RESOURCE MANAGEMENT IN CLOUD RADIO ACCESS NETWORKS

by

Lilatul Ferdouse

MASc., Ryerson University, 2015
M.S., Dhaka University, 2007

A Dissertation
presented to Ryerson University
in partial fulfilment of the requirements for the degree of Doctor of Philosophy
in the program of Electrical and Computer Engineering

Toronto, Ontario, Canada, 2019

AUTHORS DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Abstract

RESOURCE MANAGEMENT IN CLOUD RADIO ACCESS NETWORKS

© Lilatul Ferdouse, 2019
Doctor of Philosophy
Electrical and Computer Engineering
Ryerson University

This thesis focuses on resource management both in communication and computing sides of the cloud radio access networks (C-RANs). Communication and computing resources are bandwidth, power, baseband unit servers, and virtual machines, which become major resource allocation elements of C-RANs. If they are not properly handled, they create congestion and overload problems in radio access network and core network part of the backbone cellular network. We study two general problems of C-RAN networks, referred to as communication and computing resource allocation problem along with user association, base band unit (BBU) and remote radio heads (RRH) mapping problems in order to improve energy efficiency, sum data rate and to minimize delay performance of C-RAN networks.

In this thesis, we propose, implement, and evaluate several solution strategies, namely posterior probability based user association and power allocation method, double-sided auction based distributed resource allocation method, the energy efficient joint workload scheduling and BBU allocation and iterative resource allocation method to deal with the resource management problems in both orthogonal and non-orthogonal multiple access supported C-RAN networks. In the posterior probability based user association and power allocation method, we apply Bayes theory to solve the multi-cell association problem in the coordinated multi-point supported C-RANs. We also use queueing and auction theory to solve the joint communication and computing resource optimization problem. As the joint optimization problem, we investigate the delay and sum data rate performance of C-RANs. To improve the energy efficiency of C-RANs, we employ Dinkelbach theorem and propose an iterative resource allocation method. Our proposed methods are evaluated via simulations by considering the effect of bandwidth utilization percentage, different scheduling weight, signal-to-interference ratio threshold value and number of users. The results show that the proposed methods can be successfully implemented for 5G C-RANs.

Among the various non-orthogonal multiple access schemes, we consider and implement the sparse code multiple access (SCMA) scheme to jointly optimize the codebook and power allocation in the downlink of the C-RANs, where the utilization of sparse code multiple access in C-RANs to improve energy efficiency has not been investigated in detail in the literature. To solve the NP-hard joint optimization problem, we decompose the original problem into two subproblems: codebook allocation and power allocation. Using the graph theory, we propose the throughput aware sparse code multiple access based codebook selection method, which generates a stable codebook allocation solution within a finite number of steps. For the power allocation solution, we propose the iterative level-based power allocation method, which incorporates different power allocation approaches (e.g., weighted and successive interference cancellation) into different levels to satisfy the maximum power requirement. Simulation results show that the sum data rate and energy efficiency performance of non-orthogonal multiple access supported C-RANs significantly increases with the number of users when the successive interference cancellation aware geometric water-filling based power allocation is used.

Acknowledgment

I would first like to thank my supervisor, Professor Alagan Anpalagan for his support, patience and time throughout my graduate studies. I am finally able to achieve my goal with his strong support and help. He did not only guide me in the course of my thesis, but also provided me an opportunity to benefit from his vast knowledge. He was always present to help me out and to guide me whenever I needed his guidance.

I would like to thank my thesis committee members, Professor Muhammad Jaseemuddin, Professor Olivia Das and Professor Jelena Misic, for taking the time and effort to review my thesis and provide me with their insightful comments. My deep appreciation go out to Professor Ben Liang, University of Toronto, for agreeing to act as my external examiner and provide me with his valuable comments. I would also like to acknowledge Dr. Serhat Erkucuk and Dr. Waleed Ejaz for helping with my research works. Furthermore, I would also like to acknowledge my gratitude to the Department of Electrical and Computer Engineering and the School of Graduate Studies of Ryerson University, for their financial support and all the opportunities.

I would like to thank my parents and siblings for their support, love and advice. Without them, I would not be here today. Finally, and most importantly, I would like to thank my husband Mohammed Zahirul Hoq Sarker, who has been a constant source of encouragement for me throughout my studies. Without his support, courage and adorableness, I would not have been able to achieve my goal.

Contents

List of Figures

List of Tables

List of Abbreviations

BBU	Base Band Unit
C-RAN	Cloud Radio Access Network
Cloud-RAN	Cloud Radio Access Network
CN	Core Network
CoMP	Co-ordinate Multi-point Access
CAPEX	Capital Expenditure
CB	Codebook
D2D	Device-to-Device communication
DS-ADRA	Double-Sided Auction based Distributed Resource Allocation
EE	Energy efficiency
EE-JWSBA	Energy Efficient Joint Workload Scheduling and BBU Allocation
eMBB	enhanced Mobile Broadband
ILPA	Iterative Level-based Power Allocation
H-CRAN	Heterogeneous Cloud Radio Access Network
mMTC	Massive Machine Type Communications
MBS	Macro cell Base Station
MUE	Macro cell User Equipment
M2M	Machine-to-Machine communication
NOMA	Non-orthogonal Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OMA	Orthogonal Multiple Access
OPEX	Operating Expenditure

PA	Power Allocation
RA	Resource Allocation
P2UPA	Posterior Probability based User Association and Power Allocation
RRH	Remote Radio Head
RAN	Radio Access Network
RB	Resource Block
SBS	Small cell Base Station
SE	Spectrum Efficiency
SUE	Small cell User Equipment
SCMA	Sparse Code Multiple Access
SIC	Successive Interference Cancellation
SINR	Signal-to-Interference-Noise Ratio
SNR	Signal-to-Noise Ratio
TASCBS	Throughput Aware SCMA CB Selection
UEs	User Equipments
URLLC	Ultra-Reliable Low-Latency Communications

Chapter 1

Introduction

1.1 Motivation and Objective

The ever-growing use of smart phones and portable devices such as tablets, smart watches, increases the growth of cellular Internet data traffic exponentially. According to the Cisco visual networking report, the global mobile data traffic will show 53 percent compound annual growth rate from 2015 to 2020, within it 75 percent data will be video, producing 30.6 extra bytes per month by 2020 [1]. It is anticipated that within the next five years, 11.6 billion of connected devices including IoT devices, will increase network connection speeds by more than threefold and, the number of mobile-connected devices per capita will reach 1.5 by 2020 [1].

To deal with ever-increasing demand of user association and resource allocation in cellular networks, the architecture of cloud radio access networks (C-RAN) is envisioned as an attractive paradigm that takes advantage of managing large number of small cells through the centralized cloud controller, known as base band processing unit or BBU pool. Fig. 1.1 depicts a small cell based C-RAN architecture where remote radio heads (RRHs) are responsible for RF signal transmission from/to users in the small cell and to/from baseband unit (BBU) pool through fronthaul links. The access requests of users are transmitted from RRH to BBU pool for baseband processing. To satisfy the demand for large bandwidth and data rate, the optical fiber is generally considered as an ideal fronthaul link for C-RAN, whereas wired and wireless links support C-RAN backhaul [2]. The inspiring factor for such a centralized structure of C-RAN is to minimize capital

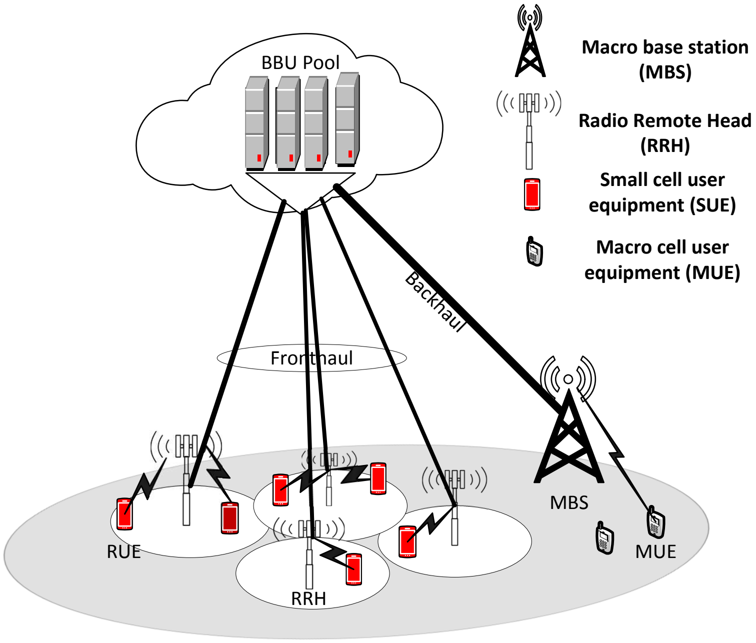


Figure 1.1: Small cell based C-RAN architecture for 5G networks.

expenditure (CAPEX) and operating expenditure (OPEX) cost as well as support scalability and flexibility of deployment of RRHs [3]. However, user association, cell activation, dynamic resource allocation based on users quality of service (QoS) requirements, workload scheduling in BBU pool, BBU-RRH mapping etc. are the major challenging issues in C-RANs.

This thesis deals with two general problems of C-RAN systems, referred to as communication and computing allocation problem along with user association, BBU-RRH mapping to improve energy efficiency, sum data rate and delay minimization of C-RAN networks. Communication and computing resources are bandwidth, power and baseband unit servers, etc. which become a major resource allocation elements of C-RANs, because, if not properly handled, they create congestion and overload problem in radio access network (RAN) and core network (CN) part of the backbone cellular network. In this thesis, we have proposed, implemented, and evaluated several solution strategies, namely posterior probability based user association and power allocation (P2UPA) method, double-sided auction based distributed resource allocation (DS-ADRA) method, energy efficient joint workload scheduling and BBU allocation (EE-JWSBA) and iterative resource allocation to deal with these problems in orthogonal multiple access and non-orthogonal multiple access supported C-RAN networks.

1.2 Major Contributions

Part I: User Association

The architecture of cloud radio access networks (C-RANs) is an attractive paradigm of 5G that takes advantages of both centralized baseband and coordinated multi-point (CoMP) processing in radio access networks. In C-RAN, the data rate provisioning can be significantly improved due to the fractional frequency reuse performed by small cells. The dense deployment of small cells, however, incurs severe inter-tier and inter-cell interference turning the user association into a more challenging problem. Moreover, multi-cell association problem occurs in CoMP and control/user planes (C/U planes) splitting based C-RAN network where users are associated with more than one cell to support joint-transmission and reception. In the first work (in Chapter 3) of this dissertation, we consider a multi-cell user association approach taking into account the data rate and aggregate interference of mobile users. We propose to use posterior probability based user association and power allocation (P2UPA) method that depends on prior knowledge of the channel state information (CSI). The objective of the proposed method is to maximize the sum data rate of small cell users while maintaining the constraints of aggregate interference, power consumption, and data rate among small cell users. Then, the sum data rate and energy efficiency performance of P2UPA are evaluated through simulations.

Part II: Computing Resource Allocation

We investigate the workload scheduling and BBU allocation problem (in chapter 4). In C-RAN, two type of resources are considered for allocation: i) computation resources (e.g., GPU, BBU servers etc.) and ii) transmission resources (e.g., spectrum, bandwidth, frequency etc.). Efficient management of cloud resources is one of the important challenges in C-RAN. We investigate a joint workload scheduling and baseband unit (BBU) allocation in C-RAN. Firstly, we consider the C-RAN queueing model. Then, we formulate an optimization problem for joint workload scheduling and BBU allocation with the aim to minimize mean response time and aggregate power. Queueing stability and workload conservation constraints are considered in the optimization problem. To solve this problem, we propose an energy efficient joint workload scheduling and BBU allocation (EE-JWSBA) algorithm using queueing theory. The EE-JWSBA algorithm is evaluated via simulations by considering three different scheduling weights (e.g., random, normalized, and upper limit). Simulation results demonstrate the effectiveness of proposed scheme using different scheduling weights.

Part III: Joint Communication and Computing Resource Allocation

In this work, we investigate joint communication and computing resource allocation along with user association, and baseband unit (BBU) and remote radio head (RRH) mapping in two-tier orthogonal frequency division multiple access (OFDMA) supported C-RAN system (in chapter 5). We initially establish a queueing model in C-RAN, followed by formulations of two optimization problems for communication (e.g., resource blocks (RBs) and power) and computing (e.g., virtual machines (VMs)) resources allocation with the aim to minimize mean response time. User association along with the RB allocation, interference and queueing stability constraints are considered in the communication resource optimization problem. The computing resource optimization problem considers BBU-RRH mapping and VM allocation for small cells, constrained to BBU server capacity and queueing stability. To solve the communication and computing resource optimization problem, we propose a joint resource allocation solution using a double-sided auction based distributed resource allocation (DS-ADRA) method, where small cell base stations and users jointly participate using the concept of auction theory. The proposed method is evaluated via simulations by considering the effect of bandwidth utilization percentage, signal-to-interference ratio threshold value and number of users.

Part IV: Energy Efficient Communication Resource Allocation

For the energy efficient communication resource (e.g., bandwidth and power) allocation, we consider orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) supported C-RANs(in chapters 6 and 7). In the OMA supported C-RAN, we investigate an underlaid approach of resource allocation for small and macro cell users to improve energy efficiency (EE). We initially provide a energy efficient resource allocation, followed by the formulation of an optimization problem for communication (e.g., resource blocks and power) resource allocation for both small and macro cell users with the aim to maximize EE in the whole C-RAN system. The RB along with power allocation, interference, quality of service (QoS) of macro cell users, and front-haul capacity constraints are considered in this optimization problem. The joint nature of RB and power allocation both in macro and small cell users turns the resource allocation problem into a computationally intractable NP-hard problem. To solve the optimization problem, we transfer the baseline problem into a relaxed problem with the time-sharing approach of resource allocation, and propose an iterative resource allocation solution using the Dinkelbach theorem. The proposed method is evaluated in terms of energy efficiency, sum data rate and Jain fairness index considering the effect of number of user association in C-RAN.

Then, in the NOMA supported C-RAN, we consider the sparse code multiple access (SCMA) scheme to jointly optimize the codebook (CB) and power allocation in the downlink of the C-RANs to improve energy efficiency. To solve the NP-hard joint optimization problem, we decompose the original problem into two sub-problems: codebook allocation and power allocation. We propose the throughput aware SCMA CB selection (TASCBS) method, which generates a stable codebook allocation solution within a finite number of steps. For the power allocation solution, we propose the iterative level-based power allocation (ILPA) method, which incorporates different power allocation approaches (e.g., weighted and NOMA successive interference cancellation (SIC)) into different levels to satisfy the maximum power requirement. Simulation results show that the sum data rate and energy efficiency performance of SCMA supported C-RANs depend on the selected power allocation approach. In terms of energy efficiency, the performance significantly increases with the number of users when the NOMA-SIC aware geometric water-filling based power allocation is used.

1.3 Thesis Organization

Chapter	Research Problem	Description
`Chapter 2`	Research Background	-Overview of Cloud-RAN -Different Cloud-RAN architectures -Optimization issues and research challenges in Cloud-RAN

Chapter	Research Problem	Description
`Chapter 3`	User/cell association problem	Addressed the following question: -Within the C-RAN, Which cells users will be associated? -Is it one-to-one or one-to-multi cells association? -For multicell association method, which BS/RRH users will select for data transfer? -RRH/BS selection method for C-RAN. **Used Methodology:** Posterior probability based multiple cell selection method.
`Chapter 4`	Workload scheduling and BBU allocation problem	Addressed the following question: -How computation resources e.g. BBU servers will be allocated among the RRHs? -How the workload are distributed among the BBU servers? **Used Methodology:** The energy-efficient joint workload scheduling and BBU allocation method.
`Chapter 5`	Joint communication and computing resource allocation problem	Addressed the following question: -Within the C-RAN, Which cells/RRH users will be associated and how power and radio resources are distributed among the C-RAN users? -How interference are managed during the resource allocation? - How computation resources e.g. BBU servers will be allocated among the RRH? **Used Methodology:** Auction based resource allocation method.

Chapter	Research Problem	Description
`Chapter 6`	Energy efficient communication resource allocation in OFDM supported C-RANs	Addressed the following question: -How to address energy efficient aspect of resource allocation in C-RANs? -In the underlaid approach of resource allocation, how power and radio resources are distributed among the small and macro cell user? **Used Methodology:** Two-step iterative resource allocation method.
`Chapter 7`	Energy efficient communication resource allocation in SCMA supported C-RANs	Addressed the following question: -Investigate energy efficient aspect of SCMA codebook and power allocation strategies in C-RANs. -How power and codebooks are distributed among the small cell C-RAN users? -How interference are managed during the resource allocation? **Used Methodology:** Throughput aware SCMA CB selection and iterative level-based power allocation method

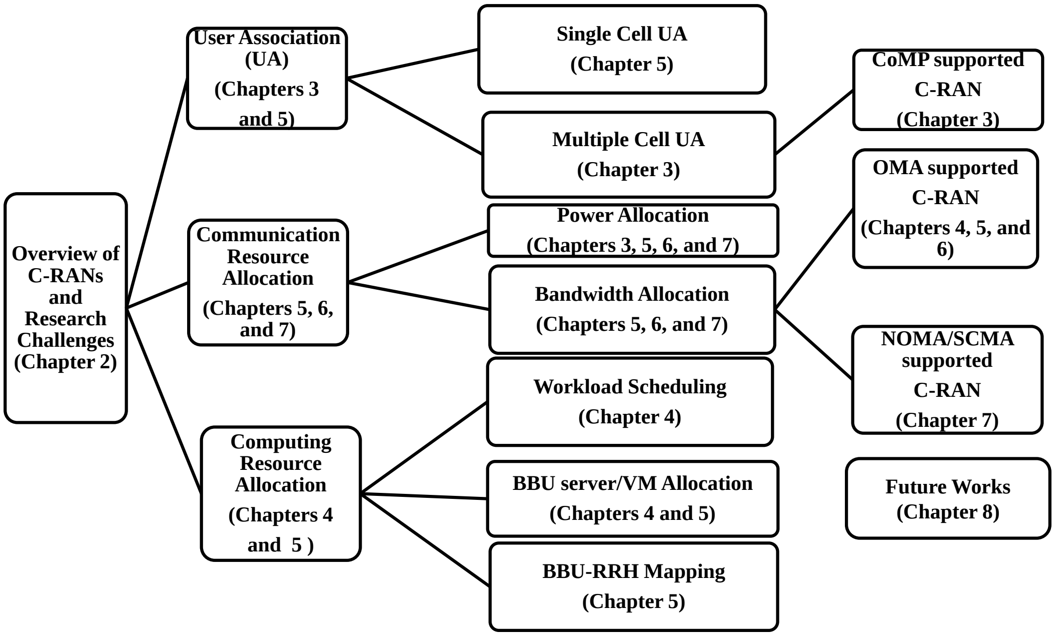

Overview of C-RANs and Research Challenges (Chapter 2)
User Association (UA) (Chapters 3 and 5)
Communication Resource Allocation (Chapters 5, 6, and 7)
Computing Resource Allocation (Chapters 4 and 5)
Single Cell UA (Chapter 5)
Multiple Cell UA (Chapter 3)
Power Allocation (Chapters 3, 5, 6, and 7)
Bandwidth Allocation (Chapters 5, 6, and 7)
Workload Scheduling (Chapter 4)
BBU server/VM Allocation (Chapters 4 and 5)
BBU-RRH Mapping (Chapter 5)
CoMP supported C-RAN (Chapter 3)
OMA supported C-RAN (Chapters 4, 5, and 6)
NOMA/SCMA supported C-RAN (Chapter 7)
Future Works (Chapter 8)

Figure 1.2: Thesis outline.

Chapter 2

Background and State of the Art

2.1 Introduction

The increasing demand on smart device applications along with the ever increasing number of cellular mobile users make the traditional cellular networks incapable of fully accommodating the over-whelming traffic loads in the future wireless networks. Utilizing small cells in cellular networks along with the overlaid macro cells improves the total network capacity due to the spatial frequency reuse achieved by the small cells. Moreover, the short transmission range between remote radio heads (RRHs) and mobile users increase the data rate experienced by users [4]. However, the dense deployment of small cells incurs high interference levels not only among small cells themselves, but also with macro cells. In addition, the capacity-limited fronthaul links act as a performance bottleneck in some cases.

To overcome the aforementioned problems, cloud radio access networks (C-RANs), which consist of remote radio heads and the baseband unit (BBU) pool, have emerged as a new paradigm that take advantages of both centralized baseband and coordinated multi-point processing (CoMP) in cloud radio access networks [5,6]. In C-RANs, the high-power nodes (HPNs) or macro base stations are responsible for providing seamless coverage, network control, and serving users with low data rates. In addition, C-RANs support high-level cooperative interference cancellation among small cells by centrally coordinating the performance of the entire network within the BBUs. On the other hand, the low-power RRHs that coexist with HPNs, provide

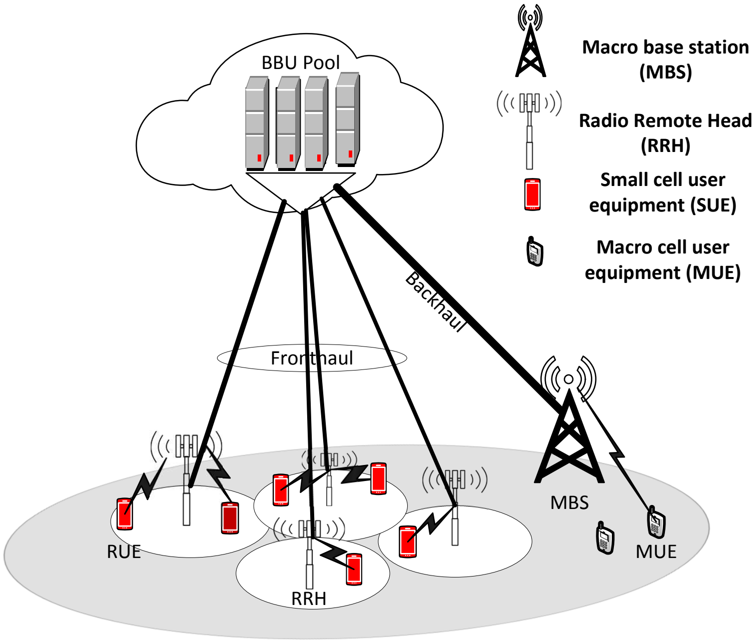


Figure 2.1: C-RAN architecture for 5G networks.

users with high data rates leaving the basic control operations such as handover management to macro cells, thus reducing the burden on the fronthaul links. Regarding the deployment costs, C-RANs are expected to reduce both capital and operation expenditures (CAPEX and OPEX).

2.2 Cloud-RAN Architecture

Cloud-RAN is regarded as architecture evolution of distributed base station system. The distributed base station system separates the base station into two parts: i) remote radio head, ii) baseband unit. Fig. 2.1 depicts a small cell-based C-RAN architecture. The generic cloud-RAN architecture consists the following three components:

Remote Radio Head (RRH): The radio function unit, also known as remote radio head, is responsible for RF signal transmission from/to users in the small cell and to/from baseband pool through fronthaul links.

Baseband Unit (BBU): The baseband unit is considered as a digital function unit. In C-RAN architecture, the BBU pool is considered as a master base station which pools the baseband processing function into a master base station to reduce the wastage of energy and resources of

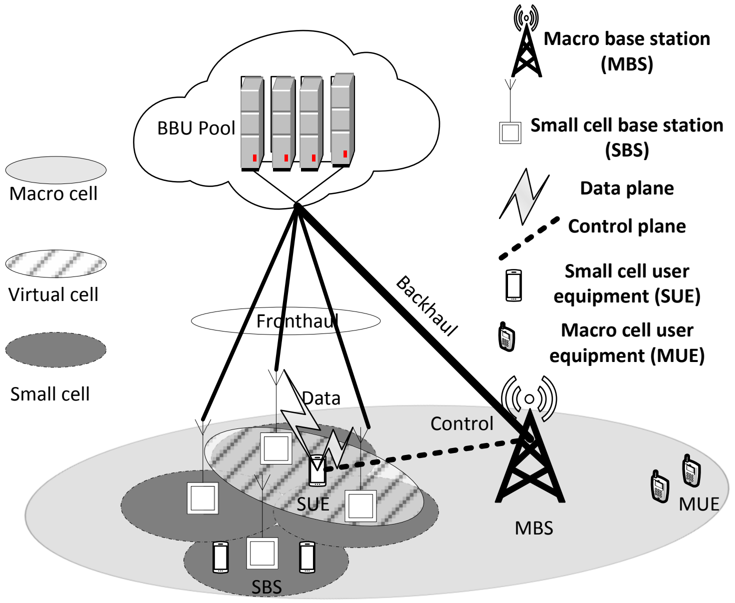


Figure 2.2: Two-tier C-RAN with separated control/data planes.

under-utilized base stations.

Fronthaul transmission link: The RRHs are connected to the BBU pool through the wire/wireless fronthaul links. To satisfy the demand for large bandwidth and data rate, the optical fiber is generally considered as an ideal fronthaul link for C-RAN, whereas wired and wireless links support C-RAN backhaul [2].

CoMP supported C-RAN: In CoMP supported C-RAN networks, user can be associated to more than one cell for data transmission. The software defied component of BBU pool handles the joint transmission and data processing with the cells and core networks. The authors in [7] considered the CoMP based C-RAN network to solve the resource allocation problem.

C/U split C-RAN: Multi-tier C-RAN networks or heterogeneous C-RAN networks (H-CRAN) sometime support separate plane in radio access network. One plane is used for data transmission, referred to as U-plane and other is used for control information transmission, referred to as C-plane. If the C-RAN supports separate data and control planes, the small cell users (SUEs) are connected with RRHs via the data plane, and with the macro base station (MBS) via the control plane. Macro BSs are assumed to remain active always to provide control and seamless coverage. Meanwhile, SUEs can establish more than one U-plane connection among SBSs to download or upload data.

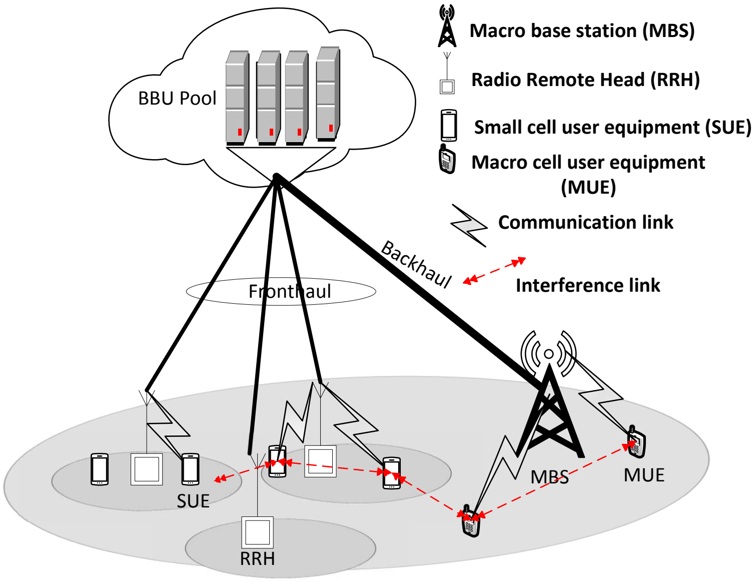


Figure 2.3: OMA or NOMA supported two-tier (C-RAN) with different interference links.

OMA supported C-RAN: It is envisioned that the 5G mobile communication networks will support 100 times more connected devices per unit area compared to 4G LTE networks. The LTE and LTE-A networks support the orthogonal multiple access (OMA) technique, which utilizes orthogonal resources for the users. Similarly, OMA supported C-RANs/H-CRANs utilize orthogonal resources for the small cell users. The radio resource blocks (RBs) or subcarrier (SCs) and power are considered as a communication resources in the OMA supported networks. The orthogonal resources are used to mitigate intra-cell interference. However, due to the limited number of orthogonal resources, an underlaid approach of orthogonal resource sharing and partial frequency reused are utilized to maximize the spectrum efficiency of the networks.

NOMA/SCMA supported C-RAN: The OMA approach supports limited number of connections due to the use of orthogonal resources. Therefore, to increase connections per unit area, non-orthogonal multiple access (NOMA) approach has been identified as a promising solution for future networks [8]. Unlike OMA, the NOMA methods utilize different power levels or overlapping signatures to provide services to multiple users. For example, the sparse code multiple access (SCMA), which is classified as one category of NOMA methods, assigns different codebooks (CBs) to different users. SCMA is regarded as the generalized low density signature (LDS), where LDS uses sparse spreading sequences and SCMA uses sparse codewords in the codebooks. Each codeword comprises of non-orthogonal resources such as sub-carriers (SCs) that are shared by different users [9]. The multiplexed signals of different users superimposed over the same sub-carrier can be decoded by the message passing algorithm [10] with low complexity. On the other hand, power domain NOMA (PD-NOMA) method uses different power levels for multiple users to provide services in the same sub-carrier and time slot [8]. The NOMA methods support massive connectivity and higher utilization of bandwidth and provide higher spectrum efficiency (SE) and energy efficiency (EE). However, the non-orthogonality in NOMA increases the mutual interference levels, therefore, successive interference cancelation (SIC) method is applied at the receiver side [11]. The details about SCMA codebook design and allocation method are presented in chapter 7.

2.3 Resource Optimization Issues in C-RANs

There are several challenges associated with C-RANs. For example, the optimization problems of radio access network (RAN) side includes user association, cell activation, communication resource allocation, RRH deployment [2], etc. and computing or cloud side includes BBU-RRH mapping [22] [18], BBU-RRH function visualization [23], BBU servers or virtual machines (VMs) allocation [24], etc. Moreover, communication and computing resources allocation are the major challenging tasks in C-RANs. Table 2.1 shows the comparative analysis of resource optimization problems, optimization types and solution approaches that are used in wireless networks, especially in cloud radio access networks.

2.3.1 User/Cell Association

The cell association or user association determines which cells or RRHs, the user will connect to for data transmission. On the other word, user association method refers to RRH or base station selection method in radio access network. The C-RAN proposed for 5G networks connects a large number of small cells from RRH through fronthaul link to BBU pool. Moreover, the CoMP supported C-RAN networks support coordinated multi cell radio access techniques where each user can be associated to more than one cell for joint transmission. Fig. 2.4 depicts the user association scenario in C-RAN networks. The efficient cell association method within the C-RAN networks address the following questions:

Table 2.1: Resource optimization problem and solution Approach

Ref.	**Research Problem**	**Networks**	**Optimization Type**	**Objective**	**Solution Approach**
[12] -2017	-User association -Resource allocation, -Power allocation	-Cognitive femtocell networks	-Mixed integer non-linear problem	-Maximize data rate	-Matching theory
[13] -2016	- Cell activation - User association - Spectrum allocation	-Heterogeneous networks (HetNet)	- l_0, l_1 norm approximation	-Energy efficiency, -Delay minimization	-Reweighed l_1 minimization approach
[14] -2016	-Cell/user association	-Cloud radio access networks (C-RAN)	-Stochastic geometric	-Maximize ergodic capacity	- Received power based approach: i) N-best, ii)Nth best, and iii)single best
[15] -2015	-Cell/user association - Cell on-off	-Heterogeneous networks (HetNet)	-0-1 Knapsack problem	-Minimize energy consumption	-Pricing based method, -Iterative algorithm
[16] -2016	-BBU allocation, -Computation resource allocation	-Cloud radio access networks (C-RAN)	-Lyapnunov optimization problem	-Energy efficiency	-Apply queueing model at remote radio head (RRH) side, -Convex solver, -Weighted mean square error (WMSE) approach
[17] -2016	-Traffic admission control, -User association -Resource block allocation, -Power allocation, -Congestion control	- Cloud radio access networks (C-RAN) with C/U-split plane	-Lyapnunov optimization problem	-Energy efficient, -Delay minimization	-Apply queueing theory, -Stochastic geometric theory
[18] -2016	-BBU-RRH mapping, -Resource block allocation, -Power allocation	-OFDMA based Cloud radio access networks (C-RAN)	-0-1 Knapsack problem, -Mixed integer nonlinear problem	-Reduce overload in fronthaul link	-Heuristic algorithm
[19] -2016	-Resource block allocation, -Admission control	-OFDMA based Cloud radio access networks (C-RAN)	-Mixed integer nonlinear problem	-Maximize the tolerable interference level	-Low complexity algorithm, -Lagrange multiplier
[20] -2016	-Resource allocation,	-Heterogenous cloud radio access networks (H-CRAN)	-Convex and non-convex problem, -Lyapunov optimization	-Energy efficiency	-Stochastic model, -Weighted energy efficient utility function
[21] -2016	-Base station sleeping, -User association	-Ultra dense small cell (UDN) networks	-Non-linear problem	-Mean delay analysis -Steady-state user association probability	-Apply queueing theory, -Markov based model
[7] -2015	-Cell activation, - User association	-Cloud assited and C/U-split, CoMP supported HetNets	-Combinational optimization problem,	-Energy efficiency	- Base station partitioning approach

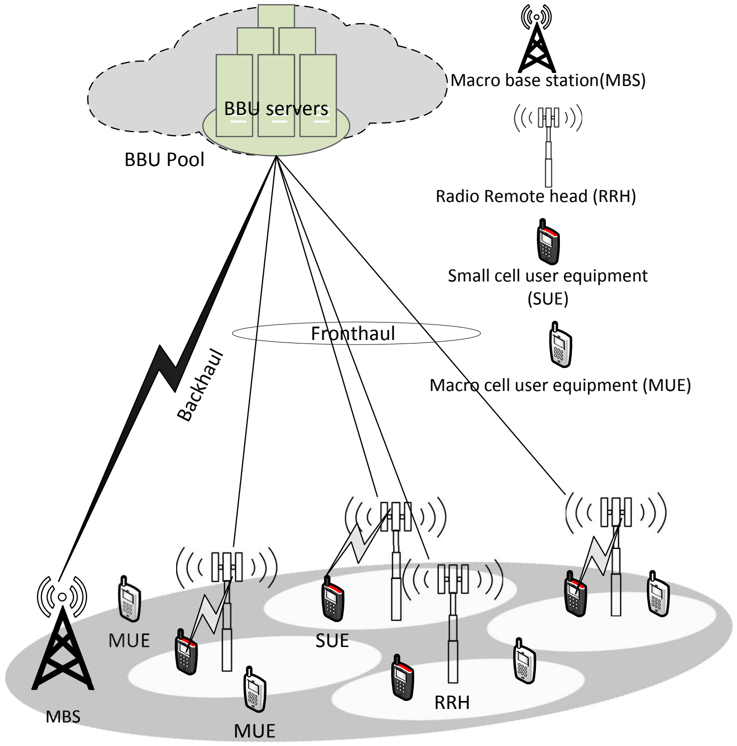


Figure 2.4: User or cell association scenario in C-RAN.

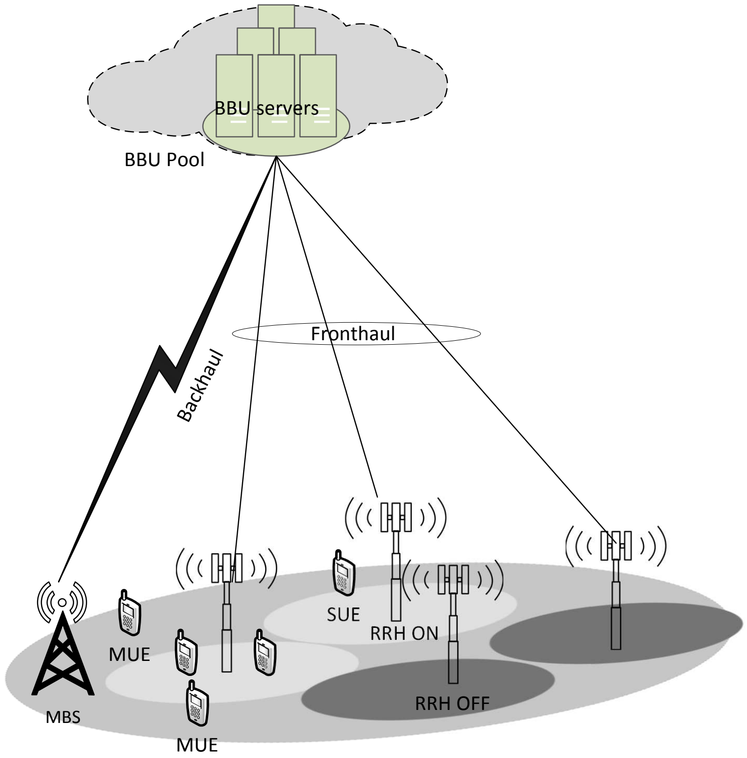


Figure 2.5: Cell activation/ON-OFF scenario in C-RAN.

-Within the C-RAN, which cells users will be associated with?

-Does the C-RAN support CoMP method? Is it one-to-one or one-to-multi-cell association?

-For multi-cell association method, which BS/RRH users will select for data transfer?

-What is the RRH/BS selection method for C-RAN?

-What are the optimization criteria and constraints considered in association process?

For user association and power allocation problem, we consider a two-tier C-RAN [6] with separated control/data planes as shown in Fig. 2.2, where S small cells are covered by a single macrocell B in underlaid manner. It is assumed that this architecture supports both C/U splitting [25] and CoMP [6] schemes. We aim at solving the multi-cell user association problem (in chapter 3) for maximum data rate provisioning under the constraints of interference and power consumption.

2.3.2 Cell On-off/Activation

Cell on-off or activation-deactivation mechanisms are applied to improve the energy efficiency of the networks. In heterogeneous networks or dense deployed C-RAN networks, it is essential to determine which cells need to be turned off to save the power. The authors in [7] [26] consider joint cell association, on-off and spectrum allocation process, whereas in [7] the authors proposed the cell deactivation method after the resource allocation and user association are done in heterogeneous networks. Fig. 2.5 shows the cell activation scenario in two-tier C-RAN networks.

2.3.3 Resources Allocation

Resource allocation is another challenging task in C-RANs. Generally, two types of resources are considered for allocation:

1) Computing resources (e.g., computing servers or BBU servers, memory, GPU, virtual machines, etc.) are considered as IT resources for computation of baseband processing signals inside the BBU pool.

2) Communication resources (e.g., bandwidth and power) are considered as wireless resources for transmission.

In chapter 4, we investigate computing resource allocation in terms of joint workload scheduling and baseband unit (BBU) allocation in C-RANs with the aim to minimize mean response time (RT) and aggregate power. Minimizing the mean response time in C-RAN with joint communication and computing resource allocation along with user association problem is investigated in chapter 5.

Deploying a large number of RRHs in C-RAN will improve network capacity and spectrum efficiency of 5G networks. However, the dense deployment of RRHs increases intra and inter-cell interference. Therefore, dynamic resource allocation is essential to control the interference level as well as to improve the energy efficiency (EE) and spectrum efficiency (SE) in C-RAN [27] [28]. Addressing the SE and EE issues, we consider an underlaid approach of communication resource allocation considering both macro and small cell users in two-tier OMA supported C-RANs in chapter 6. To handle more traffic and user loads, and higher network capacity for 5G networks, NOMA approach has been identified as a promising solution for future networks [8]. Considering NOMA in C-RAN may bring the advantages of SE, EE and massive connectivity through the centralized coordination in a BBU pool. However, the technical challenges of NOMA in the context of bandwidth and power allocation in C-RAN have not been investigated in detail in the literature. In chapter 7, we investigate in detail the energy efficient NOMA method for C-RANs in terms of codebook and power allocation.

2.4 Summary

In this chapter, we presented basic elements, some technological and architectural solutions such as CoMP supported and C/U split based cloud radio access networks. The overview of some research challenges and resource optimization problems that are considered in C-RANs have been discussed in this chapter. In the next chapter, we address the problem of the multi-cell user association along with power allocation in CoMP based C-RANs.

Chapter 3

User Association in Cloud-RAN

3.1 Introduction

As discussed in chapter 2, the cell or user association determines which cells or RRHs, the user will connect to for data transmission. User association method refers to RRH or base station selection method in radio access network. Most of the research works on the user association problem considered one-to-one allocation schemes, that is to say, one user can be associated with one base station. In this chapter, we investigate a probabilistic model for one-to-many allocation scheme in CoMP based C-RANs. The main contribution of this work is described as follows:

- An optimization problem for joint user association and power allocation is formulated with the objective of maximizing data rate provisioning for small cell users.
- We consider aggregated signal-to-noise ratio, minimum data rate requirement, and maximum power constraint for the joint user association and power allocation problem.
- To solve the problem, we propose a posterior probability based user association and power allocation method, where each user can associate with more than one SBS/RRH according to their posterior probability values.

3.2 Related Work

User association is usually formulated as a combinatorial optimization problem with high computational complexity; therefore, the commonly used algorithm is the one that associates users with the nearest RRH (or set of RRHs), thus maximizing the user data rate assuming that channel state information (CSI) is perfectly known [4]. Joint resource allocation, user association, and cell activation has been proposed in [26] using the stochastic measurements of traffic loads. These measurements can better track the slow time-scale traffic variations to model the actual small cell traffic especially in dense environments. In a similar context, joint user association and cell activation scheme was investigated in [29] considering small cells that operate at 60 GHz. The aim of using such high-frequency mmWave transmission was to achieve high data rates while keeping power consumption at low levels.

A multiple-BS user association scheme was proposed in [30, 31] to mitigate the interference and cell-edge bottlenecks in HetNets and ultra-dense networks. The problem which also involves power control and dynamic user-side interference cancellation, was formulated as a weighted-sum rate maximization and solved using the weighted minimum-mean-squared-error. Moreover, a greedy algorithm was applied in [32] to perform user association and RRH clustering. The work aimed at increasing the network sum-rate taking into account different precoding schemes such as zero-forcing and coordinated beamforming. The RRH cluster formation enables mobile users to receive data from several RRHs, and hence increases the system capacity especially for cell-edge users [33]. A dynamic RRH clustering was proposed in [34], where each RRH is allowed to join more than one cluster in order to alleviate the inter-cluster interference. Moreover, the authors in [35] proposed a recommendation-based user association scheme that enables users interact with each other and with the core network to pair with the base station that provides the highest quality of service (QoS). The association process utilized historical data analysis in the selection process. To this end, a satisfaction game was adopted, where users have to manage the tradeoff between profits and costs to get the optimum data rate.

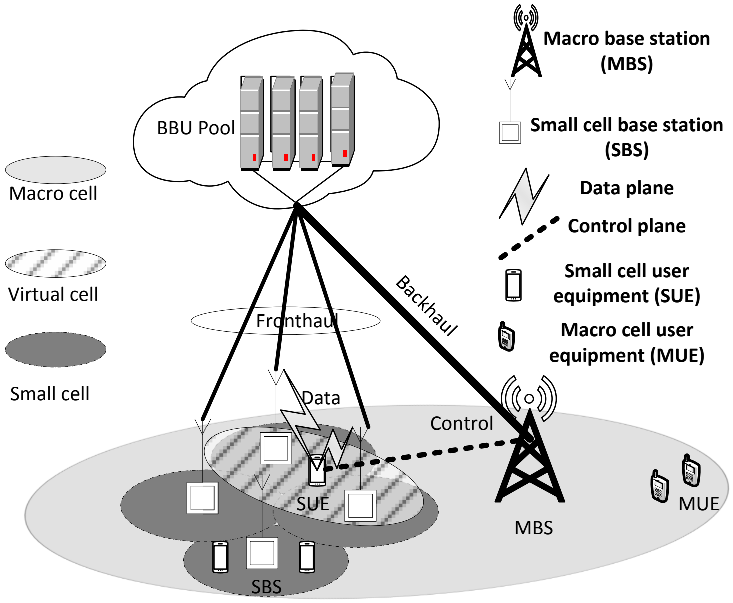


Figure 3.1: CoMP supported C-RAN with separated control/data planes.

3.3 System Model and Assumptions

For user association and power allocation problem, we consider a two-tier CoMP supported C-RAN with separated control/data planes as shown in Fig. 3.1, where S small cells are covered by a single macro cell B in underlaid manner. It is assumed that this architecture supports both C/U splitting [25] and CoMP [6] schemes. We aim at solving the multi-cell user association problem for maximum data rate provisioning under the constraints of interference and power consumption.

Each small cell in the set $\mathsf{S} = \{1, 2, ..., S\}$ is assumed to have M antennae, serving a total number of K^s SUEs. The total number of users is denoted by $K = K^s \bigcup K^m$, where K^m is the number of MUEs. Each SUE or MUE is equipped with one antenna. For downlink transmission, it is assumed that each SUE receives data from more than one SBS/RRH using the CoMP joint transmission [5] method. In addition, we assume that each SUE uses a different orthogonal channel for different data planes in order to mitigate the intra-tier interference. Let $\alpha_{i,j}$ denote the allocation matrix of SUE i and SBS/RRH j. That is,

$$\alpha_{i,j} = \begin{cases} 1, & \text{if SUE } i \text{ is associated with SBS } j, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

The channel gain from SBS j to SUE i is denoted by $h_{i,j} \in \mathcal{C}^{M\times 1}$, whereas $W_{i,j} \in \mathcal{C}^{M\times 1}$ denotes the pre-coding vector for SUE i to SBS j. The allocating power from SBS j to user i is denoted by $P_{i,j}$. Here, we can write the pre-coding vector for SUE i from all SBSs in terms of the allocating power and the allocation matrix, $W_i = \sum_{j=1}^{S} W_{i,j} = \sum_{j=1}^{S} \|\alpha_{i,j} P_{i,j}\|^2 \in \mathcal{C}^{SM\times 1}$. If SUE i is not served by SBS j, then the allocation vector sets the corresponding pre-coding vector to zero (i.e. when $\alpha_{i,j} = 0$, the pre-coding vector, $W_{i,j} = 0$).

Let x_i denote the transmitted data symbols for user i with $\mathrm{E}[|x_i|^2] = 1$. The received signal of SUE i can be written as:

$$\begin{aligned} Y_i = {} & \underbrace{\left(\sum_{j=1}^{S} h_{i,j}\alpha_{i,j}P_{i,j}\right)x_i}_{\text{desired signal}} + \underbrace{\sum_{k=1,k\neq i}^{K^s}\left(\sum_{j=1}^{S} h_{i,j}\alpha_{k,j}P_{k,j}\right)x_k}_{\text{co-tier interference signal}} \\ & + \underbrace{\sum_{m=1}^{K^m} h_{i,B}P_{m,B}x_m}_{\text{cross-tier interference signal}} + \eta_0, \\ = {} & \left(\sum_{j=1}^{S} h_{i,j}W_{i,j}\right)x_i + \sum_{k=1,k\neq i}^{K^s}\left(\sum_{j=1}^{S} h_{i,j}W_{k,j}\right)x_k \\ & + \sum_{m=1}^{K^m} h_{i,B}P_{m,B}x_m + \eta_0, \end{aligned} \tag{3.2}$$

where the first term on the right hand side is the desired signal for user i, the second term denotes the interference signal incurred by other active SUEs, and the third term indicates the interference signal from all active macro users transmission. The η_0 represents the additive white Gaussian noise (AWGN) with zero mean and unit variance. The signal to interference noise ratio

Table 3.1: List of symbols

Symbol	**Description**
S	number of SBS/RRH
K^s	number of SUE
K^m	number of MUE
K	total number of C-RAN users
$\alpha_{i,j}$	user association variable for user i to SBS j
$h_{i,j}$	channel gain of user i to SBS j
$P_{i,j}$	power allocation of user i to SBS j
r_i	data rate of user i
R	total data rate of all SUEs
γ_i	SINR of user i
C^{th}	minimum data rate of SUE
F^{max}	fronthaul maximum capacity
F_i	feasibility set of user i
$P(S_i)$	prior probability of SBS i
$P(U_i\|S_j)$	conditional probability of user i given SBS j
$P(S_j\|U_i))$	posterior probability of selecting SBS j given user i

(SINR) of SUE i can be expressed as:

$$\gamma_i = \frac{|\sum_{j=1}^{S} h_{i,j}\alpha_{i,j}P_{i,j}|^2}{\sum_{k=1,k\neq i}^{K^s} |\sum_{j=1}^{S} h_{i,j}\alpha_{k,j}P_{k,j}|^2 + \sum_{m=1}^{K^m} h_{i,B}P_{m,B} + \eta_0}, \tag{3.3}$$

where γ_i refers to the aggregated SINR of SUE i. According to the Shannon formula, the achievable data rate of SUE i will be $r_i = \triangle f \log_2(1+\gamma_i)$, where $\triangle f$ represents the channel bandwidth. The total data rate of all SUEs can be expressed as:

$$R = \sum_{i=1}^{K^s}\sum_{j=1}^{S} \alpha_{i,j} r_i. \tag{3.4}$$

All the symbols and description of the symbols is given in Table 3.1.

3.4 Problem Formulation

The objective of user association is to maximize the total data rate of all small cell users. To this point, each SUE can access and download data from more than one SBS/RRH while maintaining the interference at an acceptable level. The mathematical formulation of our user association problem can be described as follows:

$$\textbf{P1:} \max_{\alpha_{i,j}, P_{i,j}} R = \sum_{i=1}^{K^s} \sum_{j=1}^{S} \alpha_{i,j} r_i \tag{3.5}$$

Subject to:

$$\begin{aligned}
&\text{C1:} \quad \sum_{j=1}^{S} \alpha_{i,j} \geq 1, \quad \forall i \in K^s \\
&\text{C2:} \quad \sum_{i=1}^{K^s} \alpha_{i,j} P_{i,j} \leq P_j^{max}, \quad \forall j \in S, \\
&\text{C3:} \quad \gamma_i \geq \gamma^{min}, \quad \forall i \in K^s, \\
&\text{C4:} \quad \triangle f \log_2(1 + \gamma_i) \geq C^{th}, \quad \forall i \in K^s, \\
&\text{C5:} \quad \sum_{i=1}^{K^s} \alpha_{i,j} \leq F^{max}, \quad \forall j \in S, \\
&\qquad \alpha_{i,j} \in \{0, 1\}, P_{i,j} \in [P_j^{min}, P_j^{max}], \forall i \in K^s, j \in S.
\end{aligned}$$

The objective of this problem is to maximize the aggregated data rate of SUEs in the network. There are two effective parameters considered in this optimization problem: (1) the user association vector (i.e. $\alpha_{i,j} \in \{0, 1\}$) and (2) the power allocation vector (i.e. $P_{i,j} \in [P_j^{min}, P_j^{max}] \geq 0$). The constraint C1 enforces that each SUE is allowed to be connected with more than one SBSs. C2 is the power budget constraint for each small cell j which is P_j^{max}. C3 is the SINR constraint of each SUE. C4 is the minimum data rate constraint of the SUE. C5 describes the fronthaul capacity constraint, where F^{max} refers to the maximum limit of baseband signals transmitted on a fronthaul link.

This optimization problem is a mixed-integer non-linear non-convex problem and NP-hard to solve in general [36]. The optimal solution of **P1** is intractable due to the combinational

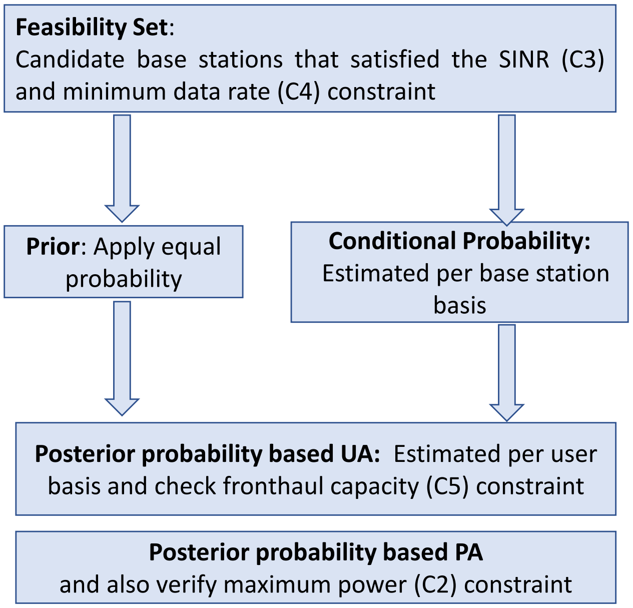

Feasibility Set:
Candidate base stations that satisfied the SINR (C3) and minimum data rate (C4) constraint
Prior: Apply equal probability
Conditional Probability: Estimated per base station basis
Posterior probability based UA: Estimated per user basis and check fronthaul capacity (C5) constraint
Posterior probability based PA and also verify maximum power (C2) constraint

Figure 3.2: Posterior probability based user association and power allocation (P2UPA) method.

nature of user association, the power allocation constraint C2, and the non-convexity nature of constraint C4.

3.5 Posterior Probability based User Association and Power Allocation (P2UPA)

We assume that SUEs are always connected to the MBS, which in turn, collects and analyses the channel state information (CSI) whenever a SUE requests to setup a data plane with small cells. Based on these CSI measurements, the SINR and the achievable data rate should be known for each of the data plane. Based on the aforementioned assumptions, in this work we consider posterior probability based user association and power allocation (P2UPA) method, as shown in Fig. 3.2, where each SUE can be associated with more than one SBS depending on the aggregate SINR and channel capacity information.

Assuming that the aggregate SINR (γ_i) and the achievable data rate (r_i) of the i SUE are known based on the CSI; then the P2UPA works as follows:

- Feasibility Set: The initial step determines which set of small cells becomes the feasible set for a user using the data plane connection. The possible SBS candidate (i.e., $F_i \in S$) for the i^{th} SUE, is obtained after applying the constraints C3 and C4.

- Prior and Conditional Probability: We estimate the prior probability from possible candidate SBSs set on a per-user basis. For example, the feasibility set of i^{th} SUE is $F_i = \{S_1, S_2, S_3, .., S_n\}$. The prior probability of i^{th} SUE to connect each SBS is $P(S_1) = P(S_2) = P(S_3) = .. = P(S_n) = \frac{1}{n}$, where n defines the maximum number of connections the SUE i can achieve without violating the constraints C3 and C4.

  Then, the conditional probability is estimated from the achievable data rates and feasibility set on a per-SBS basis. The conditional probability refers to the pobability of user i given SBS j is $P(U_i|S_j) = \frac{r_{ij}}{\sum_{i=1}^{K^s} \alpha_{i,j} r_{i,j}}$, For example, the j^{th} SBS supports two SUEs indexed by i and k. The conditional probability of users i and k given by j^{th} SBS is $P(U_i|S_j) = \frac{r_{ij}}{r_{ij}+r_{kj}}$ and $P(U_k|S_j) = \frac{r_{kj}}{r_{ij}+r_{kj}}$, respectively, where r_{ij} denotes the achievable data rate of user i from the j^{th} SBS.

- Posterior Probability: Posterior probability refers to the selection probability of SBS j given by SUE i. According to Bayes' rule, we can formulate the posterior probability of $P(S_j|U_i)$ as follows:

$$P(S_j|U_i) = \frac{P(U_i|S_j)P(S_j)}{p(x)}, \tag{3.6}$$

  where $p(x) = \sum_{n \in F_i} P(U_i|S_n)P(S_n)$.

- User Association and Power Allocation: Users can associate with small cells according to their posterior probability values. Each user selects a base station based on their maximum posterior probability (MAP) value. As long as the fronthaul capacity constraint (e.g. C5) is not satisfied, the users can associate to more than one base station based on the descending order of posterior value.

  To satisfy the maximum power constraint (e.g. C2) of each SBS, the total power is allocated to each user based on their corresponding posterior value. For example, if the user i

Table 3.2: Feasibility set and data rate of users

Users	S_1	S_2	S_3	S_4	**Feasibility Set**
U_1	20	12	5	0	$F_1 = \{S_1, S_2, S_3\}$
U_2	0	10	8	0	$F_2 = \{S_2, S_3\}$

Table 3.3: Prior and conditional probability

Users	**Prior**	$P(U_i\|S_1)$	$P(U_i\|S_2)$	$P(U_i\|S_3)$
U_1	$P(S_1) = P(S_2) = P(S_3) = \frac{1}{3}$	1	0.55	0.38
U_2	$P(S_2) = P(S_3) = \frac{1}{2}$	0	0.45	0.62

connects to base station j (e.g. if $\alpha_{i,j} = 1$), the allocated power can be expressed as follows:

$$P_{i,j} = P(S_j|U_i) \times P_j^{max}. \tag{3.7}$$

The posterior probability based user association (3.6) and power allocation schema (3.7) guarantees that the users who have the highest association probability receive maximum power for data transfer.

Example 1: To give an clear idea about the proposed user association method, let us consider a C-RAN with two users U_1 and U_2, and four small cells S_1, S_2, S_3 and S_4. Table I shows the feasibility set of U_1, U_2 and the achievable data rates after applying the constraints C3 and C4.

As mentioned before, the values of prior probability can be estimated per-user basis. According to the feasibility set of U_1, the prior probability of choosing a small cell is $P(S_1) = P(S_2) = P(S_3) = \frac{1}{3}$, whereas for U_2, the prior probability of choosing a small cell is $P(S_2) = P(S_3) = \frac{1}{2}$. Next, we estimate the conditional probability of U_1 and U_2 when the small cell is S_1, $P(U_1|S_1) = \frac{r_{11}}{r_{11}+r_{21}} = \frac{20}{20+0} = 1$, and $P(U_2|S_1) = 0$, respectively. This implies that $P(U_1|S_1) + P(U_2|S_1) = 1$. Similarly, we can get the conditional probability of U_1 and U_2 when the small cell is S_2.

Table 3.4: Posterior probability

Users	$P(S_1\|U_i))$	$P(S_2\|U_i)$	$P(S_3\|U_i)$	**Total probability**
U_1	0.52	0.28	0.20	$\sum_{S_j \in F_1} P(S_j\|U_1) = 1$
U_2	0	0.44	0.56	$\sum_{S_j \in F_2} P(S_j\|U_2) = 1$

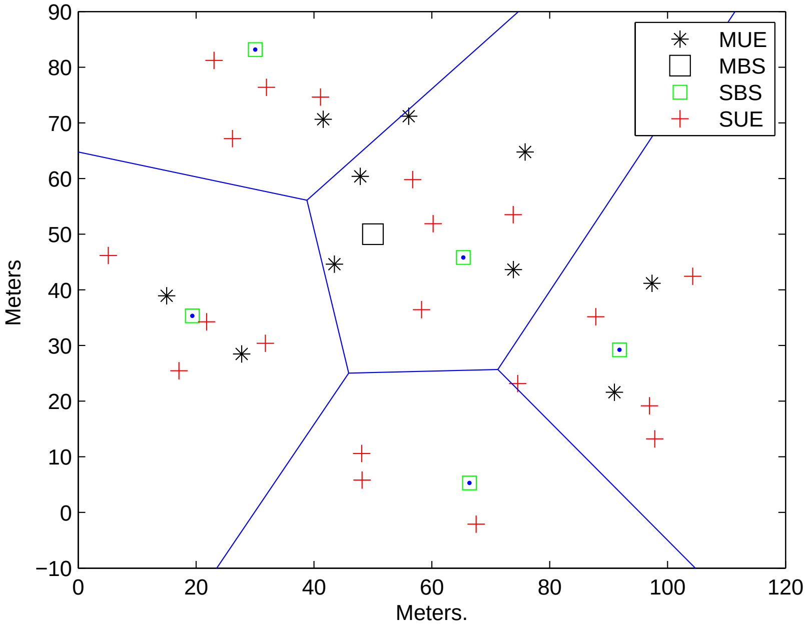


Figure 3.3: Simulation model consisting of one macro and five small cells.

By applying Bayes' rule, we can calculate the posterior probability of each user and small cell. According to Table 3.3, the small cells S_1 and S_3 are the highest priority for the users U_1 and U_2, respectively. As long as the constraint C5 is satisfied, the next possible small cell for U_1 is $\{S_2,S_3\}$ and U_2 is S_2.

3.6 Simulation Results

In this section, the performance of P2UPA is investigated and evaluated. In the simulation model, as shown in Fig. 3.3, we consider 120×100 meter area, wherein one macro base station is underlaid by five small cell base stations. The locations of SBSs are modeled using spatial Poisson point process (PPP) Φ_s with intensity λ_s, and SUEs are distributed using another independent PPP Φ_u with intensity λ_u. The settings for the simulation parameters are shown in Table 3.5.

To evaluate the performance of the proposed scheme, we consider the following UA methods: i) **Distance or location aware UA:** In the distance/location aware UA method, user $i \in \Phi_u$ selects a base station $j \in \Phi_s$ if and only if $\parallel i - j \parallel < \parallel i - k \parallel$, $\forall k \in \Phi_s$.

Table 3.5: Simulation parameters

Parameters	**Values**
Total no. of small cells	5
Total no. SUEs	$4-20$
Minimum data rate requirements	50-120 kbps
Number of MUEs	$10-20$
Transmission power of MBS	43 dBm
Transmission power of SBS	30 dBm
Path-loss exponent (λ)	4
Noise power spectrum density	-144 dBm/Hz

ii) **SINR aware UA method:** In the maximum SINR based user association method, the user i connects to a base station j when the average received SINR is maximum. In SINR based UA method, we assume the transmission power is P_t, and the noise power is σ^2. In the SINR approach, user i selects SBS j when $\underset{\forall j}{\operatorname{argmax}}\ \gamma_i = \underset{\forall j}{\operatorname{argmax}}\left[\frac{P_t * d_{ij}^{-\lambda}}{\sigma^2 + I_0}\right]$, that is:

$$\alpha_{i,j} = \begin{cases} 1, & \text{if } \underset{\forall j}{\operatorname{argmax}}\ \gamma_i = \underset{\forall j}{\operatorname{argmax}}\left[\frac{P_t * d_{ij}^{-\lambda}}{\sigma^2 + I_0}\right] \\ 0, & \text{otherwise,} \end{cases} \tag{3.8}$$

where λ and d_{ij} indicate the pathloss exponent and the distance between user i and base station j, respectively. I_0 denotes the maximum received interference power supported by C-RAN system.

iii) **MAP based UA:** In the this user association method, user i selects the SBS j depending on the maximum posterior value, that is:

$$\alpha_{i,j} = \begin{cases} 1, & \text{if } \underset{\forall j}{\operatorname{argmax}}\ P(S_j|U_i) \\ 0, & \text{otherwise,} \end{cases} \tag{3.9}$$

Fig. 3.4 shows the performance of sum data rate versus the number of small cell users for different user association method. It can be observed that the sum data rate performance of P2UPA method gives better results than the MAP, distance and SINR based single cell methods.

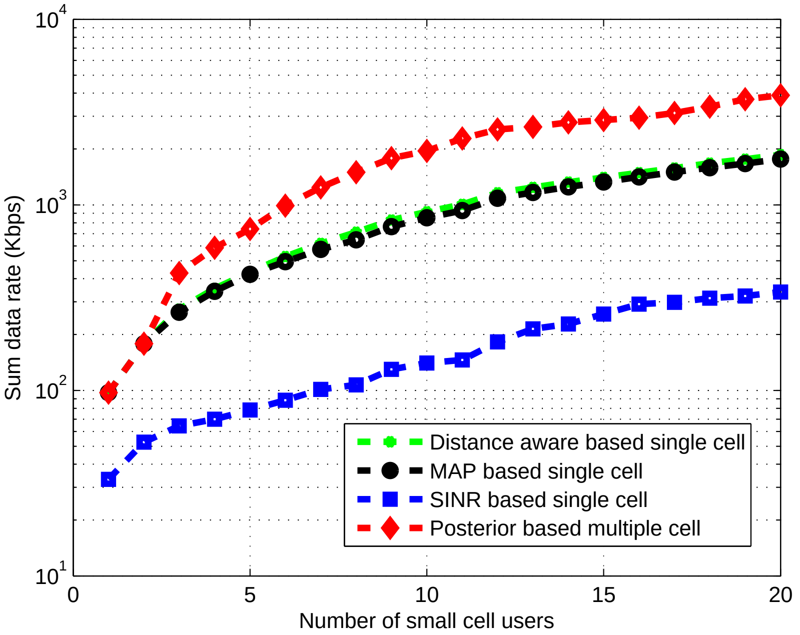


Figure 3.4: Sum data rate comparison using posterior probability based multi-cell, maximum aposterior probability (MAP) based single cell, distance, and SINR based single cell user association.

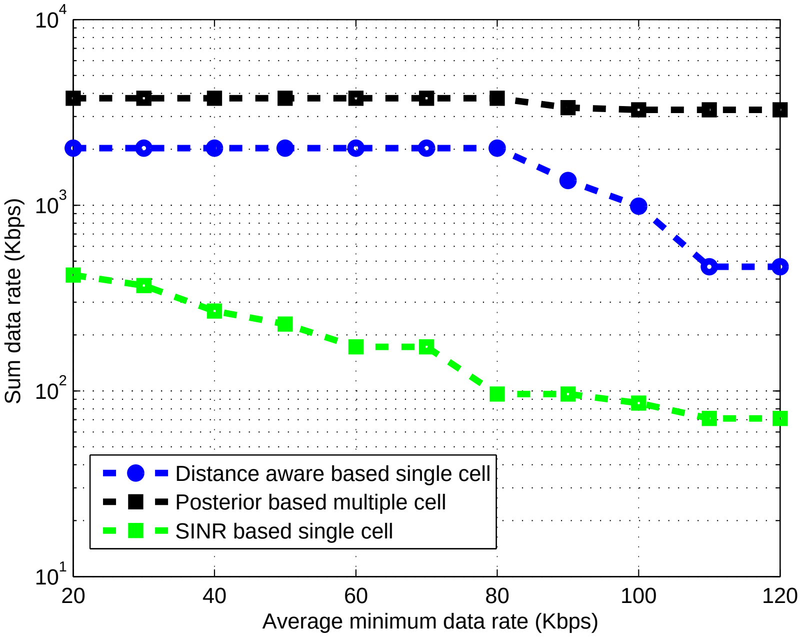


Figure 3.5: Sum data rate performance of distance-aware, SINR-based single cell and posterior probability based multi-cell user association method with different minimum data rate requirement.

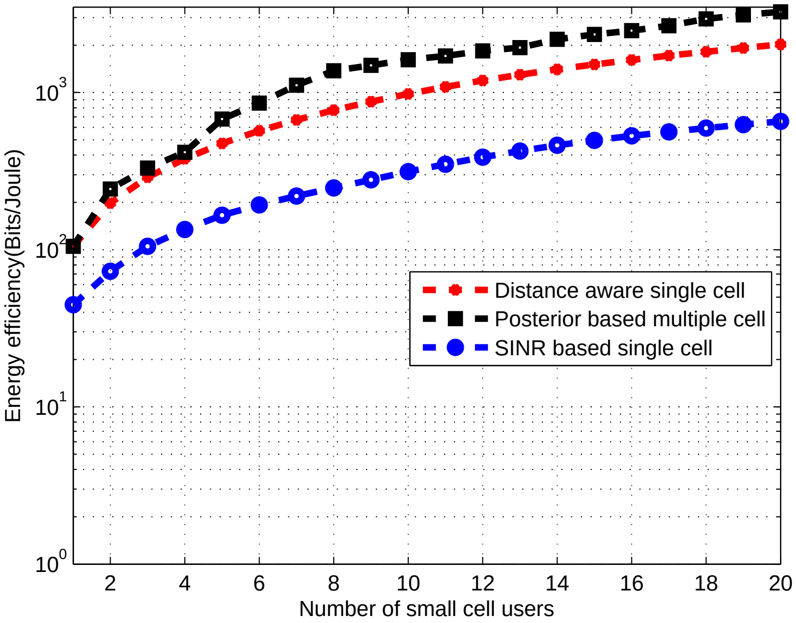


Figure 3.6: Energy efficiency vs. number of small cell users

The distance-aware and MAP based method shows equal performance. This is due to the fact that in the MAP based user association, the user i connects to base station j based on the maximum received CSI. Considering equal transmission power and noise factor, according to (3.8), the SINR of user becomes maximum when the distance between user and base station becomes minimum.

Fig. 3.5 shows that the minimum data rate requirement is an important factor for choosing the user association method. When the data rate requirement is increased, the distance-aware user association method will support fewer number of users compared to the posterior-based method. In the posterior method, users can associate with more than one base station to satisfy the minimum data rate constraint, whereas in the distance and SINR-based methods, users can choose one base station based on their relative distance or average received signal strength.

Fig. 3.6 depicts the comparison of energy efficiency performance. Energy efficiency in C-RAN represents the ratio of the total throughput to the total energy consumption. The energy efficiency performance of posterior-based user association shows the best result compared to other methods, because the total power distribute among users based on their percentage of posterior probability value whereas distance and SINR method total power are allocated equally among users.

3.7 Summary

In this chapter, we proposed a probabilistic model for user association and power allocation with prior knowledge of channel state information (CSI). The proposed P2UPA method satisfies the minimum data rate and aggregated interference constraints as well as the maximum power constraint. In addition, the P2UPA is an appropriate method for CoMP based C-RAN networks where users can associate with one or more base stations. The results show that the proposed user association scheme surpasses the distance-aware and SINR-based methods which associate users with the nearest base station. Moreover, the P2UPA method associates users to the nearest base station based on maximum aposterior value as well as finds the others candidate base stations posterior probability values taking into the account of minimum data rate, SINR, maximum power and fronthaul capacity constraints. In the next chapter, we focus on computing resource allocation and workload scheduling into BBU servers.

Chapter 4

Joint Workload Scheduling and BBU Allocation in Cloud-RAN

4.1 Introduction

In this chapter, we consider the computing resource allocation in C-RAN. The major difference between C-RAN over conventional RAN is to pool the baseband units from multiple base stations to a centralized BBU pool. In C-RAN, the pool of BBUs works servers processing requests from users in a centralized, cooperative, and coordinated way. Furthermore, centralized and cloud-based BBU pools can share resources, mitigate interference, manage mobility among users. The major benefits of C-RAN over conventional RAN are that C-RAN architecture reduces capital expenditure and operating expenditure cost as well as supports scalability and flexibility of further deployment of small cells in terms of deployment of remote radio heads [3]. In the last chapter, we addressed the user association problem in C-RAN. On the other hand, resource allocation is another challenging task in C-RAN. Generally, two types of resources are considered for allocation: 1) IT resources (e.g., computing servers or BBU, memory, GPU etc.) for computation and 2) wireless resources (e.g., bandwidth, power) for transmission.

The authors in [37] proposed a location-specific energy-efficient resource allocation in C-RAN, where computation tasks are executed in the devices or the cloud depending on the overall power budget. The framework for processing base stations in a data center is proposed in [38]. Similarly,

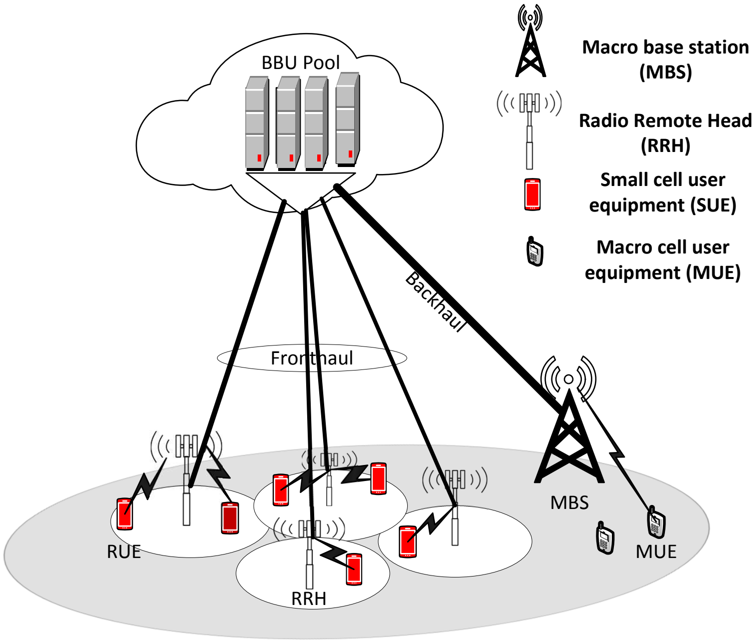


Figure 4.1: Small cell-based C-RAN architecture for 5G networks.

the authors in [39] considered bin packing scheme to match processing requirements to computing servers with the aim to minimize the number of servers or bin to be used. In [20], the authors studied energy-efficient resource allocation in queue-aware multimedia heterogeneous C-RAN, where C-RAN maintained a queue for each RRH user. However, the round trip time (RTT) or response time (RT) requirement has not been studied in the resource allocation literature. The RTT or RT is an important QoS parameter for delay sensitive applications.

In this work, we investigate a joint workload scheduling and baseband unit (BBU) allocation in C-RANs. The main contributions of this chapter are:

- We establish a queuing model in C-RAN for 5G networks, and then formulate an optimization problem for joint workload scheduling and BBU allocation with an objective to minimize mean response time and aggregate power. The C-RAN controller/scheduler distributes the workload among BBUs with optimized scheduling weights.
- We consider the queueing stability and workload conservation constraints for the joint workload scheduling and BBU allocation. Each BBU server works as queueing system with predefined service rate.

- To solve this problem, we propose an energy efficient joint workload scheduling and BBU allocation (EE-JWSBA) algorithm using the established queueing model.
- We conduct simulations in order to validate the performance of EE-JWSBA algorithm while considering three scheduling weights (e.g., random, normalized, and upper limit).

All the symbols and description of the symbols used in this chapter is given in Table 4.1.

4.2 System Model

A small cell-based C-RAN architecture for 5G networks is considered which consists of one macro base station (MBS) and N small cells/ RRHs, where the MBS covers N number of RRHs in an underlay manner. The RRHs and MBS are connected to a BBU pool using fronthaul and backhaul links, respectively. In a small cell-based C-RAN, access requests of users are transmitted from RRHs to BBU pool for baseband processing.

Figure 4.2 represents a queueing model for BBU pool where C-RAN controller or scheduler (S) distributes the incoming requests to BBU servers for computation. At a given time slot t, the scheduler receives requests from N RRHs with Poisson arrival rate $\lambda_1^{(t)}, \lambda_2^{(t)}, ..., \lambda_N^{(t)}$. The total incoming requests (i.e., $\overline{\lambda}^{(t)} = \sum_{i=1}^{N} \lambda_i^{(t)}$) at time slot t are distributed to the corresponding BBU with scheduling weight $w_i^{(t)}$, where $i = \{1, 2, ..., \Psi\}$ denotes the BBU index and Ψ denotes the maximum number of BBU servers in pool. The arrival of scheduled requests to each BBU at time slot t also follows Poisson process with an weighted average of $w_i^{(t)}\overline{\lambda}^{(t)}$ and the service rate of i^{th} BBU is $\mu_i^{(t)}$ [40, 41]. The service process of each BBU follows an M/M/1 queuing model [42].

Similar to [41], [42], the response time of i^{th} BBU can be formulated as:

$$T_i = \frac{1}{\mu_i^{(t)} - w_i^{(t)}\overline{\lambda}^{(t)}}. \tag{4.1}$$

It is assumed that the BBU servers operate in two different power states, i.e., idle state and busy state as shown in Fig. 4.3. p_i and p_b denote power consumption in idle and busy state, respectively. It is assumed that the instantaneous transition from busy to idle or idle to busy

Table 4.1: List of symbols

Symbol	**Description**
N	number of SBS/RRH
S	scheduler or cloud controller
λ_N	arrival rate of N^{th} RRH
Ψ	number of BBU servers
w_i	scheduling weight of i^{th} BBU
μ_i	service rate of i^{th} BBU
T_i	response time of i^{th} BBU
p_i	power consumption in idle state
p_b	power consumption in busy sate
P_i	total power consumption of i^{th} BBU
ρ_i	i^{th} BBU server utilization ratio
θ_i	i^{th} BBU allocation variable
η	threshold time

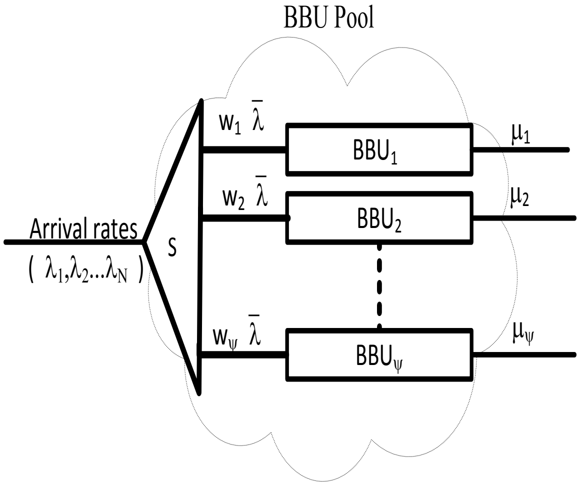


Figure 4.2: Queueing model for BBU pool in C-RAN.

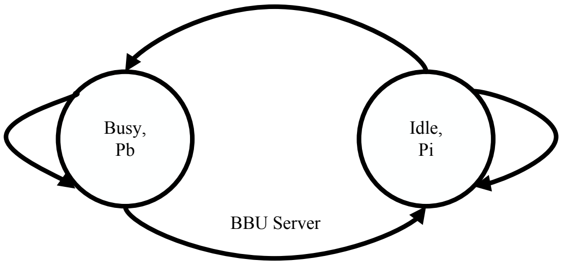


Figure 4.3: The power state transition model of a BBU server.

does not consume any power. At time instance t, the total power consumption of i^{th} BBU server is the sum of busy and idle state power weighted by the probability of busy and idle state. The total power consumption of i^{th} BBU server can be written as:

$$P_i^{(t)} = p_b \times P_i^{busy} + p_i \times P_i^{idle}, \tag{4.2}$$

where $P_i^{idle} = 1 - \rho_i^{(t)}$ and $P_i^{busy} = \rho_i^{(t)}$ are the steady-state probabilities to estimate idle state and busy state, respectively ($\rho_i^{(t)}$ represents server utilization ratio). Similar model is used in [43] for web server where server utilization ratio estimated as $\rho_i^{(t)} = \frac{\text{arrival rate}}{\text{service rate}}$. According to Fig. 4.2, the server utilization of i^{th} BBU can be estimated as, $\rho_i^{(t)} = \frac{w_i^{(t)}\overline{\lambda}^{(t)}}{\mu_i^{(t)}}$.

4.3 Problem Formulation

In order to maintain QoS requirement in terms of response time and power consumption in BBU pool, C-RAN should allocate, coordinate, and distribute workload among BBU servers. In this section, we formulate a joint workload distribution and BBU allocation optimization problem with an objective to minimize response time and power consumption.

4.3.1 Response Time

The response time in the C-RAN is defined as the duration from the time when users request arrive at the C-RAN controller to the time when the baseband processing results completely depart from the BBU server. In this work, the response time is taken as the QoS factor to measure the performance of C-RAN. To minimize the response time of C-RAN, the BBU allocation and queue stability should be considered. Let $\theta_i^{(t)}$ denote the BBU allocation variable at time instance

t given as,

$$\theta_i^{(t)} = \begin{cases} 1, & \text{if BBU } i \text{ is allocated at time instance } t \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

The service rate at each BBU will be $\theta_i^{(t)}\mu_i^{(t)}$. Thus, the response time of i^{th} BBU in (4.1) can be re-written as:

$$T_i = \frac{1}{\theta_i^{(t)}\mu_i^{(t)} - w_i^{(t)}\overline{\lambda}^{(t)}}. \tag{4.4}$$

Our objective is to minimize the mean response time by optimization of workload scheduling. The optimization problem of workload scheduling for C-RAN can be formulated as:

$$\begin{aligned} &\underset{w_i^{(t)},\theta_i^{(t)}}{\text{minimize}} : \frac{1}{\Psi}\sum_{i=1}^{\Psi} T_i, \\ &\text{Subject to : C1: } \sum_{i=1}^{\Psi} \theta_i^{(t)} \geq 1, \\ &\qquad \text{C2: } \sum_{i=1}^{\Psi} w_i^{(t)} = 1, \quad \forall w_i \geq 0, \\ &\qquad \text{C3: } w_i^{(t)}\overline{\lambda}^{(t)} \leq \theta_i^{(t)}\mu_i^{(t)}, \\ &\qquad \theta_i = \{1, 0\}, \quad \forall i \in \{1, 2, ..., \Psi\}. \end{aligned} \tag{4.5}$$

where the constraint C1 refers that at time instance t, the service requests from N RRHs are scheduled into more than one BBU. Constraint C2 refers to workload conservation constraint. At time instance t, the scheduler distributes all workload among BBUs. The constraint C3 maintains the queueing stability at each BBU, i.e., the arrival rate should not be greater than the service rate.

4.3.2 Energy Efficiency

Energy efficiency is an essential QoS parameter of C-RAN in order to reduce capital expenditure and operating expenditure cost. To optimize the power consumption in BBU pool of C-RAN, the BBU allocation, queue stability, and response time should be considered. The power consumption of i^{th} BBU can be formulated as:

$$\begin{aligned}
P_i^{(t)} &= \theta_i^{(t)} p_b P_i^{busy} + (1-\theta_i^{(t)}) p_i P_i^{idle} \\
&= \theta_i^{(t)} p_b \rho_i^{(t)} + (1-\theta_i^{(t)}) p_i (1-\rho_i^{(t)}) \\
&= \theta_i^{(t)} p_b \left(\frac{w_i^{(t)} \overline{\lambda}^{(t)}}{\mu_i^{(t)}} \right) \\
&+ (1-\theta_i^{(t)}) p_i \left(1 - \frac{w_i^{(t)} \overline{\lambda}^{(t)}}{\mu_i^{(t)}} \right).
\end{aligned} \tag{4.6}$$

The optimization problem of energy efficient joint workload scheduling and BBU allocation can be formulated as:

$$\begin{aligned}
&\underset{w_i^{(t)} \theta_i^{(t)}}{\text{Minimize}} : P^{(t)} = \sum_{i=1}^{\Psi} P_i^{(t)}, \\
&\text{Subject to} : \theta_i = \{1, 0\}, \quad \forall i \in \{1, 2, ..., \Psi\}, \\
&\quad \text{C1: } \sum_{i=1}^{\Psi} \theta_i^{(t)} \geq 1, \\
&\quad \text{C2: } \sum_{i=1}^{\Psi} w_i^{(t)} = 1, \quad \forall w_i \geq 0, \\
&\quad \text{C3: } w_i^{(t)} \overline{\lambda}^{(t)} \leq \theta_i^{(t)} \mu_i^{(t)}, \\
&\quad \text{C4: } \frac{1}{\Psi} \sum_{i=1}^{\Psi} T_i \leq \eta.
\end{aligned} \tag{4.7}$$

where the objective of the workload scheduling problem is to minimize the total power consumption in C-RAN. Constraint C1 refers that the service requests from N RRHs are scheduled into more than one BBU. Constraint C2 refers to workload distribution constraint. Constraint C3 refers to the queue stability constraint and C4 refers to the response time. At time instance t, the baseband processing in cloud-RAN should be completed before threshold time η.

4.4 Energy Efficient Joint Workload Scheduling and BBU Allocation(EE-JWSBA) Algorithm

The optimization problem (4.5) and (4.7) are mixed integer non-convex problem. These problems are generally difficult to solve directly due to the fractinal objective function in (4.5) and non-linear constraint C4 in (4.7). We consider response time optimization problem in (4.5) as a constraint to energy minimization problem (4.7). To solve the optimization problem in (4.7), we use heuristic approach and propose an energy efficient joint workload scheduling and BBU allocation (EE-JWSBA) algorithm for C-RAN. The EE-JWSBA works on the central controller (S) of C-RAN which generates the output of BBU allocation metric (θ) and scheduling weights (w) with the input that consists of number of BBUs, service rate (μ) and response time threshold (η). We also consider queue backlog for each BBU server, denoted as $Q_i^{(t)}$ where $i = 1, 2..., \Psi$. Let $Q(t) = \sum_i^{\Psi} Q_i^{(t)}$ denotes the queue backlog of the BBU pool in the time slot t. It represents how many requests are waiting in the BBU pool at time slot t. In the next time slot $(t + 1)$, the queue backlog in the BBU pool is updated as:

$$Q(t+1) = Q(t) + \sum_{i=1}^{\Psi} w_i^{(t)} \overline{\lambda}^{(t)} - \sum_{i=1}^{\Psi} \mu_i. \tag{4.8}$$

After initialing all the inputs, the EE-JWSBA described in **Algorithm 1**, works as follows:

- Line 1-2: BBU servers are sorted by multiplying of queueing backlog and service rate information, and then start from the highest value.
- Line 3-7: Estimate the scheduling weight and server utilization rate according to the following derivation:

  At each the time slot t, the arrival and service rates of i^{th} BBU are $w_i^{(t)} \overline{\lambda}_i^{(t)}$ and $\mu_i^{(t)}$, respectively. According to eq.(4.8), the queueing backlog of i^{th} BBU at the time slot $(t + 1)$ can be formulated as :

  $$Q_i^{(t+1)} = Q_i^{(t)} + w_i^{(t)} \overline{\lambda}^{(t)} - \mu_i^{(t)}.$$

- To avoid overload condition and maintain queueing stability in each BBU, the queueing backlog should be non-negative at each time instance, i.e., $Q_i^{(t+1)} \geq 0$. Therefore, the optimal scheduling weight (w_i) at each BBU can be derived as follows:

$$\begin{aligned}
&Q_i^{(t)} + w_i^{(t)}\overline{\lambda}^{(t)} - \mu_i^{(t)} \geq 0 \\
&w_i^{(t)}\overline{\lambda}^{(t)} \leq \mu_i^{(t)} - Q_i^{(t)} \\
&w_i^{(t)} \leq \frac{\mu_i^{(t)} - Q_i^{(t)}}{\overline{\lambda}^{(t)}} \\
&0 \leq w_i^{(t)} \leq \frac{\mu_i^{(t)} - Q_i^{(t)}}{\overline{\lambda}^{(t)}}.
\end{aligned} \tag{4.9}$$

The boundary of scheduling weight for each BBU is represented by (7.15).

- Line 8: Verify the queueing stability constraints and server utilization ratio.
- Line 9: Verify workload conservation constraint.
- Line 10-11: Verify the response time constraints. If all the constraints are satisfied, the BBU server is allocated and workload is distributed to that BBU.
- Line-12-13: Estimate the power utilization and update the queuing backlog. This process will work iteratively on each BBU server.

4.5 Performance Evaluation

The performance of EE-JWSBA algorithm under queue stability and scheduling weight constraints is evaluated with five BBU servers. Initially, queue backlog $Q^{(t=1)} = \{60, 30, 20, 40, 10\}$, and the service rates of BBU servers $\mu = \{25, 97, 185, 120, 50\}$ requests/time slot are considered. Further, we consider three types of scheduling weights; upper bound of scheduling weight, $w_i^{upper} = \frac{\mu_i^{(t)} - Q_i^{(t)}}{\overline{\lambda}^{(t)}}$, normalized scheduling weight, $w_i^{norm} = \frac{w_i^{upper}}{\sum w_i^{upper}}$, and randomized scheduling weights, $0 \leq w_i^{rand} \leq \frac{\mu_i^{(t)} - Q_i^{(t)}}{\overline{\lambda}^{(t)}}$. For simulations, we consider these scheduling weights to verify the performance of EE-JWSBA algorithm under given constraints.

Algorithm 1: EE-JWSBA: Energy Efficient Joint Workload Scheduling and BBU Allocation Algorithm

```
Input: No. of BBU, queue backlog $\{Q\}$ , Total no. of time slots $T$, service rate $\{\mu\}$, and response time
        threshold $\{\eta\}$.
Output: BBU allocation and scheduling weights
Sort BBU with $q_i$ in decrease order, where $q_i = Q_i\mu_i$ and $i = 1$ to No. BBU
Select largest $q_i$.
for $t \leftarrow 1$ to $T$ do
    Arrival requests at time slots $t$ is $\lambda^{(t)}$.
    for $i \leftarrow 1$ to No. BBU do
        Calculate upper bound of weight, $w_i^{upper} = w_i^{(t)} = \frac{\mu_i^{(t)} - Q_i^{(t)}}{\overline{\lambda}^{(t)}}$.
        Calculate server utilization, $\rho_i^{(t)} = \frac{w_i^{(t)}\overline{\lambda}^{(t)}}{\mu_i^{(t)}}$.
        if $\mu_i^{(t)} - (Q_i^{(t)} + w_i^{(t)}\lambda^{(t)})$ and $\rho_i^{(t)} < 1$ then
            if $\sum w_i^{(t)} \leq 1$ and $w_i^{(t)} \geq 0$ then
                Schedule user request $\lambda^{(t)}$ as long as the constraints C2,C3 and C4 are met.
                Allocate BBU $i$ at time slot $t$, $\theta_i^{(t)} \leftarrow 1$ Calculate power utilization in $i^{th}$ BBU $P_i^{(t)}$.
                Update the queue backlog according to eq.(4.8).
                $i \leftarrow i + 1$
            else
                $i \leftarrow i + 1$
return Weight $\boldsymbol{w}$,and BBU allocation $\boldsymbol{\theta}$.
```

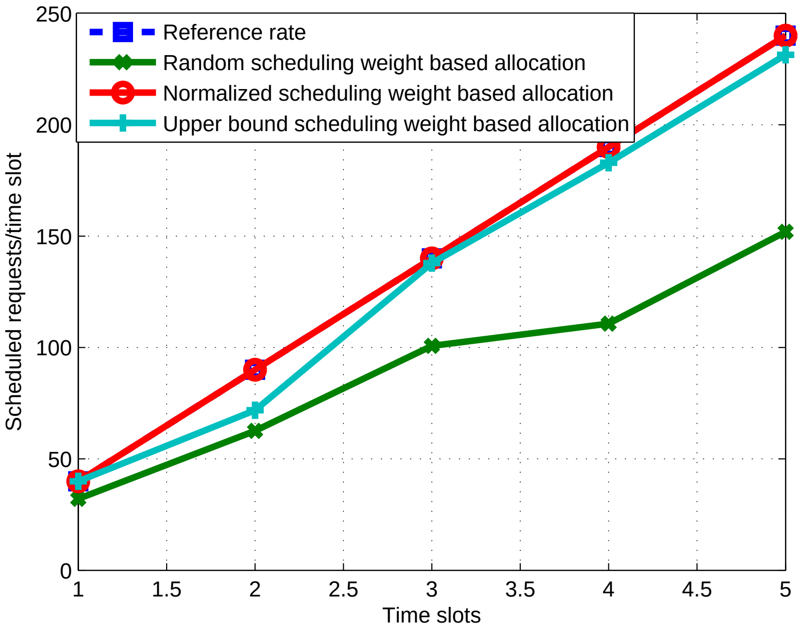


Figure 4.4: Scheduling workload based on BBU allocation using random, normalized and upper scheduling weights.

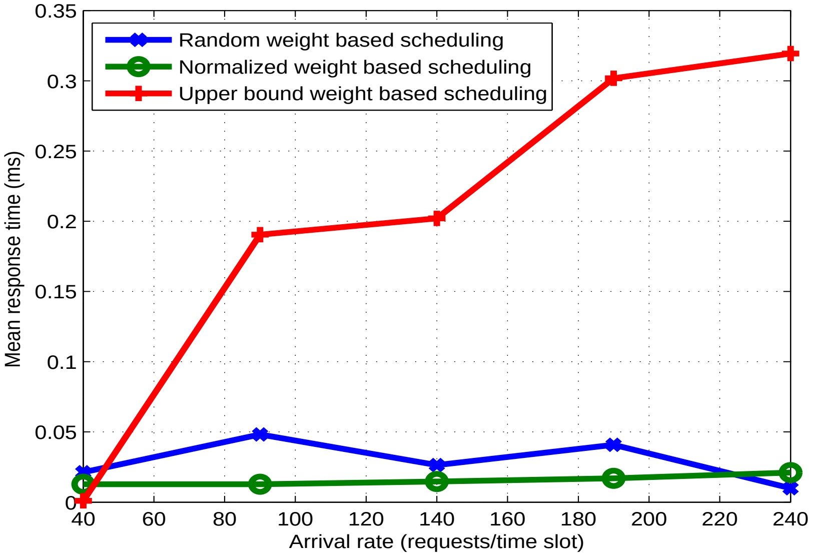
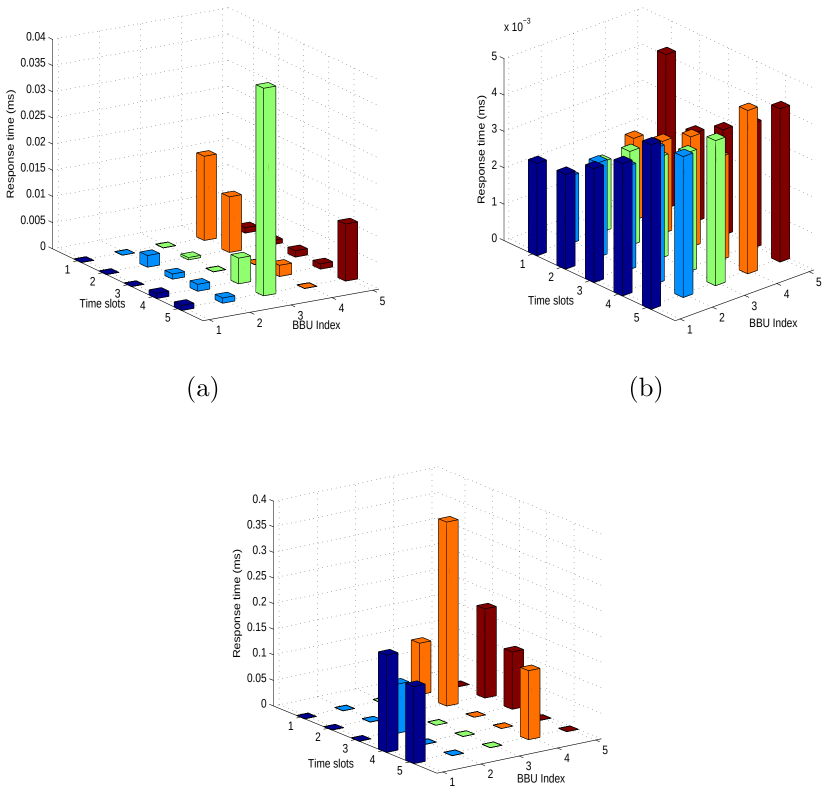


Figure 4.5: Workload scheduling among BBU based on three different scheduling weights.

Fig. 4.4 shows the scheduled requests per time slot in BBU pool by considering three scheduling weights in allocation procedure. The arrival rate per time slot depicts as a reference rate. As shown in Fig. 4.4, the normalized weight-based allocation method schedules all workload among BBU pools, whereas upper limit scheduling weight based allocation method shows some queueing backlog. For example, in time slot 2, upper limit based allocation method schedules around 70 requests out of 90, in next time slot, it schedules all arrival requests as well as previous backlog requests, however, in 4^{th} and 5^{th} time slots, it shows some backlog again. On the other hand, the best case performance of random scheduling weight based allocation method is same as upper limit based allocation because weights are chosen in the range of 0 to w^{upper}. Fig. 4.5 shows the response time performance of proposed EE-JWSBA algorithm. Among these three scheduling weights, the normalized weight based scheduling takes the lowest response time whereas upper limit based workload scheduling takes highest response time to process the requests in a BBU pool. Due to the randomness of scheduling weight, the response time of random weight based scheduling method shows fluctuating trend over arrival rates. Fig. 4.6 shows the allocated BBU index number and their corresponding response time in different time slots. Fig. 4.6 (a) shows the random weight based scheduling method which allocates three BBUs per time slot where as the normalized weight based scheduling method considers all BBUs for allocation, as depicted



(a)

(b)

(c)

Figure 4.6: BBU allocation based on (a) random scheduling weight, (b) normalized scheduling weight, and (c) upper bound of scheduling weight.

in Fig. 4.6(b). In terms of minimizing the number of BBU allocation, the upper limit based scheduling method (e.g., Fig. 4.6(c)) shows the best performance than others. This allocation method, on average, allocates one BBU per time slot whereas random weight based scheduling method allocates three BBUs per time slot.

Fig. 4.7 shows the comparison of BBU allocation on upper bound and normalized weight based method for per time slot basis. During the five time slots, the upper bound based scheduling method allocates the total number of 8 out of 25 BBU servers whereas normalized weight based method distributes all workload among all BBUs for every time slot. In terms of minimizing the mean response time, this scheduling method is the best choice. However, this scheduling method does not work for minimizing the number of BBU allocation.

Fig. 4.8 shows the utilization percentage of allocated BBU servers for upper bound and

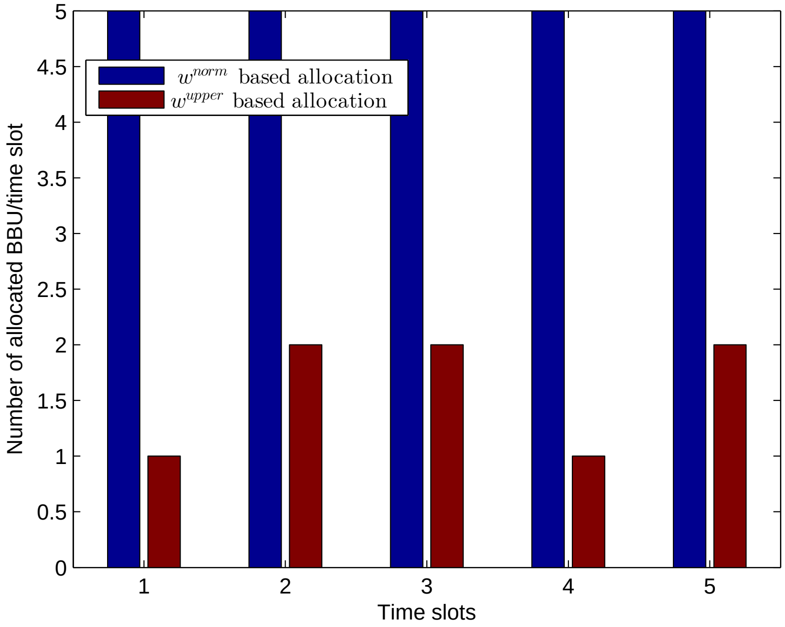


Figure 4.7: Comparison of BBU allocation on upper bound and normalized weight based method.

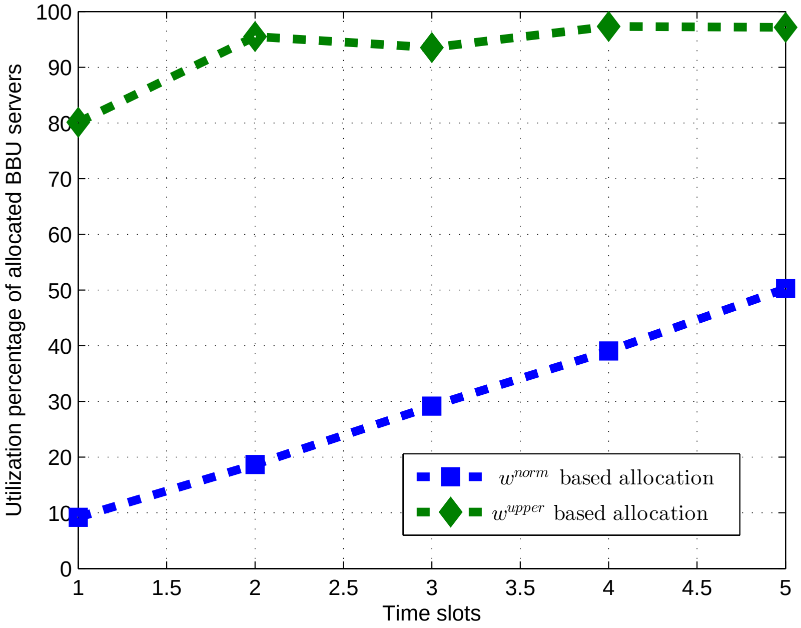


Figure 4.8: Percentage of BBU servers utilization.

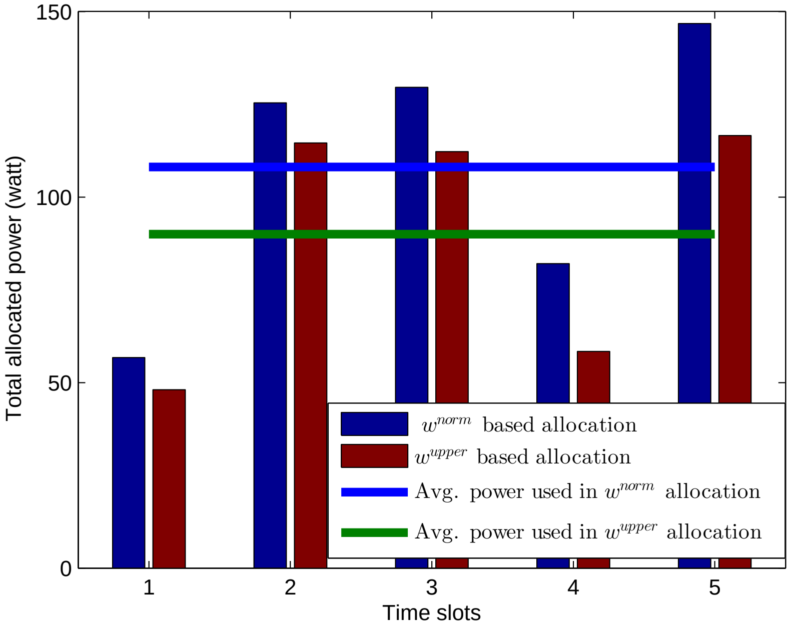


Figure 4.9: Aggregated power consumption in allocated BBU servers over time slots.

normalized weight based scheduling method. During the five time slots, the upper bound method utilizes 80% to around 98% of the allocated BBU servers. On the other hand, for each time slot, the normalized method considers all BBU servers but servers utilization increases from 10% to 50% during the five time slots.

Fig. 4.9 shows the total power consumption in the allocated BBU servers by time slots basis. In terms of power utilization in time slot basis, the upper bound based method consumes less power than normalized method. In the normalized method, the average power utilizes in BBU servers around 115 Watt, whereas the upper bound method on average saves 30 Watt during the five time slots. The choice of scheduling weight depends on the adjustment of response time requirement and server utilization or power consumption requirement. Therefore, the EE-JWSBA method provides the energy efficient when the workload are distributed according to upper bound based scheduling weight with the compromise of response time requirement.

4.6 Summary

In this chapter, we presented a joint workload distribution and BBU allocation method in C-RANs with the objective to minimize the mean response time and aggregate power. The C-RAN controller allocates scheduling weights to upcoming requests and divides the workload among BBU servers in time slot basis. Three scheduling weights (e.g. upper, normalized and random) were considered for evaluation. The BBU allocation shows a trade off between minimizing response time and minimizing power consumption. For example, the number of BBU servers are reduced if the workload is allocated based on upper limit based scheduling weight. However, this allocation method increases mean response time because all upcoming requests are scheduled to a minimum number of BBU servers. On the other hand, normalized scheduling weight provides minimum mean response time as this allocation method utilizes all BBU servers at each time instant. This scheduling method is suitable for delay sensitive scenarios whereas the former one is applicable for limited power and cost budget scenario.

In the next chapter, we focus on joint communication and computing resource allocation problem along with user association, and BBU-RRH mapping in OFDMA supported C-RANs. For communication resources, we consider resource blocks (RBs) and power allocation and for computing resources, we consider virtual machines (VMs) allocation. We consider each BBU server has fixed capacity limit, defined by the maximum number of VMs are run inside the BBU server. In this chapter, we considered the allocation of BBU servers as a computing resource allocation. In the next chapter, we consider VM allocation along with BBU-RRH mapping as a computing resource allocation in C-RANs.

Chapter 5

Joint Communication and Computing Resource Allocation in Cloud-RAN

5.1 Introduction

To deal with increasing demand of user association and resource allocation in cellular networks, the architecture of cloud radio access networks (C-RAN) is envisioned as an attractive paradigm that takes advantage of managing large number of small cells through the centralized cloud controller, known as base band processing unit or BBU pool. As mentioned earlier, the BBU pool contains BBU servers which support cloud computing technology. In the previous chapter, we investigated the energy efficient joint workload scheduling and BBU servers allocation in C-RAN. In this chapter, we investigate both communication and computing resource allocation in C-RAN. Different from the previous works, in this chapter, we apply the auction theory to solve communication and computing resource allocation along with user association and BBU-RRH mapping problem in OFDM based C-RANs with the objective to minimize delay.

The main contributions of this chapter can be summarized as follows:

- We first establish a queueing model in C-RAN. There are two queues: i) RRH transmission and ii) baseband processing queues are considered in radio access and BBU pool side, respectively. We formulate two optimization problems with the objective to minimize delay for small cell users.

- In the first optimization problem, we consider joint user association and communication resources (e.g., RB and power) allocation with the aim to minimize mean response time in RRH transmission queue. Maximum power, RB allocation, interference and queueing stability constraints are considered in this optimization problem.
- In the second optimization problem, we consider VM allocation to each small cell, constrained to one-to-one mapping with BBU and RRH, with maximum capacity limit of each BBU, and queueing stability constraint in baseband processing queue.
- To solve the communication and computing resource allocation problems jointly, we propose a double-sided distributed resource allocation method using auction theory, addressing all the aforementioned constraints. In the proposed method, the small cell users and base stations cooperatively decide transmission alignment (e.g., RB and power) and service rate. Using the communication resource allocation information for RBs and power, the centralized cloud controller, referred to as the auctioneer, decides the computing resource for each RRH using the concept of probability theory.
- The effectiveness of the proposed method is verified through Monte Carlo simulations.

5.2 Related Work

The computing resource optimization problem becomes a challenging task in cloud environment in terms of utilizing minimum physical resources (e.g, VMs, CPUs, servers), optimizing resource costs and minimizing delay. Moreover, joint optimization of computation and communication resource allocation is considered in mobile cloud computing (MCC) [51] and mobile edge computing (MEC) [44–46] environments. In [51], the authors optimized multi-task offloading decision with the help of computing access point in the MCC environment. Similarly, co-operation of users is considered in [44] to optimize energy consumption in computing and communication resource allocation in MEC. Computation offloading and fairness based MEC server selection in terms of cost minimization is considered in [45]. Similarly, in C-RANs and heterogeneous C-RANs (H-CRANs), the BBU server selection in a BBU pool and RRH-BBU mapping are considered as a computing resource allocation . On the other hand,in C-RANs and H-CRANs, the

Table 5.1: Summary of works on resource optimization problems.

Ref.	Research Problem	Objectives	Solution Approach	Solution Type
Works on MEC				
[44]	-Computation and communication resource allocation	-Minimize the energy consumption	- Lagrange dual method	Centralized
[45]	-Computation and communication resource allocation, -MEC sever selection	-Minimize cost	- Hungarian and fairness based algorithm	N/A†
[46]	-Spectrum allocation -Computation offloading -Content caching	-Maximize revenue	- Augmented Lagrangian based alternating direction method	Distributed
Works on C-RAN				
[14]	-User association	-Maximize ergodic capacity	-Received power based approach: i) N-best, ii) Nth best, and iii) single best	Centralized
[16]	-Computing resource allocation	-Energy efficiency	-Apply queueing model at remote radio head (RRH) side -Apply convex solver and WMSE approach	Distributed
[17]	-User association, -Communication resource allocation -Congestion control	-Energy efficiency, -Minimize delay	-Apply queueing theory -Stochastic geometric theory	N/A†
[18]	-BBU-RRH mapping -Communication resource allocation	-Minimize fronthaul overhead	-Heuristic algorithm	N/A†
[19]	-Communication resource allocation -Admission control	-Maximize the tolerable interference level	-Low complexity algorithm, -Lagrange multiplier	N/A†
Work utilizing auction theory				
[47]	-RB allocation	-Maximize throughput	-First price seal-bid based SINR auction method	Distributed
[48]	-Computing resource allocation	-Maximize utility	-Two Tier auction method	Distributed
[49]	-Computing resource allocation	-Maximize utility	-Combinational double auction method	Centralized
[50]	-Spectrum allocation	-Maximize revenue	-Time-line based auction method	Distributed
Proposed scheme	-User association, -Communication resource allocation -Computing resource allocation	-Minimize delay	-Lagrange multiplier -Double sided auction method -Probabilistic method for BBU-RRH maping	Distributed

† No information is available

computing and communication resource optimization problems consider delay, costs, RRH/VM utilization ratio, fairness, throughput, spectrum and energy efficiency as a QoS/QoE parameters [16]- [18], [52].

In C-RANs and H-CRANs, the data rate provisioning can be significantly improved by the fractional frequency reuse performed by small cells [26], in which specific partitions of the spectrum are shared between both RRHs and MBS to alleviate the inter-tier interference. In [53], authors studied a combinatorial optimization problem for joint resource block (RB) and power allocation in an OFDM based C-RAN system, in which the inter-tier interference is cancelled by imposing a constraint at the RRH side. Moreover, the central cooperative interference cancellation in C-RANs can significantly reduce the interference levels to provide high data rates. However, the centralized method is not scalable due to the dense deployment of small cells in a multi-tier system, turning the user association into a more challenging problem. Recently, the application of auction theory to allocate resources in a distributed way has received increasing attention amongst researchers of future wireless networks [54–57]. A comprehensive introduction and applicability of auction theory in wireless networks are provided in [54, 55]. Authors in [56], [58] proposed distributed framework for resource allocation in a multi-tier device-to-device communication enabled network, where in [56] used auction theory for distributed resource allocation. In [57], authors proposed a combinational auction algorithm to address the user association problem in 60 GHz millimeter wave wireless access network.

5.3 System Model and Assumptions

In this section, we initially present a queueing model of C-RAN in network model subsection. Then, the system models of RRH transmission and baseband processing queues are described in the subsequent subsections, respectively.

5.3.1 Network Model

In this chapter, we consider an OFDM based two tier uplink C-RAN network, as shown in Fig. 5.1, where $\mathcal{N}$ number of RRHs are covered by one macro cell in an underlay manner. Each

Table 5.2: List of symbols.

	Symbol	Description
Set	$\mathcal{N}$	Total number of RRHs
	$\mathcal{K}$	Total number of SUEs
	$\mathcal{R}$	Total number of RBs
	M	Total number of BBUs
	U	Maximum capacity or VMs of each BBU server
Index	i	Indexing for RRH
	j	Indexing for SUE
	r	Indexing for RB
	m	Indexing for BBU
	v	Indexing of VM
Queueing Parameters	λ_i^{RRH}	Average incoming requests of the i^{th} RRH, also denotes average data rate requirements of the i^{th} RRH
	λ_m^{BBU}	Average incoming requests of the m^{th} BBU
	$\mu_{i,j}$	Service rate of the i^{th} RRH for the j^{th} SUE
	μ_i^{RRH}	Mean service rate of i^{th} RRH
	μ_m^{BBU}	Mean service rate of m^{th} BBU
	$\mu_{m,v}^{BBU}$	Mean service rate of v^{th} VM
	T_i^{RRH}	Average response time of i^{th} RRH
	T_m^{BBU}	Average response time of m^{th} BBU
Optimization Parameters	$a_{i,j}$	User association parameter of i^{th} RRH for j^{th} SUE
	$b_{i,m}$	BBU-RRH mapping parameter of i^{th} RRH for m^{th} BBU
	$\beta_{i,j}^r$	RB allocation r^{th} RB to i^{th} RRH for user j
	$P_{i,j}^r$	Power allocation from i^{th} RRH to user j on r^{th} RB
Channel Parameters	$h_{i,j}^r$	The channel gain from i^{th} RRH to j^{th} SUE on r^{th} RB
	P_i^{max}	Maximum power of RRH i
	$\gamma_{i,j}^r$	SINR of j^{th} SUE connected to i^{th} RRH on r^{th} RB
	γ^{th}	SINR threshold value
Others	$\Gamma_{i,j}^r$	Parameter for both user association and RB allocation.
	$\mathcal{A}_{i,j}^r$	Parameter for power allocation
	$\sigma, \varsigma, \upsilon$	Lagrange multiplier vector
	ϕ, φ, ψ	
Auction Parameters	F_j^1,F_j^2	Bid information generated by user j
	F_i^1, F_i^2	Bid information for RRH i, generated by auctioneer
	F_i^3	Bid information generated by RRH i
	F_j^3	Acknowledgement from user j
	F_i^3	Acknowledgement from RRH i
	$\mathcal{B}_i$	Benefit of RRH i
	$\mathcal{U}_{ij}$	Utility of user j associated with RRH i

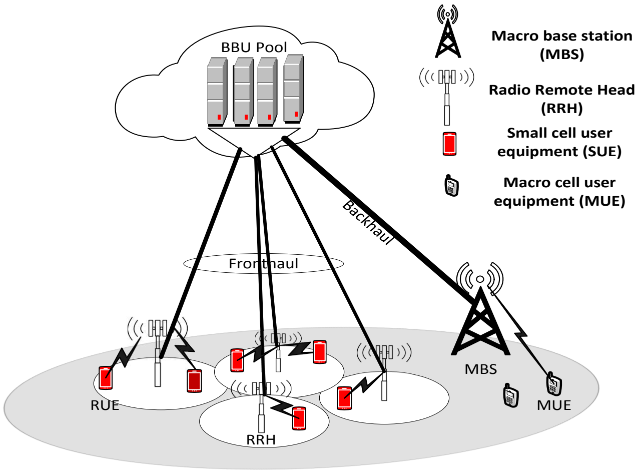


Figure 5.1: Small cell based C-RAN architecture for 5G networks.

small cell user (SUE) is equipped with one antenna and each RRH has L antenna. The system supports $\mathcal{K}$ number of users, where SUE is indexed by $j = \{1, 2, 3, ..., \mathcal{K}\}$ and RRH is indexed by $i = \{1, 2, 3, ..., \mathcal{N}\}$. For simplicity, we assume that each user is associated with one of the RRHs and one resource block is assigned between the user and RRH. The system supports $\mathcal{R}$ number of resource blocks, indexed by $r = \{1, 2, 3, ..., \mathcal{R}\}$. Fig. 5.2 represents the queueing network model of the two-tier C-RAN network. Each RRH has a transmission queue which receives requests from small cell users and processes the request at a pre-defined service rate. The RRH transmits the access requests of users to the BBU pool for baseband processing. The BBU pool is maintained by software defined C-RAN controller or scheduler which distributes the incoming requests to BBU servers for computation. Similar to [59], we assume that each BBU server runs a limited number of VMs, which refers to maximum capacity of each BBU server. For a summary of symbols and parameters, a list is provided in Table 5.2.

5.3.2 RRH Transmission Queue

Fig. 5.3 represents a queueing model for RRH. In the OFDM based C-RAN, RRHs receive requests in transmission time interval (TTI). At a given TTI, the RRH receives requests from $\mathcal{K}$ number of SUEs with Poisson arrival rate $\lambda_1^{SUE}, \lambda_2^{SUE}, ..., \lambda_{\mathcal{K}}^{SUE}$. The average incoming requests at i^{th} RRH is represented as $\lambda_i^{RRH} = \sum_{j=1}^{\mathcal{K}} a_{i,j}\lambda_j^{SUE}$, where $a_{i,j}$ denotes the user association

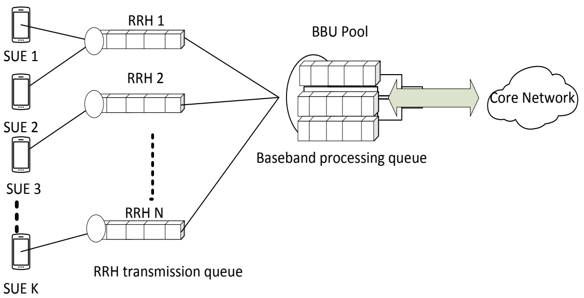


Figure 5.2: Queueing model of two tier C-RAN.

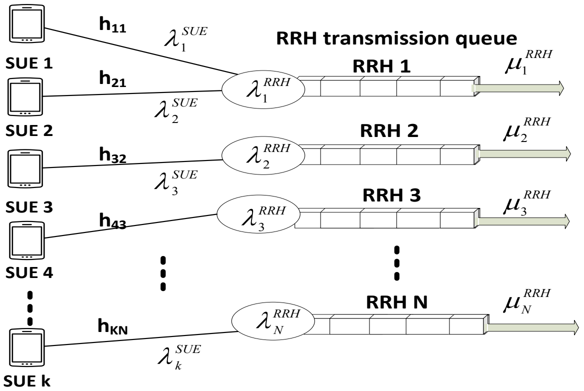


Figure 5.3: RRH transmission queue.

parameter, defined as

$$a_{i,j} = \begin{cases} 1, & \text{if user } j \text{ is associated with } i^{\text{th}} \text{ RRH}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

The arrival of scheduled requests to i^{th} RRH follows Poisson process with an average of λ_i^{RRH} and the inter-arrival service time is exponentially distributed with rate μ_i^{RRH}. The service rate of each RRH is related with transmission rate that varies with time variation of channel and states of base station [60], [21]. The service process of each RRH follows an M/M/1 queuing model. The average response time of i^{th} RRH can be formulated as:

$$T_i^{RRH} = \frac{1}{\mu_i^{RRH} - \lambda_i^{RRH}}. \tag{5.2}$$

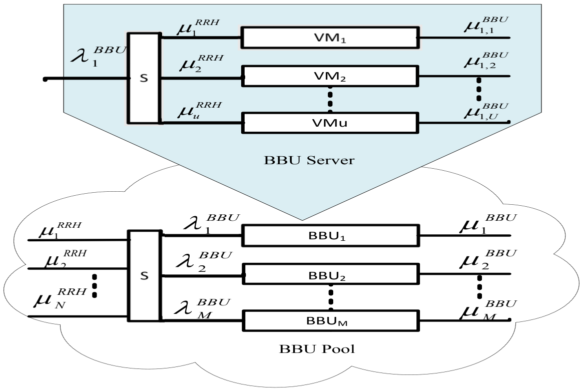


Figure 5.4: Baseband processing queue.

5.3.3 Baseband Processing Queue

Fig. 5.4 represents a queueing model for BBU pool. In C-RAN architecture the BBU pool is considered as a master base station which runs the baseband processing function into VMs in BBU servers. Assume that the BBU pool maintains M BBU servers and each server has maximum capacity U. That means each BBU server can generate a maximum number of U VMs. For BBU-RRH mapping, we assume that all the requests from one RRH is served by one VM in one BBU server. Assume that the BBU pool receives requests in TTI. At a given TTI, each BBU server receives requests from $\mathcal{N}$ number of RRHs with exponential service rate $\mu_1^{RRH}, \mu_2^{RRH}, ..., \mu_{\mathcal{N}}^{RRH}$. The average incoming request at m^{th} BBU is represented as $\lambda_m^{BBU} = \sum_{i=1}^{\mathcal{N}} b_{i,m}\mu_i^{RRH} = \sum_{i=1}^{\mathcal{N}} \sum_{v=1}^{U} b_{i,m}\lambda_{m,v}^{BBU}$, where $b_{i,m}$ denotes the BBU-RRH association parameter, defined as

$$b_{i,m} = \begin{cases} 1, & \text{if BBU } m \text{ is associated with } i^{\text{th}} \text{ RRH}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

We assume that each BBU server supports U number of VMs and each VM is allocated to one RRH. So the maximum number of U RRH is supported by one BBU server. Each VM maintains an M/M/1 queue and each BBU server is represented by a U M/M/1 queuing system, as shown in Fig. 5.4. The arrival of scheduled requests to v^{th} VM follows Poisson process with an average of $\lambda_{m,v}^{BBU}$ and the inter-arrival service time is exponentially distributed with rate $\mu_{m,v}^{BBU}$. As the service process of each VM follows an M/M/1 queuing model. The average response time of v^{th}

VM can be formulated as:

$$T_{m,v}^{BBU} = \frac{1}{\mu_{m,v}^{BBU} - \lambda_{m,v}^{BBU}}. \tag{5.4}$$

5.4 Problem Formulation

Let $\beta_{i,j}^{r}$ be the binary variable for RB allocation defined as follows:

$$\beta_{i,j}^{r} = \begin{cases} 1, & \text{if RB } r \text{ is assigned to RRH } i \text{ on SUE } j, \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

The channel gain from RRH i to SUE j on RB r is denoted as $h_{i,j}^{r} \in \mathcal{C}^{L\times 1}$. The power allocation from RRH i to user j on RB r is denoted as $P_{i,j}^{r} \in [0, P_{i}^{max}]$, where P_{i}^{max} is the maximum power of RRH i. The SINR achieved by SUE j connected to RRH i on RB r can be written as:

$$\gamma_{i,j}^{r} = \frac{|h_{i,j}^{r}|P_{i,j}^{r}}{\sum_{u\neq i, v\neq j} |h_{u,v}^{r}|P_{u,v}^{r} + \eta_0}, \tag{5.6}$$

where η_0 represents the zero mean and unit variance additive white Gaussian noise (AWGN) power. According to the Shannon's formula, the service rate for each user that is associated with i^{th} RRH can be obtained as:

$$\mu_{i,j} = \Delta B \sum_{r=1}^{\mathcal{R}} a_{i,j}\beta_{i,j}^{r}\log_2(1+\gamma_{i,j}^{r}), \tag{5.7}$$

where ΔB represents the available bandwidth of each RB. The average service rate of RRH i can be represented by:

$$\mu_{i}^{RRH} = \sum_{j=1}^{\mathcal{K}} \mu_{i,j} = \Delta B \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} a_{i,j}\beta_{i,j}^{r}\log_2(1+\gamma_{i,j}^{r}). \tag{5.8}$$

All the RRHs forward the requests to the centralized BBU pool for baseband processing. In an ideal scenario, each BBU data transmission should be equivalent to the RRH data transmission rate to satisfy SUE data rate requirements [17]. Therefore, the relation between the data rate requirements of SUE and the average BBU data transmission rate can be represented by:

$$\mu_m^{BBU} = \lambda_m^{BBU} = \mu_i^{RRH}. \tag{5.9}$$

5.4.1 Objective Function

Our objective for resource allocation is to minimize the delay of C-RAN by optimizing the average response time of each RRH and BBU server. The response time of each RRH depends on user association, RB and power allocation, and the response time of each BBU depends on BBU-RRH mapping and maximum capacity constraints. Theoretically, the delay of each RRH will be minimum when the RRH serves at a maximum rate. Therefore, the objective of each RRH can be defined as

$$\text{Minimize } T_i^{RRH} \quad \text{OR} \quad \text{Maximize } \mu_i^{RRH}. \tag{5.10}$$

Similarly the objective of each VM can be defined as

$$\text{Minimize } T_{m,v}^{BBU}. \tag{5.11}$$

5.4.2 Constraint Sets of Communication Resource Allocation

In order to ensure the minimum delay while all users are associated with RRH and received RB to transmit data without causing interference to each other, we define the following constraints set.

- The constraint (5.12) ensures that each user is associated with only one RRH, i.e.,

$$\sum_{i=1}^{\mathcal{N}} a_{i,j} = 1, \quad \forall j \in \mathcal{K}. \tag{5.12}$$

- The following constraint ensures that each associated user can use at most one RB for communication. For simplicity, we assume that each user and RRH connection utilizes one

RB for data transmission as

$$\sum_{r=1}^{\mathcal{R}} a_{i,j}\beta_{i,j}^{r} \leq 1, \quad \forall i, j. \tag{5.13}$$

- Constraint (5.14) verifies the SINR threshold value for allocated RBs. For spectrum efficiency, we consider that each SUE utilizes the macro cell users' RBs only when their instantaneous SINR exceeds a threshold value

$$a_{i,j}\beta_{i,j}^{r}\gamma_{i,j}^{r} > \gamma^{th}. \tag{5.14}$$

- The constraint (5.15) ensures that each RRH selects separate RBs for users to avoid co-tier interference. In (5.15), S_i represents the set of users that are associated with RRH i. According to this constraint, users in S_i utilize different RBs for data transmission as

$$a_{u,i}\beta_{u,i}^{r} + a_{v,i}\beta_{v,i}^{r} \leq 1, \quad \forall r \in \mathcal{R}, \forall (u,v) \in S_i. \tag{5.15}$$

- The constraint (5.16) verifies that the maximum power budget of each RRH and is given as

$$\sum_{j=1}^{\mathcal{K}}\sum_{r=1}^{\mathcal{R}} a_{i,j}\beta_{i,j}^{r}P_{i,j}^{r} \leq P_{i}^{max}, \quad \forall i \in \mathcal{N}. \tag{5.16}$$

- The constraint (5.17) maintains the queueing stability at each RRH, i.e., the arrival rate should not be greater than the service rate.

$$\lambda_{i}^{RRH} \leq \mu_{i}^{RRH}, \quad \forall i \in \mathcal{N}. \tag{5.17}$$

Considering the objective of each RRH and the aforementioned constraints, the resource optimization problem for each RRH can be formulated as:

$$\textbf{P1:} \min_{a_{i,j},\beta^r_{i,j},P^r_{i,j}} \; T_i^{RRH} \tag{5.18}$$

$$\begin{aligned}
&\text{subject to:}\\
&\text{C1:} \quad \sum_{i=1}^{\mathcal{N}} a_{i,j} = 1, \quad \forall j \in \mathcal{K},\\
&\text{C2:} \quad \sum_{r=1}^{\mathcal{R}} a_{i,j}\beta^r_{i,j} \leq 1, \quad \forall i, j,\\
&\text{C3:} \quad a_{i,j}\beta^r_{i,j}\gamma^r_{i,j} > \gamma^{th},\\
&\text{C4:} \quad a_{u,i}\beta^r_{u,i} + a_{v,i}\beta^r_{v,i} \leq 1, \quad \forall r \in \mathcal{R}, \forall (u,v) \in S_i,\\
&\text{C5:} \quad \sum_{j=1}^{\mathcal{K}}\sum_{r=1}^{\mathcal{R}} a_{i,j}\beta^r_{i,j}P^r_{i,j} \leq P_i^{max}, \quad \forall i \in \mathcal{N},\\
&\text{C6:} \quad \lambda_i^{RRH} \leq \mu_i^{RRH}, \quad \forall i \in \mathcal{N},\\
&\text{C7:} \quad a_{i,j}, \beta^r_{i,j} \in \{1, 0\}, \quad \forall j \in \mathcal{K}, \forall r \in \mathcal{R}\\
&\text{C8:} \quad P^r_{i,j} \geq 0, \quad \forall r \in \mathcal{R},
\end{aligned}$$

where C1 to C6 refer to the constraints (5.12) to (5.17), respectively. The constraint C7 represents that the decision variable for user association and RB allocation are the binary variables. Finally, C8 defines the non-negativity condition of transmit power.

Corollary 1. *The objective function in (5.18) and the constraints C6 and C7 turn the problem **P1** into a mixed integer non-linear program (MINLP) with the non-convex feasibility set. The optimization problem P1 is computationally intractable and is a NP-hard problem [56], [58].*

5.4.3 Constraint Sets of Computing Resource Allocation

In order to ensure the minimum delay in baseband processing queue, assume that each RRH is connected to only one VM in one BBU server and each BBU server can utilize maximum number of U VMs. The BBU-RRH mapping constraints can be defined as follows:

- The constraint (5.19) ensures that each RRH is associated with only one VM in one BBU server, i.e.,

$$\sum_{m=1}^{M} b_{i,m} = 1, \quad \forall i \in \mathcal{N}. \tag{5.19}$$

- The constraint (5.20) ensures the maximum capacity limits of each BBU server as

$$\sum_{i=1}^{\mathcal{N}} b_{i,m} \leq U, \quad \forall m \in M. \tag{5.20}$$

- The constraint (5.21) maintains the queueing stability at each VM, i.e., the arrival rate should not be greater than the service rate, which is

$$\lambda_{m,v}^{BBU} \leq \mu_{m,v}^{BBU}, \quad \forall m \in M, \forall v \in U. \tag{5.21}$$

Considering the aforementioned constraints, the computing resource allocation problem can be formulated as:

$$\textbf{P2:} \min_{b_{i,m}} T_{m,v}^{BBU} \tag{5.22}$$

$$\begin{aligned}
&\text{subject to:}\\
&\text{C1:} \quad \sum_{m=1}^{M} b_{i,m} = 1, \quad \forall i \in \mathcal{N},\\
&\text{C2:} \quad \sum_{i=1}^{\mathcal{N}} b_{i,m} \leq U, \quad \forall m \in M,\\
&\text{C3:} \quad \lambda_{m,v}^{BBU} \leq \mu_{m,v}^{BBU}, \quad \forall m \in M, \forall v \in U,\\
&\text{C4:} \quad b_{i,m} \in \{1, 0\}, \quad \forall i \in \mathcal{N}, \forall m \in M.
\end{aligned}$$

where $b_{i,m}$ represents that the decision variable for BBU-RRH mapping is a binary variable.

Corollary 2. *The objective function in (5.22) and the constraints C3 and C4 turn the problem **P2** into a mixed integer non-linear program (MINLP) with the non-convex feasibility set. The optimization problem **P2** is computationally intractable and is a NP-hard problem [56], [58].*

5.5 Relaxation to Fractional Resource Allocation

We relax the problem **P1** by replacing non-convex constraints with convex constraints. First, we relax the constraint C7 by assuming time sharing approach [36] of RB allocation, i.e., $0 \leq \beta_{i,j}^{r} \leq 1$. We introduce two new variables $\Gamma_{i,j}^{r} = a_{i,j} \times \beta_{i,j}^{r} \in (0,1]$ and $\mathcal{A}_{i,j}^{r} = \Gamma_{i,j}^{r} \times P_{i,j}^{r} \in (0,1]$. $\Gamma_{i,j}^{r}$ represents both user association and sharing factor of resource block. It denotes the portion of time the RB r is allocated to the user j and RRH i link. $\mathcal{A}_{i,j}^{r}$ denotes the actual transmit power of SUE j on RB r. Next, we relax the constraint C6 by assuming no interference, $\gamma_{i,j}^{r} = \frac{|h_{i,j}^{r}|P_{i,j}^{r}}{\eta_0} = \delta_{i,j}^{r}P_{i,j}^{r} = \frac{\delta_{i,j}^{r}\mathcal{A}_{i,j}^{r}}{\Gamma_{i,j}^{r}}$, where the service rate of each RRH becomes $\mu_i^{RRH} = \Delta B \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \Gamma_{i,j}^{r} \log_2\left(1 + \frac{\delta_{i,j}^{r}\mathcal{A}_{i,j}^{r}}{\Gamma_{i,j}^{r}}\right)$. The primary formulation in **P1** can be expressed in an equivalent form by including new sets of variables $\Gamma_{i,j}^{r}$ and $\mathcal{A}_{i,j}^{r}$. The relaxed problem can be represented by:

$$\textbf{P3:} \max_{\Gamma_{i,j}^{r}, \mathcal{A}_{i,j}^{r}} \mu_i^{RRH} - \lambda_i^{RRH} \tag{5.23}$$

subject to:

$$\text{C1:} \quad \sum_{i=1}^{\mathcal{N}} \frac{\Gamma_{i,j}^{r}}{\beta_{i,j}^{r}} = 1, \quad \forall j \in \mathcal{K},$$

$$\text{C2:} \quad \sum_{r=1}^{\mathcal{R}} \Gamma_{i,j}^{r} \leq 1, \quad \forall i, j,$$

$$\text{C3:} \quad -\mathcal{A}_{i,j}^{r}\delta_{i,j}^{r} + \gamma^{th} \leq 0,$$

$$\text{C4:} \quad \Gamma_{u,i}^{r} + \Gamma_{v,i}^{r} \leq 1, \quad \forall r \in \mathcal{R}, \forall (u,v) \in S_i,$$

$$\text{C5:} \quad \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \mathcal{A}_{i,j}^{r} \leq P_i^{max}, \quad \forall i \in \mathcal{N},$$

$$\text{C6:} \quad \lambda_i^{RRH} \leq \mu_i^{RRH}, \quad \forall i \in \mathcal{N},$$

$$\text{C7:} \quad \beta_{i,j}^{r} \in (0,1], \quad \forall j \in \mathcal{K}, \forall r \in \mathcal{R}.$$

As the number of resource blocks becomes relatively large, the duality gap of any optimization problem satisfying time sharing condition becomes negligible [61]. The solution of relaxed optimization problem **P3** is asymptotically optimal since it satisfies the time sharing condition [36].

Corollary 3. *The relaxed optimization problem* ***P3*** *is convex; the objective function is concave and all the constraints are affine.*

Since **P3** is a non-linear convex problem, the interior point methods can be used solve this problem [62].

To observe the nature of RB and power allocation, we formulate an equivalent problem **P3** as a base and use Karush-Kuhn-Tucker (KKT) optimality and define the following Lagrangian function:

$$
\begin{aligned}
&\mathbb{L}(\Gamma, \mathcal{A}, \sigma, \varsigma, \upsilon, \phi, \varphi, \psi) = \\
&\Delta B \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \Gamma_{i,j}^{r} \log_2 \left(1 + \frac{\delta_{i,j}^{r} \mathcal{A}_{i,j}^{r}}{\Gamma_{i,j}^{r}}\right) - \sum_{j=1}^{K} \lambda_j^{SUE} \\
&+ \sum_{j=1}^{\mathcal{K}} \sigma_j \left(1 - \sum_{i=1}^{\mathcal{N}} \frac{\Gamma_{i,j}^{r}}{\beta_{i,j}^{r}}\right) + \sum_{j=1}^{\mathcal{K}} \sum_{i=1}^{\mathcal{N}} \varsigma_{i,j} \left(1 - \sum_{r=1}^{\mathcal{R}} \Gamma_{i,j}^{r}\right) \\
&+ \sum_{r=1}^{\mathcal{R}} \upsilon_r (0 + \mathcal{A}_{i,j}^{r} \delta_{i,j}^{r} - \gamma^{th}) + \sum_{i=1}^{\mathcal{N}} \sum_{r=1}^{\mathcal{R}} \phi_{i,r} (1 - \Gamma_{u,i}^{r} - \Gamma_{v,i}^{r}) \\
&+ \sum_{i=1}^{\mathcal{N}} \varphi_i \left(P_i^{max} - \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \mathcal{A}_{i,j}^{r}\right) \\
&+ \sum_{i=1}^{\mathcal{N}} \psi_i \left(\Delta B \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \Gamma_{i,j}^{r} \log_2 \left(1 + \frac{\delta_{i,j}^{r} \mathcal{A}_{i,j}^{r}}{\Gamma_{i,j}^{r}}\right) - \sum_{j=1}^{K} \lambda_j^{SUE}\right),
\end{aligned}
\tag{5.24}
$$

where φ, and ψ are the vectors of Lagrange multipliers associated with power and queueing stability requirements for cellular and SUEs, respectively. Similarly, $\sigma, \varsigma, \upsilon, \phi$ are the Lagrange multipliers for the constraints C1-C4. Differentiating (5.24) with respect to $\mathcal{A}_{i,j}^{r}$, we obtain the following power allocation of SUE i over RB r as

$$
P_{i,j}^{r} = \frac{\mathcal{A}_{i,j}^{r}}{\Gamma_{i,j}^{r}} = \left[\xi - \frac{1}{\delta_{i,j}^{r}}\right]^{+}, \tag{5.25}
$$

where $\xi = \frac{\Delta B(1+\psi_i)}{\ln(\varphi_i - \upsilon_r \delta^r_{i,j})}$ and $[\varepsilon]^+ = \max(\varepsilon, 0)$, which is a multi-level water filling allocation [36].

Proof: The dual problem of (5.24) is :

$$\begin{aligned}
&\mathbb{D}(\sigma, \varsigma, \upsilon, \phi, \varphi, \psi) = \\
&\max_{\Gamma,\mathcal{A}} \Delta B \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \Gamma^r_{i,j} \left(1 + \sum_{i=1}^{\mathcal{N}} \psi_i\right) \log_2 \left(1 + \frac{\delta^r_{i,j} \mathcal{A}^r_{i,j}}{\Gamma^r_{i,j}}\right) \\
&- \sum_{i=1}^{\mathcal{N}} \varphi_i \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \mathcal{A}^r_{i,j} - \left(1 + \sum_{i=1}^{\mathcal{N}} \psi_i\right) \sum_{j=1}^{K} \lambda^{SUE}_j \\
&- \sum_{j=1}^{\mathcal{K}} \sigma_j \sum_{i=1}^{\mathcal{N}} \frac{\Gamma^r_{i,j}}{\beta^r_{i,j}} - \sum_{j=1}^{\mathcal{K}} \sum_{i=1}^{\mathcal{N}} \varsigma_{i,j} \sum_{r=1}^{\mathcal{R}} \Gamma^r_{i,j} + \sum_{r=1}^{\mathcal{R}} \upsilon_r \mathcal{A}^r_{i,j} \delta^r_{i,j} \\
&- \sum_{i=1}^{\mathcal{N}} \sum_{r=1}^{\mathcal{R}} \phi_{i,r} (\Gamma^r_{u,i} + \Gamma^r_{v,i}) + \sum_{i=1}^{\mathcal{N}} \varphi_i P^{max}_i + \sum_{j=1}^{\mathcal{K}} \sigma_j \\
&+ \sum_{j=1}^{\mathcal{K}} \sum_{i=1}^{\mathcal{N}} \varsigma_{i,j} - \sum_{r=1}^{\mathcal{R}} \upsilon_r \gamma^{th} + \sum_{i=1}^{\mathcal{N}} \sum_{r=1}^{\mathcal{R}} \phi_{i,r}
\end{aligned} \tag{5.26}$$

Considering only the power allocation (e.g., $\mathcal{A}^r_{i,j}$) part from (5.26):

$$\begin{aligned}
\text{Ł}(\mathcal{A}^r_{i,j}) = \Delta B \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \Gamma^r_{i,j} \left(1 + \sum_{i=1}^{\mathcal{N}} \psi_i\right) \log_2 \left(1 + \frac{\delta^r_{i,j} \mathcal{A}^r_{i,j}}{\Gamma^r_{i,j}}\right) \\
- \sum_{i=1}^{\mathcal{N}} \varphi_i \sum_{j=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} \mathcal{A}^r_{i,j} + \sum_{r=1}^{\mathcal{R}} \upsilon_r \mathcal{A}^r_{i,j} \delta^r_{i,j}
\end{aligned}$$

Maximizing **P3** for any given $\Gamma^r_{i,j}$ is equivalent to differentiating $\text{Ł}(\mathcal{A}^r_{i,j})$ with respect to $\mathcal{A}^r_{i,j}$ and setting the result to zero. That is

$$\frac{\partial \text{Ł}}{\partial \mathcal{A}^r_{i,j}} = 0$$

$$\frac{\Delta B\Gamma^{r}_{i,j}(1+\psi_i)}{\ln\left(1+\frac{\delta^{r}_{i,j}\mathcal{A}^{r}_{i,j}}{\Gamma^{r}_{i,j}}\right)}\frac{\delta^{r}_{i,j}}{\Gamma^{r}_{i,j}}-\varphi_i+\upsilon_r\delta^{r}_{i,j}=0$$

$$P^{r^*}_{i,j}=\frac{\mathcal{A}^{r}_{i,j}}{\Gamma^{r}_{i,j}}=\left[\frac{\Delta B(1+\psi_i)}{ln(\varphi_i-\upsilon_r\delta^{r}_{i,j})}-\frac{1}{\delta^{r}_{i,j}}\right]^{+}.$$

□

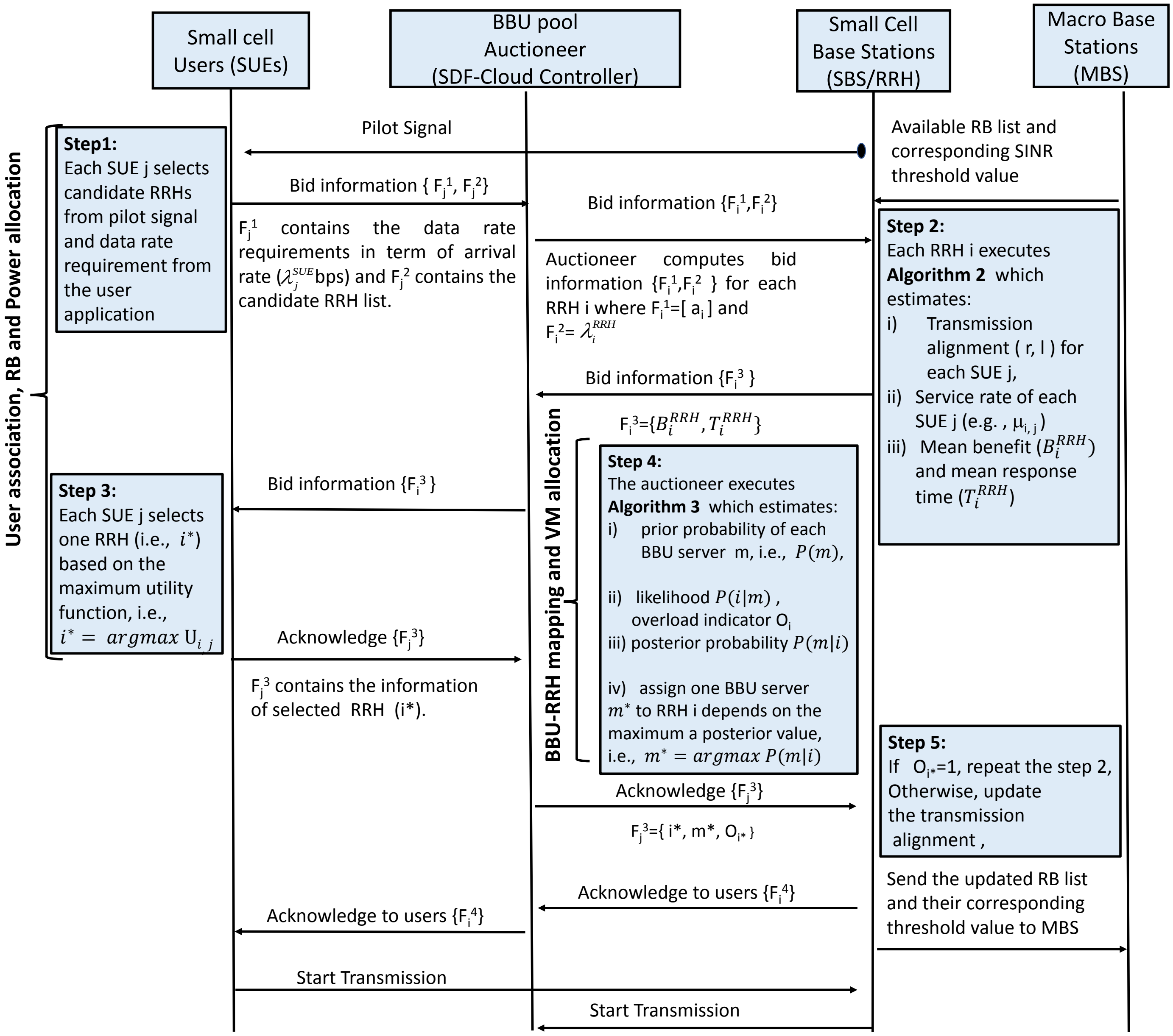
Although we have obtained a closed-form solution for optimal power allocation using Lagrange multipliers, it is still difficult to solve for optimal radio resource allocation from (5.24) due to the mathematical intractability. In the next section, we present an auction based distributed solution that satisfies all the constraints of the original problems **P1** and **P2** under the assumption that the system is feasible given the network size, number of RBs and SINR threshold value.

5.6 Double-Sided Auction based Distributed Resource Allocation (DS-ADRA)

The objective of our proposed solution is to optimize the mean response time in a C-RAN system in terms of solving communication and computing resource allocation along with user association, and BBU-RRH mapping in C-RANs. To solve the communication and computing resource optimization problem, in this section, we propose a joint resource allocation solution using a DS-ADRA method, where small cell base stations and users jointly participate using the concept of auction theory.

5.6.1 User Association and Communication Resource Allocation

In the auction based resource allocation procedure, we assume that small cell users (e.g., SUEs) and small cell base stations (e.g., SBS/RRH) are the agents. An auctioneer or cloud controller (CC) is a software defined module, which resides in BBU pool for controlling resources in C-RAN. It is assumed that all SUEs and RRHs within the C-RAN are always connected to the auctioneer using the control plane. The exchange of bidding information among the SUEs, RRHs



Small cell Users (SUEs)
BBU pool Auctioneer (SDF-Cloud Controller)
Small Cell Base Stations (SBS/RRH)
Macro Base Stations (MBS)
User association, RB and Power allocation
Step1: Each SUE j selects candidate RRHs from pilot signal and data rate requirement from the user application
Pilot Signal
Available RB list and corresponding SINR threshold value
Bid information { F_j^1, F_j^2}
F_j^1 contains the data rate requirements in term of arrival rate (λ_j^{SUE}bps) and F_j^2 contains the candidate RRH list.
Bid information {F_i^1,F_i^2}
Auctioneer computes bid information {F_i^1,F_i^2 } for each RRH i where F_i^1=[a_i] and F_i^2= λ_i^{RRH}
Step 2: Each RRH i executes Algorithm 2 which estimates:
i) Transmission alignment (r, l) for each SUE j,
ii) Service rate of each SUE j (e.g. , $\mu_{i,j}$)
iii) Mean benefit (B_i^{RRH}) and mean response time (T_i^{RRH})
Bid information {F_i^3 }
F_i^3={B_i^{RRH}, T_i^{RRH}}
BBU-RRH mapping and VM allocation
Step 3: Each SUE j selects one RRH (i.e., i^*) based on the maximum utility function, i.e., $i^* = argmax\ U_{i\ j}$
Bid information {F_i^3 }
Step 4: The auctioneer executes Algorithm 3 which estimates:
i) prior probability of each BBU server m, i.e., $P(m)$,
ii) likelihood $P(i|m)$, overload indicator O_i
iii) posterior probability $P(m|i)$
iv) assign one BBU server m^* to RRH i depends on the maximum a posterior value, i.e., $m^* = argmax\ P(m|i)$
Acknowledge {F_j^3}
F_j^3 contains the information of selected RRH (i*).
Acknowledge {F_j^3}
F_j^3={ i*, m*, O_{i*} }
Step 5: If O_{i*}=1, repeat the step 2, Otherwise, update the transmission alignment ,
Acknowledge to users {F_i^4}
Acknowledge to users {F_i^4}
Send the updated RB list and their corresponding threshold value to MBS
Start Transmission
Start Transmission

Figure 5.5: Auction based distributed resource allocation method.

and BBU pool are done through the control plane. When the communication and computing resource allocations are done, the SUEs are setup the data plane to the selected RRHs and start data transmission. The exchange of bidding information and the detailed auction procedure are illustrated in Fig. 5.5 and works as follows:

Step 1: Whenever a user j receives the pilot signal from base stations, it generates a bid information and sends this information to base stations through the auctioneer. The bid information contains two types of information, i.e., F_j^1 and F_j^2. F_j^1 represents the data rate requirement based on which applications are running on the user side, i.e., $F_j^1 = \lambda_j^{SUE}$. The data rate requirement (in bps) serves as an arrival rate to the base stations. The F_j^2 contains information about the candidate base station list. A base station becomes a candidate for users, when the following condition is satisfied:

$$\begin{gathered} F_j^2 = \{a_{i,j}\} = [\mathbf{a}_j] = \mathbf{a}_j, \forall i \in \mathcal{N} \\ a_{i,j} = 1, \text{ when } (\tfrac{\pi r_{ij}^2}{\pi R_i^2}) \leq 1, \end{gathered} \tag{5.27}$$

where r_{ij} represents the distance between SUE j and RRH i and R_i denotes the radius of i^{th} RRH. F_j^2 represents the column vector, i.e., $\mathbf{a}_j = [a_{1,j}, a_{2,j}..., a_{\mathcal{N},j}]^T$ of the user association matrix $\mathbf{A}$, defined as

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & .. & .. & a_{1,\mathcal{K}} \\ a_{2,1} & a_{2,2} & .. & .. & a_{2,\mathcal{K}} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ a_{\mathcal{N},1} & a_{\mathcal{N},2} & .. & .. & a_{\mathcal{N},\mathcal{K}} \end{pmatrix} \tag{5.28}$$

The auctioneer receives requests from all SUEs and generates two types of information for each RRH. The first information contains the initial association vector for i^{th} RRH, i.e.,

$$F_i^1 = \{F_j^1\} = \{a_{i,j}\} = [\mathbf{a}_i] = \mathbf{a}_i, \quad \forall j \in \mathcal{K}. \tag{5.29}$$

F_i^1 represents the row vector, i.e., $\mathbf{a}_i = [a_{i,1}, a_{i,2}..., a_{i,\mathcal{K}}]$ of the user association matrix $\mathbf{A}$. The second information contains the mean data rate requirements of each RRH, i.e.,

$$F_i^2 = \lambda_i^{RRH}. \tag{5.30}$$

Step 2: The base station executes **Algorithm** 2 to determine transmission alignment (r, l) (i.e., RB r and transmission power l) and the benefit for using (r, l). **Algorithm** 2 satisfies all the constraints of problem **P1** to determine a stable allocation of transmission alignment for each user. The benefit of using (r, l) for SUE j associated with RRH i is defined as $\mathcal{B}_{i,j} = v_1 \mu_{ij}$, where v_1 is the bias factor for resource allocation in small cell user. The benefit of RRH i can be defined as

$$\begin{aligned} \mathcal{B}_i^{RRH} &= \textstyle\sum_{j=1}^{\mathcal{K}} \mathcal{B}_{i,j} \\ &= v_1 \textstyle\sum_{j=1}^{\mathcal{K}} \mu_{i,j} \\ &= v_1 \mu_i^{RRH}. \end{aligned} \tag{5.31}$$

If the base station finds a feasible assignment for users, it sends the bid information, i.e., $F_i^3 = \{\mathcal{B}_i^{RRH}, T_i^{RRH}\}$ to users which contains the expected service rate and mean response time of the base station.

Proposition 1. *The transmission alignment (r, l) performed by **Algorithm** 2 leads to a stable allocation.*

Proof: Depending on the initial user association vector $\mathbf{a}_i$, the total power of RRH i is equally allocated among users at the initial stage of **Algorithm** 2 to maintain the **C5** constraints in **P1**. A RB $r \in \mathcal{R}$ is selected for the user when it satisfies the RB association and SINR constraints (Line 9 and 11 in **Algorithm** 2). Let us assume that the transmission alignment (r, l) is allocated for user j by **Algorithm** 2. This allocation is stable since the same RB r cannot be selected by another user j'. Line 9 in **Algorithm** 2 blocks the selection of r for user j'. Therefore, the transmission alignment (r, l) leads to a stable allocation. □

Proposition 2. *The communication resources allocation performed by **Algorithm** 2 terminates after some finite number of steps.*

Proof: Let the finite set $\{\beta_i\}$ denote all possible combinations of users and RBs matching for RRH i, where each element $\beta_{i,j}^r \in \{\beta_i\}$ denotes the RB r is allocated to the user j in RRH i.

Since constraints **C2** and **C4** of **P1** are applied in **Algorithm** 2, no users select the same RB more than once. Therefore, the finiteness of the set $\{\beta_i\}$ ensures the termination of **Algorithm** 2 in finite number of steps. □

Algorithm 2: Communication resource allocation for i^{th} RRH

```
Input: Receives bid information $F_i^1$ and $F_i^2$, containing the information about the initial user association
       ($\mathbf{a}_i$), average arrival rate ($\lambda_i^{RRH}$), available RBs and corresponding threshold value ($\gamma^{th}$)
Output: Returns the bid information $F_i^3$ which represents the total benefit ($\mathcal{B}_i^{RRH}$) and mean response time
        ($T_i^{RRH}$) for $i^{\text{th}}$ RRH
I. Initialization: From initial user association matrix $\mathbf{A}$, the $i^{\text{th}}$ RRH counts the total number of user
   requests and sets equal power level for each user $P_{i,j} = \frac{P_i^{max}}{\sum_{j=1}^{\mathcal{K}} a_{i,j}}$
$\mu_i^{RRH} \leftarrow \infty$
$\beta_{i,j}^r \leftarrow 0$
II. Iteration:
for $j \leftarrow 1$ to $\mathcal{K}$ do
    if $\mu_i^{RRH} > \lambda_i^{RRH}$ then // Check constraint C6 in eq.(5.17)
        if $a_{i,j} == 1$ then
            for $r \leftarrow 1$ to $\mathcal{R}$ do
                if $\beta_{i,j}^r == 0$ then // Check constraint C2 and C4 in eq.(5.13) and eq.(5.15) respectively
                    Estimate $\gamma_{i,j}^r$ using eq.(5.6)
                    if $\gamma_{i,j}^r >= \gamma^{th}$ then // Check constraint C3 in eq.(5.14)
                        $\beta_{i,j}^r \leftarrow 1$
                        $a_{i,j} \leftarrow 1$
                        Estimate service rate of each user ($\mu_{i,j}$ ) and mean service rate of RRH ($\mu_i^{RRH}$) using eqs. (5.7) and
                        (5.8)
                        Calculate response time of RRH ($T_i^{RRH}$) using eq.(5.2)
                        break;
        else
            $a_{i,j} \leftarrow 0$
            $\mu_{i,j} \leftarrow 0$
Estimate $\mathcal{B}_i^{RRH}$ and $T_i^{RRH}$ using eqs.(5.31) and (5.2)
return $\mathcal{B}_i^{RRH}$ and $T_i^{RRH}$
```

Step 3: Each SUE calculates own utility function, i.e., $\mathcal{U}_{ij} = [\mathcal{B}_i^{RRH} - \lambda_j^{SUE}]^+ = [\mu_i^{RRH} - \lambda_j^{SUE}]^+$, where $[.]^+ = \max\{0;\}$. This means if $\lambda_j^{SUE} > \mu_i^{RRH}$, then the utility set is zero ($\mathcal{U}_{ij} = 0$). Each SUE can choose one RRH (i^*) based on the maximum utility function, i.e.,

$$i^* = \arg\max \mathcal{U}_{ij}, \forall i \in \mathcal{N}. \tag{5.32}$$

The SUE j acknowledges the selected RRH (i^*) by sending acknowledgement message F_j^3.

Proposition 3: *The equilibrium assignment holds for user j when $\mathcal{U}_{ij} \geq \max_{i \neq i', i \in \mathcal{N}} \{\mathcal{U}_{i'j}\}$ is satisfied for all RRHs.*

Proof: It can be proven that $\mathcal{U}_{ij} \geq \max_{i \neq i', i \in \mathcal{N}} \{\mathcal{U}_{i'j}\}$ satisfies if and only if the conditions

$$\mathcal{U}_{ij} = C_i, \tag{5.33}$$

and

$$\mathcal{U}_{ij} \geq \max_{i \neq i', i \in \mathcal{N}} \{\mathcal{U}_{i'j}\} \tag{5.34}$$

are satisfied for all i given that $i = \{F_j^1 \in \{a_{i,j}\}\}$ and $\{a_{i,j}\} \neq 0$. If (5.33) is not satisfied, or equivalently, if there exists $i' \in F_j^1$ such that $\mathcal{U}_{i'j} \geq \mathcal{U}_{ij}$, that is $[\mathcal{B}_{i'}^{RRH} - \lambda_j^{SUE}]^+ \geq [\mathcal{B}_i^{RRH} - \lambda_j^{SUE}]^+$, where data rate requirement of user j is same for both RRHs, then user j selects i' in Step 3 in DS-ADRA procedure based on the maximum utility. □

5.6.2 Maximum A Posterior Probability (MAP) Based BBU-RRH Mapping and Computing Resource Allocation

Step 4: In the auction procedure, we assume that the auctioneer i.e., SDF-cloud controller is always connected to the RRHs, which in turn collects and analyses all the bid information among the SUEs and RRHs. Based on the bid information $F_i^3 = \{\mathcal{B}_i^{RRH}, T_i^{RRH}\}$, and the acknowledgement message $F_j^3 = i^*$, the cloud controller executes **Algorithm** 3 which determines posterior probability based BBU-RRH association value, where each RRH can be associated with one BBU server and utilizes one VM for baseband processing depending on the expected service rate and BBU capacity information. For simplicity, we assume that all the BBU servers in a BBU pool maintain the same capacity U. Therefore, each BBU server can generate U number of VMs. Also, we assume that the mean service rate of all VMs in each BBU server is same and the cloud controller knows the mean service rate of VM in each BBU server i.e. $\mu_{m,v}^{BBU}$.

Assuming that the expected service rate μ_i^{RRH} of the i^{th} RRH is known based on the benefit information of bid F_i^3, the MAP based BBU-RRH mapping operates as follows:

Algorithm 3: Computing resource allocation for i^{th} RRH

```
Input: Receives the service rate ($\mu_i^{RRH}$) and mean response time ($T_i^{RRH}$) from the bid information $F_i^3$.
Output: Returns BBU association (e.g., $[b_i] = b_{i,m}, \forall m \in M$) and overload indicator flag ($O_i$) for RRH $i$
I. Initialization: Initialize total number of BBU server($M$), maximum number of VM ($U$) and mean service rate of each VM ($\mu_{m,v}^{BBU}$)
Initialize all BBU server have same capacity ($U$) and the mean service rate of all VMs in each BBU server is same
$b_{i,m} \leftarrow 0$
II. Iteration:
for $t \leftarrow 1$ to $T$ do
    for $m \leftarrow 1$ to $M$ do // i) Prior estimation
        if $t == 1$ then
            $\mathcal{C}_m(t) = U$
            $P(m) = 1/\mathcal{C}_m(t)$
        else
            $q_m(t) = \mathcal{C}_m(t-1) - \sum(b(:,m) == 1)$
            if $q_m(t) <= 0$ then
                $P(m) = \epsilon$
            else
                $P(m) = 1/q_m(t)$
        if $\mu_i^{RRH} \leq \mu_{m,v}^{BBU}$ then // ii) Likelihood estimation
            $L(i,m) = \frac{\mu_i^{RRH}}{\mu_{m,v}^{BBU}}$
            $P(i,m) = \frac{L(i,m)}{\sum_{\forall m} L(i,m)}$ using eq.(5.36)
            $O_i = 0$
        else
            $L(i,m) = 0$
            $P(i,m) = 0$
            $O_i = 1$ // Set the overflow flag
    for $m \leftarrow 1$ to $M$ do // iii) Posterior estimation
        $T(m) = T(m) + P(i,m) * P(m)$
        $Post(m,i) = \frac{P(i,m)P(m)}{T(m)}$ using eq.(5.37)
    // iv) BBU-RRH mapping and VMs allocation $[V, M] = max(Post(:,i))$
    $m = index(M)$
    $feasibility = \mathcal{C}_m(t)$
    if $feasibility > 0$ then
        $b(i,m) = 1$
        $\mathcal{C}_m(t+1) = \mathcal{C}_m(t) - 1$
    else
        $\mathcal{C}_m(t+1) = \mathcal{C}_m(t)$
return $[b_i]$ and $O_i$
```

- Feasibility and Prior Probability: The auctioneer determines which BBU server becomes feasible for RRH i. Using the BBU capacity, U, and BBU-RRH association information, the feasibility of each BBU server is determined at time t by:

  $$q_m(t) = U - \sum_{i=1}^{\mathcal{N}} b_{i,m} \quad \forall m \in M.$$

  The prior probability of each BBU server is estimated according to their feasibility information. The prior probability of m^{th} BBU can be estimated as

  $$P(m) = \begin{cases} \frac{1}{q_m(t)}, & q_m(t) \neq 0 \\ \epsilon, & \text{otherwise.} \end{cases} \tag{5.35}$$

- Likelihood: The likelihood is estimated as per BBU server basis. For given BBU m, the likelihood of RRH i^* can be estimated as

  $$P(i^*|m) = \begin{cases} \frac{L(i^*|m)}{\sum_{\forall m} L(i^*|m)}, & \text{if } \mu_{i^*}^{RRH} \leq \mu_{m,v}^{BBU} \\ 0, & \mu_{i^*}^{RRH} > \mu_{m,v}^{BBU}, \text{ sets } O_{i^*} = 1, \end{cases} \tag{5.36}$$

  where $L(i^*|m)$ can be estimated as

  $$L(i^*|m) = \begin{cases} \frac{\mu_{i^*}^{RRH}}{\mu_{m,v}^{BBU}}, & \text{if } \mu_{i^*}^{RRH} \leq \mu_{m,v}^{BBU} \\ 0, & \mu_{i^*}^{RRH} > \mu_{m,v}^{BBU}. \end{cases}$$

  When the likelihood becomes zero ($P(i^*|m) = 0$), the auctioneer sets the overload indicator flag to one, i.e., $O_{i^*} = 1$, and adds this information to the acknowledgement message F_j^3. The overload indicator flag helps to determine which RRH is overloaded with user access requests.

- Posterior Probability: Posterior probability refers to the selection probability of BBU m given by RRH i^*. According to Bayes' rule, we formulate the posterior probability of $P(m|i^*)$ as follows:

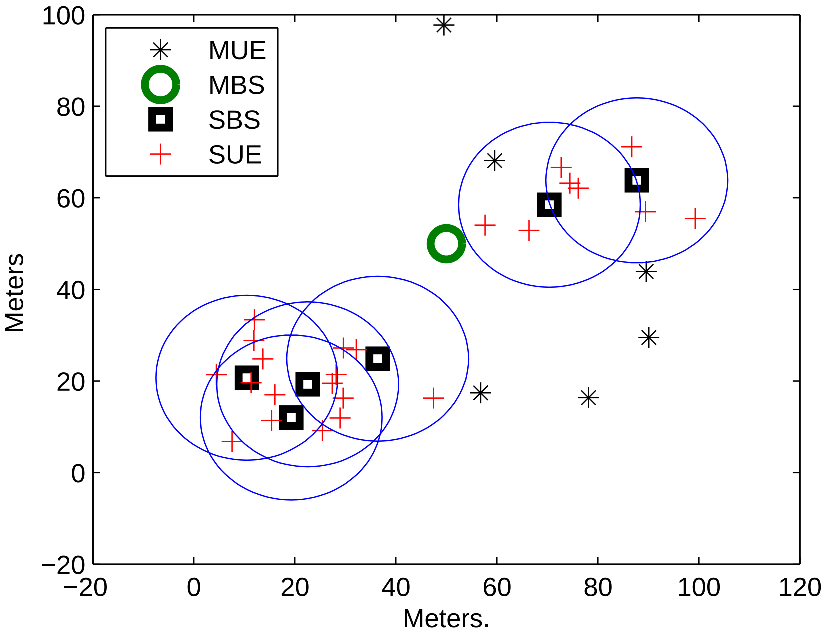


Figure 5.6: Simulation environment.

$$P(m|i^*) = \frac{P(i^*|m)P(m)}{p(x)}, \tag{5.37}$$

where $p(x)$ is $p(x) = \sum_{m \in M} P(i^*|m)P(m)$.

- BBU Association and VM Allocation: The auctioneer selects BBU m^* for RRH i^* according to their maximum a posterior probability values; that is, $m^* = \arg\max P(m|i^*)$, and assigns one VM for the RRH i^* to process baseband operation. It updates the BBU-RRH association parameter $b_{i^*,m^*} = 1$ and the feasibility of BBU server in next time slot (e.g., $q_{m^*}(t+1) = q_{m^*}(t) - 1$.
- The auctioneer adds the m^* and O_{i^*} information to the acknowledgement message F_j^3 and sends the messages to the corresponding RRH.

Step 5: If the RRH finds that the overload indicator flag is set, it repeats the step 2, otherwise sends an acknowledgement message to the users. The RRH updates the transmission alignment and sends the allocated RB list and corresponding new SINR threshold value to the MBS.

Table 5.3: Simulation parameters

Parameters	**Values**
Total no. of small cells	$6 - 10$
Total no. SUEs	$10 - 100$
Total no. of RB	50
RB bandwidth	180 kHz
System bandwidth	10 MHz
Radius of small cell	10 m
Minimum data rate requirements	50-140 kbps
Number of MUEs	$10 - 20$
Transmission power of RRH	30 dBm
Path-loss exponent	4
Noise power spectrum density	-144 dBm/Hz

5.7 Simulation Results

In this section, performance of the proposed DS-ADRA is investigated. In the simulation model, as shown in Fig. 5.6, we consider a 120 m×100 m area, where one macro base station is underlaid by 6 to 10 small cell base stations. The locations of RRHs, SUEs and MUEs are modeled using spatial Poisson point process (PPP) with predefined intensity values. The settings for the simulation parameters are shown in Table 5.3. The simulations are averaged over 100 trails. The minimum data rate requirement is considered as the arrival rate of a user. According to the service rate of the entire C-RAN system, we define Jain's fairness index and system efficiency, respectively, as

$$J = \frac{\left(\sum_{i=1}^{\mathcal{N}} \mu_i^{RRH}\right)^2}{\mathcal{N} \sum_{i=1}^{\mathcal{N}} (\mu_i^{RRH})^2},$$

and

$$E = \sum_{i=1}^{\mathcal{N}} \mu_i^{RRH}.$$

The performance of our proposed method is evaluated in terms of mean response time, system efficiency, sum data rate and Jain's fairness index for the entire C-RAN system. We also compare our proposed method with the centralized first-come-first-service (FCFS) [63] and the distributed

SINR auction based RB allocation [47] methods. The centralized FCFS is the conventional method where each user selects the nearest base station depending on the relative signal strength. The base station allocates RBs to the users based on the SINR threshold value and first-come-first-service basis. On the other hand, the working procedure of SINR auction method is as follows:

- Step 1: The users work as bidders and all the base stations work as an auctioneer. At first, for all unallocated RBs, each user proposes its bid $SINR_{RB_1}^{SUE} \geq SINR_{RB_2}^{SUE}$ where $RB_1 \neq RB_2$.
- Step 2: The base station assigns RB to the users according to these bids.
- Step 3: Each user cancel their bid to the allocated RBs, and repeat step 1 and 2 until all RBs are allocated.

However, we modified the Step 1 and 3 to make it compatible for comparison to our method, as follows:

- Step 1: For all unallocated RBs, each user proposes its bid $SINR_{RB_1}^{SUE} \geq SINR_{RB_2}^{SUE}$ where $SINR_{RB_1}^{SUE} \geq \gamma^{th}$, $SINR_{RB_2}^{SUE} \geq \gamma^{th}$ and $RB_1 \neq RB_2$.
- Step 3: Each user cancel their bid to the allocated RBs, and repeat step 1 and 2 until the user data rate requirement is satisfied.

We investigate the mean response time performance of DS-ADRA method with respect to mean arrival rate and different percentages of bandwidth utilization as shown in Fig. 5.7. In the two-tier C-RAN system, we consider that SUE utilizes the RBs of MUE in an underlaid approach under the constraints C2, C3 and C4 of problem **P1**. In DS-ADRA method, **Algorithm** 2 satisfies all these constraints and determines transmission alignment in terms of RB and power for SUEs. Here, the bandwidth utilization is referred to as frequency reuse ratio of C-RAN system. The system response time becomes the lowest when the bandwidth utilization is 100 percent compared to 60 and 80 percent utilization as shown in Fig. 5.7. It can also be observed that the mean response time of the C-RAN system increases with the mean arrival rate.

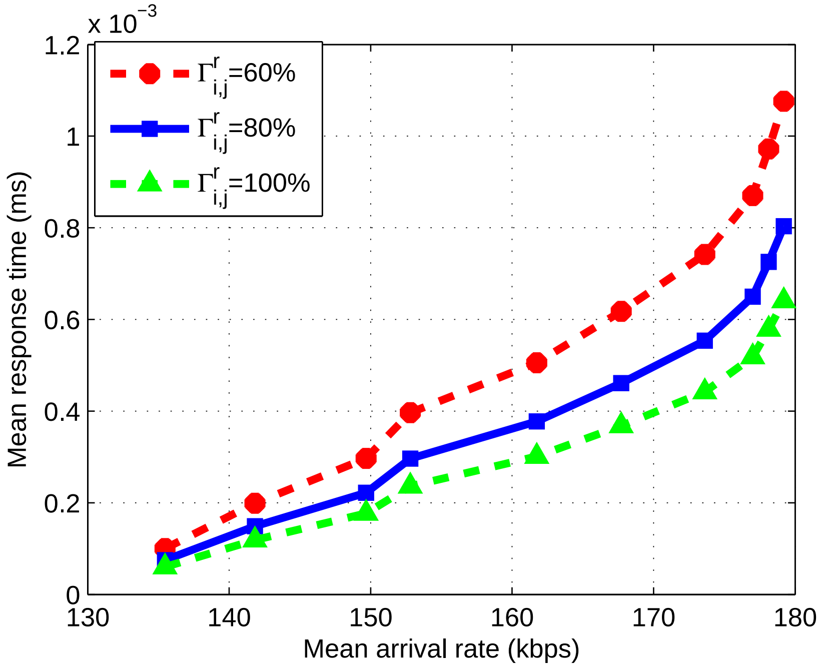


Figure 5.7: Mean response time performance of DS-ADRA method with three different percentages of bandwidth utilization.

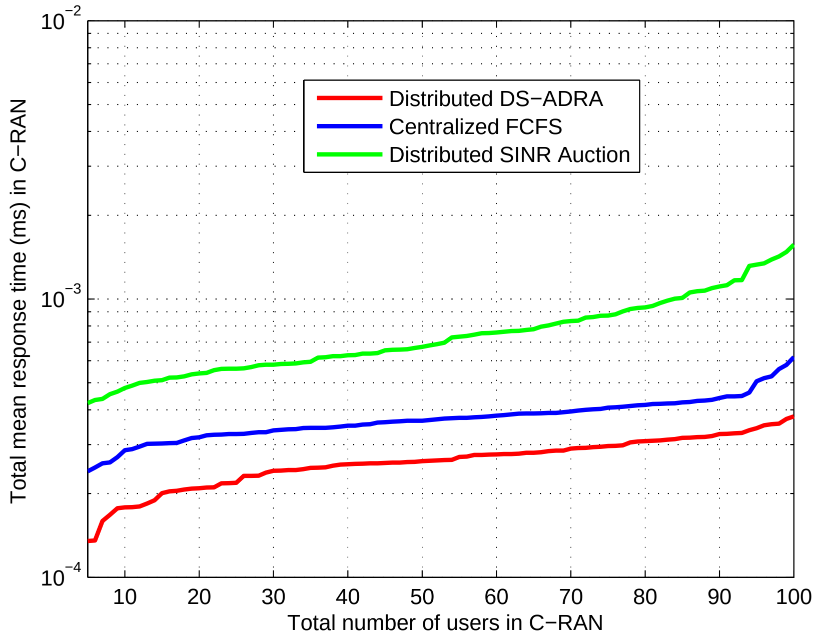


Figure 5.8: Mean response time performance among DS-ADRA, centralized-FCFS and SINR auction with different number of users when $\mathcal{N} = 6$, $\mathcal{R} = 100$ and $\gamma^{th} = 10dB$.

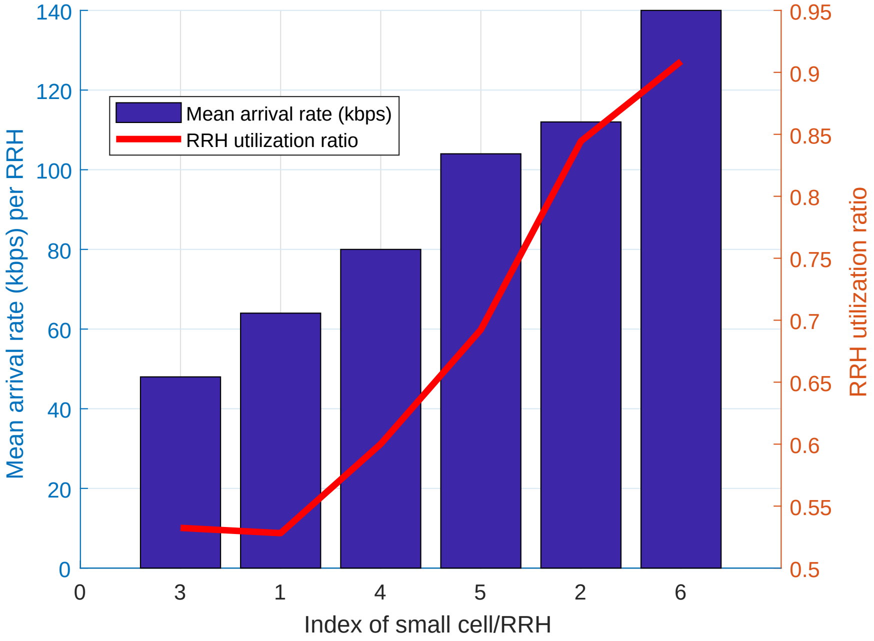


Figure 5.9: RRH utilization ratio.

Fig. 5.8 shows the mean response time performance comparison of the DS-ADRA, centralized FCFS, and SINR auction methods with respect to various number of SUEs in C-RAN. It is shown that DS-ADRA method outperforms the other methods. This due to the fact that in step 2 of **Algorithm** 2 of DS-ADRA method estimates the transmission alignment, service rate, mean benefit and mean response time for SUEs. Depending on these information, in step 3 of DS-ADRA method, each SUE selects the best RRH which gives the maximum benefit in terms of service rate and response time for data transmission. On the other hand, in the centralized FCFS method, SUEs select one of the closest proximity RRHs and the selected RRH assigns RBs to the SUEs based on the FCFS policy and SINR threshold. Also, the mean response time performance of the distributed SINR auction method worse than the others, due to the conflict choice of RBs. In this method, each user chooses RBs regardless of other users' choice. The same RBs are selected by many users, but in the end, one user becomes winner. Moreover, in the RB selection, the users only check the SINR constraint.

Fig. 5.9 shows the RRH utilization ratios and mean arrival rates in each RRH. It is evident that RRH utilization depends on the incoming requests. The simulation results in Fig. 5.9 justify that RRH utilization increases with mean arrival rate. The RRH index with 6 has the

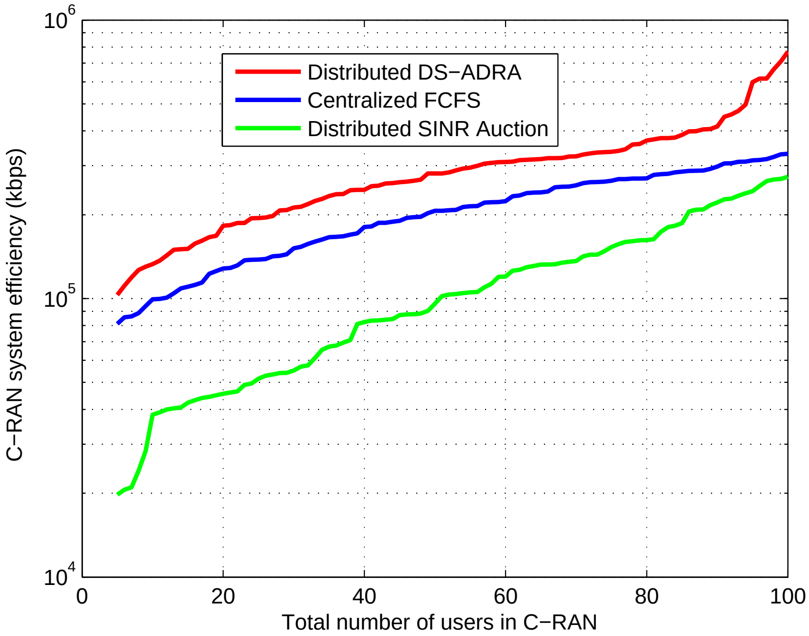


Figure 5.10: C-RAN system efficiency with different number of users when $\mathcal{N} = 6$, $\mathcal{R} = 100$ and $\gamma^{th} = 10dB$.

highest utilization since the highest arrival rate 140 kbps occurs at this base station compared to the other base stations.

Next, we investigate the C-RAN system efficiency performance of DS-ADRA, centralized FCFS, and SINR auction method with respect to different number of users as shown in Fig. 5.10. The RRHs in C-RAN utilize 100 percent bandwidth. The system efficiency of all three methods increases with increasing number of users. Among them, the DS-ADRA performs the best due to the distributed nature of RB allocation (e.g., Algorithm 1) and RRH selection (e.g., step 3 in DS-ADRA). In the DS-ADRA method, users are associated with RRHs based on the maximum utility, however, centralized method utilizes relative signal strength and closer proximity RRHs and FCFS policy for users which limits the system efficiency of C-RAN. Moreover, Fig. 5.11 shows the whole C-RAN system efficiency of DS-ADRA with respect to achievable SINR and different numbers of SUEs. In this scenario, equal SINR threshold level is considered as in **Algorithm** 2, that is 10dB. The RRHs in C-RAN are allowed to utilize 100 percent bandwidth. It is observed from the figure that as the system efficiency increases with the achievable SINR and number of SUEs. This is justifiable since the interference threshold levels that SUEs can

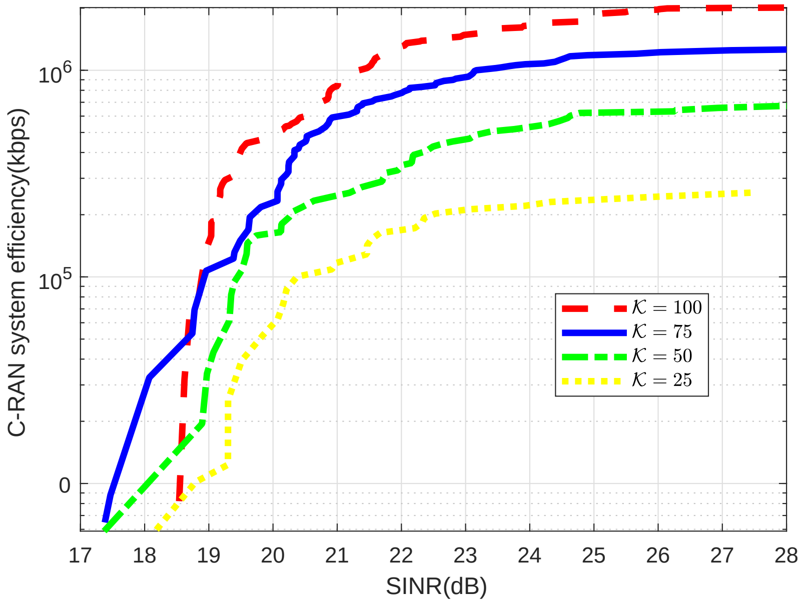


Figure 5.11: C-RAN system efficiency of DS-ADRA with respect to achievable SINR with $\Gamma^r_{ij} =$ 100% of bandwidth utilization when $\mathcal{N} = 6$, $\mathcal{R} = 100$ and $\gamma^{th} = 10dB$.

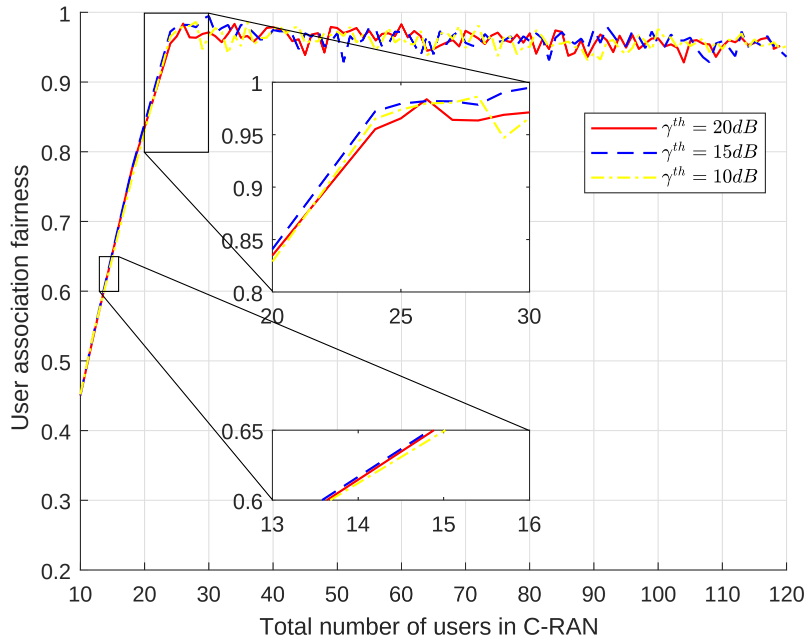


Figure 5.12: Convergence of stable allocation of Algorithm 1 shown with UA fairness considering different SINR threshold values when $\mathcal{N} = 6$, $\mathcal{R} = 100$, and $\mathcal{K} = 100$.

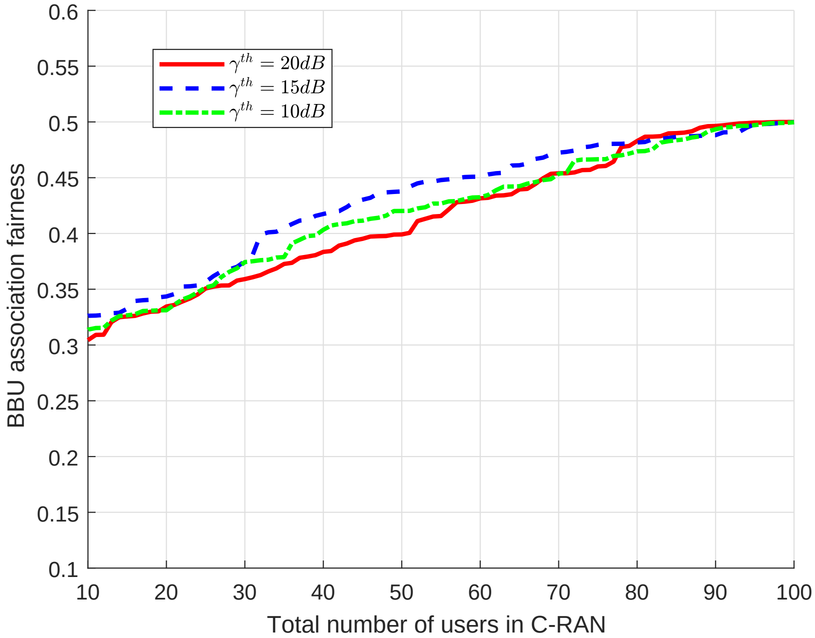


Figure 5.13: Convergence of VM allocation of Algorithm 2 considering different SINR threshold values when $\mathcal{N} = 6$, $\mathcal{R} = 100$, $\mathcal{K} = 100$, $M = 4$ and $U = 3$.

tolerate decrease as sum rate increases.

The convergence behaviors of **Algorithm** 2 and **Algorithm** 3 in DS-ADRA method are shown with fairness index in Fig. 5.12 and Fig. 5.13, respectively. According to the perspective of Jain's index, a larger J represents a fair allocation [64]. Fig. 5.12 shows that **Algorithm** 2 converges and shows fair allocation despite different threshold values. Similarly, Fig. 5.13 shows that **Algorithm** 3 finds a fair allocation of VMs for small cells within finite number of steps regardless of different SINR threshold values.

5.8 Summary

In this chapter, we proposed an auction based distributed communication and computing resource allocation method OFDM based delay aware C-RAN systems. The proposed DS-ADRA method satisfies the user association, BBU-RRH mapping, resource allocation and maximum power constraints as well as the queuing stability constraint. In addition, the DS-ADRA method associates one user with one RRH and assigns RRH to BBU pool based on the co-operative decision among users and RRH with the help of SDF C-RAN controller. In terms of mean response time and system efficiency, the proposed DS-ADRA method outperforms the centralized FCFS and SINR auction methods, due to the distributed nature of communication resource allocation (i.e., **Algorithm** 2), maximum utility based RRH selection, and posterior probability based computing resource allocation (i.e., **Algorithm** 3) being jointly executed in the DS-ADRA method. Furthermore, the proposed **Algorithm** 2 and **Algorithm** 3 ensure to converge to a stable allocation despite different arrival rates of users and SINR values of the C-RAN systems. The results show that the performance of the proposed DS-ADRA method becomes the best when the system supports 100 percent bandwidth utilization.

In the next chapter, we will focus on energy efficiency aspects of OFDM based two-tier C-RANs. First, we formulate energy efficient resource optimization problem in two tier C-RANs. We consider only communication resource allocation considering both small and macro cell users. Then, we apply Dinkelbach theorem and iterative resource allocation solution to solve the optimization problem.

Chapter 6

Energy Efficient Resource Allocation in Cloud-RAN

6.1 Introduction

In the previous chapter, we considered a queueing model in C-RANs to improve the performance of mean response time and data rate of the C-RANs in terms of both communication and computing resource allocation. In this chapter, we consider energy efficiency aspects of communication resource allocation in C-RANs. The OFDMA supported LTE and LTE-A networks utilize orthogonal resources for the users to mitigate intra-cell interference. Similarly, OFDMA supported H-CRANs utilize orthogonal resources for the small cell users [65] [66]. Considering the interference issues, authors in [65] have proposed an auction based distributed resource allocation with the aim to improve the data rate of small cell users. On the other hand, authors in [66] consider only small cell users to maximize energy efficiency H-CRAN via BBU-offloading. Due to the limited number of orthogonal resources, authors in [27] consider an underlaid approach of orthogonal resource sharing both for macro cell and small cell users with the aim to maximize the sum of the tolerable interference levels. Similarly, in this chapter, we consider an underlaid approach of communication resource allocation for both small cell and macro cell users to improve energy efficiency in two-tier C-RAN.

In this chapter, we apply the two-step iterative algorithm to allocate resource block and power

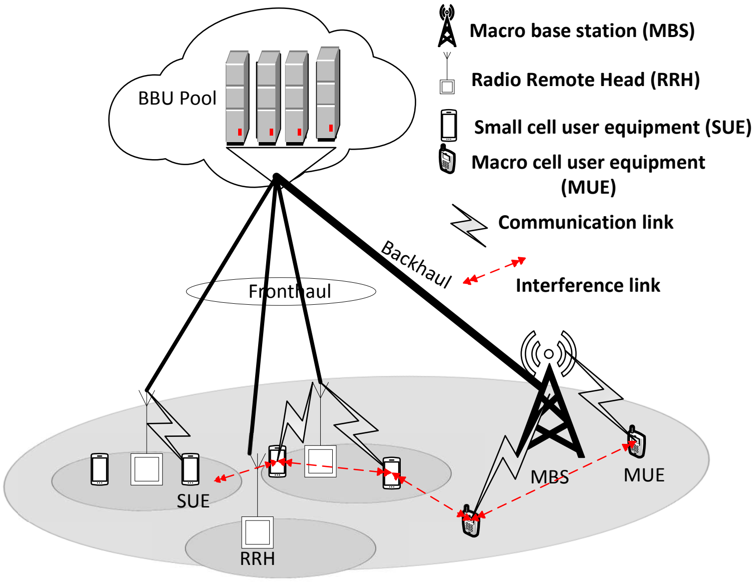


Figure 6.1: OFDMA supported two-tier C-RAN with interference links.

to the OFDMA supported C-RANs. The main contributions of this chapter are as follows:

- We consider downlink communication resources (i.e., RB and power) allocation in C-RANs considering both small cell and macro cell users. In the resource allocation (RA) optimization problem, we consider maximum power, RB allocation, minimum data date requirements constraints of macro cell users with the aim to maximize EE.
- To solve the optimization problem, we relax the original problem by replacing non-convex constraints with convex constraints considering time-sharing approach of resource allocation and transform the fractional objective function with the subtractive linear form.
- The two-step iterative algorithm is implemented with the Dinkelbach theorem to optimize resources by allocating same radio resource both small cell and macro cell users in an underlaid approach.
- The effectiveness of the proposed method is verified through Monte Carlo simulation.

6.2 System model and Assumptions

We consider an OFDMA based two-tier H-CRAN, as shown in Fig. 6.1, where $\mathcal{S}$ number of small cells that have $\mathcal{S}$ number of RRHs, indexed by $j = \{1, 2, ..., \mathcal{S}\}$, are covered by a single macro cell B with an underlaid approach. For simplicity, we assume that the system supports a total number of $\mathcal{R}$ resource blocks (RB), indexed by $r = \{1, 2, 3, ..., \mathcal{R}\}$ jointly assigned by the BBU pool to all small cell and macro cell users. Each small cell user equipment (SUE) and macro cell user equipment (MUE) are equipped with one antenna, and each small cell has ϕ antennas. The system supports a total number of users $\mathcal{K} = \mathcal{L} \bigcup \mathcal{M}$, where SUEs are indexed by $i = \{1, 2, 3, ..., \mathcal{L}\}$ and MUEs are denoted by $m = \{1, 2, 3, ..., \mathcal{M}\}$. To improve the SE of H-CRAN, we assume that SUEs and MUEs share the same set of resource blocks in an underlaid approach. For a summary of symbols and parameters, a list is provided in Table 6.1. Let $\alpha^r_{i,j}$ and $\beta^r_{m,B}$ be the binary decision variables for RB allocation, updated as follows:

$$\alpha^r_{i,j} = \begin{cases} 1, & \text{if RB } r \text{ is assigned to SUE } i \text{ on RRH } j, \\ 0, & \text{otherwise.} \end{cases}$$

$$\beta^r_{m,B} = \begin{cases} 1, & \text{if RB } r \text{ is assigned to MUE } m \text{ on MBS } B, \\ 0, & \text{otherwise.} \end{cases}$$

The channel gain from RRH j to SUE i on RB r is denoted as $h^r_{i,j} \in \mathcal{C}^{\phi \times 1}$. The allocating power from RRH j to user i on RB r is denoted as $P^r_{i,j} \in (0, P^{max}_j]$, where P^{max}_j denotes the maximum power of RRH j. The transmitted data symbol for user i is denoted by x_i, where $\mathrm{E}[|x_i|^2] = 1$. The received signal of SUE i on RB r can be written as:

Table 6.1: List of symbols

Symbol	**Description**
S	Number of small cells/RRHs
j	Indexing for RRH
$\mathcal{L}$	Number of SUEs
i	Indexing for SUE
$\mathcal{M}$	Number of MUEs
m	Indexing for MUE
$\mathcal{K}$	Total number of H-CRAN users
$\mathcal{R}$	Number of RBs
r	Indexing for RB
$\alpha_{i,j}^{r}$	RB allocation variable for r^{th} RB to j^{th} RRH for SUE i
$\boldsymbol{\alpha} = [\alpha_{i,j}^{r}]_{\mathcal{L}\times\mathcal{S}\times\mathcal{R}}$	Matrix of RB allocation for all SUEs
$\beta_{i,j}^{r}$	RB allocation variable for r^{th} RB to B^{th} MBS for MUE m
$\boldsymbol{\beta} = [\beta_{m,B}^{r}]_{\mathcal{M}\times\mathcal{R}}$	Matrix of RB allocation for all MUEs
$h_{i,j}^{r}$	Channel gain from j^{th} RRH to i^{th} SUE on r^{th} RB
$P_{i,j}^{r}$	Power allocation from j^{th} RRH to SUE i on r^{th} RB
$\boldsymbol{P_{\mathcal{L}}} = [P_{i,j}^{r}]_{\mathcal{L}\times\mathcal{S}\times\mathcal{R}}$	Matrix of power allocation for all SUEs
$P_{m,B}^{r}$	Power allocation from MBS to MUE m on r^{th} RB
$\boldsymbol{P_{\mathcal{M}}} = [P_{m,B}^{r}]_{\mathcal{M}\times\mathcal{R}}$	Matrix of power allocation for all MUEs
P_{j}^{max}	Maximum power budget of j^{th} RRH
P_{B}^{max}	Maximum power budget of MBS
$\gamma_{i,j}^{r}$	SINR of i^{th} SUE connected to j^{th} RRH on r^{th} RB
$\gamma_{m,B}^{r}$	SINR of m^{th} MUE connected to B^{th} MBS on r^{th} RB
ω_0	Additive white Gaussian noise term
σ^2	Noise power

$$Y_i^r = \underbrace{\left(h_{i,j}^r \alpha_{i,j}^r P_{i,j}^r\right) x_i}_{desired\ signal} + \underbrace{\sum_{k=1,k\neq i}^{\mathcal{L}} \left(h_{i,j}^r \alpha_{k,j}^r P_{k,j}^r\right) x_k}_{intra-cell\ interference\ signal} \tag{6.1}$$
$$+ \underbrace{\sum_{l=1,l\neq i}^{\mathcal{L}} \left(\sum_{n=1,n\neq j}^{\mathcal{S}} h_{i,n}^r \alpha_{l,n}^r P_{l,n}^r \right) x_l}_{inter-cell\ interference\ signal\ from\ other\ small\ cells} + \underbrace{\sum_{m=1}^{\mathcal{M}} h_{i,B}^r \beta_{m,B}^r P_{m,B}^r x_m}_{inter-cell\ interference\ signal\ from\ macro\ cell} + \omega_0,$$

where the first term on the right hand side is the desired signal for user i from RRH j on RB r, the second and third terms denote the interference signal coming from other active SUEs of the same and other cells using same RB r, respectively. The fourth term determines the interference signal from all active MUEs using the same RB r. The term ω_o represents the additive white Gaussian noise, where the power density of ω_0 is σ^2. The signal-to-interference-noise ratio (SINR) achieved by SUE i, attached to RRH j on RB r can be written as

$$\gamma_{i,j}^r = \frac{|h_{i,j}^r \alpha_{i,j}^r P_{i,j}^r|^2}{I_{i,j}^r + \sigma^2}. \tag{6.2}$$

Similar to [27], [67], $I_{i,j}^r$ represents the aggregated interference power, defined as:

$$I_{i,j}^r = \sum_{k=1,k\neq i}^{\mathcal{L}} h_{i,j}^r \alpha_{k,j}^r P_{k,j}^r + \sum_{l=1,l\neq i}^{\mathcal{L}} \sum_{n=1,n\neq j}^{\mathcal{S}} h_{i,n}^r \alpha_{l,n}^r P_{l,n}^r \tag{6.3}$$
$$+ \sum_{m=1}^{\mathcal{M}} h_{i,B}^r \beta_{m,B}^r P_{m,B}^r.$$

Similarly, the SINR achieved by MUE m, attached to MBS B on r can be expressed by

$$\gamma_{m,B}^r = \frac{|h_{m,B}^r \beta_{m,B}^r P_{m,B}^r|^2}{I_{m,B}^r + \sigma^2}, \tag{6.4}$$

where $I_{m,B}^r$ represents the aggregated interference power of macro cell user m observed in RB r, which can be expressed as

$$I^r_{m,B} = \underbrace{\sum_{k=1,k\neq m}^{\mathcal{M}} h^r_{m,B}\beta^r_{k,B}P^r_{k,B}}_{intra-cell\ interference\ from\ other\ macrocell\ users} + \underbrace{\sum_{l=1,l\neq m}^{\mathcal{L}} \sum_{j=1}^{\mathcal{S}} h^r_{m,j}\alpha^r_{l,j}P^r_{l,j}}_{inter-cell\ interference\ from\ all\ small\ cells}. \tag{6.5}$$

6.3 Energy Efficient Resource Allocation in H-CRANs

The objective of resource allocation in a H-CRAN is to maximize the EE (i.e., η) in terms of available spectrum and power. The EE can be measured by the total achievable data rate divided by total allocated power (bits/J), written as:

$$\begin{aligned}\eta &= \frac{R_T(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P_{\mathcal{L}}},\boldsymbol{P_{\mathcal{M}}})}{P_T(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P_{\mathcal{L}}},\boldsymbol{P_{\mathcal{M}}})}\\ &= \frac{R_{T_s}(\boldsymbol{\alpha},\boldsymbol{P_{\mathcal{L}}}) + R_{T_m}(\boldsymbol{\beta},\boldsymbol{P_{\mathcal{M}}})}{P_{T_s}(\boldsymbol{\alpha},\boldsymbol{P_{\mathcal{L}}}) + P_{T_B}(\boldsymbol{\beta},\boldsymbol{P_{\mathcal{M}}})},\end{aligned}$$

where $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P_{\mathcal{L}}}$, and $\boldsymbol{P_{\mathcal{M}}}$ are the RB association and power allocation policies to all SUEs and MUEs, respectively. According to the Shanon formula, the achievable data rate by SUE i from SBS j on RB r is $R^r_{i,j} = \triangle f \log_2(1+\gamma^r_{i,j})$, where $\triangle f$ denotes bandwidth allocated to each RB, and the total data rate of all SUEs can be expressed as

$$R_{T_s}(\boldsymbol{\alpha},\boldsymbol{P_{\mathcal{L}}}) = \sum_{i=1}^{\mathcal{L}}\sum_{j=1}^{\mathcal{S}}\sum_{r=1}^{\mathcal{R}} \alpha^r_{i,j}R^r_{i,j}.$$

Similarly, the achievable data rate of all macro cell users can be expressed as:

$$R_{T_m}(\boldsymbol{\beta},\boldsymbol{P_{\mathcal{M}}}) = \sum_{m=1}^{\mathcal{M}}\sum_{r=1}^{\mathcal{R}} \beta^r_{m,B}R^r_{m,B}.$$

Total allocated powers of small and macro cell users are denoted by

$$P_{T_s}(\boldsymbol{\alpha}, \boldsymbol{P_{\mathcal{L}}}) = \underbrace{\sum_{i=1}^{\mathcal{L}} \sum_{j=1}^{\mathcal{S}} \sum_{r=1}^{\mathcal{R}} \alpha_{i,j}^{r} P_{i,j}^{r}}_{dynamic} + \underbrace{\sum_{j=1}^{\mathcal{S}} P_j^s}_{static},$$

and

$$P_{T_B}(\boldsymbol{\beta}, \boldsymbol{P_{\mathcal{M}}}) = \sum_{m=1}^{\mathcal{M}} \sum_{r=1}^{\mathcal{R}} \beta_{m,B}^{r} P_{m,B}^{r},$$

respectively. Similar to [68], we assume that each small cell has dynamic and static power factors. Dynamic power depends on resource allocation, whereas alternative current (AC) circuit power is regarded as static power.

We define the following constraint sets in order to maximize EE in a H-CRAN. In the constraint sets, we consider that all users are associated with either RRHs or MBS, utilizing RBs and power to transmit data without causing interference to each other.

- **RB association constraint:** In constraint **C1**, each RB is associated with only one SUE. Similarly, constraint **C2** ensures that each RB is associated with only one MUE. However, for spectrum efficiency we assume that an SUE can utilize MUE's RB in an underlaid approach, i.e.,

$$\begin{aligned} &\text{C1:} \quad \sum_{j=1}^{\mathcal{S}} \alpha_{i,j}^{r} \leq 1, \quad \forall i \in \mathcal{L}, \alpha_{i,j}^{r} \in \{0,1\}, \\ &\text{C2:} \quad \sum_{m=1}^{\mathcal{M}} \beta_{m,B}^{r} \leq 1, \quad \beta_{m,B}^{r} \in \{0,1\}. \end{aligned}$$

- **Interference mitigation constraint:** Underlaid approach of RB allocation increases intra- and inter-cell interference levels in H-CRAN. To mitigate intra-cell interference, we incorporate constraints **C3** and **C4** as

$$\begin{aligned} &\text{C3:} \quad \alpha_{u,j}^{r} + \alpha_{v,j}^{r} \leq 1, \quad \forall r \in \mathcal{R}, \quad \forall (u,v) \in \mathcal{L}, \forall j \in \mathcal{S}, \\ &\text{C4:} \quad \beta_{u,B}^{r} + \beta_{v,B}^{r} \leq 1, \quad \forall (u,v) \in \mathcal{M}, \quad \forall r \in \mathcal{R}. \end{aligned}$$

The constraint **C3** ensures that users in the same RRH use separate RBs. Similarly, constraint **C4** ensures all MUEs use different RBs to avoid intra-cell interference. For inter-cell interference, we assume that the centralized BBU pool manages the interference level among RRHs.

- **Maximum power constraint:** Constraints **C5** and **C6** ensure the maximum power budget of RRH and MBS, respectively, i.e.,

$$\begin{aligned} &\text{C5:} \quad \sum_{i=1}^{\mathcal{L}} \sum_{r=1}^{\mathcal{R}} \alpha_{i,j}^{r} P_{i,j}^{r} \leq P_{j}^{max}, \quad \forall j \in \mathcal{S}, \\ &\text{C6:} \quad \sum_{m=1}^{\mathcal{M}} \sum_{r=1}^{\mathcal{R}} \beta_{m,B}^{r} P_{m,B}^{r} \leq P_{B}^{max}. \end{aligned}$$

- **Minimum data rate constraint:** As the SUE shares the same RB with the MUE, the minimum rate constraint ensures the QoS of MUE in **C7** as

$$\text{C7:} \quad \sum_{r=1}^{\mathcal{R}} \beta_{i,j}^{r} R_{m,B}^{r} \geq R_{min}, \quad \forall m \in \mathcal{M},$$

  where R_{min} is the minimum data rate threshold value for MUE.

- **Fronthaul capacity constraint:** Constraint **C8** ensures the maximum fronthaul capacity of each RRH, where $\digamma^{max}$ refers to the maximum limit of baseband signals transmitted on the fronthaul link of each RRH, and is given as

$$\text{C8:} \quad \sum_{i=1}^{\mathcal{L}} \alpha_{i,j} \leq \digamma^{max}, \quad \forall j \in \mathcal{S}.$$

The mathematical formulation of EE resource allocation problem in the H-CRAN can be described as follows:

$$\begin{aligned} \text{P1:} \quad &\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P_{\mathcal{L}}},\boldsymbol{P_{\mathcal{M}}}} \; \eta \\ &\text{subject to: C1-C8} \end{aligned} \tag{6.6}$$

In (7.12), the objective is to maximize the EE of the H-CRAN system by allocating the same radio resources to both small cell and macro cell users in an underlaid approach. There are four optimization parameters considered in this problem: i) RB allocation for both small cell and macro cell users (i.e., $\alpha^r_{i,j} \in \{0,1\}$, $\beta^r_{m,B} \in \{0,1\}$ and ii) power allocation for both small cell and macro cell users (i.e., $P^r_{i,j}$ and $P^r_{m,B}$)

Corollary 1. *The objective function in (6.6) and the constraints **C1** and **C2** turn the problem **P1** into a mixed integer non-convex fractional programming problem. The optimization problem **P1** is computationally intractable and is an NP-hard problem* [69], [36].

6.4 Problem transformation

We relax the problem **P1** by replacing non-convex constraints with the convex constraints and transform the fractional objective function of **P1** to the subtractive linear form. First, we relax the constraints **C1** and **C2** by assuming time sharing approach of RB allocation [70], i.e., $0 \leq \alpha^r_{i,j} \leq 1$ and $0 \leq \beta^r_{i,j} \leq 1$. We introduce two new variables $\Gamma^r_{i,j} = \alpha^r_{i,j} \in (0,1]$ and $\Lambda^r_{i,j} = \beta^r_{i,j} \in (0,1]$. $\Gamma^r_{i,j}$ and $\Lambda^r_{i,j}$ represent time sharing factors of resource blocks of SUE and MUE, respectively. It denotes the portion of time the RB r is allocated to the user i and RRH j link. Let $\mathcal{P}\mathcal{L}^r_{i,j} = \Gamma^r_{i,j} \times P^r_{i,j}$ and $\mathcal{P}\mathcal{M}^r_{i,j} = \Lambda^r_{i,j} \times P^r_{m,B}$, where $\mathcal{P}\mathcal{L}^r_{i,j}$ denotes the actual transmit power of SUE i on RB r and $\mathcal{P}\mathcal{M}^r_{i,j}$ is the allocated power of MUE m on RB r. Next, according to [69] and [**?**], we can transform the fractional objective function of **P1** to the subtractive linear form and solve it according to the Dinkelbach theorem [71]. The relaxed problem can be represented by:

$$\text{P2:} \max_{\mathbf{\Gamma},\mathbf{\Lambda},\boldsymbol{P_{\mathcal{L}}},\boldsymbol{P_{\mathcal{M}}}} R_T(\mathbf{\Gamma},\mathbf{\Lambda},\boldsymbol{P_{\mathcal{L}}},\boldsymbol{P_{\mathcal{M}}}) - q^* P_T(\mathbf{\Gamma},\mathbf{\Lambda},\boldsymbol{P_{\mathcal{L}}},\boldsymbol{P_{\mathcal{M}}}) \tag{6.7}$$

subject to:

$$\vec{\text{C1}}\text{:} \quad \sum_{j=1}^{\mathcal{S}} \Gamma^r_{i,j} \leq 1, \quad \forall i \in \mathcal{L}, \Gamma^r_{i,j} \in (0,1],$$

$$\vec{\text{C2}}\text{:} \quad \sum_{m=1}^{\mathcal{M}} \Lambda^r_{m,B} \leq 1, \quad \Lambda^r_{m,B} \in (0,1],$$

$$
\begin{aligned}
&\text{C}\vec{3}\text{:} \quad \forall (u,v) \in \mathcal{L}, \quad \Gamma^{r}_{u,j} + \Gamma^{r}_{v,j} \leq 1, \quad \forall r \in \mathcal{R}, \forall j \in \mathcal{S}, \\
&\text{C}\vec{4}\text{:} \quad \forall (u,v) \in \mathcal{M}, \quad \Lambda^{r}_{u,B} + \Lambda^{r}_{v,B} \leq 1, \quad \forall r \in \mathcal{R}, \\
&\text{C}\vec{5}\text{:} \quad \sum_{i=1}^{\mathcal{L}} \sum_{r=1}^{\mathcal{R}} \mathcal{P}_{\mathcal{L}}{}^{r}_{i,j} \leq P^{max}_{j}, \quad \forall j \in \mathcal{S}, \\
&\text{C}\vec{6}\text{:} \quad \sum_{m=1}^{\mathcal{M}} \sum_{r=1}^{\mathcal{R}} \mathcal{P}_{\mathcal{M}}{}^{r}_{m,B} \leq P^{max}_{B}, \\
&\text{C}\vec{7}\text{:} \quad \sum_{r=1}^{\mathcal{R}} \Lambda^{r}_{i,j} R^{r}_{m,B} \geq R_{min}, \quad \forall m \in \mathcal{M}, \\
&\text{C}\vec{8}\text{:} \quad \sum_{i=1}^{\mathcal{L}} \Gamma_{i,j} \leq F^{max}, \quad \forall j \in \mathcal{S},
\end{aligned}
$$

where q^* is the global optimal EE, i.e.,

$$
q^* = \frac{R_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}})} = \max_{\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}} \frac{R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}.
$$

Corollary 2. *The objective function in (6.7) is concave, the constraint (*$\boldsymbol{\text{C}\vec{7}}$*) is convex, and the remaining constraints in (*$\boldsymbol{\text{C}\vec{1}}$*)-(*$\boldsymbol{\text{C}\vec{6}}$*), (*$\boldsymbol{\text{C}\vec{8}}$*) are affine. Therefore, the optimization problem* ***P2*** *is convex* [36].

The relaxed problem **P2** satisfies the time-sharing approach of RB allocation. In [72], the authors showed that the duality gap is negligible when time-sharing condition is satisfied. Therefore, when the number of RBs is sufficiently large, the solution of the relaxed problem is asymptotically optimal.

Theorem 1: *For* $R_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}}) \geq 0$ *and* $P_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}}) > 0$*,* q *can reach its optimal value if and only if*

$$
\max_{\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}} R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) - q^* P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) = 0
$$

Proof: Let the feasible solution of (6.7) be denoted as $(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})$. Since $(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}})$

is the optimal solution to the problem,

$$R_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}}) - q^* P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) \geq \\ R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) - q^* P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}),$$

and $R_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}}) - q^* P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) = 0$, and $R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) - q^* P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) \leq 0$, where $P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})$ is the H-CRAN power consumption which is larger than zero. Therefore,

$$\frac{R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})} \leq q^*.$$

Hence,

$$\frac{R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})} \leq \frac{R_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}})}.$$

Thus, $(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*, \boldsymbol{P}^*_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}^*_{\boldsymbol{\mathcal{M}}})$ maximizes

$$\frac{R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}$$

while satisfying all the constraints in the relaxed optimization problem **P2**. □

6.5 Iterative Algorithm for Resource Allocation

The iterative algorithms to obtain q^* and resource allocation policy (i.e., $(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})$) are given in **Algorithm 4** and **Algorithm 5**, respectively. According to **Theorem 1** for a given q and resource allocation policy $(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})$, we can iteratively obtain the solution of q^*. The process is repeated until the EE is maximized, i.e., $R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) - q(.) P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}}) \leq \epsilon$, where $q(.) = \frac{R_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}{P_T(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{P}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}_{\boldsymbol{\mathcal{M}}})}$ is the EE at the iteration (.) and ϵ is a small-valued maximum tolerance level.

The proposed iterative resource allocation policy is summarized in Fig. 6.2 and given in **Algorithm 5**. In the H-CRAN, we assume that the resource allocation policy is centrally coordinated by the software defined controller inside the BBU pool. At the initial stage of resource allocation, the users are classified as an SUE or MUE by the location aware user

Algorithm 4: Iterative algorithm for obtaining q^*

```
Initialization: $t = 0$ and $q(t) = 0$;
Set maximum number of iterations $t_{max}$ and maximum tolerance level $\epsilon$;
Iteration:;
while not converged OR $t! = t_{max}$ do
    Solve the problem P2 using Algorithm 2 and obtain resource allocation policies $(\boldsymbol{\Gamma}', \boldsymbol{\Lambda}', \boldsymbol{P}'_{\mathcal{L}}, \boldsymbol{P}'_{\mathcal{M}})$;
    if $R_T(\boldsymbol{\Gamma}', \boldsymbol{\Lambda}', \boldsymbol{P}'_{\mathcal{L}}, \boldsymbol{P}'_{\mathcal{M}}) - q(t)P_T(\boldsymbol{\Gamma}', \boldsymbol{\Lambda}', \boldsymbol{P}'_{\mathcal{L}}, \boldsymbol{P}'_{\mathcal{M}}) \leq \epsilon$ then
        converge=true;
        return $q^*$ by Theorem 1;
        Set $(\boldsymbol{\Gamma}^*, \boldsymbol{\Lambda}^*, \boldsymbol{P}^*_{\mathcal{L}}, \boldsymbol{P}^*_{\mathcal{M}}) = (\boldsymbol{\Gamma}', \boldsymbol{\Lambda}', \boldsymbol{P}'_{\mathcal{L}}, \boldsymbol{P}'_{\mathcal{M}})$
    else
        converge=false;
        return $q(t+1) = \frac{R_T(\boldsymbol{\Gamma}^*, \boldsymbol{\Lambda}^*, \boldsymbol{P}^*_{\mathcal{L}}, \boldsymbol{P}^*_{\mathcal{M}})}{P_T(\boldsymbol{\Gamma}^*, \boldsymbol{\Lambda}^*, \boldsymbol{P}^*_{\mathcal{L}}, \boldsymbol{P}^*_{\mathcal{M}})}$
```

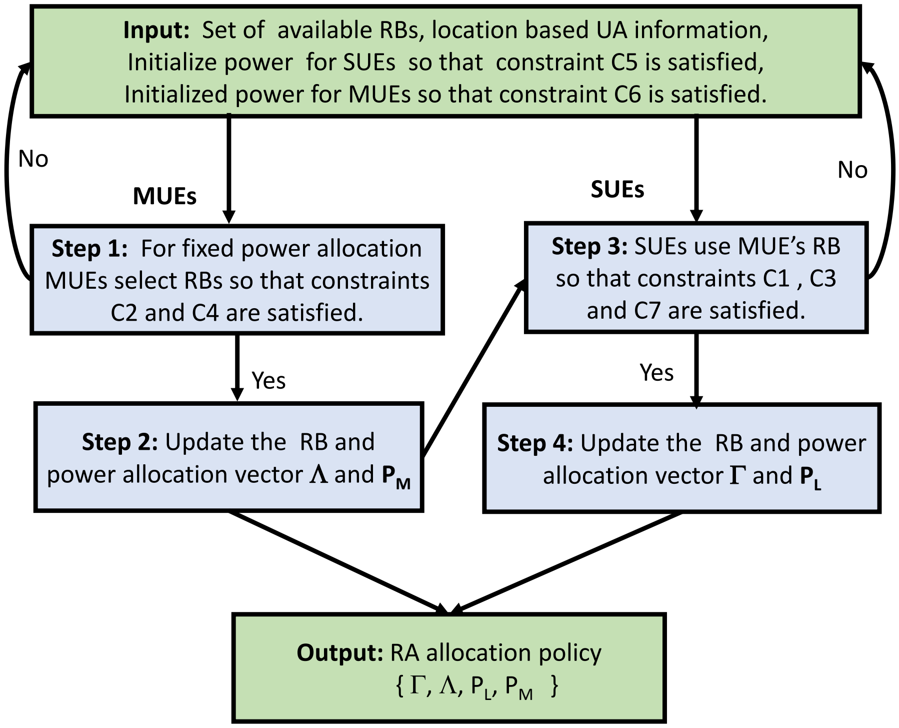

Input: Set of available RBs, location based UA information,
Initialize power for SUEs so that constraint C5 is satisfied,
Initialized power for MUEs so that constraint C6 is satisfied.
No
MUEs
SUEs
No
Step 1: For fixed power allocation MUEs select RBs so that constraints C2 and C4 are satisfied.
Step 3: SUEs use MUE's RB so that constraints C1 , C3 and C7 are satisfied.
Yes
Yes
Step 2: Update the RB and power allocation vector Λ and $\mathbf{P_M}$
Step 4: Update the RB and power allocation vector Γ and $\mathbf{P_L}$
Output: RA allocation policy
$\{\Gamma, \Lambda, P_L, P_M\}$

Figure 6.2: Resource allocation policy.

association method [73]. Depending on the number of users that are associated with macro and small cell base stations, the total power is equally allocated to users such that constraints **C5** and **C6** are satisfied. The key steps of the proposed solution are as follows:

- $\{\boldsymbol{\Lambda}$ and $\boldsymbol{P_{\mathcal{M}}}\}$-update: In the H-CRAN, we assume that SUEs are utilizing the same RBs with MUEs, so at first all MUEs in H-CRAN iteratively update $\boldsymbol{\Lambda}$ when RB association (**C2**) and intra-cell interference (**C4**) constraints are satisfied.

  The power allocation ($\boldsymbol{P_{\mathcal{M}}}$) for all MUEs is initialized during the user association phase. Step 1 in Fig. 6.2 and lines 5-11 in **Algorithm 5** show this update procedure.

- $\{\boldsymbol{\Gamma}$ and $\boldsymbol{P_{\mathcal{L}}}\}$-update: After getting the RB list from MUEs, the resource allocation method determines the RBs for SUEs verifying the constraints **C1**, **C3** and **C7**, which is shown as Step 4 in Fig. 6.2 and line 14-23 in **Algorithm 5**.

Algorithm 5: Iterative algorithm for obtaining resource allocation $\{\boldsymbol{\Gamma}', \boldsymbol{\Lambda}', \boldsymbol{P}'_{\boldsymbol{\mathcal{L}}}, \boldsymbol{P}'_{\boldsymbol{\mathcal{M}}}\}$

```
1  Initialization: Initialize the set of available RB $r = \{1, 2, 3, ...\mathcal{R}\}$;
2  Each $j^{th}$ RRH counts the total number of user requests and sets fixed power level for each user so that
   $\sum_{i=1}^{\mathcal{L}} \mathcal{P}_{\mathcal{L}i,j} \leq P_j^{max}$ is satisfied;
3  Initialized fixed power level for MUEs so that $\sum_{m=1}^{\mathcal{M}} \mathcal{P}_{\mathcal{M}m,B} \leq P_B^{max}$, is satisfied;
4  Iteration of MUEs:;
5  for $m \leftarrow 1$ to $\mathcal{M}$ do
6      for $r \leftarrow 1$ to $\mathcal{R}$ do
7          if $\Lambda_{m,B}^r == 0$ AND $R_{m,B}^r \geq R_{min}$ then
8              $\Lambda_{m,B}^r \leftarrow 1$;
9  return $\boldsymbol{\Lambda}'$ and $\boldsymbol{P}'_{\boldsymbol{\mathcal{M}}}$;
10 Iteration of SUEs:;
11 for $i \leftarrow 1$ to $\mathcal{L}$ do
12     for $r \leftarrow 1$ to $\mathcal{R}$ do
13         if $\Gamma_{i,j}^r == 0$ then // Check constraints C1 and C3
14             Find $m$ where $r$ is used;
15             if $\Lambda_{m,B}^r == 1$ AND $\Lambda_{i,j}^r R_{m,B}^r \geq R_{min}$ then // Check constraint C7
16                 $\Gamma_{i,j}^r \leftarrow 1$;
17 return $\boldsymbol{\Gamma}'$ and $\boldsymbol{P}'_{\boldsymbol{\mathcal{L}}}$
```

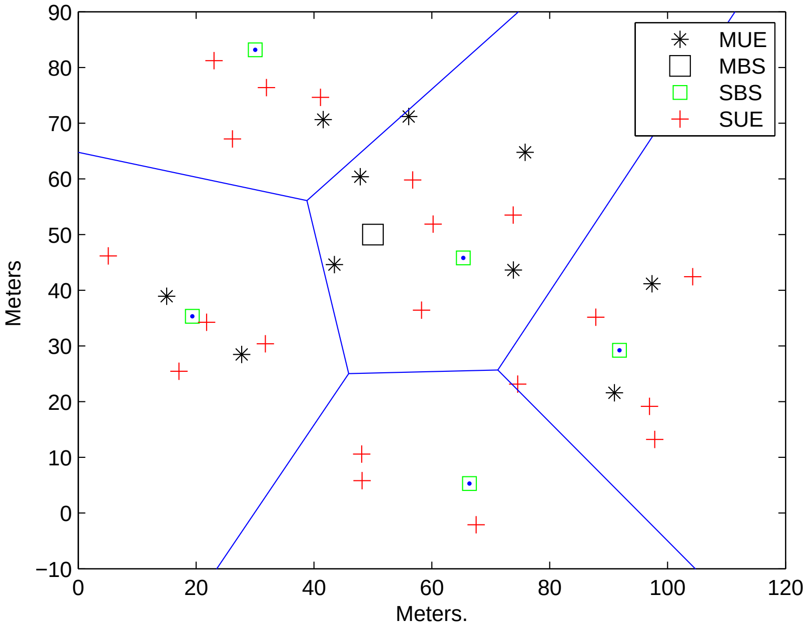


Figure 6.3: A two-tier H-CRAN model consisting of macro and small cells.

Table 6.2: Simulation parameters

Parameters	**Values**
Total number of H-CRAN users (SUEs and MUEs)	$10 - 120$
Total number of RBs	100
RB bandwidth	180 kHz
System bandwidth	20 MHz
Radius of small cell	10 m
Minimum data rate requirements	50-120 kbps
Transmission power of RRH	$10 - 20$ dB
Transmission power of MBS	$20 - 30$ dB
Path-loss exponent	4
Noise power spectral density	-144 dBm/Hz

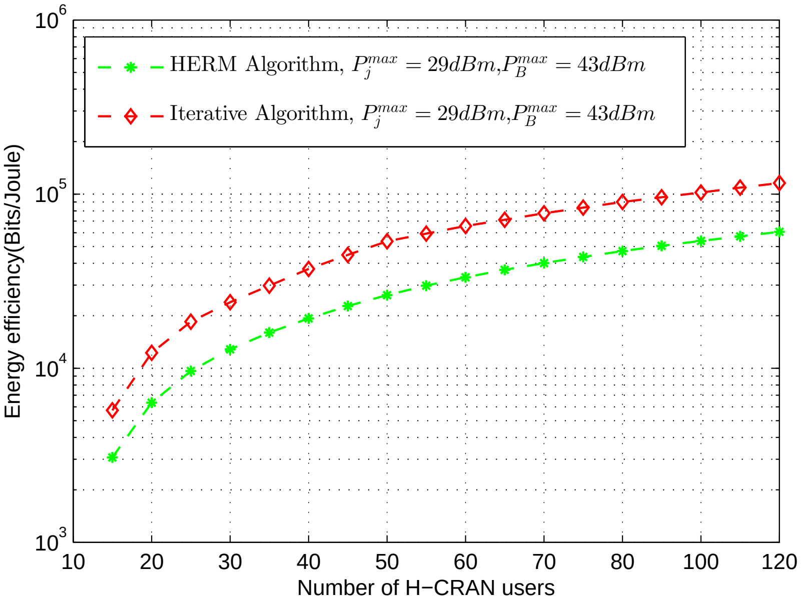


Figure 6.4: EE performance for different number of H-CRAN users.

6.6 Simulation Results

In this section, the performance of the proposed iterative solution is investigated with two different power budgets of macro and small cell base stations. In the simulation model, we consider a 120 m ×100 m area as shown in Fig. 6.3, where one macro base station is underlaid by 5 to 6 small cell base stations. The locations of RRHs, SUEs and MUEs are modeled using spatial Poisson point process with predefined intensity values. The settings for the simulation parameters are shown in Table 7.3. The simulation runs are averaged over 100 iterations. The performance of our proposed method is evaluated in terms of EE and Jain's fairness in H-CRAN. According to the EE of the entire H-CRAN system, we define Jain's fairness index as:

$$J = \frac{\left(\sum_{i=1}^{\mathcal{K}} \eta_i\right)^2}{\mathcal{K}\sum_{i=1}^{\mathcal{K}} \eta_i^2}. \tag{6.8}$$

Also, we consider the location-aware user association scheme [73], where the users are associated with the closest base station depending on relative distance and signal strength.

The EE performance of the proposed iterative method is compared with H-CRAN energy-

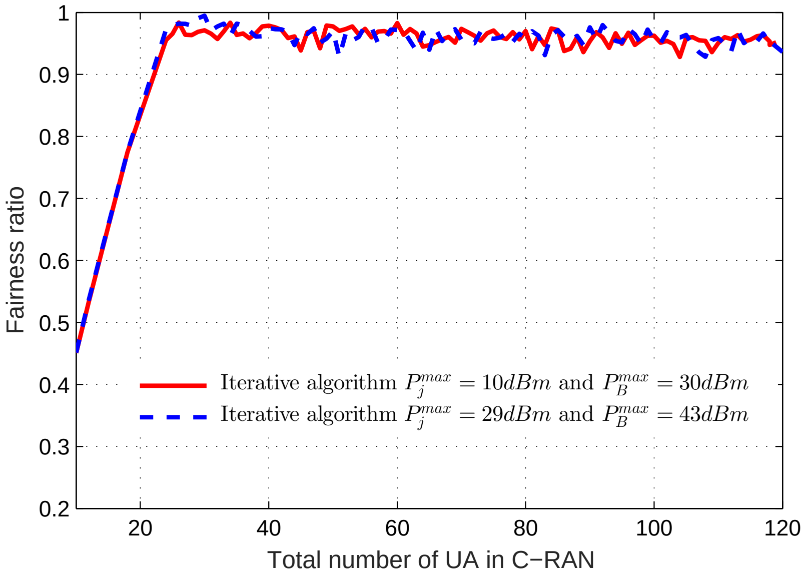


Figure 6.5: Convergence behavior of iterative algorithm.

efficient radio resource management (HERM) algorithm [28], shown in Fig. 6.4. We consider enhanced RRH (eRRH) in [28] as an MBS which consumes maximum power 43 dBm, and each RRH has 29 dBm maximum power budget [74]. The proposed iterative algorithm allocates power to the RBs during the resource allocation. In **Algorithm 5**, the MUEs use equal power level for the RBs, as SUEs use the same RB list. They optimize the power to RB so that the data rates of MUEs are not changed. On the other hand, HERM algorithm considers only energy savings at the BBU pool side. HERM uses an iterative process to optimize power on BBU pool during the low traffic load by switching off virtual machines inside the BBU pool.

The convergence behavior of iterative algorithm is shown with the fairness index in Fig. 6.5. The convergence behaviour is investigated with two different power budgets; i) $P_j^{max} = 10$ dBm, $P_B^{max} = 30$dBm , and ii) $P_j^{max} = 29$dBm, $P_B^{max} = 43$dBm. According to the definition of Jain's index in (6.8), a higher value of J represents a fair allocation of resources in H-CRAN in the perspective of EE [64]. It is apparent from the Fig. 6.5 that the iterative algorithm for both power budgets shows non-decreasing EE and converges within a fair resource allocation with increasing number of user association in H-CRAN.

Fig. 6.6 shows the performance of sum data rate performance for total number of user association in H-CRAN. It can be observed that the sum data rate performance of the iterative

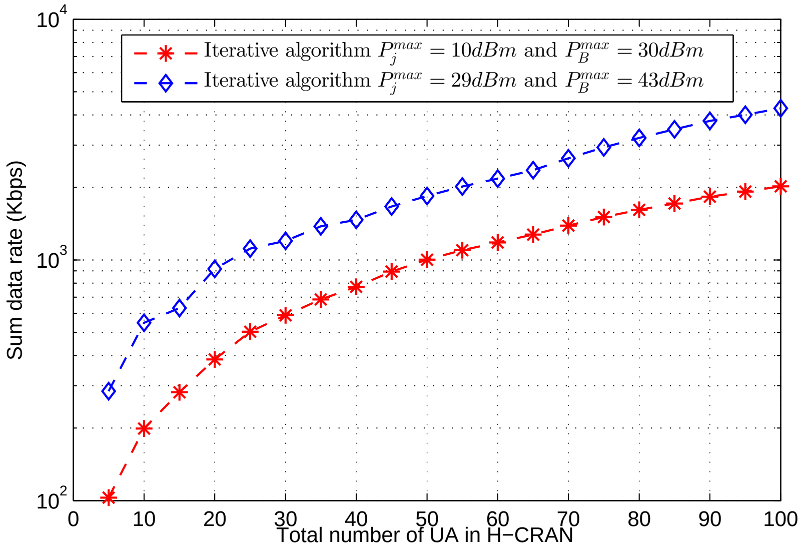


Figure 6.6: Sum data rate performance of iterative algorithm with two different power budgets.

algorithm gives better results when more power is used in RRHs and MBS. This is due to the fact that in the location based user association, the user i connects to base station j based on the maximum received channel state information (CSI). Considering equal transmission power in each small cell and the noise factor, according to the distance dependent pathloss model, the SINR of each user becomes maximum when the distance between the user and base station becomes minimum.

6.7 Summary

In this chapter, we proposed an iterative resource allocation method for two-tier OFDMA based C-RAN system where users in small cell uses the same radio resource with macro cell in an underlaid approach. The proposed iterative resource allocation method satisfied the resource allocation and maximum power constraints for both macro and small cell base stations, interference, front-haul capacity constraint as well as the QoS constraint of macro cell users. The simulation results showed that the proposed iterative algorithm converges and improves EE and sum data rate in C-RAN through the underlaid approach of resource allocation.

In this chapter, we investigated the energy efficiency aspects of resource allocation in OFMDA supported C-RANs. In the next chapter, we discuss the energy efficiency aspects of non-orthogonal multiple access (NOMA) based C-RANs. NOMA approach has been identified as a promising solution for future networks [8] to increase connections per unit area. Due to the limited number of orthogonal resources in OFDM based C-RANs support fewer number of user connections. The NOMA supported C-RANs is regarded as a new technological solution for the next generation cellular networks to handle more traffic and user loads. Authors in [75] proposed the NOMA enabled heterogeneous C-RAN architecture which brings the advantages of SE, EE and massive connectivity through the centralized coordination in a BBU pool. However, the technical challenges of NOMA in the context of bandwidth and power allocation in C-RAN have not been investigated in detail in the literature. In the next chapter, we investigate in detail the energy efficient method for NOMA supported C-RANs.

Chapter 7

Energy Efficient Resource Allocation in SCMA supported Cloud-RAN

It is envisioned that the 5G mobile communication networks will support many-fold connected devices per unit area compared to 4G LTE networks. The LTE and LTE-A networks support the orthogonal multiple access (OMA) technique, which utilizes limited number of orthogonal resources for the users. Similarly, OMA supported C-RANs utilize limited number of orthogonal resources considered as communication resources for the small cell users [65]. To improve spectrum efficiency in OMA supported C-RANs, authors in [27] and [76] have considered an underlaid approach of orthogonal resource sharing both for macro cell and small cell users. In the previous chapter, we consider an energy efficient underlaid approach of resource allocation for both small and macro cell users in C-RAN. However, this underlaid approach increases intra-cell and inter-cell interference levels which limit the data rate of users. Considering the interference issues, authors in [76] have proposed an auction based distributed resource allocation with the aim to improve the data rate of small cell users. Nevertheless, the OMA approach supports limited number of connections due to the use of orthogonal resources.

Therefore, to increase connections per unit area, non-orthogonal multiple access (NOMA) approach has been identified as a promising solution for future networks [8]. Unlike OMA, the NOMA methods utilize different power levels or overlapping signatures to provide services to multiple users. For example, the sparse code multiple access (SCMA), which is classified as

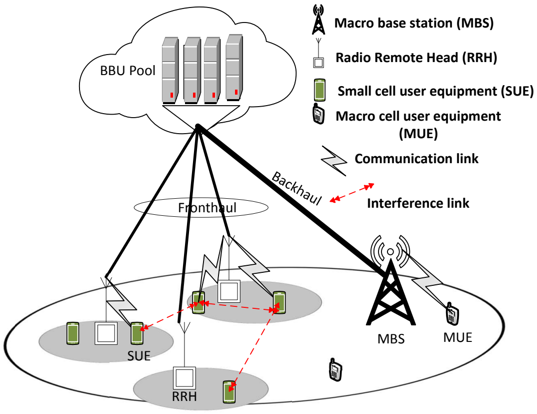


Figure 7.1: Two-tier heterogeneous cloud radio access network (H-CRAN).

one category of NOMA methods, assigns different codebooks (CBs) to different users. SCMA is regarded as the generalized low density signature (LDS), where LDS uses sparse spreading sequences and SCMA uses sparse codewords in the codebooks. Each codeword comprises of non-orthogonal resources such as sub-carriers (SCs) that are shared by different users [9]. The multiplexed signals of different users superimposed over the same sub-carrier can be decoded by the message passing algorithm [10] with low complexity. On the other hand, power domain NOMA (PD-NOMA) method uses different power levels for multiple users to provide services in the same sub-carrier and time slot [8]. The NOMA methods support massive connectivity and higher utilization of bandwidth and provide higher SE and EE. However, the non-orthogonality in NOMA increases the mutual interference levels, therefore successive interference cancelation (SIC) method is applied at the receiver side [11].

NOMA enabled heterogeneous C-RAN architecture has been proposed in [75]. Considering NOMA in C-RAN may bring the advantages of SE, EE and massive connectivity through the centralized coordination in a BBU pool. However, the technical challenges of NOMA in the context of bandwidth and power allocation in C-RAN have not been investigated in detail in the literature. In this chapter, we investigate in detail the energy efficient SCMA method for C-RANs in terms of codebook and power allocation. We apply the conflict graph and geometric water filling approaches to solve codebook and power allocation in SCMA supported C-RANs with the objective to maximize energy efficiency. To the best of our knowledge, this is the first work to consider energy efficient codebook and power optimization in SCMA supported C-RANs. The main contributions of this chapter are as follows:

- The SCMA method is implemented to jointly optimize codebook and power allocation in the downlink of the C-RANs. In the literature, many studies misuse codebook allocation of SCMA schemes and indeed implement subcarrier allocation instead. Accordingly, the structure of the SCMA codebook assignment, and the optimization formulation of codebook and power allocation with the objective to improve the energy efficiency in C-RAN are presented in detail.

- To solve the joint optimization problem, the original problem is decomposed into two optimization problems. The first optimization problem is codebook allocation (CA) with equal power allocation and the second optimization problem is power allocation (PA) with known codebook allocation.

- For the codebook allocation problem, the throughput aware SCMA codebook selection (TASCBS) method is proposed using the conflict graph theory. It is proven that the TASCBS method generates a stable codebook allocation solution within a finite number of steps.

- For the power allocation problem, the iterative level-based power allocation (ILPA) method, which incorporates different power allocation approaches (e.g., weighted and NOMA-SIC) into different levels to satisfy the maximum power requirement, is proposed.

- It is shown that the NOMA-SIC aware power allocation can be used with the geometric water filling method in the subcarrier level in a computationally efficient way and achieve higher energy efficiency compared to other power allocation approaches.

7.1 Related Works

The SCMA and PD-NOMA methods have been extensively studied in resource allocation problems in single cell networks [77–80] and multi-cell networks [81,82]. A summary of these studies and the proposed solutions are given in Table 7.1. The resource allocation in PD-NOMA systems mainly considers power and subcarrier allocation to improve system efficiency in terms of SE and EE. Similarly, in SCMA supported networks, the subcarrier allocation is referred to as CB

Table 7.1: Summary of resource optimization problems

Ref.	**Network scenario**	**Link scenario**	**Multiple access**		**Scheduling**			**Power allocation**	**Solution approach**
			SCMA	PD-NOMA	SC	CB	user		
[77]	-Single-cell	-DL	✓		✓			✓	Remove and reallocate iterative algorithm
[78]	-Single-cell	-UL		✓			✓		Optimal user grouping
[79]	-Single-cell	-UL	✓		✓				Matching theory
[80]	-Single-cell	-UL		✓	✓			✓	Matching theory, iterative water filling
[81]	-Multi-cell	-DL		✓				✓	Distributed approach
[82]	-Multi-cell	-DL	✓	✓	✓	✓		✓	SCALE and GP
Proposed	-Multi-cellC-RAN	-DL	✓		✓	✓		✓	Conflict graph based TASCBS, ILPA

or SC allocation for each user [83]. However, the joint optimization of power and SC allocation in PD-NOMA and SCMA systems leads to an NP-hard problem [80]. Therefore, the matching theory, greedy algorithm and auction method are preferred for practical implementation. The matching theory has widely used in the application of college admission, stable marriage and resource allocation problem [84]. In [80], the authors consider uplink single cell NOMA system and allocate SCs to users using geometric programming (GP) and many-to-many matching model. For the power allocation, iterative water-filling (IWF) algorithm is applied to improve data rate of the system in [80]. Similarly, matching theory based power and SC allocation in downlink single cell network is proposed in [85]. Two-tier heterogeneous network (HetNet) with downlink power allocation is studied in [81], where users receive data from multiple access points using CoMP NOMA method. Power budget and channel gain based SIC constraints are considered in the power optimization problem. Similarly, in a downlink HetNet, the performance of PD-NOMA and SCMA methods are compared in [82] in terms of maximizing the data rate. Joint SC allocation and power optimization in PD-NOMA with the maximum power and SIC constraints are considered, whereas in the SCMA method, joint codebook and power optimization problem is considered only with the base station maximum power and SC sharing constraints. To solve the nonconvex joint optimization problem in PD-NOMA and SCMA, the authors utilized GP and successive convex approximation for low complexity (SCALE) algorithm which involves a series of convex relaxations [86]. Different from the above works, in this chapter, we investigate

Table 7.2: Relationship between number of SCs (K), number of codebooks (J) and the overloading factor (α)

	$K = 4$, $N = 2$	$K = 8$, $N = 2$	$K = 12$, $N = 2$	$K = 16$, $N = 2$
Without α $J = \frac{K!}{N!(K-N)!}$	$J = 6$	$J = 28$	$J = 66$	$J = 120$
With $\alpha = 1.5$ $J = K\alpha$	$J = 6$	$J = 12$	$J = 18$	$J = 24$

the energy efficiency aspects of the SCMA method in C-RANs in terms of codebook allocation, maximum power budget both at the base station and codebook, and minimum data rate constraint for each user.

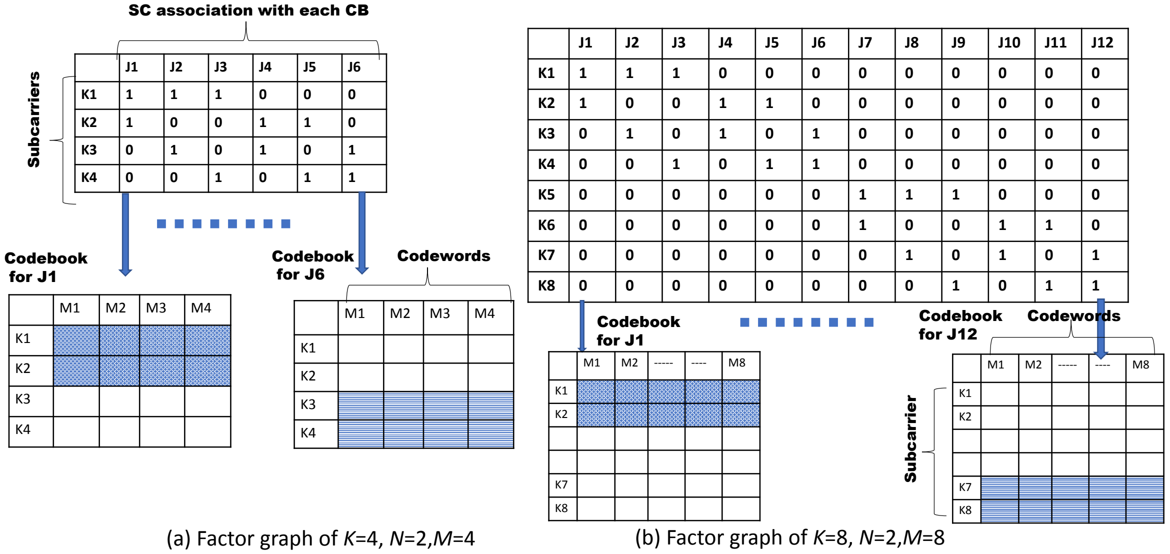
7.2 System model and Assumptions

7.2.1 SCMA Codebook

We assume that the SCMA scheme supports a total number of K subcarriers, indexed by $k = \{1, 2, 3, ..., K\}$, where each codeword uses N out of K SCs. The total number of J codebooks are available in each cell, indexed by $j = \{1, 2, ..., J\}$, where $J \leq \frac{K!}{N!(K-N)!}$. The total number of codebooks is generally limited by the overloading factor $\alpha = \frac{J}{K}$. In SCMA scheme, d_f denotes the number of users that can use the same subcarrier and d_v denotes the number of SCs used in each codeword. The relationships between SCs, codebooks and the overloading factor are shown in Table 7.2. The SC assignment in each SCMA codebook is represented by a factor graph $F = [f_{k,j}]$, where $f_{k,j}$ is the binary variable for SC allocation in each CB, updated as follows:

$$f_{k,j} = \begin{cases} 1, & \text{if SC } k \text{ is assigned to codebook } j, \\ 0, & \text{otherwise.} \end{cases} \tag{7.1}$$

The CB design follows two constraints: 1) $\sum_{j=1}^{J} f_{k,j} \leq d_f$, $\forall k$ and 2) $\sum_{k=1}^{K} f_{k,j} \leq d_v$, $\forall j$. For simplicity, we assume that the base station knows the design of the codebook. Fig. 7.2



	J1	J2	J3	J4	J5	J6
K1	1	1	1	0	0	0
K2	1	0	0	1	1	0
K3	0	1	0	1	0	1
K4	0	0	1	0	1	1

(a) Factor graph of *K*=4, *N*=2,*M*=4

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12
K1	1	1	1	0	0	0	0	0	0	0	0	0
K2	1	0	0	1	1	0	0	0	0	0	0	0
K3	0	1	0	1	0	1	0	0	0	0	0	0
K4	0	0	1	0	1	1	0	0	0	0	0	0
K5	0	0	0	0	0	0	1	1	1	0	0	0
K6	0	0	0	0	0	0	1	0	0	1	1	0
K7	0	0	0	0	0	0	0	1	0	1	0	1
K8	0	0	0	0	0	0	0	0	1	0	1	1

(b) Factor graph of *K*=8, *N*=2,*M*=8

Figure 7.2: Factor graph and codebook structure.

shows the factor graph representation and codebook assignments for two cases of J, K, M and N values. Using the factor graph, two sets of information, i.e., $\{S_{J_j}\}$ and $\{S_{K_k}\}$ can be retrieved. $\{S_{J_j}\}$ contains the set of non-zero SC information, which belongs to codebook j, i.e., $f_{k,j} = 1$ and $k \in \{S_{J_j}\}$. Similarly, $j \in \{S_{K_k}\}$ contains the information about the codebooks, where the SC k is utilized. For example, In Fig. 7.2(a), $\{S_{J_1}\} = \{K_1, K_2\}$, $\{S_{J_2}\} = \{K_1, K_3\}$,..., $\{S_{J_6}\} = \{K_3, K_4\}$ and $\{S_{K_1}\} = \{J_1, J_2, J_3\}$, $\{S_{K_2}\} = \{J_1, J_4, J_5\}$,...,$\{S_{K_4}\} = \{J_3, J_5, J_6\}$. The base station assigns one codebook to one user with the goal to optimize downlink codebook allocation and power allocation jointly to improve system efficiency in terms of data rate and power, where CA and PA will be explained in the following sections. For a summary of symbols and parameters, a list is provided in Table 7.3.

7.2.2 System Model

We consider an SCMA based two-tier C-RAN, as shown in Fig. 7.1, where $\mathcal{B}$ is the total number of small cells, indexed by $b = \{1, 2, ...B\}$, which are covered by a single macro cell. Each SUE is equipped with one antenna and each small cell has ϕ antennas. The system supports a total

Table 7.3: List of symbols.

	Symbol	**Description**
Set	J	Total number of CBs
	K	Total number of SCs
	$\mathcal{B}$	Total number of SBS
	$\mathcal{U}$	Total number of SUEs
	$\mathcal{U}_b$	Total number of SUEs in SBS b
	M	Total number of codewords in each codebook
	S_{J_j}	Set of SCs in CB j
	S_{K_k}	Set of CBs using SC k
Index	k	Indexing for SC
	j	Indexing for CB
	b	Indexing for SBS
	u	Indexing for SUE
	U_u	User notation
	J_j	CB notation
	K_k	SC notation
Optimization Parameters	$f_{k,j}$	SC association parameter of k^{th} SC for j^{th} CB
	$q^b_{u,j}$	CB-User mapping parameter of j^{th} CB to u^{th} SUE on SBS b
	$P^b_{u,j}$	Power allocation from b^{th} SBS to u^{th} SUE on CB j
	$P^b_{u,k}$	Power allocation from b^{th} SBS to u^{th} SUE on SC k
Channel Parameters	$h^b_{u,k}$	The channel gain from b^{th} SBS to u^{th} SUE on k^{th} SC
	$h^b_{u,j}$	The channel gain from b^{th} SBS to u^{th} SUE on j^{th} CB
	P^b_{max}	Maximum power of SBS b
	$\gamma^b_{u,j}$	SINR of u^{th} SUE connected to b^{th} SBS on j^{th} CB
	$\Gamma_{k,u}$	SNR of u^{th} SUE on k^{th} SC
	$g_{k,u}$	Channel gain to noise ratio of u^{th} SUE on k^{th} SC
Others	N	Total number of nonzero elements in each SC
	F	Factor graph of CB design
	α	Overloading factor
	d_f	Total number of users for each SC
	d_v	Total number of SCs for each CB
	η	Energy efficiency
	$r^b_{u,j}$	Data rate of SUE u for b^{th} SBS on j^{th} CB
	$\tilde{r}_{u,j}$	Normalized rate of u^{th} SUE on j^{th} CB
	R_T	Total achievable data rate of all SUEs
	P_T	Total allocated power of all SBS

number of small cell users $U = \sum_{b=1}^{B} U_b$, where a SUE in SBS b is indexed by $u = \{1, 2, 3, ..., U_b\}$. Let $Q = [q^b_{u,j}]$ be the CB association matrix, where $q^b_{u,j}$ is the binary variable for codebook allocation as follows:

$$q^b_{u,j} = \begin{cases} 1, & \text{if codebook } j \text{ is assigned to SUE } u \text{ on SBS } b, \\ 0, & \text{otherwise.} \end{cases} \tag{7.2}$$

The channel gain from SBS b to SUE u on SC k is denoted as $h^b_{u,k} \in \mathcal{C}^{\phi \times 1}$. The power allocated from SBS b to user u on SC k is denoted as $P^b_{u,k} \in (0, P^b_{max}]$, where P^b_{max} denotes the maximum power of SBS b. The signal-to-interference-plus noise ratio (SINR) achieved by SUE u, connected to SBS b on codebook j can be written as

$$\begin{aligned} \gamma^b_{u,j} &= \frac{q^b_{u,j} |h^b_{u,j}|^2 P^b_{u,j}}{I^b_{u,j} + \sigma^2} \\ &= \frac{q^b_{u,j} \sum_{k \in \{S_{J_j}\}} f^b_{k,j} |h^b_{u,k}|^2 P^b_{u,k}}{I^b_{u,j} + \sigma^2} \end{aligned} \tag{7.3}$$

where σ^2 is the noise power. The channel gain and power allocation from SBS b to SUE u on CB j are denoted as $h^b_{u,j}$ and $P^b_{u,j}$, respectively. Here, we denote $P^b_{u,j}$ as CB-user power allocation, which is the summation of SC-user power allocation given as

$$P^b_{u,j} = \sum_{k \in \{S_{J_j}\}} f^b_{k,j} P^b_{u,k}. \tag{7.4}$$

Similar to [67], $I^b_{u,j}$ is the aggregated interference power, defined as

$$I^b_{u,j} = \underbrace{\sum_{u' \in U_b} \sum_{k \in \{S_{J_j}\}} \sum_{\substack{j' \neq j, \\ j' \in \{S_{\mathcal{K}_k} || h^b_{u',k}|^2 > |h^b_{u,k}|^2\}}} q^b_{u',j'} f^b_{k,j'} |h^b_{u',k}|^2 P^b_{u',k}}_{\text{Intra-cell interference}} + \underbrace{\sum_{b' = B \setminus \{b\}} \sum_{u' \in U_{b'}} \sum_{k \in \{S_{J_j}\}} \sum_{\substack{j \in \{S_{\mathcal{K}_k}\} \\ |h^{b'}_{u',k}|^2 > |h^b_{u,k}|^2}} q^{b'}_{u',j} f^{b'}_{k,j} |h^{b'}_{u',k}|^2 P^{b'}_{u',k}}_{\text{Inter-cell interference}}.$$

The first term on the right hand side is the intra-cell interference signal coming from other active SUEs (i.e., $u' \in U_b$) in the same cell, utilizing same SCs (i.e., $k \in \{S_{J_j}\}$) in other codebooks, $j' \neq j$ where $j' \in \{S_{\mathcal{K}_k} || h^b_{u',k}|^2 > |h^b_{u,k}|^2\}$. Similarly, the second term denotes the inter-cell interference coming from other cells utilizing same SCs (i.e., $k \in \{S_{J_j}\}$).

7.3 Energy Efficient Codebook Allocation and Power Allocation in C-RAN

The objective of resource allocation in C-RAN is to maximize the EE in terms of codebook and power allocation. The EE can be measured by total achievable data rate divided by total allocated power (bits/J), written as

$$\eta = \frac{R_T}{P_T}. \tag{7.5}$$

According to the Shannon formula, the achievable data rate achieved by SUE u connected to SBS b using codebook j will be $r^b_{u,j} = \log_2(1 + \gamma^b_{u,j})$, and the total data rate of all SUEs can be expressed as

$$R_T = \sum_{b \in B} \sum_{u \in U_b} \sum_{j \in J} r^b_{u,j}.$$

The total allocated power of small cells is denoted by

$$P_T = \underbrace{\sum_{b \in B} \sum_{u \in U_b} \sum_{j \in J} P^b_{u,j}}_{\text{dynamic}} + \underbrace{\sum_{b \in B} P^b_s}_{\text{static}}.$$

Similar to [68], we assume that each small cell has dynamic and static power factors. Dynamic power depends on codebook allocation, whereas the circuit power is regarded as the static power P^b_s. The mathematical formulation of joint CA and PA in an SCMA based C-RAN system can be described as follows:

$$\textbf{P1:} \max_{Q,P} \quad \eta \tag{7.6}$$

subject to:

$$
\begin{aligned}
\text{C1:} \quad & \sum_{j \in J} q_{u,j}^{b} = 1, \quad \forall u \in U_b, b \in B, \\
\text{C2:} \quad & \sum_{u \in U_b} q_{u,j}^{b} = 1, \quad \forall j \in J, b \in B, \\
\text{C3:} \quad & q_{u,j}^{b} \sum_{j=1}^{J} P_{u,j}^{b} \leq P_{max}^{b}, \quad \forall u \in U_b, b \in B, \\
\text{C4:} \quad & q_{u,j}^{b} \sum_{k \in \{S_{J_j}\}} f_{k,j}^{b} P_{u,k}^{b} \leq P_{u,j}^{b}, \quad \forall u \in U_b, b \in B, \\
\text{C5:} \quad & r_{u,j}^{b} \geq r_{min}, \forall u \in U_b, b \in B, \\
\text{C6:} \quad & q_{u,j}^{b} \in \{1, 0\} \text{ and } P_{u,k}^{b} \geq 0.
\end{aligned}
$$

In (7.6), the objective is to maximize the EE of the C-RAN system by allocating the same codebook among small cells in an underlaid approach. Two optimization parameters are considered in this problem: i) codebook allocation vector for small cell users (i.e., $q_{u,j}^{b} \in \{0, 1\}$), and ii) allocated power for small cell users (i.e., $P_{u,j}^{b}$ and $P_{u,k}^{b}$). C1 and C2 enforce that each SUE is connected to one SBS using one codebook. C3 ensures that the maximum power budget constraint for each small cell b, which is P_{max}^{b}, should be satisfied. C4 is the individual power budget constraint of each SUE and codebook in terms of SCs. C5 enforces the minimum data rate constraint of each user.

The objective function in (7.6) and the constraints C5 and C6 turn the problem **P1** into a mixed integer non-linear program (MINLP) with the non-convex feasibility set. The optimization problem **P1** is computationally intractable and is a NP-hard problem [36]. Similar to [67], [87], we adopt a two-step iterative approach to solve the problem **P1**. We split **P1** into two subproblems: i) codebook allocation (**P2**) and ii) power allocation (**P3**). Assuming fixed power allocation, the codebook allocation problem can be formulated as

$$
\begin{gathered}
\textbf{P2:} \max_{Q} \quad R_T \qquad (7.7) \\
\text{subject to: C1 to C2, C5,} \\
\text{C6:} \quad q_{u,j}^{b} \in \{1, 0\}.
\end{gathered}
$$

Similarly, assuming fixed codebook allocation the power optimization problem can be formulated as

$$\begin{aligned}\textbf{P3:} \min_{P} \quad & P_T \\ \text{subject to: } & \text{C3 to C4, C5,} \\ & \text{C6:} P_{k,u}^{b} \geq 0. \end{aligned} \tag{7.8}$$

Initially, we address codebook allocation by assuming equal power allocation, no interference and minimum data rate constraints and propose the throughput aware SCMA CB Selection (TASCBS) method. Then, we adjust the power level using the channel gain to noise ratio, the effect of interference and minimum data rate information. Codebook and power allocation are explained in the following sections, respectively.

7.4 Codebook Allocation

To solve the problem **P2**, we assume that the cloud controller of C-RAN knows all the channel state information (CSI) of the users and the factor graph of the CB design. For simplicity, we assume that all the SBSs under the C-RAN use the same CB design. The proposed CB allocation is explained as follows:

7.4.1 Throughput Aware SCMA CB Selection (TASCBS)

Equal power allocation: According to the SCMA CB design, each user uses N number of non-zero SCs in each CB. Therefore, all the users receive equal power from BS according to the total number of CBs, i.e., $P_{u,j} = \frac{P_{max}^{b}}{|J|}$ and each SC receives equal portion of the received power, i.e., $P_{u,k} = \frac{P_{u,j}}{|N|}$.

SNR estimation: Considering that there is fully available CSI information and no interference, the SNR of user u on SC k can be estimated as $\Gamma_{k,u} = g_{k,u}P_{u,k}$, where $g_{k,u} = \frac{|h_{u,k}|^2}{\sigma^2}$ represents the channel gain to noise ratio (CNR) defined as

$$CNR = \begin{array}{c|ccccc|} & U_1 & U_2 & .. & .. & U_{U_b} \\ \hline K_1 & g_{1,1} & g_{1,2} & .. & .. & g_{K_1,U_b} \\ K_2 & g_{2,1} & g_{2,2} & .. & .. & g_{K_2,U_b} \\ . & .. & .. & .. & .. & .. \\ . & .. & .. & .. & .. & .. \\ K_K & g_{K_K,1} & g_{K_K,2} & .. & .. & g_{K_K,U_b} \end{array}. \tag{7.9}$$

Normalized rate estimation: Based on the CNR and fixed CB design, the data rate of user u on codebook j can be estimated as $r_{u,j} = \log_2(1 + \sum_{k \in \{S_{J_j}\}} g_{k,u} P_{u,k})$. The normalized rate of each user can be obtained as $\tilde{r}_{u,j} = \frac{r_{u,j}}{\sum_{\forall j} r_{u,j}}$, where the summation of all normalized rates per user becomes one, i.e., $\sum_{\forall j} \tilde{r}_{u,j} = 1$. In (7.10), the normalized rate of each user for all codebooks are defined as

$$\begin{array}{c|ccccc|} & J_1 & J_2 & .. & .. & J_J \\ & \{K_1, K_2\} & \{K_1, K_3\} & .. & .. & \{K_3, K_4\} \\ \hline U_1 & \tilde{r}_{1,1} & \tilde{r}_{1,2} & .. & .. & \tilde{r}_{1,J} \\ U_2 & \tilde{r}_{2,1} & \tilde{r}_{2,2} & .. & .. & \tilde{r}_{2,J} \\ . & .. & .. & .. & .. & .. \\ . & .. & .. & .. & .. & .. \\ U_{U_b} & \tilde{r}_{U_b,1} & \tilde{r}_{U_b,2} & .. & .. & \tilde{r}_{U_b,J} \end{array}. \tag{7.10}$$

Bipartite graph and CB selection: A bipartite graph $G = \{U, J, E\}$ is depicted based on the users, CBs and the normalized rate information. The vertex set U denotes the set of users and the vertex set J represents the set of CBs. Each vertex of U is connected to vertex J based on the maximum normalized rate, i.e.,

$$E(u, j^*) = \max_{\forall j \in J} \tilde{r}_{u,j} \text{ and } \tilde{r}_{u,j} \geq r_{min}.$$

This means that each user u selects the CB j^* based on the maximum normalized rate. The flow graph of the throughput aware CB selection method is given in Fig. 7.3. The step-by-step procedure of CB selection method is illustrated in Fig. 7.3. The CB selection and conflict

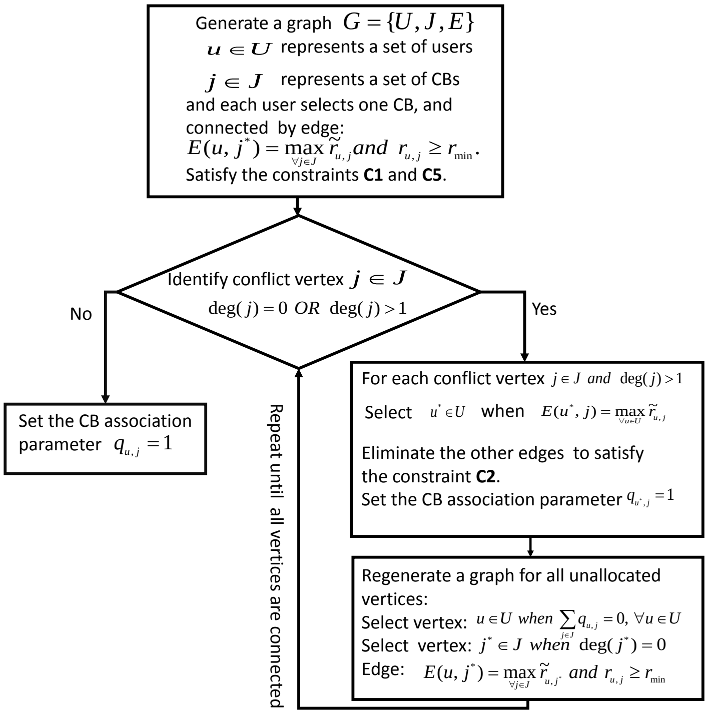

Generate a graph $G=\{U,J,E\}$
$u \in U$ represents a set of users
$j \in J$ represents a set of CBs
and each user selects one CB, and
connected by edge:
$E(u,j^*)=\max_{\forall j \in J}\tilde{r}_{u,j}$ and $r_{u,j} \geq r_{\min}$.
Satisfy the constraints **C1** and **C5**.
Identify conflict vertex $j \in J$
$\deg(j)=0$ OR $\deg(j)>1$
No
Yes
Set the CB association parameter $q_{u,j}=1$
Repeat until all vertices are connected
For each conflict vertex $j \in J$ and $\deg(j)>1$
Select $u^* \in U$ when $E(u^*,j)=\max_{\forall u \in U}\tilde{r}_{u,j}$
Eliminate the other edges to satisfy the constraint **C2**.
Set the CB association parameter $q_{u^*,j}=1$
Regenerate a graph for all unallocated vertices:
Select vertex: $u \in U$ when $\sum_{j \in J} q_{u,j}=0, \forall u \in U$
Select vertex: $j^* \in J$ when $\deg(j^*)=0$
Edge: $E(u,j^*)=\max_{\forall j \in J}\tilde{r}_{u,j^*}$ and $r_{u,j} \geq r_{\min}$

Figure 7.3: Throughput aware CB selection method.

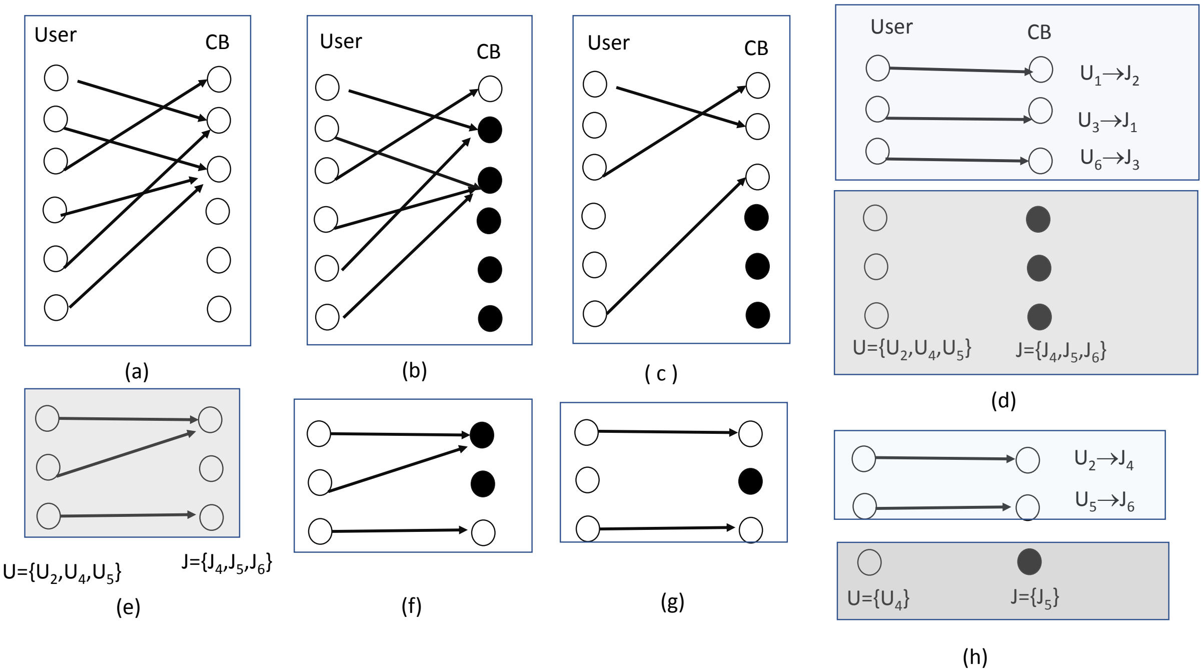


Figure 7.4: (a) Bipartite graph, (b) Identifying conflict vertices, (c) Winner selection and edge elimination, (d) Allocated and unallocated vertices, (e) Repeated case of (a), (f) Repeated case of (b), (g) Repeated case of (c), (h) Repeated case of (d).

resolution procedure are repeated until all the users are connected to their appropriate CBs. Fig. 7.4(a) shows an example of the bipartite graph representation for $U = 6$ and $J = 6$. When more than one user in the same SBS selects the same CB $j^* \in J$, then j^* is represented as a conflict vertex in the bipartite graph. In Fig. 7.4(b), all the black (dark) color vertices in the CB represent the conflict vertex.

Definition 1: *A vertex in $j \in J$ becomes a conflict vertex when it is matched by either more than one $u \in U$ or no u.*

The conflict vertices ($j \in J$) in bipartite graph G are identified using the degree information of the vertices; i) $deg(j) = 0$ or, ii) $deg(j) > 1$. The conflict vertex violates the constraint **C2** in **P1** and **P2**. For conflict resolution, we propose a winner selection and edge elimination method, which is applied to all conflict vertices as long as each of the vertices U is connected to exactly one vertex in J.

Conflict resolution: For the winner selection method, the vertex of $u^* \in U$ becomes the winner of j when its edge shows the maximum value among others. For example, conflict vertex $j \in J$, selects u^* when

$$E(u^*, j) = \max_{\forall u \in U} \tilde{r}_{u,j} \tag{7.11}$$

and eliminates other edges to satisfy the condition **C2** in **P1**. This process is repeated for all other conflict vertices which have $deg(j) > 1$. Fig. 7.4(c) shows the graphical representation after the winner selection and edge elimination method.

7.4.2 Stability and Convergence of the TASCBS method

Definition 2: *A stable allocation is defined as no conflict vertex and each vertex in $u \in U$ is connected to at most one vertex in $j \in J$ and vice versa.*

Lemma 1: *The TASCBS converges to a pair-stable allocation, when the number of conflict vertices becomes zero.*

Proof: According to the codebook selection method given in Fig. 7.3, when the proposed TASCBS converges to a stable allocation, no user $u \in U$ has a conflict with another user for the same choice of CB $j^* \in J$. If j^* is selected by more than one user, then the TASCBS resolves the conflict by picking up the best user, who will benefit the most by using the utility (7.11). Thus, the matches of user u^* must be the best choice for other users in the current situation. Hence, the terminal matching is pair-stable. □

Theorem 1: *The proposed TASCBS converges to a stable allocation after a limited number of iterations.*

Proof: In TASCBS, each user $u \in U$ selects a CB $j \in J$ without knowing other users' choices. This increases the possibility of a conflict, also named as a conflict vertex. Each iteration in TASCBS resolves the conflict vertices and eliminates the edges which violate the one-to-one matching criterion, i.e., **C1** and **C2** in **P1**. There are J CBs in each small cell, so the number of selections that each user u makes for the CB is no larger than J, and thus, the total number of iterations is no more than J. Also, after each iteration, users and CBs are categorized into two groups, i.e., allocated and unallocated. The TASCBS procedure is repeated for the unallocated groups until there are no conflict choices. Therefore, the TASCBS converges to a stable allocation according to Theorem 1. □

7.5 Power Allocation (Iterative Level-based Power Allocation (ILPA))

In each cell[1], using the known values of the CB allocation parameter Q and the factor graph F, the power allocation problem (**P3**) can be reformulated as

---

[1] Since power allocation is performed separately at each SBS, without loss of generality, we drop the superscript b.

$$\textbf{P3:}\ \min_{P_{u,k}} \quad P_T \tag{7.12}$$

subject to:

$$\text{C3:}\quad P_{max} - \sum_{j=1}^{J} P_{u,j} \geq 0,$$

$$\text{C4:}\quad P_{u,j} - \sum_{k\in\{S_{J_j}\}} P_{u,k} \geq 0,$$

$$\text{C5:}\quad \log_2(1+\gamma_{u,j}) - r_{min} \geq 0,$$

$$\text{C6:}\quad P_{u,k} \geq 0.$$

To perform the power allocation, we use Karush-Kuhn-Tucker (KKT) optimality and define the following Lagrangian function

$$\begin{aligned}\mathbb{L}(P_{u,j}, P_{u,k}, \lambda, \beta, \phi) = &\sum_{\forall u}\sum_{\forall j} P_{u,j} + P_s - \lambda\{P_{max} - \sum_{j=1}^{J} P_{u,j}\} \\ &- \sum_{\forall j}\beta_j\{P_{u,j} - \sum_{k\in\{S_{J_j}\}} P_{u,k}\} \\ &- \sum_{\forall j}\phi_j\{\log_2(1+\gamma_{u,j}) - r_{min}\}\end{aligned} \tag{7.13}$$

where λ, β, ϕ are the Lagrange multipliers for the constraints C3-C5, respectively. Differentiating (7.13) with respect to $P_{u,k}$, we obtain the following power allocation of SUE u over SC k as

$$P_{u,k} = \left[\frac{\phi_j}{\ln(1+\lambda)} - \frac{1}{\delta_{u,k}}\right]^+, \tag{7.14}$$

where $\delta_{u,k} = \frac{|h_{u,k}|^2}{I^b_{u,j}+\sigma^2}$ and $[\varepsilon]^+ = \max(\varepsilon, 0)$, which is a multi-level water filling allocation [36].

Proof: Using the relation of CB-user power and SC-user power in (7.4) and letting $\delta_{u,k} =$

$\frac{|h_{u,k}|^2}{I^b_{u,j}+\sigma^2}$, the problem of (7.13) becomes

$$\begin{aligned}\mathbb{L}(P_{u,k},\lambda,\beta,\phi) = &\sum_{\forall u}\sum_{\forall j}\sum_{k\in\{S_{J_j}} P_{u,k} + P_s \\ &- \lambda\left\{P_{max} - \sum_{j=1}^{J}\sum_{k\in\{S_{J_j}\}} P_{u,k}\right\} \\ &- \sum_{\forall j}\beta_j\left\{\sum_{k\in\{S_{J_j}\}} P_{u,k} - \sum_{k\in\{S_{J_j}\}} P_{u,k}\right\} \\ &- \sum_{\forall j}\phi_j\left\{\log_2\left(1+\sum_{k\in\{S_{J_j}\}} P_{u,k}\delta_{u,k}\right) - r_{min}\right\}.\end{aligned}$$

Minimizing **P3** for any given Q and F is equivalent to differentiating $\text{Ł}(P_{u,k})$ with respect to $P_{u,k}$ and setting the result to zero. That is

$$\begin{aligned}\frac{\partial \text{Ł}}{\partial P_{u,k}} &= 0 \\ 1+\lambda - \frac{\phi_j\delta_{u,k}}{\ln(1+P_{u,k}\delta_{u,k})} &= 0\end{aligned}$$

Therefore, $P_{u,k}$ can be obtained as

$$P_{u,k} = \left[\frac{\phi_j}{\ln(1+\lambda)} - \frac{1}{\delta_{u,k}}\right]^{+}.$$

□

In the SCMA scheme, each SC is shared among d_f number of users, therefore it is important to consider intra-cell interference during the SC-user power allocation. On the other hand, the SC-user PA in (7.14) depends on the optimal choice of λ, ϕ, and δ values and do not consider the interference cancellation during PA. Thus, we propose an iterative level based PA, which exploits the relation between users, CBs, SCs and users' channel state information during the power allocation.

Iterative Level-based Power Allocation (ILPA)

The proposed iterative level-based power allocation (ILPA) method consists of five levels to solve the problem **P3**. The levels one (L1) to three (L3) are executed with the assumption of no interference and equal power allocation based on the CNR information. In levels four (L4) and five (L5), the power level of each SC is adjusted using the NOMA SIC principle [8]. The NOMA SIC method helps to mitigate intra-cell interference. For inter-cell interference, we assume that the centralized cloud controller in the C-RAN applies the enhanced inter-cell interference cancellation (eICIC) method. The eICIC method is used to mitigate interference among the cells in heterogeneous networks, and is easily adaptable in C-RAN environment with the advancement of software defined cloud controller [2].

The ILPA method works as follows:

L1: After the CB association parameter ($Q = [q^b_{u,j}]$) is obtained from CA, the CB-user power is equally divided based on the total number of CBs available in each SBS, i.e., $P_{u,j} = \frac{P_{max}}{|J|}$.

L2: The SC power for each user is allocated based on the CB design information ($F = [f_{k,j}]$), CB association parameter and CNR information. Suppose that CB association parameters are obtained after executing the TASCBA method. An example is shown in Fig. 7.5 illustrating the case where the total number of SCs is $K = 4$, the nonzero elements of each SC is $N = 2$ and the total number of CBs is $J = 6$. Accordingly, the maximum number of users supported by the system is $U_b = 6$. Let the CB-user association matrix be as follows:

$$Q = \begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \\ U_6 \end{array} \begin{array}{c} \begin{array}{cccccc} J_1 & J_2 & J_3 & J_4 & J_5 & J_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

According to the CB-user association matrix and the CNR information in (7.9), the weight of CNR for each SC in a CB is estimated as

$$w^j_{k,u} = \frac{g_{k,u}}{\sum_{\forall k \in \{S_{J_j}\}} g_{k,j}} \tag{7.15}$$

and associated with each SC as

$$\begin{array}{c|cccccccc}
 & J_1 & & J_2 & & .. & .. & J_6 & \\
 & \{K_1, & K_2\} & \{K_1, & K_3\} & .. & .. & \{K_3, & K_4\} \\
\hline
U_1 & 0 & 0 & w^2_{1,1} & w^2_{3,1} & .. & 0 & 0 & \\
U_2 & 0 & 0 & 0 & 0 & .. & 0 & 0 & \\
U_3 & w^1_{1,3} & w^1_{2,3} & 0 & 0 & .. & 0 & 0 & \\
.. & .. & .. & .. & .. & .. & .. & .. & \\
U_6 & 0 & 0 & 0 & 0 & .. & .. & .. &
\end{array}.$$

Similar to [82], we utilize the weight-based power allocation, where the summation of all SC weights of each CB is equal to one, i.e.,

$$\sum_{\forall k \in \{S_{J_j}\}} w^j_{k,u} = 1.$$

The power in **L2** is estimated as

$$P_{u,k} = P_{u,j} \times w^j_{k,u}. \tag{7.16}$$

L3: Each SC power is estimated as:

$$P_k = q_{u,j} \sum_{j \in \{S_{\mathcal{K}_k}\}} P_{u,k}, \tag{7.17}$$

where

$$\textbf{P3.1:} \min_{P_{u,k}} \quad P_{u,j} \tag{7.18}$$
$$\text{subject to:}$$
$$\text{C4:} q_{u,j} \sum_{k \in \{S_{J_j}\}} P_{u,k} \leq P_{u,j},$$
$$\text{C6:} P_{u,k} \geq 0$$

$$\textbf{P3.2:} \min_{P_{u,j}} \quad P_T \tag{7.19}$$
$$\text{subject to:}$$
$$\text{C3:} \sum_{j=1}^{J} P_{u,j} \leq P_{max},$$
$$\text{C6:} P_{u,j} \geq 0.$$

L4: In a single cell scenario, the problem **P3** is divided into two subproblems **P3.1** and **P3.2** as shown above. The problem **P3.1** is regarded as SC-user power allocation, whereas **P3.2** is regarded as CB-user power allocation. An iterative geometric water filling (GWF) method is applied to solve the problem **P3.1**. However, the main challenge is that when the SC power (P_k) is allocated among users, they may not be causing minimum interference to each other. Thus, we consider the NOMA SIC principle, where the SC power is allocated in such a way that users with highest CNR get the lowest power.

To implement this approach, we define the step depth in GWF in such a way that the highest CNR is represented with the highest depth, so that the lowest power is allocated to it. Therefore, we represent the step depth, $d_{k,u}$, as the inverse of normalized CNR as

$$d_{k,u} = \frac{\sum_{j \in \{S_{K_k}\}} g_{k,u}}{g_{k,u}}. \tag{7.20}$$

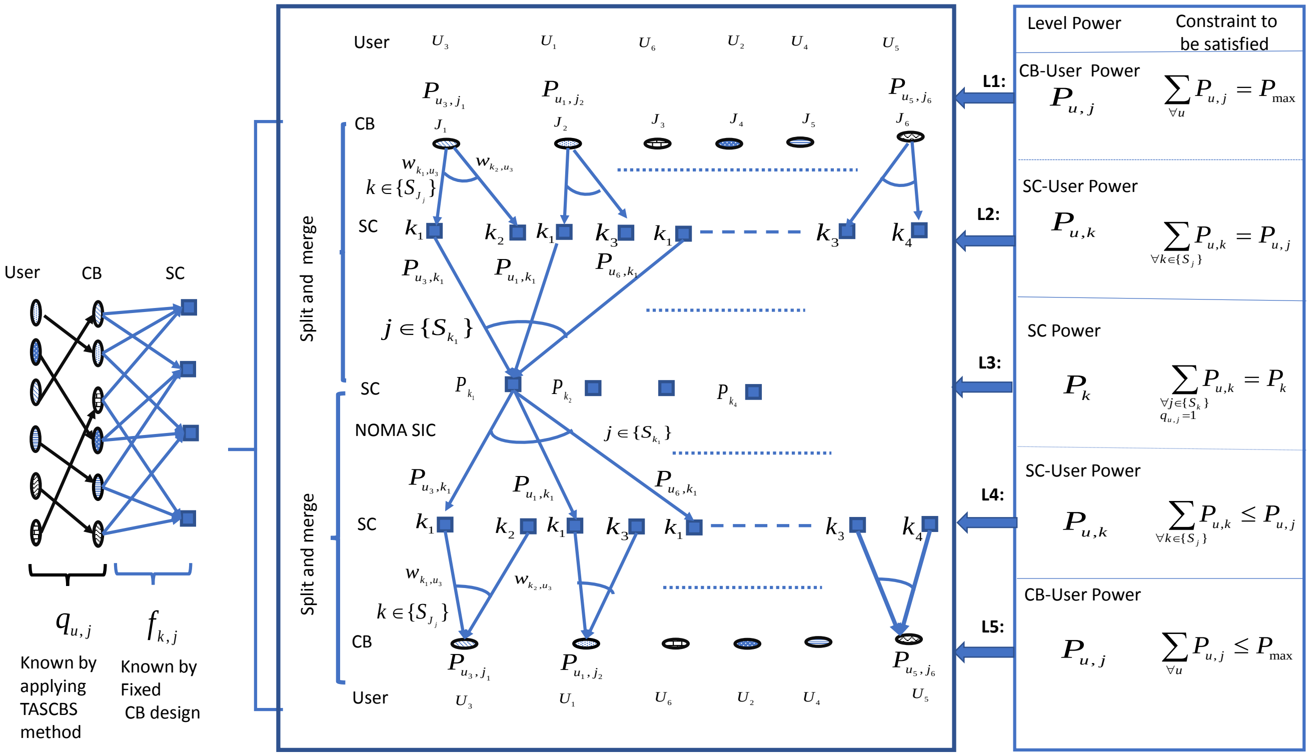

User
CB
SC
$q_{u,j}$
Known by applying TASCBS method
$f_{k,j}$
Known by Fixed CB design
Split and merge
NOMA SIC
Split and merge
User
CB
SC
SC
SC
CB
User
U_3 U_1 U_6 U_2 U_4 U_5
P_{u_3,j_1} P_{u_1,j_2} P_{u_5,j_6}
J_1 J_2 J_3 J_4 J_5 J_6
w_{k_1,u_3} w_{k_2,u_3}
$k \in \{S_{J_j}\}$
k_1 k_2 k_1 k_3 k_1 k_3 k_4
P_{u_3,k_1} P_{u_1,k_1} P_{u_6,k_1}
$j \in \{S_{k_1}\}$
P_{k_1} P_{k_2} P_{k_4}
P_{u_6,k_1}
L1:
L2:
L3:
L4:
L5:
Level Power
Constraint to be satisfied
CB-User Power
$P_{u,j}$
$\sum_{\forall u} P_{u,j} = P_{\max}$
SC-User Power
$P_{u,k}$
$\sum_{\forall k \in \{S_j\}} P_{u,k} = P_{u,j}$
SC Power
P_k
$\sum_{\substack{\forall j \in \{S_k\} \\ q_{u,j}=1}} P_{u,k} = P_k$
SC-User Power
$P_{u,k}$
$\sum_{\forall k \in \{S_j\}} P_{u,k} \le P_{u,j}$
CB-User Power
$P_{u,j}$
$\sum_{\forall u} P_{u,j} \le P_{\max}$

Figure 7.5: Iterative level-based power allocation in a single cell where $K = 4$, $N = 2$ and $J = 6$.

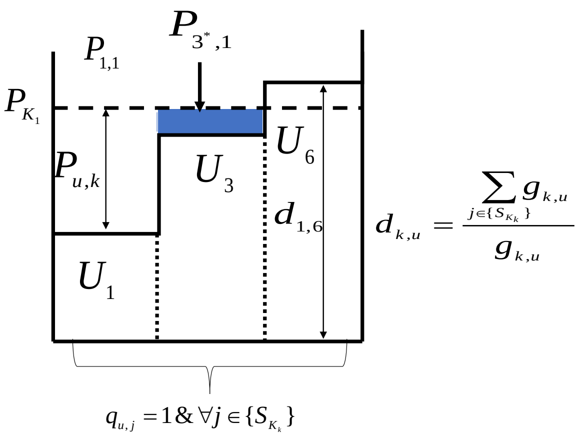

$P_{3^*,1}$
$P_{1,1}$
P_{K_1}
$P_{u,k}$
U_1
U_3
U_6
$d_{1,6}$
$d_{k,u} = \frac{\sum_{j \in \{S_{K_k}\}} g_{k,u}}{g_{k,u}}$
$q_{u,j} = 1 \& \forall j \in \{S_{K_k}\}$

Figure 7.6: NOMA SIC aware GWF procedure.

Fig. 7.6 shows the NOMA SIC aware GWF procedure. Using (7.20), the inverse of normalized CNR is represented as the largest depth, hence, we can apply the GWF method to find the explicit solution of (**P3.1**) in a computationally efficient way. According to [88], the explicit solution of (**P3.1**) is:

$$P_{u,k} = \begin{cases} P_{l^*,k} + (d_{l^*} - d_u), & \text{if } 1 < u \leq l^* \\ 0, & l^* < u < |\{S_{\mathcal{K}_k}\}|. \end{cases} \tag{7.21}$$

Here, l^* denotes the maximum water level and P_{k,l^*} denotes the allocated power in l^* level.

L5: Finally, the user adjusts the power level to its associated CB as follows:

$$P_{u,j} = \sum_{k \in \{S_{J_j}\}} P_{u,k}$$

Example: To compute **L1** to **L5** power of ILPA method, we choose the case of Fig. 7.5 where the total number of SCs is $K = 4$, the nonzero elements of each SC is $N = 2$ and the total number of CBs is $J = 6$. Accordingly, the maximum number of users supported by the system is $U_b = 6$. Assume that the total maximum power budget of the base station is $P_{max} = 12$dB. Let the CB-user association matrix be as follows:

$$Q = \begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \\ U_6 \end{array} \begin{array}{c} \begin{array}{cccccc} J_1 & J_2 & J_3 & J_4 & J_5 & J_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{7.22}$$

and the CNR information be as follows:

$$CNR = \begin{array}{c} \\ K_1 \\ K_2 \\ K_3 \\ K_4 \end{array} \begin{array}{c} \begin{array}{cccccc} U_1 & U_2 & U_3 & U_4 & U_5 & U_6 \end{array} \\ \begin{pmatrix} .7 & .2 & .6 & .2 & .1 & .8 \\ .2 & .8 & .7 & .8 & .2 & .2 \\ .5 & .5 & .5 & .2 & .5 & .2 \\ .4 & .3 & .2 & .5 & .6 & .7 \end{pmatrix} \end{array}. \tag{7.23}$$

According to the CB-user association matrix (7.22) and the CNR information in (7.23), we can apply (7.15) to estimate the weight of CNR for each SC in a CB is as follows:

$$\begin{array}{c|cccccccccccc} & J_1 & & J_2 & & J_3 & & J_4 & & J_5 & & J_6 & \\ & \{K_1, & K_2\} & \{K_1, & K_3\} & \{K_1, & K_4\} & \{K_2, & K_3\} & \{K_2, & K_4\} & \{K_3, & K_4\} \\ \hline U_1 & 0 & 0 & w^2_{1,1} = .58 & w^2_{3,1} = .42 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ U_2 & 0 & 0 & 0 & 0 & 0 & 0 & w^4_{2,2} = .62 & w^4_{3,2} = .38 & 0 & 0 & 0 & 0 \\ U_3 & w^1_{1,3} = .46 & w^1_{2,3} = .54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ U_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w^5_{2,4} = .62 & w^5_{4,4} = .38 & 0 & 0 \\ U_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w^6_{3,5} = .45 & w^6_{4,5} = .55 \\ U_6 & 0 & 0 & 0 & 0 & w^3_{1,6} = .53 & w^3_{4,6} = .47 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}.$$

L1: The L1 power are estimated as $P_{u,j} = \frac{P_{max}}{|J|} = \frac{12}{6} = 2$dB. According to the CB-User association matrix in (7.22), the CB-user power is $P_{U_1,J_2} = P_{U_2,J_4} = P_{U_3,J_1} = P_{U_4,J_5} = P_{U_5,J_6} = P_{U_6,J_3} = 2$dB

L2: Apply weight based SC-user power in (7.16), we can compute L2 power which satisfy the constraint C3 and C4 in **P1** as follows:

$$\begin{array}{c|cccccccccccc} & J_1 & & J_2 & & J_3 & & J_4 & & J_5 & & J_6 & & Constraint \\ & \{K_1, & K_2\} & \{K_1, & K_3\} & \{K_1, & K_4\} & \{K_2, & K_3\} & \{K_2, & K_4\} & \{K_3, & K_4\} & C4 : \sum_{k \in \{S_{J_j}\}} P_{u,k} \leq P_{u,j} \\ \hline U_1 & 0 & 0 & P_{1,1} = 1.16 & P_{1,3} = .84 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2dB \\ U_2 & 0 & 0 & 0 & 0 & 0 & 0 & P_{2,2} = 1.24 & P_{2,3} = .76 & 0 & 0 & 0 & 0 & 2dB \\ U_3 & P_{3,1} = .92 & P_{3,2} = 1.08 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2dB \\ U_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_{4,2} = 1.24 & P_{4,4} = .76 & 0 & 0 & 2dB \\ U_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & P_{5,3} = .9 & P_{5,4} = 1.1 & 2dB \\ U_6 & 0 & 0 & 0 & 0 & P_{6,1} = 1.06 & P_{6,4} = .94 & 0 & 0 & 0 & 0 & 0 & 0 & 2dB \\ \hline & & & & & & & & & & & C3 : \sum_{j=1}^{J} P_{u,j} \leq P_{max} & & = 12dB \end{array}.$$

L3: Apply (7.17), we can compute SC power as follows:

$P_{K_1} = P_{3,1} + P_{1,1} + P_{6,1} = .92 + 1.16 + 1.06 = 3.14$dB

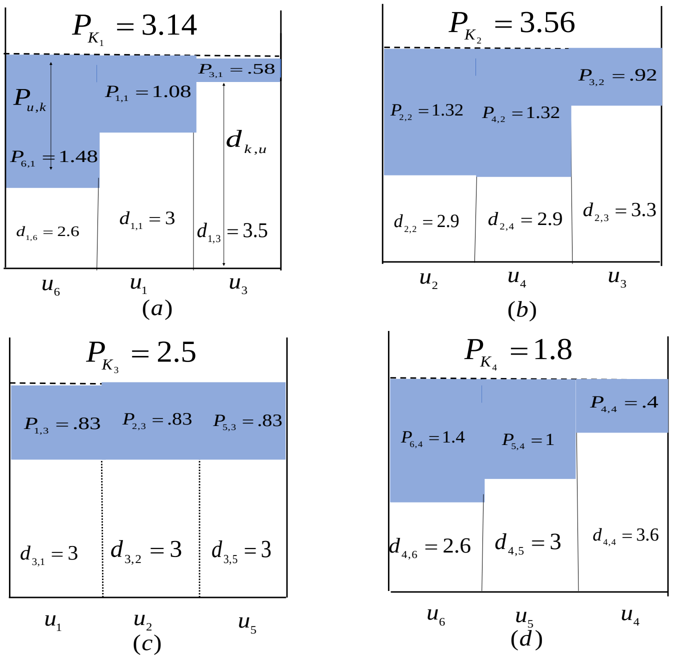


Figure 7.7: apply NOMA-SIC based GWF to estimate SC-user power in L4, (a)$P_{K_1} = 3.14$ dB (b)$P_{K_2} = 3.56$ dB (c)$P_{K_3} = 2.5$ dB (d)$P_{K_4} = 2.8$ dB.

$$P_{K_2} = P_{3,2} + P_{2,2} + P_{4,2} = 1.08 + 1.24 + 1.24 = 3.56\text{dB}$$

$$P_{K_3} = P_{1,3} + P_{2,3} + P_{5,3} = .84 + .76 + .9 = 2.5\text{dB}$$

$$P_{K_4} = P_{6,4} + P_{4,4} + P_{5,4} = 0.94 + 0.76 + 1.1 = 2.8\text{dB}$$

$$P_{K_1} + P_{K_2} + P_{K_3} + P_{K_4} = 12\text{dB}$$

L4: Using SC power in L3, we can compute the SC-user power in L4 by utilizing NOMA-SIC based GWF method. Fig. 7.7 shows the results of the SC-user power which utilizes the calculation of (7.20)(7.21).

L5: Using the L4 SC-user power information, the CB-user power is estimated as:

$$P_{U_1,J_2} = P_{U_1,K_1} + P_{U_1,K_3} = 1.08 + .83 = 1.91\text{dB}$$

$$P_{U_2,J_4} = P_{U_2,K_2} + P_{U_2,K_4} = 1.32 + .83 = 2.15\text{dB}$$

$$P_{U_3,J_1} = P_{U_3,K_1} + P_{U_3,K_2} = .58 + .92 = 1.5\text{dB}$$

$$P_{U_4,J_5} = P_{U_4,K_2} + P_{U_4,K_4} = 1.32 + .4 = 1.72\text{dB}$$

$$P_{U_5,J_6} = P_{U_5,K_3} + P_{U_5,K_4} = .83 + 1 = 1.83\text{dB}$$

$$P_{U_6,J_3} = P_{U_6,K_1} + P_{U_6,K_4} = 1.48 + 1.4 = 2.88\text{dB}$$

$$P_{U_1,J_2} + P_{U_2,J_4} + P_{U_3,J_1} + P_{U_4,J_5} + P_{U_5,J_6} + P_{U_6,J_3} = 12\text{dB}$$

7.6 Simulation Results

In this section, the sum data rate and EE performances of the proposed TASCBS and ILPA methods for SCMA supported downlink C-RANs are investigated. In the simulation model, we consider a 120m × 100m area, where one macro base station is underlaid by 3 small cell base stations. The locations of SBSs and SUEs are modeled using spatial Poisson point process (PPP) with predefined intensity values. For CB assignment, we consider the $J = 6$, $K = 4$ and $N = 2$ case for each SBS. The simulation parameters are shown in Table 7.4. The simulations are averaged over 100 trials. For performance evaluations, we initially consider equal PA, and compare SCMA and OMA bandwidth allocation with different user association (UA) schemes. Then we apply the ILPA method for the SCMA system and discuss the effect of PA on the sum data rate and EE. For the bandwidth allocation in OMA, we consider orthogonal SC allocation, whereas in the SCMA method, we consider CB allocation using the TASCBS method. Three different UA schemes, namely, i) location aware, ii) SINR based, and iii) maximum a posteriori (MAP) based schemes [73] [89] are considered for comparison.

Fig. 7.8 shows the convergence behavior of the TASCBS method. When the number of conflict vertices becomes zero, the TASCBS method results in a stable CB assignment for all the users. Here, we consider $B = 3$ SBSs with $J = 6$ CBs in each SBS, where the total number of users in each SBS is $U_b = 6$ and the total number of users in the network is $U = 18$. Note that the TASCBS method is applied to each SBS separately. It can be observed from Fig. 7.8 that initially the number of conflict vertices becomes high. Gradually, the number of conflict vertices decreases with increasing number of user-CB pairs. When the total number of stable user-CB pairs is 18, there are no more conflict vertices.

Fig. 7.9 shows the performance of sum data rate versus the number of small cell users for different user association methods in OMA and SCMA, considering equal PA. It can be observed that the sum data rate performance of SCMA with location aware UA method gives the best result among others. This can be explained by the nonorthogonal allocation of CBs and the users

Table 7.4: Simulation parameters

Parameters	**Values**
Total no. of small cells, B	3
Total no. SUEs in each SBS, U_b	6
Total no. of SCs, K	4
Total no. of CBs, J	6
The nonzero elements in each codeword, N	2
Max no. of users in each SC, d_f	3
Max no. of SCs are used by each users, d_v	2
Factor graph F	$[1, 1, 1, 0, 0, 0;$ $1, 0, 0, 1, 1, 0;$ $0, 1, 0, 1, 0, 1;$ $0, 0, 1, 0, 1, 1]$
Radius of small cell	10 m
Minimum data rate requirements, r_{min}	50-140 kbps
Maximum power of SBS, P_{max}	30 dBm
Path-loss exponent	4
Noise power spectrum density	-144 dBm/Hz

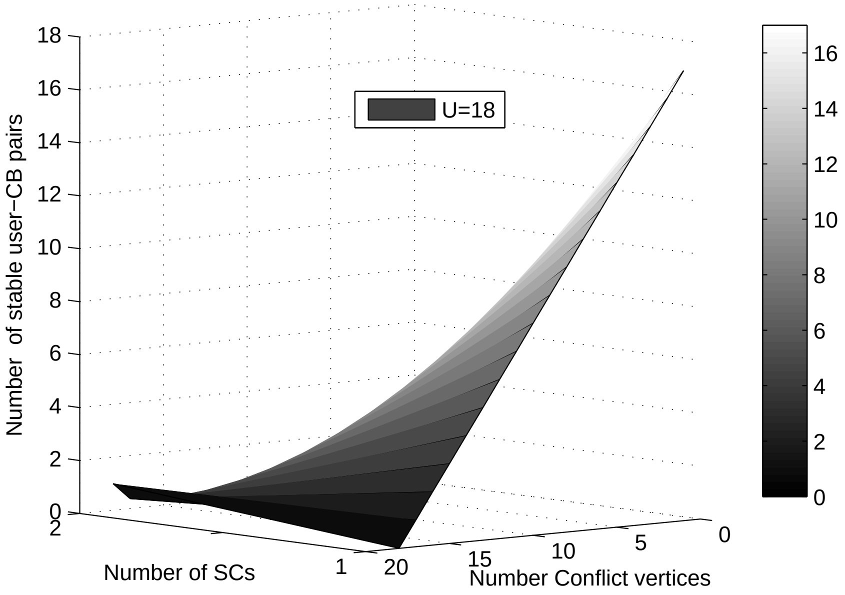


Figure 7.8: Number of stable user-CB pairs versus number of conflict vertices in TASCBS method, where $B = 3$, $J = 6$, $K = 4$ and $N = 2$, showing the convergence of TASCBS method.

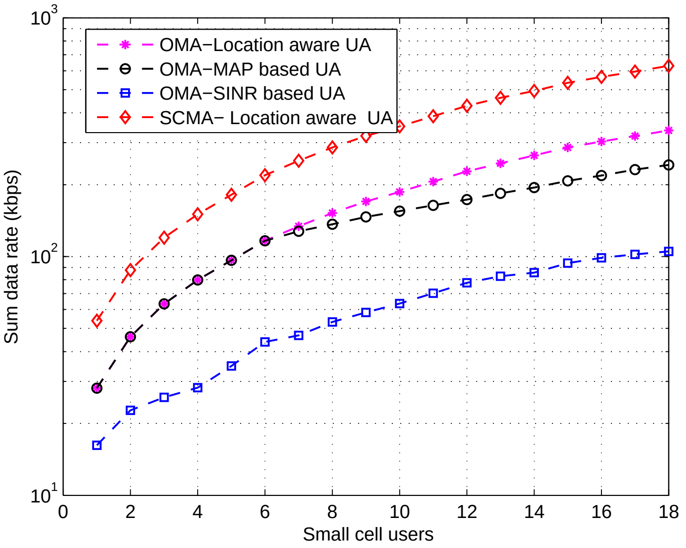


Figure 7.9: Performance of sum data rate with equal PA, where total number of SBS is $B = 3$ and each SBS supports $U_b = 6$ SUEs, $J = 6$ CBs, and $K = 4$ SCs.

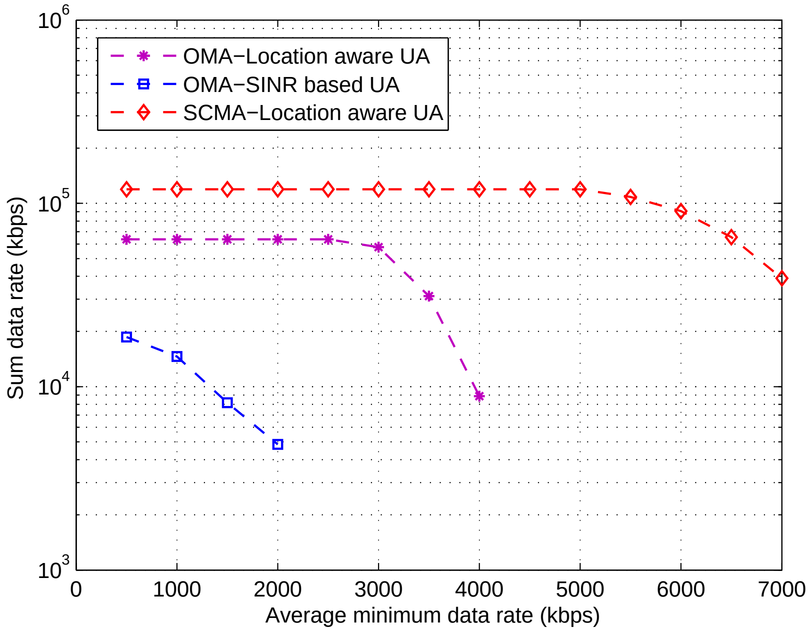


Figure 7.10: Performance of sum rate with different minimum data rate requirements.

not causing significant interference in the same SC. In OMA with location-aware UA method, users are associated with closer proximity to the base station, showing a similar performance with respect to the MAP based method. Since the MAP based UA works on the maximum received CSI information, users in close proximity to the base station receive the maximum CSI information.

7.6.1 Performance of SCMA with Minimum Data Rate Requirements

Fig. 7.10 shows the sum data rate performances of SCMA and OMA for different minimum data rate requirements. Note that the minimum data rate requirement is an important factor for choosing the multiple access and user association methods. When the minimum data rate requirement is increased, the OMA-SINR based UA method can only support fewer number of users compared to the location-aware UA method. For the SCMA location-aware UA, each user uses $d_v = 2$ SCs and each SC is shared by $d_f = 3$ users in a nonorthogonal way. Hence, it can accommodate increased number of users while satisfying the minimum data rate constraint.

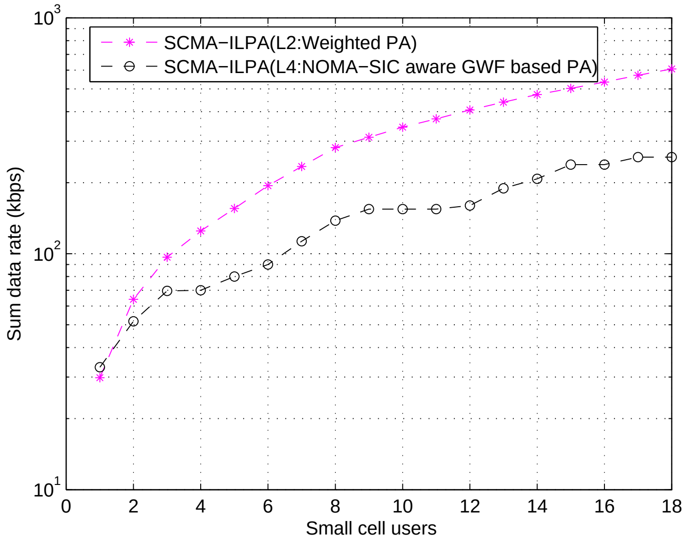


Figure 7.11: Sum data rate performance of SCMA with ILPA.

On the other hand, in OMA with location-aware and SINR-based UA methods, each user can choose only one orthogonal SC based on the relative distance or average received signal strength. Therefore, fewer users are supported and the sum data rate is lower compared to SCMA.

7.6.2 Performance of SCMA with ILPA Method

The sum data rate performance of SCMA method is compared for two power allocation approaches in the ILPA method; i) weighted PA in L2, and ii) NOMA-SIC aware GWF based PA in L4. In the weighted PA, the SC power for each user is allocated based on the CB design information, CB association parameter and CNR information. The sum data rate performance of the weighted PA is better than the NOMA SIC aware PA (L4 PA) as shown in Fig. 7.11. According to the CB design, each user uses 2 SCs and the weighted PA utilizes more power on the better channel. On the other hand, the level four in ILPA method allocates power to SCs according to the NOMA SIC principle. According to this principle, among the 3 users in the same SC, the one which has the highest channel gain utilizes less power to avoid intra-SC interference, hence, the sum data rate performance of this level becomes less than the weighted PA. However, L4 PA uses less power for each user, which makes it suitable for EE. The EE performance of L4 PA increases with the number of users and L4 PA shows better performance

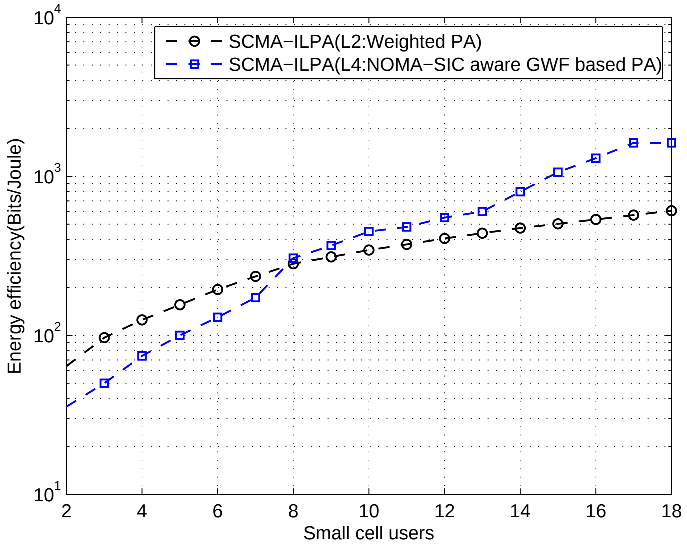


Figure 7.12: EE performance of SCMA with ILPA.

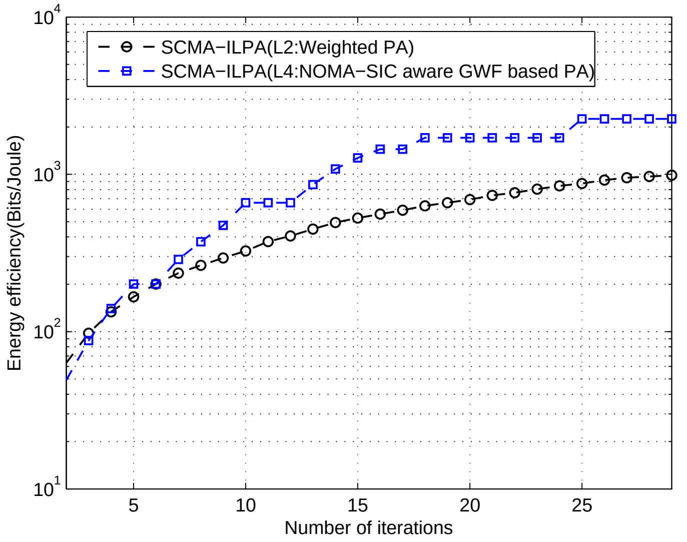


Figure 7.13: Convergence behavior of ILPA.

compared to the weighted PA as shown in Fig. 7.12. Therefore, it can be concluded that the EE performance of SCMA becomes more significant when the NOMA SIC is incorporated into the SCMA PA. Finally, Fig. 7.13 shows the convergence behavior of L2 and L4 of the ILPA method, where $J = 6$, $K = 4$ and $N = 2$. It is apparent from the figure that the weighted PA in L2 and NOMA-SIC aware GWF based PA in L4 show non-decreasing energy efficiency and converge within a limited number of iterations.

7.7 Summary

In this chapter, we considered the EE performance of SCMA supported C-RANs. We implemented the SCMA method to jointly optimize the codebook and power allocation in the downlink of C-RANs. To solve the optimization problem, we proposed the throughput aware SCMA codebook selection and iterative level-based power allocation methods. From the implementation perspective, the software defined cloud controller executes the TASCBS at each SBS, which results in a stable codebook allocation solution within a finite number of steps. After obtaining the codebook allocation solution, the ILPA method helps to optimize the power allocation for the whole C-RAN. Simulation results show that EE performance of NOMA SIC aware GWF based power allocation becomes better than the weighted power allocation scheme with increasing number of users. On the other hand, the weighted power allocation scheme provides higher sum data rate for increasing number of users. The sum data rate and EE performances mainly depend on dynamic codebook design, which is a future research topic of investigation. The proposed approaches TASCBS and ILPA are important for practical implementation of SCMA supported downlink C-RANs.

Chapter 8

Conclusions and Future Work

The ever increasing number of mobile users, smart devices and applications increase the device density and mobile traffic loads, and eventually make the traditional cellular networks incapable of handling such high demand with high quality of service. Therefore, researchers and engineers, both in academia and industry, have been working on new technological and architectural solutions for the next generation cellular networks, e.g., in 5G to handle 100 times more traffic and user loads, 1000 times higher network capacity and 1 ms latency. Moreover, spectrum efficiency (SE) and energy efficiency (EE) are the main focus in future wireless networks. The cloud radio access network (C-RAN) is regarded as a promising solution to utilize small cells in cellular networks along with macro cells to improve the total network capacity due to the spatial frequency reuse achieved by the flexible centralization of small cells through cloud computing. The C-RAN consisting of remote radio heads (RRHs), the centralized baseband unit (BBU) pool and optical or wireless fronthaul links is regarded as a new architectural paradigm that provides centralized and coordinated multi-point (CoMP) processing solutions to achieve higher capacity, SE, EE, seamless coverage, network control and cost efficient operation. In C-RANs, the macro base stations (MBS) provide seamless coverage and network control, whereas the small cell base stations, regarded as RRHs, provide users with high data rates while leaving the basic control operations such as interference and handover management to macro cells. In addition, all the BBUs in a BBU pool are considered as computing servers to perform baseband signal processing through cloud computing technologies. However, user association, cell activation, dynamic resource allocation based on users QoS requirements, workload scheduling in BBU pool, BBU-RRH mapping etc. are the major challenging issues in C-RANs. In this thesis, we considered user association, communication and computing resource, BBU-RRH mapping problems in C-RANs and proposed effective solutions to best leverage the achievable data rate, EE and minimize delay in C-RANs.

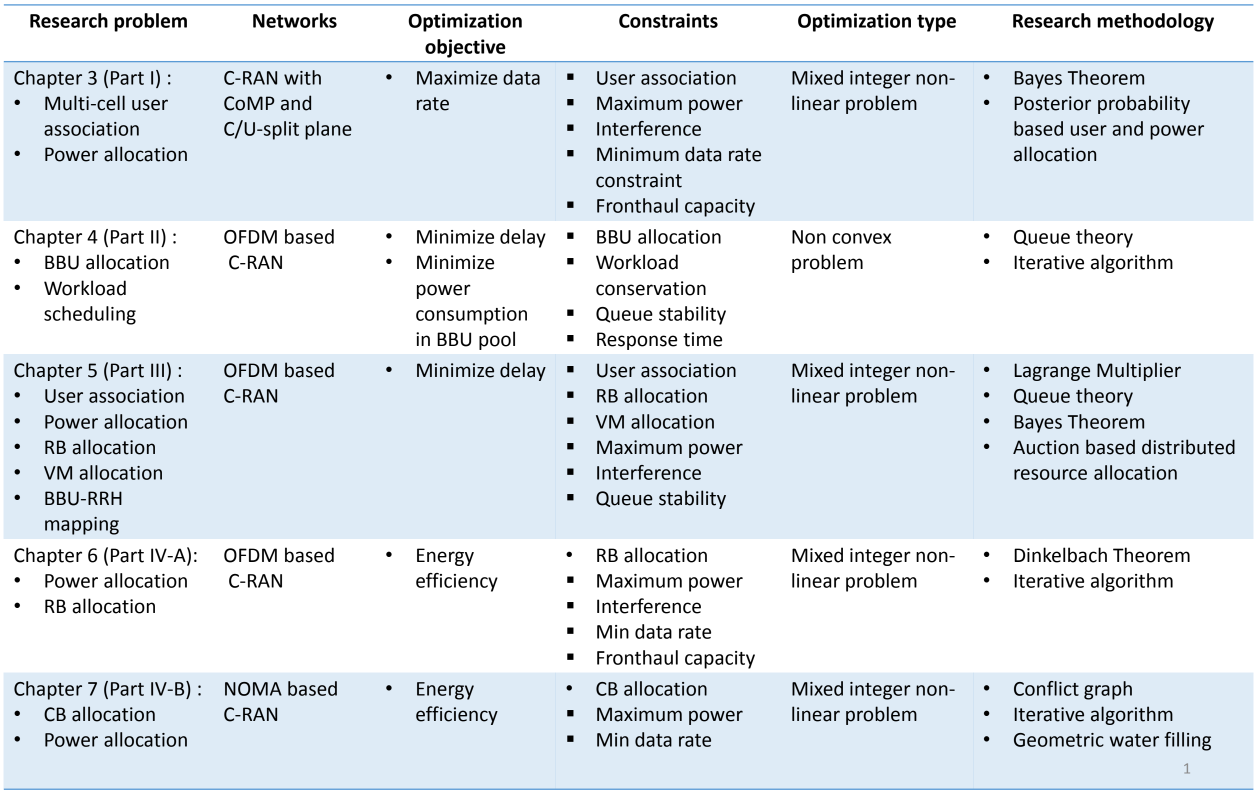
8.1 Conclusions

In this thesis, we focused on the resource management techniques for cloud radio access networks to increase data rate, energy efficiency and minimize delay. In our research, we mainly focused on two important resource management issues for C-RANs which are communication (i.e., resource blocks, codebooks, power) and computing (i.e., BBU servers, VMs) resources in the presence of practical limitations. On the other hand, user association, cell on-off, workload scheduling in BBU pool, RRH clustering etc. are the challenging problems in C-RAN. Among these problems, user/cell association, power and bandwidth allocation are the major challenging problems to optimize resources in the radio access part of C-RANs.

In chapter 3, we addressed multi-cell user association for CoMP supported C-RANs, taking into account of data rate and aggregated interference of mobile users. We proposed the posterior probability based user association and power allocation (P2UPA) method that depends on prior knowledge of the channel state information (CSI). The objective of the proposed method is to maximize the sum data rate of small cell users while maintaining the constraints of aggregated interference, power consumption, and data rate among small cell users. The sum data rate and energy efficiency performance of P2UPA are evaluated through simulations.

In chapter 4, joint BBU allocation and workload scheduling among BBU servers were analysed in terms of queueing theory with the aim to minimize mean response time and aggregate power. Queueing stability and workload conservation constraints are considered in this optimization problem. To solve this problem, we propose an energy efficient joint workload scheduling and BBU allocation (EE-JWSBA) algorithm. The EE-JWSBA algorithm is evaluated via simulations by considering three different scheduling weights (e.g., random, normalized, and upper limit). Simulation results demonstrate the effectiveness of proposed scheme using different scheduling weights. The summary of these works are depicted in the Table 8.1.

Table 8.1: Research contributions



Research problem	Networks	Optimization objective	Constraints	Optimization type	Research methodology
Chapter 3 (Part I) : • Multi-cell user association • Power allocation	C-RAN with CoMP and C/U-split plane	• Maximize data rate	▪ User association ▪ Maximum power ▪ Interference ▪ Minimum data rate constraint ▪ Fronthaul capacity	Mixed integer non-linear problem	• Bayes Theorem • Posterior probability based user and power allocation
Chapter 4 (Part II) : • BBU allocation • Workload scheduling	OFDM based C-RAN	• Minimize delay • Minimize power consumption in BBU pool	▪ BBU allocation ▪ Workload conservation ▪ Queue stability ▪ Response time	Non convex problem	• Queue theory • Iterative algorithm
Chapter 5 (Part III) : • User association • Power allocation • RB allocation • VM allocation • BBU-RRH mapping	OFDM based C-RAN	• Minimize delay	▪ User association ▪ RB allocation ▪ VM allocation ▪ Maximum power ▪ Interference ▪ Queue stability	Mixed integer non-linear problem	• Lagrange Multiplier • Queue theory • Bayes Theorem • Auction based distributed resource allocation
Chapter 6 (Part IV-A): • Power allocation • RB allocation	OFDM based C-RAN	• Energy efficiency	• RB allocation ▪ Maximum power ▪ Interference ▪ Min data rate ▪ Fronthaul capacity	Mixed integer non-linear problem	• Dinkelbach Theorem • Iterative algorithm
Chapter 7 (Part IV-B) : • CB allocation • Power allocation	NOMA based C-RAN	• Energy efficiency	• CB allocation ▪ Maximum power ▪ Min data rate	Mixed integer non-linear problem	• Conflict graph • Iterative algorithm • Geometric water filling

In chapter 5, we addressed the joint communication and computing resource allocation problem along with user association, and baseband unit and remote radio head mapping in C-RANs. We initially established a queueing model in C-RAN, followed by formulations of two optimization problems for communication (e.g., resource blocks and power) and computing (e.g., virtual machines) resources allocation with the aim to minimize mean response time. User association along with the resource block allocation, interference and queueing stability constraints are considered in the communication resource optimization problem. The computing resource optimization problem considered BBU-RRH mapping and virtual machines allocation for small cells, constrained to BBU server capacity and queueing stability. To solve the communication and computing resource optimization problem, we proposed a joint resource allocation solution using a double-sided auction based distributed resource allocation (DS-ADRA) method, where small cell base stations and users jointly participate using the concept of auction theory. The proposed method is evaluated via simulations by considering the effect of bandwidth utilization percentage, signal-to-interference ratio threshold value and number of users.

In chapter 6, we reviewed the communication resource allocation problem in C-RANs in the perspective of energy efficiency. We established an energy efficient resource allocation in C-RANs, followed by the formulation of an optimization problem for communication (e.g., resource blocks and power) resources allocation considering both small and macro cell users with the aim to maximize energy efficiency in C-RANs. The resource block along with power allocation, interference, quality of service of macro cell users, and front-haul capacity constraints are considered in this optimization problem. Due to the joint nature of resource block and power allocation both in macro and small cell users, turned the resource allocation problem into a computationally intractable and NP-hard problem. To solve this optimization problem, we transferred the baseline problem into a relaxed problem with the time-sharing approach of resource allocation, and proposed an iterative resource allocation solution using the Dinkelbach theorem. The proposed method is evaluated in terms of energy efficiency, sum data rate and Jain fairness index considering the effect of number of user association in C-RAN.

Different from the above works, in chapter 7, we considered non-orthogonal multiple access based C-RAN systems. The non-orthogonal multiple access scheme is regarded as an attractive solution to support multi-user resource sharing in order to improve spectrum and energy efficiency in 5G wireless networks. In this chapter, among various NOMA schemes, we considered the sparse code multiple access scheme to jointly optimize the codebook and power allocation in the downlink of the C-RANs. To solve the NP-hard joint optimization problem, we decomposed the original problem into two sub-problems: codebook allocation and power allocation. Using the conflict graph, we proposed the throughput aware sparse code multiple access based codebook selection (TASCBS) method, which generates a stable codebook allocation solution within a finite number of steps. For the power allocation solution, we propose the iterative level-based power allocation (ILPA) method, which incorporates different power allocation approaches (e.g., weighted and NOMA successive interference cancellation) into different levels to satisfy the maximum power requirement. Simulation results showed that the sum data rate and energy efficiency performances of sparse code multiple access supported C-RANs depend on the selected power allocation approach. In terms of energy efficiency, the performance significantly increases with the number of users when the NOMA-SIC aware geometric water-filling based power allocation is used.

8.2 Future Work

Throughout this thesis, we proposed several algorithms that contributed to the efficient resource management for cloud radio access networks. However, there are some relevant and recent issues that warrant further consideration in the future work. For instance, 3GPP release 15 in 2018 and ITU has divided 5G network services into three categories [90]:

i) enhanced Mobile Broadband (eMBB) or handsets;

ii) Ultra-Reliable Low-Latency Communications (URLLC), which includes industrial applications and autonomous vehicles; and

iii) Massive Machine Type Communications (mMTC) or sensors.

Initial 5G deployments with cloud radio access networks focus on eMBB and fixed wireless, which makes use of many of the same capabilities as LTE/LTE-A. However, there are several research scopes to investigate the resource management in the C-RANs in the following areas:

- **C-RANs with URLLC:** The main challenge of ultra-reliable low-latency communication is 1 msec one-way latency in the radio access network with 99.999 percent reliability [91].

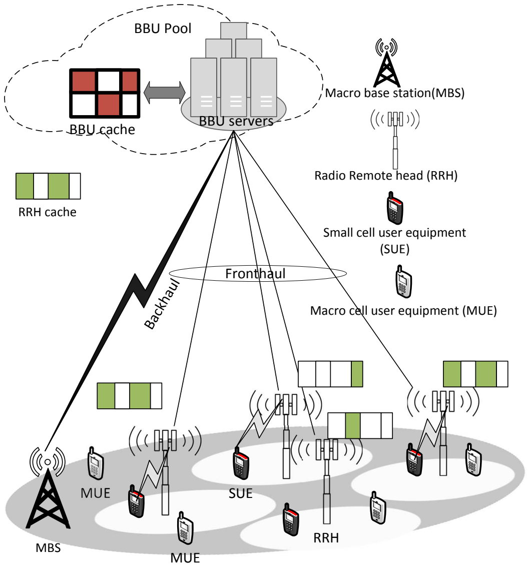


Figure 8.1: Cache enabled C-RAN architecture for 5G networks.

C-RANs with optical front-haul connection becomes a viable option for URLLC. Authors in [92] studied C-RANs with multi-cell scheduling algorithms to overcome the challenges for supporting URLLC in the 5G new radio (5G NR) networks. The 5G NR is designed to support the IMT 2020 requirements, being able to support a diverse set of services with different characteristics and quality-of-service (QoS) targets. Similarly, we can incorporate the 5G new radio framework to our C-RAN queueing model by studying the distributed resource management procedures to overcome some of the challenges for supporting URLLC.

- **C-RANs with mMTC and D2D:** According to IMT-2020 5G specifications, total number of devices per unit area is $106/km^2$ for mMTC traffic [91] [93]. To handle such connection density, NOMA enabled C-RAN with software define network will be the suitable choice. To manage massive access and resource allocation in C-RAN for mMTC will have a significant impact on future research for capacity enhancement. Enhancing C-RAN with device-to-device (D2D) network has been studied in [94]. In D2D based C-RANs, the challenges are to solve the problem of jointly optimizing communication and computing resources. We have studied the joint optimization of communication computing resources in C-RANs in chapter 5. We can extended our work to investigate the techniques and methodologies applied in resource allocation procedure to support D2D communication underlaid by C-RAN networks.

- **C-RANs with caching and edge computing:** C-RANs with caching capacities can store data in RRHs and BBU pool. Fig. 8.1 shows the C-RAN with caching architecture. RRHs with large storage capabilities proactively downloading the data that will be requested with a high probability and then caching it. When the requests for the cached data arrive, they will be served by RRHs, which means the requested data need not be fetched via fronthaul links. Similarly, for edge computing, users can offload data to the nearest RRHs to do the computing tasks. Both caching and edge computing improve quality of experience (QoE) and reduce transmission delay to users since the data is more closer to users. One of the challenging issues in C-RANs with edge caching is RRH association and data fetching strategy. An efficient data fetching strategy should be developed to help to decide where to fetch the data. In CoMP supported C-RANs, one UE is served by multiple RRHs simultaneous. We have studied the multi-cell association problem in chapter 3. We can further exploit edge cache with the RRH association strategy in future.

- **C-RANs with container supported BBU pool:** In the BBU pool, we studied energy efficient VM allocation and workload scheduling among the BBU servers. In term of delay minimization and URLLC communication, we can incorporate container technology in BBU pool. Containers are different from server virtual machine. Each VM can run an OS in an independent environment and take up more space because they need a guest OS to run. On the other hand, containers do not consume as much space because each container shares the host's OS. Moreover, VMs require substantial resource overhead, such as memory, disk and network input/output because each VM runs an independent OS that take longer time to create than containers [95]. Since containers share the OS kernel, only one instance of an OS can run many isolated containers. Therefore, it is potential for allocating containers in BBU pool instead of VMs for minimizing delay to support URLLC communication in C-RANs .

Bibliography

[1] Cisco, "Visual Networking Index," *White paper [Online], Available:www.Cisco.com*, Feb. 2015.

[2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud-RAN for mobile networks - a technology overview," *IEEE Commun. Surveys and Tut.*, vol. 17, pp. 405 – 426, Sep. 2015.

[3] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5g signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Processing Magazine*, vol. 31, pp. 35 – 44, Oct. 2014.

[4] J. Zuo, J. Zhang, C. Yuen, W. Jiang, and W. Luo, "Energy Efficient User Association for Cloud Radio Access Networks," *IEEE Access*, vol. 4.

[5] A. Davydov, G. Morozov, I. Bolotin, and A. Papathanassiou, "Evaluation of joint transmission comp in c-ran based lte-a hetnets with large coordination areas," in *2013 IEEE Globecom Workshops (GC Wkshps)*, pp. 801–806, Dec 2013.

[6] V. N. Ha, L. B. Le, and N. D. o, "Coordinated multipoint transmission design for cloud-rans with limited fronthaul capacity constraints," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 7432–7447, Sept 2016.

[7] G. K. Tran, H. Shimodaira, R. Rezagah, K. Sakaguchi, and K. Araki, "Dynamic cell activation and user association for green 5G heterogeneous cellular networks," pp. 2364 – 8, 2015.

[8] Y. Tao, L. Liu, S. Liu, and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," *China Commun.*, vol. 12, pp. 1–15, Oct 2015.

[9] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *IEEE Vehic. Tech. Conf.*, pp. 1–5, Sep. 2014.

[10] A. Ghaffari, M. Leonardon, Y. Savaria, C. Jego, and C. Leroux, "Improving performance of SCMA MPA decoders using estimation of conditional probabilities," in *IEEE Int. New Circuits and Syst. Conf.*, pp. 21–24, June 2017.

[11] X. Yue, Z. Qin, Y. Liu, S. Kang, and Y. Chen, "A unified framework for non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 66, pp. 5346–5359, Nov. 2018.

[12] T. LeAnh, N. H. Tran, W. Saad, L. Le, D. Niyato, T. Ho, and C. S. Hong, "Matching theory for distributed user association and resource allocation in cognitive femtocell network," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[13] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 823–831, April 2016.

[14] S. Yan, W. Wang, Z. Zhao, and A. Ahmed, "Investigation of cell association techniques in uplink cloud radio access networks," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 8, pp. 1044–1054, 2016.

[15] E. Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy consumption analysis and minimization in multi-layer heterogeneous wireless systems," *IEEE Transactions on Mobile Computing*, vol. 14, pp. 2474–2487, Dec 2015.

[16] K. Wang, W. Zhou, and S. Mao, "Energy efficient joint resource scheduling for delay-aware traffic in cloud-ran," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2016.

[17] J. Li, M. Peng, Y. Yu, and Z. Ding, "Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 9873–9887, Dec 2016.

[18] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic resource allocation for cloud-ran in lte with real-time bbu/rrh assignment," pp. 1–6, May 2016.

[19] 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, pp. 2837–2850, Nov 2016.

[20] M. Peng, Y. Yu, H. Xiang, and H. Poor, "Energy efficient resource allocation optimization for multimedia heterogenegeous cloud radio access networks," *IEEE transactions on Multimedia*, vol. 18, pp. 879– 892, May 2016.

[21] L. Pei, J. Huilin, G. Shen, D. Fei, and P. Zhiwen, "Impact of bs sleeping and user association scheme on delay in ultra-dense networks," in *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*, pp. 1–6, Oct. 2016.

[22] M. Khan, R. Alhumaima, and H. Al-Raweshidy, "Quality of service aware dynamic BBU-RRH mapping in cloud radio access network," *2015 International Conference on Emerging Technologies (ICET). Proceedings*, pp. 1–5, Dec. 2015.

[23] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wubben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a flexible functional split for cloud-RAN networks," *EuCNC 2014 - European Conference on Networks and Communications*, pp. 1–5, June 2014.

[24] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, and S. Gosselin, "Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks," *IEEE journal of lightwave technology*, vol. 34, pp. 1963– 1970, April 2016.

[25] Q. Liu, G. Wu, Y. Guo, Y. Zhang, and S. Hu, "Energy efficient resource allocation for control data separated heterogeneous-cran," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2016.

[26] B. Zhuang, D. Guo, and M. Honig, "Energy-Efficient Cell Activation, User Association, and Spectrum Allocation in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 823 – 831, Apr. 2016.

[27] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA Cloud-RAN of small cells underlaying a macrocell," *IEEE Trans. Mobile Comput.*, vol. 15, pp. 2837–2850, Nov. 2016.

[28] Q. Liu, T. Han, N. Ansari, and G. Wu, "On designing energy-efficient heterogeneous cloud radio access networks," *IEEE Trans. Green Commun. and Net.*, vol. 2, pp. 721–734, Sep. 2018.

[29] G. Tran, H. Shimodaira, R. Rezagah, K. Sakaguchi, and K. Araki, "Dynamic Cell Activation and User Association for Green 5G Heterogeneous Cellular Networks," in *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 2364 – 2368, Sep. 2015.

[30] A. L. V. K. Lau, "Joint BS-User Association, Power Allocation, and User-Side Interference Cancellation in Cell-free Heterogeneous Networks," *IEEE Transactions on Signal Processing*, vol. 65, pp. 335 – 345, Jan. 2017.

[31] M. Kamel, W. Hamouda, and A. Youssef, "Performance analysis of multiple association in ultra-dense networks," *IEEE Transactions on Communications*, vol. PP, no. 99, pp. 1–1, 2017.

[32] M. Rahman, H. Ghauch, S. Imtiaz, and J. Gross, "RRH Clustering and Transmit Precoding for Interference-Limited 5G CRAN Downlink," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 1 – 7, Dec. 2015.

[33] H. Zhang, C. Jiang, J. Cheng, and V. C. M. Leung, "Cooperative Interference Mitigation and Handover Management for Heterogeneous Cloud Small Cell Networks," *IEEE Wireless Communications*, vol. 22, pp. 92 – 99, June 2015.

[34] A. Hajisami and D. Pompili, “DJP: Dynamic Joint Processing for Interference Cancellation in Cloud Radio Access Networks,” in *IEEE 82nd Vehicular Technology Conference*, pp. 1 – 5, Sep. 2015.

[35] Y. Meng, C. Jiang, L. Xu, Y. Ren, and Z. Han, “User Association in Heterogeneous Networks: A Social Interaction Approach,” *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 9982 – 9993, Dec. 2016.

[36] M. Hasan and E. Hossain, “Resource allocation for network-integrated device-to-device communications using smart relays,” in *2013 IEEE Globecom Workshops (GC Wkshps)*, pp. 591–596, Dec. 2013.

[37] H. Zhang, H. Ji, X. Li, K. Wang, and W. Wang, “Energy efficient resource allocation over cloud-ran based heterogeneous network,” *IEEE 7thInternational Conference on Cloud Computing technology and science*, pp. 483– 486, 2015.

[38] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, “Cloudiq: A framework for processing base stations in a data center,” pp. 125–136, 2012.

[39] T. Sigwele, A. Alam, P. Pillai, and Y. Hu, “Evaluating energy-efficient cloud radio access networks for 5G,” *2015 IEEE International Conference on Data Science and Data-Intensive Systems (DSDIS)*, pp. 362 – 369, 2015.

[40] D.Gross, *Fundamentals of Queueing theory, Second Edition.* India: Wiley, 2008.

[41] X. Nan, Y. He, and L. Guan, “Optimal resource allocation for multimedia cloud in priority service scheme,” *IEEE InternationalConference on Distributed Computing Systems*, pp. 268– 277, May 2011.

[42] Z. Tan, C. Yang, J. Song, Y. Liu, and Z. Wang, “Energy consumption analysis of c-ran architecture based on 10g epon front-haul with daily user behaviour,” *IEEE International Conference on Optical Communications and Networks (ICOCN)*, July 2015.

[43] C. H. Lien, Y. W. Bai, M. B. Lin, C. Y. Chang, and M. Y. Tsai, "Web server power estimation,modeling and management," *IEEE International Conference on Networks*, pp. 1–6, Sept. 2006.

[44] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for mobile edge computing," in *Int. Symp. Modeling and Opt. in Mobile, Ad Hoc, and Wireless Networks)*, pp. 1–6, May 2018.

[45] X. Lin, H. Zhang, H. Ji, and V. C. M. Leung, "Joint computation and communication resource allocation in mobile-edge cloud computing networks," in *IEEE Int. Conf. on Network Infrastructure and Digital Content*, pp. 166–171, Sept 2016.

[46] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Joint computation offloading, resource allocation and content caching in cellular networks with mobile edge computing," in *IEEE Int. Conf. on Comm.*, pp. 1–6, May 2017.

[47] Y. Gai, P. Gong, J. Lv, and W. Wu, "Auction-based radio resource allocation for ofdma systems," in *2009 5th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1–4, Sept 2009.

[48] L. Chen, J. Wu, X. X. Zhang, and G. Zhou, "Tarco: Two-stage auction for d2d relay aided computation resource allocation in hetnet," *IEEE Transactions on Services Computing*, vol. PP, no. 99, pp. 1–1, 2018.

[49] X. Zhang, H. Qian, K. Zhu, R. Wang, and Y. Zhang, "Virtualization of 5g cellular networks: A combinatorial double auction approach," in *IEEE Global Communi. Conf.*, pp. 1–6, Dec 2017.

[50] C. Yi, J. Cai, and G. Zhang, "Spectrum auction for differential secondary wireless service provisioning with time-dependent valuation information," *IEEE Trans. on Wireless Commun.*, vol. 16, pp. 206–220, Jan 2017.

[51] M. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for multi-user multi-task mobile cloud," in *IEEE Int. Conf. on Comm.*, pp. 1–6, May 2016.

[52] Q. Liu, T. Han, N. Ansari, and G. Wu, “On designing energy-efficient heterogeneous cloud radio access networks,” *IEEE Trans. on Green Commun. and Net.*, vol. 2, pp. 721–734, Sep. 2018.

[53] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, “Resource allocation and admission control in ofdma-based cloud-ran,” in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2016.

[54] Y. Zhang, C. Lee, D. Niyato, and P. Wang, “Auction approaches for resource allocation in wireless systems: A survey,” *IEEE Communications Surveys Tutorials*, vol. 15, pp. 1020–1041, March 2013.

[55] Y. Zhang, D. Niyato, P. Wang, and E. Hossain, “Auction-based resource allocation in cognitive radio systems,” *IEEE Communications Magazine*, vol. 50, pp. 108–120, Nov. 2012.

[56] M. Hasan and E. Hossain, “Distributed resource allocation in d2d-enabled multi-tier cellular networks: An auction approach,” in *2015 IEEE International Conference on Communications (ICC)*, pp. 2949–2954, June 2015.

[57] G. Athanasiou, P. C. Weeraddana, and C. Fischione, “Auction-based resource allocation in millimeterwave wireless access networks,” *IEEE Communications Letters*, vol. 17, pp. 2108–2111, Nov. 2013.

[58] M. Hasan and E. Hossain, “Distributed resource allocation for relay-aided device-to-device communication: A message passing approach,” *IEEE Trans. on Wire. Commun.*, vol. 13, pp. 6326–6341, Nov 2014.

[59] K. Wang, W. Zhou, and S. Mao, “On joint bbu/rrh resource allocation in heterogeneous cloud-rans,” *IEEE Internet of Things Jour.*, vol. 4, pp. 749–759, June 2017.

[60] F. Kong, X. Sun, V. Leung, and H. b. Zhu, “Delay-optimal biased user association in heterogeneous network,” *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[61] W. Yu and R. Lui, “Dual methods for nonconvex spectrum optimization of multicarrier systems,” *IEEE Trans. on Commun.*, vol. 54, pp. 1310–1322, July 2006.

[62] S. Boyd and L. Vandenberghe, “Convex optimization,” in *New York, Cambridge University Press,Chapter-11*, 2004.

[63] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, “User association for load balancing in heterogeneous cellular networks,” *IEEE Trans. on Wire. Commun.*, vol. 12, pp. 2706–2716, June 2013.

[64] D. C. R. Jain and W. Hawe, “A quantitative measure of fairness and discrimination for resource allocation in shared systems,” in *Digital Equipment Corporation, Tech. Rep. DEC-TR-301*, Sep 1984.

[65] L. Ferdouse, O. Das, and A. Anpalagan, “Auction based distributed resource allocation for delay aware OFDM based Cloud-RAN system,” in *IEEE Global Commun. Conf.*, pp. 1–6, Dec. 2017.

[66] N. Amani, H. Pedram, H. Taheri, and S. Parsaeefard, “Energy-efficient resource allocation in heterogeneous cloud radio access networks via BBU offloading,” *IEEE Trans. on Vehic. Tech.*, vol. 68, pp. 1365–1377, Feb. 2019.

[67] B. Xu, Y. Chen, J. R. Carrion, and T. Zhang, “Resource allocation in energy-cooperation enabled two-tier NOMA HetNets toward green 5G,” *IEEE Journ. Select. Areas Commun.*, vol. 35, pp. 2758–2770, Dec. 2017.

[68] L. Ferdouse, W. Ejaz, A. Anpalagan, and A. M. Khattak, “Joint workload scheduling and BBU allocation in cloud-ran for 5G networks,” in *Proceedings of the Symposium on Applied Computing*, SAC ’17, (New York, NY, USA), pp. 621–627, ACM, 2017.

[69] Z. Chang, Z. Wang, X. Guo, C. Yang, Z. Han, and T. Ristaniemi, “Distributed resource allocation for energy efficiency in ofdma multicell networks with wireless power transfer,” *IEEE Journ. on Select. Areas in Commun.*, vol. 37, pp. 345–356, Feb. 2019.

[70] E. H. Monowar Hasan and D. I. Kim, "Resource Allocation Under Channel Uncertainties for Relay-Aided Device-to-Device Communication Underlaying LTE-A Cellular Networks," *IEEE Transaction on Wireless communication*, vol. 13, pp. 2322–2338, Mar. 2014.

[71] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. 492–498, 1967.

[72] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. on Commun.*, vol. 54, pp. 1310–1322, July 2006.

[73] L. Ferdouse, A. Alnoman, A. Bulzacki, and A. Anpalagan, "Energy efficient multiple association in CoMP based 5G cloud-ran systems," in *IEEE Vehic. Techn. Conf.*, pp. 1–5, Sept. 2017.

[74] X. Chu, D. Lopez-Perez, Y. Yang, and F. Gunnarsson, *Heterogeneous Cellular Networks: Theory, Simulation and Deployment.* New York, NY, USA: Cambridge University Press, 2013.

[75] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," *IEEE Network*, vol. 32, pp. 152–160, March 2018.

[76] L. Ferdouse, A. Anpalagan, and S. Erkucuk, "Joint communication and computing resource allocation in 5G cloud radio access networks," *(under review) IEEE Trans. Vehic. Tech.*, submitted, Jan. 2019.

[77] W. Zhu, L. Qiu, and Z. Chen, "Joint subcarrier assignment and power allocation in downlink SCMA systems," in *IEEE Vehic. Tech. Conf.*, pp. 1–5, Sep. 2017.

[78] M. A. Sedaghat and R. R. Muller, "On user pairing in uplink NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 3474–3486, May 2018.

[79] B. Di, L. Song, and Y. Li, "Radio resource allocation for uplink sparse code multiple access SCMA networks using matching game," in *IEEE Int. Conf. Comm.*, pp. 1–6, May 2016.

[80] R. Ruby, S. Zhong, H. Yang, and K. Wu, "Enhanced uplink resource allocation in non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 1432–1444, Mar. 2018.

[81] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, pp. 3982–3998, Sep. 2018.

[82] M. Moltafet, N. M. Yamchi, M. R. Javan, and P. Azmi, "Comparison study between PD-NOMA and SCMA," *IEEE Trans. Vehic. Tech.*, vol. 67, pp. 1830–1834, Feb 2018.

[83] Z. Li, W. Chen, F. Wei, F. Wang, X. Xu, and Y. Chen, "Joint codebook assignment and power allocation for SCMA based on capacity with gaussian input," in *IEEE Int. Conf. Commun. in China*, pp. 1–6, July 2016.

[84] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[85] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE Jour. Select. Areas Commun.*, vol. 35, pp. 2744–2757, Dec. 2017.

[86] J. Papandriopoulos and J. S. Evans, "Scale: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Trans. Info. Theory*, vol. 55, pp. 3711–3724, Aug 2009.

[87] Y. Li, M. Sheng, Z. Sun, Y. Sun, L. Liu, D. Zhai, and J. Li, "Cost-efficient codebook assignment and power allocation for energy efficiency maximization in SCMA networks," in *IEEE Vehic. Tech. Conf.*, pp. 1–5, Sep. 2016.

[88] P. He, L. Zhao, S. Zhou, and Z. Niu, "Water-filling: A geometric approach and its application to solve generalized radio resource allocation problems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3637–3647, 2013.

[89] N. Wang, E. Hossain, and V. K. Bhargava, "Backhauling 5g small cells: A radio resource management perspective," *IEEE Wire. Commun.*, vol. 22, pp. 41–49, Oct. 2015.

[90] "System architecture for the 5G system," *document 3GPP Technical Specication 23.501, Release 15*, Dec. 2017.

[91] "IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond," *document M.2083, ITU Radio communication Study Groups*, Feb. 2015.

[92] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72253–72262, 2018.

[93] "Study on scenarios and requirements for next generation access technologies," *document 3GPP 38.913, Version 14.1.0*, Mar. 2016.

[94] K. M. S. Huq, S. Mumtaz, J. Rodriguez, P. Marques, B. Okyere, and V. Frascolla, "Enhanced C-RAN using D2D network," *IEEE Communications Magazine*, vol. 55, pp. 100–107, March 2017.

[95] Y. C. Tay, K. Gaurav, and P. Karkun, "A performance comparison of containers and virtual machines in workload migration context," in *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pp. 61–66, June 2017.