharing algorithms between the different cells' RRHs to limit the interference and improve the quality of indoor coverage [37].

Additionally, C-RAN is challenged by the design of resource allocation algorithms with respect to traffic demand, interference management and BBU pool capacity [38]. Such algorithms are needed not only to improve the spectral efficiency of users served by specific RRHs, but also to allow efficient BBU-RRH mappings. In fact, by properly assigning RRHs to BBUs based on traffic load and BBU capacity, a single BBU can be assigned to manage multiple few-loaded RRHs instead of one-one mapping. Hence, the BBU pool power consumption can be minimized with fewer BBUs being instantiated, and the OPEX costs can also be lessened.

Moreover, to optimize energy savings and meet the power limitations of an eco-friendly C-RAN, some RRHs need to be selected to be dynamically turned on/off based on the traffic evolution. By doing so, the total RRHs transmission power can be minimized as well as the number of active BBUs by less overhead in the transport network. This is a major issue for RRM algorithms for allocating resources to users while accounting the limits of C-RAN's power consumption, which does not only springs from the RRHs transmission, but also from the associated fronthaul links and BBUs.

1.3.4 BBU virtualization

BBU virtualization technologies will be massively integrated in 5G [10]. Nevertheless, the actual implementation of virtualization is more difficult, despite its conceptual simplicity. In fact, wireless communication systems are distinct from Information Technology (IT) data centers, in that wireless communications have extremely strict requirements for real-time processing and data transmission. Building a new BBU platform based on a virtualized technology different from IT data centers is crucial for C-RAN. Some efforts have been done in the direction of building a new virtualized RAN ecosystem, such as the OpenAirInterface Software Alliance [39], which is an open source implementation of fully real-time stack (eNodeB, UE and core network) on general purpose processors, combined with SDN and NFV tools to bring efficiency in C-RAN design. However, the actual implementation in OpenAir is not always easy due to a tangled development platform.

Virtualization techniques implementation for group processing between BBUs is also an important challenge as it can enable resource sharing between multiple BBU entities and UE data exchange. Besides, the interface interconnecting all BBUs should support a reliable, high bandwidth and low latency switching network for UL and DL CSI exchange. In the same context, we can cite other challenges for BBU virtualization in C-RAN, that reside in:

- Real-time data processing algorithms implementation,
- Dynamic BBU processing capacity allocation, to face the dynamic loads of cells,
- Group processing between multiple BBUs in central office to share radio resources incoming from multiple operators,
- Security issue: co-location of multiple BBUs from different cloud providers requires special security and resilience mechanisms.

Furthermore, the SDN and NFV technologies are considered as major candidates for C-RAN virtualization application and for dynamic allocation of resources between BBUs. A challenge consists then, of cost-effectively integrating these technologies in C-RAN.

1.4 Problem statement

In this thesis, we address the problematic of dynamic resource allocation and power minimization in downlink for C-RAN, considering dynamic BBU-RRH assignment. This is motivated by the existence of several constraints for designing a centralized resource allocation strategy such as:

- Mobile users traffic demands, profiles and QoS requirements,
- RRHs transmission power limitations,

- Interference handling,
- Fronthaul links capacity,
- BBU virtualization,
- Cost optimization.

What is striking, dynamic resource allocation algorithms are needed to cope with significantly fluctuating and time-varying traffic loads at different RRHs [40]. A joint optimization algorithm can help cater to the traffic demand of mobile users located in different cells and with different bandwidth requests. Such approach should also dynamically select the optimal RRHs to be turned on/off based on the fluctuating traffic, so as to minimize the total C-RAN transmission power. Incidentally, by setting the BBU-RRH connections and allocating resources to RRHs according to users' traffic profiles and volumes, the number of instantiated BBUs in the cloud can effectively be reduced, which will lead to more power and OPEX savings.

Furthermore, if we consider functional splits separating user and cell related functions instead of CPRI, the fronthaul links become constrained in terms of the baseband traffic that can be conveyed from the cloud to the RRHs. In fact, functional splits reduce the fronthaul links bandwidth to few hundreds of Mbps instead of CPRI's 10-20 Gbps. Consequently, the C-RAN capacity becomes limited to satisfy all users demands and an admission control scheme is crucial to help maximize the number of satisfied UEs while taking into account their QoS requirements, interference handling and fronthaul links capacity constraints.

The more RRHs are assigned to BBUs, the higher is the C-RAN multiplexing gain [20] [38] and the more cost savings are for operators. In this respect, the BBU-RRH assignment problem should be further investigated while jointly considering several aspects such as BBU pool resiliency, operational costs, BBU processing power, and constraints on cost budget. In fact, the problem of BBU pool resiliency constitutes a major requirement for MNOs in order to guarantee limited disruptions in network availability during the day, while respecting the budget and traffic load handling requirements. What is more, considering that BBUs are provided from different Cloud Service Providers (CSPs), each with distinct failure probabilities and processing costs, a careful decision for BBUs selection must be made to meet the operators' network and budget constraints.

1.5 Thesis contribution

In this section, we summarize the significant contributions of this thesis:

1.5.1 A survey of C-RAN resource allocation and BBU-RRH assignment strategies

We will provide an in-depth overview of the resource allocation and BBU-RRH assignment schemes found in literature. The majority of proposed policies in the literature can be categorized into either: i) addressing exclusively the C-RAN power minimization problem with a limit on the total transmission power; ii) focusing on the resource allocation and admission control of mobile users with QoS requirements subject to fronthaul links capacity constraint; or iii) proposing solutions to the BBU-RRH mappings problem subject to traffic and BBU pool capacity. A taxonomy of this kind of strategies can be based on the main following criteria: i) static or ii) dynamic approaches, and also based on the iii) achieved UEs data rates and iv) computational complexity of the used algorithm. It is worth noting that a majority of related strategies are based on offline optimization approaches. In other words, they are based on network snapshots models and do not measure the UEs time arrivals in their propositions. We will highlight some existent C-RAN dynamic optimization schemes that are applicable for online optimization thanks to their low complexity design. However such approaches usually find limitations for large-scale networks and are only applicable for low users data-rate and QoS requirements, which is impractical for future 5G networks.

1.5.2 Proposed algorithms for dynamic resource allocation in LTE downlink for C-RAN

We will present in this contribution two optimization models for i) the resource allocation and power minimization problem, and ii) the BBU-RRH assignment problem in C-RAN, considering constraints on transmission power and Signal-to-Interference-plus-Noise-Ratio (SINR) for UEs. Specifically, we will investigate how to dynamically optimize the set of active RRHs and baseband resources to serve dynamic traffic flows of incoming UEs. Due to the problem's combinatorial nature, the computational complexity is NP-hard if an exact optimal solution is to be obtained for a large-scale system [41]. To handle this issue, we will propose a meta-heuristic algorithm, based on the Simulated Annealing (SA) framework, for providing quick sub-optimal solutions to the first-stage problem at a much reduced complexity. The objective of the second stage of our contributions consists to properly assign RRHs to BBUs based on RRHs instantaneous loads and the BBUs capacity. We formulate this BBU-RRH assignment problem as a Multiple Knapsack Problem (MKP), which can efficiently be solved by commercial solvers such as IBM CPLEX [42]. This contribution is the object of a conference publication in the International Conference on Communications (ICC 2016) [43] and the forthcoming publication [44].

1.5.3 Proposed algorithms for QoS-based admission control in C-RAN with functional split

In this contribution, we will extend our study to a C-RAN with capacity-constrained fronthaul links. We will propose a joint downlink Resource Allocation and Admission Control (RAAC) algorithm, considering two types of user profiles: Guaranteed-service users – also known as Gold users, and Best Effort ones. The considered problem takes into consideration i) each type of mobile users QoS requirements in terms of Physical Resource Blocks (PRBs) (i.e., bandwidth), ii) RRHs transmission power limitations, and iii) fronthaul links capacity constraints for a user-cell functional split. The problem is formulated as a Mixed Integer Non-Linear Program (MINLP) with a non-linear logarithmic constraint, induced by the fronthaul data-rates. We will propose then a two-stage framework to solve the problem. Our design will be based on problem decomposition and solving two inter-related problems using a fixed time branch-and-cut technique. A fast greedy algorithm will also be proposed for addressing the fronthaul admission control task, This contribution is the object of a conference publication in the Global communication conference (GLOBECOM 2016) [45] and the forthcoming publication [46].

1.5.4 Proposed algorithms for cost-resilience BBU selection in C-RAN

In this final contribution, we will consider the case where a MNO has to select BBU equipments from several CSPs to run its virtualized BBU pool. We assume that each CSP's BBU is characterized by a distinct failure probability [47] and a capacity cost, that can be equivalent to content delivery network prices [48] for the services required from CSPs. We will propose in this context, a framework addressing the problem of optimal BBUs selection from several CSPs. The instantiated BBU pool should meet the MNO's expectations in terms of reliability, cost efficiency, processing power optimization and traffic load catering. We will formulate our selection problem, named Cost-Resilience BBU Selection (CRBS), as an ILP problem, designed with a weighted objective function focusing on three optimization goals: i) minimizing the BBU pool processing power, ii) maximizing its resiliency and iii) increasing the RRHs traffic load handling, subject to the virtualization's capacity and budget constraints. To solve the ILP CRBS problem, we will propose to employ the Branch-and-Price (B&P) algorithm [49], which is a combination of the Branch-and-Bound and Column Generation methods for efficiently solving large-scale ILP problems. Simulation results will demonstrate the good performance of our approach to solve the BBU selection problem for different scenarios. Additionally, our analysis will evaluate several BBU selection policies and will provide general guidelines that can be used by MNOs to decide the best BBU optimization strategy according to their needs. Note that this contribution is the object of a conference publication in IEEE 86th Vehicular Technology Conference (VTC-Fall 2017) [50] and the forthcoming publication [51].

1.6 Thesis outline

This thesis is organized as follows. In Chapter 2, we will outline the related work dealing with resource allocation strategies in C-RAN and will classify them into different categories. In Chapter 3, we will describe our Dynamic Resource Allocation in C-RAN based on Simulated Annealing (DRAC-SA). Afterwards, we will present our Resource Allocation and Admission Control (RAAC) scheme in Chapter 4. In Chapter 5, we will describe our Cost-Resilience BBU Selection (CRBS) framework and present our Branch-and-Price (B&P)-based approach. Finally, Chapter 6 will conclude the thesis and will present our ongoing and future work in this research field.

Chapter 2

Overview of resource allocation approaches in C-RAN

Contents

2.1 Introduction

Thanks to C-RAN's centralized architecture, operators can serve dynamic flows of mobile traffic with more efficient use of baseband resource and lesser operation costs than the traditional Distributed RAN (D-RAN) architecture. For this reason, the implementation of dynamic resource allocation and power minimization algorithms in the BBU pool to cope with the time-varying traffic loads of RRHs is one of the most motivating challenges in C-RAN. An interesting issue that occurs in C-RAN resource allocation is when the fronthaul links become capacity-constrained due to functional splitting. Admission control strategies can come into play to optimize the number of accepted UEs while taking into account their QoS requirements and the fronthaul links capacity constraints. On another note, as virtualization of the C-RAN advances with the emergence of 5G technologies and the migration to LTE-A, Mobile Network Operators (MNOs) have to upgrade their infrastructure to not only support higher processing capacities but also to be more resilient.

Still, a common question that arises when a MNO is faced with choosing virtualized BBUs for its cloud is: What is the best BBU selection strategy when not only resiliency, but also virtualization cost and RRHs traffic handling all come into play?

In this Chapter, we will give an in-depth overview of the different C-RAN resource allocation optimization strategies proposed in literature. The next section will present resource allocation approaches with a focus on transmission power minimization strategies in downlink. Section 2.3 will pin the problematic of admission control in C-RAN and its major related work. Afterwards, Section 2.4 will provide an overview of the BBU-RRH traffic-based assignment strategies with an emphasize on the cost and resilience dilemma in C-RAN. Finally, Section 2.5 will conclude this Chapter by summarizing all main strategies with defined metrics and performance characteristics.

2.2 Overview of resource allocation and power minimization approaches in C-RAN

The D-RAN sub-optimality

It is undeniable that the research community has known a profusion of works on resource allocation in traditional D-RAN. However it is also commonly acknowledged that distributed solutions remain sub-optimal due to their distributed nature. In fact, the sub-optimality of these resource allocation algorithms comes from the absence of a global network information of the network's cells, which forces them to only rely on a local network information at each serving eNodeB or small-cell (femtocell or picocell) [6] [52]. Some proposals have tackled the issue through information sharing [53], small-cells self-organization [54] and cognitive approaches [55]. Nevertheless, in the novel C-RAN architecture, many tasks such as resource allocation, power minimization and admission control, which were previously solved sub-optimally by distributed solutions, with only local network information, can now benefit from the cost-effectiveness brought by the centralized global perspective.

C-RAN resource allocation and power minimization proposals

There have been many active research efforts from the research community to design efficient resource allocation strategies in C-RAN system. Resource allocation algorithms are essentially designed to exploit wireless channels variations by dynamically allocating the available baseband resources to mobile users so as they could run their mobile applications [56]. Besides, the problem is often coupled with power minimization issues, as operators do not target to only serve users the best way they can, but also to lessen the power consumption and antenna power radiation to create a green and eco-friendly network [57] [58]. A common problem that arises is when we target to minimize the total C-RAN power consumption subject to users resource requests and

Algorithm 1: `Successive RRH Selection Algorithm [59]`

1: Initialize the set of inactive RRHs $\mathcal{N}^{(0)} = \emptyset$, the set of active RRHs $\mathcal{A}^{(0)} = \{1, ..., L\}$ and $i = 0$.

2: Solve the power minimization to optimality and obtain the optimal beamforming gains for each RRH r.

   a) If the problem is feasible, find the RRH $r^{(i)}$ with lowest achieved beamforming gain and update $\mathcal{N}^{(i+1)} = \mathcal{N}^{(i)} \bigcup \{r^{(i)}\}$ and $i = i + 1$, and then go to **Step 2**.

   b) If it is infeasible, denote $I = i$ and go to **Step 3**.

3: Obtain the minimum network power consumption $\mathcal{P}^{\star} = \min\{P^{(0)}, P^{(1)}, ..., P^{(I-1)}\}$.

SINR constraints. This issue has been deeply tackled in many recent C-RAN papers that we will highlight.

- For instance, authors in [59] presented a Group Sparse-based Beamforming approach (GSB), that can minimize the C-RAN total power transmitted from the RRHs as well as the power consumed in the transport network. In fact, as RRHs are deployed in cell sites, they must be connected to the BBU pool through fronthaul transport links. Consequently, the more RRHs are active, the more significant the power consumption in the transport network becomes. The authors formulated the problem as a joint RRH selection and transmit-plus-transport-network power minimization beamforming problem, with a SINR constraint at each user. They have addressed this problematic by sorting the L active RRHs in their system in an increasing order of their beamforming gain (i.e., transmitting power), and then proposing a successive RRH selection algorithm. The latter, which pseudo-code is detailed in Algorithm 1, is based on successively switching off RRHs with minimum beamforming gain until the problem becomes infeasible to minimize the whole transmission power consumption. Besides, by switching off a set of RRHs, their associated fronthaul links will turn into idle mode with a low static power, which will result in the minimization of the whole transport network power consumption.

- The GSB approach has been very successful as a benchmarking scheme and has found a lot of echo in oder works tackling C-RAN power minimization. A notable work in view of this, is [60]'s joint DL and UL power minimization through UE-RRH association and beamforming. The authors, building on the highly stringent UL requirements for supporting interactive user applications (e.g., network gaming, full high-definition video calling and real-time broadcasting), considered jointly the power minimization of DL and UL transmissions in C-RAN. In fact, they proposed a joint DL and UL UE-RRH association and beamforming design to optimize the energy consumption tradeoffs between the active RRHs and UEs, sub-

ject to UEs DL and UL QoS requirements. Their proposal is based on transforming the two DL and UL problems into two DL inter-related subproblems by leveraging on the celebrated DL-UL duality result [61]. Their proposal is based on the following:

  - The DL power minimization problem was addressed through [59]' GSB approach, where the DL UE-RRH association problem is solved by optimally assigning UEs to be served by the minimal subset of active RRHs and finding the beamforming gains to transmit from the selected RRHs to all UEs.
  - Regarding the UL problem, it was addressed by leveraging on the so-called UL-DL duality result [61], where a virtual DL transmission is established to simulate the original UL problem. The UL optimization was then converted into an equivalent DL transmission problem in C-RAN, solved using integer programming and GSB.

  [60]'s work was the first attempt to unify the DL and UL UE-RRH association and beamforming design problems. Their numerical results exhibited improvements in the network energy efficiency and power consumption tradeoffs between RRHs and UEs. However, the GSB algorithms used in [59] and [60] could not exhibit the number of BBUs required to manage the entire system. In fact, the studies were carried out separately from the cloud, without taking into account the reconfigurations that can be performed at the BBU pool level, nor the online optimization delays.

- What is more, beamforming schemes may find limitations for computing-resource sharing, as highlighted by the authors of [62]. The latter paper introduced a novel dynamic radio cooperation strategy for C-RAN to maximize the downlink weighted sum-rate system utility, which outperforms beamforming schemes used in [59] and [60]. In fact, due to the combinatorial nature of the resource allocation problems in C-RAN and the non-convexity of the cooperative beamforming design, the underlying optimization problems are generally NP-hard, and extremely difficult to solve for a large-scale network. The authors in [62] proposed a transformation approach of the original beamforming problem into a Mixed-Integer Second-Order Cone Program (MI-SOCP) with no computing-resource constraint on the accumulated data rate $R(u, w)$ of all users u in the system with beamforming gain w. They then presented a fast iterative algorithm, detailed in Algorithm 2, for finding suboptimal solutions to the original problem.

- Other attempts towards centralized resource allocation have been previously made. For instance, Wang *et al.* proposed a graph-based approach for dynamic frequency reuse in C-RAN [63]. The authors presented a load-based graph coloring method to allocate spectrum resource to each RRH dynamically depending on traffic demands as well as reducing the inter-cell interference. Their simulation results demonstrated good achieved throughputs for

Algorithm 2: `Cooperative Resource Sharing (CRS) design [62]`

1: Solve the relaxed SOCP problem to find the optimal beamformers weights $\bar{w}$.
2: Verify the dropped maximum computing-resource constraint.

- If the dropped constraint is verified, return the optimal solution $w^* = \bar{w}$.
- Otherwise: Drop users' rates using a greedy algorithm.
  - *Repeat*: Update Rate of UE u $R(u, \bar{w}) = R(u, \bar{w}) - \tau$, where τ is a small decreasing step.
  - Go to the next UE when $R(u, \bar{w}) = 0$.
  - *Until*: the maximum data rate computing-resource threshold is achieved.
  - Return $w^* = \bar{w}$.

both cell-center and cell-edge UEs. Furthermore, in [64], authors proposed to minimize the whole energy cost in C-RAN while of optimizing the end-to-end performance of Mobile Cloud Computing (MCC) UEs. By adopting a decision-theoretic approach, they formulated the response latency experienced by A MCC UE as a constraint to solve the problem with respect to C-RAN timing requirement. Their proposal yielded to generally good performances in cloud energy savings for a small-scale network.

- In [36], authors presented the QoS-based resource allocation framework, which is a centralized approach for resource and power allocation in femtocells networks. In their proposal, cooperation between neighboring RRHs is exploited to improve resource allocation and throughput satisfaction via power minimization. Within each cluster of RRHs, a joint resource and power allocation is centralized at a cluster-head that periodically optimizes the throughput satisfaction. Later in [34], the authors presented a novel approach based on cooperative game theory to address the problem of interference mitigation. In their proposal, cooperation between neighbouring RRHs is exploited based on interference maps detection. Solutions from coalitional game theory, such as Nucleolus and the Shapley value, were adopted to solve the problem to optimality. However, in the latter, dynamic resource allocation considering online optimization delay was not taken into account, and the algorithm's computational time was fairly high.

2.3 Overview of resource allocation and admission control approaches in C-RAN

In the centralized C-RAN deployment presented in the survey [17], BBUs are housed in a centralized macro-cell room owned by the operator and interconnected to RRHs at the cell sites via dark point-to-point fiber. The protocol used over the fiber to meet the stringent latency and jitter requirements of LTE HARQ is usually CPRI [16] or OBSAI [65]. This poses several challenges regarding the fronthaul deployment of C-RAN solution, as highlighted by [11] and [20] since it may increase the global CAPEX expenses for expanding a fiber-access and high-capacity fronthaul links from the BBU pool to the cell site. A solution would be based on defining a tradeoff between the required centralized processing of the BBU pool and the fronthaul capacity [38], by splitting the functionality of the BBU at different layers. The partially-centralized C-RAN introduced by [22] has presented a family of flexible functional splits where parts of the BBU's PHY and MAC layers are distributed at the RRH end.

For instance, functional splits, such as the UE-Cell and PDCP-RLC examples in Section 1.3.1 of the previous Chapter, can help relax the stringent latency requirement of C-RAN fronthaul infrastructure [21]. They can enable traffic dependent bandwidth adaption to the actual traffic served in the cell, unlike the CPRI protocol, which is constant bit rate. However, the fronthaul network is limited in terms of bandwidth it can actually convey for the required latency. What is striking, the study of fronthaul-constrained C-RANs is steadily coupled with admission control problems, due to the induced fronthaul links capacity limit to support all UEs data throughputs. Therefore, the design of C-RAN wireless communications schemes should account for the fronthaul links capacity constraints.

- In view of this, Abdelnasser *et al.* [66] proposed a joint multi-channel allocation and admission control optimization framework for a two-tier cellular C-RAN under fronthaul limitations. In their work, the authors considered the DL of a cellular two-tier network composed by a single macrocell overlaying a number of RRHs, as illustrated in Figure 2.1. The resource allocation problems for each tier of the macrocell and RRHs are formulated as optimization problems described in the following:

    1. **Macrocell tier:** Being conscient of the C-RAN RRHs presence, the macrocell maximizes the sum of interference levels that it can tolerate from the lower tier, subject to macrocell maximum power and QoS constraints of the Macrocell UEs (MUEs), which are the users served by it.
    2. **C-RAN RRHs tier:** The RRHs try to maximize the number of its supported users, called Small cell UEs (SUEs), while minimizing the total DL transmission power subject to RRHs power budget, SUEs' QoS requirements, interference thresholds for MUEs,

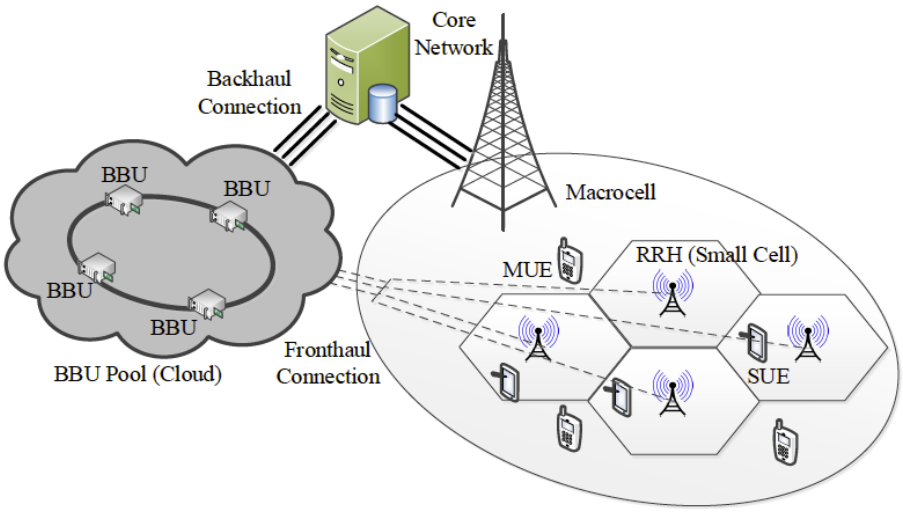


Figure 2.1 – Two-tier network with small cells deployed in a C-RAN architecture within the coverage area of a macrocell [66].

and fronthaul constraints.

Both tiers problems are shown to be MINLPs and NP-hard. To address the issue, the authors first present a matrix reformulation of the original macrocell tier problem and propose to drop the induced rank-one constraint that hinders the convexity of the reformulation. The resulting problem is a Semi-Definite Program (SDP), which is efficiently solved by interior point method [67]. They then present, for the second C-RAN tier, a low complexity algorithm based on allocating the remaining resources to SUEs that have high channel gains and that will cause low interference to MUEs first. The pseudo-code of the Sub-Channel Allocation and Admission Control (SCAAC) algorithm is detailed in Algorithm 3.

The authors further improved their proposition in [68] by working on the relaxation of the SDP problem and presenting a successive convex approximation algorithm. Even though their proposals in [66] and [68] both provided low complexity solutions for the addressed problems in DL transmission, their experimental results revealed their algorithms' limitation to only low QoS threshold regime and users data-rate requirements. In fact, the percentage of total admitted users showcased in [66] is less than 50% for a maximum QoS threshold of 11 dB and resource request of 3 sub-channels. This is quite underwhelming for current LTE cellular systems, where the QoS requirements are expectedly higher (not to speak of future 5G data rates). A similar observation can be made in [68], where the target user data-rate

Algorithm 3: `SCAAC algorithm [66]`

Given the sets $\mathcal{N}$ and $\mathcal{N}'$ of total sub-channels and sub-channels to that are not allocated to MUEs, respectively.
for each SUE f **do**
  Initialize $\mathcal{N}_f = \emptyset$, $\mathcal{A} = \mathcal{N}'$
  repeat
    Find best sub-channel with highest gain n
    $\mathcal{N}_f = \mathcal{N}_f + \{n\}$, $\mathcal{A} = \mathcal{A} - \{n\}$
  until All sub-channels in the set $\mathcal{A}$ are allocated or the QoS requirement for SUE f is satisfied
  if All SUEs have their QoS requirement satisfied **then**
    Terminate
  else
    given the set $\mathcal{N}'' = n : n \in \mathcal{N} - \mathcal{N}'$
    for all SUEs f **do**
      Initialize $\mathcal{A} = \mathcal{N}''$
      repeat
        Find best sub-channel with highest gain **and** lowest interference on MUEs $\dot{n}$
        $\mathcal{N}_f = \mathcal{N}_f + \{\dot{n}\}$, $\mathcal{A} = \mathcal{A} - \{\dot{n}\}$
      until The QoS requirement for SUE f is satisfied
    end for
  end if
end for

requirements remain fairly low for next-generation wireless standards.

- In [69], the authors considered the problem of joint rate and fronthaul resource allocation for UL C-RAN transmission. Their goal is to determine the best UEs transmission rates that have to be conveyed, over a capacity-limited fronthaul network, from the RRHs to the BBU pool. Specifically, the authors presented a Fronthaul-and-Computation-Constrained Sum Rate Maximization (FCCRM) design, aiming at maximizing the system sum-rate through optimal allocation of baseband resources and fronthaul capacity to RRHs. To this end, they have proposed a two-stage approach to optimize the integer variables related to the rate-fronthaul allocation:
  - In the first stage, a relaxation of the integer variables was done to attain a relaxed version of the problem, which is then solved efficiently via a pricing-based method.
  - The second stage consists of developing an iterative algorithm in which the pricing parameter is updated and a then a new version of the problem is solved at each iteration.

  A comparison of the achieved sum-rate gains was done with respect to a standard greedy al-

Algorithm 4: `Fast Greedy Algorithm (FGA) [70]`

1: Initialization: Set removal set $\mathcal{S} = \emptyset$
2: Solve the admission control problem for all UEs without the fronthaul constraint.
3: If it is infeasible to maintain all SINR constraints, remove the UE having the weakest channel gain to its nearest RRH, update $\mathcal{S}$, and go back to step 2.
4: Otherwise, if all SINR constraints are verified, then

a) Calculate δ for all RRHs.

b) Drop the weakest links from the RRH to its active UEs that achieves the smallest δ.

c) Solve the problem with the current allocation of RRHs for active UEs

d) If the new solution can maintain SINR constraints for all active UEs then terminate.

e) Otherwise, remove the UE requiring the highest transmission power, update set $\mathcal{S}$, and go back to step 2.

gorithm. However, the performance gains for the achieved data rates were not really sizable: only 9% of marginal gap is between FCCRM and the greedy approach.

- On the other hand, authors in [70] considered the DL C-RAN admission control by proposing a DL coordinated beamforming scheme for fronthaul-constrained network. Their proposition aims at minimizing the total transmission power subject to user QoS and maximum RRH power constraints. They introduced a convex relaxation of the admission control problem into a SDP that is solved iteratively until convergence. Moreover, since solving SDP problems generally takes a fair amount of time, they presented a Fast Greedy Algorithm (FGA), which is based on calculating the relative contribution δ of a channel link between a RRH and UE to the SINR of this UE, and then gradually removing users from the system until all constraints can be satisfied. The pseudo-code of the FGA scheme is detailed in Algorithm 4.

- FGA schemes have been commonly used in literature to address the problem of admission control in C-RAN with fronthaul constraint, due to the computation effort required to find quick solutions for these types of problem. This approach has not only been used in [70], but also in works such as [71] and [72] who presented similar approaches for C-RAN admission control and total DL power minimization over a capacity-limited fronthaul network. Recently in [73], authors addressed the sum-rate maximization problem in C-RAN, with joint QoS and fronthaul constraints. Under 5G data-rate practical assumptions and by ignoring some constraints, they proposed a fast algorithm based on Quasi-Geometric-Programming (QGP) that can solve the problem with low complexity compared to conventional convex optimization solvers. However, their solution can hardly be applied in resource allocation problems, where

the number of variables is fairly high, and thus increase the computation time. Besides, works such as [66], [59], [68] and [70] are based on snapshots models of wireless cellular networks and do not measure the network's traffic time arrivals. Their physical layer designs must include stochastic and time-varying constraints to consider the fluctuating traffic variation of mobile users and to be applicable in a real-time context.

2.4 Overview of BBU-RRH assignment approaches in C-RAN

In the detailed survey presented by [20], the authors stressed out C-RAN's agility benefits for operators compared to Distributed RAN BSs. What is striking, while more BSs are needed in D-RAN to increase mobile traffic coverage and handling, some of them are under-utilized during certain hours in the day that correspond to low traffic load (day time for residential areas and night time for office ones), resulting in ineffective use of baseband resources.

- C-RAN can address this issue, as motivated by the authors in [38], by maximizing the so-called Statistical Multiplexing Gain (SMG), that is the ratio of sum of single BBU capacity to the capacity required in the BBU pool. In [38]'s C-RAN scenario with 200 RRHs - 100 covering office areas and 100 residential ones, - the authors presented the BBUs savings that can be achieved through SMG maximization compared to D-RAN. In fact, by dynamically setting the BBU-RRH connections and allocating resources to RRHs according to users' traffic profiles and volumes, cost reduction and energy savings can be realized due to the reduction in the number of BBUs and effective utilization of the baseband resources. This is highlighted in Figure 2.2, where only 71 BBUs are needed to handle both types of areas' RRHs during the day; saving up in total 129 BBUs than the D-RAN setup.

- There have been other approaches tackling the BBU-RRH assignment problem in respect to traffic demand:

  - Based on the traffic analysis of Tokyo's metropolitan area, the authors in [74] proposed a new "Colony" RAN design that can dynamically change the BBUs-RRHs connections with respect to traffic demand. Their proposal showed that, by setting up dynamic BBU-RRH mappings, the number of BBUs can be lessened by maximum 75% compared to the traditional RAN architecture. This, nevertheless, remains a rough estimation by the authors themselves.
  - In [75], the same authors as in [74] further enhanced their proposal by introducing two BBU-RRH switching schemes for C-RAN: Semi-Static (SS) and Adaptive. The SS algorithm, which pseudo-code is described in Algorithm 5, determines the combinations of BBUs and RRHs to accommodate peak hour traffic load for all RRHs within a large

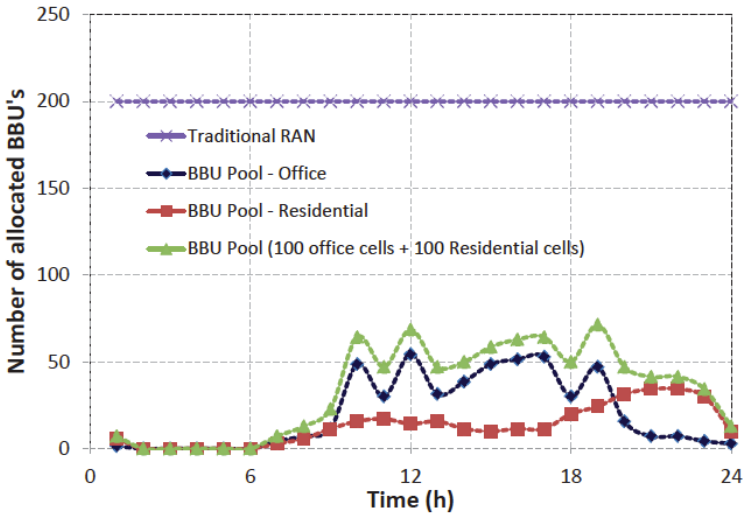


Figure 2.2 – Number of BBUs allocated during a day, for 100 office and 100 residential RRHs in C-RAN compared to D-RAN. Source: Checko *et .al* [38].

time interval. In contrast, their adaptive scheme maps RRHs to BBUs based on the SS algorithm as well as neighboring RRHs loads and BBUs resource usage limits within a short time interval (one hour). The authors demonstrated that under a traffic distribution in an office area, the number of BBUs can be reduced by 26% and 47% for the semi-static and adaptive schemes, respectively.

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. Besides, it is unequivocal that very few research works have addressed the problem of cost-resilience BBU selection in C-RAN, or more generally the problem of resiliency in C-RAN (not to speak of the cost of instantiating the BBU pool). Some related work have adumbrated these issues by far, for instance:

- In [76], the authors studied the problem of cloud resilient BBU-RRH assignment with virtualized BBU placement in metropolitan infrastructures. To solve the problem, they presented three approaches for resilient assignments:
  - *one-to-one* BBU-RRH protection;
  - *one-to-one* BBU-RRH protection and virtualized BBU protection;
  - *partial* BBU-RRH protection.

Algorithm 5: `Semi-Static (SS) BBU-RRH switching algorithm [75]`

Initialization:

- i: ith RRH $(1 \leq i \leq I)$
- j: jth BBU $(1 \leq j \leq J)$
- $R(i)/R(j)$: Used resources in RRH i or assigned resources to BBU j
- $j = 0$

for each RRH i **do**
  if no BBU is assigned to RRH i **then**
    Insert RRH i in queue.
    while queue is not empty **do**
      Search RRH t that used resources is the most in the queue.
      if $R(j) + R(t) \leq UpperLimit$ of the queue **then**
        RRH t is assigned to BBU j.
        Update resources of BBU j: $R(j) \leftarrow R(j) + R(t)$.
        Insert neighboring RRHs to RRH t that have not been inserted yet in the queue.
      end if
    end while
    Remove RRH t from queue.
  end if
  $j \leftarrow j + 1$
end for

They formulated all three propositions through ILP formalisation and compared them with non-resilient BBU-RRH mapping. The results showed that the *one-to-one* BBU-RRH and virtualized BBU protection scheme provides the highest resiliency for processing and network failures, followed by the low-cost *partial* BBU-RRH protection approach.

- Meanwhile, authors in [77] presented a BBU virtualization scheme that aims at minimizing the total BBU pool power consumption subject to processing capacity constraint and with a linear computational complexity order. They tackled the problem via a simulated annealing-based heuristic to find near-optimal results in minimum time. However, the resiliency and processing cost aspects were not considered, and, if taken into account, the simulated annealing heuristic may find limitations for such scenario.

- In [78], Bouet *et al.* proposed a novel scheme for virtualized BBUs placement in a cloud infrastructure to meet the operational cost constraints such as licence fees and/or power consumption. In fact, the placement of virtualized BBUs in cloud hotels realizes a tradeoff between possibly conflicting goals and a careful optimization design is needed. To address

the issue, the authors proposed first to cast the problem as a multi-commodity flow ILP problem. They then devised a centrality-based greedy heuristic that proved to run in polynomial time, and assessed its validity by comparing its performance gains in terms of processing time and cost with respect to the ILP optimal solution.

- In [79], authors proposed a similar scheme to model the problem of BBUs placement in COs as an ILP optimization problem, subject to traffic demand and a Multi-Stage WDM-Passive Optical Network (PON) backhauling, which architecture is described in [80]. Their goal is to decide what are the best COs to place within BBUs, with respect to the routing latencies and baseband assignment of traffic demands, so as the number of COs can be minimized. The authors mentioned a random multi-stage tree topologies generation algorithm to solve the underlying ILP problem, however, the algorithm's details were skipped in their paper. Besides, both [78] and [79] proposals consist only of early work studies, which have not fully exploited the flexibility of centralized C-RAN architecture regarding efficient baseband resource pooling and its ability to benefit from cost-efficient optimization.

2.5 Summary

Table 2.1 presents a comprehensive survey of the aforementioned C-RAN resource allocation algorithms found in literature. A taxonomy of these strategies in terms of: i) objective functions ii) constraints, iii) used algorithm, iv) application, iv) achieved UE throughput and v) computational complexity are highlighted.

Table 2.1 – Comparison of C-RAN resource allocation strategies

Ref.	Objective	Constraints	Solution	Application	Achieved UEs data rates	Computational complexity
[59] [60]	Total DL power minimization	- Max power - SINR	GSB	Static	Low	High
[62]	DL Sum rate maximization	- Max power - Computing res.	CRS	Dynamic	High	Low
[63]	Spectrum allocation	- Traffic demand - SINR	Graph-coloring	Dynamic	High	-
[64]	Total DL power minimization	- MCC UE response latency	Decision theoritic	Static	Low	-
[36] [34]	Total DL power minimization	- Max power - SINR	Game theory	Static	High	-
[66] [68]	- Admission Control - Maximizing RRHs interf. levels	- Fronthaul - UE QoS - Max Power	-SDP -SCAAC	Static Static	Low Low	High Low
[69]	UL Sum rate maximization	- Fronthaul - BBU capacity	FCCRM	Static	Low	Low
[70] [71] [72]	- Admission Control -Total DL power minimization	Fronthaul - UE QoS - Max Power	- SDP - FGA	Static Static	Low Low	High Low
[73]	DL Sum rate maximization	- Max power - Fronthaul	QGP	Static	High	Low
[38]	BBU-RRH assignment	- BBU capacity	SMG	Dynamic	-	-
[74]	BBU-RRH assignment	- BBU capacity	Colony-RAN	Dynamic	-	-
[75]	BBU-RRH assignment	- BBU capacity	- SS - Adaptive	Dynamic	-	-
[76]	Resilient BBU-RRH mapping	- links protection & survivability	Simplex algorithm	Dynamic	-	Low
[77]	-BBU processing power minimization -BBU-RRH assignment	- Single BBU capacity - BBU pool total capacity	Simulated annealing	Dynamic	-	Linear
[79]	Virtualized BBUs hoteling	- Traffic demand - backhaul latency	-	Static	-	-
[78]	Virtualized BBUs hoteling	- processing cost - Max power	Greedy heuristic	Static	-	Polynomial

2.6 Conclusion

This Chapter provided an overview of C-RAN resource allocation strategies in literature. First, we described the C-RAN transmission power minimization schemes. Next, we outlined the resource allocation and admission control approaches for fronthaul-constrained C-RANs. Then, we presented the different existing solutions for BBU-RRH mappings. Afterwards, we summarized all the discussed related C-RAN strategies, while outlining the main metrics for evaluating the performance of the proposed solutions in terms of algorithm application, achieved UEs data rates and computational complexity.

Considering all the above criteria clearly makes the problem of dynamic resource allocation in C-RAN atypical and challenging. Unfortunately, a majority of the surveyed works consist of offline optimization algorithms. In fact, no dynamic resource allocation strategy has been proposed including stochastic and time-varying constraints to consider the fluctuating traffic variation of UEs and their high resource demands, inherent to 4G/5G networks. Consequently, to the best of our knowledge, our study is the first attempt to present a centralized approach combining dynamic resource allocation, transmission power minimization and BBU-RRH assignment into one framework, which does not only support 5G higher rates, but is also dynamic, has low-complexity and is applicable for large-scale networks optimization.

Chapter 3

Dynamic resource allocation and power minimization in LTE DL for C-RAN

Contents

3.1 Introduction

In this Chapter, we present a novel two-stage framework to address the issues of dynamic resource allocation, transmission power minimization and BBU-RRH assignment in DL C-RAN. Specifically, we investigate how to dynamically optimize the set of active RRHs and allocated resources to

serve dynamic traffic flows of incoming UEs with varying resource demands. To do so, we propose in the first stage a Dynamic Resource Allocation in C-RAN algorithm based on Simulated Annealing (DRAC-SA), that aims at dynamically associating the best spectrum set of frequency/time resources to incoming UEs, with proper UE-RRH attachment. Based on the results of this first stage, the second one consists in computing the optimal number of BBUs required to manage the system, and appropriately assigning them to RRHs in order to handle the whole traffic load. We model this second problem using a Multiple Knapsack Problem (MKP) formulation, that can efficiently be solved using commercial standard solvers such as IBM CPLEX [42].

To gauge the effectiveness of our proposal, we compare our resource allocation design with two main existing strategies in centralized resource allocation and power control: QP-FCRA [36] and GSB [59] schemes. We also include comparisons to a greedy approach, as well as to the optimal solution returned from an offline optimization run. Additionally, we compare our second-stage solutions in terms of BBU-RRH attachment to the Semi-Static (SS) and Adaptive schemes proposed in [75].

The remainder of this Chapter is organized as follows. In section 3.2, we present the mathematical characterization of our two-stage system model. In section 3.3, we describe the DRAC-SA algorithm used to solve the resource allocation problem. Numerical results are presented in section 3.4 to illustrate the performance of DRAC-SA. Finally, section 3.5 concludes this Chapter.

3.2 System model

We describe in this section our two "Centralized-Resource Allocation & Power Minimization" (C-RAPM) and MKP optimization models. 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.

3.2.1 Centralized resource allocation and power minimization 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 i from $\mathcal{S}$. Each UE $u \in \{1, .., N\}$ requests from its serving RRH a number of PRBs N_u to run its applications [81]. 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 [33]. Each resolution can lead to different transmission powers. We define our UE-RRH attachment, PRB allocation and transmit power variables:

$$x_i^u = \begin{cases} 1, & \text{if UE } u \text{ is attached to RRH } i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.1}$$

$$y_{ik}^u = \begin{cases} 1, & \text{if PRB } k \text{ is allocated to UE } u \text{ on RRH } i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.2}$$

$$p_{ik}^u = \begin{cases} p \in \{p_1, ..., p_L\}, & \text{if } y_{ik}^u = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.3}$$

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

$$\gamma_{ik}^u = \frac{p_{ik}^u g_{ik}^u}{\sum_{j \neq i} \sum_{v \neq u} p_{jk}^v g_{jk}^u + \sigma^2} \tag{3.2.4}$$

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 multipath, frequency selectivity and domain scheduling [34]. Our objective is to find the optimal resource allocation strategy (i.e., to find the serving RRHs and PRBs allocation in downlink), to serve in a best effort way the existing UEs, while minimizing the total RRHs transmitted power and attaining the UEs requested SINRs level at each used PRB. Our C-RAPM optimization problem can be written as follows:

$$\underset{x^u, y^u, p^u}{\text{minimize}} \quad \sum_{u=1}^{N} \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} \left(\epsilon \frac{p_{ik}^u}{P_{max}} - (1-\epsilon)\frac{x_i^u y_{ik}^u}{K}\right) \tag{3.2.5}$$

$$\text{subject to} \quad \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} x_i^u y_{ik}^u \leqslant N_u, \ \forall u \tag{3.2.6}$$

$$\sum_{u=1}^{N} \sum_{k \in \mathcal{K}} p_{ik}^u \leqslant p_{max}, \ i \in \mathcal{S} \tag{3.2.7}$$

$$\gamma_{ik}^u \geqslant y_{ik}^u \Gamma_k^u, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \tag{3.2.8}$$

$$p_{ik}^u \geqslant y_{ik}^u p_{min}, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \tag{3.2.9}$$

$$\sum_{u=1}^{N} y_{ik}^u \leqslant 1, \ i \in \mathcal{S}, k \in \mathcal{K} \tag{3.2.10}$$

$$y_{ik}^u \leqslant x_i^u, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \tag{3.2.11}$$

$$x_i^u, y_{ik}^u \in \{0, 1\}, \ i \in \mathcal{S}, k \in \mathcal{K}, \forall u \tag{3.2.12}$$

We outline in the objective function (3.2.5) 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 $(3.2.6)$ stresses out the fact that the total allocated resources for each UE cannot exceed its requested demand N_u. Conditions $(3.2.7)$ and $(3.2.9)$ are the power constraints on RRH and UE, respectively. Condition $(3.2.8)$ ensures that the received SINR is equal to the required one Γ_k^u when the PRB k is in use (i.e., $y_{ik}^u = 1$) [36]. Constraint $(3.2.10)$ ensures that two UEs attached to the same RRH cannot use the same PRB, $(3.2.11)$ imposes all $y_{ik}^u = 0$ if $x_i^u = 0$ (i.e., the RRH i is not transmitting any PRBs), and finally $(3.2.12)$ refers that y_{ik}^u and x_i^u are binary variables.

It is worth noting that the optimization problem $(3.2.5 - 3.2.12)$ is an Integer Non-Linear Program (INLP), which is NP-hard due to the quadratic objective function and the non-convex SINR constraint $(3.2.8)$ [41]. To simplify the resolution of this problem, we reformulate it as an ILP thanks to the well-known big-M method [82]. In fact, we can replace the product of the two binary variables y_{ik}^u and x_i^u by a new binary variable z_{ik}^u and add the new following constraints:

$$z_{ik}^u \leqslant x_i^u, \tag{3.2.13}$$

$$z_{ik}^u \leqslant y_{ik}^u, \tag{3.2.14}$$

$$z_{ik}^u \geqslant x_i^u + y_{ik}^u - 1. \tag{3.2.15}$$

regarding the SINR constraint (3.2.8), we find it convenient to reformulate it as follows:

$$(1 + \frac{1}{\Gamma_k^u}) p_{ik}^u g_{ik}^u \geqslant y_{ik}^u \Upsilon_k^u + y_{ik}^u \sigma^2 \tag{3.2.16}$$

where Υ_k^u is equal to $\sum_j \sum_v p_{jk}^v g_{jk}^u$. The non-convex product between binary variable y_{ik}^u and continuous variable Υ_k^u can also be linearized using the big-M modeling, as long as Υ_k^u has explicit lower and upper bounds. From the variable definition in $(3.2.3)$ and the constraint $(3.2.9)$, we can easily deduce Lwr and $Uppr$, the lower and upper bounds of Υ_k^u, respectively. Thus, the product $y_{ik}^u \Upsilon_k^u$ can be replaced by a new continuous variable w_{ik}^u and the corresponding constraints can be rewritten as:

$$y_{ik}^u Lwr \leqslant w_{ik}^u \leqslant y_{ik}^u Uppr \tag{3.2.17}$$

$$(1 - y_{ik}^u) Lwr \leqslant \Upsilon_k^u - w_{ik}^u \leqslant (1 - y_{ik}^u) Uppr \tag{3.2.18}$$

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

$$\underset{x^u, y^u, p^u}{\text{minimize}} \quad \sum_{u=1}^{N} \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} \alpha \frac{p_{ik}^u}{P_{max}} - (1 - \alpha) \frac{z_{ik}^u}{K} \tag{3.2.19}$$

$$\text{subject to} \quad \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} z_{ik}^u \leqslant N_u, \ \forall u \tag{3.2.20}$$

$$(3.2.7), (3.2.9 - 3.2.15), (3.2.17 - 3.2.18) \tag{3.2.21}$$

$$(1 + \frac{1}{\Gamma_k^u}) p_{ik}^u g_{ik}^u \geqslant w_{ik}^u + y_{ik}^u \sigma^2 \tag{3.2.22}$$

3.2.2 Multiple knapsack problem formulation for BBU-RRH assignment

In a Distributed RAN system, one BBU is entirely assigned to a single RRH to handle the total traffic load. Thanks to C-RAN's centralization, the resources of one BBU can be shared between different connected RRHs that have few traffic load [38]. 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 N_{BBU} to manage the S loaded RRHs as follows:

$$N_{BBU} = \lceil \frac{\textit{Sum of all RRHs traffic charges}}{K} \rceil \tag{3.2.23}$$

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 is to properly assign RRHs to the N_{BBU} BBUs using a MKP formulation [83], 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 one if RRH i is attached to BBU j and zero otherwise. After solving the first problem, we can compute the weight of RRH i c_i as follows:

$$c_i = \sum_{k \in \mathcal{K}} y^{\star}_{ik} / K \tag{3.2.24}$$

where $y^{\star}$ is the returned solution 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). Our BBU-RRH MKP problem can be formulated as follows:

$$\underset{r}{\text{maximize}} \quad \sum_{j=1}^{N_{BBU}} \sum_{i=1}^{S} r_{ij} \tag{3.2.25}$$

$$\text{subject to} \quad \sum_{i=1}^{S} c_i r_{ij} \leqslant 1, \; j \in \{1, ..., N_{BBU}\}, \tag{3.2.26}$$

$$\sum_{j=1}^{N_{BBU}} r_{ij} \leqslant 1, \; i \in \{1, ..., S\}, \tag{3.2.27}$$

$$r_{ij} \in \{0, 1\}, \; i \in \{1, ..., S\}, j \in \{1, ..., N_{BBU}\} \tag{3.2.28}$$

where constraint $(3.2.27)$ denotes that one RRH cannot be managed by more than one BBU. This formulated problem is a linear program, which can be efficiently solved by commercial standard solvers such as IBM's ILOG CPLEX.

3.3 Proposal: DRAC-SA Algorithm

We have previously attempted to address the C-RAPM problem formalized in $(3.2.19) - (3.2.22)$, by proposing an optimal DRAC approach based on the branch-and-cut algorithm in [43]. 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.

In this section, we will present our Dynamic Resource Allocation in C-RAN based on Simulated Annealing (DRAC-SA) algorithm with defined neighborhood search program to find near-optimal solutions to the C-RAPM in minimum time.

3.3.1 Algorithm overview

The Simulated Annealing (SA) algorithm [84] is a powerful stochastic algorithm used to solve many combinatorial optimization problems in a fixed amount of time. It has already been applied in several works in LTE wireless networks context, such as for downlink multiuser scheduling [85] and for joint transmitter and receiver optimization in multiuser MIMO-OFDM environment [86]. 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. The global optimal solution is found when the maximum temperature is sufficiently high and the cooling is done sufficiently low to explore many system states. 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 [87], 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 accepted with a probability given by: $\exp(\frac{-\Delta E}{T})$, 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 that decrease the energy.

3.3.1.1 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. What is more, the local search can be further lessened by focusing our resolution on a smaller optimization space, generated by the "core" variable z_{ik}^u. In fact, variables x_i^u and y_{ik}^u can be derived from the big-*M* reformulation constraints (3.2.13), (3.2.14), (3.2.15), and p_{ik}^u comes as a "sub-core" optimization variable, which

can be deduced from (3.2.7) and (3.2.9). We denote E_0 the cost function (or energy) of this initial solution and T_{max} the maximum annealing temperature. We also define TSR_u, the throughput satisfaction rate of UE u, which represents the ratio of the number of its allocated PRBs on its initial demand N_u.

3.3.1.2 Neighborhood search structure

Here, we define our specific neighborhood search program 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 $\hat{x}^u_i$ to other UEs untouched. For the available PRBs ($y^u_{ik} = 0$), we select the eligible ones that can be allocated to UE u based on the SINR constraint $\gamma^u_{ik} \geq \Gamma^u_k$, while determining for each one the minimal power that satisfies that constraint.

- *Step 3:* For the eligible PRBs that satisfy $\gamma^u_{ik} \geq \Gamma^u_k$, they are allocated to UE u following a Bernoulli distribution with parameter $TSR_u \times \frac{\gamma^u_{ik}}{SNR_{max}}$, where $SNR_{max} = p_{max} g^u_{ik}/\sigma^2$ 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).

3.3.1.3 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:

$$\exp(-\frac{E_n - E_0}{T_n}) \geq \delta \tag{3.3.29}$$

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

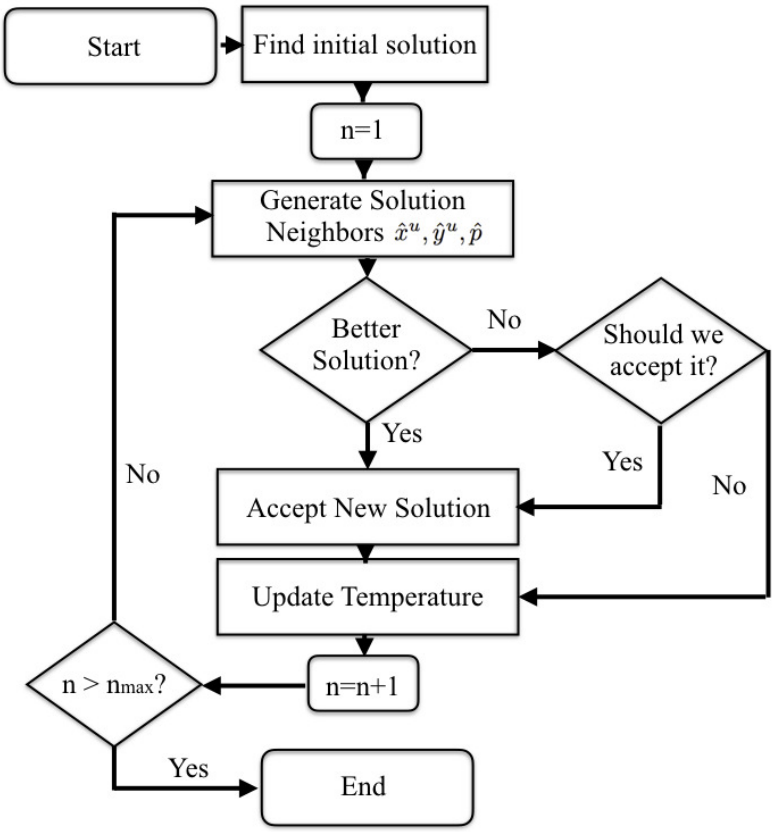


Figure 3.1 – DRAC-SA Flow Chart

.

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:

$$T_n \leftarrow \frac{T_n}{\ln(n)} \tag{3.3.30}$$

3.3.1.4 Stopping condition

Figure 3.1 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.

3.3.2 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 required BBUs N_{BBU} to handle the total traffic load (3.2.23). Since the MKP problem (3.2.25)-(3.2.28) is a linear program, we used CPLEX to compute its solution. The latter was able to find optimal results with very low 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.

3.4 Performance evaluation

Table 3.1 – 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 α	0.5
Path loss model	148.1 + 37.6$\log_{10}(d)$, d in Km
Shadowing standard deviation	5 dB
Fading model	Normal distribution $\mathcal{N}(0, \mathbf{I})$
Thermal noise	-174 dBm/Hz
Transmit antenna power gain	8 dBi
Poisson Parameter	$\lambda \in [1, 10]$ (default 5)
Departure rate	$\mu = 0.1$
UE's PRB demand	Uniform distribution $\mathcal{U}(1, 25)$
BBU capacity W	1 (100%)
Initial hot temperature	$T_{max} = 1000$
Max. number of iterations	1000

In this section, we evaluate the performance of our proposed DRAC-SA algorithm and compare the benefits of our solution with respect to state-of-the-art schemes: QP-FCRA [36] and Iterative GSB [59] algorithms for solving the C-RAPM problem. We also include comparisons to the greedy approach, which consists of the generated initial solutions of the DRAC-SA algorithm. Furthermore, we also compare to the optimal solution returned from previous DRAC approach in [43]. The latter was run in offline mode due to its high computation time for the chosen system parame-

ters. On another hand, we also compare the Semi-Static and Adaptive switching algorithms in [75] to the returned solutions of our MKP regarding the BBU-RRH assignment problem.

3.4.1 Simulation environment

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 [43]: $h_i^u = 10^{-L(d_i^u)/20}\sqrt{\phi_i^u s_i^u} g_i^u$, 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. Note that UEs' positions are randomly generated at each run and remain fixed during their whole stay time in the network. 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 3.1 reports the simulation parameters.

3.4.2 Performance metrics

In the next sections, we present the simulation results derived from our approach in terms of the following performance metrics:

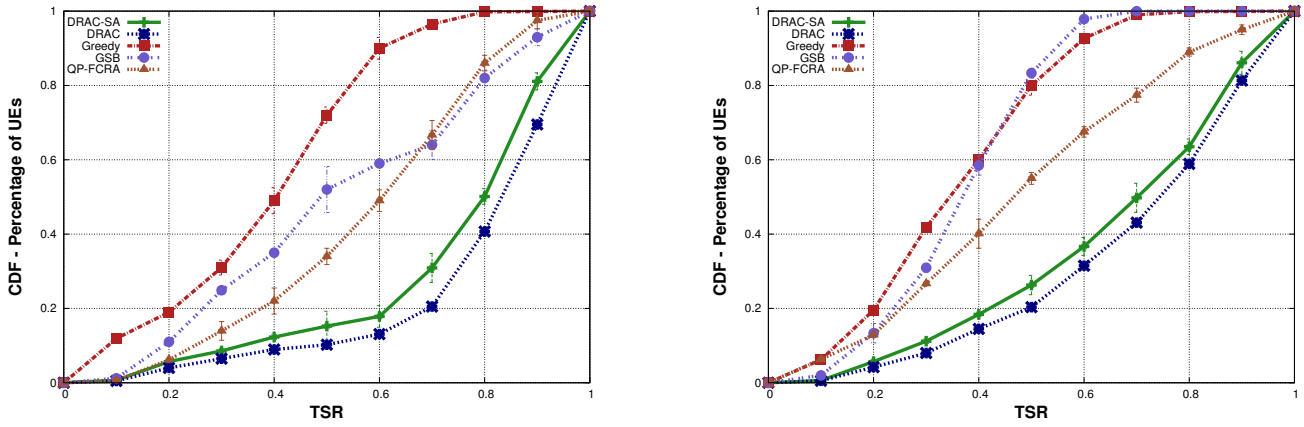
- Throughput Satisfaction Rate (TSR): As defined earlier, it denotes the satisfaction degree of a UE with respect to the initial demand N_u. For a UE u attached to RRH i, it can be expressed as follows:

$$TSR_u^i = \sum_{k \in \mathcal{K}} y_{ik}^u / N_u \tag{3.4.31}$$

- Spectrum Spatial Reuse (SSR): Denotes the average portion of RRHs using the same PRB and can be expressed as follows:

$$SSR = \frac{1}{S \times K} \sum_{u=1}^{N} \sum_{i \in \mathcal{S}} \sum_{k \in \mathcal{K}} y_{ik}^u \tag{3.4.32}$$

- Transmitted power per RRH: We compute the discrete transmission power allocated to UEs on each PRB. This latter parameter is computed to help emphasize the number of active RRHs transmitting on each power level, the porting of inactive RRHs, and the total system transmission power.



(a) SINR = 10 dB

(b) SINR = 25 dB

Figure 3.2 – Throughput Cumulative Density Function

- Number of BBUs (N_{BBU}) calculated through (3.2.23), and the corresponding number of assigned RRHs.

3.4.3 Simulation results

a) Throughput Satisfaction Rate (TSR)

Figure 3.2 shows the Cumulative Distributed Function (CDF) of the TSR. 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 Figure 3.2(a) - at low SINR threshold, and to 47% and 35%, respectively, in Figure 3.2(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 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.

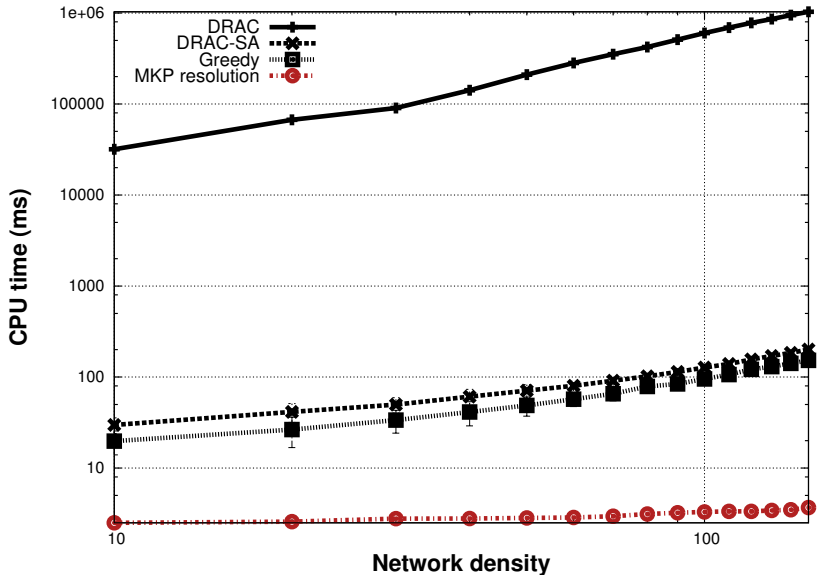


Figure 3.3 – CPU time vs number of UEs

Table 3.2 – Mean Spectrum Spatial Reuse

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

b) CPU time vs network density

As can be seen in Figure 3.3, the complexity evolution of DRAC-SA is polynomial in terms of network density, and is visibly 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 milliseconds 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.

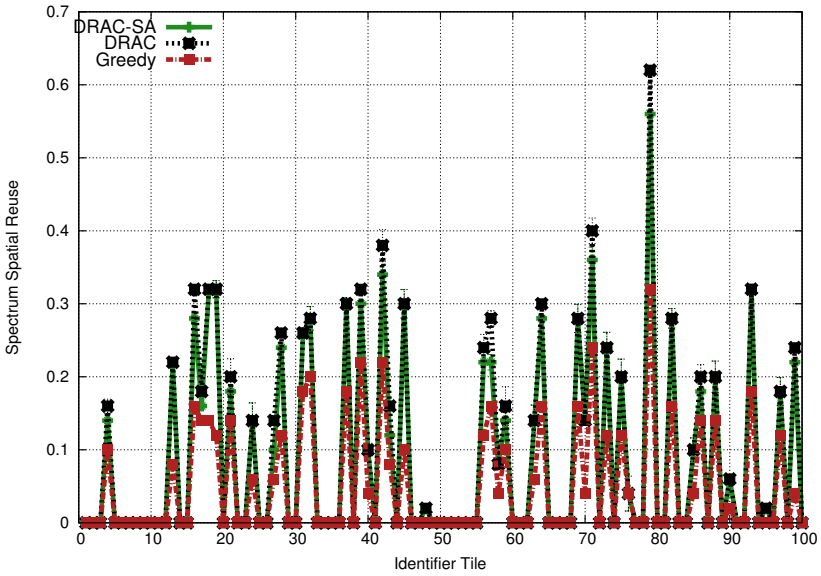


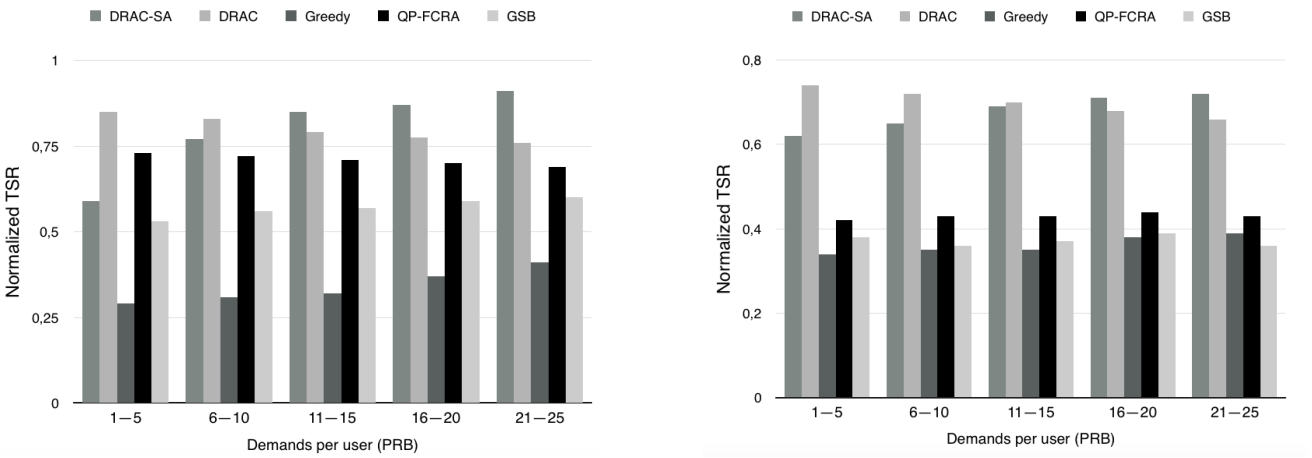
Figure 3.4 – SSR per PRB for SINR = 25 dB

c) Spectrum Spatial Reuse (SSR)

Table 3.2 reports the SSR of all aforementioned approaches. The more a PRB is reused, the better is the performance. Table 3.2 clearly shows that our proposal DRAC-SA fosters more 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. Figure 3.4 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. The confidence intervals also indicate that the DRAC-SA reuse factor can reach the optimal for most PRBs.

d) Normalized throughput vs UEs demand

In order to illustrate how the allocated resources are affected by UEs' demand volume, Figure 3.5 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 Figure 3.5(a), with an emphasis on low PRB and high PRB demand, respectively, which may imply that their resource allocation is done independently of UEs' PRBs demand. On the other hand,



(a) SINR = 10 dB

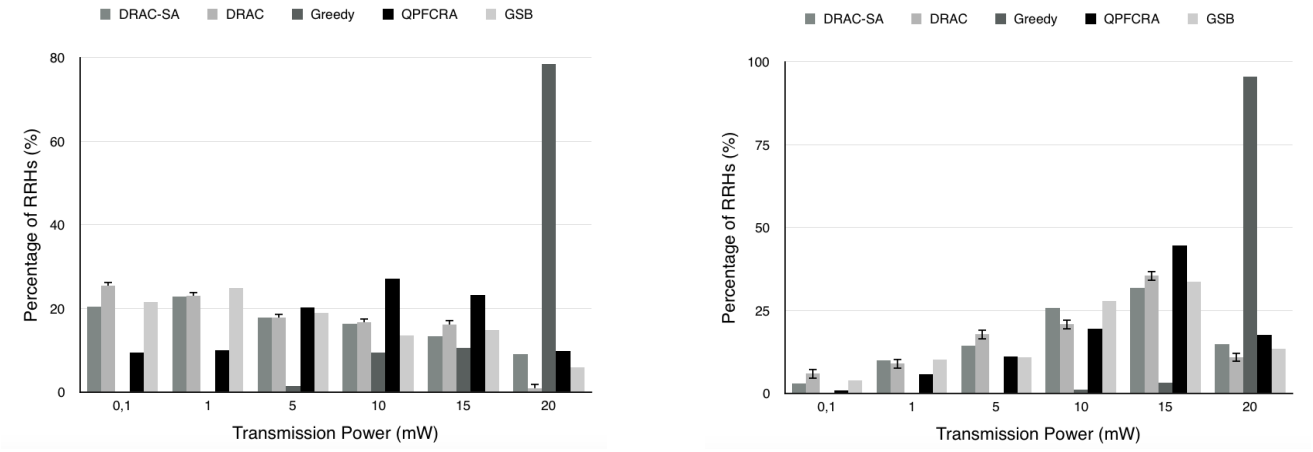
(b) SINR = 25 dB

Figure 3.5 – Throughput distribution as a function of user demands

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 (Figure 3.5(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.

e) Transmitted Power per RRH

Figure 3.6 illustrates the percentage of RRHs transmitting on each transmission power. We remark that at low SINR regime (see Figure 3.6(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 Figure 3.6(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 Figure 3.7, which presents the total C-RAN transmitted power. We can remark that the GSB scheme realizes minimum power consumption, after DRAC, by switching off RRHs based on their successive RRH selection algorithm. However, this is negatively reflected on the TSR of mobile users, as seen in Figure 3.2, since they are less satisfied by their allocated PRBs. In fact,



(a) SINR = 10 dB

(b) SINR = 25 dB

Figure 3.6 – Percentage of RRHs vs transmission power levels

while the iterative GSB method fosters more RRHs switching off, it results in less transmission power consumption in the system, but severely impacts the throughput satisfaction for mobile users due to the dynamic scaling between power p_{ik}^u and PRB allocation y_{ik}^u variables. The QP-FCRA approach, on the other hand, supposes all RRHs are turned on, which leads to a higher power consumption but to a good TSR. As observed in Figure 3.2 and Figure 3.7, our proposed DRAC-SA scheme performs a good tradeoff between satisfaction rate and overall power consumption for both SINR threshold levels.

f) Number of BBUs and on RRHs

Figure 3.8 illustrates the number of BBUs needed per time as well as the needed number of on RRHs to manage the traffic load at each instant, when the SINR threshold is equal to 25 dB. The one-one mapping in conventional RAN imposes as many BBUs as deployed RRHs to handle the radio site coverage and the fluctuating traffic load. This imposes heavy investments from operators to manage their network and increase their total BBU equipment costs. As shown in the curve, the number of BBUs calculated from the output of DRAC-SA can achieve up to 86% and 92% BBUs savings compared to a RRH-based RAN scenario and a traditional D-RAN one.

For the BBU-RRH assignment problem, we solve the MKP in (26) using CPLEX, which was able to find optimal results with very low computation time (at average 3 ms at each instant). Table 3.3 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 realizes more BBUs savings to handle the same volume of traffic load with reduced number of RRHs. This not only improves the network capacity, since many RRHs can be handled by the same BBU, but also helps maximizing

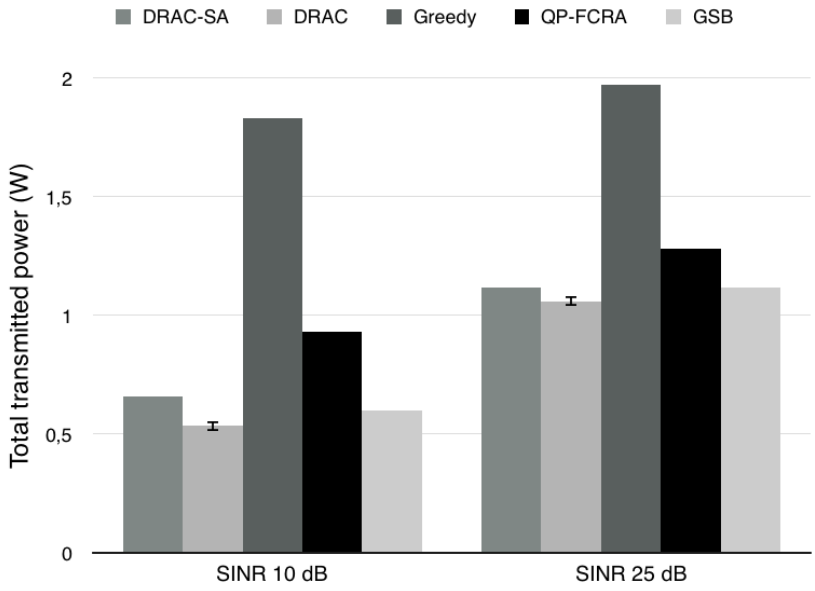


Figure 3.7 – Total RRHs transmitted power

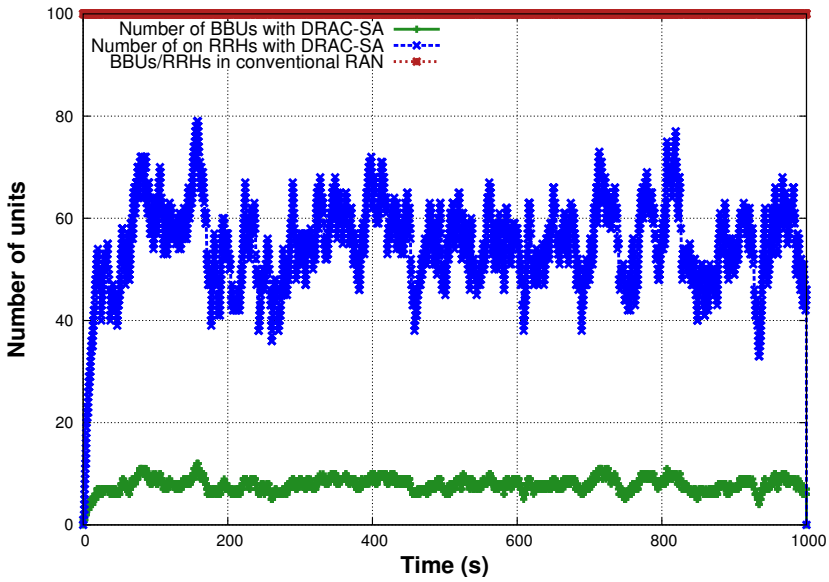


Figure 3.8 – Number of needed BBUs and on RRHs per time

the efficiency of BBUs within the virtual pool.

Table 3.3 – Number of BBUs and RRHs

Scheme	Mean N_{BBU}	Mean on RRHs	Max RRHs/BBU
MKP	7.97 ± 0.06	55.4 ± 2.2	39
Semi-static	15.1	59.7	27
Adaptive	13.5	62.4	28

g) Global TSR vs arrival rates

We vary next the arrival rates of mobile users in the system, where λ takes values in $[1, 10]$. Figure 3.9 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 Figure 3.10, 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).

h) Number of BBUs vs arrival rates

In the following, we present the variation of the number of BBUs as a function of the network density for each arrival rate. Figure 3.11 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 represents the number of on RRHs, which imposes the same number of BBUs due to the one-one mapping in D-RAN deployment. As we can observe, 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 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 his 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.

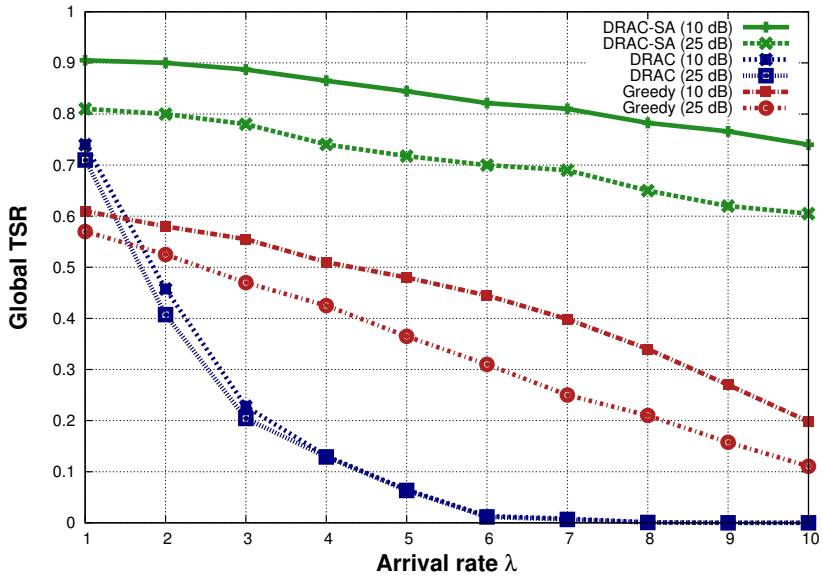


Figure 3.9 – Global TSR vs arrival rate

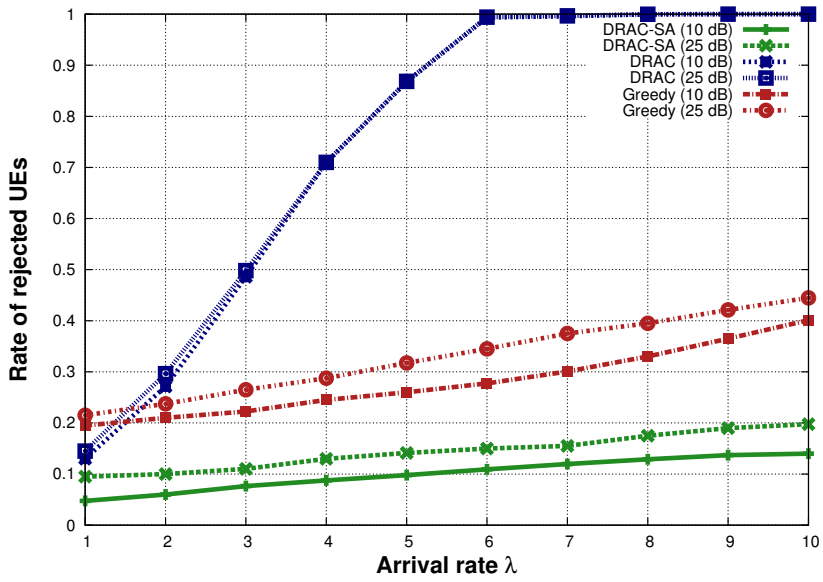


Figure 3.10 – Rejected users rate vs arrival rate

3.5 Conclusion

In this Chapter, we have presented a novel approach based on simulated annealing to address the problem of dynamic resource allocation and power minimization in C-RAN for a dynamic flow

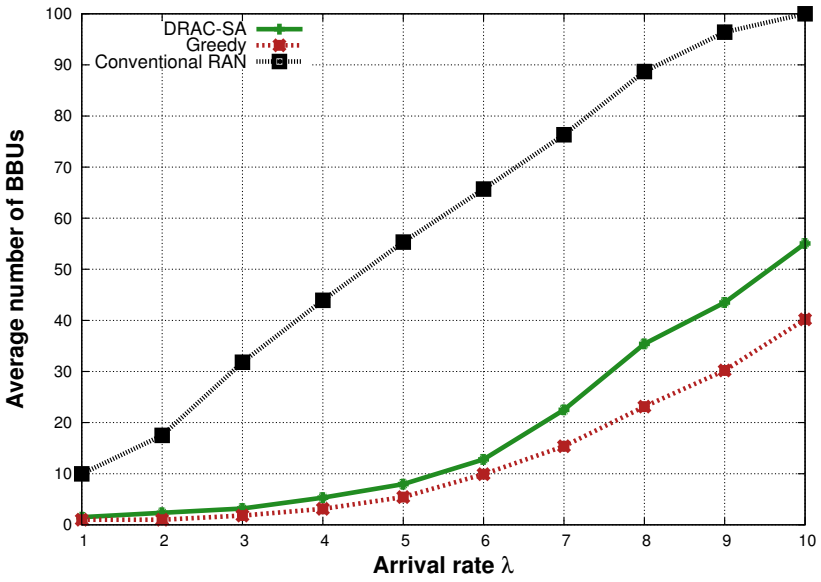


Figure 3.11 – Number of BBUs vs arrival rate

of mobile UEs. Specifically, our newly improved DRAC-SA framework can quickly find the best PRB allocation and transmission power strategy to cater to traffic demand, while satisfying individual SINR constraints and maximum power limitations. Besides, our approach can dynamically determine the optimal number of RRHs to be turned on and the number of needed BBUs to handle the whole traffic load. 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 D-RAN scenarios.

In the next Chapter, we will build on the work carried out in this first stage by integrating two user profiles in our problem: Gold and Best Effort, and considering a capacity-limited fronthaul network between BBUs and RRHs. We will present a joint resource allocation and admission control scheme that can maximize the number of accepted users in the system subject to multi-user QoS requirements, transmission power and fronthaul links capacity constraints.

Chapter 4

QoS-based resource allocation and admission control in C-RAN

Contents

4.1 Introduction

In this Chapter, we address the problem of DL resource allocation and admission control for a fronthaul-constrained C-RAN. Specifically, we consider the Resource Allocation and Admission Control (RAAC) of two sets of mobile users profiles: Gold (Guaranteed-service) and Best Effort. We define our optimization problem subject to constraints on mobile users resource demands, maximum transmission power and fronthaul links capacity. By dropping the non-linear logarithmic

constraint induced by the data-rate constraint on the fronthaul links, we propose a two-stage framework to efficiently solve the problem. We have previously attempted to solve this problem in [45] with Best effort mobile users, however the power minimization aspect was not taken into account. In this Chapter, we enhance the problem solutions by including the transmission power minimization in our design, without further increasing complexity. Numerical results from event-based simulation demonstrate the superior performance of our RAAC proposal in terms of computation time, number of accepted users and total transmission power, when compared with state-of-the-art methods used for the admission control task in C-RAN.

This Chapter is organized as follows: Section 4.2 presents the physical layer assumptions and the mathematical formulation of our system model. In section 4.3, we describe our framework and proposed algorithms to solve the optimization problem. Performance evaluation of our proposed algorithms is discussed in section 4.4, followed by section 4.5 that concludes the Chapter.

4.2 System Model

In this section, we firstly discuss the physical layer assumptions and constraints of our considered downlink OFDMA fronthaul-constrained C-RAN system model. Then, we formally describe the problem under consideration and present two formulations of it.

4.2.1 Problem formulation

A C-RAN architecture similar to Figure 4.1 with UE-cell split is deployed in an area consisting of S RRHs. Each RRH i, from the set $\mathcal{S} = \{i | 1 \leq i \leq S\}$, is responsible for the radio coverage of one hexagonal-sized cell. We denote by K the number of available baseband resources, known as PRBs, assigned by the BBU pool to all S RRHs. The cloud jointly assigns to all RRHs the same pool of resources, which means that each RRH in the system has the same capacity in terms of PRBs that can be delivered to its attached UEs. We suppose that each RRH i handles one cell in a delimited area, and that a UE u can only be served by one RRH that covers the area the UE is positioned within. We denote by $\mathcal{U}$ the set of UEs in the whole system and $\mathcal{U}^i$ the set of UEs in the coverage area of RRH i. We have $|\mathcal{U}| = \sum_{i=1}^{S} |\mathcal{U}^i|$, where $|\mathcal{U}^i|$ is the cardinality of the set $\mathcal{U}^i$.

All UEs are categorized into two different sets of user profiles: Gold users, denoted by the set $\mathcal{U}_G$, and Best Effort user, denoted by the set $\mathcal{U}_{BE}$. Each UE $u \in \mathcal{U}$ (i.e., $\mathcal{U}_G \cup \mathcal{U}_{BE}$) requests from its serving RRH a number of resources n_u^G or n_u^{BE}, for a Gold or Best-effort user, respectively. As a rule of thumb, Gold users must be served exactly the same number of requested PRBs, n_u^G, in order to run their mobile applications. Meanwhile, Best-effort users are served the best the system can regarding the other transmission constraints imposed on the C-RAN. We assume that each user device, be it Gold or Best Effort, can only be connected and served by at most one RRH based on its geographical position. Let y_k^u be the binary decision variable, which is equal to 1 if PRB k is

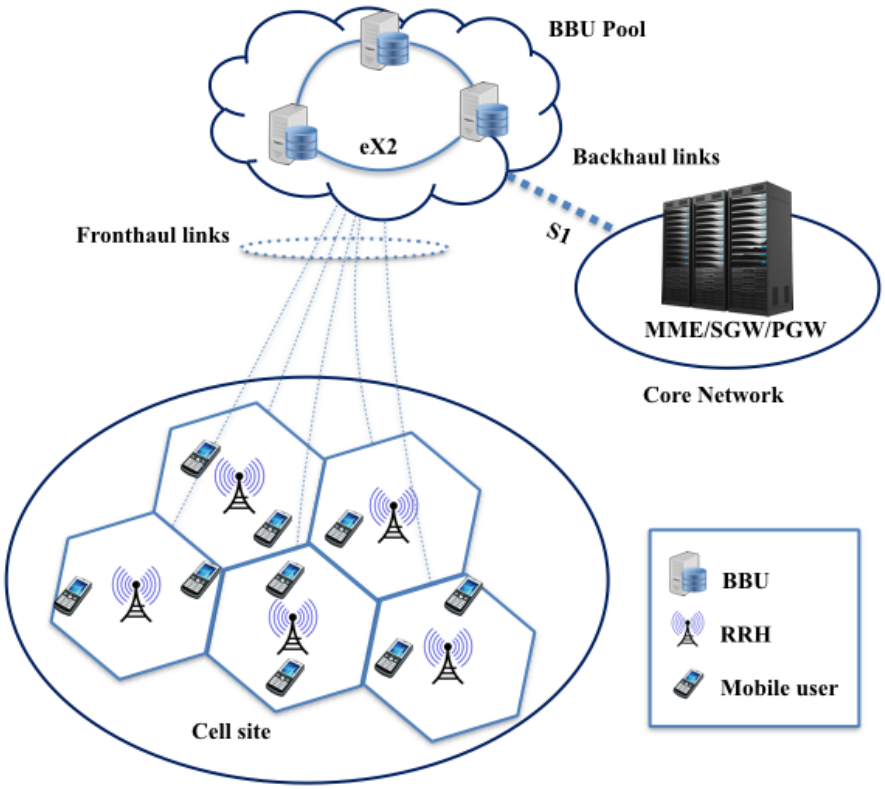


Figure 4.1 – Considered C-RAN Architecture

allocated to UE $u \in \mathcal{U}$, or 0 otherwise. We define another new binary variable z_u, which is equal to 1 if Gold user u is accepted, (i.e., it received all of the requested PRBs $\sum_{k=1}^{K} y_k^u = n_u^G$), or 0 otherwise. The transmitted power regarding the resource allocation, independently of the user's profile, is defined as follows:

$$p_k^u = \begin{cases} p \in [p_{min}, p_{max}^i], & \text{if } y_k^u = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2.1}$$

where p_{min} is the minimum power that UE u can receive, and p_{max}^i is the maximum power allowed to be transmitted from a RRH i. We define the received Signal-to-Interference-plus-Noise Ratio (SINR) for UE u on PRB k as follows:

$$\gamma_k^u = \frac{p_k^u g_k^u}{\sum\limits_{u' \neq u} p_k^{u'} g_k^u + \sigma^2} \tag{4.2.2}$$

where g_k^u is the path gain coefficient of UE u and PRB k, and σ^2 is the noise power.

We formulate our optimization problem for resource allocation in C-RAN under fronthaul constraint as the following weighted objective function problem $(\mathcal{P}_\epsilon)$:

$$\max_{z,y,p} \quad \epsilon \sum_{u \in \mathcal{U}_G} z_u + (1-\epsilon) \sum_{u \in \mathcal{U}_{BE}} \sum_{k=1}^{K} \frac{y_k^u}{n_u^{BE}} \tag{4.2.3}$$

$$\text{subject to :} \quad \sum_{u \in \mathcal{U}^i} \sum_{k=1}^{K} p_k^u \leq p_{max}^i, \forall i \in \mathcal{S} \tag{4.2.4}$$

$$y_k^u p_{min} \leq p_k^u \leq y_k^u p_{max}^i, \forall u, i, k \tag{4.2.5}$$

$$\sum_{k=1}^{K} y_k^u = z_u n_u^G, \forall u \in \mathcal{U}_G \tag{4.2.6}$$

$$\sum_{k=1}^{K} y_k^u \leq n_u^{BE}, \forall u \in \mathcal{U}_{BE} \tag{4.2.7}$$

$$\gamma_k^u \geq y_k^u \Gamma_k^u, \forall u \in \mathcal{U} \tag{4.2.8}$$

$$\sum_{u \in \mathcal{U}^i} y_k^u \leqslant 1, \forall i, k \tag{4.2.9}$$

$$\sum_{u \in \mathcal{U}^i} \sum_{k=1}^{K} B \log_2(1+\gamma_k^u) \leqslant c_{max}^i, \forall i \in \mathcal{S} \tag{4.2.10}$$

$$p_k^u \in \mathbb{R}^+, y_k^u, z_u \in \{0,1\}, \forall u, k \tag{4.2.11}$$

where ϵ is a constant optimization weight that allows to combine the utility function of the two sets of user profiles into one utility function. The utility function of the set of Gold users $\mathcal{U}_G$ is to maximize the number of accepted users who have their resource requests entirely fulfilled. On the other hand, the utility function of the set of Best Effort users $\mathcal{U}_{BE}$ is to maximize the overall throughput satisfaction rate, which is the ratio of allocated resources on the whole PRB demand. The value of the weight ϵ can be chosen arbitrarily to determine whether the operator should direct its admission control strategy to increase the number of gold users on behalf of the Best effort throughput satisfaction, or the opposite, or to possibly achieve fairness between the two sets in terms of QoS. Regarding problem $(\mathcal{P}_\epsilon)$'s constraints, constraint $(4.2.4)$ refers to the maximum power constraint on each RRH i; constraint (4.2.6) means that a Gold user must receive the exact number of requested PRBs, meanwhile constraint (4.2.7) stresses the fact that the number of allocated PRBs to a Best-effort user cannot exceed its requested number. Constraint (4.2.8) ensures that the received SINR is at least equal to the required one Γ_k^u for UE u when PRB k is in use (i.e., $y_k^u = 1$). The constraint (4.2.9) is the OFDMA constraint, which imposes that one PRB cannot be shared by multiple users served by the same RRH i. Finally, constraint (4.2.10) is the fronthaul constraint for each RRH i, characterized by a limited capacity c_{max}^i in terms of data-rate that can be conveyed through it. The total data-rate observed on RRH i is expressed using the

Shanon formula: $\sum_{u\in\mathcal{U}^i}\sum_{k=1}^{K} B\log_2(1+\gamma_k^u)$, where B is the downlink channel bandwidth. On another note, with the total accumulated traffic load observed at each RRH, we can also estimate here the number of needed BBUs in the BBU pool to manage the whole C-RAN traffic, using the same formula in (3.2.23).

4.2.2 MILP formulation

For simplification purpose, it will be helpful to rewrite the SINR constraint (4.2.8) as follows:

$$(1+\frac{1}{\Gamma_k^u})p_k^u g_k^u \geq y_k^u(\sum_{v\in\mathcal{U}} p_k^v g_k^u + \sigma^2), \forall u \in \mathcal{U} \tag{4.2.12}$$

Since variable p is explicitly bounded, the product between binary and continuous variables in (4.2.12) can be linearized using the big-*M* modelling, as performed in the previous Chapter. With this reworking, everything in problem $(\mathcal{P}_\epsilon)$ is linear, except the fronthaul constraint (4.2.10). Besides, problem $(\mathcal{P}_\epsilon)$ is a MINLP, whose solution is intractable due to its combinatorial nature. What is more, problem $(\mathcal{P}_\epsilon)$ is NP-hard [41]. We propose next, to drop constraint (4.2.10), which will be taken into account later in the fronthaul admission control algorithm, and to formulate the following MILP problem $(\mathcal{P}_\epsilon{}')$:

$$\max_{z,y,p} \quad \epsilon\sum_{u\in\mathcal{U}_G} z_u + (1-\epsilon)\sum_{u\in\mathcal{U}_{BE}}\sum_{k=1}^{K}\frac{y_k^u}{n_u^{BE}} \tag{4.2.13}$$

$$\text{subject to :} \quad (4.2.4)-(4.2.7), (4.2.9), (4.2.11), (4.2.12) \tag{4.2.14}$$

We detail in the next section, the proposed methodology to solve our MILP problem.

4.3 RAAC Proposal

In this section, we present our two-stage framework to solve problem $(\mathcal{P}_\epsilon{}')$. We present in subsection 4.3.1 the first stage of our Resource Allocation and Admission Control (RAAC) proposal. The latter is based on two combined algorithms to solve the MILP problem $(\mathcal{P}_\epsilon{}')$ while minimizing the transmission power. Afterwards, we detail in subsection 4.3.2 the second stage of our proposal, where a fast greedy heuristic is used to solve the fronthaul admission control problem.

4.3.1 RAAC algorithm

To find solutions to the MILP problem formalized in $(\mathcal{P}_\epsilon{}')$, we design an algorithm based on the Branch-and-Cut (B&C) framework to solve the relaxed MILP, which includes the objective function (4.2.14) and the constraints (4.2.4)-(4.2.7), (4.2.9),(4.2.11),(4.2.12). The proposed modified B&C method, whose pseudo-code is presented in Algorithm 6, is considered as a powerful algorithm that merges the advantages of both the Gomory Cutting Planes and the Branch-and-Bound

Algorithm 6: `Modified B&C algorithm`

Inputs: RRHs S, PRBs K, channel gain g, power $\mathbf{P}_{max}$, p_{min}, SINR threshold $\mathbf{\Gamma}$, noise σ^2, PRB demands $\mathbf{n}^{BE}$, $\mathbf{n}^{BE}$, ϵ.
Outputs: incumbent $(\boldsymbol{y}^*, \mathbf{p}^*, \mathbf{z}^*)$.
I. Initialization:
Denote the initial problem $\mathcal{P}^0$ and the set of active problem nodes to be $\mathcal{L} = \{\mathcal{P}^0\}$.
Let the initial value set of variables $y^* = \emptyset$, $p^* = \emptyset$ and the initial lower bound $LB = -\infty$. Set $f^* = -\infty$ the initial value of objective function.
II. Iteration: do
while Number of iterations $\leq I_{max}$ **do**
    Select and delete a problem $\mathcal{P}^l$ from $\mathcal{L}$.
    Solve $\mathcal{P}^{lR}$, relaxed version of $\mathcal{P}^l$, where y takes continuous values between 0 and 1.
    if $\mathcal{P}^{lR}$ is infeasible, go back to step 3. **else** denote the optimal solution y^{lR} and p^{lR} with objective function value f.
    if $f \leq f^*$ go back to step 3.
    if y^{lR} is integer, set $f^* \leftarrow f$ and $y^* \leftarrow y$. Go back to step 3.
    If desired, search for cutting planes from previous dropped constraints that are violated by y^{lR}; if any are found, add them to the relaxation and return to step 3.
    Branch to partition the problem into new problems with restricted feasible regions. Add these problem to $\mathcal{L}$ and go back to step 3.
    Go to the next iteration.
end while

schemes into one design [88]. This leads to an algorithm that is not only more reliable, but also much faster than the conventional Branch-and-Bound alone. As detailed, in Algorithm 6, the B&C algorithm is based on a linear relaxation of the integer variables y into continuous ones, while adding Cutting Planes to enhance the problem's relaxation. This is essential to come closer to approximate integer solutions and speed up the convergence time. We also add an upper bound limit I_{max} for the algorithm's maximum number of iterations. In fact, according to [89], UEs connected to a RRH wait in maximum 10 s to receive their requested resources before disconnecting. If no resource transmission has been fulfilled within this time window, the BBU, and consequently the RRH handling the user, will release the connection and disconnect from the user. Therefore, the number of iterations I_{max} should be scalable based on the processing machine so as to not exceed the 10 s waiting period [89].

If found, we call the best solution (i.e., $(\boldsymbol{y}^*, \mathbf{p}^*, \mathbf{z}^*)$), returned by the B&C algorithm, the incumbent. This solution is however not optimal with respect to the minimization of the total C-RAN transmission power. In order to improve the solution and without adding further complexity to the problem, we propose to attempt to solve, after each resolution of the MILP problem $(\mathcal{P}_\epsilon{}')$, the following Linear Program (LP) problem:

Algorithm 7: `RAAC algorithm`

Inputs: RRHs S, PRBs K, channel gain g, power $\mathbf{p}_{max}$, p_{min}, SINR threshold $\mathbf{\Gamma}$, noise σ^2, PRB demands $\mathbf{n}^{BE}$, $\mathbf{n}^{BE}$, ϵ, tolerance δ.
Outputs: near optimal solution $(\boldsymbol{y}^*, \mathbf{p}^{'}, \mathbf{z}^*)$.
Initialization:
Set $L \leftarrow 0$ and construct MILP $(\mathcal{P}_\epsilon{}')$;
repeat
  solve MILP problem $(\mathcal{P}_\epsilon{}')$ with B&C (see Alg. 6);
  if MILP is infeasible **then**
    output: "*Problem is infeasible*" and quit;
  end if
  let $(\boldsymbol{y}^*, \mathbf{p}^*, \mathbf{z}^*)$ be the optimal solution to the MILP and U the associated upper bound;
  solve the LP problem (15) using ILOG CPLEX to find the best p for given $\boldsymbol{y}^*$;
  if LP is feasible **then**
    let $\mathbf{p}^{'}$ be the optimal LP solution;
    $L^{'}$ the associated lower bound;
  end if
  if $L^{'} \geq L$ **then**
    $L \leftarrow L^{'}$;
    save new incumbent $(\boldsymbol{y}^*, \mathbf{p}^{'}, \mathbf{z}^*)$;
  end if
until $L \geq 0$ and $(U - L)/L \leq \delta$;

$$\min_{p} \quad \sum_{u \in \mathcal{U}} \sum_{k \in \mathcal{K}_u} p_k^u \tag{4.3.15}$$

$$\text{subject to :} \quad \sum_{u \in \mathcal{U}^i} \sum_{k \in \mathcal{K}_u} p_k^u \leq p_{max}^i, \forall i \in \mathcal{S} \tag{4.3.16}$$

$$(1 + \frac{1}{\Gamma_k^u}) p_k^u g_k^u \geq \sum_{v \in \mathcal{U}} p_k^v g_k^u + \sigma^2, \forall u, \forall k \in \mathcal{K}_u \tag{4.3.17}$$

$$p_k^u \in \mathbb{R}^+ \tag{4.3.18}$$

where $\mathcal{K}_u = \{k \in \mathcal{K} | y_k^u = 1\}$ is the set of PRBs that have been allocated to UE u in the initial MILP solution. Since it is a LP problem with finite number of variables, it can be solved very rapidly using modern LP solvers such as ILOG CPLEX. Solving this LP problem, given the resource allocation solution $\boldsymbol{y}^*$ of the MILP, helps to ensure that the power minimization is also achieved, while respecting the SINR requirements of all users. Furthermore, we define in our RAAC algorithm a tolerance parameter δ as the acceptable gap between U and L, respectively the upper and lower bounds of the objective function of problem $(\mathcal{P}_\epsilon{}')$. The overall RAAC algorithm for solving both problem $(\mathcal{P}_\epsilon{}')$ and the power minimization problem is summarized in Algorithm 7. Besides, as

Algorithm 8: FFAC

Inputs: $\mathcal{S}$, data-rates $\mathbf{c}$, RRHs fronthaul constraint $\mathbf{c}_{max}$
Outputs: number of admitted Gold and Best Effort users; fronthaul links supported data-rate of each RRH.
for each RRH $i \in \mathcal{S}$ **do**
  sort in an increasing order elements of $\mathbf{c}^G$ and $\mathbf{c}^{BE}$, achieved data-rates of Gold and Best Effort users, respectively;
  $l \leftarrow 0$; $u \leftarrow 1$; $v \leftarrow 1$;
  while $l \leq c^i_{max}$ and $u < |\mathbf{c}^G| + 1$ **do**
    if $l + \mathbf{c}^G(u) \leq c^i_{max}$ **then**
      $l \leftarrow l + \mathbf{c}^G(u)$;
      $u \leftarrow u + 1$;
      if $u = |\mathbf{c}^G| + 1$ or $l + \mathbf{c}^G(u) > c^i_{max}$ **then**
        while $l \leq c^i_{max}$ and $v < |\mathbf{c}^{BE}| + 1$ **do**
          if $l + \mathbf{c}^{BE}(v) \leq c^i_{max}$ **then**
            $l \leftarrow l + \mathbf{c}^{BE}(v)$;
            $v \leftarrow v + 1$;
          else
            break;
          end if
        end while
      end if
    end if
  end while
end for
Return $u - 1$ number of admitted Gold and $v - 1$ number of admitted Best effort users.

will be detailed next section in the simulation results, it turns out to be an exact algorithm, which is capable of solving realistic size instances of the MILP problem to proven optimality in reasonable time for different network sizes.

4.3.2 Fast greedy heuristic for fronthaul admission control

The following phase of our proposal is to find a feasible solution to the initial problem $(\mathcal{P}_\epsilon)$, including the dropped fronthaul constraint (4.2.10). Towards this end, we propose a Fast greedy heuristic for Fronthaul Admission Control (FFAC) of each RRH, based on the achieved data-rate of the accepted UEs. The FFAC heuristic's pseudo-code is described in Algorithm 8. It is based on taking the returned values of $(\boldsymbol{y}^*, \mathbf{p}^{'})$ from Algorithm 7 to compute the achieved data-rate c_u of UE u over its allocated set of baseband resources $\mathcal{K}_u = \{k \in \mathcal{K} | y^u_k = 1\}$ and received SINR γ^u_k.

The achieved data-rate can be expressed by the Shanon formula as follows:

$$c_u = \sum_{k \in \mathcal{K}_u} B \log_2(1 + \gamma_k^u), u \in \mathcal{U} \tag{4.3.19}$$

The algorithm starts by sorting separately the data-rates of Gold ($\mathbf{c}^G$) and Best effort users ($\mathbf{c}^{BE}$). The algorithm works similarly to a greedy knapsack algorithm [90] where it admits, firstly, as many Gold users as possible until it reaches the fronthaul capacity c^i_{max} of RRH i on the u^{th} ordered user $\mathbf{c}^G(u)$. If the capacity is still not achieved, the algorithm moves to accept, in an increasing order of data-rate $\mathbf{c}^{BE}$, the remaining Best Effort users linked to the same RRH. We remarked that, with a realistic network size and a relatively loose fronthaul constraint, instance of the problem can be solved in a fraction of a second, which is 10 times lesser than the maximum computation period [89]. Complexity analysis will be presented with more details in the next section.

4.4 Performance evaluation

In this section, we evaluate the performance of our proposed RAAC and FFAC algorithms, and compare the benefits of our solution with respect to the Semi-Definite Positive Relaxation based Algorithm (SDPRA) and Fast Greedy Algorithm (FGA), which have been presented in [70] for admission control in C-RAN. We also compare our approach with the optimal solution returned by CPLEX, when the software is used to solve the MILP problem $(\mathcal{P}_\epsilon{}')$.

4.4.1 Simulation environment

Table 4.1 reports the simulation parameters that have been used for our event-base simulations. We simulate a wireless LTE C-RAN environment similar to Figure 4.1, consisting of 19 hexagonal-sized RRHs. Each RRH has the same radius R_{radius} and the distance between two nearest RRHs is $2d$, where $R_{radius} = 2d/\sqrt{3}$, for all RRHs in $\mathcal{S}$. We simulated two poissonian flows of mobile users; λ_G and λ_{BE} the arrival rates of Gold and Best Effort users, respectively. We vary at each simulation instance, the users stay time and service demand following an exponential and uniform laws, respectively. Note that UEs positions are randomly generated at each simulation instance, and remain fixed during their whole stay time in the network. We suppose that the required SINR threshold for all UEs is the same, denoted by γ_{th}.

For each simulation instance, we run our algorithms at each user arrival until either the problem is solved, or the gap between U and L, measured by a percentage of L is below δ, or the problem was proved to be infeasible, or we exceeded the maximum limit of number of iterations I_{max}. The results are obtained over 30 simulation instances for each scenario of SINR threshold γ_{th} and C_{max}, with a confidence interval of 95%. Repeated simulations helped us to fix the maximum number of iterations I_{max} to 500 for Algorithm 6, which allowed us to find close to optimal solutions while respecting the maximum time budget.

Table 4.1 – Simulation Parameters

Parameters	Values
Number of hexagonal cells	$R = 19$
Arrival rates of Gold UEs λ_G	1 arrivals per second
Arrival rates of Best Effort UEs λ_{BE}	2 arrivals per second
Departure rate (for both profiles)	Exponential distribution $\mu = 0.1$
Time window	3000 seconds
Bandwidth	$B = 5$ MHz
Total number of PRBs	$K = 24$
Gold UE PRB demand	Uniform distribution in $[1, 15]$
Best Effort UE PRB demand	Uniform distribution in $[1, 10]$
Path loss between RRH i and UE u	43.8 + 36.8$\log_{10}(d_{i,u})$
Carrier frequency	2.5 GHz
Thermal noise	-174 dBm/Hz
p^i_{max}	10 $W, \forall i \in \mathcal{S}$
p_{min}	1 mW
Distance d	250 m
Weight constant ϵ	0.5
Tolerance parameter δ	0.1%

4.4.2 Performance metrics

In what follows, we present the corresponding numerical results in terms of:

- Convergence time t to solve problem $(\mathcal{P}_\epsilon{}')$.
- Number of accepted Gold UEs N_{aGUE}: This metric denotes the percentage of Gold users admitted in the network during the resolution period. Recall that, All Gold users will have their resource demand 100% satisfied.
- Throughput Satisfaction Rate of Best Effort UEs TSR_{BE} (note that the TSR_{GUE} of Gold UEs is equal to 1): This metric is similar to the one seen previously in Chapter 3 and represents the ratio of allocated PRBs to Best Effort users on their requests.
- Total transmission power in the C-RAN T_{power}: It represents the sum of all RRHs transmission powers.

$$T_{power} = \sum_{u \in \mathcal{U}} \sum_{k \in \mathcal{K}_u} p^u_k \tag{4.4.20}$$

- Number of BBUs N_{BBU} in the cloud: Calculated through the formula in (3.2.23) in previous Chapter.

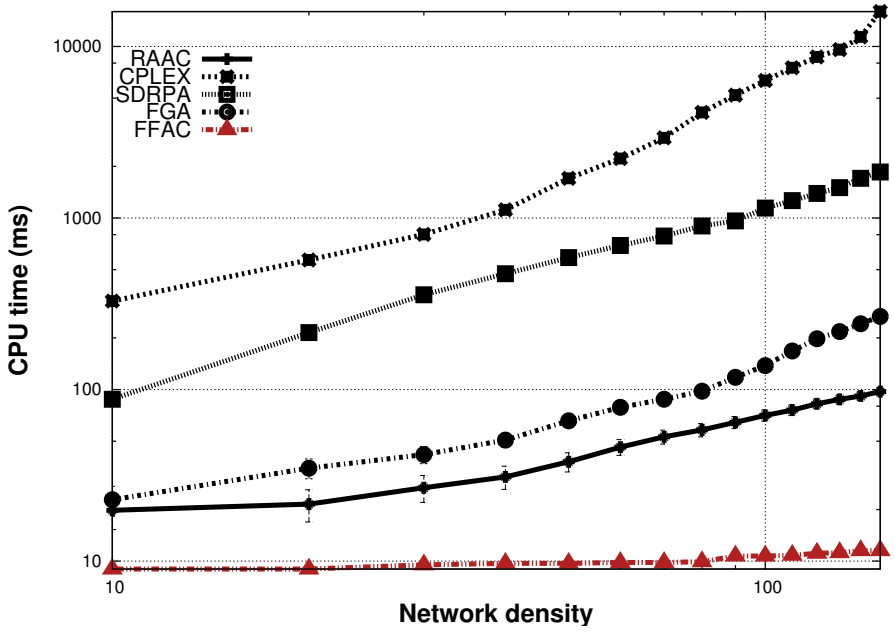


Figure 4.2 – CPU time vs. network density

Table 4.2 – Average CPU time comparison

	RAAC	CPLEX	t_2/t_1
Average computation time	$t_1 = 37.9$ ms	$t_2 = 1703.5$ ms	44.9

4.4.3 Simulation results

a) Convergence time

First of all, we find it is interesting to start presenting the convergence time of the different studied schemes as well as IBM's ILOG CPLEX. Considering our event-based simulation, we start the problem resolution each time a new user enters the network. Each algorithm is thus executed to solve the problem while considering all the existing users at that time instant a new UE enters the system. In Figure 4.2, the evolution of convergence time as a function of the network density is depicted in a logarithmic scale for all approaches. The overall results indicate that the proposed RAAC method (along with the FFAC, which has very small computation time) can solve the resource allocation and admission control problem much faster than any other scheme. In fact, for the highest network density of 150 UEs, it can return the problem's solutions in less than 100 ms. Besides, our approach achieves 44.9 more CPU time savings compared to CPLEX, as presented in

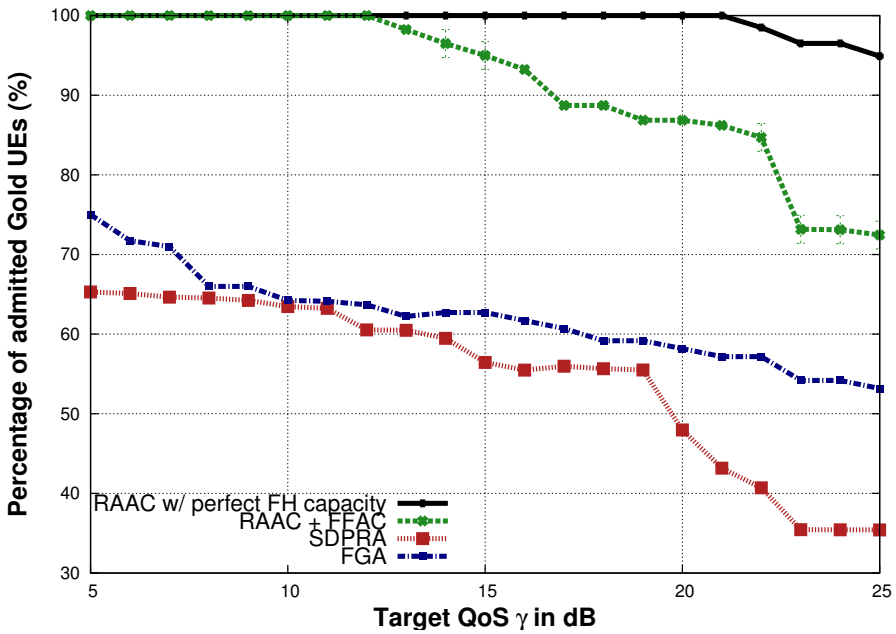


Figure 4.3 – Percentage of admitted UEs versus target QoS with fixed fronthaul capacity $C_{max} = 500$ Mbps.

Table 4.2. Besides, IBM's ILOG CPLEX appears to be unpractical for being used, since it takes significantly high time to return the optimum solution (up to 1.6×10^4 ms for the highest network density), which severely impacts the rate of served users and the global TSR as will be described later on.

b) Number of admitted Gold users

Figure 4.3 shows the evolution of the percentage of N_{aGUE} versus the QoS target for the three studied schemes, when the fronthaul capacity is fixed at $C_{max} = 500$ Mbps for each RRH. For the sake of comparison, we also add the curve for the perfect fronthaul links capacity scenario (i.e., $C_{max} = \infty$). As we can observe, the number of admitted Gold UEs decreases with the increase of the target QoS level. Our proposed RAAC scheme with fixed fronthaul capacity results in increased number of supported users; up to 44% and 36.4% more admitted UEs compared to the SDPRA and FGA schemes, respectively. The gap of N_{aGUE} between RAAC's fixed and perfect fronthaul links capacity is at maximum 23.4%. We can remark that [70]'s FGA results surprisingly in more accepted UEs than the superior SDPRA scheme. This is due to the latter's higher convergence time, as highlighted in Figure 4.2, which resulted in discarded users during the problem resolution. We vary next the fronthaul capacity C_{max} (which can be equivalent to different functional splits),

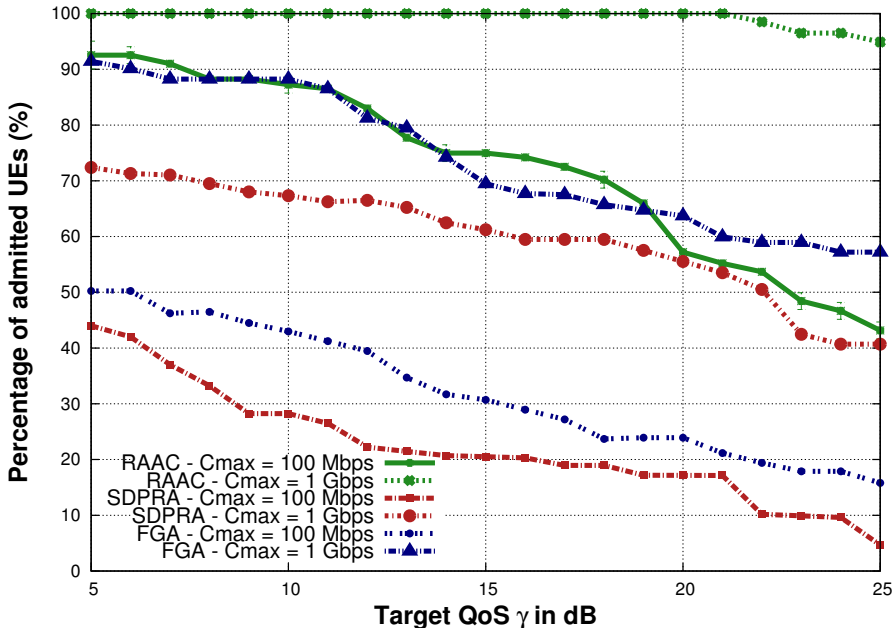


Figure 4.4 – Percentage of admitted UEs versus target QoS.

and study its impact on the percentage of accepted UEs. Figure 4.4 illustrates the percentage of N_{aGUE} versus γ_{th} for low ($C_{max} = 100$ Mbps) and high ($C_{max} = 1$ Gbps) fronthaul capacities. We remark that the number of supported UEs increases when the fronthaul network allows more data-rate to be conveyed from the BBU cloud to the RRHs. As a result, this enables the C-RAN to allocate more PRBs to mobile users and fully satisfy their QoS requirement. Moreover, we notice that RAAC outperforms both SDPRA and FGA, as it achieves relatively, when $C_{max} = 100$ Mbps, the same results as the $C_{max} = 1$ Gbps SDPRA and FGA outputs. Besides, we can remark that RAAC accepts approximately the same number of users for $C_{max} = 1$ Gbps as the perfect capacity scenario. This assesses the scalability of our proposal regarding increasing network traffic and fronthaul capacity.

c) Throughput satisfaction rate of Best Effort users

Figure 4.5 shows the Cumulative Distributed Function (CDF) of the average TSR_{BE}, which is the ratio of the average number of allocated PRBs to the total initial demands of Best Effort UEs. In this case, we fixed the fronthaul capacity at $C_{max} = 500$ Mbps and considered one QoS threshold $\gamma_{th} = 15$ dB for all Best effort users. We can observe by comparing the CDF of the three methods, that RAAC yields more UEs satisfaction compared to SDPRA and FGA. In fact, we can observe that more than 50% of UEs have their TSR greater than 55% with RAAC combined with FFAC.

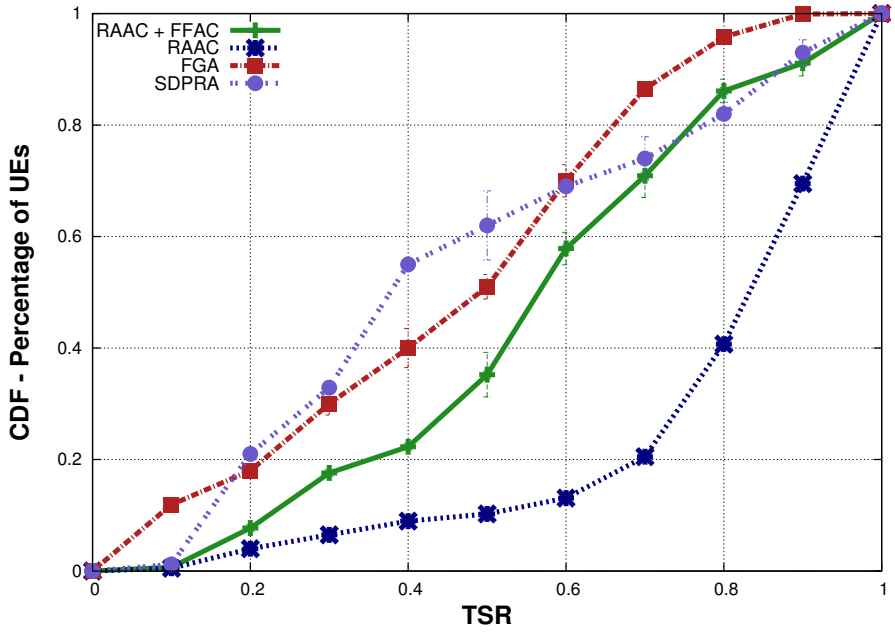


Figure 4.5 – Throughput Cumulative Density Function for Best Effort UEs.

This throughput is lessened to 46% and 34% with FGA and SDPRA, respectively. It is no surprise to see again FGA outperforming SDPRA due the difference in their computation time. On another note, we can remark that the TSR returned by RAAC alone (corresponding to perfect fronthaul links capacity) is greater than any other scheme. In fact, there is a gap of 27% between the two fronthaul capacity scenarios, since some users have been discarded due to the admission control mechanism, which fosters on Gold users admission.

d) Sensibility analysis of TSR

In Figure 4.6, we study the evolution of TSR_{BE} while variating the optimization weight ϵ. The value of ϵ allows us to combine the utility function of the two sets of user profiles into one utility function instead of two. As can be seen, setting ϵ to 1 forces the average TSR_{BE} to zero, since the problem has focused on only maximizing the number of accepted Gold users N_{aGUE}, and *vice-versa*. A decrease of ϵ will lessen the average N_{aGUE} to the profit of TSR_{BE}. However, as can be observed, the latter can hardly reach 100% even for the highest fronthaul capacity (i.e., $C_{max} = 1$ Gbps) due to power and interference limitations. A tradeoff between TSR_{BE} and N_{aGUE} can be met by setting ϵ in the set $[0.4; 0.5]$.

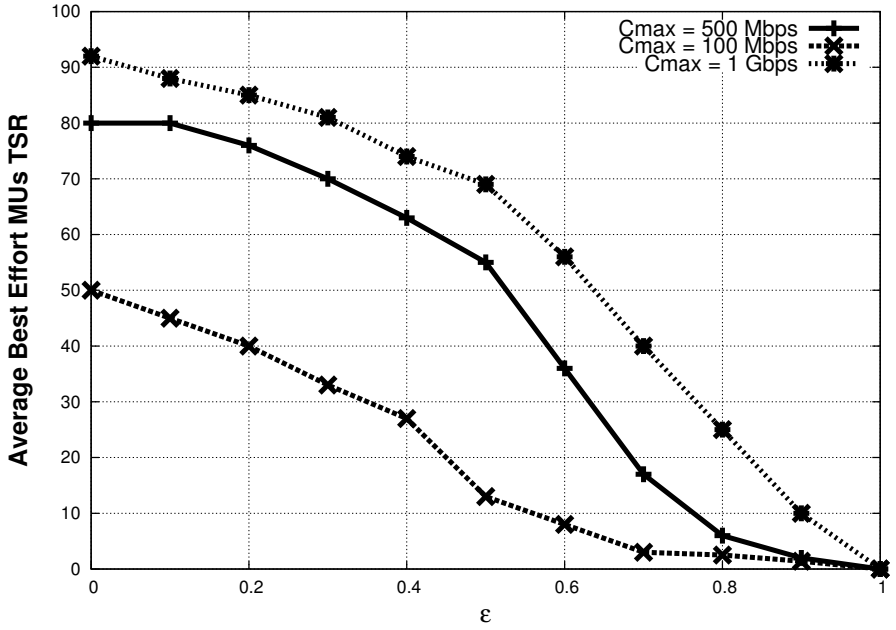


Figure 4.6 – Average TSR vs. weight constant.

e) Total transmission power

We study next the variations of the total transmission power T_{power} for all RRHs versus the QoS target γ_{th}. As shown in Figure 4.7, our method generates less transmission power than the SDPRA and FGA schemes for the same performance in terms of N_{aGUE} and TSR_{BE}. Furthermore, we notice that the total transmission power in both SDPRA and FGA present an unscaled behavior to the threshold regime γ_{th}, which implies that the power allocation is done irrespectively to UEs' QoS target. Our RAAC algorithm, on the other hand, shows that the total transmission power increases along with the growing QoS demand γ_{th}. We can also remark that the average transmission power is lessened for RAAC combined with FFAC, since some UEs will be removed from the system after the fronthaul admission control procedure. We fix next the QoS target γ_{th} to 15 dB for all users and study the total transmission power evolution versus the fronthaul capacity C_{max}. Figure 4.8 illustrates this evolution, where we observe that the total transmission power decreases as the fronthaul network capacity increases and allows more baseband resources to be transmitted. Besides, RAAC achieves more than 54.1% and 64% in power savings for large fronthaul capacity ($C_{max} \geq 500$ Mbps) compared to SDPRA and FGA, respectively. This, once again, assesses the stability and good performance and scalability of our approach.

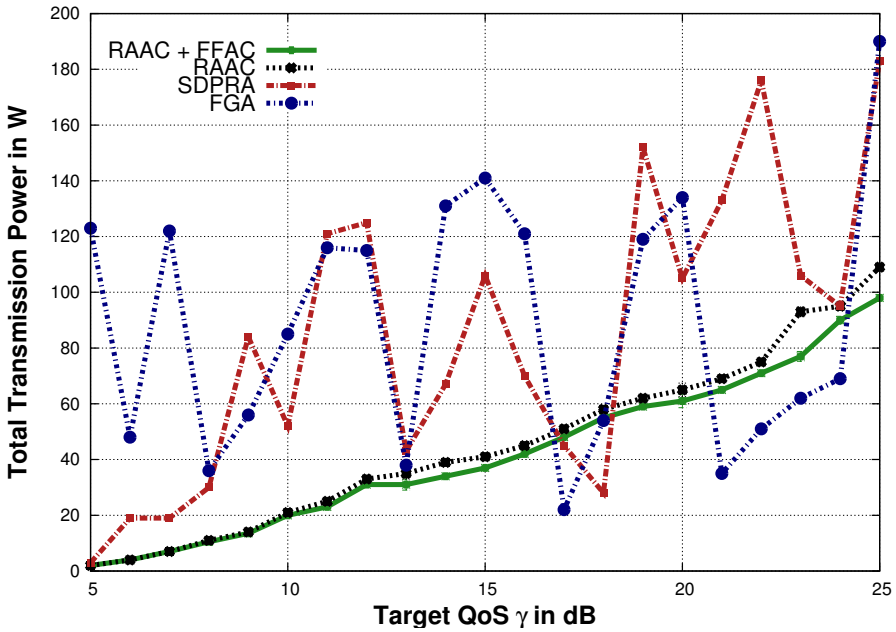


Figure 4.7 – Total transmission power versus target QoS with fixed fronthaul capacity $C_{max} = 500$ Mbps.

f) BBUs average utilization

Figure 4.9 illustrates the number of BBUs needed per time unit as well as the needed number of on RRHs to manage the traffic load at each UE's arrival, when the SINR threshold is equal to 15 dB and the fronthaul links capacity to $C_{max} = 500$ Mbps for all RRHs. We can remark that the number of RRHs is set dynamically at each instant to serve the overall traffic load. What is more, only a small number of BBU (lesser than one) is needed to handle it. For further investigation, we present in Figure 4.10 the variation of number of needed BBUs for (RAAC+FFAC) and D-RAN scenario versus the QoS target, for fixed fronthaul capacity $C_{max} = 500$ Mbps. Provided the fronthaul links have sufficient capacity, one BBU can handle the traffic load of different RRHs, while meeting the existing users' data rate requirements. We see in Figure 4.10 that, at maximum, only 80% of a full BBU usage is needed to handle the traffic load of all RRHs. After the fronthaul limit is reached for $\gamma_{th} = 12$ dB, the percentage of BBU usage is lessened since more users will be rejected from the system because of their high QoS demand and/or their interference to other UEs. N_{BBU} stabilizes then to 58%-60% of a single BBU usage, which represents 76% savings compared to D-RAN scenario.

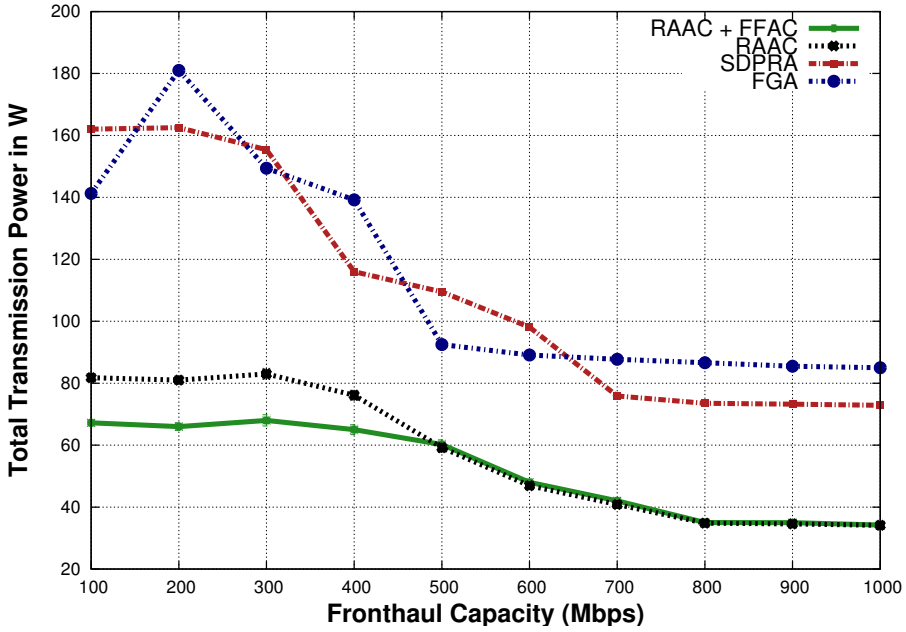


Figure 4.8 – Total transmission power versus fronthaul capacity C_{max}

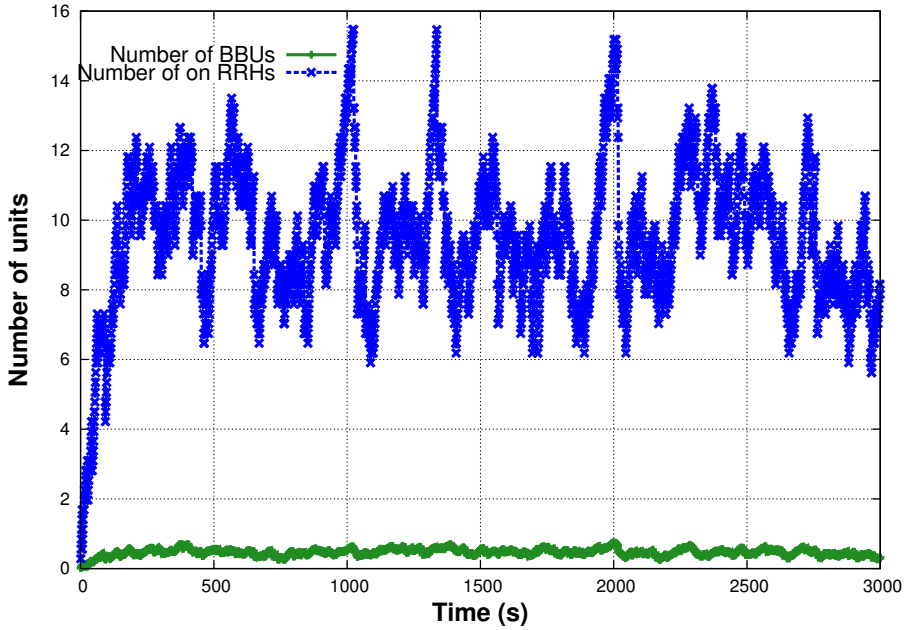


Figure 4.9 – Number of BBUs vs. time

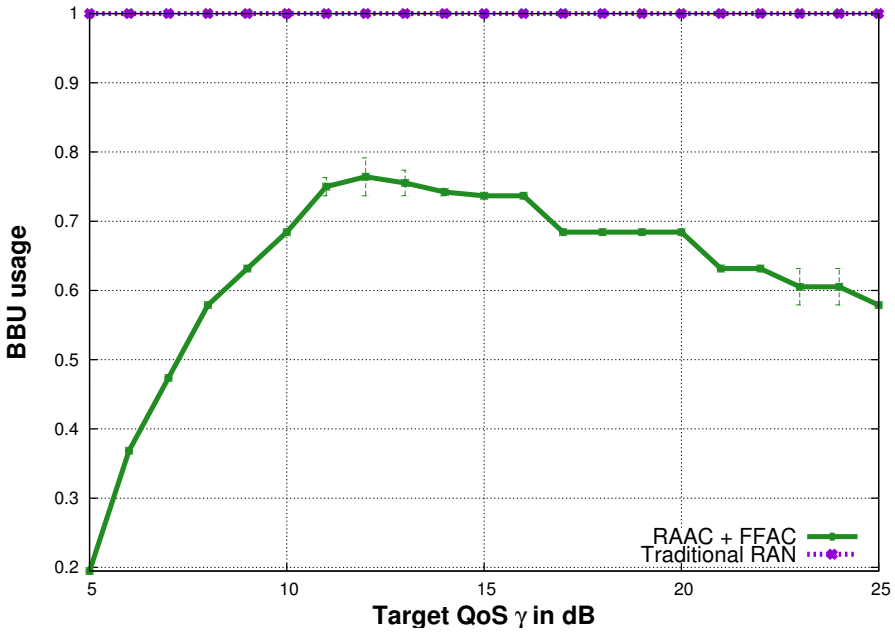


Figure 4.10 – Number of BBUs versus target QoS for $C_{max} = 500$ Mbps.

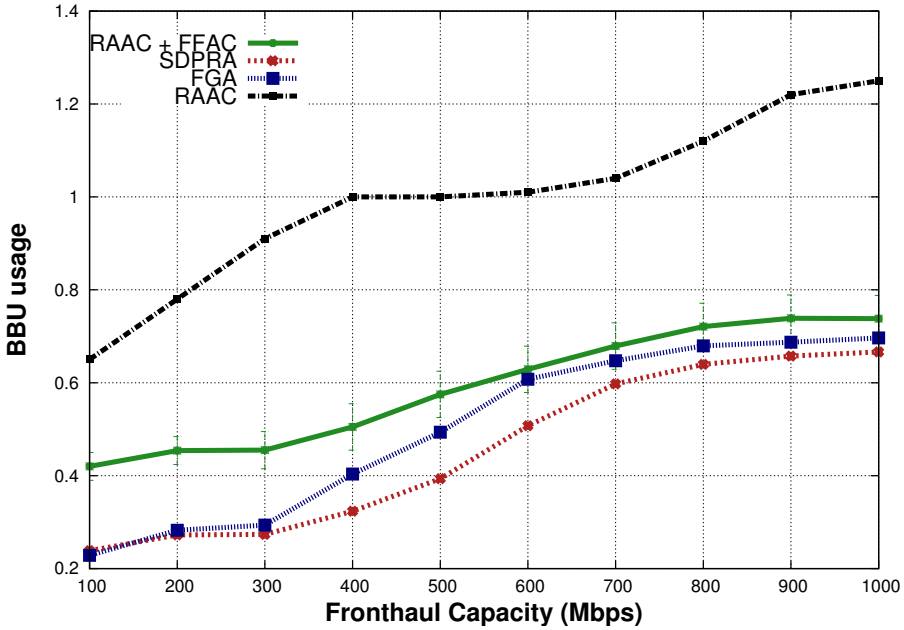


Figure 4.11 – Number of BBUs versus fronthaul capacity C_{max}.

g) BBUs usage versus fronthaul capacity

We present next the evolution of N_{BBU} when the fronthaul capacity C_{max} varies for a fixed QoS target $\gamma_{th} = 15$ dB. Figure 4.11 depicts this evolution for RAAC+FFAC, RAAC alone and the two state-of-the-art schemes. We can remark that RAAC+FFAC realizes the minimum number of instantiated BBUs usage in the cloud compared the other schemes. When the fronthaul capacity limit grows, it is shown that RAAC+FFAC, SDPRA and FFAC all converge to approximately the same value of 75% of BBU usage. On another hand, if we consider an unconstrained fronthaul network (i.e., CPRI), the RAAC scheme alone can exhibit the optimal percentage of needed BBU usage in the cloud, up to 1 and 25% of a second BBU, that can accept all mobile users and guarantee their QoS.

4.5 Conclusion

This Chapter has presented a novel framework for jointly addressing the resource allocation, admission control and power minimization problems in C-RAN, considering multi-users QoS requirements, power and fronthaul network limitations constraints. We presented a dynamic algorithm to solve the problem in real-time. Numerical results have confirmed the superior performances of our proposed RAAC approach, which increases users' admission by 44% and 36%, and saves 54% and 64% more C-RAN transmission power compared to the SDPRA and FGA schemes, respectively. We also highlighted the benefits of our proposal regarding the number of BBUs' reduction in the cloud, when compared to the previous approaches and the Distributed RAN scenario.

In the next Chapter, we will investigate the BBU selection problem from different CSPs while jointly considering several important challenges such as BBU processing power to handle time-varying traffic loads, BBU pool resiliency, processing and cost budgets.

Chapter 5

An optimization scheme for cost-resilience BBU selection in C-RAN

Contents

5.1 Introduction

In this Chapter, we consider a C-RAN architecture where the Mobile Network Operator (MNO) has to select BBU equipments to instantiate from different Cloud Service providers (CSPs) in order to run its virtualized baseband pool. We assume that each CSP's BBU is characterized by a failure probability [47] and a capacity cost, that can be equivalent to content delivery network prices [48] for the services required from CSPs. We propose in this context, a novel framework addressing the problem of optimal BBUs selection from several CSPs. The instantiated BBU pool should meet the MNO's expectations in terms of reliability, cost efficiency, processing power optimization and

traffic load catering. To the best of our knowledge, this work is the first attempt to present an optimization design for C-RAN BBU selection based on resiliency and virtualization price.

We formulate our selection problem, named Cost-Resilience BBU Selection (CRBS), as an ILP problem, designed with a weighted objective function focusing on three optimization goals: i) minimizing the BBU pool processing power, ii) maximizing its resiliency and iii) increasing the RRHs traffic load handling, subject to the virtualization's capacity and budget constraints. Additionally, we consider that each RRH is characterized by its hourly traffic, depending on the type of area it covers (e.g., business or residential area). It is worth noting that the third optimization goal focuses on maximizing the overall percentage of traffic that can be handled from the lowest traffic load RRHs at a given hour, subject to ensuring the management of the high-load ones. In fact, we consider that the traffic from the highest traffic load RRHs will not only generate a highest average traffic volume, but also highest peaks. Therefore, since there are more RRHs distributed throughout the radio site, such RRHs will be closer to the end user, and thus, will be able to cater best to the user's capacity demand. We assume that an overlaying macro cell will handle the remaining traffic from lowest traffic load RRHs if the BBU pool's capacity is reached at that specific hour.

To solve the ILP CRBS problem, we propose to employ the Branch-and-Price (B&P) algorithm [49], which is a combination of the Branch-and-Bound and Column Generation methods for efficiently solving large-scale ILP problems. Our analysis evaluates several BBU selection policies and provides general guidelines that can be used by operators to decide the best optimization strategy according to their needs: BBU processing power minimization, resiliency, traffic handling or all. Simulation results demonstrate the good performance of the B&P algorithm to solve the BBU selection problem for different scenarios, while also emphasizing the advantages of a particular one that can realize more than 48% in virtualization cost savings.

The remainder of this Chapter is organized as follows: section 5.2 presents the mathematical formulation of our system model. In section 5.3, we present our B&P algorithm design to solve the CRBS problem. Performance evaluation of our proposal is discussed in section 5.4, followed by section 5.5, which concludes the Chapter.

5.2 System model and problem formulation

We consider a two-tier C-RAN architecture with a cell-layout composed of a macrocell, overlaying a number of S RRHs [68]. We denote by N the number of BBU candidates that the MNO is inclined to instantiate in its BBU pool $\mathcal{B} = \{j|1 \leq j \leq N\}$ to handle the traffic load of all S RRHs. The number N of needed BBUs can be deduced from the overall existent traffic load on all RRHs and the downlink capacity of the operator, as has been highlighted in previous chapters. Each CSP's BBU candidate j is characterized by its per Mbps pricing m_j and its failure probability p_j. We denote by r_{ji} the binary variable, which is equal to 1 if BBU j handles RRH i, and 0 otherwise.

Without loss of generality, we assume that each BBU is limited by a fixed capacity C in terms of traffic it can handle [77]. We define the average utilization b_j of a BBU j as follows:

$$b_j = \sum_{i=1}^{S} r_{ji} l_j / C \tag{5.2.1}$$

where l_i is the current traffic in RRH i. The BBU-RRH dependent traffic is realized by a functional split separating user and cell related functions [21]. Besides, we consider that the baseband processing power consumption in a single BBU is linear with its average utilization thanks to the UE-Cell split. Thus, the processing power P_j consumed at BBU j can be expressed as:

$$P_j = P_0 + \Delta P_{max} b_j \tag{5.2.2}$$

In particular, P_0 represents the power in the BBU when the latter is in idle mode, and P_{max} when in full usage mode. Δ is a constant between 0 and 1, which represents the slope of the equivalent linear power model. On the other hand, we suppose that the failure probability of the BBU pool $p(\mathcal{B})$ is equal to 1 if $\mathcal{B} = \emptyset$, and $\prod_j p_j$, otherwise. In order to optimize this term as one of the MNO's intended goals for enhancing the cloud's resiliency, we transform the product of probabilities into linear summation by defining the following function I_j, which is equal to the negative value of the logarithmic function on the failure probability of BBU j, i.e. $I_j = -log(p_j)$. Consequently, we can write:

$$I_{\mathcal{B}} = -log(p(\mathcal{B})) = -\sum_{j} log(p_j) = \sum_{j} I_j \tag{5.2.3}$$

Furthermore, we consider two sets of RRHs in the deployed architecture denoted as $\mathcal{S}_{\mathcal{H}}$ and $\mathcal{S}_{\mathcal{L}}$, which represent the sets of RRHs with high (maximum) traffic load and low (minimum) traffic load, respectively. We formulate in the following our mathematical CRBS optimization problem $(\mathcal{P})$,

expressed as a generic weighted optimization problem with three homogenised objective terms:

$$\underset{r}{\text{minimize}} \quad \alpha \sum_{j=1}^{N} \frac{x_j P_0 + \Delta . P_{max} b_j}{P_{max}} - \beta \sum_{j=1}^{N} x_j \frac{I_j}{max(I)} \qquad (\mathcal{P}) \qquad (5.2.4)$$

$$- \gamma \frac{\sum_{j=1}^{N} \sum_{i \in \mathcal{S}_\mathcal{L}} r_{ji} l_i}{\sum_{i \in \mathcal{S}_\mathcal{L}} l_i}$$

$$\text{subject to :} \quad \sum_{i=1}^{S} r_{ji} l_i \leq C, \forall j \qquad (5.2.5)$$

$$\sum_{j=1}^{N} r_{ji} \leq 1, \forall i \qquad (5.2.6)$$

$$\sum_{j=1}^{N} m_j \sum_{i=1}^{S} r_{ji} l_i \leq M \qquad (5.2.7)$$

$$\sum_{j=1}^{N} \sum_{i \in \mathcal{S}_\mathcal{H}} r_{ji} l_i = \sum_{i \in \mathcal{S}_\mathcal{H}} l_i \qquad (5.2.8)$$

$$r_{ji} \in \{0, 1\}, \forall j, i \qquad (5.2.9)$$

The proposed objective function consists to minimize the total BBU pool processing power, while maximizing the resiliency (or minimizing the failure probability) of the pool as well as the traffic load that can be handled from the low-traffic RRHs by the instantiated BBUs. $max(I)$ represents the maximum value of $\{I_1, ..., I_N\}$. We define α, β and γ as constant weights between 0 and 1, and whose total sum is equal to 1 (i.e., $\alpha + \beta + \gamma = 1$). We consider that these constants are set in advance by the MNO in order to fix the optimization strategy depending on its most prevailing focuses. We define a binary indicator x_j, which represents the statue of BBU j, ($x_j = 0$, if $\sum_{i=1}^{S} r_{ji} = 0$, i.e., BBU j is inactive, and 1 otherwise). (5.2.5) is the total BBU resource usage limitation constraint, and constraint (5.2.6) implies that a RRH cannot be shared by more than one BBU. Meanwhile, constraint (5.2.7) denotes that the capacity costs of all instantiated BBUs should be less than or equal to the MNO's virtualization budget M. Also, constraint (5.2.8) ensures that the traffic of high-load RRHs is 100% handled and, finally, (5.2.9) indicates that variable r_{ji} is binary.

5.3 Proposed B&P algorithm for solving the CRBS problem

The CRBS problem formalized in $(\mathcal{P})$ is ILP and, contrarily to Linear Programs (LPs), cannot be solved directly using convex optimization techniques. Besides, problem $(\mathcal{P})$ is NP-hard [77] and can only be solved by exhaustively figuring out all N^S possible combinations of the BBU-RRH assignment variable, which is impracticable for large-scale networks. To find solutions to our

problem, we propose to make use of the B&P framework, which combines the branch-and-bound and column generation approaches to compute the optimal solution of ILP problems [49].

The algorithm is based on solving by column generation the linear relaxation in each node of a branch-and-bound tree. In the B&P algorithm, a sets of columns are left out of the LP relaxation in order to handle the problem more efficiently by decreasing the computational difficulty. Columns are then "*priced*" and added back to the LP relaxation as needed. To decide which column will be added, a sub-problem called the "pricing problem" is created to identify which columns should enter the basis so as to increase the objective function (in case of maximization problem). If such columns are found, the LP is then re-optimized. We detail next, the steps of designing each of the Master and Pricing problems for our B&P algorithm.

The first step consists in reformulating the original problem by applying the well-known Dantzig-Wolfe's reformulation [49], that sub-divides the problem into a Master ($\mathcal{MP}$) and a Pricing Problem ($\mathcal{PP}$). However, before applying the problem reformulation, we found that it is convenient to transform problem ($\mathcal{P}$) by considering two binary variables v and w instead of single variable r, which represent low and high-traffic load RRHs assignment variables to BBUs, respectively. The new CRBS problem ($\mathcal{P}'$) can be expressed as follows:

$$\begin{aligned}
\underset{v,w}{\text{maximize}} \quad & \sum_{j=1}^{N} \Phi_j x_j + \Psi \sum_{j=1}^{N} \sum_{i \in \mathcal{S}_\mathcal{H}} w_{ji} l_i^H \qquad (\mathcal{P}') \\
& + \Omega \sum_{j=1}^{N} \sum_{i \in \mathcal{S}_\mathcal{L}} v_{ji} l_i^L \\
\text{subject to :} \quad & \sum_{i \in \mathcal{S}_\mathcal{H}} w_{ji} l_i^H + \sum_{i \in \mathcal{S}_\mathcal{L}} v_{ji} l_i^L \leq C, \forall j && (5.3.10) \\
& \sum_{j=1}^{N} v_{ji} \leq 1, \forall i \in \mathcal{S}_\mathcal{L} && (5.3.11) \\
& \sum_{j=1}^{N} w_{ji} \leq 1, \forall i \in \mathcal{S}_\mathcal{H} && (5.3.12) \\
& \sum_{j=1}^{N} m_j \Big(\sum_{i \in \mathcal{S}_\mathcal{H}} w_{ji} l_i^H + \sum_{i \in \mathcal{S}_\mathcal{L}} v_{ji} l_i^L \Big) \leq M && (5.3.13) \\
& \sum_{j=1}^{N} \sum_{i \in \mathcal{S}_\mathcal{H}} w_{ji} l_i^H = \sum_{i \in \mathcal{S}_\mathcal{H}} l_j^H && (5.3.14) \\
& v_{ji}, w_{ji} \in \{0,1\}, \forall j, i && (5.3.15)
\end{aligned}$$

where: $\Phi_j = \beta I_j / max(I) - \alpha P_0 / P_{max}$; $\Psi = -\alpha\Delta/C$; $\Omega = \gamma / \sum_{i \in \mathcal{S}_\mathcal{L}} l_i^L - \alpha\Delta/C$. We denote by l_i^L and l_i^H the traffic in low and high-traffic load RRH i, respectively. Also, $x_j = 0$, if

$\sum_{i\in\mathcal{S}_\mathcal{H}} w_{ji} + \sum_{i\in\mathcal{S}_\mathcal{L}} v_{ji} = 0$, and 1 otherwise. We apply next Dantzig-Wolfe's reformulation. Let $\mathcal{K}_j^L = \{v_1^j, v_2^j, ..., v_{k_j}^j\}$ and $\mathcal{K}_j^H = \{w_1^j, w_2^j, ..., w_{k_j}^i\}$ be the sets of possible feasible assignments of low and high-traffic load RRHs to BBU j, respectively. In this case, $v_k^j = \{v_{1k}^j, v_{2k}^j, ..., v_{Sk}^j\}$ and $w_k^j = \{w_{1k}^j, w_{2k}^j, ..., w_{Sk}^j\}$ constitute a feasible solution to problem $(\mathcal{P}')$. Let $t_k^j = (\dot{t}_k^j, \ddot{t}_k^j)$ be a new variable, which is equal to $(1,1)$ if feasible solution (v_k^j, w_k^j) is selected, and $(0,0)$ otherwise. We express in the following the Master Problem:

$$\underset{t}{\text{maximize}} \quad \sum_{j=1}^{N}\sum_{k=1}^{k_j}(\Phi_j x_j + \Psi \sum_{i\in\mathcal{S}_\mathcal{H}} (w_{ik}^j l_j^H)\ddot{t}_k^j \qquad (\mathcal{MP})$$
$$+ \Omega \sum_{i\in\mathcal{S}_\mathcal{L}} (v_{ik}^j l_i^L)\dot{t}_k^j)$$

$$\text{subject to :} \quad \sum_{j=1}^{N}\sum_{k=1}^{k_j} v_{ik}^j \dot{t}_k^j \leq 1, \forall i \in \mathcal{S}_\mathcal{L} \tag{5.3.16}$$

$$\sum_{j=1}^{N}\sum_{k=1}^{k_j} w_{ik}^j \ddot{t}_k^j \leq 1, \forall i \in \mathcal{S}_\mathcal{H} \tag{5.3.17}$$

$$\sum_{k=1}^{k_j} \dot{t}_k^j \leq 1, \sum_{k=1}^{k_j} \ddot{t}_k^j \leq 1, \forall j \tag{5.3.18}$$

$$\sum_{j=1}^{N} m_j \sum_{k=1}^{k_j} (\sum_{i\in\mathcal{S}_\mathcal{H}} w_{ik}^j \ddot{t}_k^j l_i^H + \sum_{i\in\mathcal{S}_\mathcal{L}} v_{ik}^j \dot{t}_k^j l_i^L) \leq M \tag{5.3.19}$$

$$\sum_{j=1}^{N}\sum_{k=1}^{k_j}\sum_{i\in\mathcal{S}_\mathcal{H}} w_{ik}^j \ddot{t}_k^j l_i^H = \sum_{i\in\mathcal{S}_\mathcal{H}} l_i^H \tag{5.3.20}$$

$$\dot{t}_k^j, \ddot{t}_k^j \in \{0,1\}, \forall j,k \tag{5.3.21}$$

In the Master Problem $(\mathcal{MP})$, t_k^j represents a feasible assignment of RRHs to BBU j. Note that $(\mathcal{MP})$ cannot be solved directly due to its exponential number of columns, this is why we consider only a subset of columns that constitutes the Restricted Master Problem (RMP), where the values of the variables that do not appear are padded to zero. The observation is, for large-scale ILPs, most columns will have their associated variables equal to zero in any optimal solution anyway. Let $t^\star$ be the corresponding dual solution of the RMP. The next step consists in adding a number of columns with positive reduced cost that are found by solving the two following sub-problems:

$$\underset{1\leq j\leq N}{\text{maximize}} \qquad \{u^j - t^{\star j}\} \tag{5.3.22}$$

where $u^j = (\dot{u}^j, \ddot{u}^j)$ is the optimal solution of the following Pricing Problem $(\mathcal{PP})$:

$$\begin{aligned}
\underset{v^j, w^j}{\text{maximize}} \quad & \Phi_j x_j + \Psi \sum_{i \in \mathcal{S}_\mathcal{H}} (l_i^H - w_i^\star) w_i^j \qquad (\mathcal{PP}) \\
& + \Omega \sum_{i \in \mathcal{S}_\mathcal{L}} (l_i^L - v_i^\star) v_i^j \\
\text{subject to :} \quad & \sum_{i \in \mathcal{S}_\mathcal{H}} w_i^j l_i^H + \sum_{i \in \mathcal{S}_\mathcal{L}} v_i^j l_i^L \leq C \qquad (5.3.23) \\
& v_{ji}, w_{ji} \in \{0, 1\}, \forall i \qquad (5.3.24)
\end{aligned}$$

where $v_i^\star$ and $w_i^\star$ correspond to the optimal dual price from the solution of the RMP associated with the partitioning constraints of low and high-traffic load RRH i, respectively. In the Pricing Problem $(\mathcal{PP})$, we generate the best feasible low and high-traffic load RRH assignments from all the feasible ones for each BBU j. After that, we look for the best BBU-RRH assignments over all BBUs, which is precisely done by problem (5.3.22). Figure 5.1 summarizes the B&P algorithm's flow-chart. It is worth noting that branching in the B&P occurs when no columns have been "*priced* out" to enter the basis and the LP solution does not satisfy constraints. Furthermore, it is not required to solve $(\mathcal{PP})$ to optimality; in fact, any column with a positive reduced cost can be accepted. Hence, if the value of the objective function to the column generation sub-problem is less or equal to zero, then the current optimal solution for the RMP is also optimal for $(\mathcal{MP})$.

5.4 Performance Evaluation

In this section, we evaluate the benefits of the B&P algorithm to solve the CRBS problem, while comparing the results for different scenarios of the optimization weights (α, β, γ). We have tested different combinations and outline, in the following, the most representative of the other weights values:

- $(\alpha, \beta, \gamma) = (1, 0, 0)$: Total Power Minimization Scheme (TPMiS), where the MNO intends to exclusively minimize the BBU processing power;
- $(\alpha, \beta, \gamma) = (0, 1, 0)$: Resilience Maximization Scheme (RMaS), where the MNO intends to exclusively maximize the system resiliency;
- $(\alpha, \beta, \gamma) = (0, 0, 1)$: Low Traffic Maximization Scheme (LTMaS), where the MNO intends to exclusively maximize the number of handled low-traffic RRHs;
- $(\alpha, \beta, \gamma) = (1/3, 1/3, 1/3)$: Equal-Weighted Optimization Scheme (EWOS), a tradeoff option where all weights are a priori equal;

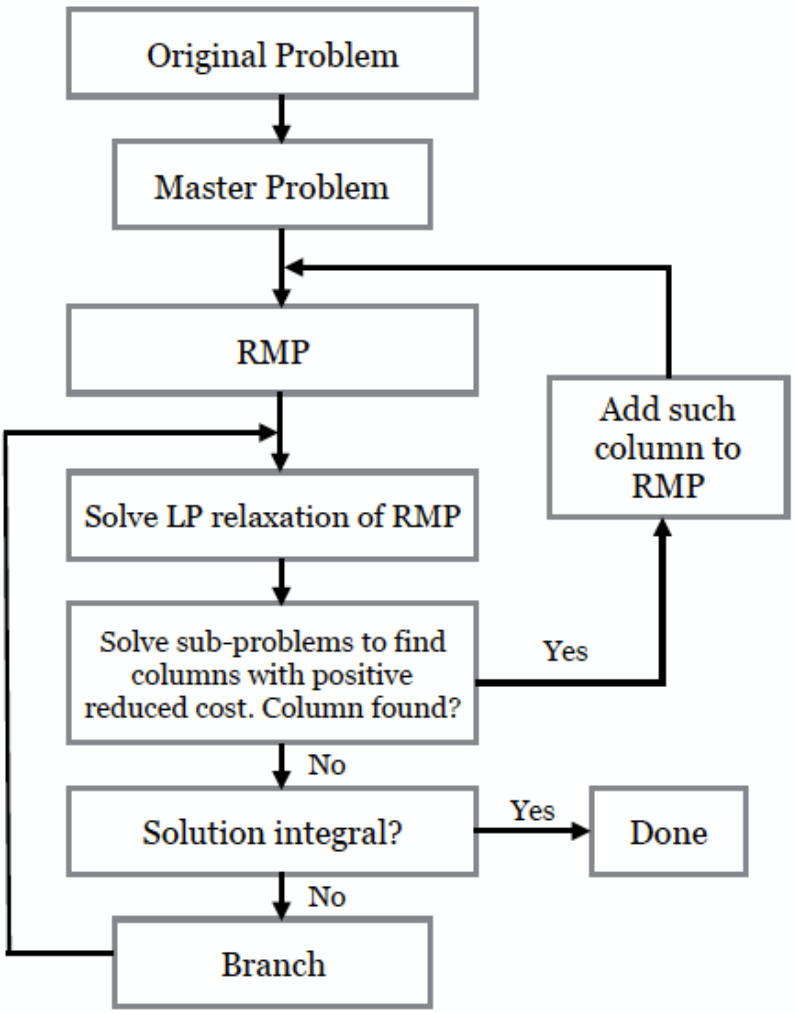


Figure 5.1 – B&P algorithm Flow Chart

- $(\alpha, \beta, \gamma) = (0.4, 0.2, 0.4)$: 424-Scheme, which was chosen in consideration that reliability may be twice less important for a MNO than to serve the whole traffic and minimize the total C-RAN BBU power (numerical results will latter exhibit the benefits of this latter's weights choice).

5.4.1 Simulation environment

In all our simulation scenarios, we consider four CSPs, with their corresponding failure probabilities, I-function values and price per Mbps that are detailed in Table 5.1 (data of existent commercial content delivery network that can be closely equivalent to CSPs can be found in [48]). The rest of

Table 5.1 – CSP inputs from [47] [48]

CSP-i	Failure probab. p_i	I_j	pricing m_j
CSP-1	0.01	2	$0.8\$/Mbps$
CSP-2	0.05	1.30103	$0.9\$/Mbps$
CSP-3	0.1	1	$1\$/Mbps$
CSP-4	0.01	2	$1.1\$/Mbps$

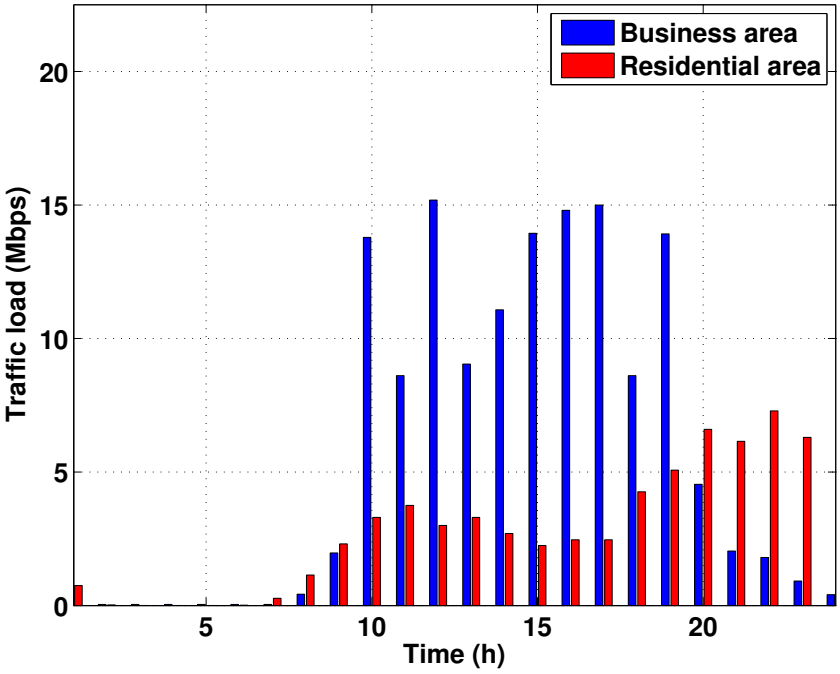


Figure 5.2 – Daily traffic load observed on business and residential area RRHs during a workday [5]

the default simulation parameters are described in Table 5.2. We consider a total of $S = 19$ RRHs, with 15 business and 4 residential RRHs, differentiated by an hourly traffic load given by [5] and illustrated in Figure 5.2. The observation is the number of RRHs that can be within the low traffic set $\mathcal{S}_{\mathcal{L}}$ or the highest one $\mathcal{S}_{\mathcal{H}}$ varies depending on the hour. We use the solver IBM ILOG CPLEX [42] to solve both the LP relaxation of the RMP and the Pricing Problem.

Table 5.2 – Simulation parameters

Parameters	Values
Number of RRHs S	19
Number of business RRHs	15
Number of residential RRHs	$\in [1; 19]$ (default 4)
$P_0(W)/\Delta/P_{max}(W)$	1.25 / 1 / 3.75
Number of BBUs N	8 (2 from each CSP)
BBU capacity $C(Mbps)$ [21]	200
Bandwidth	10 MHz
Physical Resource Blocks	50
Virtualization budget M [48]	230$/hour (2M$/year)

5.4.2 Performance metrics

In what follows, we present the evaluation of our approach through the means of the following performance metrics:

- Time complexity, which represents the global computation time to solve the ILP problem.
- BBU pool processing power: Denotes the sum of all processing powers of each active BBU in $\mathcal{B}$ (i.e., $\sum_j x_j P_j$).
- Resiliency: It is measured by the failure probability $\prod_j x_j p_j$ of the active BBUs in $\mathcal{B}$.
- Number of handled RRHs by each BBU j, calculated by $\sum_{i=1}^{S} r_{ji}$.
- Virtualization cost: it represents the hourly OPEX to instantiate the BBU pool with the BBUs to handle the traffic load. It is computed by $\sum_{j=1}^{N} x_j m_j \sum_{i=1}^{S} r_{ji} l_i$.
- The handled C-RAN traffic load: We emphasize on the evolution of handled traffic load's percentage while varying the number of residential RRHs in the network.

For the sake of comparison to a static benchmark, we compare the results to a Static Selection Scheme (SSS) where the MNO targets to satisfy the maximum achieved network load at all hours, so as to ensure maximum users quality of service, while contracting with only one CSP (CSP-4) and with no budget constraint.

Table 5.3 – Time complexity analysis

	B&P		Exhaustive Search	
# (BBU,RRH)	# nodes	CPU	# nodes	CPU
(8,19)	7	5.9	14	59.3
(10,35)	7	6.4	72	174.1
(12,50)	8	7.6	145	849.3
(15,75)	8	8.2	509	2668.9
(18,100)	7	10.6	836	7055.9
(35,200)	15	62.9	NA	NA

5.4.3 Simulation results

a) Time complexity analysis

Before comparing the performances of the different weights scenarios, we find it interesting to start studying the time complexity of our developed B&P algorithm to solve the CRBS problem, with respect to the exhaustive search method. The optimal solution in the latter is obtained by searching all possible N^S BBU-RRH assignment combinations. Table 5.3 shows the time complexity of the B&P and exhaustive search methods for different network sizes of BBUs and RRHs. We can remark that the average overall computation time of the B&P algorithm is less than $6\ ms$ for the default network size, and less than $10\ ms$ for a 100-RRHs based network. On another note, we can clearly see that the computation complexity of the exhaustive method is very big as the network size increases. Besides, it finds limitations for a large network consisting of 200 RRHs and 35 BBUs, whereas the B&P returned the optimal solution in less than $63\ ms$. Therefore, we can assess the scalability of our approach to efficiently solve large-scale versions of the CRBS problem. In what follows, we consider the default system parameters of Table 5.2 and the study the evaluation of all five weights scenarios.

b) BBU pool processing power

We compare in Figure 5.3 the hourly BBU pool processing power returned from the six approaches. We can remark an adaptive behaviour to the fluctuating traffic load for all weights scenarios schemes, with TPMiS having the minimum power consumption. In fact, the latter instantiates the least number of BBUs compared to the others, which lessens the total BBU pool processing power. LTMaS comes second in power minimization, since it has to handle more traffic coming from low traffic load RRHs. Besides, we can remark that 424-S consumes less processing power than LTMaS at certain peak traffic hours, such as at $h = 12:00$, $h = 14:00$ and $h = 15:00$. On the other hand, RMaS has the second highest power consumption, since it tends to maximize the number of in-

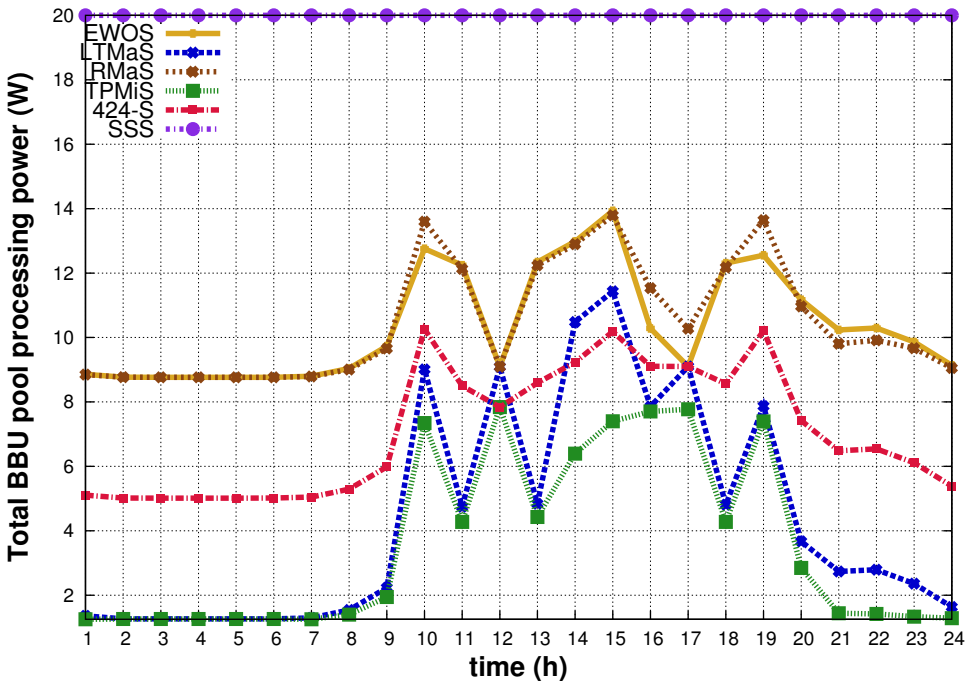


Figure 5.3 – BBU pool processing power vs time

Table 5.4 – Average BBU pool failure probability and number of BBUs

Scheme	TPMiS	RMaS	LTMaS	EWOS	424-S
$p(\mathcal{B})$	$\leq 10^{-2}$	$\leq 5.10^{-6}$	$\leq 10^{-2}$	$\leq 5.10^{-6}$	$\leq 10^{-4}$
Min.	1	3	1	3	2
Max.	2	8	5	7	4

voked BBUs to increase the resiliency. Furthermore, we measured that EWOS and 424-S consume at maximum 30% and 37.5% less processing power than SSS, respectively.

c) Resiliency

Table 5.4 presents the different values of the BBU pool's highest failure probability $p(\mathcal{B})$ during the day, as well as the minimum and maximum number of instantiated BBUs. Since $p(\mathcal{B})$ is the product of all invoked BBUs failure probabilities, the more BBUs are instantiated, the smaller is the failure probability and the more resilient is the BBU pool. This can be seen in both the RMaS and EWOS schemes as they achieve the maximum resiliency throughout the day by invoking at least more than three BBUs from CSP-2 and CSP-1 and/or CSP-3. On the other hand, TPMiS and LTMaS usually

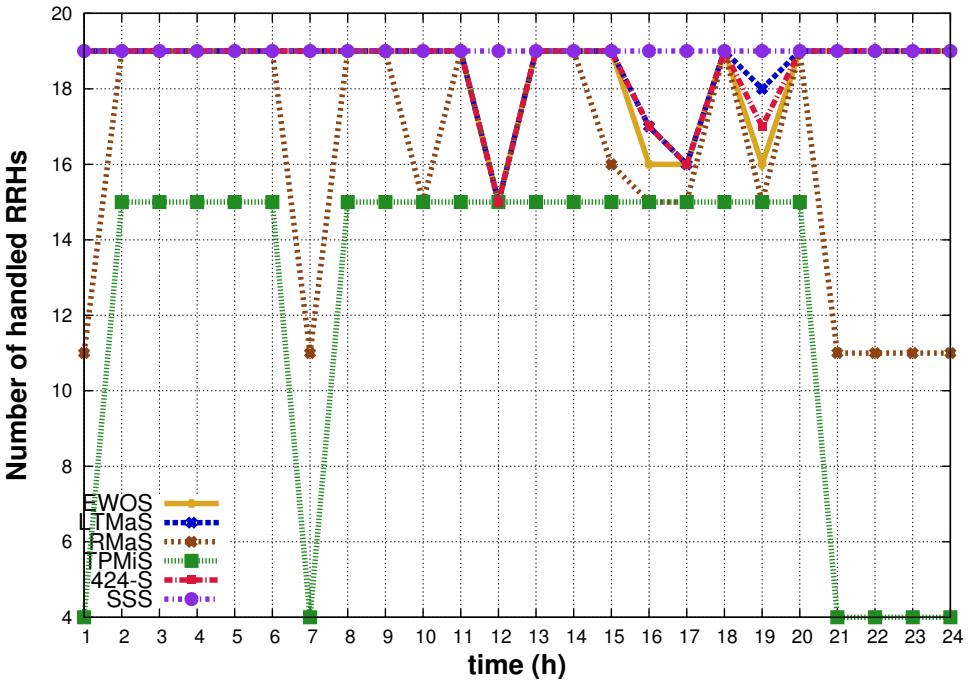


Figure 5.4 – Number of active RRHs vs time

start with few number of BBUs, then instantiate as many as possible to accommodate to peak traffic at high-traffic load hours, and the extra coming from low traffic load RRHs for LTMaS. Regarding 424-S, it invokes fewer number of BBUs than EWOS but, as will be seen next, can handle more RRH traffic.

d) Number of handled RRHs

The evolution of the number of active RRHs is depicted in Figure 5.4, where it is shown that EWOS, 424-S and LTMaS maximize the total number of handled RRHs. As for TPMiS and RMaS, they cater exclusively to the load of high-traffic cells, since they tend to minimize and maximize, respectively, the number of BBUs ($\gamma = 0$). However, we remark that at peak traffic hours such as at $h = 12:00$, $h = 16:00$, $h = 17:00$, and $h = 19:00$, not all RRHs could be handled by EWOS, 424-S and LTMaS, unlike SSS, due to restricted BBU capacity and budget (as will be detailled in the next figure). Meaning that, all (or a major part) of the traffic from residential areas during office hours will be handled by the macro base station. We can also note how 424-S performs better than EWOS at $h = 16:00$ and $h = 19:00$ as it handles more RRHs traffic.

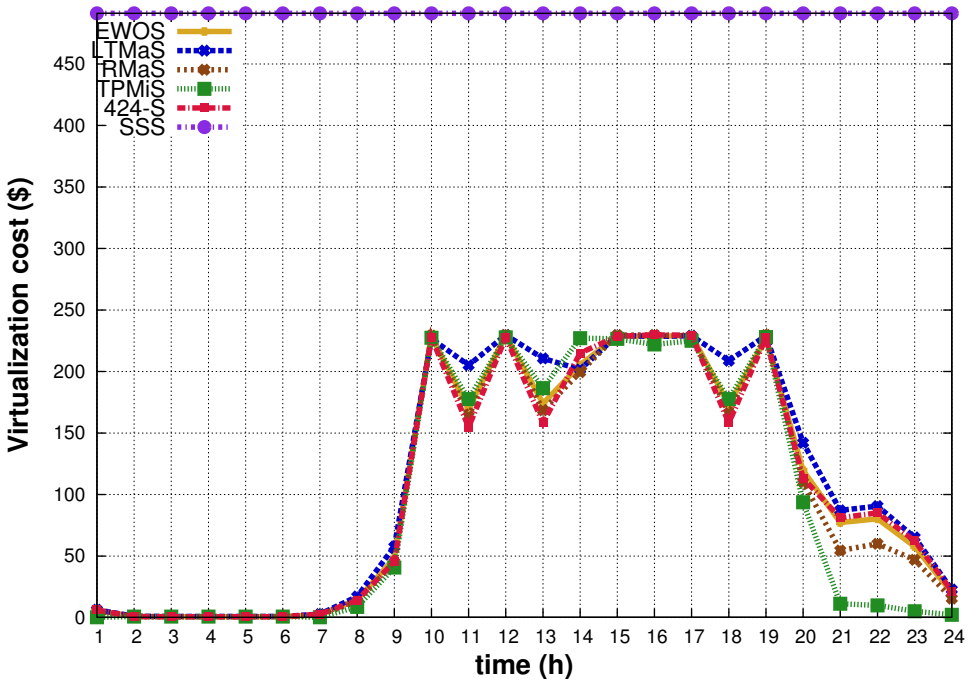


Figure 5.5 – Virtualization cost vs time

Table 5.5 – MNO annual OPEX and cost savings

Scheme	TPMiS	RMaS	LTMaS	EWOS	424-S
Total cost (K$)	907	943	1,048	991	1,042
Annual savings	54.65%	52.85%	47.6%	50.45%	47.9%
Savings to SSS	80.85%	80.09%	77.87%	79.08%	78%

e) Virtualization cost

The study of the virtualization cost is presented in Figure 5.5. We can remark that at peak traffic hours, the MNO's budget limitation B of 230$ per hour is reached for most schemes, with TPMiS realizing minimum cost from $h = 20:00$ to $h = 09:00$. This can be explained as TPMiS serves less cells during those times compared to schemes like LTMaS and 424-S. The latter on the other hand achieves the lowest virtualization cost at peak traffic hours from $h = 10:00$ to $h = 13:00$ and from $h = 17:00$ to $h = 19:00$. In Table V, we detail the annual expenditure and cost savings to both the annual budget and SSS for all five schemes. We can remark that all methods achieve great reductions compared to SSS, from 77% up to almost 81% of cost savings.

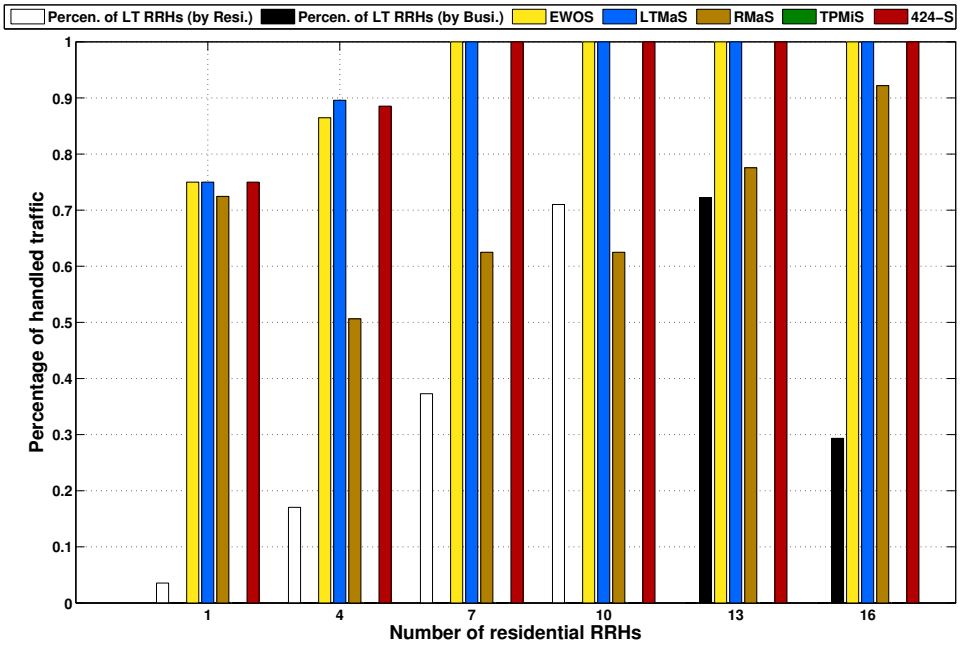


Figure 5.6 – Average percentage of handled low traffic RRH vs number of residential cells

LTMaS on the other hand, realizes the least cost reduction with 47.6% of savings of the total annual budget, since it handles more cells and thus uses more budget to cater to their traffic. Consequently, to allow more traffic handling, the MNO should either boost its budget limit to reach the SSS benchmark; or extend its contractual budget with more CSPs; or increase the BBU capacity C, with the repercussion of decreasing its total cost savings. A special mention goes to the 424-S approach, which achieves almost 48% of annual savings and 78% to the static approach.

f) Percentage of handled low traffic load

In Figure 5.6, we study the evolution of the average percentage of handled low-traffic RRHs, while varying the number of residential RRHs in the network. The average is taken from the 24 values of the day. With the increase of the number of residential RRHs, the load of the business ones become less important in the network and then switches to become the set $\mathcal{S}_{\mathcal{L}}$. This allows the BBU pool to handle 100% of all the RRHs load by EWOS, LTMaS and 424-S. Besides, since RMaS tends only to maximize the number of BBUs while satisfying constraint $(C4)$, it consequently increases the capacity of the BBU pool to handle extra traffic from the residential cells, but still not achieving 100%. We can note the absence of TPMiS since $\gamma = 0$. On the other hand, it is shown in Figure 5.7 that the number of instantiated BBUs generally increases with the increase of residential cells. For

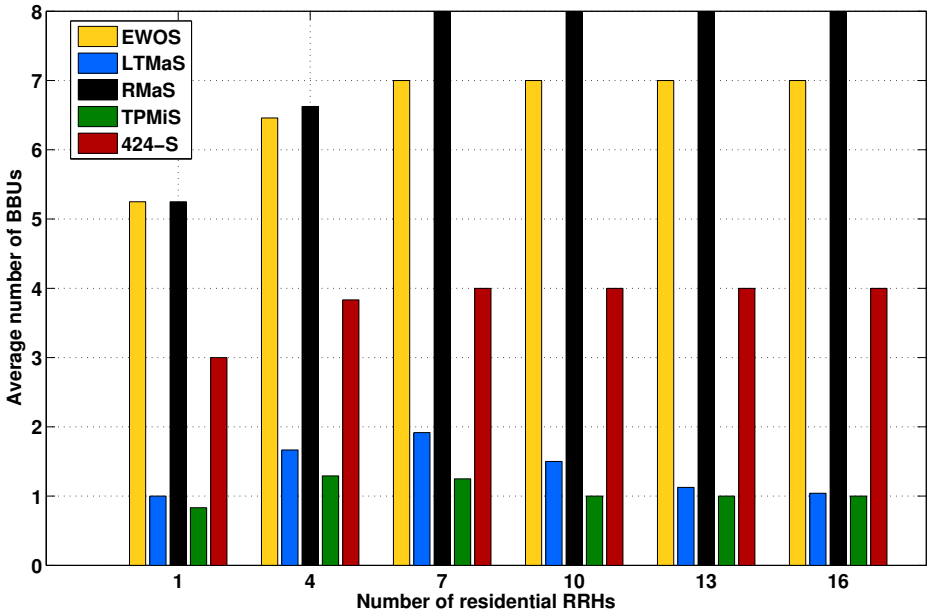
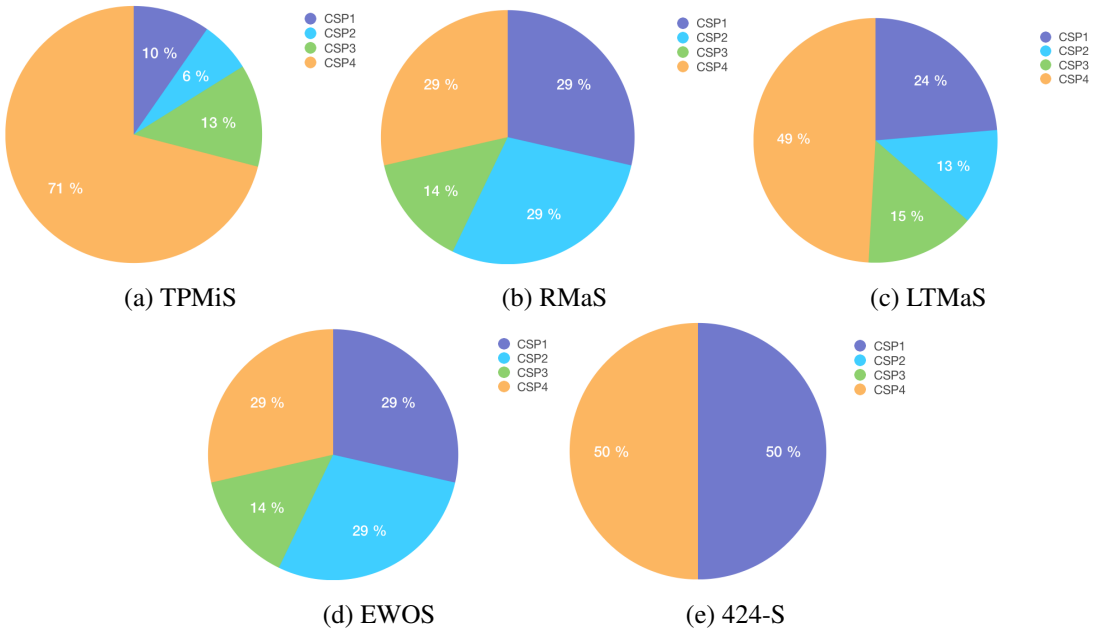


Figure 5.7 – Average number of instantiated BBUs vs number of residential cells

TPMiS and LTMaS however, the number of BBUs decreases at a certain point due the residential cells outnumbering the business ones, which causes less total generated traffic, less processing power and, consequently, less BBUs. We can also remark how 424-S invokes 50% less BBUs than EWOS for handling the same amount of traffic.

g) Most used CSPs

In Figure 5.8, we present the percentage share of the four commercial CSPs from Table I that have been used for each scheme. As we can observe, RMaS and EWOS have the same share of CSPs, with an equal distribution between CSP-1, 2 and 4 of 29%. BBUs from CSP-3 however, are the least solicited ones for these two schemes; this can be explained due to having a higher failure probability ($p_3 = 0.1$) than the other suppliers. Meanwhile, we can remark that TPMiS and LTMaS both solicit CSP-4 as a main provider with a share of 71% and 49%, respectively. Although they are the ones that instantiate the minimum number of BBUs, TPMiS and LTMaS mainly contract with CSP-4, which is the most expensive supplier, to reach the budget constraint B. On the other hand, 424-S contracts with only two CSPs (1 and 4) with an equal share of 50%. Surprisingly, despite having lesser resiliency weight ($\beta = 0.2$), 424-S uses the CSPs that provide the lowest failure probability ($p_1 = p_4 = 0.01$). These results can give an insight to MNOs in order to help



(a) TPMiS (b) RMaS (c) LTMaS

(d) EWOS (e) 424-S

Figure 5.8 – Most used CSPs for each scheme

them better direct their CSPs budget and find the ones that suit them the most.

Final remarks

From our analysis of all the performance metrics, we can deduce that forgoing a part of resiliency weight for selecting BBUs in favor of the processing power and traffic handling can yield the best results. This was assessed in the 424-S case which outperformed the tradeoff scheme EWOS in practically all aspects. In fact, 424-S proved to be a good strategy choice for MNOs to satisfy 100% of traffic with the minimum number of BBUs, provided they can tolerate its failure probability of 10^{-4} (which is faultless) and meet the total expenditure of the scheme. Besides, we have seen that 424-S realizes 48% savings of the annual budget M, and 78% when compared to a static approach.

5.5 Conclusion

In this Chapter, we have evaluated several BBU selection policies and provided general guidelines for MNOs to decide the best optimization strategy according to their needs: BBU processing power minimization, resiliency, traffic handling or all. More precisely, the presented EWOS and 424-S approaches have shown their ability to adapt to most traffic load scenarios and answer MNOs' major constraints. Especially, the 424-S may be the best option if a MNO targets to instantiate 50% less BBUs, and realize 37.5% of BBU power savings and 48% of OPEX economy.

Chapter 6

Conclusions

Contents

In this Chapter, we conclude the thesis and present some perspectives of the future work to be carried out. In section 6.1, we summarize our proposals outlined in this thesis. Then, in section 6.2, we discuss the future research directions that we are planning to consider as extension to our works, from short and long term views. Finally, section 6.3 gives an overview of the list of publications that have been achieved during this thesis.

6.1 Summary of contributions

In this thesis, we have addressed the problematic of real-time resource allocation and power minimization in DL communications for C-RAN. Our research has mainly focused on how to dynamically allocate baseband resources to time-varying flows of mobile users, while properly assigning UEs to RRHs and RRHs to BBUs in order to accommodate the traffic demands. We have studied the cases where the fronthaul capacity is both constrained and unconstrained, and when the MNO's BBU pool processing budget is limited to handle all traffic loads. The exposed optimization problems are non-linear and NP-hard, encompassing many constraints on mobile users' resources demands, QoS requirements, interference management, BBU pool capacity, transmission power

limitations, fronthaul links capacity and resiliency. To cope with the problem intractability, we have presented three algorithms for real-time resource allocation in C-RAN, while making use of different theoretical approaches. Hereafter, we will summarize our main contributions.

The first contribution provided a detailed overview of C-RAN resource allocation strategies in literature. First, we described the C-RAN transmission power minimization schemes. Next, we outlined the resource allocation and admission control approaches for fronthaul-constrained C-RANs. Then, we presented the different existing solutions for BBU-RRH mappings. Afterwards, we summarized all the discussed related C-RAN strategies, while exposing different metrics for evaluating the performance of the proposed solutions: i) algorithm application, ii) achieved UEs data rates and iii) computational complexity.

In the second contribution, we have jointly studied the problems of resource allocation and power minimization in C-RAN subject to transmission power and SINR constraints, as well as the underlying BBU-RRH assignment problem. Our two-stage framework can determine at each UE's arrival the best: i) PRBs allocation to efficiently satisfy UEs resource requests, ii) the number of RRHs to turned on and iii) the number of needed BBUs to handle the whole traffic load. Through extensive event-based simulations, we outlined the performance gains achieved by our DRAC-SA method in terms of transmit power minimization, TSR for mobile users and BBU savings compared to state-of-the-art schemes. We hence believe that our approach represents a promising solution for centralized resource allocation in future C-RAN deployments.

In the third contribution, our focus was on addressing jointly the resource allocation and admission control tasks in a fronhaul-constrained C-RAN, considering mobile users QoS requirements, interference and fronthaul network limitation constraints. We presented a three-level algorithm design to solve the resource allocation problem and the admission control task. Numerical results have confirmed the performance of our proposed RAAC approach, which increases users' admission and saves more C-RAN transmission power compared to related strategies. We also highlighted the benefits of our proposal regarding the number of BBUs' reduction in the cloud. We believe that our proposal can be readily used by operators for network dimensioning and for leveraging the C-RAN's benefits for dynamic multi-user resource allocation.

In the final contribution, we have provided a novel C-RAN BBU selection framework with different optimization schemes to help operators choose, from a variety of CSPs, the best BBUs to instantiate in their virtual pool. We have presented a powerful algorithm based on the Branch-and-Price scheme to optimally solve the problem in minimum time for different network sizes. The different introduced schemes in part of this framework have shown adaptiveness and flexibility to MNOs strategies, based on their prevailing requirements in terms of processing power minimization, resiliency, virtualization budget and adjustment to cells traffic profiles. More precisely, the presented EWOS and 424-S approaches have shown their ability to adapt to most traffic load scenarios and answer MNOs' major strategies and constraints. Especially, the 424-S may be the best

option if a MNO targets to instantiate 50% less BBUs, and realize 37.5% of BBU power savings and 48% of virtualization expenditure economy.

6.2 Future work

Hereafter, we expose the future work related to our thesis. In fact, we categorize the work perspectives into three groups: i) short-term, ii) medium-term and iii) long-term views.

6.2.1 Short-term perspectives

As a short-term planned work, we aim to consolidate our proposals by considering the problem of BBUs placement in centralized virtual pools, subject to MNOs functional splittings. In fact, as different functional splits can yield to different BBU-RRH distances, latencies and fronthaul links rates [21], they impose on the downside a cost/availability versus performance tradeoff, because the split limits the ability of the RAN to manage interference. MNOs with specific RAN splittings must then decide the best placement strategies of BBUs in their geographical-fixed BBU hotels, that can meet the fronthaul and latency requirements of their RAN splittings while handling RRHs traffic loads. A work in this direction would be to find the minimum number of BBUs to instantiate in each BBU hotel so as to maximize the RRHs traffic loads catering, for both central and edge traffic, subject to cell site fronthaul distance and latency constraints.

6.2.2 Medium-term perspectives

In this thesis, all performance evaluations have been performed through simulations. However, despite the quality of obtained results depicting the efficiency of our algorithms, testbed experiments are required to validate these algorithms and to evaluate their performance in real LTE environment. C-RAN can benefit from Software-Defined Radio (SDR) implementations [37], that deport most signal processing functions on software packages for emulating PHY/MAC functions without the need of using dedicated hardware [91] [92]. Currently, we have started building a SDR testbed [93] based on OpenAirInterface package softwares [39] and Ettus Research's Universal Software Radio Platform (USRP) devices [94]. The goal is to emulate a real-time LTE resource scheduler by leveraging the openness brought by SDR platforms using General Purpose Platforms (GPP) PC. The drawback however of using GPP is the latency requirement that can lead to occasional loss of baseband processing packets. Hence, as future medium-term works, we envisage to build a SDR-based C-RAN that can implement and validate our propositions with proper timing requirements, while also working on other algorithms for emulating LTE UL resource allocation, centralized handover management and BBUs collaborative radio processing.

6.2.3 Long-term perspectives

One vision of 5G is to create a cell-less environnement, where a user is not only connected to a single cell, but to a cloud of cells. In this regard, C-RAN is sure to be a cornerstone of 5G deployments to address its performance needs and optimize deployment costs. However, the centralized model may not be the best solution due to the current fronthaul design not meeting the delay requirement in retransmission, thereby lowering application data throughput. Meanwhile, the decentralized approach, using site-by-site computing, may have some merit in 5G to enable reality-time, IoT and Smart Vehicles [95] applications that will come with ultra-high capacity and incredibly low latency requirements. Therefore, as long-term perspective work, we aim to study the balance between the centralized and distributed architectures in 5G, with respect to resource allocation. In fact, our proposed resource allocation algorithms do not cover uplink scenarios, whose traffics may be exceeding the downlink ones in 5G. A more generalized uplink-downlink algorithm for resource allocation, with multiple BBU pools, and possibly other services present in the network, will be an exciting future research topic for 5G C-RAN.

Moreover, as we have studied BBU virtualization and baseband resource allocation, methods for allocating other resource elements, like fronthaul or cell site equipments, will be of great importance in 5G too. What is striking, the topic of network sharing is connected with hybrid fronthaul and backhaul optimization. Future mobile networks will most likely consist of standalone 5G deployments as well as LTE/LTE-A C-RANs. Joint capacity and control plane optimization for both fronthaul networks will enable more efficient usage of resources, thereby lowering network deployment and operation costs.

6.3 Publications

This section summarizes the publications that have resulted from the work undertaken in this thesis.

- **Journals**
  - M. Y. Lyazidi, L. Giupponi, J. Mangues, N. Aitsaadi and R. Langar, “An Optimization Scheme for Cost-Resilience BBU Selection in Cloud-Radio Access Network”, **submitted** in IEEE Transactions on Wireless Communications 2017.
  - M. Y. Lyazidi, N. Aitsaadi, R. Langar, P. Rubin, and R. Boutaba, “Dynamic Resource Allocation and Admission Control in Downlink Cloud Radio Access Network with Fixed Fronthaul Capacity”, **submitted** in IEEE Transactions on Mobile Computing 2017.
  - M. Y. Lyazidi, N. Aitsaadi, and R. Langar, “A Dynamic Resource Allocation Framework in LTE Downlink for Cloud-Radio Access Network”, **submitted** in IEEE Transactions on Vehicular Technology 2017.

- **Conferences**
  - M. Y. Lyazidi, L. Giupponi, J. Mangues, N. Aitsaadi and R. Langar, "A Novel Optimization Framework for C-RAN BBU Selection based on Resiliency and Price", IEEE Vehicular Technology Conference (VTC-Fall'17), Toronto, Canada, September 2017.
  - M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Resource Allocation and Admission Control in OFDMA-based Cloud-RAN", IEEE Global Communication Conference (GLOBECOM'16), Washington D.C., U.S., December 2016.
  - M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic Resource Allocation in Cloud-RAN with Real-Time BBU/RRH Assignment", IEEE International Conference on Communications (ICC'16), Kuala Lumpur, May 2016.

List of figures

List of tables

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast 2015-2020, white paper, available online at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf," 2016.

[2] Ericsson, "Vestberg Foresees Industry Shift, press release, http://www .ericsson.com/news/1384303," 2010.

[3] MarketingCharts, "Mobile network operators face cost crunch, available online: http://www.marketingcharts.com/wp/direct/mobile-network-operators-face-cost-crunch-17700/," 2011.

[4] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward Green and Soft: a 5G perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, 2014.

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

[6] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1888–1906, 2013.

[7] N. Alliance, "Further study on critical C-RAN technologies, available online at: https://www.ngmn.org/publications/all-downloads/article/project-ran-evolution-further-study-on-critical-c-ran-technologies.html," 2015.

[8] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "EASE: EPC as a service to ease mobile core network," *IEEE Network Magazine, Vol. 29, No. 2, pp.78 - 88*, 2015.

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

[10] Nokia, "Multi-layer and cloud-ready radio evolution towards 5G, 2016, white paper, available online at: https://gsacom.com/paper/multi-layer-cloud-ready-radio-evolution-towards-5g," 2016.

[11] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. W¸bben, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, 2014.

[12] IJOIN, "Interworking and JOINt Design of an Open Access and Backhaul Network Architecture for Small Cells based on Cloud networks, website: http://www.ict-ijoin.eu," 2012.

[13] MCN, "Mobile Cloud Networking (MCN) project. [online]. available: https://www.mobile-cloud-networking.eu/," 2013.

[14] HARP, "FP7 project High capacity network Architecture with Remote radio heads and Parasitic antenna arrays (HARP). [online]. available: http://www.fp7-harp.eu/," 2014.

[15] NTT-DoCoMo, "Special articles on 5G technologies toward 2020 deployment," *Technical Journal, Vol. 17, No. 4*, 2016.

[16] CPRI, "Common public radio interface (CPRI): Interface specification, version 6.1. online at http://www.cpri.info," 2014.

[17] C.-L. I, J. Huang, R. Duan, C. Cui, J. X. JIiang, and L. Li, "Recent Progress on C-RAN Centralization and Cloudification," *Special section on recent advances in Cloud Radio Access Networks*, 2014.

[18] N. J. Gomesa, P. Chancloub, P. Turnbullc, A. Mageec, and V. Jungnickeld, "Fronthaul evolution: From CPRI to ethernet," *Optical Fiber Technology - Next Generation Access Networks*, vol. 26, Part A, pp. 50 – 58, 2015.

[19] 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.

[20] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks - a Technology Overview," *IEEE Communications Surveys and Tutorials*, 2013.

[21] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings," *Journal of Communications and Networks*, vol. 18, no. 2, pp. 162–172, 2016.

[22] R. Agrawal, A. Bedekar, T. Kolding, and V. Ram, "Cloud RAN challenges and solutions," *International ICIN Conference - Innovations in Clouds Internet and Networks*, 2016.

[23] "Small cell virtualization functional splits and use cases," technical report, Small Cell Forum, 2015.

[24] R. Agrawal, A. Bedekar, S. Kalyanasundaram, T. Kolding, H. Kroener, and V. Ram, "Architecture principles for cloud-ran," in *IEEE Vehicular Technology Conference (VTC-Spring)*, 2016.

[25] 3GPP, "TR 36.819 study on coordinated multi-point operation for LTE," *available at www.3gpp.org*, 2011.

[26] ——, "RP-111454: Status report to TSG; enhanced ICIC for non-CA based deployments of heterogeneous networks for LTE - performance part," *TSG RAN meeting n.54*, 2011.

[27] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, "The Case for Reconfigurable Backhaul in Cloud-RAN based Small Cell Networks," *IEEE INFOCOM*, 2013.

[28] G. T. 36.211, "Physical channels and modulation, 3GPP," *www.3gpp.org*, 2008.

[29] Q. Wang, D. Jiang, J. Jin, G. Liu, Z. Yan, and D. Yang, "Application of BBU+RRU based comp system to lte-advanced," in *IEEE International Conference on Communications (ICC) Workshops*, 2009.

[30] EASY-C, "Enablers for ambient services and systems part c - wide area coverage, [online]. available: http://www.easy-c.com."

[31] ARTIST4G, "Artist4g project website, [online]. available: www.ict-artist4g.eu."

[32] FUTON, "Fibre optic networks for distributed, extendible heterogeneous radio architectures and service provisioning [online]. available: http://www.ict-futon.eu."

[33] 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.

[34] 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.

[35] Y. Yu, E. Dutkiewicz, X. Huang, and M. Mueck, "Downlink Resource Allocation for Next Generation Wireless Networks with Inter-Cell Interference," *IEEE transactions on Wireless Communications, Vol. 12, No. 4*, 2013.

[36] 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.

[37] R. J. Y. Dayene Beyene and K. Ruttik, "Cloud-RAN Architecture for Indoor DAS," *Special section on recent advances in Cloud Radio Access Networks*, 2014.

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

[39] "Openairinterface software alliance, EURECOM, China Mobile, Orange, BUPT, IITH, OPNFV – Openairinterface an upstream project for OPNFV. online at http://www.openairinterface.org/," 2016.

[40] 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.

[41] 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.

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

[43] 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.

[44] ——, "A dynamic resource allocation framework in LTE downlink for cloud-radio access network," *submitted to IEEE Transactions on Vehicular Technology*, 2017.

[45] ——, "Resource allocation and admission control in OFDMA-based Cloud-RAN," *IEEE Global Communications Conference (GLOBECOM)*, 2016.

[46] M. Y. Lyazidi, N. Aitsaadi, R. Langar, P. Rubin, and R. Boutaba, "Dynamic resource allocation and admission control in downlink cloud radio access network with fixed fronthaul capacity," *submitted to IEEE Transactions on Mobile Computing*, 2017.

[47] "100% availability for cloud and data center applications. [online]. available: http://www.cedexis.com/."

[48] D. Rayburn, "Current state of the CDN market: DIY, pricing trends, competitive dynamics, may 2016, available online at: www.cdnpricing.com," 2016.

[49] C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. Savelsbergh, and P. H. Vance, "Branch-and-price: Column generation for solving huge integer programs," *Operations research*, vol. 46, no. 3, pp. 316–329, 1998.

[50] M. Y. Lyazidi, L. Giupponi, J. Mangues-Bafalluy, N. Aitsaadi, and R. Langar, "A novel optimization framework for C-RAN BBU selection based on resiliency and price," in *IEEE Vehicular Technology Conference (VTC 2017-Fall)*, 2017.

[51] ——, "An optimization scheme for cost-resilience BBU selection in cloud-radio access network," *submitted to IEEE Transactions on Wireless Communications*, 2017.

[52] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications magazine*, vol. 46, no. 9, 2008.

[53] C. Nam, C. Joo, S. G. Yoon, and S. Bahk, "Resource allocation in full-duplex ofdma networks: Approaches for full and limited csis," *Journal of Communications and Networks*, vol. 18, no. 6, pp. 913–925, 2016.

[54] D. Lopez-Perez, A. Ladanyi, A. Jüttner, H. Rivano, and J. Zhang, "Optimization method for the joint allocation of modulation schemes, coding rates, resource blocks and power in self-organizing lte networks," in *IEEE INFOCOM Proceedings*, 2011.

[55] J. Xiao, R. Q. Hu, Y. Qian, L. Gong, and B. Wang, "Expanding lte network spectrum with cognitive radios: From concept to implementation," *IEEE Wireless Communications, Vol. 20, No.2, pp.12 - 19*, 2013.

[56] K. Fazel and S. Kaiser, "Multi-carrier and spread spectrum systems," *Wiley*, 2008.

[57] S. K. Gond and A. Sai Prasad, "Green antenna and radio over fiber technology for a cellular wireless," *International journal of Engineering and Innovative Technology (IJEIT), ISSN*, pp. 2277–3754, 2012.

[58] C. Rowell, S. Han, Z. Xu, and I. C. Lin, "Green RF technologies for 5G networks," in *IEEE International Wireless Symposium (IWS)*, 2014.

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

[60] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 494–508, 2015.

[61] M. Schubert and H. Boche, "Iterative multiuser uplink and downlink beamforming under sinr constraints," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2324–2334, 2005.

[62] T. X. Tran and D. Pompili, "Dynamic radio cooperation for downlink cloud-rans with computing resource sharing," in *IEEE International Conference onMobile Ad Hoc and Sensor Systems (MASS)*, 2015.

[63] K. Wang, M. Zhao, and W. Zhou, "Graph-based dynamic frequency reuse in cloud-ran," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2014.

[64] K. Wang, K. Yang, and C. Magurawalage, "Joint energy minimization and resource allocation in c-ran with mobile cloud," *IEEE Transactions on Cloud Computing*, 2016.

[65] OBSAI, "Reference point 3 specification, version 4.2. online at http://www.obsai.com."

[66] A. Abdelnasser and E. Hossain, "Two-tier ofdma cellular cloud-ran: Joint resource allocation and admission control," in *IEEE Global Communications Conference (GLOBECOM)*, 2015.

[67] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[68] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA cloud-RAN of small cells underlaying a macrocell," *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2837–2850, 2016.

[69] V. N. Ha and L. B. Le, "Resource allocation for uplink OFDMA C-RANs with limited computation and fronthaul capacity," in *IEEE International Conference on Communications (ICC)*, 2016.

[70] V. Nguyen and L. B. Le, "Joint Coordinated Beamforming and Admission Control for Fronthaul Constrained Cloud-RAN," *IEEE Global Communications Conference (GLOBECOM)*, 2014.

[71] V. N. Ha, L. B. Le, and N. D. Dao, "Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2014.

[72] ——, "Energy-efficient coordinated transmission for cloud-RANs: Algorithm design and trade-off," in *Annual Conference on Information Sciences and Systems (CISS)*, 2014.

[73] J. Zhao, T. Q. S. Quek, and Z. Lei, "Fast algorithm for utility maximization in C-RAN with joint QoS and fronthaul rate constraints," in *IEEE International Conference on Communications (ICC)*, 2016.

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

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

[76] C. Colman-Meixner, G. B. Figueiredo, M. Fiorani, M. Tornatore, and B. Mukherjee, "Resilient cloud network mapping with virtualized BBU placement for Cloud RAN," *IEEE International Conference on Advanced Networks and Telecommunications Systems ANTS*, 2016.

[77] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband processing units virtualization for cloud radio access networks," *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 189–192, 2015.

[78] M. Bouet, J. Leguay, and V. Conan, "Cost-based placement of vdpi functions in nfv infrastructures," in *IEEE Conference on Network Softwarization (NetSoft)*, 2015.

[79] N. Carapellese, M. Tornatore, and A. Pattavina, "Placement of base-band units (bbus) over fixed/mobile converged multi-stage wdm-pons," in *International Conference on Optical Networking Design and Modeling (ONDM)*, 2013.

[80] A. Checko, G. Kardaras, C. Lanzani, D. Temple, C. Mathiasen, L. Pedersen, and B. Klaps, *OTN Transport of Baseband Radio Serial Protocols in C-RAN Architecture for Mobile Network Applications*. Altera, 2014.

[81] 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 *Online Conference on Green Communications (OnlineGreenComm)*, 2013.

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

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

[84] 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.

[85] M. E. Aydin, R. Kwan, J. Wu, and J. Zhang, "Multiuser scheduling on the LTE downlink with simulated annealing," in *IEEE Vehicular Technology Conference (VTC-Spring)*, 2011.

[86] A. P. Iserte, A. I. Perez-Neira, and M. A. L. Hernandez, "Joint transceiver optimization in wireless multiuser MIMO-OFDM channels based on simulated annealing," in *European Signal Processing Conference*, 2002.

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

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

[89] N. S. Networks, "LTE smartphones measurements," 2011. [Online]. Available: http://networks.nokia.com/system/files/document/lte_measurements_final.pdf

[90] A. R. Kan, L. Stougie, and C. Vercellis, "A class of generalized greedy algorithms for the multi-knapsack problem," *Discrete applied mathematics*, vol. 42, no. 2-3, pp. 279–290, 1993.

[91] "OpenLTE, an open source implementation of the 3GPP LTE specifications. online at: http://openlte.sourceforge.net."

[92] "srsLTE, an open source 3GPP LTE library. online at https://github.com/srslte/srslte."

[93] "SDR LAB, a software defined radio lab. Online at: https://sdr-lab.lip6.fr."

[94] "Ettus research, a National Instrument company. URL: https://www.ettus.com."

[95] T. N. Pham, M. F. Tsai, D. B. Nguyen, C. R. Dow, and D. J. Deng, "A cloud-based smart-parking system based on internet-of-things technologies," *IEEE Access*, vol. 3, pp. 1581–1591, 2015.