Energy-Efficient Resource Allocation in Limited Fronthaul Capacity Cloud-Radio Access Networks

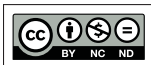
by

Phuong Thi Thu LUONG

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, "DECEMBER 4, 2018"

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC



BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

M. Charles Despins, Thesis Supervisor
Département de Génie Électrique, École de Technologie Supérieure

M. François Gagnon, Co-supervisor
Département de Génie Électrique, École de technologie supérieure

M. Stéphane Coulombe, President of the Board of Examiners
Département de Génie Logiciel et des TI, École de technologie supérieure

M. Michel Kadoch, Member of the jury
Département de Génie Électrique, École de technologie supérieure

M. Yang Yang, External Independent Examiner
ShanghaiTech University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON "NOVEMBER 8, 2018"

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

Throughout my graduate studies, many people have supported and helped me in one or other ways. This dissertation could not have been accomplished without their advices and guidances. I would like to extend my sincere appreciation to all of them.

First and foremost, I would like to express my deepest appreciation and thanks to my supervisor, Professor Charles Despins from École de technologie supérieure (ÉTS), for his unconditional encouragement, insights and inspirations during my studies at ÉTS. I have benefited tremendously from his efforts and motivation to my research and personal growth. As well, I would like to specially thank my co-supervisor, Professor François Gagnon from ÉTS, who taught me with generous support and helpful advices for my work. I really appreciate the opportunity he gave me to pursue my Ph.D study and shape my research career. Moreover, I would like to express my heartfelt gratitude to Professor Le-Nam Tran at University College Dublin in Ireland with many fruitful discussions and important advices for my research. Without his concrete guidances, I would not have been able to conduct my precise results.

I also would like to thank to all the jury members, Professor Stéphane Coulombe, Professor Michel Kadoch and Professor Yang Yang for accepting to supervise my research. I am very grateful to have them as my jury members. I would like to acknowledge the Department of Electrical Engineering at ÉTS for providing the great learning experience.

I am forever thankful to the NSERC-PERSWADE program and Dr. Paul Marinier for supporting me a research internship at InterDigital in Montreal. I have learned immeasurable knowledge in the 3GPP standards and 5G wireless design from his elaborate instructions. I also want to deliver my thanks to my colleague, Dr. Patrick Tooher who supports and collaborates to improve my work at InterDigital. I would like to express my gratitude to all my colleagues, especially Dr. Minh Au, Dr. Mouna Hajir, Dr. Moshiur Rahman, Mr. Azzouz Omar Zayen, Dr. Sara Lakani, Dr. Edenalisoa Rakotomanana and other members of LACIME laboratory, for their helpful support and unforgettable time at ÉTS.

To all my friends, thank you for sharing the hard time and joyful moments with me. Special thanks to my friends for giving me your time, energy and expertise over the years in Montreal. I also want to thanks to all my friends scattered around the world for their inspiring, wishes and visits, and supporting me whenever I needed a friend.

As always, from bottom of my heart I would like to give a grateful thank to my family and my in-laws for inspiring and supporting me to pursue my study. Their love, encouragement and belief are my power and motivation to go through the difficult times in my doctoral work. Lastly, words cannot express how grateful I am to my husband, Nguyen Minh Tri, and my daughter, Emma Ngoc Nguyen, for all of their love and the sacrifices they have made on my behalf. I will never have been able to finish the long jouney of the doctoral study without them.

ALLOCATION DE RESSOURCES À EFFICACITÉ ÉNERGÉTIQUE DANS DES RÉSEAUX D'ACCÈS RADIO-CLOUD À CAPACITÉ LIMITÉE DU FRONTHAUL

Phuong Thi Thu LUONG

RÉSUMÉ

Au cours des dernières années, les réseaux d'accès radio cloud (C-RAN) ont continuellement démontré leur rôle à titre de candidat technologique fort prometteur pour relever les nombreux défis liés à l'avènement des réseaux mobiles de cinquième génération (5G). Dans les C-RANs, les modules capables de traiter les données et les signaux radio sont physiquement séparés en deux groupes fonctionnels principaux : le groupe BBU (Baseband Unit) constitué de plusieurs BBUs sur le cloud et les réseaux d'accès radio (RAN) composés de plusieurs têtes de radiocommande à faible puissance (RRH) dont la fonctionnalité est simplifiée grâce à la transmission / réception radio. Grâce à la capacité de calcul centralisée du cloud computing, les C-RANs permettent une coordination entre les RRHs afin d'améliorer significativement l'efficacité spectrale réalisable, permettant ainsi de satisfaire la demande croissante des utilisateurs. Par surcroît, cette performance améliorée peut être obtenue en mode d'économie d'énergie, créant ainsi une perspective de C-RAN écoénergétique. On notera qu'une telle amélioration peut être réalisée dans une condition idéale de fronthaul avec une capacité très élevée et stable. Cependant, en pratique, les liens fronthaul dédiés doivent être substantiellement divisés pour connecter une grande quantité de RRH au cloud, conduisant à un scénario de capacité limitée de fronthaul pour chaque RRH. Cela impose une limite supérieure à l'efficacité spectrale de chaque utilisateur, limitant ainsi la performance des C-RANs. Pour tirer pleinement profit des C-RANs écoénergétiques tout en respectant leurs caractéristiques strictes en matière de capacité de transmission fronthaul, il est ainsi essentiel de concevoir un réseau plus approprié et plus efficace.

L'objectif principal de cette thèse est d'optimiser l'empreinte écologique des C-RANs en terme d'efficacité énergétique dans des conditions de capacité de fronthaul non idéales, à savoir une conception économe en énergie dans des C-RANs à capacité de fronthaul limitée. En déterminant conjointement la formation du faisceau d'émission, la sélection RRH et l'association RRH- utilisateur, notre étude cible les trois problèmes de conception essentiels suivants : le compromis optimal entre la maximisation du débit et la minimisation de la consommation énergétique totale, l'efficacité énergétique maximale dépendante du modèle de puissance du débit adaptatif, la conception optimale de l'informatique virtuelle efficace concernant l'économie d'énergie ainsi que l'allocation des ressources radio dans les C-RANs virtualisés. Les contributions significatives et les éléments novateurs de ces travaux sont décrits dans les chapitres suivants.

Premièrement, nous présentons au chapitre 3 la conception conjointe de la formation de faisceau d'émission, la sélection RRH et l'association RRH-utilisateur pour optimiser le compromis entre la maximisation du débit et la minimisation de la consommation énergétique totale dans les liaisons descendantes des C-RANs. Nous avons développé un algorithme puissant

avec une haute complexité et deux nouveaux algorithmes de faible complexité pour résoudre respectivement les solutions optimales globales et celles sous-optimales de haute qualité. Les résultats de ce chapitre montrent que les algorithmes proposés, en plus de surmonter la résolution du problème non convexe difficile dans un temps polynomial, surpassent également les techniques proposées dans la littérature sur le plan de la convergence et de la performance du réseau.

Deuxièmement, nous avons proposé dans le chapitre 4 un nouveau modèle reflétant la dépendance de la puissance consommée sur le débit de données de l'utilisateur et met en évidence son impact à travers divers métriques d'efficacité énergétique dans les C-RANs. La performance dominante du modèle des résultats du chapitre 4, comparée au modèle conventionnel sans puissance adaptée en fonction du débit, corrobore l'importance du nouveau modèle proposé pour conserver de manière appropriée la puissance du système et obtenir la performance C-RAN la plus économe en énergie.

Enfin, nous avons proposé dans le chapitre 5 un nouveau modèle sur le centre du cloud qui permet la virtualisation et l'allocation adaptative des ressources informatiques selon le trafic significatif de données pour conserver plus de puissance. Un problème de conception conjointe de la ressource informatique virtuelle avec le beamforming, la sélection RRH, et l'association RRH-utilisateur qui maximise l'efficacité énergétique C-RAN virtualisée est considéré. Pour faire face au problème de taille complexe sur l'optimisation formulée et aboutir à la solution, un nouvel algorithme efficace avec une plus faible complexité par rapport au travail précédent a été développé. À partir de différentes évaluations, les résultats obtenus démontrent la supériorité des concepts proposés par rapport aux travaux conventionnels.

Mots-clés: Beamforming, réseau d'accès radio cloud, fronthaul limité, optimisation

ENERGY-EFFICIENT RESOURCE ALLOCATION IN LIMITED FRONTHAUL CAPACITY CLOUD-RADIO ACCESS NETWORKS

Phuong Thi Thu LUONG

ABSTRACT

In recent years, cloud radio access networks (C-RANs) have demonstrated their role as a formidable technology candidate to address the challenging issues from the advent of Fifth Generation (5G) mobile networks. In C-RANs, the modules which are capable of processing data and handling radio signals are physically separated in two main functional groups: the baseband unit (BBU) pool consisting of multiple BBUs on the cloud, and the radio access networks (RANs) consisting of several low-power remote radio heads (RRH) whose functionality are simplified with radio transmission/reception. Thanks to the centralized computation capability of cloud computing, C-RANs enable the coordination between RRHs to significantly improve the achievable spectral efficiency to satisfy the explosive traffic demand from users. More importantly, this enhanced performance can be attained at its power-saving mode, which results in the energy-efficient C-RAN perspective. Note that such improvement can be achieved under an ideal fronthaul condition of very high and stable capacity. However, in practice, dedicated fronthaul links must remarkably be divided to connect a large amount of RRHs to the cloud, leading to a scenario of non-ideal limited fronthaul capacity for each RRH. This imposes a certain upper-bound on each user's spectral efficiency, which limits the promising achievement of C-RANs. To fully harness the energy-efficient C-RANs while respecting their stringent limited fronthaul capacity characteristics, a more appropriate and efficient network design is essential.

The main scope of this thesis aims at optimizing the green performance of C-RANs in terms of energy-efficiency under the non-ideal fronthaul capacity condition, namely energy-efficient design in limited fronthaul capacity C-RANs. Our study, via jointly determining the transmit beamforming, RRH selection, and RRH–user association, targets the following three vital design issues: the optimal trade-off between maximizing achievable sum rate and minimizing total power consumption, the maximum energy-efficiency under adaptive rate-dependent power model, the optimal joint energy-efficient design of virtual computing along with the radio resource allocation in virtualized C-RANs. The significant contributions and novelties of this work can be elaborated in the followings.

Firstly, the joint design of transmit beamforming, RRH selection, and RRH–user association to optimize the trade-off between user sum rate maximization and total power consumption minimization in the downlink transmissions of C-RANs is presented in Chapter 3. We develop one powerful with high-complexity and two novel efficient low-complexity algorithms to respectively solve for a global optimal and high-quality sub-optimal solutions. The findings in this chapter show that the proposed algorithms, besides overcoming the burden to solve difficult non-convex problems within a polynomial time, also outperform the techniques in the literature in terms of convergence and achieved network performance.

Secondly, Chapter 4 proposes a novel model reflecting the dependence of consumed power on the user data rate and highlights its impact through various energy-efficiency metrics in C-RANs. The dominant performance of the results form Chapter 4, compared to the conventional work without adaptive rate-dependent power model, corroborates the importance of the newly proposed model in appropriately conserving the system power to achieve the most energy-efficient C-RAN performance.

Finally, we propose a novel model on the cloud center which enables the virtualization and adaptive allocation of computing resources according to the data traffic demand to conserve more power in Chapter 5. A problem of jointly designing the virtual computing resource together with the beamforming, RRH selection, and RRH–user association which maximizes the virtualized C-RAN energy-efficiency is considered. To cope with the huge size of the formulated optimization problem, a novel efficient with much lower-complexity algorithm compared to previous work is developed to achieve the solution. The achieved results from different evaluations demonstrate the superiority of the proposed designs compared to the conventional work.

Keywords: Beamforming, cloud radio access network, limited fronthaul, optimization

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABREVIATIONS

5G	Fifth Generation
ADC	Analog to digital conversion
AWGN	Additive white gaussian noise
BBU	Baseband unit
BCD	Block coordinate descent
BnB	Branch and bound
BRB	Branch and reduce and bound
BS	Base station
CAPEX	Capital expenditure
CDF	Cumulative distribution function
CoMP	Coordinated multipoint
CPRI	Common public radio interface
CPU	Computer processing unit
C-RAN	Cloud-Radio access network
CSI	Channel state information
CWDM	Coarse Wavelength Division Mutiplexing
DAC	Digital to analog conversion
DCA	Difference of convex algorithm
DL	Downlink

DPC	Dirty-paper coding
EE	Energy efficiency
H-CRAN	Heterogeneous C-RAN
IFFT	Inverse fast Fourier transform
I/Q	In-phase/Quadrature
LOS	Line of sight
MAC	Media access control
MBS	Macro base station
MC	Multicell
MIMO	Multiple-input multiple-output
MINLP	Mixed integer nonlinear program
MI-SOCP	Mixed integer-second order cone program
MMSE	Minimum mean square error
MU	Multiuser
NFV	Network function virtualization
NLOS	Non-line of sight
NP-hard	Non-deterministic polynomial-time hard
OPEX	Operating expenditure
PDCP	Packet data convergence Control
PHY	Physical layer

PPP	Poisson Point Process
PtP	Point-to-Point
PtmP	Point-to-multipoint
QoS	Quality of service
RAN	Radio access network
RF	Radio frequency
RLC	Radio link control
RRH	Remote radio head
RRC	Radio resource control
RRH-UE	Remote radio head-user association
SCA	Successive convex approximation
SDN	Software define network
SE	Spectral efficiency
SINR	Signal-to-interference-plus noise ratio
SNR	Signal-to-noise
SOCP	Second order cone program
UDN	Ultra dense heterogeneous netrworks
UE	User
UL	Uplink
VCRA	Virtual computing resource allocation

VM	Virtual machine
WMMSE	Weighted minimum mean square error
xDSL	DSL technologies
ZF	Zero-forcing

LIST OF ALGORITHMS

INTRODUCTION

Recently, the demand to serve ubiquitously an incredibly huge amount of wireless devices from the proliferation of *Internet of Things (IoT)* has been exponentially increasing. This drives wireless technology and its vertical businesses as one of the most growing industries in this decade (Ghanbari *et al.*, 2017). With more than 50 billions connected devices at the end of 2020 (Peng *et al.*, 2015), wireless services are pertinent to providing high-speed, ultra-high reliability, ultra-low latency, and ubiquity to cope with a wide classifications of use–case applications. These new uses of wireless applications are found through various domains such as augmented and virtual reality, high definition video streaming, social networks, machine-to-machine communications, automatic driving and flying, tactile Internet, etc (c.f. Fig. 0.1). While technological efforts in dealing with the data–hunger phenomenon are underway, the issues of envisaging much more power consumption scaling with the tremendous surge of wireless devices is obviously inevitable. Not only that this is the dominant source of environmental pollution via emitting CO_2 in the air, but also it raises additional expensive cost to the existing expenses of network capital expenditure (CAPEX) and operating expenditure (OPEX). Therefore, achieving greener communications in terms of energy-efficient solution along with the development of future technologies is utmost imperative towards the Fifth Generation (5G) implementation.

To satisfy all the angles of 5G's requirements, recent updates in all 5G's technology candidates are able to radically deliver the superior enhancement of system capacity, data rate, latency and more importantly, energy-efficiency. Among the key enablers of 5G such as massive multiple-input multiple-output (MIMO), millimeter wave (mmWave) communications, dense heterogeneous networks (HetNets), or full-duplex transceivers, cloud radio access networks (C-RANS) emerge as a tangible technology candidate to achieve such plausible green solution. By bridging the conventional radio access networks (RANs) and cloud computing technology using fronthaul connections, C-RANs possess all the advantages of centralized computing

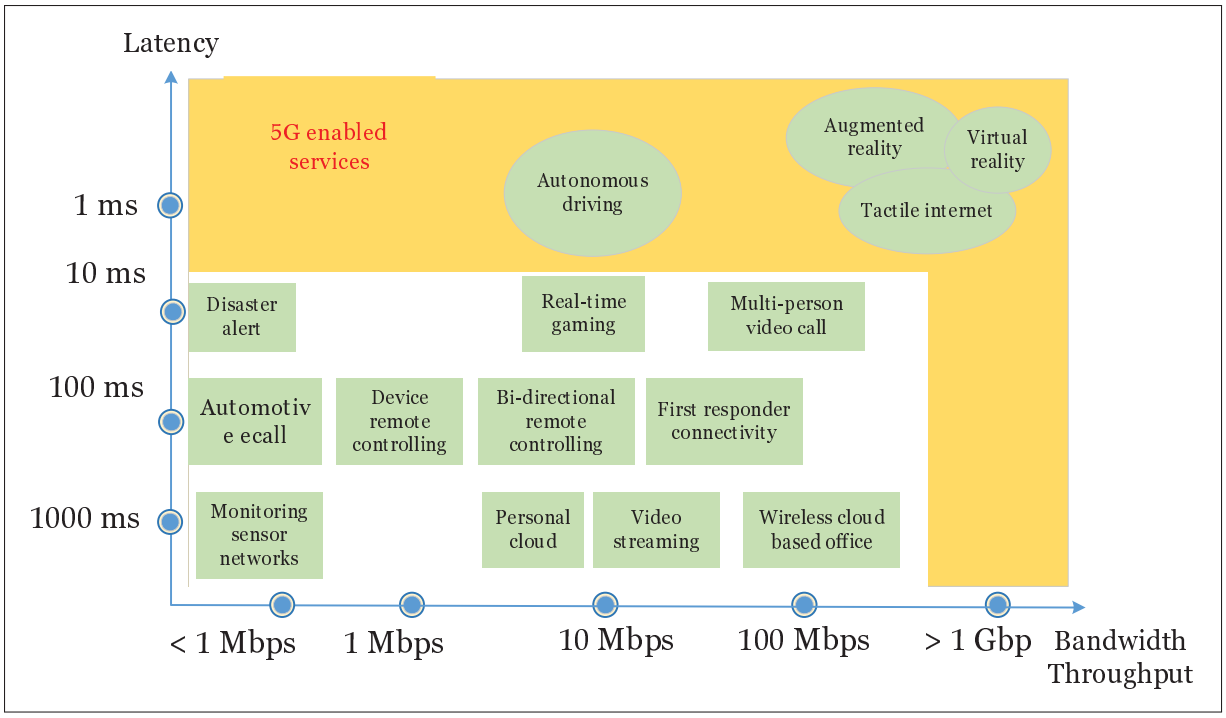


Figure 0.1 Different use-case applications in 5G (Intelligence, 2014).

and virtualization capabilities in cloud computing and virtualization technology together with the autonomy, ubiquity, low-power operation in small cells. Thus, it can leverage the system energy-efficiency by improving system spectrum efficiency and reducing overall power consumption via non-trivial radio resource allocation design, thanks to the centralized coordination on the cloud. One bottleneck challenge of C-RANs is that, in practice, the high capacity fronthaul links are greatly divided into finite capacity ones to support a dense deployment of RAN, e.g., remote radio head (RRH). This leads to a limited data rate flow to/from the cloud center, which could degrade or jeopardize the C-RAN's operation if network resource is under- or over-utilized, respectively. In light of this, this thesis mainly focuses at the designs of three major components of C-RANs, namely transmit beamforming, RRH idle/active selection, RRH–user association, which maximizes the C-RAN energy-efficiency.

The organization of this thesis, which contains 6 chapters, is as follows. Chapter 1 introduces the overview of C-RANs concerning about the energy-efficiency issues, the motivations, the problem objectives together with its novel contributions, and the proposed methodology. Chapter 2 surveys the literature review of prior work on C-RANs. Then, each following chapter presents an article which has been published, or submitted in a peer-reviewed journal.

Specifically, Chapter 3 presents the first article's work on a joint design of RRH-user association, RRH selection and beamforming in a downlink limited fronthaul C-RAN which simultaneously maximizes the achievable sum rate and minimizes the total power consumption. The formulated problem is a mixed integer non-linear program due to the joint appearance of binary variables and non-convex functions, which is generally difficult to solve. Based on this, various algorithmic approaches at different complexity levels are proposed to solve for solution. First, a high-complexity branch-and-bound based algorithm is developed to compute a global optimal solution. Second, a method based on novel transformation techniques, successive convex approximations, and Lipschitz continuity property is proposed to drive the problem into a series of approximated mixed integer second order cone programming (MI-SOCP), which can be sequentially solved by modern dedicated solver until convergence. The third approach relaxes the binary variables into continuous ones, iteratively solves a sequence of relaxed continuous SOCP problems, and performs a post-processing procedure to seek for a feasible value of the binary variable, until convergence. Towards the end of this chapter, the framework of sparsity-inducing regularization is used to develop another low-complexity efficient algorithm based on SOCP, which also iteratively solves the approximated problems until convergence. Extensive numerical results show that the developed algorithms always outperform the conventional solution approaches in terms of algorithm's convergence speed and achieved system performance.

Chapter 4 presents the second article's work on a joint design of RRH-user association, RRH selection, and beamforming in a downlink limited fronthaul C-RAN which maximizes each of the following three metrics:

- Global energy-efficiency: defined as the ratio of sum achievable rate over the total power consumption.
- Weighted sum of energy-efficiency: defined as the weighted sum of all individual RRH's energy-efficiency, where each individual energy-efficiency term is computed as the ratio of achievable rate over the power consumption at that RRH.
- Minimum of all individuals' energy-efficiency (fairness): defined as the smallest individual energy-efficiency among all considered RRH.

In all problems, a novel rate-dependent model, which more accurately characterizes the behavior of total power consumption than conventional work, is proposed. Based on the developed solution framework in Chapter 3, Chapter 4 investigates the impact of using a precise power consumption's model on improving the achieved energy-efficiency in limited fronthaul capacity C-RANs.

Chapter 5 presents the third article on energy-efficient resource allocation of limited fronthaul capacity C-RANs with virtualization. In particular, this chapter proposes a novel virtual computing resource allocation (VCRA) scheme, in which the incoming user traffic is adaptively split into smaller workload fractions and parallelly processed by virtual in the BBU pool. Each virtual machine can be accordingly active (or idle) and allocated with a sufficient amount of virtual computing resource to conserve more power. Based on this scheme, a problem of jointly designing the virtual computing resource along with the radio resource allocation which maximizes the system energy-efficiency in the virtualized limited-fronthaul C-RAN is formulated. Then, a novel algorithm, which is based on the difference of convex (D.C.) technique combined

with Lipschitz continuity, with much lower complexity than the previous solution approaches in Chapter 3 and 4 is developed to solve the formulated problem at the best efficient way. Numerical results show that the proposed VCRA model together with the newly developed algorithm significantly outperforms other conventional designs in terms of algorithm's efficiency and achieved system performance.

CHAPTER 1

CLOUD RADIO ACCESS NETWORKS: OVERVIEW, MOTIVATIONS, CHALLENGES, RESEARCH OBJECTIVE, METHODOLOGY

1.1 Overview

Towards the end of current decade, mobile wireless networks will experience several updates along the announcement of 5G standardization (Andrews *et al.*, 2014; Boccardi *et al.*, 2014). The emergence of 5G, as being well-known for its promises on the enhancement over several domains (Qualcomm, 2016), as depicted in Fig. 1.1, has become the most anticipated topic for the last few years. Many technologies such as massive MIMO, mmWave, or dense networks (Chin *et al.*, 2014) have been lending themselves as the promising candidates to fulfill the aforementioned goals by enhancing the conventional transmission quality with advanced hardware upgrades and groundbreaking technologies, mainly at the serving base station (BS) side. In particular, the BSs can either be equipped with a large scale number of antenna array (Larsson *et al.*, 2014; Nguyen *et al.*, 2015; Nguyen & Le, 2015; Hoydis *et al.*, 2013), migrate wireless transmissions to a much higher mmWave frequency band (Ayach *et al.*, 2014; Hur *et al.*, 2013), be densely deployed over the cellular coverage (Bhushan *et al.*, 2014), or even join these techniques altogether to suit future network performance with the 5G's requirement. Beside the rate-prioritized expectations, improving the network energy-efficiency has been recently highlighted as an important issue and rapidly drawn significant attention from academic and industrial research. According to a report in (Peng *et al.*, 2015), the total power consumption contributed by the cellular network sector is roughly 60 billions kWh per year, where 80% of energy demand is from the BS side (Auer *et al.*, 2011). This subsequently causes hundreds of millions of polluted carbon dioxide emission annually, which is envisioned to be doubled by 2020. Obviously, an intense increase of BS deployment can tremendously raise consumed power and environmental responsibility cost (OPEX) to an incredible figure. Note that this is in addition to the inherent expensive expenditure CAPEX, which handles the network planning, site acquisition, radio frequency (RF) hardware, baseband hardware, software licenses, leased

line connections, installation, civil cost and site support, like power and cooling. Achieving an energy-efficient network in 5G is therefore seeking for a practical solution which harmonizes the improvement of system throughput at the most plausible cost of power consumption.

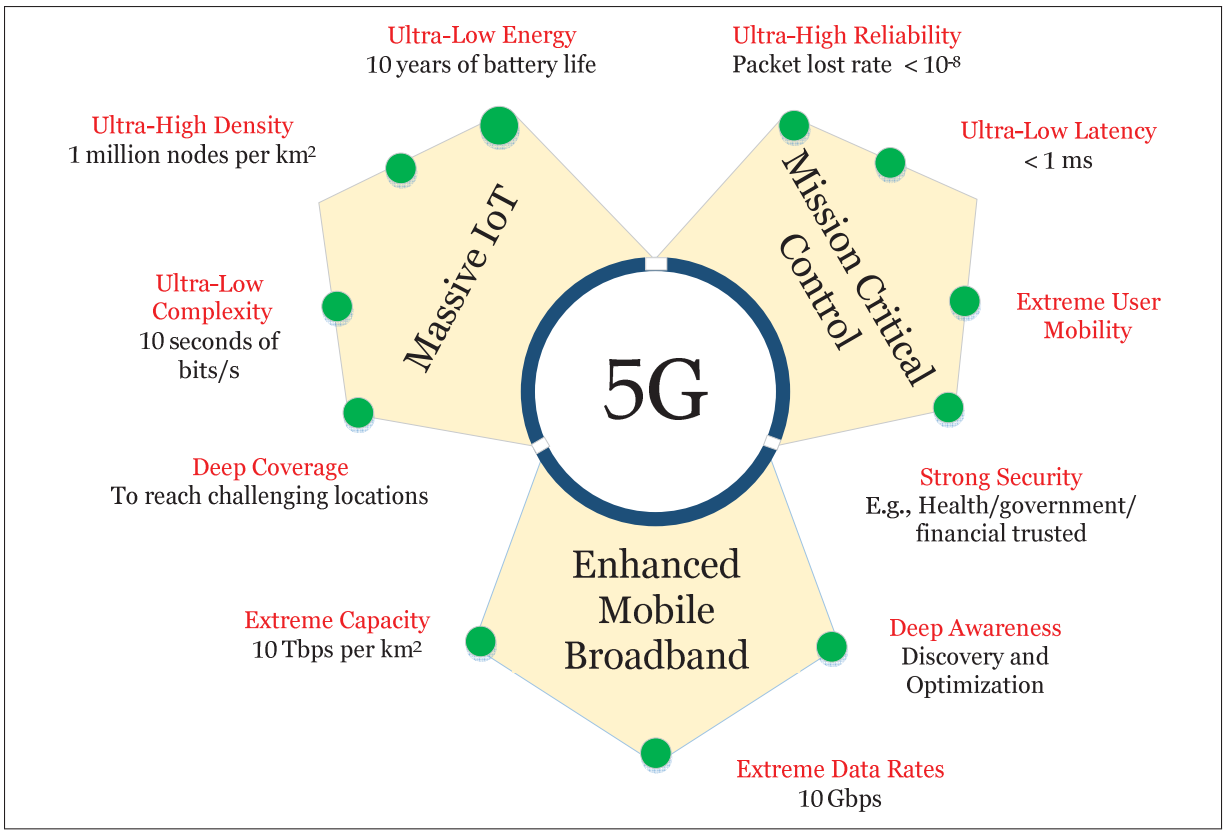


Figure 1.1 Expected enhancement of 5G of different use cases (Qualcomm, 2016).

C-RAN emerges as a prominent technology which can efficiently overcome this energy-efficient barrier to achieve a greener future network (Rost *et al.*, 2014; Abid *et al.*, 2011). Unlike traditional cellular networks, the general structure of C-RANs aims at physically separating the computing and radio transmission modules. The computing components are gathered on the cloud center to form a pool of base band unit (BBU), where using virtualization technique to control computing resources strengthens the centralized coordination's capability for heavy signal processing tasks. On the other hand, each radio transmission unit, as known as remote

radio head (RRH), is typically simplified with a RF module to transmit/receive signal to/from users. By either using fiber or wireless fronthaul links to connect these two cloud-based and radio-based component types, C-RANs are successful in fusing the fruitful advantages of joint transmission coordination from Coordinated Multi-Point Communications (CoMP) (Irmer *et al.*, 2011; Lee *et al.*, 2012) and inexpensive deployment of low-power RRHs at a enormous density from Ultra-Dense heterogeneous Networks (UDN) (Ge *et al.*, 2016), thus advances as a pragmatic energy-efficient solution.

1.1.1 C-RAN Structure

A basic architecture of C-RAN including the RRHs connected to the BBU pool via multiple fronthaul connections (Wu *et al.*, 2015) is depicted in Fig. 1.2. The three components' features can be further detailed as follows:

- **RRH:** In contrast to the traditional cellular BS whose BBUs are internally integrated, RRH is only equipped with radio units and spatially separated far from the BBUs, which are located in the cloud center. In fact, each RRH only accounts for the compression and transmission/reception of radio signals to/from users, which has similar functionalities of small cell BSs in heterogeneous networks (Nguyen *et al.*, 2013, 2012a,b; Nguyen & Le, 2014b,a; Chandrasekhar *et al.*, 2008). In particular, a RRH can operate on both uplink (UL) and downlink (DL) to communicate with its served users where its main tasks include the digital processing, digital to analog conversion (DAC), analog to digital conversion (ADC), power amplification, filtering, RF transmission and hnadling interface of fronthaul link for the data transportation to/from the BBU pool.

- **BBU Pool:** BBU pool consists of multiple BBUs which collocate in the cloud center. BBU pool's main functionalities include the radio resource control (RRC) in Layer 3[1], transport medium access control (MAC) in Layer 2 and channel coding/decoding, quantization, antenna mapping, resource block mapping, sampling, modulation, inverse fast Fourier trans-

[1] Please refer to the 7 layers of the OSI system.

form (IFFT) in Layer 1 (c.f. Fig. 1.3). By centrally coordinating all the central processing units (CPUs), a BBU pool can handle the sophisticated baseband signal processing and even compute optimal radio resource allocation such as transmit beamforming or power for the RRHs' transmissions/receptions. In general, a BBU pool often applies two transpiration mechanisms: compression-after-precoding and compression-before-precoding. In this first mechanism, the BBU pool precodes the beamforming vectors intended for the RRHs, then compresses the precoded data and forwards it to RRHs via the corresponding fronthaul links. In the second one, the BBU pool forwards the compressed beamforming vectors to RRHs, and the RRHs precode the received signals and transmit them to the UEs.

- **Fronthaul link:** A fronthaul link can be a wired or wireless connection that connects the BBU pool to the corresponding RRHs using a specific interface. The in-phase/quadrature (I/Q) data transmission protocol between RRH and the BBU pool in the fronthaul link can be bidirectional such as Common Public Radio Interface (CPRI) (Doc, 2013), Open Base Station Architecture Initiative (OBSAI) (Doc, 2006) and Open Radio equipment Interface (ORI) (Doc, 2011). Fronthaul links can be generally categorized into two types: ideal and non-ideal limited fronthaul. In the ideal fronthaul case, optical fiber cables of high capacity are used to guarantee fronthaul data transportation with low latency and high reliability. In contrast, the non-ideal limited fronthaul often refers to wireless communications to offer an inexpensive and flexible solution for fronthaul data transmissions. Due to the challenges of wireless channel attenuation and wireless transmissions' concurrency, non-ideal fronthaul capacity is limited at some finite value, which set a remarkable upper-bound for the potential performance of C-RANs. Table 1.1 summarizes the up-to-date wireless technologies for the non-ideal fronthaul used in practical C-RANs.

1.1.2 C-RAN Functional Split

A typical C-RAN contains a certain set of functionalities belonging to either the BBU pool or RRH. However, these functionalities are fixed at one place and can be adaptively equipped at the BBU or RRH depending on the operator's demand. Based on the quality of the fronthaul

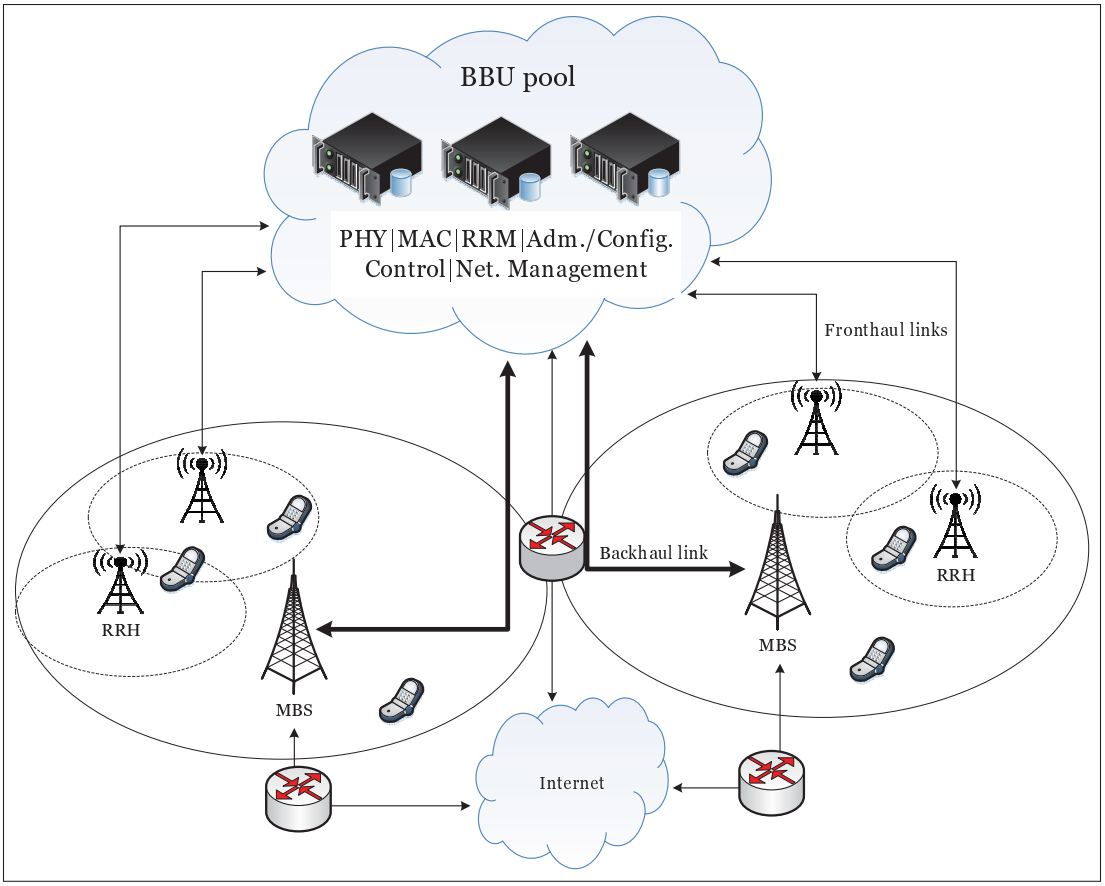


Figure 1.2 System architecture for the deployment of Cloud radio access network (C-RAN) into the existing cellular network.

connections and the functionalities equipped at the BBU pool and each RRH, C-RAN operation can be categorized, according to Fig. 1.3, into three types in term of the network coordination level:

- **Full centralization:** The BBU pool is responsible for the functions of baseband physical layer, MAC layer, and network layer which manage the operation of all RRHs. The BBUs that operate in this coordination level contains all the tasks of signal processing and resource management similar to the traditional cellular BS but suffers heavy burden from the limited fronthaul conditions since extremely high data traffic is transmitted between the RRH and the BBU pool via the fronthaul links. The required fronthaul data rate, denoted as R_{FH},

Table 1.1 Different Fronthaul Technology

Fronthaul Technology	**Latency (per hop)**	**Throughput**	**Topology**
mmWave 60GHz Unlicensed	≤ 5 ms ≤ 200 μs	≤ 800 Mbps ≤ 1 Gbps	PtP (LOS) PtP (LOS)
mmWave 70-80GHz Light licensed	≤ 200 μs	≤ 2.5 Gbps	PtP (LOS)
Microwave 28-42 GHz licensed	≤ 200 μs ≤ 10 ms	≤ 1 Gbps ≤ 1 Gbps	PtP (LOS) PtmP (LOS)
Sub-6 GHz Unlicensed or licensed	≤ 5 ms ≤ 10 ms ≤ 5 ms	≤ 500 Mbps ≤ 500 Mbps ≤ 1 Gbps	PtP (LOS) PtmP (NLOS) PtmP (NLOS)
Dark Fibre	5 μs/km $\times$ 2	≤ 10 Gbps	PtP
CWDM	5 μs/km $\times$ 2	$\leq$ 10N Gbps (N ≤ 8)	Ring
Metro Optical Network	250 μs	≤ 1 Gbps	Mesh/Ring
PON (Passive Optical Networks)	≤ 1 ms	100 Mbps–2.5 Gbps	Ptmp
xDSL	5-35 ms	10 Mbps–100 Mbps	PtP

which is responsible for I/Q data transportation, can be computed as (Wübben *et al.*, 2014)

$$R_{\mathrm{FH}} = 2N_0 f_{\mathrm{s}} N_{\mathrm{Q}} N_{\mathrm{R}} \tag{1.1}$$

where $N_0, f_{\mathrm{s}}, N_{\mathrm{Q}}$, and N_{R} are the oversampling factor, sampling frequency, quantization bits per I/Q and number of receive antenna, respectively. This coordination level embraces the existence of low-cost RRH, where no digital processing unit is required. However, utilization of signal processing and beamforming techniques is necessarily optimized to reduce the traffic within the fronthaul link and harness the advantages of this setting.

- **Partial centralization:** The RRH is functioned with some additional baseband signal processing functions beside the RF capability. This means that the BBU pool tasks are facilitated with less operations on the baseband signal processing, while still include the MAC and network layer tasks. At this level, the signal processing tasks such as forward-error-correction or decoding are executed at the RRH side, resulting in the pure MAC payload inside the fronthaul link to/from the BBU. Depending on modulation and coding scheme,

the resulting fronthaul data rate is approximated as

$$R_{\mathrm{FH}} = \frac{N_{\mathrm{SC}} \eta S}{T_{\mathrm{s}}} \tag{1.2}$$

where $N_{\mathrm{SC}}, T_{\mathrm{s}}$, η, and S are number of used sub-carriers, symbol duration, resource element utilization, and spectral efficiency, respectively. This category can potentially provide great reduction of fronthaul signaling overhead and alleviation of limited fronthaul constraint. However, its implementation is very complicated and requires further research to be conducted in industrial operation.

- **Hybrid centralization:** This is considered a special case of full centralization, where a subset of base band processing physical layer is avoided at the BBU to be assembled into a new separated processing unit, which can flexibly be a part of the BBU pool. The benefit of this structure is its flexibility to allocated resource to support the joint operation of BBU and RRH and thus embrace the network reconfiguration and reduce energy consumption in the BBUs to achieve higher system energy-efficiency. If the cyclic prefix and fast Fourier transformation units are equipped to RRH to process the baseband signal, the required fronthaul data rate is significantly decreased as

$$R_{\mathrm{FH}} = \frac{2 N_{\mathrm{SC}} N_{\mathrm{Q}} N_{\mathrm{R}}}{T_{\mathrm{s}}} \tag{1.3}$$

1.1.3 Virtualization in C-RANs

Virtualization technology basically separates resources into virtual entities from the underlying physical hardware (Liang & Yu, 2015; Hawilo *et al.*, 2014). In C-RANs, virtualization is used to create a virtualized BBU pool that is operated on different commercial servers. In fact, the virtualized network functions and protocols can be deployed in a virtualized network that connects several virtual machines through virtual links, enabling the division of computing-related resources existing in the BBU pool such as data storage, memory size, CPU capacity to run

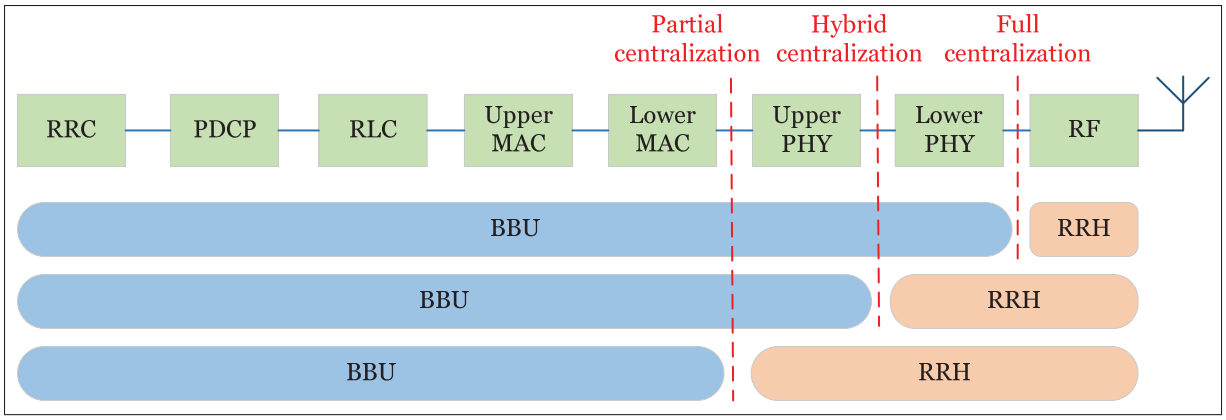


Figure 1.3 Three types of centralization in C-RAN.

different applications, operating systems and controls. Thus, C-RAN network virtualization reduces the cost of hardware utilization and increases the scalability of dynamically activating/deactivating the virtual resources. It allows C-RAN to inherit the technologies related to cloud computing and virtualization such as the Network Function Virtualization (NFV) and Software Defined Network (SDN) in order to flexibly support a broad range of services and multiple operators that are not achievable in traditional RANs.

Softwared Defined Network

SDN is the technology which splits the control and data planes. It enables the programmability and global management of network control via the external applications and the underlying infrastructure. These network controls can be reprogrammed, abstracted, and adjusted to suit with various applications and network services. Moreover, SDN based controller can offer available open interfaces between the devices and network controllers of different protocols from various vendors. With these capabilities, SDN is the key technique to provide an effective control plane in C-RAN's development towards 5G. Although SDN is not the topic focused in this thesis, it is briefly dicussed in this paragraph to complete the C-RAN concept.

Network Function Virtualization

NFV allows the functions specially developed for the networking purposes to be virtualized and complete general computing tasks. By using the virtualization technique, multiple network functions can operate through the software installed in the data center. This means that the different network devices such as packet gateways, routers, switches, and hubs can be replaced by virtualized functions operated on the common off-the-shelf servers. Thus, NFV provides network to achieve flexibility and scalability to deal with the changes of network's environment as well as enhances the hardware utilization by using software update rather than hardware update. In C-RAN, with resource cloudification in the centralized fashion, it is more beneficial to develop NFV. By using virtualization techniques to transform the physical CPUs into multiple virtual machines, the computing capability improves significantly and efficiently. In particular, it enables a more dynamic and scalable system operation to cope with the temporal and spatial traffic fluctuations. Consequently, NFV can reduce the network operation cost, exposes more software services, and provide flexible solutions.

1.2 Motivations

1.2.1 Benefits of C-RANs

Throughput Enhancement

From the basic architecture of C-RANs, it is obvious that numerous RRHs can be centrally coordinated in the cloud center, thanks to the fronthaul connections. Depending on how the centralization level is authorized in Section 1.1.2, C-RANs can accordingly allow a suitable coordination between RRHs by allocating appropriate resource allocation such as transmit power (or beamforming) so that users sharing the same time-frequency resources are simultaneously served at higher achievable rate. In fact, having multiple coordinated BSs significantly increases the degrees of freedom of transmissions similar to the MIMO scenario, so that beam-

foming techniques can be leveraged to improve cell-edge users while maintaining the good performance of the cell-centered users.

In the literature, the technique which exploits MIMO and interference management through BS cooperation is inspired by the CoMP strategy (Lee *et al.*, 2012). In CoMP, there are two levels of coordination, which are also widely used in the context of C-RANs. The first approach is the multi-point coordination, where the cooperative BSs only share CSI, but not the transmit signals. Hence, the resulted CoMP system is similar to the multi-cell networks, where each BS transmits data to its users and often suffers from the neighboring inter-cell interference. By sharing CSI, the set of radio resources such as power allocation (or transmit beamforming) is jointly computed for all the considered BSs at a central server to control interference and optimize the overall system performance.

The second approach is multi-point joint transmission, where BSs now share all the CSI together with transmit data. The BSs are now in full-cooperation mode. Hence, the resulted system is similar to the MISO multi-user networks, where the cooperative BSs cooperatively forms a centralized multi-antenna BS to simultaneously beamform data to all users. Fig. 1.4 illustrates the differences between the multi-point coordination and multi-point joint transmission. Note that users in this scenario still suffer from co-channel inter-user interference, but now with multi-antenna beamforming technique, interference can be better suppressed or mitigated to achieve a much higher gain in system performance compared to the first approach. One main drawback of CoMP joint transmission is that the signaling overhead to backhaul data between BSs towards the central server is excessive which drives the system viability far from practical implementation.

Power Conservation

The radio part of C-RANs mainly includes the low-power short-range cost-efficient RRH, which greatly reduces the overall network power consumption. Under the centralized control of BBU pool, some RRHs can be adaptively switched between active and idle modes to

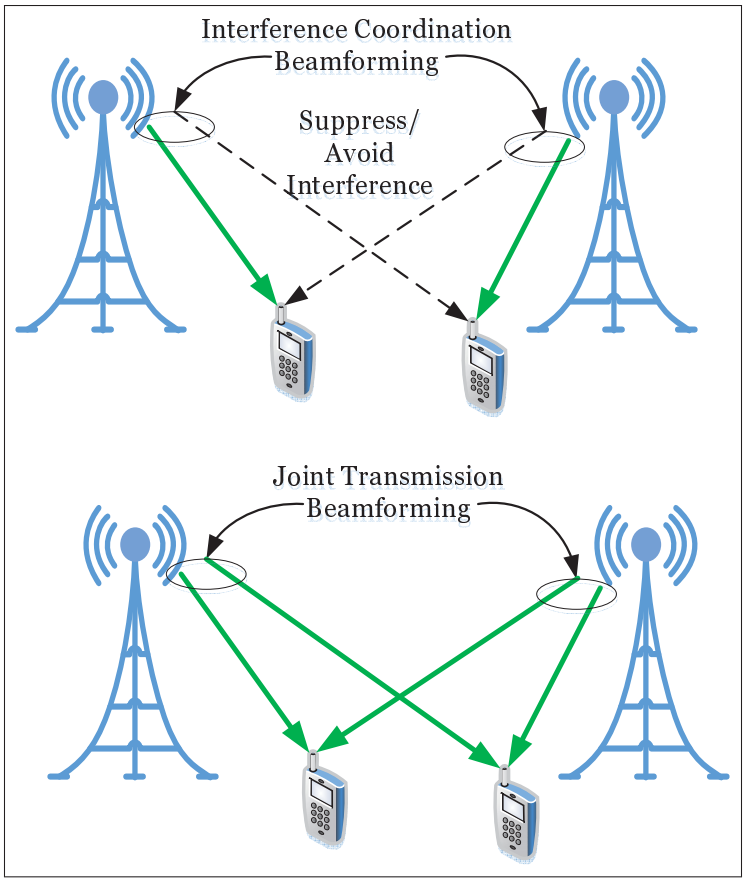


Figure 1.4 Coordinated beamforming and joint transmission in CoMP.

save more power when the channel condition is not good to communicate with their users. Another advantage is that since BBUs are moved to the cloud, the power cooling systems can be smartly controlled by the server. Further, with virtualization on the cloud, the computing unit in the BBU pool can be virtualized and adjusted to adapt with the traffic variations. In this way, the unnecessary BBUs can be turned off when data traffic is low to conserve more energy.

User Traffic Adaptability

User data traffic often varies randomly in time and space rather than remains deterministic. For example, traffic in the shopping mall or football stadium is often higher than in a rural area, and it is higher at date time rather than at night. As discussed above, C-RAN enabled with virtualization technology on the cloud can dynamically allocate the computational resource to

be active or idle to adapt with the traffic fluctuation. In particular, the amount of VM capacity[2] can be sufficiently reconfigured in the physical servers (PSs) of the same or different BBUs pool according to the required amount of baseband signal processing. Note that reducing the amount of active VMs also decreases the set of active RRHs to serve each user. This shows an efficient way of utilization BBUs and scheduling RRHs in C-RANs with virtualization compared to the traditional RANs of fixing the amount of BBU resources.

In C-RANs, RRHs positioned at the different areas can either connect to the same or different BBU pool. With fast and high-quality connections, these BBU pool can be considered as a single cloud base station to more flexibly offload data between RRHs or associate users with suitable RRHs according to the variations of data traffic and channel conditions. Besides, RRHs also can be easily added to the BBU pool to extend the network coverage extension and BBU pool can also add more servers, hardwares and install new softwares to cope with the network demand.

1.2.2 Challenges

Limited Fronthaul Capacity

Although it is known that C-RANs can offer many promising benefits, in practice, C-RAN do not have ideal fronthaul connection of infinite capacity. Since hardware resources and investment budgets are limited, fronthaul capacity must be greatly divided to connect to an enormous number of RRHs. Therefore, practical C-RANs are constrained by limited fronthaul capacity Peng *et al.* (2015); Bernardos *et al.* (2013). Since each connection is often reserved for signaling overhead, which includes the CSI estimates, raw baseband signal on the uplink, and quantized I/Q data on the downlink, etc., a limited connection leads to the following problems:

[2] VM capacity is mostly represented by three components: (i) computing capacity in CPU cycles/second, (ii) memory and storage capacity in bytes, and (iii) network interface speed in bps.

- The achievable data rate at each RRH must not exceed its given fronthaul link's capacity threshold. This explicitly imposes a certain bound on the system throughput, which can potentially degrade C-RAN performance compared to the ideal fronthaul condition (Peng *et al.*, 2015).

- Higher lossy quantization schemes must be used in order to compress the amount of information to suit with the capacity threshold so that the precoded I/Q data can be transfered to the RRH. This creates in unreliable transmitting sources, which also reduces the system performance.

- Signaling overhead exchanges often require excessively high rate from the fronthaul link (Biermann *et al.*, 2013). One common approach is to reduce the amount of signaling information over the fronthaul link. However, reducing the amount of CSI to/from the centralized BBU pool by just sending a part of the complete CSI set could cause insufficient and inaccurate data input to compute the expected good solution for the system performance.

- Limited fronthaul capacity also results in system delay, asynchronization, imperfect CSI.

Clustering Issues

From the discussion of C-RAN functional split, it is obvious that a dense scale of RRH supported by full centralization can achieve very high cooperative signal processing gain, which can significantly enhance C-RAN throughput performance. However, large-scale full centralization in C-RANs implies the processing of multiple large channel matrices, which subsequently leads to high computational complexity and channel estimation overhead. To reduce the high capacity burden incurred at the limited fronthaul link, clustering is proposed as an effective technique to restrain the number of cooperative RRHs per cluster that serves a typical user. In general, there exists three RRH clustering strategies, whose pros and cons are described as follows:

- **Disjoint clustering:** The entire RRHs are divided into non-overlapping clusters. Each cluster contains a subset of RRHs that jointly serve all UEs within the clustered coverage. Although its low complexity implementation is favored, this scheme has one drawback that the cluster-edge UEs must suffer from remarkable inter-cluster interference.

- **User-centric clustering:** Each UE is served by a selected subset of neighboring RRHs which forms a cluster; and different clusters serving each individual UE may overlap. Unlike the previous case, user-centric clustering gains more benefit than that of the disjoint scheme since there exists no explicit cluster edge. Thus, inter-cluster interference can be well mitigated if optimal user-centric clustering design is attained. Moreover, this approach can be further categorized into dynamic and static RRH clustering implementation. The dynamic user-centric clustering allows the serving RRHs for each user to change over different time slots. This enables more flexible association between user and serving RRHs to avoid the short term blockage due to the user mobility or the occurrence of obstacles. However, this causes a large signaling overhead when the new user-RRH associations are established. On the contrary, static user-centric clustering fixes the serving RRHs to each user over time and may be only updated when the user location changes.

- **Content-centric clustering:** Each RRH in the C-RAN system is equipped with a cache server to prefetch and store popular data and use these data to directly transmit to the UEs. According to the similarities between the content stored in neighboring RRHs and the content requested from nearby UEs, a problem of designing appropriate clusters that accounts for content-centric factor is considered to achieve the best formulated network objective.

It is now obvious that clustering technique can help to reduce the heavy burden on the fronthaul capacity. However, this inevitably results in less cooperative gains and reduce the potential of full centralization. Determining which RRH should be clustered changes the amount of data rate necessary for each fronthaul link and affect the overall network throughput. In addition, the RRH clustering also removes unnecessary RRHs out of the clusters. These unused RRHs can be switched into idle mode to save more power consumption and reduce the inter-cell

interference. Moreover, the cluster size directly impacts to the density of scheduled users per area unit and the redefines the system diversity gains.

Virtualization Challenges

Virtualization, beside bringing many benefits to C-RAN, also entails the following challenges:

- The schemes to virtualize the computing resources from the physical servers into VMs to process the varying traffic demand and ensure the QoS requirements must be appropriate, dynamic, efficient. Beside, parallel computing capability for delay-sensitive applications should be highlighted when designing the virtualization.
- BBUs are represented by a large amount of VMs, which results in a highly complicated cloud data network topology. Therefore, supporting high bandwidth, low latency condition for BBUs connections and flexible topology of RRHs-BBUs inter-connection must always be taken in account.
- The overhead signaling issues such as control information, data exchange between the virtual networks and between virtual machines are important to be addressed in the BBU pool. VM consolidation and VM-RRH association strategy should be appropriately designed to prevent overloading the overhead in such the virtualized BBU pool.
- To apply NFV in the BBU pool, new additional virtual functions for computation intensive physical layer functions must be designed to peform the radio signal processing tasks with strict real-time requirements.

1.2.3 Motivations

The future of information technology, in order to be sustained and further developed, should take into account human-centric issues. Now and then, building a future system with energy-efficient perspective to endure human-living environment is the most important priority. Therefore, this thesis is motivated by overcoming the fundamental challenges and harness all the

benefits of energy-efficient C-RAN solution. To achieve such goal, a joint design of many coupled system parameters, such as transmit beamforming, RRH idle/active selection, RRH-user association, which optimizes the energy-efficiency-based objective while respecting the limited fronthaul capacity condition is necessary.

In general, the mathematical constrained optimization problem of limited fronthaul capacity C-RANs, while explicitly imposing all the technical issues as constraints, often falls into a category of NP-hard problem. This means that solving for an optimal solution is very difficult, while no method or off-the-shelf algorithms can optimally solve this problem type within a polynomial time. This thesis's motivations are first to formulate this problem in the most appropriate and elegant way and develop a framework which can solve for a high-quality solution by a low-complexity and efficient algorithm. Not only that, the developed framework is also multi-disciplinary, hence it can also be employed for various problems in other research topics and fields as well.

1.3 Thesis Objective and Methodology

1.3.1 Thesis Objective

The objectives of this thesis aim at constructing a unified framework to develop a series of practical algorithmic approaches which optimize various energy-efficiency-like metrics, in the limited fronthaul capacity C-RANs. The generic energy-efficiency metric throughout this thesis is defined as the ratio between the spectral efficiency, computed in bps/Hz and the power consumption, computed in Watts. Some mathematical optimization problems are formulated via the joint design of the important system parameters, such as transmit beamforming, RRH idle/active selection, RRH-user association (clustering), which optimize the following objectives of interest:

- The trade-off between achievable sum rate maximization and total power consumption minimization.

- Various energy-efficiency maximization under rate-dependent power model.
- System energy-efficiency maximization in virtualized limited fronthaul capacity C-RANs. Note that in this work, the formulated problem considers the design of VM ON/OFF status in addition to the design of transmit beamforming, RRH idle/active selection, RRH-user association is considered.

The goals of these works are to exploit the benefits of the centralized computation capability in C-RANs to develop centralized efficient algorithms which can attain a close-to-optimal solution and consolidates the potentials of C-RANs for future green networks.

1.3.2 Highlighted Contributions and Novelty

The main theme of this thesis' contribution lies in the explicit consideration of the limited fronthaul capacity constraint in all of the optimization problems. As analyzed in Section 1.2.2, mathematically imposing this constraint introduces several new obstacles to solve the formulated problems, thus remarkably raises the computational complexity in attaining the optimal solution. This thesis, beside other particular contributions presented in each work, aims at providing a novel approach to radically handle this type of constraints and enjoy the application of this approach to solve the problems visited throughout Chapter 3–5.

In particular, by introducing the explicit per-fronthaul capacity constraint in the optimization problems in Chapter 3, a novel design with more appropriate consideration of joint RRH-user association, RRH selection and beamforming design is highlighted. Unlike (Dai & Yu, 2014; Ha *et al.*, 2016) where the authors assign a predetermined achievable rate to overcome the non-convex fronthaul constraint, Chapter 3 directly tackles it by proposing a novel transformation to arrive at an equivalent but more tractable form. To evaluate the superiority of this approach, Chapter 3 considers a multi-objective optimization problem including both achievable sum rate maximization and total power consumption minimization and jointly design the transmit beamforming, RRH selection and RRH-user association in the downlink C-RANs. The binary selection variables for RRH selection and RRH-user association factors are introduced into the

problem formulation and a global optimal solution is derived by using the monotonic optimization based branch-and-bound method. In addition, a new method which approximates the non-convex parts in the considered per-fronthaul capacity constraints and the objective functions into convex SOC ones is effectively utilized to drive the original non-convex problem into mixed integer second order cone programming (MI-SOCP). Furthermore, by relaxing the binary variables and then invoking a post-processing procedure on these relaxed variables to search for a high-performance solution, a more pragmatic, and much faster algorithm (than the developed MI-SOCP based algorithm) based on SOCP is developed. Finally, we also solve the considered problem using another approach of sparsity-inducing regularization. With the novel contributions, the enhancement in terms of convergent behavior and overall network performance is achieved.

In Chapter 4, various energy-efficiency objectives, namely global, weighted sum, and fair energy-efficiency, are considered in the optimization problem under the limited fronthaul capacity C-RANs. Note that a joint design of the transmit beamforming, RRH selection and RRH-user association similar to the previous chapter is visited. Here, the significance of this work is the proposal of rate-dependent power model. This new formula can more precisely reflect the adaptation of each RRH's power consumption according to its achievable rate and is employed in all of the energy-efficiency formulation. Despite the different structure of the visited energy-efficiency metrics, a unified framework based on a similar principle to the previous chapter, such as equivalent transformation and SCA techniques, is presented to develop low-complexity algorithms to solve for each problem's solution. Numerical results are extensively studied to validate the importance of the proposed rate-dependent power model in achieving more energy-efficiency compared to existing work.

Recently, virtualization for wireless communications has been shown to be a powerful technology fully and flexibly exploiting the computing resources of C-RANs. Chapter 5 targets to study an energy-efficient design of a virtualized C-RAN under limited fronthaul capacity condition. More advanced than (Pompili *et al.*, 2016; Tang *et al.*, 2015; Guo *et al.*, 2016b; Wang *et al.*, 2016a; Saxena *et al.*, 2016), this chapter proposes a novel virtual computing re-

source allocation (VCRA) scheme, in which the users' arrival workload can be split into smaller fractions and processed by VMs in parallel. Specifically, the virtual computing resources are allocated to the VMs according to the user's traffic demand and VMs are assigned to the physical servers in such a way that the minimum physical servers are necessary to conserve more power consumption. Based on this, joint virtual computing and radio resources are designed to improve the considered virtualized limited fronthaul capacity C-RANs. Another significant contribution lies in the novel solution approach based on difference of convex algorithm (DCA) and Lipschitz continuity to solve the formulated non-convex problem. This proposed method has been shown to significantly reduce the computational complexity in solving the difficult problem, and is reflected by its enhanced convergent speed and network performance compared to the existing algorithms.

My journal and conference publications are listed as follows

1. P. Luong, F. Gagnon, C. Despins, L.-N. Tran, "Joint Virtual Computing and Radio Resource Allocation in Limited Fronthaul Green C-RANs", IEEE Trans. on Wireless Commun., Vol. 17, No. 4, Apr. 2018, pp. 2602 - 2617.

2. P. Luong, F. Gagnon, C. Despins, L.-N. Tran, "Optimal Joint Remote Radio Head Selection and Beamforming Design for Limited Fronthaul C-RAN", IEEE Trans. on Sig. Proces., Vol. 65, No. 21, Nov. 2017, pp. 5605-5620.

3. P. Luong, F. Gagnon, C. Despins, L.-N. Tran, "Optimal Energy-Efficient Beamforming Designs for Cloud-RANs with Rate-Dependent Power", Submitted to IEEE Trans. on Commun. in August 2018.

4. P. Luong, C. Despins, F. Gagnon, L.-N. Tran, "A Novel Energy-Efficient Resource Allocation Approach in Limited Fronthaul Virtualized C-RANs", Proc. IEEE Veh. Tech. Conf. (VTC-Spring 2018), Porto, Portugal, pp. 1-6.

5. P. Luong, C. Despins, F. Gagnon, L.-N. Tran, "Designing Green C-RAN with Limited Fronthaul via Mixed-Integer Second Order Cone Programming", Proc. IEEE Int. Conf. on Commun. (ICC 2017), Paris, France, pp. 1-6.

6. P. Luong, C. Despins, F. Gagnon, L.-N. Tran, "A Fast Converging Algorithm for Limited Fronthaul C-RANs Design: Power and Throughput Trade-off", Proc. IEEE Int. Conf. on Commun. (ICC 2017), Paris, France, pp. 1-6.

7. P. Luong, C. Despins, F. Gagnon, L.-N. Tran, "Joint beamforming and remote radio head selection in limited fronthaul C-RAN", Proc. IEEE Veh. Tech. Conf. (VTC-Fall 2016), Montreal, Canada, pp. 1-6.

1.3.3 Methodology

In most cases, the formulated optimization problem is a mixed binary non-convex problem, which is generally NP-hard and difficult to solve (Boyd & Vandenberghe, 2004). Throughout this thesis, the following methods are employed to achieve the solution of the formulated problem.

First, an exhaustive search approach is necessary to find the optimal solution. In particular, a generic framework to develop a high-complexity solution based on Branch-and-Bound exhaustive search algorithm is proposed to numerical arrive at an optimal solution. Note that this approach is impractical for the real-time wireless applications since its complexity exponentially scales with the problem size. The motivation of providing this solution approach is only to serve as a benchmark to compare with other low-complexity and efficient algorithms.

The second method, which is more pragmatic than the Branch-and-Bound approach due to its applicability and efficiency, is to develop more realistic algorithm which can solve for a solution within a polynomial time. To achieve such goal, two obstacles must be addressed: the integer constraint and the non-convex characteristic. The basic steps of the second method can be organized in the following order:

1. Equivalent transformations of the originally formulated optimization problem into a more tractable form, where the non-convex factors which cause the problem non-convex are revealed.

2. Relax the condition on binary variable into continuous ones bounded by the interval from 0 to 1 to turn the equivalent problem into continuous (smooth) non-convex problem.

3. Invoke the idea of successive convex approximation (SCA) technique to approximate the continuous non-convex problem into a series of convex approximated problems, where each of them is a convex upper-bound of the non-convex original problem. Note that this approximation technique introduces a few more parameters into each of the convex approximated problems.

4. Conduct some non-trivial algebraic manipulations and conic approximations to equivalently rewrite each of the convex approximated problems into a standard form of second order cone programming (SOCP).

5. Develop an algorithm to successively solve the convex approximated problems and update the parameters with the achieved optimal solution until convergence.

6. Employ the post-processing procedure based on the Inflation-Deflation algorithm (Cheng *et al.*, 2013) to sequentially refine (rounding) the value of relaxed binary variable into the binary values to finally achieve a feasible solution which satisfy all the constraints of originally formulated optimization problem.

In some cases, the problem can be formulated as a sparsity-inducing ℓ–norm non-convex optimization problem and it is categorized as a non-smooth non-convex problem. One common way to solve this problem type is to directly replace the non-smooth ℓ_0–norm function by its continuous approximation ℓ_1/ℓ_p, and then employ similar principle from Step 3–5 to solve for solution.

Finally, it is important to note that solving the constrained convex problems in Step 4, MATLAB software integrated with the YALMIP platform embedded modern dedicted solvers such

as SDPT3, SEDUMI, and MOSEK (mos, 2014) are used as the main simulation environment and tool.

CHAPTER 2

LITERATURE REVIEW

This chapter aims to cover the state-of-the-art of the existing analytical results based on tractable system design and radio resource allocation based on optimization framework techniques for the development of C-RAN with limited fronthaul capacity. More specifically, in Section 2.1, we present a survey of analytical results and statistical expression of network achievable rate for general CoMP and C-RAN in random small scale and large scale scenarios. In Section 2.2, we present a literature review of the radio resource designs based on solving an optimization problem that is used to achieve the best network performance in the considered limited fronthaul C-RAN. In Section 2.3, we introduce the literature review in which the virtual computing in the BBU pool is taken into account the radio resource design when optimizing the C-RAN system performance.

2.1 Tractable Analytical Framework for C-RAN

In C-RAN with the ideal fronthaul condition, CoMP is seen as a promising technique that enables effective interference elimination and fully harnesses the potential cooperative processing gains by allowing the coordination among all RRHs at the BBU pool (Jungnickel *et al.*, 2014; Irmer *et al.*, 2011). In the literature, there have been an extensive number of works that aim to characterize the network performance under various CoMP transmission schemes. In particular, the achievable per-cell rate that relies on the Wyner-type channel model was derived in (Simeone *et al.*, 2009) by studying the cooperative decoding scheme at the BS under the finite capacity backhaul assumption. In (Jing *et al.*, 2008), the authors also relied on the similar channel model with clustered cell-edge UEs and BS cooperation assumption to analyze the performance of several precoding schemes including dirty-paper coding (DPC), cophasing, zero-forcing (ZF) and MMSE precoders on the DL of multicell (MC) multiuser (MU) cooperative networks. Different from these works, in (Wen *et al.*, 2014), the message passing algorithm that requires only the local communications and computation of neighboring BSs

were devised to obtain the regularized ZF beamforming in the cooperative DL transmission. Further, Wu & Liang (2015) analyzed the sum-rate capacity and SE for MC cooperative cellular networks by applying the two-dimensional nested co-array, which enables the calculation of all elements in the covariance matrix of channel fading coefficients. The analytical results can be extended to the nest distributed BSs in the non-fading and Rayleigh fading channel. Lozano *et al.* (2013) presented the limitation of the cooperation within a limited size clusters of BSs by deriving the upper bound of the SE. Moreover, this paper showed the ratio of user data sharing among several BSs and provided more complex precoding and decoding scheme that scales with size of BS cooperation via examining the MC MIMO cooperation from different perspectives of coding, signal processing and information theory.

Unlike ideal C-RAN, by considering the non-ideal fronthaul which is limited by finite capacity and strict time latency, limited C-RAN is under-exploited from full cooperation among RRHs and maximum performance might be degraded when improper system parameters are chosen. Thus, it is vital to conduct more research on different CoMP strategies to determine the feasible operating regime of the limited C-RAN system. In particular, the analysis that studies the impact of limited fronthaul capacity on C-RAN performance under the various RRH cooperation strategies can be found in (Zakhour & Gesbert, 2011; Zhang *et al.*, 2013; Sanderovich *et al.*, 2009; Marsch & Fettweis, 2011; Peng *et al.*, 2014a). In (Zakhour & Gesbert, 2011), the achievable rate region for two-cell CoMP scenario was analyzed under the limited fronthaul capacity. The work in (Zhang *et al.*, 2013) studied the switching schemes between CoMP-joint processing and CoMP-coordinated beamforming with limited capacity fronthaul. The theoretical UL CoMP achievable rates analysis for classical Wyner model was presented in (Sanderovich *et al.*, 2009), which was based on decompression and decoding technique. Similarly, other works on theoretical UL schemes including decode-and-forward and compress-and-forward strategies were also analyzed under the constrained fronthaul infrastructure and imperfect CSI in (Marsch & Fettweis, 2011; Zhou *et al.*, 2013). Peng *et al.* (2014a) presented an analytical closed form of ergodic capacity for single nearest and N-nearest RRH association schemes.

These analytical results show that there should be a finite number of RRH associations in order to achieve the balance between the performance gain and implementation cost.

Despite achieving many benefits from the above tractable frameworks and results, these works only consider the random small scale fading channel model in which BS locations are fixed. In practical C-RAN with dense deployment of RRHs, these transmitting nodes are likely to be randomly positioned in space, where in most cases, the distribution of the RRH locations are assumed to follow the Poisson Point Process (PPP) with a given RRH intensity. It is important to say that since the RRHs are coordinated to mitigate the interference in C-RAN, conventional stochastic geometry analysis without considering the cooperation is unable to straightforwardly be applied to analyze the C-RAN performance. Thus, it is more challenging to propose the analytical framework and derive the closed-form expression for SE and EE metrics in large scale PPP C-RAN. By applying the stochastic geometry to cater the random large scale pathloss and small scale fading gain, the works in (Ding & Poor, 2013) analyzed the performance of C-RAN and the trade-off between reliability and system complexity with distributed antenna array. In addition, Lee *et al.* (2013) analyzed the performance of dynamic BS clustering scheme based on the stochastic geometry in dense C-RAN, in which the closed-form cumulative distribution function (CDF) of the achieved SINR was attained. On the contrary, Zhao *et al.* (2015) studied the fixed cluster formation for DL transmissions in C-RAN, where the successful access probability was derived by applying the stochastic geometry tool. However, these closed-form capacity expressions considering stochastic geometry were only analyzed for SE while the EE analysis in C-RAN is more challenging and still left open for further research.

2.2 Optimization-Based Radio Resource Allocation Design in C-RAN

2.2.1 Power minimization

Over the recent years, designs for network power consumption minimization in MIMO cellular networks has evolved to be adopted in different variant of wide cooperative networks such as C-RAN. In C-RAN, the cooperative BSs are represented by the RRHs that are allowed to densely

coordinate with each other to jointly serve the regarding UEs. This raises several fundamental issues in maintaining appropriate network configuration to meet the desired performance. With dense number of active RRHs and UEs, high power budget are inevitably required to satisfy the given minimum rate requirement quality-of-service (QoS) due to severe inter-user and inter-cluster interference. On the other hand, distributing limited fronthaul capacity among the large amount of RRHs and reducing signaling overhead carried over the fronthaul link shared by these RRHs is important. It is indeed beneficial to switch OFF a few RRHs or route an appropriate subset of RRHs to serve a corresponding group of UEs so that more network power are conserved with the least cost of fronthaul capacity. In fact, compared to the low complexity conventional disjoint clustering, an attempt to approach the user-centric or content-centric clustering by fully considering the joint design of power control, RRH selection and RRH-UE association potentially provides more significant power conservation with higher stability of limited fronthaul constraint satisfactory.

Denote a set of RRHs as $\mathscr{I} = \{1, \ldots, I\}$ and a set of single antenna users as $\mathscr{K} = \{1, \ldots, K\}$. Each RRH is equipped with M_i antennas. In a beamforming design framework, the baseband transmit signals from the ith RRH are of the form:

$$\mathbf{x}_i = \sum_{\forall k \in \mathscr{K}} \mathbf{w}_{i,k} s_k, \quad \forall i \in \mathscr{I} \tag{2.1}$$

where s_k is a complex scalar denoting the data signal intended to the kth user and $\mathbf{w}_{i,k} \in \mathbb{C}^{M_i}$ is the transmit beamforming vector at the ith RRH for the kth user. Without loss of generality, we assume that $\mathbb{E}[|s_k|^2] = 1$ and s_k's are independent with each other. The baseband received signal at the kth user is given by

$$y_k = \sum_{\forall i \in \mathscr{I}} \mathbf{h}_{i,k}^{\mathrm{H}} \mathbf{w}_{i,k} s_k + \sum_{\forall i \in \mathscr{I}} \sum_{\forall j \neq k} \mathbf{h}_{i,k}^{\mathrm{H}} \mathbf{w}_{i,j} s_j + z_k, \quad \forall k \in \mathscr{K} \tag{2.2}$$

where $\mathbf{h}_{i,k} \in \mathbb{C}^{M_i}$ is the channel vector from the ith RRH to the kth user and z_k stands for the additive white Gaussian noise (AWGN) at the kth user with mean zero and variance σ_k^2, $\forall k \in \mathscr{K}$. By treating the interference as noise, the received SINR and achievable rate at the kth

user is given by

$$\Gamma_k(\mathbf{w}) = \frac{|\sum_{\forall i \in \mathscr{I}} \mathbf{h}_{i,k}^{\mathrm{H}} \mathbf{w}_{i,k}|^2}{\sum_{\forall j \neq k} |\sum_{\forall i \in \mathscr{I}} \mathbf{h}_{i,k}^{\mathrm{H}} \mathbf{w}_{i,j}|^2 + \sigma_k^2} \tag{2.3}$$

$$R_k(\mathbf{w}) = \log_2(1 + \Gamma_k((\mathbf{w}))) \tag{2.4}$$

To describe the RRHs clustering scheme in which the k user is served by a set of RRHs, the indicator function $\mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right)$ for the association status of the kth user and ith RRH is introduced as

$$\mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right) = \begin{cases} 1 & \text{, if } \left\|\mathbf{w}_{i,k}\right\|^2 > 0 \\ 0 & \text{, otherwise} \end{cases} \quad \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{2.5}$$

The cluster which serves the kth user now can be expressed as $\mathscr{V}_k = \left\{ i \middle| \mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right) = 1, \forall i \in \mathscr{I} \right\}$. It can be seen that if the ith RRH serves the kth user, the BBU pool needs to send the user data message s_k along with the beamforming coefficient $\mathbf{w}_{i,k}$ to the ith RRH through the corresponding fronthaul. Hence, the total accumulated data rates of those users served by the ith RRH must satisfy the maximal capacity allowed at the ith fronthaul link in the following constraint

$$\sum_{\forall k \in \mathscr{K}} \mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right) R_k(\mathbf{w}) \leq C_i \tag{2.6}$$

In (Ha *et al.*, 2016), the problem that jointly optimizes the set of RRHs serving each user and precoding in C-RAN to minimize overall power consumption under the rate requirement QoS

constraint and explicit per-fronthaul capacity constraint is formulated as

$$\min_{\mathbf{w}} \sum_{\forall i \in \mathscr{I}} \sum_{\forall k \in \mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \tag{2.7a}$$

$$\text{s.t.} \sum_{k \in \mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq P^{\max} \tag{2.7b}$$

$$\Gamma_k(\mathbf{w}) \geq \Gamma_k^{\max} \tag{2.7c}$$

$$\sum_{\forall k \in \mathscr{K}} \mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right) R_{i,k}^{\text{fh}} \leq C_i \tag{2.7d}$$

It is worth mentioning that the difficulty of the presented problem in (2.7) lies in the non-convexity of fronthaul capacity constraint in (2.7d) due to the nonsmooth indicator function. To overcome this challenge, the authors in (Ha *et al.*, 2016, 2014) approximated the indicator function into a linear function by applying the conjugate function concept. Particularly, the indicator function $\mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right)$ is replaced by $\nabla f\left(\left\|\mathbf{w}_{i,k}^{(n)}\right\|^2\right) - f^{\star}\left(\left\|\mathbf{w}_{i,k}^{(n)}\right\|^2\right)$ at the $(n+1)$th iteration of the proposed algorithm where $f^{\star}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right)$ is the conjugate function of $f\left(\left\|\mathbf{w}_{i,k}\right\|^2\right)$. Note that, the rate function in the per-fronthaul capacity constraint in (2.7d) is set to a constant $R_{i,k}^{\text{fh}}$, $\forall i \in \mathscr{I}$ and $\forall k \in \mathscr{K}$. Using the constant rates results in the upper bound of the fronthaul transmission rate and does not exactly describe the fronthaul capacity compared to the constraint in (2.6) in the C-RAN system.

To further save the power consumption, the RRH selection is considered in (Dai & Yu, 2016; Pan *et al.*, 2017; Cheng *et al.*, 2013). Particularly, the power consumption at each RRH is categorized into two types: data-dependent power, which is related to the transmitted signal, and data-independent power. The data-independent power can be further sub-categorized into two types: power to keep each ith RRH active, denoted as P_i^{ra}, and power to keep each ith RRH idle, denoted as P_i^{ri}. Denote $\mathbf{w}_i = \left[\mathbf{w}_{i,1}^{\text{T}}, \ldots, \mathbf{w}_{i,K}^{\text{T}}\right]^{\text{T}} \in \mathbb{C}^{(K \times M_i) \times 1}$ the transmit beamforming vector at the ith RRH. By defining the indicator function $\mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right), \forall i \in \mathscr{I}$ to represent the operation mode of each ith RRH where $\mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right) = 0$ indicates that the ith RRH is in sleep mode if $\left\|\mathbf{w}_i\right\|^2 = 0$ and $\mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right) = 1$ otherwise if $\left\|\mathbf{w}_i\right\|^2 > 0$, the total power consumption at

the *i*th RRH is calculated as

$$P_i^{\mathrm{RRH}}(\mathbf{w}) = \frac{1}{\eta_i}\sum\nolimits_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 + \mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right) P_i^{\mathrm{ra}} + \left(1-\mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right)\right) P_i^{\mathrm{ri}} \tag{2.8}$$

Thus, designing the set of active RRHs $\mathscr{I}^a = \left\{i|\mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right) = 1, \forall i \in \mathscr{I}\right\}$, the RRH clusters $\mathscr{V}_k, \forall k \in \mathscr{K}$ and the transmit beamforming to minimize the total power consumption while still guaranteeing the required QoS is of the interest. For example, the work in (Dai & Yu, 2016) aimed at minimizing the total power consumption of RRHs and corresponding fronthauls, in which a more accurate analysis of non-linear power consumption that takes into account the amount of fronthaul power usage was considered and a low complexity algorithm was called to solve the following problem

$$\min_{\mathbf{w}} \sum_{\forall i\in\mathscr{I}} \left(P_i^{\mathrm{RRH}}(\mathbf{w}) + \rho_i \sum_{\forall k\in\mathscr{K}} \mathbb{I}\left(\left\|\mathbf{w}_{i,k}\right\|^2\right) R_k(\mathbf{w})\right) \tag{2.9a}$$

$$\text{s.t.} \sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq P^{\max} \tag{2.9b}$$

$$\Gamma_k(\mathbf{w}) \geq \Gamma_k^{\max} \tag{2.9c}$$

The authors stated that the minimum SINR constraints in (2.9c) are necessarily met with equality at the optimal solution. Therefore, the rate function in the fronthaul power consumption part of the objective is replaced by the minimum rate and the difficulty only remaining in the said problem is the non-convexity of the indicator function. The authors in (Dai & Yu, 2016) proposed to approximate the nonconvex indicator function by using the reweighted convex ℓ_1-norm as

$$\mathbb{I}\left(\left\|\mathbf{w}_i\right\|^2\right) = \left\|\sum_{\forall k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|^2\right\|_0 = \frac{c}{\sum_{\forall k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}^{(n)}\right\|^2 + \tau} \sum_{\forall k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|^2 \tag{2.10}$$

where c and τ are small and positive constants and $\mathbf{w}_{i,k}^{(n)}$ are parameters obtained from the previous iteration of the proposed algorithm. Similarly, the authors in (Pan *et al.*, 2017) adopted

the reweighted ℓ_1-norm approximation to solve the two stage problem of user-centric clustering and RRH selection. In particular, a user selection algorithm was developed to maximize the subset of admitted users in the first stage and the RRH selection was performed by an iterative algorithm in the second stage.

In another approach, by using the big-M formulation with binary variables introduction $b_i \in \{0,1\}$ and $a_{i,k} \in \{0,1\}$ which represent for the ith RRH active/sleep and association status between the ith RRH and kth user $\forall i \in \mathscr{I}$, $\forall k \in \mathscr{K}$, respectively, the authors in (Cheng *et al.*, 2013) addressed the total power consumption minimization problem as the follows

$$\min_{\mathbf{w}} \sum_{\forall i \in \mathscr{I}} \left(\frac{1}{\eta_i} \sum_{\forall k \in \mathscr{K}} \left\| \mathbf{w}_{i,k} \right\|^2 + b_i P_i^{\mathrm{ra}} + (1 - b_i) P_i^{\mathrm{ri}} + \sum_{\forall k \in \mathscr{K}} a_{i,k} P_{i,k}^{\mathrm{fh}} \right) \tag{2.11a}$$

$$\text{s.t.} \sum_{k \in \mathscr{K}} \left\| \mathbf{w}_{i,k} \right\|^2 \leq b_i P^{\max} \tag{2.11b}$$

$$\Gamma_k(\mathbf{w}) \geq \Gamma_k^{\max} \tag{2.11c}$$

$$\left\| \mathbf{w}_{i,k} \right\|^2 \leq a_{i,k} P^{\max} \tag{2.11d}$$

$$a_{i,k} \leq b_i \tag{2.11e}$$

$$\sum_{\forall i \in \mathscr{I}} a_{i,k} \geq 1 \tag{2.11f}$$

where $P_{i,k}^{\mathrm{fh}}$ is the fixed fronthaul power consumption for the ith fronthaul to convey the data message for the kth user, $\forall i \in \mathscr{I}$ and $\forall k \in \mathscr{K}$. The above problem arrives at MI-SOCP, which can be solved optimally by applying branch and cut method. In addition, the relaxed-integer programming approach has been adopted to solve the MI-SOCP problem with much reduced the computational complexity, requiring the inflation or deflation methods to gradually update the set of active RRHs or RRH-user association based on the relaxed solutions. The similar method to formulate the problem of RRH selection depending on the traffic density conditions is applied in (Zhao & Wang, 2016) to minimize the power consumption in C-RAN.

The works in (Shi *et al.*, 2016b, 2014; Zhao *et al.*, 2013; Luo *et al.*, 2015) also developed a joint beamforming and BS selection design to minimize the power consumption in C-RANs so

that the related fronthaul capacity required to transport data was implicitly minimized. Unlike the approach in (Cheng *et al.*, 2013), the mixed ℓ_1/ℓ_p norm method was used to induce the sparsity of the beamforming vectors in the cooperative C-RAN in (Shi *et al.*, 2016b, 2014; Luo *et al.*, 2015). Basically, when the *i*th RRH is switched off, the corresponding coefficients in the vector $\tilde{\mathbf{w}}_i = \left[\mathbf{w}_{i,1}^T, \ldots, \mathbf{w}_{i,K}^T\right]^T \in \mathbb{C}^{(K\times M_i)\times 1}$ will be set to zero simultaneously. Similarly, the vector $\mathbf{w} = \left[\tilde{\mathbf{w}}_1^T, \ldots, \tilde{\mathbf{w}}_I^T\right]^T$ has a group sparsity structure where multiple RRHs are switched off when the corresponding block of varibles $\tilde{\mathbf{w}}_i$'s are zeros. Thus, to induce the group sparsity for the beamformers, the weighted mixed ℓ_1/ℓ_2-norm is aplied and the total power minimization problem of RRH selection and beamforming can be expressed as

$$\min_{\mathbf{w}} \sum_{\forall i\in\mathscr{I}} \beta_i \left\|\tilde{\mathbf{w}}_i\right\|_{\ell_2} \tag{2.12a}$$

$$\text{s.t. } (2.9b), (2.9c) \tag{2.12b}$$

where β_i is a weight parameter associated to the *i*th RRH, $\forall i \in \mathscr{I}$. The above formulated problem is SOCP which can solved efficiently. Then, the authors proposed the iterative algorithm that sorts the best candidate RRHs based on the obtained beamformers to achieve the best objective value. In the same way, Zhao *et al.* (2013) used reweighted ℓ_1-norm approximation to minimize the total number users counted in the fronthaul links while maintaining the user QoS requirements.

Inspired by the works in (Cheng *et al.*, 2013; Shi *et al.*, 2016b, 2014), (Luo *et al.*, 2015) further addressed the coupling factor of UL and DL transmissions in C-RAN to resolve the problem of (Shi *et al.*, 2016b, 2014) by exploiting the UL-DL duality. Luo *et al.* (2015) proposed both group sparsity optimization and relaxed integer programming approaches to solve the problem of UL-DL power minimization. The optimal active BSs selection and transceiver design was also considered for the purpose of operational overhead reduction in (Lia *et al.*, 2014). The power beamforming minimization problem was proposed in (Tolli *et al.*, 2011; Dhifallah *et al.*, 2015) by relying on the limited information exchange between BSs. Although the fronthaul power consumption and fronthaul overhead signaling are minimized, however the fronthaul

capacity constraints were not explicitly considered into the optimization problems in these works (Dai & Yu, 2016; Cheng *et al.*, 2013; Shi *et al.*, 2016b, 2014; Zhao *et al.*, 2013; Luo *et al.*, 2015; Lia *et al.*, 2014; Tolli *et al.*, 2011; Dhifallah *et al.*, 2015). This may result in the infeasible solutions in the practical system with limited fronthaul capacity.

While the aforementioned works mainly design the system parameters to obtain the beamforming and formation of user-centric clustering, recent research on context-aware wireless communication has stated the importance of content-centric clustering in the regards of the C-RAN development. In this content-centric context, it is assumed that each RRH is equipped with a cache server to prefetch and store popular download content data locally so that their near-by UEs are more conveniently served. For this caching scheme, only unsaved data requested by the UEs is required to be transmitted from the cloud server via fronthaul to the RRH, and thus, fronthaul traffic was intuitively reduced. To study the effect of the proposed caching scheme compared to conventional C-RAN system, Zhou *et al.* (2015) formulated a problem that jointly designs the multi-cast beamforming and content-centric clustering to minimize the overall system cost, where each fronthaul link cost here is proportional to the associated UE achievable rate and the status of the requested data from that UE in the cache. Thus, the existing solution approaches for this problem category can be adopted. To further investigate the system behavior with caching, the authors in (Li *et al.*, 2016) proposed a criteria to observe the data queuing stability. Then, a problem based on Lyapunov optimization framework was formulated and a low complexity algorithm was developed to attain the desired solution. More works on content-centric cached C-RAN can be found in (Dhifallah *et al.*, 2015; Chen *et al.*, 2016; Ugur *et al.*, 2016) and the references therein.

2.2.2 Rate maximization

Akin to the conventional rate maximization problem in homogeneous and multi-tier heterogeneous networks (Luong *et al.*, 2016c,b, 2014a,b) and by inheriting the design property from the previous power minimization category, the problem of rate maximization in limited fronthaul capacity constrained C-RAN aims to maximize the overall network achievable rate by

jointly (or partly) designing the transmit beamforming and RRH-UE association. Note that in this problem, the non-convex limited fronthaul capacity constraint can no longer be relaxed like in (Zhao *et al.*, 2013; Ha *et al.*, 2016, 2014), so that solving it is highly complicated since it is a non-convex combinatorial optimization problem, which is also NP-hard. More specifically, even if we relax the integer constraints included in the RRH-UE association problem, solving the remaining problem is still intractable due to the existence of the non-convex non-concave rate formulas as the functions of beamforming variables that lies in the objective and the fronthaul constraints. Being restrained by these difficulties, an attempt to design a viable and efficient algorithm with polynomial-time complexity is more reasonable and realistic than achieving the optimal solution via exhaustive search. The survey of rate maximization problem in limited fronthaul C-RAN is summarized in (Dai & Yu, 2014; Ha *et al.*, 2015; Park *et al.*, 2016, 2013b,c,a; Zhou & Yu, 2014; Nguyen *et al.*, 2018b; Nguyen *et al.*).

In the literature, the method mostly used to tackle the non-convex weighted sum-rate maximization is relaxing it to a weighted sum-mean square error (MSE) minimization problem, named the weighted minimum mean square error (WMMSE) method. Based on this observation, the first works of weighted sum-rate maximization in limited fronthaul C-RAN were studied in (Dai & Yu, 2014; Ha *et al.*, 2015) by jointly optimizing the transmit beamforming and UE scheduling with the imposition of explicit per-fronthaul capacity constraints, which is formulated as

$$\max_{\mathbf{w}} \sum_{\forall k \in \mathscr{K}} \alpha_k R_k(\mathbf{w}) \tag{2.13a}$$

$$\text{s.t.} \sum_{k \in \mathscr{K}} \left\| \mathbf{w}_{i,k} \right\|_2^2 \leq P^{\max} \tag{2.13b}$$

$$\sum_{\forall k \in \mathscr{K}} \mathbb{I}\left(\left\| \mathbf{w}_{i,k} \right\|^2 \right) R_k(\mathbf{w}) \leq C_i \tag{2.13c}$$

where α_k, $\forall k \in \mathscr{K}$ is the priority weight associated with the kth user and the achievable rate $R_k(\mathbf{w})$ is computed as $R_k(\mathbf{w}) = \log\left(1 + \mathbf{w}_k^H \mathbf{H}_k^H \left(\sum_{j \neq k} \mathbf{H}_j \mathbf{w}_j \mathbf{w}_j^H \mathbf{H}_k^H + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{H}_k \mathbf{w}_k\right)$. Here, denote $\mathbf{H}_k = \left[\mathbf{H}_{1,k}, \ldots, \mathbf{H}_{I,k}\right] \in \mathbb{C}^{N \times M}$ as the channel matrix composing the channel matrix from all RRHs to the kth user, where $\mathbf{H}_{i,k} \in \mathbb{C}^{N \times M_i}$ is the channel state information matrix

from M_i antennas in the ith RRH to the kth user equipped with N antennas and $M = \sum_{\forall i \in \mathscr{I}} M_i$. In this work, the authors took advantage of the ℓ_0-norm to expose the relationship between beamforming design and the RRH-UE connection to redirect the problem into non-smooth non-convex group sparsity optimization problem. In particular, they proposed an iterative low complexity algorithm based on the reweighted ℓ_1-norm approximation of the ℓ_0-norm as calculated in (2.10), which is used to overcome the nonconvexity of indicator function in the fronthaul capacity constraints in (2.13c). Additionally, to tackle the nonconvex rate function in the fronthaul capacity constraint in (2.13c), achievable user rate $R_k(\mathbf{w})$ is assigned to a fixed value $\tilde{R}_k$ which is computed from the beamforming values obtained in the previous iteration of the proposed algorithm. By doing this way, the authors in Dai & Yu (2014); Ha *et al.* (2015) applied the WMMSE method to sequentially solve the convex approximated problem and update respective parameters until convergence as below

$$\max_{\mathbf{w},\{\rho_k,\mathbf{u}_k\}_{\forall k\in\mathscr{K}}} \sum_{\forall k\in\mathscr{K}} \alpha_k \left(\log \rho_k - \rho_k e_k\right) \tag{2.14a}$$

$$\text{s.t.} \sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq P^{\max} \tag{2.14b}$$

$$\sum_{\forall k\in\mathscr{K}} \frac{c\left\|\mathbf{w}_{i,k}\right\|_2^2}{\left\|\mathbf{w}_{i,k}^{(n)}\right\|_2^2 + \tau} \tilde{R}_k \leq C_i \tag{2.14c}$$

where $\mathbf{u}_k \in \mathbb{C}^N$ and ρ_k are denoted for a receive beamforming vector and the MSE weight at the kth user. Note that, the MSE e_k is given by

$$\begin{aligned} e_k &= \mathbb{E}\left[\left\|\mathbf{u}_k^H \mathbf{y}_k - s_k\right\|_2^2\right] \\ &= \mathbf{u}_k^H \left(\sum_{\forall j\in\mathscr{K}} \mathbf{H}_k \mathbf{w}_j \mathbf{w}_j^H \mathbf{H}_k^H + \sigma^2 \mathbf{I}\right) \mathbf{u}_k - 2\Re\left\{\mathbf{u}_k^H \mathbf{H}_k \mathbf{w}_k\right\} + 1 \end{aligned} \tag{2.15}$$

The problem (2.14) can be solved by using a block coordinate descent (BCD) method that allows to iteratively optimize over ρ_k, $\mathbf{u}_k$, $\mathbf{w}$ as the following. For a given $\mathbf{w}^{(n)}$, the receive

beamforming $\mathbf{u}_k^{(n)}$ to minimize MSE at the kth user can be obtained as

$$\mathbf{u}_k^{(n)} = \arg\min_{\mathbf{u}_k} \mathbb{E}\left[\left\|\mathbf{u}_k^H \mathbf{y}_k - s_k\right\|_2^2\right] = \left(\sum_{\forall j \in \mathscr{K}} \mathbf{H}_k \mathbf{w}_j^{(n)} \mathbf{w}_j^{(n)H} \mathbf{H}_k^H + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{H}_k \mathbf{w}_k^{(n)}, \forall k \in \mathscr{K} \tag{2.16}$$

Given fixed beamforming $\mathbf{w}^{(n)}$ and receive beamforming $\mathbf{u}_k^{(n)}$, the MSE weight $\rho_k^{(n)}$ can be determined by

$$\rho_k^{(n)} = \left(e_k^{(n)}\right)^{-1} \tag{2.17}$$

Then, the optimal beamforming $\mathbf{w}$ can be found under fixed ρ_k and $\mathbf{u}_k$, $\forall k \in \mathscr{K}$, by solving the following quadratic problem

$$\max_{\mathbf{w}} \sum_{\forall k \in \mathscr{K}} \mathbf{w}_k^H \left(\sum_{\forall j \neq k} \alpha_j \rho_j^{(n)} \mathbf{H}_j^H \mathbf{u}_j^{(n)} \mathbf{u}_j^{(n)H} \mathbf{H}_j\right) \mathbf{w}_k - 2 \sum_{\forall k \in \mathscr{K}} \alpha_k \rho_k^{(n)} \Re\left\{\mathbf{u}_k^{(n)H} \mathbf{H}_k \mathbf{w}_k\right\} \tag{2.18a}$$

$$\text{s.t. } (2.14b), (2.14c) \tag{2.18b}$$

However, the convergence proof of proposed WMMSE algorithm in (Dai & Yu, 2014; Ha *et al.*, 2015) could not be justified.

In contrast to the WMMSE approach, the successive convex approximation (SCA) method is another approach proved to be very useful in improving the convergent speed in the view of optimization. Inspired by (Nguyen *et al.*, 2014a), Tran & Pompili (2017a) proposed the SCA based algorithm to solve the sum rate maximization problem with the computing-capacity constraint for the C-RAN system. Particularly, the fixed cluster decision is assumed to be known and an iterative SOCP algorithm is developed to solve the fixed cluster based problem without the computing-capacity constraint. The resulting solution is verified against the computing-capacity constraint to obtain finally the beamforming solution of the original problem. The SCA method was again applied in (Parsaeefard *et al.*, 2017) to solve the sum rate maximization problem, considering a single pair of RRH and BBU assinged to serve each user and effects of pilot contaminations in C-RAN.

Alternatively, the join design of fronthaul and radio access links for C-RAN with wireless fronthauling was studied to maximize the weighted sum rate in (Park *et al.*, 2016; Nguyen *et al.*, 2016b). The difference of convex (DC) programming approach is applied to the problem to achieve a sequence of monotonically nondecreasing objective values at each iteration. Park *et al.* (2013b) studied the impact of different multi-rate compression strategies on the DL transmissions of C-RAN and proposed an iterative majorization algorithm to solve for the suboptimal precoding solution. Then, the authors continued to study the compression scheme of the UL of C-RAN in (Park *et al.*, 2013c) by considering distributed source coding strategies to effectively exploit the correlation of the received signal and proposed a similar iterative approach to solve for solution. In (Zhou & Yu, 2014), the quantization noise levels were optimized for the weighted sum achievable rate maximization problem through the Wyner-Ziv model and single-user compression scheme. Results in this work showed that the near-optimal solution can be achieved when the quantization noise level were set proportional to the background noise level at high signal-quantization-noise ratio regime.

2.2.3 Energy Efficiency Maximization

Along with the demand of meeting higher network throughput performance, achieving greener communication is also considered as an active research trend in the limited C-RAN (Peng *et al.*, 2014b; Luong *et al.*, 2011, 2018a). In particular, how green the communication can be relies on the number of data bits over the energy unit Joule, which is also known as energy efficiency (EE). According to the report in (Auer *et al.*, 2011), almost 80% the total network energy is spent at BS sites; thus, in a dense C-RAN deployment, saving more energy by operating less power at the RRHs to maintain an acceptable throughput leads to greener and more economical communications. Motivated by the need of EE amendment, which also helps in lowering operational costs for mobile network operators and contributes to the decrease of CO_2 emissions, optimally managing the radio resource is essential to attain the best system EE.

Devising a radio resource allocation to optimally maximize the EE in a limited C-RAN is much more difficult and complicated than a general EE in cellular wireless communication. Although

we can transform the conventional nonlinear fractional form of EE into a linear subtractive form to reduce the problem complexity (Dinkelbach, 1967), solving the EE maximization in C-RAN must abide by the limited fronthaul capacity. On the other hand, by involving in the design of active clustering and UE association that allows to switch the RRH on/off to conserve more power and select the "good" UE served by a subset of coordinated RRHs to achieve better throughput, this EE problem normally falls into a combinatorial non-convex optimization problem. Because of these challenges, it is important to reconstruct the problem of interest into a more tractable form where the combination of methods used to solve the power minimization and sum rate maximization can be flexibly adopted to develop the polynomial time complexity algorithm and obtain a "high-quality" solution. In fact, to facilitate the EE solution, the work in (Dai & Yu, 2016; Shi *et al.*, 2014, 2016b) proposed to turn it into a problem of minimizing the system cost or power consumption with additional QoS rate requirement constraint, that were presented to the section 2.2.1.

In (Peng *et al.*, 2016c; Li *et al.*, 2016; Luong *et al.*, 2012), the authors visited the EE maximization problem in limited C-RAN with queue-aware assumption. In this case, the objective function of weighted EE utility at time slot t is defined by equivalent EE metric as follows

$$\eta_{EE}(t) = \frac{\alpha}{K} \sum_{\forall k \in \mathscr{K}} \frac{R_k(t)}{\omega_k} - \frac{1-\alpha}{I} \sum_{\forall i \in \mathscr{I}} \frac{P_i(t)}{\mu_i} \tag{2.19}$$

where α is the weighting factor representing the ratio of the achievable rate to the power consumption. ω_k (b/s/Hz) and μ_i (Watts) represent the transmission rate and power consumption reference, respectively. Thus, the stochastic optimization is applied to solve the problem as $\max_{\mathbf{w}} \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\eta_{EE}(\tau)\}$. An additional constraint that covers the queue stability behavior is proposed together with the upper bound limited fronthaul capacity. The optimization problem is reformulated as an equivalent sum-MSE minimization problem, where a low complexity algorithm based on reweighted ℓ_1-norm and WMMSE method (cf. presented in section 2.2.2) is proposed to sequentially solve the a block of variables while fixing the other blocks and update corresponding parameter until convergence.

2.3 Optimization-Based Virtual Computing and Radio Resource Allocation Design in C-RAN

As the C-RAN definition that merges the cloud computing and RAN together, studies on the integration of powerful computing resources and wireless network resources to boost the overall C-RAN network performance recently emerged. Various works as in (Tang *et al.*, 2015, 2017; Guo *et al.*, 2016b; Wang *et al.*, 2016b) that jointly optimize the computing resources and radio resources have been proposed in the literature. Tang *et al.* (2015, 2017) stated that applying the virtualization technique and resource sharing to the centralized BBU pool can reduce the computing power consumption and improve the hardware utilization. Through virtualizing the computing resources in the physical servers into many virtual machines (VMs), the VMs provisioning schemes that allow to elastically scale service capacities depending on the traffic requirements in the cloud-based BBU pool were proposed in (Tang *et al.*, 2015, 2017). Compared to the previous works where minimize only the RRH power consumption, the total system cost consisting of computational power consumption and RRH power consumption was minimized in Tang *et al.* (2015, 2017) with the slightly modified objective function as

$$m\phi + \sum_{\forall i \in \mathscr{I}} P_i^{\mathrm{RRH}}(\mathbf{w}) \tag{2.20}$$

where m is a variable representing the number of VMs needed to process the users's traffic and ϕ is the cost associated with a VM. Since this kind of optimization problem is formulated as MINLP, the authors proposed the integer search approach to perform a search for number of VMs m which minimizes the objective function value. Under fixed m, the problem is reduced to the RRH power minimization problem which can be solved by adopting the presented methods such as reweighted ℓ_1-norm relaxation. A similar idea of VM assignment provisioning in combination with a hybrid clustering scheme was considered in (Guo *et al.*, 2016b) to minimize the total power consumption in C-RAN. Particularly, the power consumption at the VM m determined by VM assignment $x_{m,k} \in \{0,1\}$ between VM m and user k is expressed as $\sum_{\forall k \in \mathscr{K}} x_{m,k}\phi_m$ where ϕ_m is the cost associated to VM m. The hybrid clustering scheme is defined by variables $t_{i,k} \in \{0,1\}$, e.g., $t_{i,k} = 1$ states the ith RRH transmits to the kth user and

versus, and variables $a_{i,k} \in \{0,1\}$, e.g., $a_{i,k} = 1$ states the ith RRH does not transmit the desired data to the kth user but avoid interfering with the kth user, otherwise $a_{i,k} = 0$, $\forall k \in \mathscr{K}$, $\forall i \in \mathscr{I}$. Thus, the system power consumption minimization proposed in Guo *et al.* (2016b) is given by

$$\min_{\mathbf{t},\mathbf{a},\mathbf{x}} \sum_{\forall m \in \mathscr{M}} \sum_{\forall k \in \mathscr{K}} x_{m,k}\phi_m + \eta \sum_{\forall i \in \mathscr{I}} \{ \sum_{\forall k \in \mathscr{K}} p_{i,k} + \left\| \sum_{\forall k \in \mathscr{K}} t_{i,k} \right\|_0 P_i^{\mathrm{ra}} \} \tag{2.21}$$

where the transmit power $p_{i,k}$ can be calculated with equal transmit power as $P^{\max}/S_i$ where S_i is the maximum number of users served by the ith fronthaul.

2.4 Conclusion

The state-of-the-art of recent results on C-RAN limited fronthaul capacity has been surveyed in this chapter. In particular, we present a survey of analytical results and statistical expressions of network achievable rate for general CoMP and C-RAN in random small scale and large scale scenarios. In 2.2 and 2.3, we review the virtual computing and radio resource designs based on solving an optimization problem that are used to achieve the best network performance in the considered limited fronthaul C-RAN. The survey has shown that the existing approaches were not efficient and fully exploited the potential performance of C-RAN. This motivates us to develop more effective and practical methods, which will be presented in the subsequent chapters.

CHAPTER 3

OPTIMAL JOINT REMOTE RADIO HEAD SELECTION AND BEAMFORMING DESIGN FOR LIMITED FRONTHAUL C-RAN

Phuong Luong [1], François Gagnon[1], Charles Despins[1], Le-Nam Tran[2]

[1] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3,
[2] School of Electronical and Electronic Engineering, University College Dublin
4 Dublin, Ireland

This article was published in *IEEE Transaction on Signal Processing* in August 2017 (Luong *et al.*, 2017c).

3.1 Introduction

Recently, cloud radio access networks (C-RANs) have been considered as a key technology to significantly enhance network performance in order to cope with the explosive demand expected in the foreseen 5G networks (Rost *et al.*, 2014). By merging the capability of cloud computing and radio frequency (RF) transmissions, C-RAN architectures are anticipated to use low-cost low-power base stations for radio services while embracing coordinated and centralized computational tasks at the cloud center to achieve higher network performance. Generally, C-RAN systems contain several low-power RRHs that are all connected to a baseband unit (BBU) pool through fronthaul links (Wu *et al.*, 2015), as illustrated in Fig. 3.1. In C-RANs, RRHs equipped with RF modules only account for compression and transmission/reception of radio signals to/from user equipment (UEs). The fronthaul links connecting RRHs and the BBU pool play a role as a data signal transportation media towards/backwards the BBU pool from/to RRHs. On the BBU side, the joint centralized processing task powered by multiple advanced computer processing units (CPUs) is executed to handle all the relevant baseband signals. With this architectural advantage, C-RANs are able to cater both effective interference management and cooperative gains, thereby increasing system capacity. However, the performance of a C-RAN is heavily restricted by the limited fronthaul capacity between RRHs and

the BBU pool (Peng *et al.*, 2015; Bernardos *et al.*, 2013). This creates a fundamental bottleneck on the network operation, which requires appropriate management on the selection and transmit power design at RRHs to attain the optimal performance.

Achieving the maximal achievable sum rate with minimal amount of available resources is a vital problem in wireless networks in general and in C-RANs in particular. The number of RRHs, together with the associated fronthaul links, in C-RANs can be very high, which results in huge power consumption. In this regard, RRH selection and RRH-user association problem is of particular interest. This should be done in accordance with limited fronthaul capacity constraints, which present a new challenge in the design of C-RANs. Consequently, the existing design techniques for conventional wireless communication networks are no longer applicable and thus new design methods for C-RANs are required.

There have been several pioneer works that study the joint design of RRH-user (UE) association and beamforming in C-RAN with limited fronthaul capacity. For example, the work in (Park *et al.*, 2013b; Zhou & Yu, 2014) proposed various compression techniques to minimize the transmitted data delivered over the fronthaul network. In (Fan *et al.*, 2016), Fan *et. al.* developed a low-complexity and efficient algorithm to form clusters of RRHs so that the number of centralized computational processing tasks at the BBU pool was greatly reduced. In (Shi *et al.*, 2014; Zhao *et al.*, 2013), the authors employed a sparsity inducing-norm to develop a joint beamforming and base station (BS) selection design to minimize the power consumption in C-RANs so that the related fronthaul capacity required to transport data was implicitly minimized. Inspired by these works, the authors in (Luo *et al.*, 2015) further addressed the coupling factor of uplink (UL) and downlink (DL) transmissions in C-RANs to resolve the problem of (Shi *et al.*, 2014) by exploiting the UL-DL duality and MI-SOCP framework. In (Ramamonjison *et al.*, 2014), the authors employed a generalized Bender decomposition (GBD) method to develop a decentralized algorithm that jointly optimizes the beamforming and BS clustering under the limited message exchange assumption in cognitive radio networks. The work of (Ng & Schober, 2015) considered the limited backhaul constraint and formulated a power minimization problem as a combinatorial non-convex problem, where different resource allocation

algorithms based on GBD combined with semidefinite programming and difference of convex programming were derived. In (Niu *et al.*, 2016), an increment-based greedy allocation algorithm was proposed to solve the problem of resource allocation and user association through a user-centric resource sharing scheme for a C-RAN with fronthaul capacity constraint. In addition, the authors in (Dai & Yu, 2014; Ha *et al.*, 2016) explicitly incorporated the per-fronthaul capacity constraints in their optimization problems and applied different methods based on group sparsity inducing norms to attain their designs. In (Peng *et al.*, 2016a), the authors assigned the fixed rates in the previous iteration to overcome the non-convexity of fronthaul capacity constraints and applied a generalized WMMSE method to solve the problem of energy efficiency maximization in queue-aware H-CRAN. Sun *et al.* (2016); Zhao *et al.* (2016) developed the coalitional formation game based algorithm to form an RRH cluster, while a contract game based interference coordination was proposed in (Peng *et al.*, 2015b). Using the approximation Bellman equation, the authors in (Lau *et al.*, 2013) derived a close-form approximation function for the problem of power-delay trade-off for MU-MIMO systems. To develop an optimal algorithm for resource allocation in C-RANs, branch and bound method was used in (Cheng *et al.*, 2013; Tang *et al.*, 2015) and the dual decomposition method was exploited in (Peng *et al.*, 2015a; Li *et al.*, 2016). An exhaustive search was adopted in (Ha *et al.*, 2016; Guo *et al.*, 2016a) to find the optimal RRH cluster. However, the authors in (Ramamonjison *et al.*, 2014; Luo *et al.*, 2015; Cheng *et al.*, 2013; Dai & Yu, 2016; Tang *et al.*, 2015; Peng *et al.*, 2015a; Li *et al.*, 2016; Peng *et al.*, 2015b) did not explicitly consider the fronthaul capacity constraints, while the user rates were set to be constant to overcome the non-convex fronthaul capacity constraints in (Dai & Yu, 2014; Ha *et al.*, 2016; Ng & Schober, 2015; Niu *et al.*, 2016; Guo *et al.*, 2016a; Peng *et al.*, 2016a).

From a network optimization perspective, total power consumption minimization and achievable sum rate maximization are the two most common performance metrics when designing wireless communications. However, these two design criteria have been often considered separately as their goals are conflicting. Note that, by weighing the two objectives, we can find the whole rate and power region of the system (Boyd & Vandenberghe, 2004; Wu *et al.*, 2014).

This is in close relation to energy-efficient transmission strategies (Vu *et al.*, 2016). The research on transmit power-throughput trade-off was considered in (Manosha *et al.*, 2014; Yadav *et al.*, 2016b,a), where the convex hull of the entire achievable power-rate region of MIMO heterogeneous networks was obtained. By an MI-SOCP approach, a mechanism to find the optimal trade-off between the overall BS power consumption and power consumption overhead associated with CoMP transmission was proposed in (Cheng *et al.*, 2013). The work in (Luong *et al.*, 2016a) was the first to employ the MI-SOCP approach to study the power minimization in limited C-RANs. The works of (Dai & Yu, 2016) adopted the reweighted ℓ_1-norm to study the trade-off between total power consumption and fronthaul capacity for data sharing and compression strategies in C-RANs.

It is worth mentioning that the studies in relation to the overall power consumption minimization in C-RANs implicitly imply the minimization of the fronthaul capacity usage. This also helps the cloud center to use the least computational resource to satisfy QoS requirements. Investigations on how C-RANs can benefit from cloud computing capabilities have been reported recently. For example, the works of (Tang *et al.*, 2015), (Pompili *et al.*, 2016) proposed a joint design of virtual machine computation capacity, RRH selection and beamforming to minimize the total power consumption in C-RANs. In this paper, we focus on the communications part of C-RANs rather than the cloud computing capabilities.

In this paper, we consider the downlink transmission of a C-RAN with limited fronthaul capacity. In the considered system model, digital data is transmitted from the BBU pool to RRHs using fronthaul links of finite capacity, and beamforming technique is used to send data to UEs. Under this context, we study a joint design of RRH selection, RRH-UE association and beamformer that simultaneously maximizes the achievable sum rate and minimizes the total power consumption. The main motivation for jointly designing beamforming with RRH selection, and RRH-UE association is due to the design goal. It is true that for spectral efficiency maximization, we do not need to consider the RRH-UE association and RRH selection since maximum degree of freedom is achieved if all RRHs are allowed to serve all the UEs in the system. We note that the objective function in our problem strikes the balance between spectral

efficiency maximization and total power minimization. Thus, RRH-UE association and RRH selection are particularly relevant. Intuitively, the optimization of RRH-UE association and RRH selection is important because there exists a situation where some RRHs of severe fading conditions can be switched off and each UE can be served by a small subset of active RRHs to save power.

To deal with two conflicting targets, we formulate the problem of interest as a multi-objective (or vector) optimization problem, directly solving which is cumbersome. To overcome this difficulty, we propose to employ the scalarization approach for the formulated problem by linearly combining each weighted element of the vector objective function to result in a standard scalar optimization problem. As shown later on, two challenges arise in the considered problem: (i) the non-convexity of per-fronthaul capacity constraints, and (ii) the combinatorial nature of the selection procedure. To deal with the latter one, we naturally introduce binary selection variables to represent the selection status of RRHs and associated users. The formulated problem is basically a combinatorial one, which is generally NP-hard. Moreover, another problem is that even if these binary selection variables are relaxed to be continuous, the resulted problem is still non-convex because of the non-convexity of the objective function and per-fronthaul capacity constraints. This attribute makes the considered problem much more difficult to solve, and the methods presented in previous studies such as those in (Ramamonjison *et al.*, 2014; Luo *et al.*, 2015; Cheng *et al.*, 2013; Dai & Yu, 2016; Tang *et al.*, 2015; Peng *et al.*, 2015a; Li *et al.*, 2016; Peng *et al.*, 2015b) are no longer applicable. Moreover, different from (Dai & Yu, 2014; Ha *et al.*, 2016; Ng & Schober, 2015; Niu *et al.*, 2016; Guo *et al.*, 2016a; Peng *et al.*, 2016a) where the authors simply assign a predetermined achievable rate to overcome the non-convex fronthaul constraint, we directly tackle it by proposing novel transformations to arrive at an equivalent but more tractable form. Based on that, we develop a new framework using SCA method (Beck *et al.*, 2010) to solve the considered problem efficiently. The main contributions of this paper are summarized as follows.

- We formulate the joint design of RRH selection, RRH-UE association and beamforming for achievable sum rate-total power maximization problem by employing the concept of

multi-objective optimization (Boyd & Vandenberghe, 2004). The problem is formulated as a mixed integer nonlinear program. We then present a novel transformation to rewrite the design problem in a form that facilitates a customized branch-and-reduce-and-bound (BRB) algorithm to find a globally optimal solution based on monotonic optimization.

- To overcome the high complexity inherently in a global optimization method, we propose novel transformations and convex approximation techniques to derive two suboptimal low-complexity algorithms aiming at attaining a high-quality feasible solution. More specifically, in the first method, we iteratively approximate the continuous non-convex constraints by convex ones using SCA framework. By using a quadratic bound of the logarithm function, we are able to arrive at a sequence of MI-SOCPs, for which dedicated solvers are available and efficient. The second method is a simplified version of the first one where we further relax the binary variables in each iteration to be continuous. That is to say, each iteration of the second method merely requires solving an SOCP. After convergence, we then perform a post-processing procedure on the relaxed selection variables to search for a high-performance solution.

- From a different viewpoint, we reformulate the considered problem under the concept of sparsity-inducing regularization. The connection status of a particular pair of RRH and UE is represented by the norm of the associated beamforming vector, which is encouraged to be zero if doing so improves the objective. By exploiting a ℓ_2-norm based logarithm approximation, the new optimization problem basically shares the same non-convex structure as the previous one. Applying similar steps in the proposed methods mentioned above, we arrive at an SOCP, but of smaller size, in each iteration. Then, RRH selection and RRH-UE association can be decided by ignoring the zero elements in the obtained sparse solution, after the convergence of the iterative algorithm.

- Extensive numerical results are presented to show the efficiency of our proposed algorithms, compared to known solutions in the literature, especially for the cases of sum achievable rate maximization and power minimization. In particular, compared to the WMMSE approach

in (Dai & Yu, 2014), our proposed SCA-based methods converge much faster, while still achieving a better performance.

The rest of this chapter is organized as follows. Section 3.2 introduces the system model and formulates our joint RRH selection, RRH-UE association and transmit beamformers into an achievable sum rate-total power consumption optimization problem. Section 3.3 provides the global optimal algorithm. In Section 3.4, we introduce our proposed low complexity algorithms. Section 3.5 presents our numerical results and insight discussions under different simulation setups. Finally, the concluding remark of the this work is given in Section 3.6.

3.2 System Model and Problem Formulation

3.2.1 Transmission Model

We consider the DL of C-RAN consisting of I RRHs and K single antenna UEs. For notational convenience, we denote $\mathscr{I} = \{1,\ldots,I\}$ and $\mathscr{K} = \{1,\ldots,K\}$ as the set of RRHs and UEs, respectively. We assume that the ith RRH is equipped with M_i antennas,$\forall i \in \mathscr{I}$. As shown in Fig. 4.1, we assume that all the RRHs are connected to BBU pool via the fronthaul links, e.g., high-speed optical ones, where the ith link has a predetermined maximum capacity C_i. Each UE is served by a specific group of RRHs but one RRH can serve more than one users simultaneously. Let us denote by s_k the signal with unit power, i.e., $\mathbb{E}\left\{s_k s_k^*\right\} = 1$, intended for the kth UE and by $\mathbf{w}_{i,k} \in \mathbb{C}^{M_i \times 1}$ the transmit beamforming vector from the ith RRH to the kth UE. The vector of channel coefficients encompassing small-scale fading and pathloss from the ith RRH to the kth UE is represented by $\mathbf{h}_{i,k} \in \mathbb{C}^{M_i \times 1}$. For notational convenience we denote the set of beamforming vectors intended for the kth UE as $\mathbf{w}_k \triangleq [\mathbf{w}_{1,k}^T, \mathbf{w}_{2,k}^T, \ldots, \mathbf{w}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, and the vector including the channels from all RRHs to the kth UE as $\mathbf{h}_k \triangleq [\mathbf{h}_{1,k}^T, \mathbf{h}_{2,k}^T, \ldots, \mathbf{h}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, where $M = \sum_{i \in \mathscr{I}} M_i$. Using these notations, the received signal at the kth UE is given by

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \sum_{j \in \mathscr{K} \setminus k} \mathbf{h}_k^H \mathbf{w}_j s_j + z_k \tag{3.1}$$

where $z_k \sim \mathscr{CN}(0, \sigma_0^2)$ is the additive white Gaussian noise (AWGN) and σ_0^2 is the noise power. Note that in (3.1), we have assumed that the kth UE is connected to all the RRHs, but the ith RRH serves the kth UE only if $\left\|\mathbf{w}_{i,k}\right\|_2^2 > 0$. By treating interference as noise, the achievable rate in b/s/Hz at the kth UE is given by

$$R_k(\mathbf{w}) = \log_2(1 + \Gamma_k(\mathbf{w})) \tag{3.2}$$

where

$$\Gamma_k(\mathbf{w}) = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j \in \mathscr{K} \setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2} \tag{3.3}$$

and $\mathbf{w} \triangleq [\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_k^T]^T \in \mathbb{C}^{(KM)\times 1}$ is vector stacking the beamformers for all users.

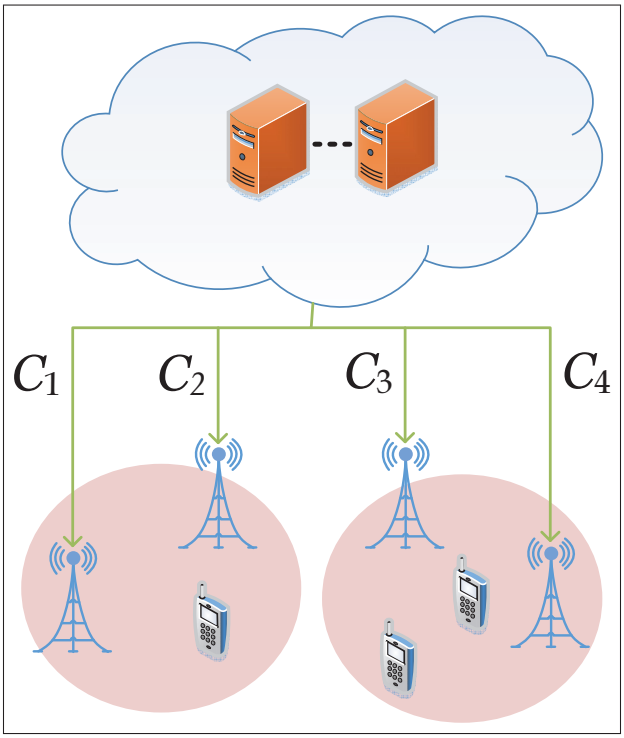


Figure 3.1 Limited fronthaul C-RAN.

3.2.2 Fronthaul Capacity Constraint

After the BBU pool performs a relevant radio resource management algorithm to determine the beamforming vectors, data for the k^{th} UE is routed from the BBU pool to the i^{th} RRH via the i^{th} fronthaul link only if $\left\|\mathbf{w}_{i,k}\right\|_2 > 0$. For the transmission to be feasible, the capacity of the i^{th} fronthaul link should be ξ_i times greater than or equal to the total achievable rate at the i^{th} RRH where $\xi_i \geq 1, \forall i \in \mathscr{I}$ (Peng *et al.*, 2016b). Herein, we assume that the channel conditions are slow varying. Thus, the transportation of CSI via the fronthaul link occurs less frequently than that of data. As a result, conveying CSI consumes much less fronthaul capacity than conveying the users' data, and thus can be neglected for the sake of simplicity. For the purpose of problem formulation, let us introduce binary variables $a_{i,k} \in \{0,1\}, \forall i \in \mathscr{I}$ and $k \in \mathscr{K}$ to represent the association status between the i^{th} RRH and the k^{th} UE, i.e., $a_{i,k} = 1$ implies that the k^{th} UE is served by the i^{th} RRH and $a_{i,k} = 0$, otherwise. Then, the per-fronthaul capacity constraints can be written as

$$\sum_{k \in \mathscr{K}} a_{i,k} R_k(\mathbf{w}) \leq \frac{C_i}{\xi_i}, \forall i \in \mathscr{I}. \tag{3.4}$$

3.2.3 Power consumption

In this subsection, we present a power consumption model that accounts for the power consumption at RRHs as well as for transmitting digital data from the BBU pool to the corresponding RRHs. According to Shi *et al.* (2014), the power consumption at a RRH consists of two types, namely, data dependent power and data independent power. The former is the power dispatched at the power amplifiers in an RRH which is a function of transmitted signals, while the latter is mostly due to electronic components. The data independent power can be sub-categorized into two types, one, denoted by P_i^{ra}, representing the fixed amount of power when the ith RRH is in active mode, and one, denoted by P_i^{ri}, accounting for the power required to keep the ith RRH in sleep mode. More specifically, P_i^{ra} and P_i^{ri} are the power that is consumed by the circuit and to maintain the operation of the fronthaul optical link in the active and sleep mode of the ith RRH, respectively. The power consumption for forwarding

information data and beamformers related to the transmission from the ith RRH to the kth UE via fronthaul tranmission is denoted by $P_{i,k}^{\mathrm{FH}}$. From the introduction of $a_{i,k}$, it is obvious that when $a_{i,k}=0$, then $P_{i,k}^{\mathrm{FH}}=0$. To represent the operation mode of the ith RRH, we introduce a binary variable $b_i=\{0,1\},\forall i\in\mathscr{I}$. In particular, $b_i=0$ states that the ith RRH is in sleep mode and $b_i=1$ means otherwise. In summary, the sum power consumption at all RRHs and corresponding fronthaul links can be written as

$$P^{\mathrm{tot}}(\mathbf{w},\mathbf{a},\mathbf{b})=\frac{1}{\eta_i}\sum_{i\in\mathscr{I}}\sum_{k\in\mathscr{K}}\|\mathbf{w}_{i,k}\|_2^2+\underbrace{\sum_{i\in\mathscr{I}}b_iP_i^{\mathrm{ra}}+\sum_{i\in\mathscr{I}}(1-b_i)P_i^{\mathrm{ri}}+\sum_{i\in\mathscr{I}}\sum_{k\in\mathscr{K}}a_{i,k}P_{i,k}^{\mathrm{FH}}}_{P^{\mathrm{circ}}(\mathbf{a},\mathbf{b})} \tag{3.5}$$

where $\eta_i\in[0,1]$ is the power amplifier efficiency, $\mathbf{b}=[b_1,\ldots,b_I]^T$ and $\mathbf{a}=\left[\mathbf{a}_1^T,\ldots,\mathbf{a}_K^T\right]^T$ where $\mathbf{a}_k=\left[a_{1,k}\ldots,a_{I,k}\right]^T$. For simplicity, we denote $P^{\mathrm{circ}}(\mathbf{a},\mathbf{b})=\sum_{i\in\mathscr{I}}b_iP_i^{\mathrm{ra}}+\sum_{i\in\mathscr{I}}(1-b_i)P_i^{\mathrm{ri}}+\sum_{i\in\mathscr{I}}\sum_{k\in\mathscr{K}}a_{i,k}P_{i,k}^{\mathrm{FH}}$.

3.2.4 Problem Formulation

We are now ready to formulate the problem of simultaneously maximizing the achievable sum rate of K users and minimizing the total power consumption of the considered C-RAN model. These are the most two common design criteria in wireless networks. By optimizing the two performance measures in a single framework, we can achieve maximal sum rate with minimal total power consumption, and also easily trade-off between the two conflicting objectives. In general, this problem is categorized as a multi-objective optimization one, where the objective is a vector-valued function. A common method to solve it is to apply the scalarization method by taking a linear combination of individual components Boyd & Vandenberghe (2004). Motivated by this method, we consider a joint design of beamforming, RRH selection and RRH-UE

association given by

$$\max_{\mathbf{b},\mathbf{a},\mathbf{w},\nu} \ \alpha\frac{R^{\text{tot}}(\mathbf{w})}{R_0} - (1-\alpha)\frac{P^{\text{tot}}\left(\mathbf{w},\mathbf{a},\mathbf{b}\right)}{P_0} \tag{3.6a}$$

$$\text{s.t.} \ \ \Gamma_k(\mathbf{w}) \geq \Gamma_k^{\min}, \forall k \in \mathscr{K} \tag{3.6b}$$

$$\sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq b_i P^{\max}, \forall i \in \mathscr{I} \tag{3.6c}$$

$$\left\|\mathbf{w}_{i,k}\right\|_2^2 \leq a_{i,k}\nu_{i,k}, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{3.6d}$$

$$\nu_{i,k} \leq a_{i,k}P^{\max}, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{3.6e}$$

$$a_{i,k} \leq b_i, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{3.6f}$$

$$\sum_{i\in\mathscr{I}} a_{i,k} \geq 1, \forall k \in \mathscr{K} \tag{3.6g}$$

$$\sum_{k\in\mathscr{K}} a_{i,k}R_k(\mathbf{w}) \leq C_i, \forall i \in \mathscr{I} \tag{3.6h}$$

$$b_i \in \{0,1\}, a_{i,k} \in \{0,1\}, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{3.6i}$$

where $R^{\text{tot}}(\mathbf{w}) \triangleq \sum_{k\in\mathscr{K}} R_k(\mathbf{w})$. In (3.6), we have introduced the weight $\alpha \in [0,1]$ to strike the balance between sum rate maximization and total power minimization. It is worth mentioning that if $\alpha = 1$ (or $\alpha = 0$), we arrive at the sum rate maximization problem (or total power minimization). In addition, due to the different physical meaning of rate and power in the objective, we divide $R^{\text{tot}}(\mathbf{w})$ and $P^{\text{tot}}\left(\mathbf{w},\mathbf{a},\mathbf{b}\right)$ by a reference value R_0 (b/s/Hz) and P_0 (Watt), respectively. The values of R_0 and P_0 are provided in Section 3.5. Before proceeding further, we note that there exist other scalarization techniques such as weighted Tchebycheff Ng *et al.* (2016), weighted exponential and other methods introduced in Marler & Arora (2004) to solve a multi-objective problem. However, the weighted Tchebycheff method is inefficient to the considered problem in this paper because optimizing individual objectives is already intractable. The weighted exponential and other methods in Marler & Arora (2004) essentially lead to a formulation similar to (3.6), and thus the proposed solutions in the subsequent sections are still applicable. In this paper, we adopt the linear scalarization method for its popularity and simplicity.

The introduction of the set of auxiliary variables $\nu = \left\{\nu_{i,k}, \forall i \in \mathscr{I}, k \in \mathscr{K}\right\}$ and the constraints in (3.6) deserve further explanation. Intuitively, $\nu_{i,k}$ represents the *soft* power transmitted from the ith RRH to UE k. Constraint (3.6b) is to ensure the QoS requirement for the kth user, where $\Gamma_k^{\min}$ is the predetermined SINR requirement for the kth user. Moreover, constraint (3.6c) implies that the total transmit power at each RRH is limited by a given budget power $P^{\max}$. The constraints (3.6c) and (3.6d) are to make sure that when the ith RRH is in sleep mode, e.g., $b_i = 0$, no power will be transmitted from it. This can be easily seen as $b_i = 0$, then $a_{i,k} = 0$ for all $k \in \mathscr{K}$ and $\sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 = 0$. Similarly, in (3.6d) we guarantee that the transmit power $\left\|\mathbf{w}_{i,k}\right\|_2^2$ from the ith RRH to the kth user is zero if $a_{i,k} = 0$. The constraint in (3.6e) means that the soft power from the ith RRH to the kth user should not exceed $P^{\max}$. We impose the constraint (3.6g) to ensure that each user is served by at least one RRH. Finally, the per-fronthaul capacity constraint is explicitly presented in (3.6h).

We remark that problem (3.6) includes, as a special case, RRH clustering Ramamonjison *et al.* (2014); Dai & Yu (2014); Cheng *et al.* (2013). Specifically, a dynamic cluster of RRHs can be formed by posing an extra constraint on the variable $\{a_{i,k}\}_{\forall i,k}$, i.e., $\sum_{i\in\mathscr{I}} a_{i,k} \leq \kappa$ where $\kappa \leq I$ to require that each user can only connect to at most κ RRHs instead of all RRHs. In this way, dynamic RRH cluster formation can be optimized through binary variables $a_{i,k} \in \{0,1\}$, $\forall i \in \mathscr{I}$ and $\forall k \in \mathscr{K}$ in each scheduling slot. Exploring the potential gains offered by dynamic RRH clustering deserves a thorough study, and thus is left as future work.

Towards solving (3.6) optimally, we note that the constraint (3.6d) is called a rotated second order cone Cheng *et al.* (2013); Boyd & Vandenberghe (2004). It is trivial to see that (3.6d) can be rewritten as $\left(\frac{a_{i,k}+\nu_{i,k}}{2}\right)^2 - \left(\frac{a_{i,k}-\nu_{i,k}}{2}\right)^2 \geq \left\|\mathbf{w}_{i,k}\right\|_2^2$. Thus (3.6d) is equivalent to the following SOC constraint

$$\frac{a_{i,k}+\nu_{i,k}}{2} \geq \left\| \left[\frac{a_{i,k}-\nu_{i,k}}{2}, \mathbf{w}_{i,k}^T\right]^T \right\|_2, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{3.7}$$

3.3 Global Optimization Method

In this section we present a solution to solve (3.6) optimally. Before proceeding further, we provide some comments on the complexity of (3.6). First, problem (3.6) is a mixed integer non-linear program (MINLP) due to binary variables **a** and **b**, which is generally NP-hard. Moreover, even when **a** and **b** are relaxed to be continuous, the obtained problem is still non-convex because of the non-convexity of the objective function (3.6a) and the constraint (3.6h). In mathematical programming, (3.6) is categorized as a mixed integer non-convex problem for which the method in Cheng *et al.* (2013) is not applicable to find a globally optimal solution. To the best of our knowledge, there is no off-the-shelf solver for (3.6). In what follows, we present an equivalent formulation of (3.6), based on which a BRB algorithm using monotonic optimization is customized to solve it optimally.

3.3.1 Equivalent Formulation

Consider the following problem

$$\max_{\mathbf{b},\mathbf{a},\mathbf{w},\nu,\mathbf{u}} \; f(\mathbf{u}) \triangleq \tilde{\alpha} \sum_{k\in\mathscr{K}} u_k - \bar{\alpha} u_0^{-1} \tag{3.8a}$$

$$\text{s.t. } R_k(\mathbf{w}) \geq u_k, \forall k \in \mathscr{K} \tag{3.8b}$$

$$u_k \geq \log(1+\Gamma_k^{\min}) \tag{3.8c}$$

$$P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \leq u_0^{-1} \tag{3.8d}$$

$$\sum_{k\in\mathscr{K}} a_{i,k} u_k \leq C_i \tag{3.8e}$$

$$u_k \geq 0, k = 0,1,2,\ldots,K \tag{3.8f}$$

$$(3.6c), (3.6e) - (3.6g), (3.6i), (3.7). \tag{3.8g}$$

where $\tilde{\alpha} \triangleq \alpha/R_0$ and $\bar{\alpha} \triangleq (1-\alpha)/P_0$. The key to the development of our proposed optimal solution is due to the following lemma.

Lemma 1. *The formulations in* (3.6) *and* (3.8) *are equivalent in the sense that they have the same optimal objective.*

Proof: The equivalence is due to the observation that at optimality of (3.8), the inequalities $R_k(\mathbf{w}) \geq u_k$ and $P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \leq u_0^{-1}$ must hold with equality. The details of the proof are presented in Appendix 1.

3.3.2 Proposed BRB Solution

While the formulation in (3.8) does not reduce the non-convexity of the considered problem, it facilitates the development of an optimal design based on monotonic optimization. More specifically, it is easy to see that the objective in (3.8) monotonically increases with respect to each entry of $\mathbf{u}$. Thus we can apply a BRB method to solve (3.8) optimally as done in Tervo *et al.* (2015); Tuy & khayya-and P. Thach (2005). We refer the interested reader to Tervo *et al.* (2015); Tuy & khayya-and P. Thach (2005) for a detailed description of a monotonic optimization-based BRB. Herein we present the customized steps required for solving the considered problem. For this purpose, we reuse the definitions and concepts in Tervo *et al.* (2015); Tuy & khayya-and P. Thach (2005) relevant to the development of the proposed BRB. Specifically, we define the compact normal set $\mathscr{Q} = \{\mathbf{u} \in \mathbb{R}_+^{K+1} | (3.8\text{b}) - (3.8\text{g})\}$ and $\mathscr{U} = [\underline{\mathbf{u}}, \bar{\mathbf{u}}]$ as the box that contains all $\mathbf{u}$ feasible to (3.8). The values of $\underline{\mathbf{u}}$ and $\bar{\mathbf{u}}$ are can be computed as follows. From (3.8c), it holds that $u_k \geq \log(1+\Gamma_k^{\min}) = \underline{u}_k,\ \forall k = \{1,\ldots,K\}$. Moreover, we have

$$u_0 \geq \frac{1}{\frac{1}{\eta_i} I \times P^{\max} + I \times P^{\text{ra}} + I \times K \times P_{i,k}^{\text{FH}}} = \underline{u}_0. \tag{3.9}$$

Similarly, an upper bound of $\mathbf{u}$ can be given by

$$u_0 \leq \frac{1}{I \times P^{\mathrm{ri}}} = \overline{u}_0 \tag{3.10}$$

$$u_k \overset{(a)}{\leq} \log(1+\frac{\left|\mathbf{h}_k^H \mathbf{w}_k\right|^2}{\sigma_0^2}) \overset{(b)}{\leq} \log\left(1+\frac{\|\mathbf{h}_k\|_2^2 \|\mathbf{w}_k\|_2^2}{\sigma_0^2}\right)$$

$$\overset{(c)}{\leq} \log\left(1+\frac{I \times P_{\max}\|\mathbf{h}_k\|_2^2}{\sigma_0^2}\right) = \overline{u}_k, \forall k \in \mathscr{K}. \tag{3.11}$$

where (a) is due to omitting the inter-user interference, (b) is the result of applying Cauchy–Schwarz inequality, and (c) is obvious from the power constraint for each $\mathbf{w}_{i,k}$. The main problem in a BRB algorithm using monotonic optimization framework is to check if a given $\mathbf{u}$ belongs to $\mathscr{Q}$ or not. Mathematically we need to solve the following feasibility problem for a given $\mathbf{u}$

$$\text{find } \mathbf{w}, \mathbf{a}, \mathbf{b}, \nu \tag{3.12a}$$

$$\text{s.t. } (3.8b), (3.8c), (3.8d), (3.8e), (3.8f), (3.8g). \tag{3.12b}$$

Similar to (3.26) we can equivalently rewrite (3.8b) as

$$c'\Re(\mathbf{h}_k^H \mathbf{w}_k) \geq \left\|\mathbf{h}_k^H \mathbf{w}_1, \ldots, \mathbf{h}_k^H \mathbf{w}_K, \sigma_0\right\|_2 \tag{3.13}$$

where $c' = \sqrt{\frac{1}{2^{\overline{u}_k}-1}+1}$. Furthermore (3.8d) is equivalent to

$$\frac{u_0^{-1} - P^{\mathrm{circ}}(\mathbf{a},\mathbf{b}) + 1}{2} \geq \left\| \left[\frac{\mathbf{w}_{1,1}^T}{\sqrt{\eta_1}}, \ldots, \frac{\mathbf{w}_{I,K}^T}{\sqrt{\eta_I}}, \frac{u_0^{-1} - P^{\mathrm{circ}}(\mathbf{a},\mathbf{b}) - 1}{2} \right]^T \right\|_2 \tag{3.14}$$

From the above transformations, it is clear that when $\mathbf{u}$ is fixed, (3.12) is a MI-SOCP feasibility problem, which can be solved optimally by (3.12) dedicated solvers such as MOSEK MOS and Gurobi Gurobi Optimization (2015). Despite exponential worst-case complexity, these mixed integer solvers can solve (3.12) reasonably fast in practice, especially when leveraging the distributed and parallel optimization capability in a cloud computing platform.

Based on the above analysis, problem (3.8) can now be expressed as $\max\{f(\mathbf{u})|\mathbf{u}\in\mathscr{Q}\subset\mathscr{U}\}$. First, we check whether $\underline{\mathbf{u}}$ is feasible or not. If so, we apply the proposed BRB algorithm, which is outlined in Algorithm 3.1, to find a globally optimal solution to (3.8). In principle, the proposed BRB algorithm recursively branches a box $\mathscr{B}$ into two smaller boxes, checks the feasibility of each new box, update the current upper and lower bounds by the box reduction and bound computation process, and removes the boxes that do not contain an optimal solution. The details of these operations can be found in Tervo *et al.* (2015), and thus omitted here for space limitation. For our particular problem, the upper and lower bound of a box $\mathscr{B}=[\underline{\mathbf{u}},\bar{\mathbf{u}}]$ is computed $UB(\mathscr{B})=f(\bar{\mathbf{u}})$ and $LB(\mathscr{B})=f(\underline{\mathbf{u}})$, respectively. According to Tuy & khayya-and P. Thach (2005), the proposed BRB algorithm is bound improving and terminates after finitely many iterations for a given desired accuracy level ε.

Algorithm 3.1: Proposed BRB algorithm.

Apply box reduction to $\mathscr{U}$ to obtain $\text{redu}(\mathscr{U})$
$n=1$; $\mathscr{B}_1=\text{redu}(\mathscr{U})$; $\mathscr{D}_1=\{\mathscr{B}_1\}$;$\zeta_1=LB(\mathscr{B}_1)$;
repeat
  Select the box with the largest upper bound to branch: $\mathscr{B}_n=\arg\max_{\mathscr{B}_i\subset\mathscr{D}_n}UB(\mathscr{B}_i)$;
  Branch the box $\mathscr{B}_n$ into two small boxes $\mathscr{B}_n^{(1)}$ and $\mathscr{B}_n^{(2)}$; // *Box Branching* //
  for $j=1:2$ **do**
    Compute lower bound set of $\mathscr{B}_n^{(j)}$, denoted as $\underline{\mathscr{X}}_n^{(j)}=\left\{\underline{\mathbf{u}}_n^{(j)}\right\}$;
    if $\underline{\mathrm{X}}_n^{(j)}$ is feasible **then**
      Apply box reduction to $\mathscr{B}_n^{(j)}$ to obtain $\text{redu}\left(\mathscr{B}_n^{(j)}\right)$; // *Box Reduction* //
      $\underline{\mathscr{X}}_n^{(j)}=\emptyset$;
    end if
    Compute lower bound $LB\left(\text{redu}\left(\mathscr{B}_n^{(j)}\right)\right)$, upper bound $UB\left(\text{redu}\left(\mathscr{B}_n^{(j)}\right)\right)$ from the reduced box; // *Bound Computation* //
  end for
  Update the lower bound: $\zeta_{n+1}=\max\left(LB\left(\text{redu}\left(\mathscr{B}_n^{(1)}\right)\right),LB\left(\text{redu}\left(\mathscr{B}_n^{(2)}\right)\right),\zeta_n\right)$;
  Update the set of boxes: $\mathscr{D}_{n+1}=\left\{\mathscr{D}_n,\mathscr{B}_n^{(1)},\mathscr{B}_n^{(2)}\right\}$;
  Delete the box that do not contain optimal solution:
  $\mathscr{D}_{n+1}=\mathscr{D}_n\setminus\{\mathscr{B}_i|\zeta_{n+1}>UB(\text{redu}(\mathscr{B}_i)),\forall i=1,\ldots,\text{cardinal}(\mathscr{D}_n)\}$; // *Pruning* //
  $n=n+1$;
until $\left|\max_{\mathscr{B}_i\subset\mathscr{D}_n}UB(\text{redu}(\mathscr{B}_i))-\zeta_n\right|\leq\varepsilon$;

3.3.3 Convergence analysis

Algorithm 3.1 is guaranteed to compute an optimal solution to (3.8) and its convergence can be proved using the same arguments as those in Tuy & khayya-and P. Thach (2005), which can be explained as follows. First, the branching rule improves the lower and upper bounds of the objective (3.8a) after every iteration. Specifically, by the updating rule in Step 14, the lower bound is non-decreasing after each iteration. Due to the box reduction and bound computation rule, the upper bound is non-increasing. After a finite number of iterations, Algorithm 3.1 will create a set of boxes that contain an optimal solution, and the gap between the upper bound and lower bound is less than or equal to ε, where ε is a predetermined desired accuracy level.

3.4 Low-complexity Algorithms

In general, computing a globally optimal solution to (3.6) is very difficult and even if possible, it is of little practical use in wireless communications since the channel conditions can change quickly. Thus, the proposed optimal solution presented in the preceding section are mostly useful for benchmarking purposes. For more practically appealing methods, we derive in this section three iterative low-complexity approaches to find a high-quality feasible solution to (3.6). In the first approach, we employ successive convex approximation (SCA) method to convexify the non-convex continuous part of problem (3.6). In this way the problem at each iteration of the proposed algorithm is still an MI-SOCP. However, the number of MI-SOCPs that needs to be solved is significantly reduced, compared to the optimal BRB method, since the SCA-based convexification converges rapidly. In the second approach, we further lower the computational complexity of the first approach by relaxing the binary variables into continuous ones. This results in a series of SOCP being solved until convergence. For a continuous relaxation method, it is generally known that the obtained solution may not produce a high-performance (or even a feasible) solution. To this end, we carry out a post-processing procedure over the obtained solution to search for a high-quality solution. In the final method, the problem is reformulated from the viewpoint of sparsity-inducing regularization by applying reweighted ℓ_1-norm in combination with the SCA-based approximation.

3.4.1 New Equivalent Transformation

We first remark that, although (3.6) and (3.8) are equivalent as shown in Lemma 1, their feasible sets are different. It means a feasible solution of (3.8) might be infeasible to (3.6). This can be verified by observing that in (3.8b), we can find a feasible solution $\bar{\mathbf{w}}, \bar{u}_k, \bar{a}_{i,k}, \forall i,k$ such that $R_k(\bar{\mathbf{w}}) > \bar{u}_k$ and $\sum_{k\in\mathscr{K}} \bar{a}_{i,k} R_k(\bar{\mathbf{w}}) > \frac{C_i}{\xi_i}$, violating constraint (3.6h). In this section we are about to apply SCA optimization to find low-complexity algorithms that provide suboptimality of (3.6). Thus a new transformation with an equivalent feasible set is necessary. To this end, we consider the following formulation

$$\max_{\substack{\mathbf{w},\mathbf{a},\mathbf{b},\\ \nu,\mu,\gamma}} \ \tilde{\alpha} \sum_{k\in\mathscr{K}} \mu_k - \bar{\alpha} P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{3.15a}$$

$$\text{s.t.} \ \log(1+\gamma_k) \geq \mu_k \tag{3.15b}$$

$$\Gamma_k(\mathbf{w}) \geq \gamma_k \tag{3.15c}$$

$$\gamma_k \geq \Gamma_k^{\min} \tag{3.15d}$$

$$(3.6\text{c}), (3.6\text{e}) - (3.6\text{i}), (3.7). \tag{3.15e}$$

It is easy to see that a solution feasible to (3.15) is also feasible to (3.6). Moreover, all the constraints (3.15b)–(3.15c) are active at optimality. Thus (3.15) is an equivalent formulation of (3.6) that serves the purpose mentioned above. From the previous discussions, it is clear that the continuous nonconvexity of (3.15) is due to (3.6h) and (3.15c). To proceed further, we first rewrite (3.6h) as

$$\sum_{k=1}^{K} a_{i,k} \log(1+t_k) \leq \frac{C_i}{\xi_i}, \tag{3.16a}$$

$$\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j\neq k}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2} \leq t_k, \tag{3.16b}$$

where $\mathbf{t} = \{t_k \geq 0, \forall k \in \mathscr{K}\}$ is the set of newly introduced variables. Moreover, with the introduction of additional auxiliary variables $\mathbf{z} = \{z_k \geq 0, \forall k \in \mathscr{K}\}$, we can rewrite (3.16a) as

$$\sum_{k\in\mathscr{K}} a_{i,k}^2/z_k \leq \frac{C_i}{\xi_i}, \tag{3.17}$$

$$1 + t_k \leq e^{1/z_k}. \tag{3.18}$$

A subtle point should be made here. In fact, to arrive at (3.17), we have used the fact that $a_{i,k} = a_{i,k}^2$ for $a_{i,k} \in \{0,1\}$. This maneuver has two purposes. Firstly, (3.17) is SOC representable. Secondly, if $a_{i,k}$ is allowed to be continuous on $[0,1]$, then it holds that $a_{i,k} \geq a_{i,k}^2$. Thus, if $a_{i,k}$ satisfies (3.17), then it also does for (3.16a). This important observation will be exploited to derive a high-performance solution based on the continuous relaxation. To summary, we can equivalently rewrite (3.6) as

$$\max_{\substack{\mathbf{w},\mathbf{a},\mathbf{b},\\ \nu,\mu,\gamma\\ \mathbf{t},\mathbf{z}}} \ \tilde{\alpha} \sum_{k\in\mathscr{K}} \mu_k - \bar{\alpha} P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{3.19a}$$

$$\text{s.t. } \log(1+\gamma_k) \geq \mu_k \tag{3.19b}$$

$$\sum_{j\in\mathscr{K}\setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2 \leq \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\gamma_k} \tag{3.19c}$$

$$1 + t_k \leq e^{1/z_k}, \tag{3.19d}$$

$$\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{t_k} \leq \sum_{j\neq k}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2, \tag{3.19e}$$

$$(3.6e) - (3.6g), (3.6i), (3.6c), (3.7), (3.15d), (3.17). \tag{3.19f}$$

We remark that problem (3.19) is still non-convex but its nonconvexity is easier to handle in light of SCA as demonstrated in the following.

3.4.2 SCA-MISOCP Algorithm

In the first iterative method we preserve the Boolean variables, and only approximate the continuous nonconvex parts of (3.19). In particular, we do so by applying the framework of successive convex optimization. Explicitly, at iteration n of the proposed algorithm, the right side of (3.19c) is simply replaced by its first order Taylor approximation around the points $\mathbf{w}_k^{(n)}$ and $\gamma_k^{(n)}$

$$H\left(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)}\right)=\frac{2\Re\left(\mathbf{w}_k^{(n)H}\mathbf{H}_k\mathbf{w}_k\right)}{\gamma_k^{(n)}}-\frac{\left|\mathbf{h}_k^H\mathbf{w}_k^{(n)}\right|^2}{\gamma_k^{(n)2}}\gamma_k \tag{3.20}$$

where $\mathbf{H}_k\triangleq\mathbf{h}_k\mathbf{h}_k^H$, and we have denoted $\mathbf{w}_k^{(n)H}=\left(\mathbf{w}_k^{(n)}\right)^H$ and $\gamma_k^{(n)2}=\left(\gamma_k^{(n)}\right)^2$ to lighten the notation. In the same way we convexify the right sides of in (3.19d) and (3.19e) by the first order Taylor approximation as

$$F\left(z_k;z_k^{(n)}\right)=e^{1/z_k^{(n)}}-\frac{e^{1/z_k^{(n)}}}{z_k^{(n)2}}\left(z_k-z_k^{(n)}\right) \tag{3.21}$$

$$G\left(\mathbf{w};\mathbf{w}^{(n)}\right)=\sum_{j\neq k}^{K}2\Re\left(\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j\right)-\sum_{j\neq k}^{K}\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j^{(n)}+\sigma_0^2 \tag{3.22}$$

By applying these approximations into the non-convex constraints (3.19c),(3.19d) and (3.19e), we can formulate the mixed integer convex approximation of problem (3.19) at iteration $n+1$ as below

$$\max_{\substack{\mathbf{w},\mathbf{a},\mathbf{b},\\ \nu,\mu,\gamma\\ \mathbf{t},\mathbf{z}}}\ \tilde{\alpha}\sum_{k\in\mathscr{K}}\mu_k-\bar{\alpha}P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{3.23a}$$

$$\text{s.t.}\ \log\left(1+\gamma_k\right)\geq\mu_k \tag{3.23b}$$

$$\sum_{j\in\mathscr{K}\setminus k}\left|\mathbf{h}_k^H\mathbf{w}_j\right|^2+\sigma_0^2\leq H\left(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)}\right) \tag{3.23c}$$

$$1+t_k\leq F\left(z_k;z_k^{(n)}\right) \tag{3.23d}$$

$$\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2/t_k\leq G\left(\mathbf{w};\mathbf{w}^{(n)}\right) \tag{3.23e}$$

$$(3.6\text{e})-(3.6\text{g}),(3.6\text{i}),(3.6\text{c}),(3.7),(3.15\text{d}),(3.17). \tag{3.23f}$$

where $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \gamma^{(n)}$ are the parameters to be updated at the $(n+1)$th iteration.

Remark 1. *Note that all the continuous constraints in* (3.23)*, except* (3.23b)*, are convex quadratic representable. Thus* (3.23) *is recognized as a generic convex mixed-integer program for which dedicated solvers are quite limited. In an effort to preserve the convexity, while still able to avail of more efficient solvers, the authors in Tervo* et al. *(2015) approximate* (3.23b) *by a system of SOC constraints. In this way,* (3.23) *reduces to an MI-SOCP for which dedicated solvers such as MOSEK have proved to be very efficient. However, the number of SOC constraints required to approximate the exponential cone in* (3.23b) *increases quickly with the accuracy.*

In this work we propose a novel approach to transform (3.23) into an MI-SOCP. To this end we first present the following inequality. For any $\gamma_k \geq 0$ it holds that

$$\log(1+\gamma_k) \geq U\left(\gamma_k;\gamma_k^{(n)}\right) = \log\left(1+\gamma_k^{(n)}\right) + \frac{1}{1+\gamma_k^{(n)}}\left(\gamma_k-\gamma_k^{(n)}\right) - \frac{1}{2}\left(\gamma_k-\gamma_k^{(n)}\right)^2 \tag{3.24}$$

In fact $U\left(\gamma_k;\gamma_k^{(n)}\right)$ is a quadratic lower bound of $\log(1+\gamma_k)$ around $\gamma_k^{(n)}$, which is derived from the Lipschitz continuity of the derivative of $\log(1+\gamma_k)$. The proof is given in Appendix 2. In the MI-SOCP formulation of (3.23) we replace (3.23b) by

$$U(\gamma_k;\gamma_k^{(n)})) \geq \mu_k \tag{3.25}$$

which is conic quadratic representable. The first proposed algorithm, referred to as the SCA-MISOCP based Algorithm, is outlined in Algorithm 3.2.

Convergence Analysis

We now prove that Algorithm 3.2 is guaranteed to converge. This can be established by showing that the sequence of objectives returned by Algorithm 3.2 is monotonically convergent. Towards this end, let $\theta^{(n)}$ and $\Theta^{(n)}$ denote the optimal objective value and the achieved optimal solution at the n^{th} iteration of Algorithm 3.2, respectively. Due to the first order approximation

Algorithm 3.2: SCA-MISOCP based Algorithm.

1: Initialize starting points of $\mathbf{w}^{(n)}, z^{(n)}, \gamma^{(n)}$;
2: Set $n := 0$;
3: **repeat**
4: Solve the approximated problem (3.23) with the SOC approximation (3.25) at $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \gamma^{(n)}$ to achieve the optimal solution $\mathbf{a}^\star, \mathbf{b}^\star, \gamma^\star, \mathbf{t}^\star, \mu^\star, \nu^\star, \mathbf{w}^\star, \mathbf{z}^\star$;
5: Set $n := n+1$;
6: Update $\mathbf{w}^{(n)} = \mathbf{w}^\star, \mathbf{z}^{(n)} = \mathbf{z}^\star, \gamma^{(n)} = \gamma^\star$;
7: **until** Convergence;

in (3.20), (3.21) and (3.22), it holds that equalities occur at (a) when $(\mathbf{w}_k^{(n)}, \gamma_k^{(n)}) = (\mathbf{w}_k, \gamma_k)$, at (b) when $z_k^{(n)} = z_k$, and (c) when $\mathbf{w}^{(n)} = \mathbf{w}$, respectively. Then, the updating rule in Algorithm 2 (cf. Step 5 in Algorithm 3.2) ensures that $\Theta^{(n)}$ is also feasible to problem (3.23) at the $(n+1)^{\text{th}}$ iteration. This subsequently leads to $\theta^{(n+1)} \geq \theta^{(n)}$, meaning that Algorithm 3.2 generates a non-decreasing sequence of objective function values. Due to the power budget constraint (3.6c), the sequence of objectives $\{\theta^{\{n\}}\}$ is upper bounded and thus, is convergent.

Generation of Initial Point

To start the iterative process in Algorithm 3.2, it is essential to find a feasible point in Step 1 of Algorithm 3.2. For this purpose, we can simply set $t_k = \Gamma_{\min}$, $\gamma_k = \Gamma_{\min}$, $\mu_k = \log(1+\Gamma_{\min})$ for all $k \in \mathscr{K}$, and then solve the following feasibility problem $(\text{Pini}) = \text{find}\{\mathbf{a}, \mathbf{z} | z_k \leq 1/\log(1+\Gamma_{\min}), (3.17), (3.6\text{g}), a_{i,k} \in \{0,1\}\}$. We remark that the problem (Pini) is a feasibility MI-SOCP program which can be solved optimally by off-the-shelf solvers such as MOSEK or GUROBI. Next, from the obtained value of $\mathbf{a}, \mathbf{t}, \mathbf{z}, \gamma, \mu$, we consider the following mixed integer program $(\text{Pmin}) = \min_{\mathbf{w},\mathbf{b},\nu}\{P^{\text{tot}}(\mathbf{w}, \mathbf{a}, \mathbf{b}) | (3.6\text{b}) - (3.6\text{f}), b_i \in \{0,1\}\}$. Note that the constraints (3.6c)-(3.6f) are SOC representable as discussed earlier. In fact, (3.6b) can also be reformulated by a SOC constraint as shown inTervo *et al.* (2015); Wiesel *et al.* (2006), which can be briefly explained as follows. It is easy to check that if $\mathbf{w}_k, \forall k \in \mathscr{K}$ is feasible to (3.6), then a phase rotation on $\mathbf{w}_k$ (i.e., replace $\mathbf{w}_k$ by $\mathbf{w}_k e^{j\phi_k}$ for some $\phi_k \in [0, 2\pi]$) creates another feasible solution of the same objective value. Therefore, without loss of optimality, $\mathbf{w}_k$ can be chosen such that $\mathbf{h}_k^H \mathbf{w}_k$ is real and non-negative $\forall k \in \mathscr{K}$. As a result, (3.6b) is equivalent to the following two

constraints

$$c\mathrm{Re}\left(\mathbf{h}_k^H\mathbf{w}_k\right) \geq \left\| \left[\mathbf{h}_k^H\mathbf{w}_1,\ldots,\mathbf{h}_k^H\mathbf{w}_K,\sigma_0^2\right]^T \right\|_2 \tag{3.26a}$$

$$\mathrm{Im}\left(\mathbf{h}_k^H\mathbf{w}_k\right) = 0 \tag{3.26b}$$

where $c = \sqrt{\left(\Gamma_{\min}+1\right)/\Gamma_{\min}}$. Now, it is clear that (Pmin) is a MI-SOCP problem, and thus can be solved optimally. The obtained values of $\mathbf{w},\mathbf{z},\boldsymbol{\gamma}$ by solving (Pini) and (Pmin) are then used to start Algorithm 3.2. Alternative option is to initialize Algorithm 3.2 from a feasible solution that can be found by the suboptimal algorithms presented in the subsequent subsections.

3.4.3 Continuous relaxation and inflation based algorithm

To develop a more practically appealing algorithm, we further consider the continuous relaxation of (3.23), i.e., $0 \leq b_i \leq 1, 0 \leq a_{i,k} \leq 1$ for $\forall i \in \mathscr{I}, \forall k \in \mathscr{K}$. As a result, the continuous relaxation of (3.23), denoted as $(\mathscr{P}^r)$, becomes an SOCP which can be solved in polynomial time by modern conic solvers. The second proposed iterative method combines two stages: (i) continuous relaxation and (ii) post-processing. In the first stage, we follow an iterative algorithm similar to Algorithm 3.2, but simply solve $(\mathscr{P}^r)$ in Step 3. The post-processing process is then used to map the obtained b_i's and $a_{i,k}$'s to the binary values, which is required due to the continuous relaxation. Towards this end, we apply the inflation procedure in Cheng *et al.* (2013) to refine the achieved solution. In particular, we rely on the solution to the continuous relaxation at convergence as an incentive measure to make a decision on the binary value of $\mathbf{a}$ and $\mathbf{b}$. Let us denote $\tilde{\mathbf{a}},\tilde{\mathbf{b}}$ and $\tilde{\mathbf{w}}$ as the solution achieved after the first stage. Intuitively, the connection between the ith RRH and the kth UE is more likely if the channel of the link is in better condition and the power consumed to transmit fronthaul data $P_{i,k}^{\mathrm{FH}}$ is smaller than the others. Consequently, solving the continuous relaxation would possibly yield higher $\widetilde{b}_i$ for the ith RRH and higher $\widetilde{a}_{i,k}$ for the connection between the ith RRH and the kth UE. Based on the above intuitive observations, we propose an iterative procedure to determine the set of active RRHs and RRH-UE association based on $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$. The process starts by assuming that all the RRHs are off and there is no association between RRH and UE. In each iteration, $(\mathscr{P}^r)$

is solved with a set of remaining inactive RRHs and RRH-UE association that is not connected. The RRH-UE association with the largest $\tilde{a}_{i,k}$ will be made connected and the resulting RRH will be set active, following the relationship in (3.6f). The overall algorithm is presented in Algorithm 3.3.

Algorithm 3.3: Inflation based algorithm

1: Set $m := 0$, $\pi^{(m)}$ is significantly small, and initialize the set $\mathscr{R}_{\text{off}}^{(m)} = \{(i,k) \times i \in (\mathscr{I}, \mathscr{K}) \times \mathscr{I}\}$.
2: **repeat**
3: Set $m := m + 1$;
4: Solve $(\mathscr{P}^r)$ with $a_{i',k'} = 1$ and $b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathscr{R}_{\text{off}}^{(m-1)}$;
5: Update $\mathscr{R}_{\text{off}}^{(m)} = \mathscr{R}_{\text{off}}^{(m-1)} \setminus \left\{ (i',k') \times i' = \arg\max_{i,k \in \mathscr{R}_{\text{off}}^{(m-1)}} \widetilde{a}_{i,k} \right\}$;
6: Solve (3.23) with (3.25) given $a_{i',k'} = 1, b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathscr{R}_{\text{off}}^{(m)}$ and $a_{i,k} = 0, b_i = 0, \forall \{(i,k) \times i\} \in \mathscr{R}_{\text{off}}^{(m)}$, denoted as $\left(\mathscr{P}^{int}\right)$. If $\left(\mathscr{P}^{int}\right)$ is feasible, set $\pi^{(m)}$as the value of objective function achieved at the convergence. If not, set $\pi^{(m)} = \pi^{(0)}$.
7: **until** $(\mathscr{P}^r)$ starts to be infeasible or $\left(\mathscr{P}^{int}\right)$ is feasible and $\pi^{(m)} < \pi^{(m-1)}$;
8: Solve (3.23) with (3.25) given $a_{i',k'} = 1, b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathscr{R}_{\text{off}}^{(m-1)}$ and $a_{i,k} = 0, b_i = 0, \forall \{(i,k) \times i\} \in \mathscr{R}_{\text{off}}^{(m-1)}$ to obtain $\mathbf{w}^\star, \nu^\star, \mathbf{t}^\star, \mathbf{z}^\star, \mu^\star, \gamma^\star$;

Convergence Analysis

Algorithm 3.3 is provably convergent due to two facts. First, the SCA-based algorithm to solve $(\mathscr{P}^r)$ is guaranteed to converge and this can be proved in the same way as done for Algorithm 3.2. Second, the post-processing procedure is executed $(I-1)K$ times in the worst case. In the last step when all the binary variables have been fixed, the SCA-based algorithm is applied to solve (3.23) until convergence. Note that in this case, we deal with a continuous optimization problem and a stronger convergence result can be achieved. Specifically, every limit point of the SCA-based algorithm is a stationary solution to the continuous optimization problem. However, we remark that a stationary point is not necessarily a locally optimal solution. Exploring further properties of the obtained stationary solution is beyond the scope of the paper.

Generation of Initial Point

To apply the SCA-based algorithm to solve $(\mathscr{P}^r)$ in the first iteration of Algorithm 3.3 (i.e, when $0 \leq a_{i,k} \leq 1$ and $0 \leq b_i \leq 1$), we may need a feasible point. However, as mentioned earlier, the challenge is that $(\mathscr{P}^r)$ (cf. (3.23)) is nonconvex with the remaining other continuous variables, making it difficult to find a feasible point. There is in fact a penalty method to allow the SCA-based procedure to start from an infeasible point, which is described in Lipp & Boyd (2016). The idea is to introduce slack variables into each constraint as the violations and penalizing the sum of these violations in the objective. In this way, first iterations of Algorithm 3.3 may be infeasible to (3.23), but violations are forced to be zero as the iterative process progresses. We refer the interested reader to (Lipp & Boyd, 2016, Algorithm 3.1) for a complete description of this initialization method.

3.4.4 Sparsity-inducing Norm Approach

In the final low-complexity approach, we reformulate the sum rate-power maximization from a viewpoint of group sparsity. Note that the i^{th} RRH will not be selected if the vector $\tilde{\mathbf{w}}_i = \left[\mathbf{w}_{i,1}^H, \ldots, \mathbf{w}_{i,K}^H\right]$ which includes all beamformers related to the i^{th} RRH is a zero vector. Let us rewrite the total power consumption as

$$P_{\text{sparse}}^{\text{tot}}(\mathbf{w}) = \frac{1}{\eta_i} \sum_{\forall i \in \mathscr{I}} \left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \sum_{\forall i \in \mathscr{I}} \chi\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2\right) \left(P_i^{\text{ra}} - P_i^{\text{ri}}\right) + \sum_{\forall i \in \mathscr{I}} P^{\text{ri}} + \sum_{\forall k \in \mathscr{K}} \sum_{\forall i \in \mathscr{I}} \chi\left(\left\|\mathbf{w}_{i,k}\right\|_2^2\right) P_{i,k}^{\text{FH}}. \quad (3.27)$$

The sum rate–power optimization can now be written as

$$\max_{\mathbf{w}} \quad \tilde{\alpha} \sum_{k=1}^{K} R_k(\mathbf{w}) - \bar{\alpha} P_{\text{sparse}}^{\text{tot}}(\mathbf{w}) \tag{3.28a}$$

$$\text{s.t. } \Gamma_k(\mathbf{w}) \geq \Gamma_k^{\min}, \forall k \in \mathscr{K} \tag{3.28b}$$

$$\sum_{\forall k \in \mathscr{K}} \chi\left(\left\|\mathbf{w}_{i,k}\right\|_2^2\right) R_k(\mathbf{w}) \leq \frac{C_i}{\xi_i}, \forall i \in \mathscr{I}, \tag{3.28c}$$

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 \leq P^{\max}, \forall i \in \mathscr{I}. \tag{3.28d}$$

In fact we can impose sparsity on the soft power vector ν to derive the sparsity-inducing norm method. However our idea is to impose sparsity directly on the beamforming vector $\mathbf{w}$ to arrive at (3.28). Thus, all slack variables are not introduced to reduce the complexity of the resulting formulation. However, problem (3.28) is still non-convex due to the presence of the indication functions, which are intractable. To deal with this problem, we will replace $\chi(x)$ by $\log(\tau + x)$ for a small $\tau > 0$, following the result in Candes *et al.* (2008). In this way we can approximate

$$\chi\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2\right) \approxeq \log\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1\right) \tag{3.29}$$

$$\chi\left(\left\|\mathbf{w}_{i,k}\right\|_2^2\right) \approxeq \log\left(\left\|\mathbf{w}_{i,k}\right\|_2^2 + \tau_2\right) \tag{3.30}$$

where $\tau_1, \tau_2 > 0$ are small positive parameters. Consequently, we can obtain a continuous approximation of (3.28) as

$$\max_{\mathbf{w}} \quad \tilde{\alpha} \sum_{k=1}^{K} R_k(\mathbf{w}) - \bar{\alpha} \tilde{P}_{\text{sparse}}^{\text{tot}}(\mathbf{w}, \mathbf{p}, \mathbf{q}) \tag{3.31a}$$

$$\text{s.t.} \quad \sum_{\forall k \in \mathscr{K}} q_{i,k}^2 R_k(\mathbf{w}) \leq \frac{C_i}{\xi_i}, \forall i \in \mathscr{I}, \tag{3.31b}$$

$$\log\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1\right) \leq p_i, \forall i \in \mathscr{I}, \tag{3.31c}$$

$$\log\left(\left\|\mathbf{w}_{i,k}\right\|_2^2 + \tau_2\right) \leq q_{i,k}^2 \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{3.31d}$$

$$(3.28\text{b}), (3.28\text{d}) \tag{3.31e}$$

where we have introduced $\mathbf{q} = \left\{q_{i,k} \geq 0, \forall k \in \mathscr{K}, \forall i \in \mathscr{I}\right\}$ and $\mathbf{p} = \{p_i \geq 0, \forall i \in \mathscr{I}\}$ and defined

$$\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w},\mathbf{p},\mathbf{q}) = \frac{1}{\eta_i}\sum_{\forall i\in\mathscr{I}}\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \sum_{\forall i\in\mathscr{I}} p_i\left(P_i^{\text{ra}} - P_i^{\text{ri}}\right) + \sum_{\forall i\in\mathscr{I}} P_i^{\text{ri}} + \sum_{\forall k\in\mathscr{K}}\sum_{\forall i\in\mathscr{I}} q_{i,k}^2 P_{i,k}^{\text{FH}}. \quad (3.32)$$

Note that $\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w},\mathbf{p},\mathbf{q})$ is convex with the involving variables, and that the purpose of using the second order on $q_{i,k}$ is to reuse the approximations presented previously, as we will show shortly. The constraint in (3.31c) can be equivalently rewritten as

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1 \leq e^{p_i}, \forall i \in \mathscr{I} \quad (3.33)$$

and thus can be approximated by

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1 \leq e^{p_i^{(n)}} + e^{p_i^{(n)}}(p_i - p_i^{(n)}) \triangleq \tilde{F}(p_i; p_i^{(n)}). \quad (3.34)$$

In the same way (3.31d) can be approximated as

$$\left\|\mathbf{w}_{i,k}\right\|_2^2 + \tau_2 \leq e^{q_{i,k}^{(n)2}} + 2q_{i,k}^{(n)} e^{q_{i,k}^{(n)2}}(q_{i,k} - q_{i,k}^{(n)}) \triangleq \bar{F}(q_{i,k}; q_{i,k}^{(n)}). \quad (3.35)$$

Here we write $e^{q_{i,k}^{(n)2}}$ instead of $e^{\left(q_{i,k}^{(n)}\right)^2}$ to lighten the notation. Unlike to approach that fixes the rate function $R_k(\mathbf{w})$ in each iteration in Dai & Yu (2014); Ha *et al.* (2016), here, we deal with the nonconvexity in (3.31b) by equivalently rewriting it as

$$\begin{cases} \displaystyle\sum_{k\in\mathscr{K}} \frac{q_{i,k}^2}{z_k} \leq \frac{C_i}{\xi_i}, \\ (3.16\text{b}), (3.18). \end{cases} \quad (3.36)$$

where $\mathbf{t}$ and $\mathbf{z}$ are introduced as done similarly in (3.16b)–(3.18). Now the approximations used to deal with (3.16b)–(3.18) can be applied, which results in the following convex approximated

Algorithm 3.4: Sparsity-inducing norm-based algorithm.

1: Set $n := 0$ and initialize starting points of $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \mathbf{p}^{(n)}, \mathbf{q}^{(n)}$;
2: **repeat**
3: Solve the approximated problem (3.37) at $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \mathbf{p}^{(n)}, \mathbf{q}^{(n)}$ to achieve the optimal solution $\mathbf{w}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\gamma}^\star, \mathbf{t}^\star, \mathbf{z}^\star, \mathbf{p}^\star, \mathbf{q}^\star$;
4: Update $\mathbf{w}^{(n+1)} = \mathbf{w}^\star, \mathbf{z}^{(n+1)} = \mathbf{z}^\star, \boldsymbol{\gamma}^{(n+1)} = \boldsymbol{\gamma}^\star, \mathbf{p}^{(n+1)} = \mathbf{p}^\star, \mathbf{q}^{(n+1)} = \mathbf{q}^\star$;
5: Set $n := n+1$;
6: **until** Convergence;

problem (3.28) at the $(n+1)^{\text{th}}$ iteration of the sparsity-based iterative algorithm

$$\max_{\substack{\mathbf{w},\boldsymbol{\mu},\boldsymbol{\gamma},\mathbf{t}\\ \mathbf{z},\mathbf{p},\mathbf{q}}} \tilde{\alpha} \sum_{k\in\mathscr{K}} \mu_k - \bar{\alpha}\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w};\mathbf{p};\mathbf{q}) \tag{3.37a}$$

$$\text{s.t.} \left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1 \le \tilde{F}\left(p_i; p_i^{(n)}\right) \tag{3.37b}$$

$$U\left(\gamma_k; \gamma_k^{(n)}\right) \ge \mu_k \tag{3.37c}$$

$$\gamma_k \ge \Gamma_k^{\min} \tag{3.37d}$$

$$\sum_{j\in\mathscr{K}\setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2 \le H\left(\mathbf{w}_k, \gamma_k; \mathbf{w}_k^{(n)}, \gamma_k^{(n)}\right) \tag{3.37e}$$

$$1 + t_k \le F\left(z_k; z_k^{(n)}\right) \tag{3.37f}$$

$$\left\|\mathbf{w}_{i,k}\right\|_2^2 + \tau_2 \le \bar{F}\left(q_{i,k}; q_{i,k}^{(n)}\right) \tag{3.37g}$$

$$\left|\mathbf{h}_k^H \mathbf{w}_k\right|^2 / t_k \le G\left(\mathbf{w}; \mathbf{w}^{(n)}\right) \tag{3.37h}$$

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 \le P^{\max}, \forall i \in \mathscr{I}, \tag{3.37i}$$

$$\sum_{k\in\mathscr{K}} \frac{q_{i,k}^2}{z_k} \le \frac{C_i}{\xi_i}, \forall i \in \mathscr{I} \tag{3.37j}$$

where $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \mathbf{p}^{(n)}, \mathbf{q}^{(n)}$ are the parameters that are updated at the $(n+1)^{\text{th}}$ iteration. The proposed iterative approach to solve problem (3.28) is given in Algorithm 3.4. Note that the convergence of Algorithm 3.4 can be established following the same arguments as those in Algorithms 3.2 and 3.3 above. Also, the generation of an initial point for Algorithm 3.4 can be carried out the same way as for Algorithm 3.3.

3.4.5 Convergence and Complexity Analysis

We now discuss the complexity of each proposed algorithm in this section. For the optimal design based on a BRB method, i.e., Algorithm 3.1, the complexity is extremely high since the number of the boxes needs to be considered increases exponentially with the problem dimension. Moreover, in each iteration, an MI-SOCP feasibility problem is solved. For Algorithm 3.2, the overall complexity mainly depends on that of solving the MI-SOCP problem in (3.23) which is indeed a combinatorial optimization problem. In particular, there are IK binary variables $a_{i.k}$'s and I binary variables b_i's, resulting in 2^{IK+I} combinations for all the binary variables. Given fixed $\mathbf{a}$ and $\mathbf{b}$, all the constraints in problem (3.23) approximately consist of a total number of $KM+2IK+4K+1$ variables and a number of $3IK+2I+4K+1$ SOC constraints of dimension $KM+1$. Thus, the worst-case complexity of Algorithm 3.2 in each iteration can be written as $\mathscr{O}\left(2^{IK+K}(K^4M^3I)\right)$. Compared to Algorithm 3.1, Algorithm 3.2 has less complexity due to the continuous approximation converging rapidly.

Next, we analyze the complexity of Algorithms 3.3 and 3.4. First we remark that in the worst case, Algorithm 3.3 must iteratively solve and update the resulting parameters for the SOCP problem $(\mathscr{P}^r)$ and $\left(\mathscr{P}^{int}\right)$ for $(I-1)K$ times. In each step, the complexity of solving $(\mathscr{P}^r)$ and $\left(\mathscr{P}^{int}\right)$ is approximately $\mathscr{O}\left(K^4M^3I\right)$, resulting the overall complexity of $\mathscr{O}\left(2(I-1)K\left(K^4M^3I\right)\right)$ for Algorithm 3.3. In Section 3.5, the numerical results show that Algorithm 3.3 yields a performance very close to that of Algorithm 3.2 but with much lower computation time. Finally, for Algorithm 3.4, the worst-case complexity is given by $\mathscr{O}\left(K^4M^3I\right)$.

3.5 Numerical Results

In this section, we numerically evaluate the performance of the proposed algorithms. For most numerical experiments, we use the simulation parameters listed in Table 3.1. In particular, the parameters in the RRH and fronthaul power consumption model are taken from Cheng *et al.* (2013). For the spatial model, we assume a network consisting of I RRHs that are uniformly located around the considered coverage and K UEs are randomly scattered across the considered

Table 3.1 Simulation parameters in Chapter 3

Description	Notation	Value
Number of RRHs	I	6
Number of users	K	4
Number of antennas per RRH	M_i	2
Power amplifier efficiency	η_i	0.35
Maximum transmit power	$P^{\max}$	10 dBW
Active power for RRH and fronthaul	P_i^{ra}	12.5 dBW
Sleep power for RRH and fronthaul	P_i^{ri}	2.5 dBW
Reference rate	R_0	1 b/s/Hz
Reference power	P_0	0 dBW
Noise power	σ_0^2	-143 dBW
Fronthaul power	P_i^{FH}	0 dBW
Maximum fronthaul capacity	$C_i = C, \forall i$	500 b/s/Hz
Fronthaul capacity factor	$\xi_i, \forall i$	10
Reweighted parameter	τ_1, τ_2	10^{-3}

network coverage. Moreover, we assume Rayleigh fading channel and the pathloss component is calculated as $(d_{ik}/d_0)^{-3}$where d_{ik} is the distance between the i^{th} RRH and the k^{th} user and $d_0 = 100$ m is the reference distance. To simplify the notations, we also assume the fronthaul link capacity $C_i = C, \forall i \in \mathscr{I}$, can be achieved up to 10 Gbps over 20 MHz bandwidth, which is equivalent to 500 b/s/Hz. In our simulations, Algorithms 3.2, 3.3, and 3.4 are terminated when the increase in the objective between two consecutive iterations is less than 10^{-5}.

In Fig. 3.2, we show the convergence of the objective function in (3.6) of Algorithms 3.1 and 3.2 for a set of random channel realizations. In this numerical experiment, we set $\alpha = 0.7$ and consider a small network setting with $I = 4$, $K = 3$. For Algorithm 3.2, we show the convergence behavior of the objective function with two different initial points $\mathbf{w}^{(0)}, \mathbf{z}^{(0)}, \boldsymbol{\gamma}^{(0)}$. As expected, Algorithm 3.1 requires much more iterations to update the upper and lower bounds, and thus converges after many iterations. On the other hand, Algorithm 3.2 converges much faster, just after a few iterations, and achieves the same objective value as return by the optimal algorithm despite the choice of initial points. This clearly demonstrates the effectiveness of Algorithm 3.2 which is used for benchmarking in the next experiments.

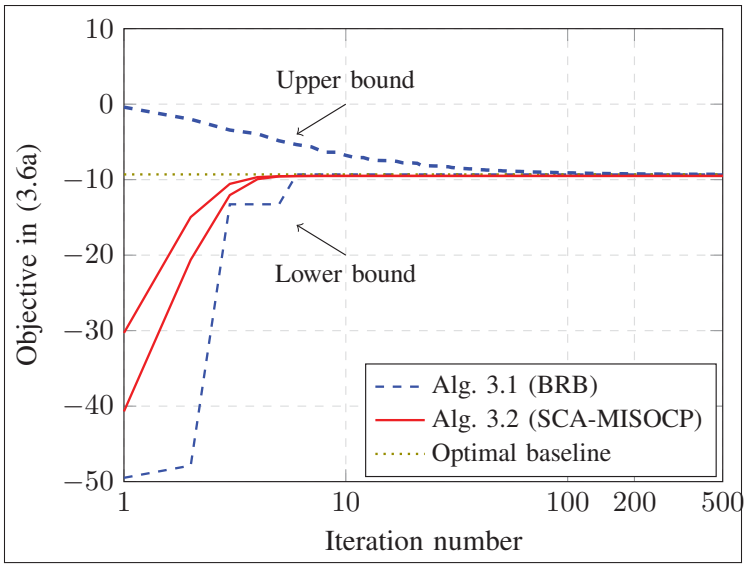


Figure 3.2 The convergence of our proposed algorithms for a set of random channel realizations.

For onwards, we will consider the network setting as mentioned in Table 3.1. In Figs. 3.3 and 3.4, we compare the convergence performance of our proposed low-complexity algorithms with the iterative WMMSE-reweighted ℓ_1-norm algorithm introduced in Dai & Yu (2014) and coalitional game based algorithm in Sun *et al.* (2016) for $\alpha = 0.9$, in terms of both number of iterations and the overall runtime. Although only the sum rate maximization problem was studied in Dai & Yu (2014) and Sun *et al.* (2016), we can easily modify their algorithm to solve the sum rate-power maximization problem considered in this paper. As can be clearly seen, our proposed solutions need a much smaller number of iterations to converge (possibly to different objectives), compared the reweighted ℓ_1-norm algorithm and coalitional game based algorithm. We note that in Fig. 3.3, the convergence of each SOCP during the inflation process is plotted, which explains the uphill and downhill effect in the figure. As can be seen, Algorithm 3.2 just takes a few iterations to stabilize. However its overall runtime is very high since the problem in each iteration is an MI-SOCP. On the other hand, Algorithms 3.3 and 3.4 require more iterations to converge but the per-iteration problem is an SOCP, which can be solved with much computational effort. Thus their eventual computation time is much lower than that of Algorithm 3.2. Due to the fast converging property, Algorithms 3.3 and 3.4 outperform

the reweighted ℓ_1-norm method, which is shown in Fig. 3.4. In Fig. 3.3, we can also see that the reweighted ℓ_1-norm and coalitional game based optimization methods converge to a smaller value, compared to our proposed solutions. This will be elaborated in the following experiments.

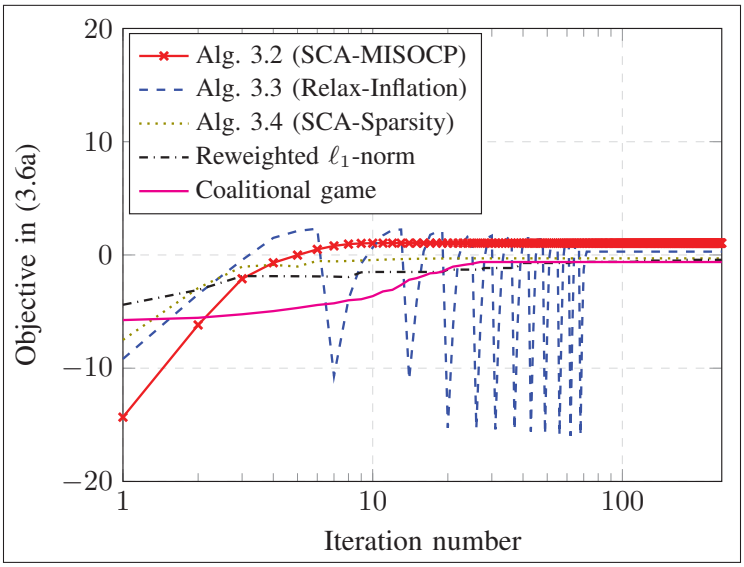


Figure 3.3 The convergence comparison between different low complexity algorithms.

In Figs. 3.5 and 3.6, we study the trade-off between achievable sum rate (ASR) and total power consumption (TPC) by varying the parameter α over the interval $[0,1]$ for the algorithms of comparison. The end points on bottom left of Figs. 3.5 and 3.6 show the smallest possible value of TPC without any consideration of the ASR (i.e., $\alpha = 0$). On the contrary, the end points on top right represent the largest possible ASR that can be obtained without any consideration of the TPC (i.e., $\alpha = 1$). As expected and shown in Figs. 3.5 and 3.6, when the TPC increases, so does the ASR. Moreover, it can be clearly seen that our proposed algorithms outperform the reweighted ℓ_1-norm and coalitional game based algorithm. Algorithm 3.2 is shown to attain the best performance among all the algorithms. Noticeably, the differences in the TPC between our proposed algorithms and the reweighted ℓ_1-norm algorithm as well as coalitional game based algorithm are significant. The reason is that, the method in Dai & Yu (2014) does not take into

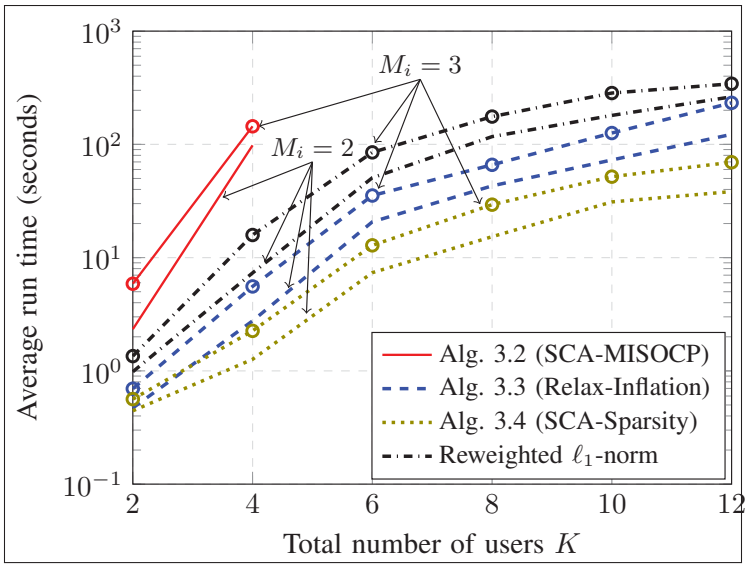


Figure 3.4 The average run time comparision between different algorithms with number of antenna per RRH $M_i = 2, 3$.

the fronthaul power while it becomes significant for large P^{FH} and the method in Sun *et al.* (2016) does not consider the RRH selection in their coalition formation algorithm.

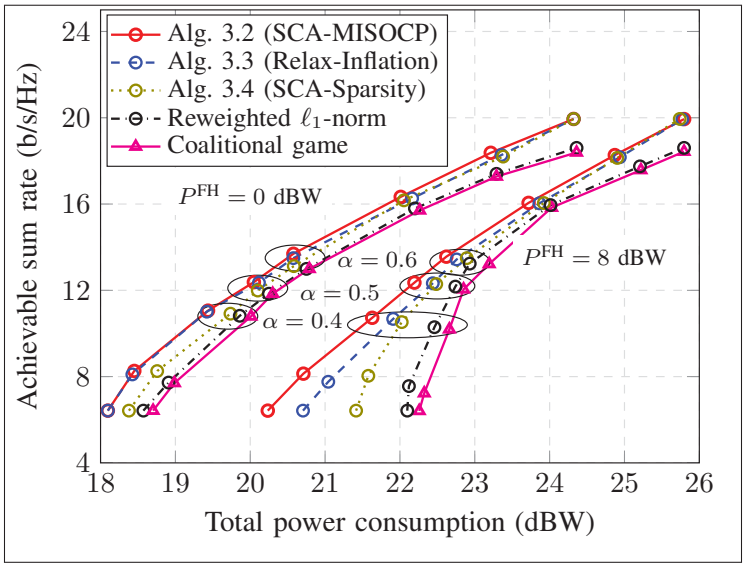


Figure 3.5 Trade-off between achievable sum rate and sum power consumption.

In Fig. 3.6, we investigate the trade-off between ASR and TPC for different values of fronthaul capacity factor $\xi_i = 50, 100, \forall i \in \mathscr{I}$. We first observe that there exists a "strong" trade-off between ASR and TPC in the high power regime when the fronthaul capacity is small. That is to say, a large amount of TPC is consumed just for a negligible improvement in the ASR. For high transmit power, the ASR over the wireless medium may be high but the small fronthaul capacity will act as a bottleneck. The practical guidance here is to avoid transmitting at full power for small fronthaul capacity to improve the network energy efficiency. Fig. 3.6 also demonstrates the increase of the ASR with the fronthaul capacity. To achieve the same ASR, more TPC is required for the networks with smaller fronthaul capacities. Intuitively, for smaller fronthaul capacities, the number of cooperative RRHs is reduced (cf. Table 3.2 for further insights). In fact, when each fronthaul capacity limit is small, common data shared by cooperative RRHs are limited to be transported via the fronthaul link to the RRHs, which in turn allows lesser number of cooperative RRHs. This results in more concurrent transmissions from the non-cooperative RRHs, which increases interference at each UE and subsequently leads to an increase in TPC to achieve the desired ASR. Furthermore, the results in Fig. 3.6 again show that the proposed algorithms achieve an improvement in the sum rate by up to 3.2 bits/s/Hz for the same TPC in the case $\xi_i = 100$, compared to the reweighted ℓ_1-norm and coalitional game based algorithms. In addition, the performance of coalitional game method is worse than other methods since RRHs are formed disjoint coalitions, thereby increasing the cell-edge interference and decreasing the ASR.

To gain more insights into the considered problem, we list the average number of active RRHs and number of RRH-UE associations versus the fronthaul power and fronthaul capacity in Table 3.2. In this table, when $P^{\mathrm{FH}} = 0, 8$ dBW, we choose $C = 500$ b/s/Hz and when $C = 60, 100$ b/s/Hz, we choose $P^{\mathrm{FH}} = 0$ dBW. We can see that when the fronthaul power consumption increases, fewer RRH-UE associations are active and more RRHs are turned on. We observe that our proposed algorithms switch on only 50% of RRHs and 29.17% of user-RRHs associations to further reduce the total transmit power, while the referred algorithm switches on 66.67% of RRHs as well as RRH-UE associations.

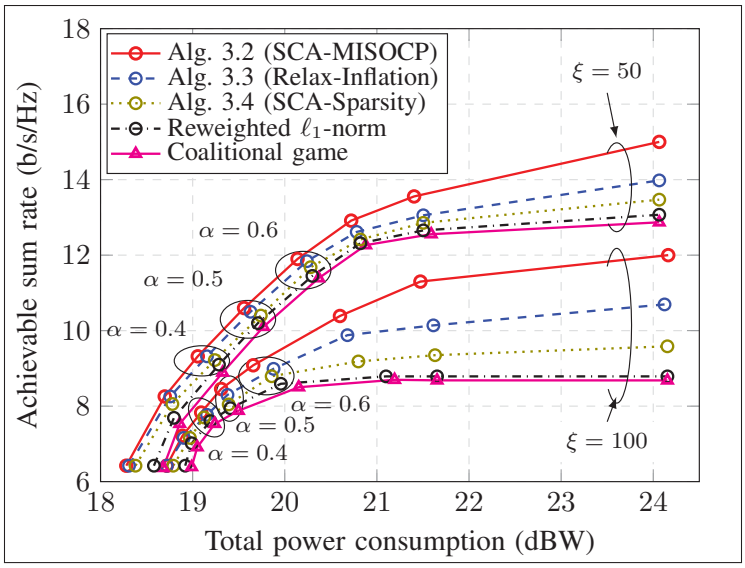


Figure 3.6 Trade-off between sum achiveable rate and total power consumption.

Table 3.2 Average number of active RRH-UE associations (Avr.RRH-UE) and active RRHs (Avr.RRHs).

P^{FH}/C		Alg. 3.2	Alg. 3.3	Alg. 3.4	Dai & Yu (2014)
P^{FH}=0 dB	Avr.RRH-UE	0.4167	0.4167	0.5	0.5
	Avr.RRHs	0.5	0.5	0.5	0.5
P^{FH}=8 dB	Avr.RRH-UE	0.2917	0.4167	0.5	0.6667
	Avr.RRHs	0.5	0.5	0.6667	0.6667
C=6 b/s/Hz	Avr.RRH-UE	0.375	0.4583	0.5	0.5
	Avr.RRHs	0.5	0.6667	0.6667	0.6667
C=10 b/s/Hz	Avr.RRH-UE	0.4583	0.4583	0.5	0.5
	Avr.RRHs	0.5	0.5	0.5	0.5

In Fig. 3.7, we compare the achievable sum rate at $\alpha = 1$ with respect to the fronthaul capacity C beween different algorithms. Note that problem (3.6) with $\alpha = 1$ is equivalent to the problem of maximizing the achievable sum rate. From the figure, when C increases, the achievable sum rates obtained by all the algorithms increase accordingly. This can be explained as when C increases, more data information transported via the fronthaul link, which enables more cooperation between RRHs. This cooperation has the impact of reducing inter-RRH interference and thus improves the overall achievable rate of the system. However, at high regime of fronthaul capacity, these achievable sum rates saturate at a fixed value. This

is because there always exists interference between multi-user transmissions even as more co-operation is enabled between RRHs. This interference creates a upper limit to the achievable rate for all users so that increasing more fronthaul capacity do not provide more benefit to the system performance. In addition, we observe that at high value of C, our proposed algorithms achieve a better achievable sum rate than the WMMSE-reweighted l_1-norm algorithm, which again proves the superiority of our algorithms compared to traditional work Dai & Yu (2014).

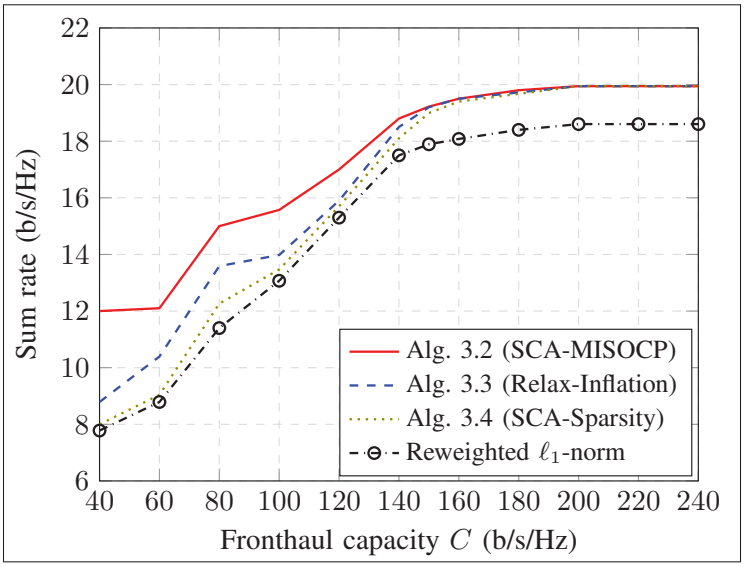


Figure 3.7 Sum achievable rate of different algorithms for sum rate maximization problem ($\alpha = 1$).

In Fig. 3.8 and Fig. 3.9, we investigate the power consumption minimization problem by attaching the weight $\alpha = 0$ to the two–dimensional achievable sum rate-power maximization problem. Here the achievable sum rate term is completely not considered. In Fig. 3.8 and Fig. 3.9, we compare the total power consumption of our proposed algorithms to that of the linear-relaxed based algorithm in (Ha *et al.*, 2016). In Fig. 3.8, it shows the total transmission power versus the fronthaul transmission power P^{FH} is achieved by applying different algorithms at two different $\Gamma_k^{\min} = 2, 6$ dB. It is observed that when P^{FH} increases, more power is required to transport the data via the fronthaul link, thus resulting in an increment of P^{tot}. At higher $\Gamma_k^{\min}$,

the system consumes more power to achieve the required target SINR. It can also be seen that our algorithms outperform the linear-relax algorithm in (Ha *et al.*, 2016).

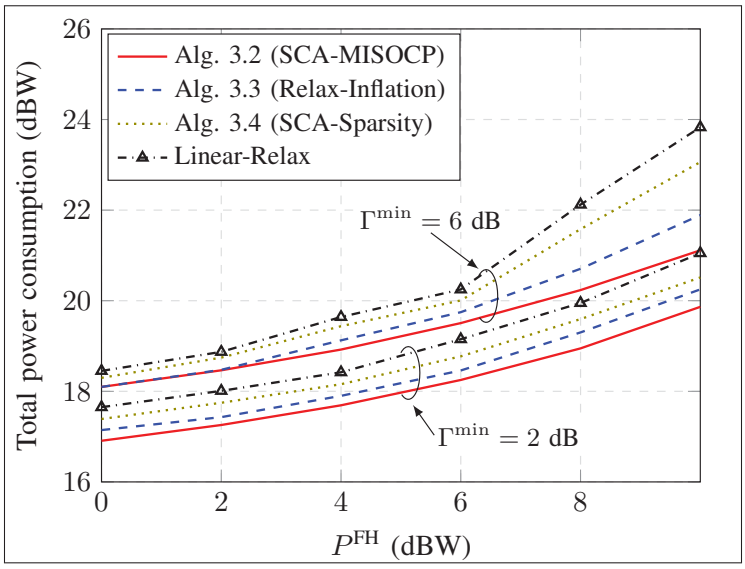


Figure 3.8 Total power consumption of different algorithms versus the fronthaul power consumption for power minimization problem ($\alpha = 0$).

In Fig. 3.9, we show the total transmission power versus the fronthaul link capacity $C_i = C, \forall i \in \mathscr{I}$ with $\Gamma_k^{\min}$ =2, 6 dB and $P^{\mathrm{FH}} = 0$ dB. As shown in the figure, total power consumption decreases with the increment of fronthaul maximum capacity C and it increases with the increment of target SINR. This can be explained by the fact that when the fronthaul capacity becomes higher, the number of UEs which are served by each RRH is larger, resulting in less power consumption in each RRH.

In Figs. 3.10 and 3.11, we compare the performance of the optimal solution attained by Algorithm 3.1 and the suboptimal solution by Algorithm 3.2 versus the required minimum SINR $\Gamma_k^{\min} = \Gamma^{\min}$, $\forall k \in \mathscr{K}$ with parameter $\alpha = 0.2$, 0.5, 0.7 and 0.9. Here, we consider a small network setting with $I = 4$, $K = 3$. As can be seen, when $\Gamma^{\min}$ increases, the ASR and TPC increase for all values of α. Particularly, the ASR and the TPC rapidly increase in the regime of low $\Gamma^{\min}$ and slightly increase in the high regime of $\Gamma^{\min}$. The increase of the ASR when

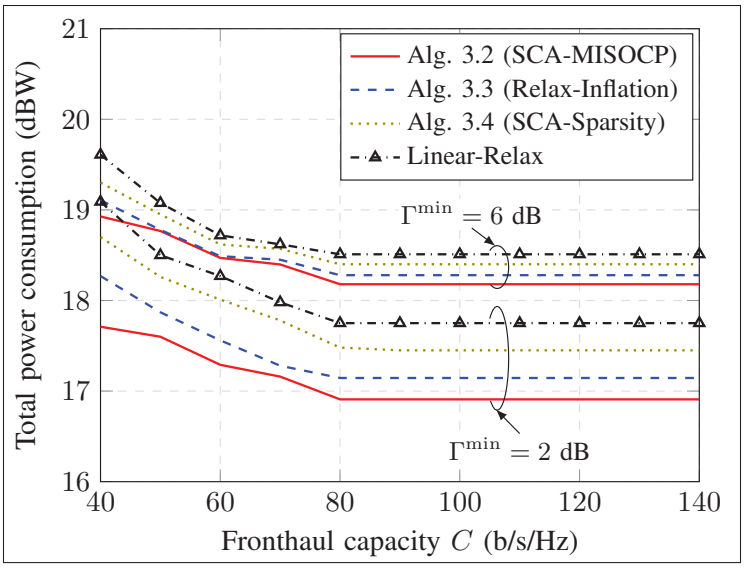


Figure 3.9 Total power consumption of different algorithms versus maximal fronthaul capacity for power consumption minimization problem ($\alpha = 0$).

$\Gamma^{\min}$ grows in Fig. 3.10 can be explained as follows. At low value of α, problem (3.6) has more priority to minimize TPC under the minimum rate constraint. Each user's rate achieved by solving the optimization (3.6) in this case is almost equal to the minimum rate, so that when $\Gamma^{\min}$ grows, the ASR increases proportionally. However, at high value of α, the problem of sum rate maximization is dominant. As a result, increasing $\Gamma^{\min}$ has less impact on the ASR performance. Similar explanation can be applied for the increase of the TPC at different α when $\Gamma^{\min}$ increases in Fig. 3.11. Moreover, in Fig. 3.12, the ASR is shown with three different values of the noise power $\sigma_0^2 = -143$, -140, and -130 dBW for $\alpha = 0.8$. It is obvious that when the noise power increases, the ASR decreases since the SINR of all users is eventually reduced. Regarding the optimality of the proposed suboptimal solutions, it is shown numerically in Fig. 3.10, 3.11 and 3.12 that the suboptimal solution achieved by Algorithm 3.2 is very close to the optimal solution obtained by Algorithm 3.1. This again demonstrates the effectiveness of Algorithm 3.2.

In Fig. 3.13, we compare the objective of different algorithms for different values of α. We note that the variation of the objective in (3.6a) depends not only on α but also on the values of R_0 and P_0. For the chosen R_0 and P_0 stated in Table 3.1, we observe that when α increases, the

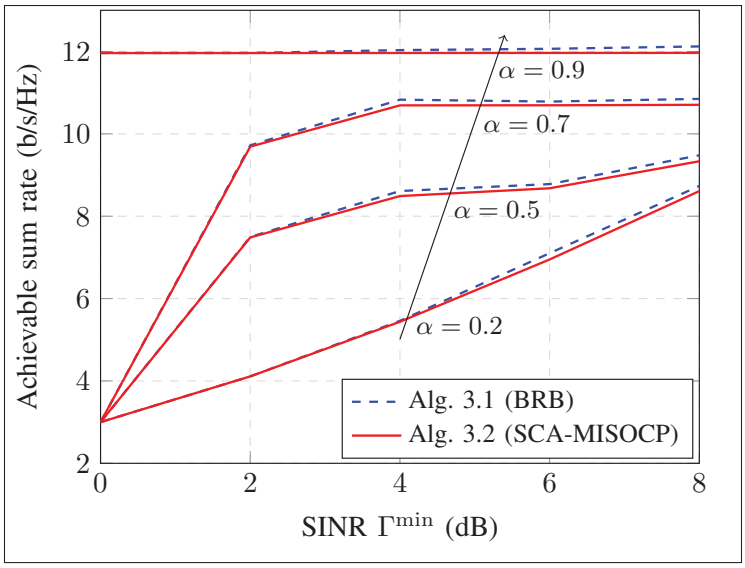


Figure 3.10 ASR versus required SINR $\Gamma^{\min}$ with parameter $\alpha = 0.2, 0.5, 0.7$ and 0.9.

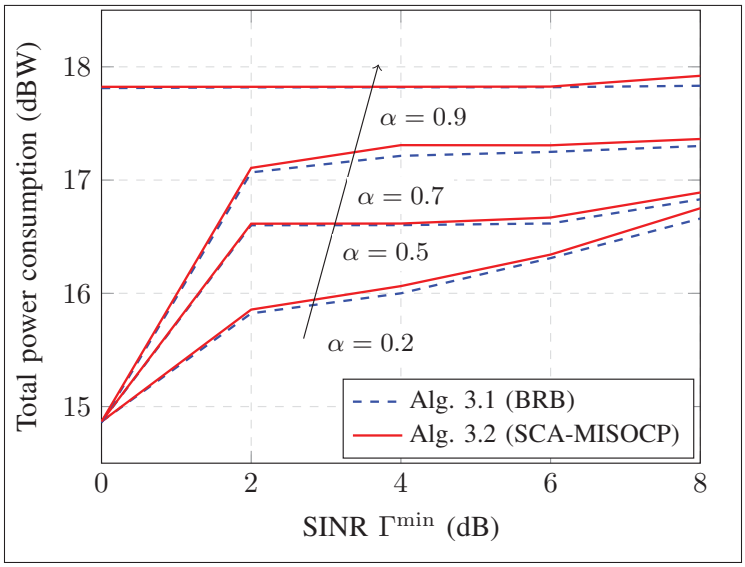


Figure 3.11 TPC versus required SINR $\Gamma^{\min}$ with parameter $\alpha = 0.2, 0.5, 0.7$ and 0.9.

objective first decreases and then increases. From the results shown in Figs. 3.5 and 3.6, it is clear that when α increases, the TPC increases, and this make the objective decrease. However, after a certain point, the term $\alpha R^{\text{tot}}(\mathbf{w})/R_0$ will become dominant $(1-\alpha)P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b})/P_0$ since the weight associated with power consumption is small, resulting in the objective increasing.

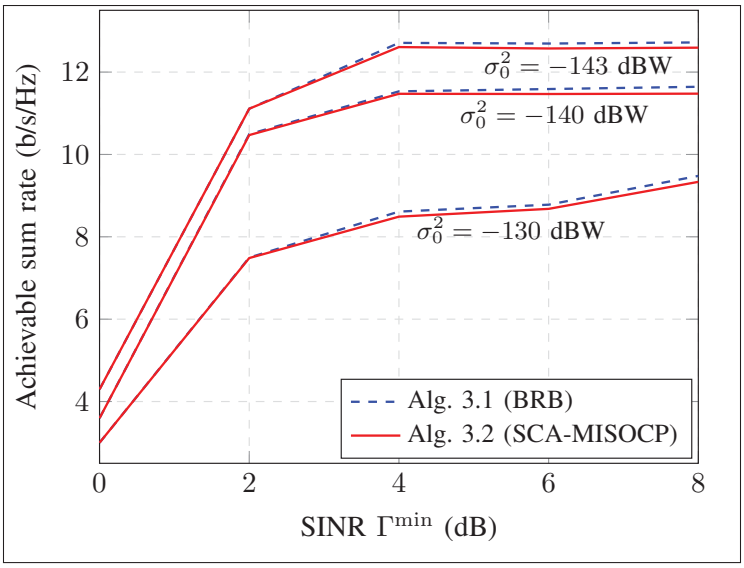


Figure 3.12 ASR versus required SINR $\Gamma^{\min}$ with some different values of parameter $\sigma_0^2 = -143, -140, -130$ dBW.

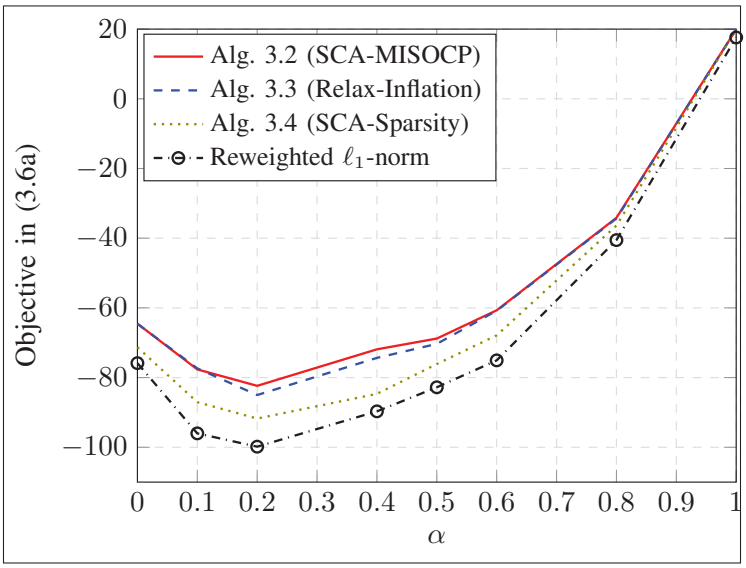


Figure 3.13 Objective in (3.6a) versus parameter α.

Fig. 3.13 again demonstrates that our proposed algorithms outperform the reweighted ℓ_1-norm algorithm.

In the final numerical experiment, we consider a relatively large network setting with the number of RRHs $I = 60$ for the number of UEs $K = 50$ and $K = 60$. In Fig. 3.14, the trade-off

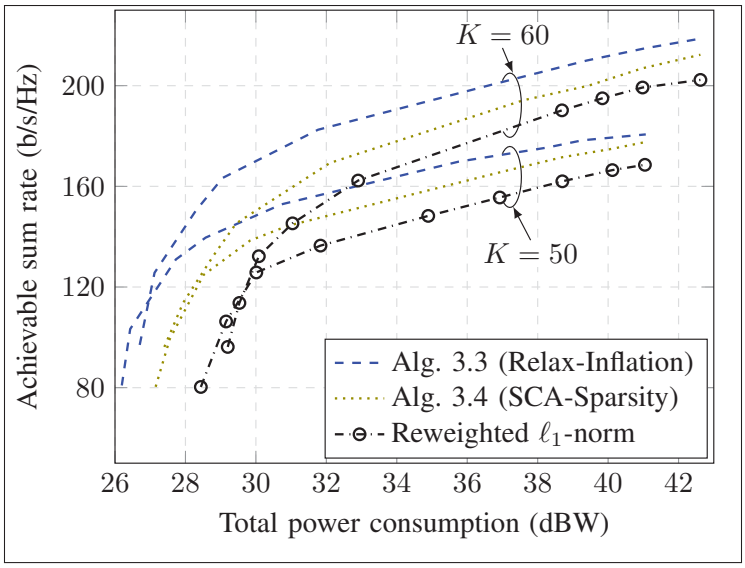


Figure 3.14 Trade-off curves with $K = 50$ and $K = 60$.

between the ASR and TPC of low-complexity algorithms is plotted by varying $\alpha \in [0, 1]$. As shown in Fig. 3.14, the ASR and TPC increase when the weight associated with sum achievable rate (i.e., α) increases. The reasons is that in this case, the objective is in favor of sum rate maximization rather than power consumption minimization. This leads to more power consumption needed to obtain the higher ASR. Moreover, when the number of UEs increases, so do the ASR and TPC. It can be clearly explained that for higher number of UEs, more RRHs should be active to provide sufficient degree of freedom, leading to an increase in the TPC and also ASR. Again, Fig. 3.14 shows that Algorithms 3.3 and 3.4 attain a better performance compared to reweighted ℓ_1-norm algorithm. This demonstrates the effectiveness of our proposed framework.

3.6 Concluding remarks

In this chapter, joint beamforming, RRHs selection and RRH-UE association design has been proposed to maximize achievable sum rate and minimize total power consumption in the DL of C-RAN with limited capacity fronthaul links. In order to solve this multi-objective optimization problem, we have employed the scalarization method to form a scalar weighted sum

objective function. Then, by novel transformations, we have transformed the combinatorial optimization problem into a more tractable form based on which a BRB algorithm has been customized to find an optimal solution. To overcome the high computational complexity of a global optimization method, we have also proposed three low-complexity algorithms by introducing novel transformations and approximation in light of the SCA framework. In the first approach dubbed as SCA-MISOCP algorithm, we have approximated the continuous non-convex part of the design problem and solved a sequence of MI-SOCP problems. Numerical results have shown that the SCA-MISOCP algorithm can attain a performance close to that of the optimal one achieved by the BnRnB algorithm within a few iterations. The second method has been developed based on the continuous relaxation of binary optimization variables and post-processing. This inflation–based SCA-SOCP algorithm has much lower complexity, while still achieves a performance that has been numerically shown close to that of the SCA-MISOCP algorithm. In the final low complexity method, we have reformulated the considered problem from a perspective of sparsity-inducing regularization and utilized the reweighted ℓ_1-norm technique, in combination with SCA method, to solve the resulting problem iteratively. The extensive numerical results have confirmed that our proposed algorithms achieve a good convergence rate under various simulation settings and significantly outperform other reference algorithms.

CHAPTER 4

OPIMAL ENERGY-EFFICIENT BEAMFORMING DESIGNS FOR CLOUD-RANS WITH RATE-DEPENDENT POWER

Phuong Luong [1], François Gagnon[1], Charles Despins[1], Le-Nam Tran[2]

[1] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3,
[2] School of Electronical and Electronic Engineering, University College Dublin
4 Dublin, Ireland

This article was submitted to *IEEE Transactions on Communications* in August 2018
(Luong *et al.*, submitted).

4.1 Introduction

Towards the fifth generation (5G) wireless networks, the exponential growth of data demand from numerous high-speed applications along with seamless area coverage requirement have been drawing significant attention in wireless communications (Andrews *et al.*, 2014). This demand excessively increases the energy consumption due to the additional deployed base stations and transporting network. According to the estimation in (Auer *et al.*, 2011), 80% of total power consumption of a wireless telecommunication network comes from the base station side. This is also the main source of CO_2 emission and tremendous electrical costs for the operators. Consequently, the important of energy efficient design for wireless mobile network has been raised during the past five years.

Emergence of cloud-radio access networks (C-RANs) is foreseen as an essential solution to significantly enhance not only system spectral but also global network energy efficiency (EE) (Wu *et al.*, 2015). Particularly, C-RAN is composed of multiple low-power low-cost RRHs which typically simplified with only radio frequency (RF) functions to handle the transmission/reception of radio signal to/from the users, and clouds of BBUs which are connected with RRHs through fronthauls and centrally execute the sophisticated baseband signals processing tasks instead of RRHs (Rost *et al.*, 2014). This novel architecture enables the huge computing

resources of BBU pools in the cloud exploiting centralized computing capability for baseband signals processing whereas interference suppression and large scale coordinated signal processing can be efficiently achieved, thereby increasing the throughput (Simeone *et al.*, 2016). Further, RRHs are distributed closely to users to furnish less power consumed and higher bit data rate. Thus, the overall EE network is greatly enhanced.

Despite these benefits, C-RAN entails some certain challenges on the radio resource allocation. Firstly, the RRHs cooperation scale depends on the capacity of the fronthauls that are limited in practice, directly imposing a constraint to the potential EE performance of the C-RANs system (Peng *et al.*, 2015). Secondly, the power consumption of fronthauls increases proportionally with the number of associated users, and were not treated appropriately when the user number becomes large (Buzzi *et al.*, 2016). Consequently, it is critically required to consider an energy efficient design of beamforming, RRH-user association and RRH selection along with a precise fronthaul power consumption model to attain the optimal EE performance of C-RANs.

The radio resource allocation for maximizing EE has been deeply investigated in the various wireless communication networks (Tervo *et al.*, 2015; Shi *et al.*, 2016a; Xiong *et al.*, 2016; Pan *et al.*, 2016). For example, an efficient iterative algorithm was proposed in (Tervo *et al.*, 2015) by applying the successive convex approximation (SCA) method to maximize EE in multi-user multiple input single output (MISO) system while the problem of EE maximization in two-tiers wireless backhaul HetNets was considered in (Nguyen *et al.*, 2017a, 2016a,c). The EE maximization problem of joint beamforming and power splitting design for MISO SWIPT systems was studied in (Shi *et al.*, 2016a), where a Lagrangian relaxation coupled with Dinkelbach method was proposed. Furthermore, a low-complexity approximation that each designed factor is optimized given others was developed in (Xiong *et al.*, 2016) to obtain the maximal EE in multi-relay OFDM networks. To overcome the weighted sum EE maximization problem for MISO interference channels, an efficient distributed beamforming algorithm based pricing mechanism was proposed in (Pan *et al.*, 2016).

To address the weighted energy efficiency (EE) maximization problem in C-RANs, a generalized weighted minimum mean square error (WMMSE) approach was used in (Peng *et al.*, 2016) under the Lyapunov framework. By exploiting the ℓ_1/ℓ_2-norm approximation and the block coordinate descent (BCD) method, an iterative mechanism based algorithm was proposed to maximize the minimum EE in an user-centric green C-RAN (Lin *et al.*, 2016). For energy efficient D2D communications in C-RAN, a joint design of channel selection and power allocation was presented in (Zhou *et al.*, 2016), which is solved by applying the Dinkelbach method. In (Pompili *et al.*, 2016), the authors proposed virtual base station architectures for improving EE in C-RAN. In (Dai & Yu, 2016), an energy consumption minimization algorithm was developed via the design of beamforming and user association in C-RAN by utilizing the reweighted ℓ_1–norm approximation method while Shi *et al.* (2016b) employed smoothed ℓ_p minimization approach for solving the beamforming and RRH selection solution of power consumption minimization problem. Motivated by the reweighted ℓ_1-norm approximation technique in (Dai & Yu, 2016), Ariffin *et al.* (2017) proposed the joint design of beamforming and energy to minimize the total energy cost in C-RAN where RRHs are equipped with renewable energy resources. Guo *et al.* (2016b) studied the framework for green C-RAN by optimizing computation provisioning in the BBU pool coupled with hybrid clustering. For the analytical framework, Liu & Yang (2016) analyzed the maximal EE gain given a cache strategy at the base stations under the condition of limited capacity backhaul.

In this paper, we study various EE metrics of the limited fronthaul capacity C-RAN via solving the resource allocation optimization probems. Compared to existing work in (Shi *et al.*, 2014; Luo *et al.*, 2015; Lin *et al.*, 2016), we consider a practical model where the power consumed by the fronthaul links depends on the associated user's rate served by the corresponding RRH (Dai & Yu, 2016). The formulated problems are generally combinatorial non-convex, which leads to the following challenges: (i) the fractional and non-convex nature of the objective functions, (ii) the non-convexity of the per-fronthaul capacity constraints and (ii) the combinatorial nature of the introduced binary variables. Another problem is that even if these binary association variables are relaxed to be continuous, the resulted problem is still non-convex.

A significant departure from previous works of EE maximization where Dinklebach method is applied to convert the fractional objective function into the subtraction form, we directly tackle the considered problem by proposing novel transformations to arrive at an equivalent but more tractable form. Based on this, we develop an iterative algorithm using SCA method (Marks & Wright, 1977) to arrive at a MI-SOCP in each iteration, which can be solved optimally by dedicated MI-SOCP solvers to compute a high-quality feasible solution. Our contributions are expressed as follows:

- We study three different EE objectives: (i) GEE which quantifies the EE of overall network, (ii) WSEE which focuses on controlling the EE of each individual RRH, (iii) and FEE which provides the best EE fairness among all RRHs. Unlike the existing models where the fronthaul power is a quadratic or linear function of the respective resource variables, we consider the rate-dependent power model where the power consumed by the fronthaul depends on the rate served by the corresponding RRH. Compared to the literature (Peng *et al.*, 2016; Lin *et al.*, 2016; Zhou *et al.*, 2016; Pompili *et al.*, 2016; Dai & Yu, 2016; Shi *et al.*, 2016b; Ariffin *et al.*, 2017; Guo *et al.*, 2016b), our work is the first one consider the problems of GEE, WSEE, and FEE maximization problems which takes into account the this rate-dependent power model.

- We first transform the three EE problems into the equivalent forms which are more suitable with the customized branch-and-reduce-and-bound (BRB) algorithm, where global optimal solutions can be attained. To develop a practical and low-compelxity solution approach which can generally be applied to all the problems, we propose some unified transformation techniques to transform all the problems into the forms which are more amenable to the SCA method. Then, we employ the continuous relaxation on the binary variable together with appropriate convex approximation techniques to approximate the non-convex problem into a sequence of convex approximated one. Finally, we present an unified algorithm based on the SCA and Relaxation method to iterative solve the the sequence of approximiated problems and compute the high-quality sub-opitmal solution at convergence.

- Extensive numerical results are presented to show the effectiveness of our proposed algorithms in terms of convergent speed and achieved near-to-optimal EE for three different EE metrics. Finally, the impact of proposed rate-dependent power consumption on the EE of the C-RANs is demonstrated as it outperforms the linear fronthaul power consumption model which was used in the literature.

4.1.1 Organization

The rest of the paper is organized as follows. Section 4.2 introduces the system model. Section 4.3 formulates our joint transmit beamformers, RRH selection, and RRH-UE association design into three different energy efficiency optimization problems. Section 4.4 provides the globally optimal algorithm. In Section 4.5, we introduce our proposed SCA based low complexity algorithms. Section 4.6 presents our numerical results and insight discussions under different simulation scenarios. Finally, the conclusion of the paper is given in Section 4.7.

4.2 System Model

4.2.1 Transmission Model

We consider the DL of C-RAN consisting of I RRHs and K single antenna UEs. For notational convenience, we denote $\mathscr{I} = \{1,\ldots,I\}$ and $\mathscr{K} = \{1,\ldots,K\}$ as the set of RRHs and UEs, respectively. We assume that the ith RRH is equipped with M_i antennas,$\forall i \in \mathscr{I}$. As shown in Fig. 4.1, we assume that all the RRHs are connected to BBU pool via the fronthaul links, e.g., high-speed optical ones, where the ith link has a predetermined maximum capacity $C_i^{\max}$. Each UE is served by a specific group of RRHs but one RRH can serve more than one users simultaneously. Let us denote by s_k the signal with unit power, i.e., $\mathbb{E}\left\{s_k s_k^*\right\} = 1$, intended for the kth UE and by $\mathbf{w}_{i,k} \in \mathbb{C}^{M_i \times 1}$ the transmit beamforming vector from the ith RRH to the kth UE. The vector of channel coefficients encompassing small-scale fading and pathloss from the ith RRH to the kth UE is represented by $\mathbf{h}_{i,k} \in \mathbb{C}^{M_i \times 1}$. For notational convenience we denote the set of beamforming vectors intended for the kth UE as $\mathbf{w}_k \triangleq [\mathbf{w}_{1,k}^T, \mathbf{w}_{2,k}^T, \ldots, \mathbf{w}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, and

the vector including the channels from all RRHs to the kth UE as $\mathbf{h}_k \triangleq [\mathbf{h}_{1,k}^T, \mathbf{h}_{2,k}^T, \ldots, \mathbf{h}_{I,k}^T]^T \in \mathbb{C}^{M\times 1}$, where $M = \sum_{i\in\mathscr{I}} M_i$. Using these notations, the received signal at the kth UE is given by

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \sum_{j\in\mathscr{K}\setminus k} \mathbf{h}_k^H \mathbf{w}_j s_j + z_k \tag{4.1}$$

where $z_k \sim \mathscr{CN}(0,\sigma_0^2)$ is the additive white Gaussian noise (AWGN) and σ_0^2 is the noise power. Note that in (4.1), we have assumed that the kth UE is connected to all the RRHs, but the ith RRH serves the kth UE only if $\left\|\mathbf{w}_{i,k}\right\|_2 > 0$. By treating interference as noise, the achievable rate in b/s/Hz at the kth UE is given by

$$R_k(\mathbf{w}) = \log_2(1+\Gamma_k(\mathbf{w})) \tag{4.2}$$

where

$$\Gamma_k(\mathbf{w}) = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j\in\mathscr{K}\setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2} \tag{4.3}$$

and $\mathbf{w} \triangleq [\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_k^T]^T \in \mathbb{C}^{(KM)\times 1}$ is vector stacking the beamformers for all users.

4.2.2 Fronthaul Capacity Constraint

After the BBU pool performs a relevant radio resource management algorithm to determine the beamforming vectors, data for the kth UE is routed from the BBU pool to the ith RRH via the ith fronthaul link only if $\left\|\mathbf{w}_{i,k}\right\|_2 > 0$. For the transmission to be feasible, the capacity of the ith fronthaul link should be greater than or equal to the total achievable rate at the ith RRH. For the purpose of problem formulation, let us introduce binary variables $a_{i,k} \in \{0,1\}, \forall i \in \mathscr{I}$ and $k \in \mathscr{K}$ to represent the association status between the ith RRH and the kth UE, i.e., $a_{i,k} = 1$ implies that the kth UE is served by the ith RRH and $a_{i,k} = 0$, otherwise. Then, the per-fronthaul capacity constraints can be written as

$$\sum_{k\in\mathscr{K}} a_{i,k} R_k(\mathbf{w}) \leq C_i^{\max}, \forall i \in \mathscr{I}. \tag{4.4}$$

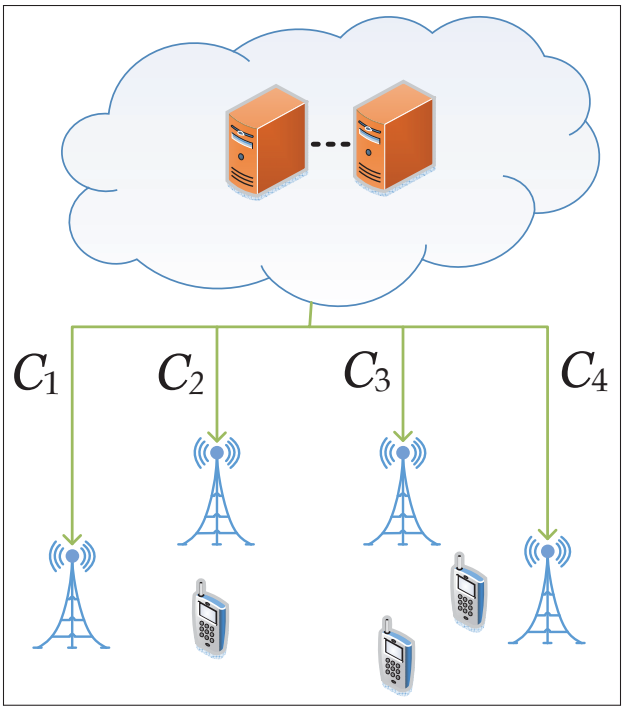


Figure 4.1 Limited fronthaul C-RAN.

4.2.3 Power consumption

1. According to (Shi *et al.*, 2014), the power consumption at a RRH consists of two parts: (i) the power dispatched at the power amplifiers in each RRH which is a function of transmitted signals; (ii) the static power dued to electronic components which are powers required to keep the ith RRH and its corresponding fronthau link in active and sleep mode, denoted by P_i^{ra} and P_i^{ri}, respectively. To represent the operation mode of the ith RRH, we introduce a binary variable $b_i = \{0,1\}, \forall i \in \mathscr{I}$. In particular, $b_i = 0$ states that the ith RRH is in sleep mode and $b_i = 1$ means otherwise. The sum power consumption at the ith RRHs can be expressed as below

$$P_i^{\text{RRH}}(\mathbf{w}, b_i) = \frac{1}{\eta_i} \sum_{k \in \mathscr{K}} \left\| \mathbf{w}_{i,k} \right\|_2^2 + b_i P_i^{\text{ra}} + \left(1 - b_i\right) P_i^{\text{ri}} \tag{4.5}$$

where $\eta_i \in [0,1]$ is the power amplifier efficiency at the ith RRH.

2. In the fronthaul links, the power consumption for forwarding information data and beamformers related to the transmission to the ith RRH from the BBU pool is denoted by P_i^{FH}. More importantly, the term of fronthaul power consumption P_i^{FH} should be correctly characterized in C-RAN. We model that each ith fronthaul transfers the total transmission rates of users served by the corresponding RRH, thus the power consumed for the ith fronthaul is proportional to the sum achievable rates at each ith RRH Dai & Yu (2016). Hence, we have the rate dependent fronthaul power consumption model expressed as

$$P_i^{\mathrm{FH}}(\mathbf{w},\mathbf{a}_i) = \rho_i \sum_{k\in\mathscr{K}} a_{i,k} R_k(\mathbf{w}) \tag{4.6}$$

where ρ_i is a constant scaling factor and $\mathbf{a}_i = [\mathbf{a}_{i,1},\ldots,\mathbf{a}_{i,K}]^T$. In summary, the total power consumption of all RRHs and fronthaul links is given by

$$P^{\mathrm{tot}}\big(\mathbf{w},\mathbf{a},\mathbf{b}\big) = \sum_{i\in\mathscr{I}} \left(\underbrace{P_i^{\mathrm{FH}}(\mathbf{w},\mathbf{a}_i) + P_i^{\mathrm{rrh}}(\mathbf{w},b_i)}_{P_i(\mathbf{w},\mathbf{a}_i,\mathbf{b}_i)} \right) \tag{4.7}$$

where we denote $\mathbf{b} = [b_1,\ldots,b_I]^T$, and $\mathbf{a} = [\mathbf{a}_1^T,\ldots,\mathbf{a}_I^T]^T$.

4.3 Problem Formulation

In this paper, we study various different EE metrics of designing a green communication. First, global energy efficiency (GEE) is referred as the ratio of network sum rate and the total network power consumption, which is presented as

$$f_{\mathrm{GEE}}(\mathbf{w},\mathbf{a},\mathbf{b}) = \frac{R^{\mathrm{tot}}(\mathbf{w})}{P^{\mathrm{tot}}\big(\mathbf{w},\mathbf{a},\mathbf{b}\big)} \tag{4.8}$$

where the total rate over all RRHs is given by

$$R^{\mathrm{tot}}(\mathbf{w}) = \sum_i \sum_k a_{i,k} R_k(\mathbf{w}). \tag{4.9}$$

GEE is the common EE objective in the literature used to quantify the system EE performance, however GEE is not able to control the maximization of individual EEs which have different priorities. To handle this issue, the weighted sum of individual energy efficiency (WSEE) over all RRHs is studied and given by

$$f_{\text{WSEE}}(\mathbf{w},\mathbf{a},\mathbf{b}) = \sum_{i\in\mathscr{I}} \lambda_i \frac{\sum_{k\in\mathscr{K}} a_{i,k} R_k(\mathbf{w})}{P_i(\mathbf{w},\mathbf{a}_i,\mathbf{b}_i)} \tag{4.10}$$

where $\lambda = [\lambda_1^T,\ldots,\lambda_I^T]^T$ is the vector of the weights associated to the corresponding RRHs. It is worth mentioning that GEE and WSEE fail to guarantee the fairness between RRHs, leading the introduction of EE fairness metric which aims at maximizing the minimum EE across all RRHs and is mathematically stated as

$$f_{\text{FEE}}(\mathbf{w},\mathbf{a},\mathbf{b}) = \min_{i\in\mathscr{I}} \{ \frac{\sum_{k\in\mathscr{K}} a_{i,k} R_k(\mathbf{w})}{P_i(\mathbf{w},\mathbf{a}_i,\mathbf{b}_i)} \} \tag{4.11}$$

We are now ready to formulate a joint design of beamforming, RRH-UE association and RRH selection that maximizes the three different EE merits in the considered C-RAN system as

$$(\mathscr{E}_{\text{X}}): \max_{\mathbf{b},\mathbf{a},\mathbf{w},\nu} \; f_{\text{X}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{4.12a}$$

$$\text{s.t.} \;\; \Gamma_k(\mathbf{w}) \geq \Gamma_k^{\min}, \forall k \in \mathscr{K} \tag{4.12b}$$

$$\sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq b_i P^{\max}, \forall i \in \mathscr{I} \tag{4.12c}$$

$$\left\|\mathbf{w}_{i,k}\right\|_2^2 \leq a_{i,k}\nu_{i,k}, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{4.12d}$$

$$\nu_{i,k} \leq a_{i,k} P^{\max}, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{4.12e}$$

$$a_{i,k} \leq b_i, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{4.12f}$$

$$\sum_{k\in\mathscr{K}} a_{i,k} R_k(\mathbf{w}) \leq C_i, \forall i \in \mathscr{I} \tag{4.12g}$$

$$a_{i,k} \in \{0,1\}, b_i \in \{0,1\}, \forall i \in \mathscr{I}, \forall k \in \mathscr{K} \tag{4.12h}$$

where X can be GEE, WSEE, and FEE representing for the performance metrics of EE introduced in (4.8), (4.10), and (4.11), respectively. In addition, the introduction of the set of

auxiliary variables $\nu = \left\{ \nu_{i,k} \geq 0, \forall i \in \mathscr{I}, k \in \mathscr{K} \right\}$ and the constraints in (4.12) deserve further explanation. Intuitively, $\nu_{i,k}$ represents the *soft* power transmitted from the ith RRH to UE k. Constraint (4.12b) is to ensure the QoS requirement for the kth user, where $\Gamma_k^{\min}$ is the predetermined SINR requirement for the kth user. Moreover, constraint (4.12c) implies that the total transmit power at each RRH is limited by a given budget power $P^{\max}$. The constraints (4.12c) and (4.12d) are to make sure that when the ith RRH is in sleep mode, e.g., $b_i = 0$, no power will be transmitted from it. This can be easily seen as $b_i = 0$, then $a_{i,k} = 0$ for all $k \in \mathscr{K}$ and $\sum_{k \in \mathscr{K}} \left\| \mathbf{w}_{i,k} \right\|_2^2 = 0$. Similarly, in (4.12d) we also guarantee that the transmit power $\left\| \mathbf{w}_{i,k} \right\|_2^2$ from the ith RRH to the kth user is zero if $a_{i,k} = 0$. The constraint in (4.12e) means that the soft power from the ith RRH to the kth user should not exceed $P^{\max}$. Finally, the per-fronthaul capacity constraint is explicitly presented in (4.12g). Note that, the constraint (4.12d) is called a rotated SOC which can be reformulated as an SOC

$$\left(a_{i,k} + \nu_{i,k}\right)/2 \geq \left\| \left[\left(a_{i,k} - \nu_{i,k}\right)/2, \mathbf{w}_{i,k}^T \right]^T \right\|_2 \tag{4.13}$$

4.4 Globally Optimal Solution

In this section we present a solution to solve $(\mathscr{E}_{\mathrm{X}})$ optimally. Before proceeding further, we provide some comments on the complexity of $(\mathscr{E}_{\mathrm{X}})$. First, as we mentioned above that problem $(\mathscr{E}_{\mathrm{X}})$ is generally NP-hard. Moreover, even when binary variables **a** and **b** are relaxed to be continuous, the obtained problem is still non-convex because of the non-convexity of the objective function (4.12a) and the constraint (4.12g). More precisely, in mathematical programming, $(\mathscr{E}_{\mathrm{X}})$ is categorized as a mixed integer non-convex problem for which such a method in (Tang *et al.*, 2015; Guo *et al.*, 2016b; Cheng *et al.*, 2013) is not applicable to find a globally optimal solution. To the best of our knowledge, there is no off-the-shelf solver for $(\mathscr{E}_{\mathrm{X}})$. In what follows, we present an equivalent formulation of $(\mathscr{E}_{\mathrm{X}})$, based on which a binary branch and reduce and bound (BRB) algorithm using monotonic optimization (MO) is customized to solve $(\mathscr{E}_{\mathrm{X}})$ optimally.

4.4.1 Equivalent Formulation

Let us introduce the slack variables $\tau = [\tau_1, \ldots, \tau_K]^T$, $\mathbf{t} = [t_0, t_1, \ldots, t_I]^T$, and compact the variables $\mathbf{a}$, τ and $\mathbf{t}$ into a vector denoted as $\mathbf{s} = [\mathbf{a}^T, \tau^T, \mathbf{t}^T]^T \in R_+^S$ with dimension $S = N + K + I + 1$ where $N = IK$ is the length of vector $\mathbf{a}$. Thus, we are able to rewrite $(\mathscr{E}_{\mathrm{X}})$ into the following equivalent problem

$$\max_{\mathbf{a},\mathbf{b},\mathbf{w},\nu,\tau,\mathbf{t}} \; f_{\mathrm{X}}(\mathbf{s}) \tag{4.14a}$$

$$\text{s.t. } R_k(\mathbf{w}) \geq \tau_k, \forall k \in \mathscr{K} \tag{4.14b}$$

$$\tau_k \geq \log_2(1 + \Gamma_k^{\min}), \forall k \in \mathscr{K} \tag{4.14c}$$

$$\begin{cases} \tilde{P}^{\mathrm{tot}}(\mathbf{w}, \mathbf{a}, \mathbf{b}, \tau) \leq 1/t_0, & \text{if X is GEE} \\ \tilde{P}_i(\mathbf{w}, \mathbf{a}, \mathbf{b}, \tau) \leq 1/t_i, \forall i \in \mathscr{I} & \text{otherwise} \end{cases} \tag{4.14d}$$

$$\sum_{k \in \mathscr{K}} a_{i,k} \tau_k \leq C_i, \forall i \in \mathscr{I} \tag{4.14e}$$

$$(4.12c), (4.12e), (4.12f), (4.12h), (4.13). \tag{4.14f}$$

where three different EE objectives are expressed in the form of $f_{\mathrm{X}}(\mathbf{s})$ as below

$$f_{\mathrm{X}}(\mathbf{s}) = \begin{cases} t_0 \sum_i \sum_k a_{i,k} \tau_k & \text{if X is GEE} \\ \sum_{\forall i \in \mathscr{I}} \lambda_i t_i \sum_k a_{i,k} \tau_k & \text{if X isWSEE} \\ \min_{\forall i \in \mathscr{I}} \{ t_i \sum_k a_{i,k} \tau_k \} & \text{if X is FEE} \end{cases} \tag{4.15}$$

According to the slack variable introduction of τ and $\mathbf{t}$ and the equivalent transformation, the total power consumption function $P^{\mathrm{tot}}(\mathbf{w}, \mathbf{a}, \mathbf{b})$ and power consumption function at each RRH and corresponding fronthaul $P_i(\mathbf{w}, \mathbf{a}_i, \mathbf{b}_i)$ in the denominator of objective function in (4.12a) are respectively rewritten in the newly additional constraints in (4.14d) as

$$\tilde{P}^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b},\tau) = \sum_{i\in\mathscr{I}} \left(\rho_i \sum_{k\in\mathscr{K}} a_{i,k}\tau_k + P_i^{\text{rrh}}(\mathbf{w},b_i) \right) \tag{4.16}$$

$$\tilde{P}_i(\mathbf{w},\mathbf{a},\mathbf{b},\tau) = \rho_i \sum_{k\in\mathscr{K}} a_{i,k}\tau_k + P_i^{\text{rrh}}(\mathbf{w},b_i), \forall i \in \mathscr{I} \tag{4.17}$$

The following lemma is to characterize the property of problem (4.14).

Lemma 2. *The formulations in the problem* (4.12) *and* (4.14) *are equivalent in the sense that they have the same optimal solution set and objective.*

Proof. The equivalence between (4.12) and (4.14) is due to the observation that at optimality of (4.14), the inequalities (4.14b) and (4.14d) must hold with equalities. In addition, an optimal solution set to (4.14) is also optimal solution set to (4.12). The detail of proof are presented in Appendix 3. □

4.4.2 Optimal Solution based BRB Algorithm for Problem (4.14)

In this subsection, we aim at solving the problem (4.14) optimally using the BRB algorithm based on MO framework. This is possible due to two following important observations.

- The objective in (4.15) monotonically increases with respect to (w.r.t) each entry of $\mathbf{s}$. Particularly, we can observe that the objective increases if we keep increasing each of $\mathbf{a}$, $\mathbf{t}$ or τ as long as it is still feasible to (4.14).
- For given s, the resulted problem of (4.14) becomes a feasibility checking problem, which is recognized as mixed-integer second order cone program (MI-SOCP) feasibility problem and can be solved optimally by dedicated solvers such as MOSEK.

Herein we present the customized steps required for solving the considered problem (4.14). Specifically, we define the compact normal set $\mathscr{S} = \{s \in \mathbb{R}_+^L | (4.14\text{b}) - (4.14\text{f})\}$ as the feasible

set of (4.14) and $\mathscr{U} = [\underline{s}; \bar{s}]$ as the box that contains all s feasible to (4.14). The calculation of box $\mathscr{U}$ is presented detailed in Appendix 4. The general idea to solve the problem (4.14) optimally in a BRB algorithm using MO framework is to check if a given $s \in \mathscr{U}$ belongs to $\mathscr{S}$ or not. Mathematically we need to solve the following feasibility problem for a given s

$$\text{find } \mathbf{b}, \mathbf{w}, \nu \tag{4.18a}$$

$$\text{s.t. } (5.10\text{b}) - (5.10\text{i}). \tag{4.18b}$$

It is easy to check that when τ is fixed, (4.14b) can be reformulated as SOC constraint as

$$c\Re(\mathbf{h}_k^H \mathbf{w}_k) \geq \left\| \mathbf{h}_k^H \mathbf{w}_1, \ldots, \mathbf{h}_k^H \mathbf{w}_K, \sigma_0 \right\|_2 \tag{4.19}$$

where $c = \sqrt{\frac{1}{2^{\tau_k}-1} + 1}$. Similarly, for given $\mathbf{s}$ the constraint (4.14d) is SOC representable as

$$\begin{cases} \frac{1/t_0 - \hat{P}^{\text{tot}}(\mathbf{b})+1}{2} \geq \left\| \left[\frac{\mathbf{w}_{1,1}^T}{\sqrt{\eta_1}}, \ldots, \frac{\mathbf{w}_{I,K}^T}{\sqrt{\eta_I}}, \frac{1/t_0 - \hat{P}^{\text{tot}}(\mathbf{b})-1}{2} \right]^T \right\|_2 & \text{if X is GEE} \\ \frac{1/t_i - \hat{P}_i(\mathbf{b})+1}{2} \geq \left\| \left[\frac{\mathbf{w}_{1,1}^T}{\sqrt{\eta_1}}, \ldots, \frac{\mathbf{w}_{I,K}^T}{\sqrt{\eta_I}}, \frac{1/t_i - \hat{P}_i(\mathbf{b})-1}{2} \right]^T \right\|_2, \forall i \in \mathscr{I} & \text{otherwise} \end{cases} \tag{4.20}$$

where

$$\hat{P}^{\text{tot}}(\mathbf{b}) = \sum_{i \in \mathscr{I}} \rho_i \sum_{k \in \mathscr{K}} a_{i,k} \tau_k + \sum_{i \in \mathscr{I}} \left(b_i P_i^{\text{ra}} + \left(1 - b_i\right) P_i^{\text{ri}} \right) \tag{4.21}$$

$$\hat{P}_i(\mathbf{b}) = \rho_i \sum_{k \in \mathscr{K}} a_{i,k} \tau_k + b_i P_i^{\text{ra}} + \left(1 - b_i\right) P_i^{\text{ri}}, \forall i \in \mathscr{I} \tag{4.22}$$

Thus, the feasibility checking problem (4.18) is in fact a MISOCP problem.

Based on the above analysis, problem (4.14) can now be expressed as $\max\{f_{\text{X}}(s) | s \in \mathscr{S} \subset \mathscr{U}\}$. First, we check whether $\underline{s}$ is feasible or not. If so, we apply the proposed BRB method to find a globally optimal solution to (4.14). The proposed method recursively branches a box $\mathscr{U}$ into two smaller boxes, checks the feasibility of each new box, update the current upper and lower bounds by the box reduction and bound computation process, and removes the boxes that do not

contain an optimal solution. Note that, the variable of interest **s**comprises both binary variables **a** and continuous variables τ and **t**, thus the branching and reduction procedures need to be adjusted to guarantee the exact solution of binary variables. The operations of BRB algorithm is presented as follows.

- Box branching: At each iteration, the box which currently contains the largest upper bound is selected to branch. This box is split into two smaller boxes by a standard bi-partition along the longest edge (Tuy & khayya-and P. Thach, 2005). In particular, supposed that the box $\mathscr{B} = [\mathbf{x}, \mathbf{y}]$ is chosen to branch, then it is divided into two following subboxes

$$\mathscr{B}^{(1)} = \begin{cases} [\mathbf{x}, \mathbf{y} - \mathbf{e}_l] & \text{if } l \leq N \\ [\mathbf{x}, \mathbf{y} - \mathbf{e}_l(y_l - x_l)/2] & \text{if } l > N \end{cases} \text{ and } \mathscr{B}^{(2)} = \begin{cases} [\mathbf{x} + \mathbf{e}_l, \mathbf{y}] & \text{if } l \leq N \\ [\mathbf{x} + \mathbf{e}_l(y_l - x_l)/2, \mathbf{y}] & \text{if } l > N \end{cases} \tag{4.23}$$

where $l = \arg\max_{l \in \{0,\ldots,S\}} (y_l - x_l)$ is the index of the longest edge of $\mathscr{B}$ and $\mathbf{e}_l$ is a $S \times 1$ vector where all entries are zero except that the lth entry is 1. By using the above branching rules, the branched binary variables are guaranteed to lie in binary cutting plane.

- Box reduction: Supposed that we perform the reduction process on the box $\mathscr{B}^{(1)} = [\mathbf{p}, \mathbf{q}]$ to obtain the reduced box redu$(\mathscr{B}^{(1)})$ without any loss of optimality. Note that a similar argument can be applied to $\mathscr{B}^{(2)}$ as well. First, we check whether $\mathscr{B}^{(1)}$ contains at least one feasible solution to problem (5.10) or not by checking the feasibility of (5.11) for the given lower bound set **p**. If (5.11) is infeasible, it means that $\mathbf{p} \in \mathscr{U} \setminus \mathscr{S}$ and $\mathscr{B}^{(1)}$ does not contain any feasible solution to (5.10) so that it can be discarded. Otherwise, the reduction process is performed to find redu$(\mathscr{B}^{(1)}) = [\mathbf{p}', \mathbf{q}']$. In particular, we calculate $\mathbf{p}' = \mathbf{q} - \sum_{l \in \{0,\ldots,S\}} \lambda_l (q_l - p_l)\mathbf{e}_l$, where $\lambda_l = \sup\{\lambda | \lambda \in [0,1], \mathbf{q} - \lambda(q_l - p_l)\mathbf{e}_l \in \mathscr{U} \setminus \mathscr{S}\}$, $\forall l \in \{0,\ldots,S\}$ and $\mathbf{q}' = \mathbf{p}' + \sum_{l=\{0,\ldots,S\}} \beta_l (q_l - p'_l)\mathbf{e}_l$, where $\beta_l = \sup\{\beta | \beta \in [0,1], \mathbf{p}' + \beta(q_l - p'_l)\mathbf{e}_l \in \mathscr{S}\}$, $\forall l \in \{0,\ldots,S\}$. The problem of finding λ_l and β_l can be solved easily using a bisection procedure over $\lambda \in [0,1]$

and $\beta \in [0,1]$, respectively. To guarantee p_l' , $q_l' \in \{0,1\}$, $\forall l \leq N$ while reducing, we do

$$p_l' = \begin{cases} 1 & \text{if } \mathbf{q} - \mathbf{e}_l \in \mathscr{U} \setminus \mathscr{S} \\ 0 & \text{otherwise} \end{cases} \text{ and } q_l' = \begin{cases} 1 & \text{if } \mathbf{p}' + \mathbf{e}_l \in \mathscr{S} \\ 0 & \text{otherwise} \end{cases}. \tag{4.24}$$

It is shown in (Tuy & khayya-and P. Thach, 2005) that if $\mathscr{B}^{(1)}$ contains an optimal solution, then redu$(\mathscr{B}^{(1)})$ also contains this optimal solution.

- Bounding and Pruning: due to the monotone objective function, the upper and lower bound of a box $\mathscr{B} = [\underline{s}, \overline{s}]$ is simply computed $U(\mathscr{B}) = f(\overline{s})$ and $L(\mathscr{B}) = f(\underline{s})$, respectively. After updating the current best lower bound ζ_n, the pruning is performed to delete the boxes whose upper bounds are smaller than ζ_n. According to (Tuy & khayya-and P. Thach, 2005), the proposed algorithm is bound improving and terminates after finitely many iterations for a given desired accuracy level ε.

We remark that BRB algorithm requires very high computational complexity due to the MIS-OCP feasiblity checking problem and a large number of iterations to terminate. Thus, BRB algorithm is practically used as benchmark for the low-complexity algorithms that are proposed in the following sections.

4.5 Proposed SCA-based Low Complexity Algorithms

4.5.1 General SCA method

Before elaborating the SCA method to solve the EE maximization problems, we take an opportunity to present how SCA method can address a general nonconvex problem. Let us consider

the general nonconvex optimization problem in the following

$$\max_{\mathbf{x}\in\mathscr{X}} f(\mathbf{x}) \tag{4.25a}$$

$$\text{s.t. } g_i(\mathbf{x}) \leq 0,\, i = 1,\ldots,L_1 \tag{4.25b}$$

$$p_j(\mathbf{x}) \leq q_j(\mathbf{x}),\, j = 0,\ldots,L_2 \tag{4.25c}$$

where $g_i(\mathbf{x}) : \mathbb{C}^n \to \mathbb{R}$,$i = 1,\ldots,L_1$, $p_j(\mathbf{x}) : \mathbb{C}^n \to \mathbb{R}$,$j = 1,\ldots,L_2$ and $q_j(\mathbf{x}) : \mathbb{C}^n \to \mathbb{R}$,$j = 1,\ldots,L_2$ are convex functions w.r.t $\mathbf{x} \in \mathbb{C}^N$, respectively. The idea to deal with the nonconvex constraint (4.25c) is to apply the approximation on the nonconvex part of (4.25c) into the convex one. In particularly, assuming $q_j(\mathbf{x})$ is differentiable (which is mostly true for the constraints in wireless communications), SCA linearizes $q_j(\mathbf{x})$ around the current iteration parameters $\mathbf{x}^{(n)}$ to arrive at the following constraint

$$p(\mathbf{x}) - q(\mathbf{x}^{(n)}) - \left\langle \nabla q(\mathbf{x}^{(n)}), \mathbf{x} - \mathbf{x}^{(n)} \right\rangle \leq 0 \tag{4.26}$$

Note that (4.26) implies (4.25c) as a concave function (i.e., $-q(\mathbf{x})$ as mentioned above) is upper bounded by its linearization. In other words, SCA arrives at an inner approximation of the feasible set of the nonconvex program in (4.25), which is expressed by the following convex subproblem

$$\max_{\mathbf{x}\in\mathscr{X}} \{ f(\mathbf{x}) | \, (4.25\text{b}), (4.26) \} \tag{4.27}$$

at the $n+1$th iteration and updates the operating point $\mathbf{x}^{(n)}$ until convergence. The SCA based algorithm to solve the problem (4.25) is presented in Algorithm 4.1.

Algorithm 4.1: SCA based Algorithm

1: Set $n := 0$ and initialize starting points of $\mathbf{x}^{(n)}$;
2: **repeat**
3: Solve the approximated convex problem (4.27) at $\mathbf{x}^{(n)}$ to achieve the optimal solution $\mathbf{x}^\star$;
4: Update $\mathbf{x}^{(n+1)} = \mathbf{x}^\star$;
5: Set $n := n + 1$;
6: **until** Convergence;

In the following subsection, we present how to apply the SCA based algorithm presented in this subsection 4.5.1 to three different formulated EE problems in (4.12). It is worth mentioning that it is not amenable to directly apply this presented methods to three different EE problems in (4.12) since they are not well formed as (4.25). Therefore, it is necessary to propose the new transformations for each EE problem in (4.12) to equivalently transform it into more tractable form so that the nonconvex constraints are revealed and can be handled by the light of SCA method.

4.5.2 SCA-based Algorithm for GEE Maximization Problem

In what follows, we present SCA approach to provide a sub-optimal solution of (4.12) for GEE maximization problem. The proposed method is developed based on the transformations introduced in (Luong *et al.*, 2017a).

4.5.2.1 Equivalent transformations

First, let us introduce the new slack variables $\psi \geq 0$, $\mu \geq 0$, $\varphi = \{\varphi_{i,k} \geq 0\}_{\forall i \in \mathscr{I}, \forall k \in \mathscr{K}}$, $\phi = \{\phi_k \geq 0\}_{\forall k \in \mathscr{K}}$, $\gamma = \{\gamma_k \geq 0\}_{\forall k \in \mathscr{K}}$ and $\mathbf{z} = \{z_k \geq 0\}_{\forall k \in \mathscr{K}}$, and consider the following

equivalent formulation

$$\max_{\substack{\mathbf{b},\mathbf{a},\mathbf{w},\nu\\ \mu,\gamma,\phi,\mathbf{z},\psi,\varphi}} \quad \psi \tag{4.28a}$$

$$\text{s.t.} \quad \frac{\left(\sum_i \sum_{k\in\mathscr{K}} \varphi_{i,k}^2\right)}{\mu} \geq \psi \tag{4.28b}$$

$$\tilde{P}^{\text{tot}}\left(\mathbf{w},\mathbf{a},\mathbf{b},\mathbf{z}\right) \leq \mu \tag{4.28c}$$

$$a_{i,k}\phi_k \geq \varphi_{i,k}^2 \tag{4.28d}$$

$$\log\left(1+\gamma_k\right) \geq \phi_k \tag{4.28e}$$

$$\Gamma_k(\mathbf{w}) \geq \gamma_k \tag{4.28f}$$

$$\gamma_k \geq \Gamma_k^{\min} \tag{4.28g}$$

$$\sum_{k\in\mathscr{K}} \left(a_{i,k}/z_k\right) \leq C_i \tag{4.28h}$$

$$R_k(\mathbf{w}) \leq 1/z_k \tag{4.28i}$$

$$(4.12\text{c}), (4.12\text{e}), (4.12\text{f}), (4.12\text{h}), (4.13). \tag{4.28j}$$

where we denote $\tilde{P}^{\text{tot}}\left(\mathbf{w},\mathbf{a},\mathbf{b},\mathbf{z}\right) = \sum_i P_i^{\text{rrh}}(\mathbf{w},b_i) + \sum_{i\in\mathscr{I}} \rho_i \sum_{k\in\mathscr{K}} \left(a_{i,k}/z_k\right)$. Note that the meaning of $\varphi_{i,k}^2$ is considered as the transmission rate of user k transported through the fronthaul link i. It is easy to see that all the constraints (4.28b) and (4.28c) are active at optimality. Moreover, a feasible solution of (4.28) is also feasible to (4.12). Thus, (4.28) is equivalent to (4.12).

It is recognized that the newly additional constraint (4.28d) is SOC representable as $\frac{\phi_k + a_{i,k}}{2} \geq \left\| \varphi_{i,k}, \frac{\phi_k - a_{i,k}}{2} \right\|_2$. To proceed further, we use a subtle observation that $a_{i,k} = a_{i,k}^2$ for $a_{i,k} \in \{0,1\}$ to replace $a_{i,k}$ by $a_{i,k}^2$ and rewrite the nonconvex constraint in (4.28h) into two equivalent inequalities as

$$a_{i,k}^2 \leq \theta_{i,k} z_k, \tag{4.29}$$

$$\sum_{k\in\mathscr{K}} \theta_{i,k} \leq C_i, \tag{4.30}$$

with newly introduced variables $\theta = \{\theta_{i,k} \geq 0\}_{\forall i \in \mathscr{I}, k \in \mathscr{K}}$. By taking advantage of the transformation in (4.29), we can rewrite (4.28c) by a convex constraint as

$$\sum_i P_i^{\text{rrh}}(\mathbf{w}, b_i) + \sum_{i \in \mathscr{I}} \rho_i \sum_{k \in \mathscr{K}} \theta_{i,k} \leq \mu \tag{4.31}$$

Finally, to reveal the hidden convexity of (4.28i), we can equivalently rewrite it as

$$1 + \xi_k \leq \exp(1/z_k) \tag{4.32}$$

$$\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\xi_k} \leq \sum_{j \neq k}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2, \tag{4.33}$$

where $\xi = \{\xi_k \geq 0, \forall k \in \mathscr{K}\}$ is the set of newly introduced variables. By denoting the set of variables as $\Psi_{\text{GEE}} = \{\mathbf{b}, \mathbf{a}, \mathbf{w}, \nu, \theta, \mu, \gamma, \phi, \mathbf{z}, \psi, \xi, \varphi\}$, the final equivalent problem after being transformed is presented as

$$\max_{\Psi_{\text{GEE}}} \ \psi \tag{4.34a}$$

$$\text{s.t. } \psi \leq \frac{\left(\sum_i \sum_{k \in \mathscr{K}} \varphi_{i,k}^2\right)}{\mu} \tag{4.34b}$$

$$\sum_i P_i^{\text{rrh}}(\mathbf{w}, b_i) + \sum_{i \in \mathscr{I}} \rho_i \sum_{k \in \mathscr{K}} \theta_{i,k} \leq \mu \tag{4.34c}$$

$$a_{i,k} \phi_k \geq \varphi_{i,k}^2 \tag{4.34d}$$

$$\log(1 + \gamma_k) \geq \phi_k \tag{4.34e}$$

$$\sum_{j \in \mathscr{K} \setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2 \leq \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\gamma_k} \tag{4.34f}$$

$$1 + \xi_k \leq \exp(1/z_k) \tag{4.34g}$$

$$\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\xi_k} \leq \left(\sum_{j \neq k}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2\right) \tag{4.34h}$$

$$(4.12c), (4.12e), (4.12f), (4.12h), (4.13), (4.28g), (4.29), (4.30). \tag{4.34i}$$

where the constraint (4.34f) is rewritten from the constraint (4.28f). We remark that problem (4.34) is still nonconvex but its nonconvex constraints in (4.34f), (4.34c), (4.34g) and (4.34h) are ready to be handled in light of the SCA method as the follows.

4.5.2.2 SCA-GEE Algorithm

In this method we preserve the Boolean variables, and only approximate the continuous nonconvex parts of (4.34). In particular, we do so by applying the framework of SCA. Let us first consider the nonconvex constraint (4.34b) and (4.34f). These functions of $\frac{\left(\sum_i \sum_{k\in\mathscr{K}} \varphi_{i,k}^2\right)}{\mu}$ and $\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\gamma_k}$ have the same form of function $h(p,q) = \frac{|p|^2}{q}$, $\forall p \in \mathbb{C}$,$q \in \mathbb{R}_+$, i.e., in (4.34b) it is $h(\varphi,\mu)$ and in (4.34f), it is $h(\mathbf{w}_k,\gamma_k)$. Thus, at iteration $n+1$ th the proposed algorithm we apply the first order Taylor approximation to $h(p,q)$ around the point of $p^{(n)}$, $q^{(n)}$ by

$$H\left(p,q;p^{(n)},q^{(n)}\right) = \frac{2\Re\left(p^{(n)}p\right)}{q^{(n)}} - \frac{\left|p^{(n)}\right|^2}{q^{(n)2}}q \tag{4.35}$$

where we have denoted $q^{(n)2} = \left(q^{(n)}\right)^2$ to lighten the notation. Particularly, $H(\varphi,\mu;\varphi^{(n)},\mu^{(n)})$ is concave upper bound function of $h(\varphi,\mu)$ around the point of $\varphi^{(n)}$ and $\mu^{(n)}$ given by

$$H(\varphi,\mu;\varphi^{(n)},\mu^{(n)}) = \frac{\sum_i \sum_{k\in\mathscr{K}} 2\varphi_{i,k}^{(n)}\varphi_{i,k}}{\mu^{(n)}} - \frac{\sum_i \sum_{k\in\mathscr{K}} \varphi_{i,k}^{(n)2}}{\mu^{(n)2}}\mu$$

In the same way, we have $H(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)})$ is simply a linearization of function $h(\mathbf{w}_k,\gamma_k) = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\gamma_k}$ around the point of $\mathbf{w}_k^{(n)}$ and $\gamma_k^{(n)}$ given by

$$H(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)}) = \frac{2\Re\left(\mathbf{w}_k^{(n)H}\mathbf{H}_k\mathbf{w}_k\right)}{\gamma_k^{(n)}} - \frac{\left|\mathbf{h}_k^H\mathbf{w}_k^{(n)}\right|^2}{\gamma_k^{(n)2}}\gamma_k \tag{4.36}$$

Here, we have denoted $\mathbf{H}_k \triangleq \mathbf{h}_k\mathbf{h}_k^H$, $\mathbf{w}_k^{(n)H} = \left(\mathbf{w}_k^{(n)}\right)^H$ to lighten the notation. The nonconvex constraints (4.34g) and (4.34h) can also be approximated by its concave upper bound as

$$1+\xi_k - F\left(z_k; z_k^{(n)}\right) \leq 0 \tag{4.37}$$

$$\frac{\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2}{\xi_k} - G\left(\mathbf{w};\mathbf{w}^{(n)}\right) \leq 0 \tag{4.38}$$

where $F(z_k; z_k^{(n)})$ given in (4.39) is lower bound concave approximation of $f(z_k) = \exp(1/z_k)$ around the point of $z_k^{(n)}$ and $G\left(\mathbf{w};\mathbf{w}^{(n)}\right)$ given in (4.40) is lower bound concave approximation of $g(\mathbf{w}) = \sum_{j\neq k}^{K} |\mathbf{h}_k^H\mathbf{w}_j|^2 + \sigma_0^2$ around the point of $\mathbf{w}^{(n)}$.

$$F\left(z_k; z_k^{(n)}\right) = e^{(1/z_k^{(n)})} - \frac{e^{(1/z_k^{(n)})}}{z_k^{(n)2}}\left(z_k - z_k^{(n)}\right) \tag{4.39}$$

$$G\left(\mathbf{w};\mathbf{w}^{(n)}\right) = \sum_{j\neq k}^{K} 2\Re\left(\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j\right) - \sum_{j\neq k}^{K} \mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j^{(n)} + \sigma_0^2 \tag{4.40}$$

By applying these approximations we can obtain a mixed integer convex approximation of problem (4.34) at iteration $n+1$ of the algorithm as

$$\max_{\Psi_{\text{GEE}}} \ \psi \tag{4.41a}$$

$$\text{s.t.} \psi - H(\varphi,\mu;\varphi^{(n)},\mu^{(n)}) \leq 0 \tag{4.41b}$$

$$\sum_i P_i^{\text{rrh}}(\mathbf{w}, b_i) + \sum_{i\in\mathscr{I}} \rho_i \sum_{k\in\mathscr{K}} \theta_{i,k} \leq \mu \tag{4.41c}$$

$$\frac{a_{i,k}+\phi_k}{2} \geq \left\| \varphi_{i,k}, \frac{a_{i,k}-\phi_k}{2} \right\|_2 \tag{4.41d}$$

$$\log(1+\gamma_k) \geq \phi_k \tag{4.41e}$$

$$\sum_{j\in\mathscr{K}\setminus k} |\mathbf{h}_k^H\mathbf{w}_j|^2 + \sigma_0^2 - H\left(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)}\right) \leq 0 \tag{4.41f}$$

$$1+\xi_k - F\left(z_k; z_k^{(n)}\right) \leq 0 \tag{4.41g}$$

$$\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2/\xi_k - G\left(\mathbf{w};\mathbf{w}^{(n)}\right) \leq 0 \tag{4.41h}$$

$$(4.12c), (4.12e), (4.12f), (4.12h), (4.13), (4.28g), (4.29), (4.30) \tag{4.41i}$$

where $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \gamma^{(n)}, \mu^{(n)}, \varphi^{(n)}$ are the parameters to be updated at the $(n+1)$th iteration.

Remark 2. *Note that all the continuous constraints in* (4.41)*, except* (4.41e)*, are convex quadratic representable. However, due to the presence of the exponential cone in* (4.41e)*,* (4.41) *is still recognized as a generic convex mixed-integer program for which dedicated solvers are quite limited. To avail of powerful modern MI-SOCP solvers such as MOSEK or GUROBI, our idea is to approximate* (4.41e) *by a conic constraint and this should be done in light of SCA. More specifically, a conic lower bound of the left hand side of* (4.41e) *is desired.*

The key is due to the following inequality. For any $\gamma_k \geq 0$ it holds that

$$\log(1+\gamma_k) \geq U\big(\gamma_k;\gamma_k^{(n)}\big) = \log\big(1+\gamma_k^{(n)}\big) + \frac{1}{1+\gamma_k^{(n)}}\big(\gamma_k-\gamma_k^{(n)}\big) - \frac{1}{2}\big(\gamma_k-\gamma_k^{(n)}\big)^2. \quad (4.42)$$

In fact $U\big(\gamma_k;\gamma_k^{(n)}\big)$ is a quadratic lower bound of $\log(1+\gamma_k)$ around $\gamma_k^{(n)}$, which is derived from the Lipschitz continuity of the derivative of $\log(1+\gamma_k)$. The proof is given in Appendix 2. To obtain an MI-SOCP formulation of (4.41), we replace (4.41e) by

$$U(\gamma_k;\gamma_k^{(n)})) \geq \phi_k \quad (4.43)$$

which is SOC representable. The problem (4.41) becomes MI-SOCP program, which can be solved by SCA based algorithm outlined in Algorithm 4.1.

4.5.3 SCA-based Algorithm for WSEE Maximization Problem

In this section, we present the method to apply SCA to solve the problem WSEE maximization problem in (4.12). By introducing the similar slack variables $\psi = \{\psi_i \geq 0\}_{\forall i\in\mathscr{I}}$, $\mu = \{\mu_i \geq 0\}_{\forall i\in\mathscr{I}}$, $\varphi = \{\varphi_{i,k} \geq 0\}_{\forall i\in\mathscr{I},\forall k\in\mathscr{K}}$, $\phi = \{\phi_k \geq 0\}_{\forall k\in\mathscr{K}}$, $\gamma = \{\gamma_k \geq 0\}_{\forall k\in\mathscr{K}}$, $\theta = \{\theta_{i,k} \geq 0\}_{\forall i\in\mathscr{I},\forall k\in\mathscr{K}}$, $\xi = \{\xi_k \geq 0\}_{\forall k\in\mathscr{K}}$ and $\mathbf{z} = \{z_k \geq 0\}_{\forall k\in\mathscr{K}}$, as presented in the GEE maximization and denote $\Psi_{\text{WSEE}} = \{\mathbf{b}, \mathbf{a}, \mathbf{w}, \nu, \theta, \mu, \gamma, \phi, \mathbf{z}, \psi, \xi, \varphi\}$, we are able to equiva-

lently transform the maximization problem $(\mathscr{E}_{\text{WSEE}})$ as the following

$$\max_{\Psi_{\text{WSEE}}} \sum_i \lambda_i \psi_i \tag{4.44a}$$

$$\text{s.t.} \quad \psi_i \leq \frac{\sum_k \varphi_{i,k}^2}{\mu_i} \tag{4.44b}$$

$$P_i^{\text{rrh}}(\mathbf{w}, b_i) + \rho_i \sum_{k \in \mathscr{K}} \theta_{i,k} \leq \mu_i \tag{4.44c}$$

$$(4.12c), (4.12e), (4.12f), (4.12h), (4.13), (4.28g), (4.29), (4.30) \tag{4.44d}$$

$$(4.34d), (4.34e), (4.34f), (4.34g), (4.34h). \tag{4.44e}$$

It is noted that μ_i is the variable representing the upper bound for the ith RRH and its fronthaul power consumption and ψ_i represents the lower bound for the EE of ith RRH. We have an observation that the problem (4.44) shares the similar form with the problem (4.34). For instance, it is also interesting to show that the function $\frac{\sum_k \varphi_{i,k}^2}{\mu_i}$ in the righ hand side of constraint (4.44b) has the form of $h(p,q)$. Thus, nonconvex constraint (4.44b) can be approximated into a convex one by applying the linearization to the $h(\varphi_i, \mu_i)$ as

$$H(\varphi_i, \mu_i; \varphi_i^{(n)}, \mu_i^{(n)}) = \frac{\sum_{k \in \mathscr{K}} 2\varphi_{i,k}^{(n)} \varphi_{i,k}}{\mu_i^{(n)}} - \frac{\sum_{k \in \mathscr{K}} \varphi_{i,k}^{(n)2}}{\mu_i^{(n)2}} \mu_i \tag{4.45}$$

where we denote $\varphi_i = \{\varphi_{i,k}\}_{\forall k \in \mathscr{K}}$. Similarly, we apply the same transformations and convex approximations in the previous section to obtain the MI-SOCP problem of (4.44) at the $n+1$th iteration as

$$\max_{\Psi_{\text{WSEE}}} \sum_i \lambda_i \psi_i \tag{4.46a}$$

$$\text{s.t.} \quad \psi_i - H(\varphi_i, \mu_i; \varphi_i^{(n)}, \mu_i^{(n)}) \leq 0, \tag{4.46b}$$

$$P_i^{\text{rrh}}(\mathbf{w}, b_i) + \rho_i \sum_{k \in \mathscr{K}} \theta_{i,k} \leq \mu_i \tag{4.46c}$$

$$(4.12c), (4.12e), (4.12f), (4.12h), (4.13), (4.28g), (4.29), (4.30) \tag{4.46d}$$

$$(4.41d), (4.41f), (4.41g), (4.41h), (4.43). \tag{4.46e}$$

where $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \gamma^{(n)}, \varphi^{(n)}, \mu^{(n)}$ are the parameters that are updated at the $(n+1)$th iteration. The proposed iterative approach to solve problem (4.44) for WSEE maximization problem is given in Algorithm 4.1, where the MISOCP convex problem (4.46) is applied in the step 3.

4.5.4 SCA-based Algorithm for FEE Maximization Problem

Inspired by the SCA method for solving the problems of GEE and WSEE maximization, here we provide the main steps to tackle the FEE maximization problem. First, with the set of variables $\Psi_{\text{FEE}} = \{\mathbf{b}, \mathbf{a}, \mathbf{w}, \nu, \theta, \mu, \gamma, \phi, \mathbf{z}, \psi, \xi, \varphi\}$ where newly similar slack variables $\psi \geq 0$, $\mu = \{\mu_i \geq 0\}_{\forall i \in \mathscr{I}}$, $\varphi = \{\varphi_{i,k} \geq 0\}_{\forall i \in \mathscr{I}, \forall k \in \mathscr{K}}$, $\phi = \{\phi_i \geq 0\}_{\forall i \in \mathscr{I}}$, $\gamma = \{\gamma_k \geq 0\}_{\forall k \in \mathscr{K}}$, $\theta = \{\theta_{i,k} \geq 0\}_{\forall i \in \mathscr{I}, \forall k \in \mathscr{K}}$, $\xi = \{\xi_k \geq 0\}_{\forall k \in \mathscr{K}}$ and $\mathbf{z} = \{z_k \geq 0\}_{\forall k \in \mathscr{K}}$ are introduced, the FEE maximization problem in (4.12) can be equivalently transformed into the following problem

$$\max_{\Psi_{\text{FEE}}} \ \psi \tag{4.47a}$$

$$\text{s.t.} \ \ \psi \leq \frac{\sum_{k \in \mathscr{K}} \varphi_{i,k}^2}{\mu_i} \tag{4.47b}$$

$$P_i^{\text{rrh}}(\mathbf{w}, b_i) + \rho_i \sum_{k \in \mathscr{K}} \theta_{i,k} \leq \mu_i \tag{4.47c}$$

$$(4.12c), (4.12e), (4.12f), (4.12h), (4.13), (4.28g), (4.29), (4.30) \tag{4.47d}$$

$$(4.34d), (4.34e), (4.34f), (4.34g), (4.34h). \tag{4.47e}$$

where ψ represents for the minimum EE across all RRHs. The formulation of problem (4.47) now is exactly as same as the problem (4.44). Thus, the same transformations and approximations as in the previous sections can be applied to (4.47). As a result, the problem (4.47) can be approximated to a MISOCP problem at the $n+1$th iteration as the following

$$\max_{\Psi_{\text{FEE}}} \left\{ \psi \,\middle|\, \psi - H(\varphi_i, \mu_i; \varphi_i^{(n)}, \mu_i^{(n)}) \leq 0, (4.46c), (4.46d), (4.46e) \right\} \tag{4.48}$$

where $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \boldsymbol{\varphi}^{(n)}, \boldsymbol{\mu}^{(n)}$ are the parameters that are updated at the $(n+1)$th iteration. The proposed iterative approach to solve problem (4.47) for FEE maximization problem is given in Algorithm 4.1 where the problem (4.48) is applied to the step 3.

4.5.5 Relaxed based Algorithms

To develop an algorithm with polynomial time, we further consider the continuous relaxation of binary variables, i.e., $0 \leq b_i \leq 1, 0 \leq a_{i,k} \leq 1$ for $\forall i \in \mathscr{I}, \forall k \in \mathscr{K}$. As a result, the continuous relaxation of (4.41), (4.46), and (4.48) denoted as $(\mathscr{E}^r_{\mathrm{X}})$, becomes an SOCP which can be solved in polynomial time, with X representing for GEE, WSEE, and FEE, respectively. The relaxed based algorithm generally combines two stages: (i) continuous relaxation and (ii) post-processing. In the first stage, we follow Algorithm 4.1, but simply solve $(\mathscr{E}^r_{\mathrm{X}})$ in Step 3. The post-processing process is then used to map the obtained b_i's and $a_{i,k}$'s to the binary values, which is required due to the continuous relaxation. In particular, we rely on the solution to the continuous relaxation at convergence as an incentive measure to make a decision on the binary value of $\mathbf{a}$ and $\mathbf{b}$. Let us denote $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}$ and $\tilde{\mathbf{w}}$ as the solution achieved after the first stage. Intuitively, the connection between the ith RRH and the kth UE is more likely if the channel of the link is in better condition and the power consumed to transmit fronthaul data $P_i^{\mathrm{FH}}(\mathbf{w}, \mathbf{a}_i)$ is smaller than the others. Consequently, solving the continuous relaxation would possibly yield higher $\widetilde{b}_i$ for the ith RRH and higher $\widetilde{a}_{i,k}$ for the connection between the ith RRH and the kth UE. Based on the above intuitive observations, we propose an iterative procedure to determine the set of active RRHs and RRH-UE association based on $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$. The process starts by assuming that all the RRHs are off and there is no association between RRH and UE. In each iteration, $(\mathscr{E}^r_{\mathrm{X}})$ is solved given a set of active RRHs and RRH-UE association that is connected. The RRH-UE association with the largest $\tilde{a}_{i,k}$ will be made connected and the resulting RRH will be set active, following the relationship in (4.12f). The overall algorithm is presented in Algorithm 4.2.

Algorithm 4.2: Relaxed algorithm

1: Set $m := 0$, $\boldsymbol{\pi}^{(m)}$ is significantly small, and initialize the set $\mathscr{R}_{\text{off}}^{(m)} = \{(i,k) \times i \in (\mathscr{I}, \mathscr{K}) \times \mathscr{I}\}$.
2: **repeat**
3: Set $m := m + 1$;
4: Algorithm 4.1 is used to solve $(\mathscr{E}_{\text{X}}^{r})$ given $a_{i',k'} = 1$ and $b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathscr{R}_{\text{off}}^{(m-1)}$;
5: Update $\mathscr{R}_{\text{off}}^{(m)} = \mathscr{R}_{\text{off}}^{(m-1)} \setminus \left\{ (i',k') \times i' = \arg\max_{i,k \in \mathscr{R}_{\text{off}}^{(m-1)}} \widetilde{a}_{i,k} \right\}$;
6: Algorithm 4.1 is used to solve $(\mathscr{E}_{\text{X}}^{r})$ given $a_{i',k'} = 1, b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathscr{R}_{\text{off}}^{(m)}$ and $a_{i,k} = 0, b_i = 0, \forall \{(i,k) \times i\} \in \mathscr{R}_{\text{off}}^{(m)}$. If it is feasible, set $\boldsymbol{\pi}^{(m)}$as the value of objective function achieved at the convergence. If not, set $\boldsymbol{\pi}^{(m)} = \boldsymbol{\pi}^{(0)}$.
7: **until** $(\mathscr{E}_{\text{X}}^{r})$ starts to be infeasible or it is feasible but $\boldsymbol{\pi}^{(m)} < \boldsymbol{\pi}^{(m-1)}$;
8: Algorithm 4.1 is used to solve $(\mathscr{E}_{\text{X}}^{r})$ given $a_{i',k'} = 1, b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathscr{R}_{\text{off}}^{(m-1)}$ and $a_{i,k} = 0, b_i = 0, \forall \{(i,k) \times i\} \in \mathscr{R}_{\text{off}}^{(m-1)}$ to obtain final solution $\Psi_{\text{X}}^{\star}$;

4.5.6 Convergence and Complexity Analysis

The convergence of SCA-based algorithms has been well studied Marks & Wright (1977). In particular, due to the use of convex approximations, the optimal solution obtained at iteration n is feasible to the convex problems (4.41), (4.46), and (4.48) at iteration $n+1$ for GEE, WSEE, and FEE maximization problems, respectively. This results in a non-decreasing sequences of objectives. Since the objective is upper bounded due to the power budget, the iterative algorithms are probably convergent. We now discuss the complexity of the proposed algorithms in this section. For Algorithm 4.1, the overall complexity mainly depends on that of solving the MI-SOCP problem in (4.41), (4.46), and (4.48), respectively, which is indeed a combinatorial optimization problem. In particular, there are IK binary variables $a_{i,k}$'s and I binary variables b_i's, resulting in 2^{IK+I} combinations for all the binary variables. Thus, the worst-case complexity of Algorithm 4.1 in each iteration can be written as $\mathscr{O}\left(2^{IK+K}(K^4M^3I)\right)$. For Algorithm 4.2, first we remark that in the worst case, Algorithm 4.2 must iteratively solve and update the resulting parameters for the SOCP problem $(\mathscr{E}_{\text{X}}^{r})$ for $(I-1)K$ times, resulting the overall complexity of $\mathscr{O}\left(2(I-1)K\left(K^4M^3I\right)\right)$. In Section 4.6, Algorithm 4.2 is shown to yield

a performance very close to that of SCA based algorithms but with polynomial computation time.

4.6 Numerical Results

In this section, the extensive numerical results are presented to evaluate the performance of the proposed algorithms. For most numerical experiments, we use the simulation parameters listed in Table 4.1 (Cheng *et al.*, 2013; Dai & Yu, 2016). For the spatial model, we assume a network consisting of I RRHs that are uniformly located around the considered coverage and K UEs are randomly scattered across the considered network coverage. Moreover, we assume shadowing channel and the path-loss component is calculated as $(d_{ik}/d_0)^{-3}$where d_{ik} is the distance between the ith RRH and the kth user and $d_0 = 100$ m is the reference distance. In our simulations, Algorithms 4.1 is terminated when the increase in the objective between two consecutive iterations is less than 10^{-5}.

Table 4.1 Simulation parameters in Chapter 4

Notation	Value	Notation	Value
M_i	2	η_i	0.35
$P^{\max}$	10 dBW	$C_i^{\max} = C, \forall i$	20 b/s/Hz
P_i^{ra}	38.5 dBW	P_i^{ri}	36.5 dBW
$\rho_i, \forall i$	1	λ	$[1/I, \ldots, 1/I]$

4.6.1 Convergence and achieved EE performance

Fig. 4.2, Fig. 4.3 and Fig. 4.4 show the convergence of lower and upper bound of BRB algorithms and proposed low-complexity algorithms for GEE, WSEE and FEE maximization problems, respectively. In general for all cases of GEE, WSEE and FEE maximization problems, it can be seen that the lower and upper bound of BRB algorithm require more than 10^4 iterations to converge at the optimal solution while the SCA based algorithms need few iterations to converge to the objective value that is very close to the optimal value returned by the optimal

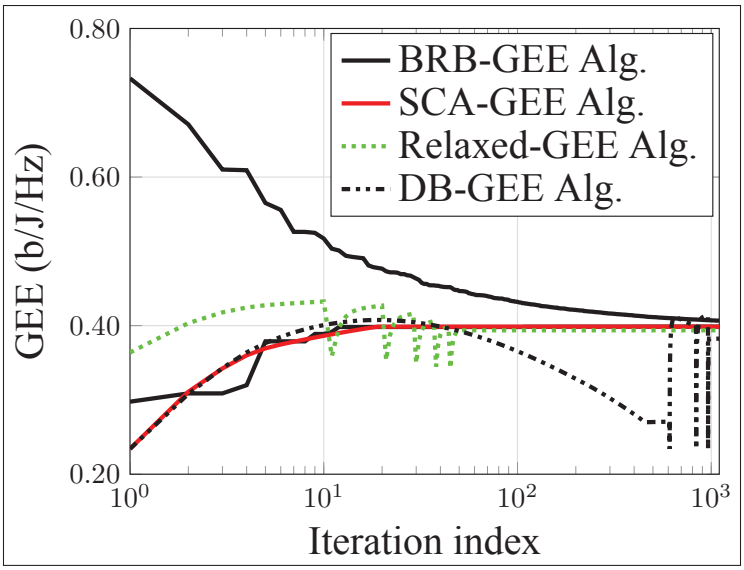


Figure 4.2 Convergent behavior of different algorithms for GEE maximization problems.

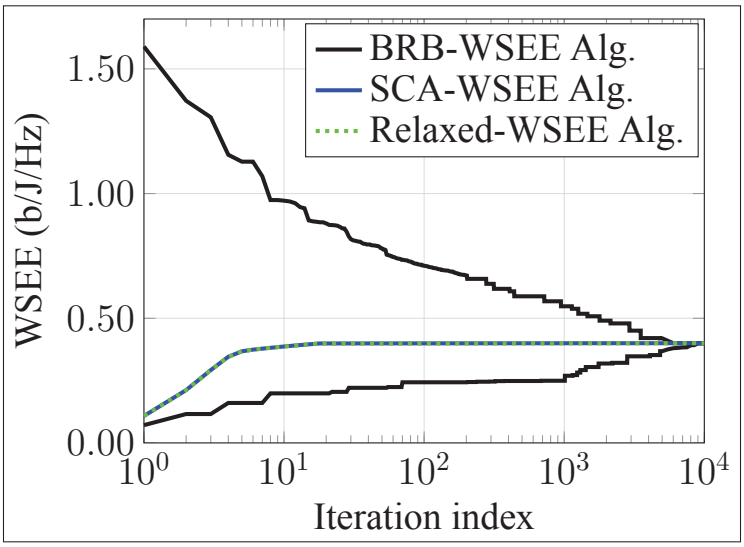


Figure 4.3 Convergent behavior of different algorithms for WSEE maximization problems.

BRB algorithm. In Fig. 4.2, the convergence of each SOCP $\mathscr{E}^{r}_{\mathrm{GEE}}$ during the relaxed processes is plotted for the relaxed-GEE algorithm, which illustrates the uphill and downhill effect in the figure. For example, at some first iterations, the GEE objective achieved in the relaxed-GEE algorithm is higher than in BRB and SCA based algorithms because of the relaxation continuous

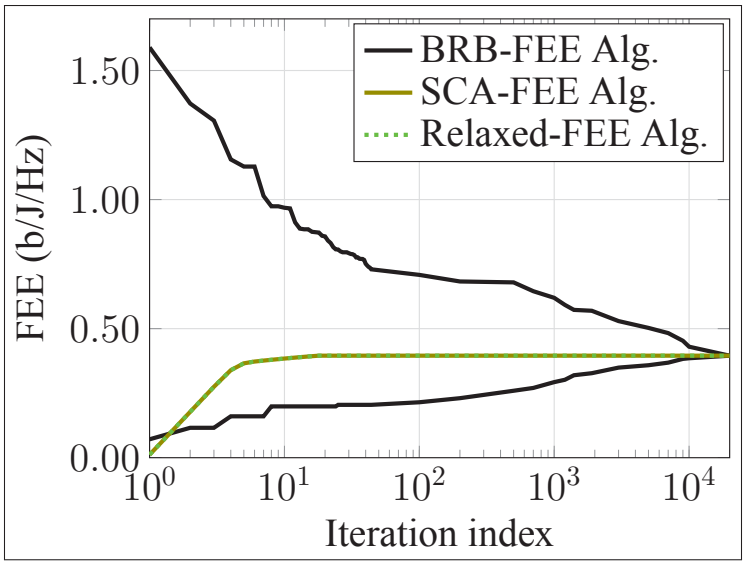


Figure 4.4 Convergent behavior of different algorithms for FEE maximization problems.

of binary variables and as expected, it eventually converges to the near-optimal objective value after all relaxed binary variables are mapped into feasible binary values. In addition, to demonstrate the performance gains of our proposed GEE algorithms, we compare our GEE algorithms with the Dinklebatch method, named DB-GEE algorithm. Especially, in DB-GEE algorithm, Dinklebatch approach is used to transform the fractional GEE objective function in (4.8) into the subtraction form associated with a fixed parameter, then the SCA method is applied to solve this subtraction formulation problem and the parameter in the objective can be updated until convergence. It is shown that DB-GEE algorithm requires more iterations to stabilize. In addition, in Fig. 4.3 and Fig. 4.4, the performance of relaxed-WSEE and relaxed-FEE algorithms are exactly same as that of SCA-WSEE and SCA-FEE algorithms, respectively. This can be explained as solving the relaxed problem $\mathscr{E}^r_{\text{WSEE}}$ and $\mathscr{E}^r_{\text{FEE}}$ at the first stage of corresponding relaxed algorithms result the binary values of all relaxed variables. Thus, Fig. 4.2, Fig. 4.3 and Fig. 4.4 prove the effectiveness of our proposed algorithms in both terms of convergence and achieved EE performance.

Fig. 4.5, Fig. 4.6 and Fig. 4.7 evaluate the performance of SCA-based algorithms with respect to achieved GEE, WSEE and FEE metrics in (4.8), (4.10) and (4.11), respectively, versus the maximum fronthaul capacity $C_i^{\max} = C^{\max}$, for $\forall i \in \mathscr{I}$. Our first observation is that the

achieved GEE, WSEE and FEE performance increases in the low regime of $C^{\max}$ and becomes saturated in the large regime of $C^{\max}$. The reason is that the multi-user interference always exists even as more cooperation can be attained among all RRHs. For interference limited situations, there is an upper bound on the achievable rate for all users so that increasing more fronthaul capacity basically provides no benefit to the system performance. It is obvious that the resource allocation solution obtained from GEE maximization algorithm results the best GEE performance than the WSEE and FEE maximization algorithms in Fig. 4.5. Similarly, WSEE maximization algorithm outperforms the GEE and FEE maximization algorithms in term of the achieved WSEE performance in Fig. 4.6. Meanwhile, FEE maximization algorithm achieves the best minimum EE value compared to the GEE and WSEE maximization algorithms in Fig. 4.7, but incur the small loss of GEE and WSEE performance shown in Fig. 4.5 and Fig. 4.6. These observations imply that the FEE metric yields the better minimum EE than GEE and WSEE criteria but still achieves the good performance in terms of GEE and WSEE.

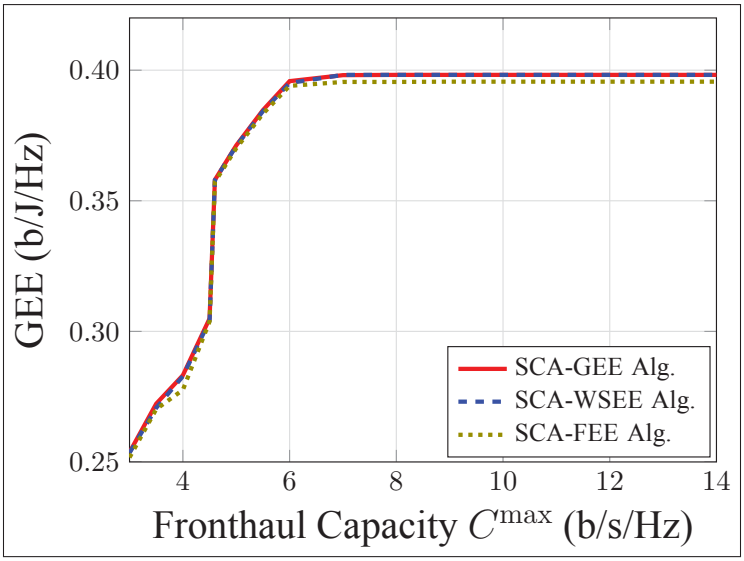


Figure 4.5 GEE objective in (4.8) calculated from the solutions obtained by applying the different algorithms versus $C^{\max}$.

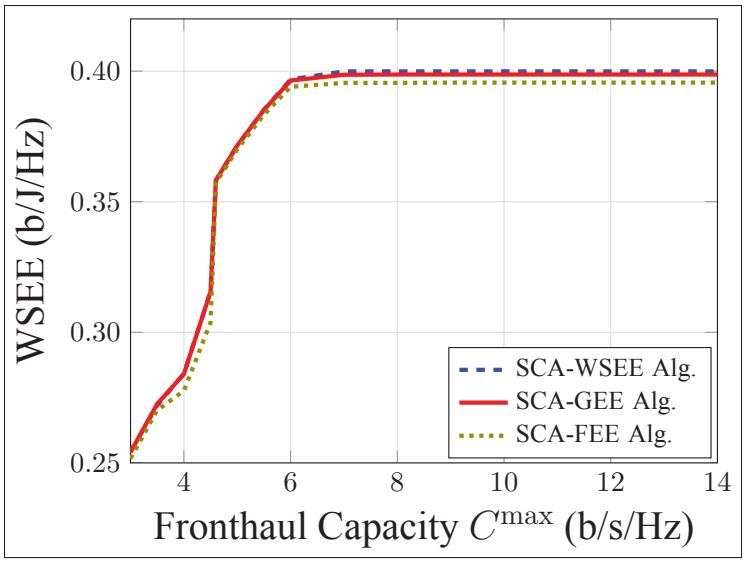


Figure 4.6 WSEE objective in (4.10) calculated from the solutions obtained by applying the different algorithms versus $C^{\max}$.

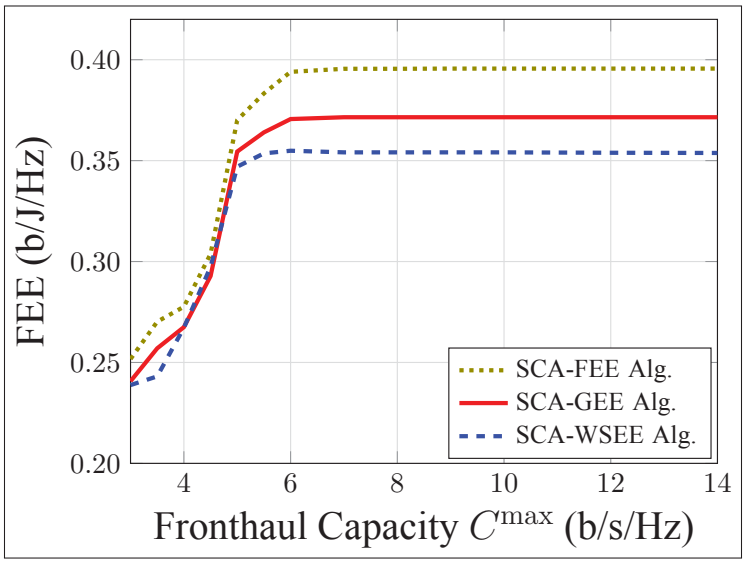


Figure 4.7 FEE objective in (4.11) calculated from the solutions obtained by applying the different algorithms versus $C^{\max}$.

4.6.2 Advantages of proposed rate-dependent power model

In this numerical result, we demonstrate the advantages of the GEE, WSEE and FEE performance achieved by our proposed rate dependent power consumption model compared to the

linear power consumption model used mostly in the recent literature. Instead of using our proposed rate dependent fronthaul power model in (4.6), the linear fronthaul power consumption model (LFP) is expressed as

$$P_i^{\mathrm{FH}}(\mathbf{a}_i) = \sum_{k \in \mathscr{K}} a_{i,k} P_{i,k}^{\mathrm{fh}} \tag{4.49}$$

where $P_{i,k}^{\mathrm{fh}}$ is the fixed power consumption, i.e., $P_{i,k}^{\mathrm{fh}} = 2$ Watts Tang *et al.* (2017); Guo *et al.* (2016b), used for data transmission between the kth user and the ith fronthaul. Then our proposed SCA-based algorithm is employed to optimize the GEE, WSEE and FEE metrics with the linear power consumption model, called GEE-LFP, WSEE-LFP and FEE-LFP algorithms, respectively.

Fig. 4.8 and 4.9 plot the GEE and WSEE objective values, respectively, achieved from the solutions of the GEE, WSEE, FEE, GEE-LFP, WSEE-LFP and FEE-LFP maximization algorithms. As expected, achieved GEE and WSEE performance increase and reach a plateau along with the growth of $P^{\max}$. The fact is that the sum achieved rate of users first increases due to the increase of transmit power consumption, which leads to an increase in the achieved GEE and WSEE performance. However, when power budget becomes sufficient large, the gain of sum achieved data rate can not compensate for the quick increase in the power consumption. Thus, GEE and WSEE maximization algorithms will not use the excess transmit power to further increase the rate to maintain the high values of GEE and WSEE. Another important observation is that our proposed rate dependent power consumption model outperforms the linear power consumption model. This is obvious as the linear power model consumes the fixed power consumption for each active transmission between user and fronthaul, which is not precise in practice.

Fig. 4.10 shows the FEE objective values in (4.11) calculated from the solutions obtained by applying different algorithms versus $P^{\max}$. Clearly, FEE maximization algorithm that focuses on maximizing the minimum EE, achieves the best FEE objective values than GEE and WSEE maximization algorithms. Moreover, it can be seen that for GEE and WSEE maximization algorithms, FEE value increases and then slightly decreases and eventually get saturated when

$P^{\max}$ increases. This can be explained similarly to the explanation of Fig. 4.8 and Fig. 4.9 in the previous paragraph. The achieved GEE and WSEE first increase with the increase of $P^{\max}$ because of the increase of users' data rates. However, at a certain point of $P^{\max}$, the increase of sum achieved data rate gain lead to larger transmit beamforming power forced to increase at the RRH which has the worst channel conditions to its served UEs. This results in the decrease of EE at this RRH. Furthermore, at the sufficient large $P^{\max}$ increasing more transmit power provides no benefit to the system throughput, thus GEE and WSEE algorithms will not continue to increase the transmit beamforming at the RRHs to maintain the still good achieved EE performance. Last but not least, the performance in term of FEE achieved from our proposed power model is larger than that of linear power model. This not only proves the benefits of our proposed rate dependent power consumption model, but also emphasizes the important in correctly characterizing the power consumption in C-RAN.

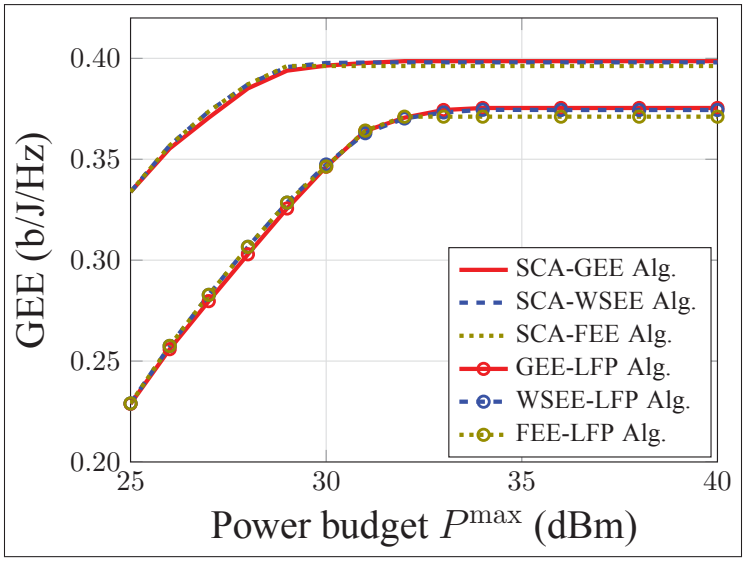


Figure 4.8 The comparison of GEE performance between our proposed rate dependent power consumption model and the linear fronthaul power consumption model versus $P^{\max}$.

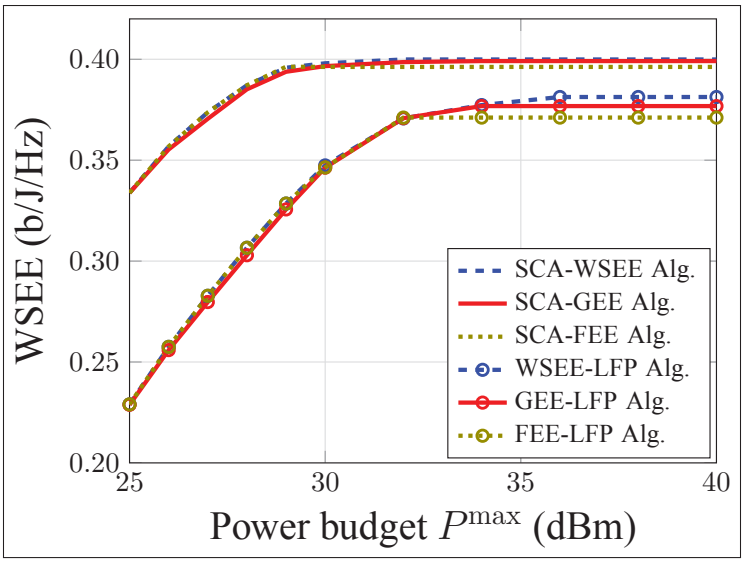


Figure 4.9 The comparison of WSEE performance between our proposed rate dependent power consumption model and the linear fronthaul power consumption model versus $P^{\max}$.

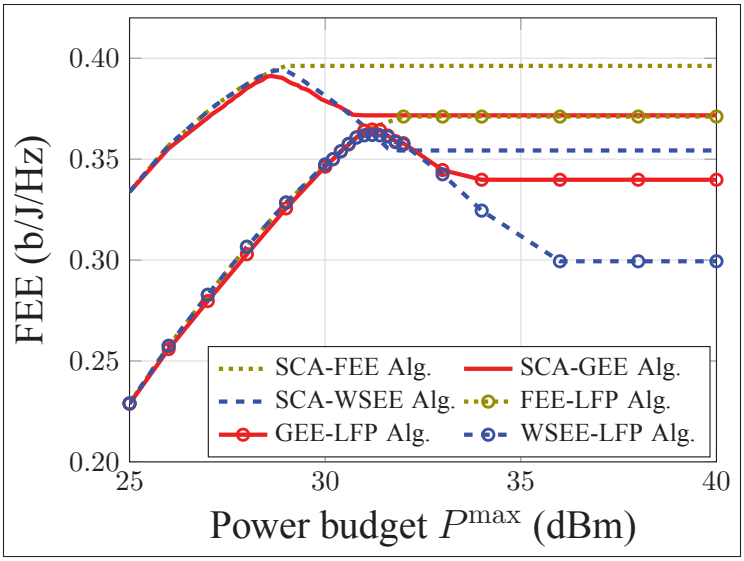


Figure 4.10 The comparison of FEE performance between our proposed rate dependent power consumption model and the linear fronthaul power consumption model versus $P^{\max}$.

4.7 Concluding Remarks

We investigated energy efficient resource allocation in the downlink of a limited fronthaul capacity C-RAN with the constraints where the rate-dependent power model was proposed. Three different EE metrics, namely GEE, WSEE of all RRHs, and FEE across all RRHs were considered. In this work, we customize the BRB algorithm to find the globally optimal solutions for all the visited problems. For the low-complexity solution approach, we presented some novel techniques to transform the combinatorial and non-convex EE problems into more tractable forms. Further, the unified framework based on the SCA method and the relaxation method was developed to approximate the EE problems into a sequence of SOCP problems. Then, a low-complexity algorithm based on SCA and relaxation framework was conducted to iteratively compute the resource allocation solution at convergence. Numerical results show that our proposed SCA- and relaxed-based algorithm significantly outperform all the existing methods in terms of convergent speed and can achieve near-to-optimal compared to the BRB algorithm. It was shown that WSEE provided more freedoms on the individual EE of each RRH for resource allocation design compared to popular GEE metric. Also, FEE yielded the best balanced EE performance on system design compared to others. Additionally, we have numerically illustrated the significance of our proposed rate-dependent power model in achieve higher EE compared to the existing power model in the literature.

CHAPTER 5

JOINT VIRTUAL COMPUTING AND RADIO RESOURCE ALLOCATION IN LIMITED FRONTHAUL GREEN C-RANS

Phuong Luong [1], François Gagnon[1], Charles Despins[1], Le-Nam Tran[2]

[1] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3,
[2] School of Electronical and Electronic Engineering, University College Dublin
4 Dublin, Ireland

This article was published in *IEEE Transaction on Wireless Communications* in April 2018 (Luong *et al.*, 2018b).

5.1 Introduction

The development of the next generation wireless network, commonly referred to as 5G, is underway. In general, 5G is expected to meet a set of challenging requirements that can solve many problems in existing wireless systems (Andrews *et al.*, 2014). Some key technologies for 5G are introduced in (Wong *et al.*, 2017). From a network architecture viewpoint, cloud-radio access networks (C-RANs) have received growing attention as a powerful candidate to implement 5G standards. In particular, C-RANs can significantly enhance both system spectral and energy efficiency (EE), and satisfy other quality-of-service (QoS) requirements. In C-RANs, signals transmitted or received by low-power remote radio heads (RRHs) are processed by the centralized baseband unit (BBU) pool, comprising several physical servers (PSs) on a cloud-computing platform (Rost *et al.*, 2014; Andrews *et al.*, 2014). A RRH is typically simplified with only radio frequency (RF) functions to handle the transmission/reception of radio signals to/from the users. On the other hand, sophisticated baseband signal processing tasks are migrated to the BBU pool, i.e., on the cloud. In this way, the virtualization and network slicing technology can be used to deliver dynamic and powerful resource allocation. As a main feature of cloud-computing, virtualization at the BBU pool enables each PS to dynamically split the dedicated computing resources into various virtual machines (VMs) sizes,

depending on data traffic. Some noticeable results in this regard were reported in Simeone *et al.* (2016).

Although the benefits of C-RAN technology, with the use of distributed RRHs, are relatively convincing from a viewpoint of centralized resource management, it still raises a serious concern over the energy power consumption. In fact, there are many power consumption sources in a C-RAN, ranging from the circuit powers and RF transmission power to the fronthaul data transportation and processing, fronthaul maintenance power, computational power from the cloud, etc. In light of this, designing greener resource allocation which maximizes the EE of C-RANs has become an essential criteria for viable C-RAN solutions. A variety of our studies on the energy efficiency maximization, the power consumption minimization as well as the maximization of the trade-off between total power consumption and throughput in the C-RAN by jointly optimizing the radio resource allocation given a fixed setting of cloud computing capability have been carried out in (Luong *et al.*, 2017a, 2016a, 2017b,c). Despite many potential benefits, the understanding on energy-efficiency of virtualized C-RANs with limited-capacity fronthaul is far from comprehensive, this paper should contribute to thoroughly clarify the issue. From the virtualization standpoint, an energy-efficient design must be able to adapt computing resources to elastic traffic (Pompili *et al.*, 2016; Saxena *et al.*, 2016; Tang *et al.*, 2015). From the energy-efficiency maximization perspective, it should maximize the system spectral efficiency while consuming the least power. Thus, PSs on the cloud can be activated wisely, otherwise the power consumption for signal processing taking place in the BBU pool dominates other power consumption sources in the network, leading to a poor energy-efficiency performance (Lin *et al.*, 2011). More explicitly, some PSs should be switched OFF to save power while others must be active to maintain the system quality-of-service (QoS). Moreover, the number of RRHs and its associated fronthaul links can be very large in a dense C-RAN, generating a huge pressure on the total network power consumption. As a result, designing a green C-RAN must also consider the RRH selection together with the RRH-user association problem, concerning the limited-capacity fronthaul (Peng *et al.*, 2015). These motivations call for new radio resource management methods to design energy-efficient C-RANs.

Energy-efficient approaches for C-RANs have been presented in the recent literature, especially for joint design of RRH selection and RRH clustering (Zhuang *et al.*, 2016; Zhao & Wang, 2016; Shi *et al.*, 2016b; Niu *et al.*, 2016; Luo *et al.*, 2015; Shi *et al.*, 2014; Peng *et al.*, 2016; Saxena *et al.*, 2016). Specifically, to minimize network power consumption, the authors in (Zhuang *et al.*, 2016) considered a joint design of cell activation and spectrum allocation, using the reweighted ℓ_1-norm approximation technique. Similarly, a smoothed ℓ_p-norm minimization method was involved to deal with the RRH selection and user admission problem in a multicast C-RAN (Shi *et al.*, 2016b). The work of in (Luo *et al.*, 2015) studied joint optimization of beamforming and user association for green C-RANs. In (Li *et al.*, 2015a), stochastic optimization was applied to solve the problem of queue-aware joint RRH selection and beamforming to minimize the network power consumption. Queue-aware energy efficiency maximization was also studied for heterogeneous C-RAN in (Peng *et al.*, 2016). A stochastic game approach was proposed in (Saxena *et al.*, 2016) to allow virtual base stations to learn the cellular traffic variation and thus enable to switch OFF some RRHs to reduce the overall energy consumption in C-RANs. Unlike these, a recent and pragmatic technology which deploys a cached server at each base station to alleviate the fronthaul congestion and improve the system energy efficiency was considered in (Liu & Yang, 2016). Given a cache strategy, the authors developed a framework to analyze and derive the EE closed-form expression on top of achieving an optimal cache policy which maximizes the network EE (Liu & Yang, 2016). The energy-efficient resource allocation considering limited fronthaul capacity was studied in (Wang *et al.*, 2017; Ng *et al.*, 2012). Specifically, in (Wang *et al.*, 2017), the authors aimed to maximize the sigmoidal function of user's SINR under imperfect CSI condition and limited backhaul capacity. In (Ng *et al.*, 2012), the authors studied an EE resource allocation in multi-cell limited backhaul OFDMA downlink networks, where zero-forcing beamforming and semi-orthogonal user selection policies were employed prior to the EE problem formulation. The trade-off between EE and spectral efficiency was investigated in (Wu *et al.*, 2014). There also exists the trade-off between the EE and the system delay as shown in (Li *et al.*, 2015b).

In recent years, virtualization for wireless communications has proved to be a powerful technology to fully and flexibly exploit the computing resources of C-RANs (Simeone *et al.*, 2016). In (Tang *et al.*, 2014), the authors considered a framework of dynamic request allocation of VMs to minimize the VM computing cost. A heuristic algorithm was developed in (Xu & Fortes, 2010) to design the VM placement and network element switching ON/OFF to conserve more computing power. In (Nejad *et al.*, 2015), an integer linear program was formulated for the problem of dynamic VM provisioning and allocation. For a joint design of radio resource management and network virtualization for C-RANs, the work of (Pompili *et al.*, 2016) proposed a virtual base station (VBS) cluster associated with a subset of RRHs which can adapt to traffic dynamics. Likewise, the authors in (Wang *et al.*, 2016b) leveraged the VBS concept and optimized VBS formation using an mixed-integer linear program. A joint design of VM computation capacity, RRH selection, and beamforming to minimize the total power consumption in C-RANs was proposed in (Tang *et al.*, 2015; Guo *et al.*, 2016b). Similarly, the authors in (Wang *et al.*, 2016a) considered the joint computing capacity and beamforming design for energy minimization problem in a mobile cloud computing network. Throughput maximization in C-RAN was studied in (Tran & Pompili, 2017b), taking into account the constraint of computing resource capacity of VBS pool.

In this paper we study an energy-efficient design of a virtualized C-RAN with limited-capacity fronthaul. Unlike (Wang *et al.*, 2017; Ng *et al.*, 2012), where the virtual computing resource at the central unit or BBU pool is not accounted for, we consider a joint optimization of transmit beamforming, virtual computing resource allocation, RRH selection, and the RRH-user association to maximize the global network energy efficiency. Compared to the recent literature (Pompili *et al.*, 2016; Tang *et al.*, 2015; Guo *et al.*, 2016b; Wang *et al.*, 2016a; Saxena *et al.*, 2016), we first propose a novel virtual computing resource allocation (VCRA) scheme. In particular, to best exploit the VMs in each PS, the proposed VCRA method splits the users' workload into smaller pieces that can be served by different VMs in parallel. The distinguishing features of the proposed VCRA scheme are as follows: (i) data traffic of a user can be processed by heterogeneous VMs; (ii) VM's computing capacity is dynamically allocated ac-

cording to the traffic condition; (iii) the assignment of VMs to PSs is done in such a way that the number of unused PSs is maximal. Furthermore, to quantify the power consumption more accurately, we introduce a new power consumption model which includes the rate dependent fronthaul power consumption. More specifically, the fronthaul power consumption model is computed based on the total transmission rates served by the corresponding RRH, which is more realistic and different from that in (Luo *et al.*, 2015; Shi *et al.*, 2016b; Guo *et al.*, 2016b), where fronthaul power is a quadratic or linear function of involved variables.

We formulate the problem as a mixed-integer non-convex program, for which is generally difficult to find an optimal solution. Even if possible, the complexity is prohibitively high since a mixed-integer program is commonly known to be NP-hard. Solving the problem is far more challenging for several reasons: (i) the non-convexity of the cost function, (ii) the non-convexity of the limited-capacity fronthaul and the cross-layer delay constraints, and (iii) the combinatorial nature of the selection and assignment procedure. The said non-convexity actually implies that the continuous relaxation of the problem is non-convex. In fact, this attribute makes known mixed-integer optimization solvers of no use to solve the considered problem, which motivates us to develop an optimal algorithm. To this end we first propose multiple novel transformations to reformulate the original problem into a form that amenable to monotonic optimization (Tervo *et al.*, 2015). For a more practically appealing method, we develop a low complexity algorithm based on difference of convex algorithm (DCA) (Pham & Thi, 2014), approximating the original problem by a series of convex quadratic programs, using Lipschitz continuity (Parikh & Boyd, 2014). Our contributions are the following:

- We propose a novel VCRA strategy at the BBU pool and consider a joint optimization problem of the beamforming, VCRA, RRH selection and RRH-user association. To formulate the problem of interest we introduce several binary preference variables. The objective is to maximize the overall network energy efficiency under explicitly limited-capacity fronthaul constraints. We customize a branch-and-reduce-and-bound (BnRnB) algorithm to compute a globally optimal solution to the formulated problem, which is a mixed-integer non-convex program.

- To find a high-quality low-complexity solution, we first deal with the continuous relaxation of the problem and then propose a post-processing procedure to recover binary variables. This way seems to be a standard mixed-integer programming but, as mentioned earlier, the relaxed problem is non-convex and still difficult to solve. Our aim is to solve the relaxed problem using a local optimization method called DCA (Pham & Thi, 2014), which has been shown to be very effective in may applications. The proposed DCA method can cope with the non-convex limited fronthaul constraints without assigning a fixed rate as done in (Wang *et al.*, 2017) or employing zero forcing beamforning and user selection as in (Ng *et al.*, 2012) before solving the optimization problem. To this end, we invoke the concept of Lipschitz continuity to rewrite the relaxed problem as a DC program. This reformation offers two benefits. First, no slack variable is introduced in the DC program, which is different from previous publications in the similar context (Luong *et al.*, 2017a; Nguyen *et al.*, 2014b). Second, the resulting DC program can be easily approximated by a sequence of convex quadratic optimization problems using DCA. Finally, a post-processing algorithm is then carried out to search for a high-performance binary solution.

- Extensive numerical results are presented to show the efficiency of our proposed algorithms in terms of the convergent rate and achievable energy efficiency performance, compared to other existing methods. In particular, the numerical results also demonstrate that the proposed VCRA scheme significantly outperform the known methods.

The rest of the paper is organized as follows. Section 5.2 introduces the system model. Section 5.3 formulates the joint design of the energy efficiency maximization problem and Section 5.4 presents a global optimization algorithm. In Section 5.5, we propose a low-complexity algorithm to find a high-quality feasible solution. Section 5.6 presents numerical results and insight discussions under different simulation setups. Finally, the concluding remarks of this chapter is given in Section 5.7.

5.2 System Model

5.2.1 Transmission Model

We consider the downlink of a C-RAN consisting of I RRHs and K single-antenna user equipments (UEs). We denote by $\mathscr{I} = \{1,\ldots,I\}$ and $\mathscr{K} = \{1,\ldots,K\}$ the set of RRHs and UEs, respectively. The ith RRH is equipped with M_i antennas,$\forall i \in \mathscr{I}$. As shown in Fig. 5.1, we assume that all the RRHs are connected to BBU pool via the fronthaul links, e.g., high-speed optical ones, where the ith link has a maximum capacity C_i^{FH}. Each UE is served by a specific group of RRHs but one RRH can serve more than one UEs simultaneously. Let s_k be the signal with unit power, i.e., $\mathbb{E}\left\{s_k s_k^*\right\} = 1$, intended for the kth UE and $\mathbf{w}_{i,k} \in \mathbb{C}^{M_i \times 1}$ be the beamforming vector from the ith RRH to the kth UE. The vector of channel coefficients from the ith RRH to the kth UE is represented by $\mathbf{h}_{i,k} \in \mathbb{C}^{M_i \times 1}$. In this work, we assume perfect channel state information (CSI) between the RRHs and the UEs.[1] For notational convenience, we denote the set of beamforming vectors intended for the kth UE as $\mathbf{w}_k \triangleq [\mathbf{w}_{1,k}^T, \mathbf{w}_{2,k}^T, \ldots, \mathbf{w}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, and the vector including the channels from all RRHs to the kth UE as $\mathbf{h}_k \triangleq [\mathbf{h}_{1,k}^T, \mathbf{h}_{2,k}^T, \ldots, \mathbf{h}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, where $M = \sum_{i \in \mathscr{I}} M_i$. Using these notations, the received signal at the kth UE is given by

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \sum\nolimits_{j \in \mathscr{K} \setminus k} \mathbf{h}_k^H \mathbf{w}_j s_j + z_k \tag{5.1}$$

where $z_k \sim \mathscr{CN}(0, \sigma_0^2)$ is the additive white Gaussian noise (AWGN) and σ_0^2 is the noise power. We normalize the noise power factor to 1 for the sake of notational simplicity in the rest of the paper. Note that in (5.1), we have assumed that the kth UE is connected to all the RRHs, but the ith RRH serves the kth UE only if $\left\|\mathbf{w}_{i,k}\right\|_2^2 > 0$. By treating interference as noise, the achievable rate in b/s/Hz for a given set of channel realizations at the kth UE is given by

$$R_k(\mathbf{w}) = \log_2\left(1 + \Gamma_k(\mathbf{w})\right), \tag{5.2}$$

[1] In practice, CSI between RRHs and UEs is estimated by exploiting the channel reciprocity between the UL and DL transmissions in the time division duplexing system or by the feedback channels in the frequency division duplexing system.

where

$$\Gamma_k(\mathbf{w}) = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j \in \mathscr{K} \setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2} \tag{5.3}$$

where $\mathbf{w} \triangleq [\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_k^T]^T \in \mathbb{C}^{(KM)\times 1}$ is vector stacking the beamformers for all UEs.

In C-RAN systems, CSI between RRHs and UEs is exchanged between RRHs and the BBU pool via the fronthaul links. The cost of the incurred overhead scales with the amount of CSI fed back to the cloud via fronthaul links. Thus, a certain portion of fronthaul capacity is reserved for this overhead information transportation to and from the BBU pool. This reduces each effective fronthaul capacity budget C_i^{FH} for the data transmission, which may degrade the overall network performance in terms of network throughput and EE. For simplicity, we assume that this overhead fronthaul capacity reservation is done prior to our problem formulation. Each limited fronthaul capacity budget C_i^{FH} is now reserved for the data transportation. The benefit of RRHs coordination is limited by the overhead of pilot-assisted channel estimation (Fan *et al.*, 2016). Estimating a subset of channel coefficients, rather than all the channel coefficients from all UEs can simply reduce the overhead of CSI and signaling, but restricts the cooperation within a limited number of RRHs, resulting in the loss of system throughput. We note that data for the kth UE is routed from the BBU pool to the ith RRH via the ith fronthaul link only if $\left\|\mathbf{w}_{i,k}\right\|_2^2 > 0$. Let binary variables $a_{i,k} \in \{0,1\}, \forall i \in \mathscr{I}$ and $k \in \mathscr{K}$ represent the association status between the ith RRH and the kth UE, i.e., $a_{i,k} = 1$ implies that the kth UE is served by the ith RRH and $a_{i,k} = 0$, otherwise. Then, the per-fronthaul capacity constraints can be

$$\sum_{k \in \mathscr{K}} a_{i,k} R_k(\mathbf{w}) \leq C_i^{\mathrm{FH}}, \forall i \in \mathscr{I}. \tag{C1}$$

5.2.2 Proposed Virtual Machine Computing Model

We consider a BBU pool consisting a set of $\mathscr{S} = \{1, \ldots, S\}$ physical servers (PSs). The proposed VCRA scheme is described as follows. Assume that each PS is capable of creating multiple virtual machines (VMs) to process the incoming packets in parallel. Unlike the work in (Guo *et al.*, 2016b), we consider a VM assignment in which one VM can only process

the packets to one user but one user's packets can be served by several VMs with different computing capacities. To model this assignment scheme, we introduce the binary variables $c_{s,k} \in \{0,1\}, \forall k \in \mathscr{K}$ and $\forall s \in \mathscr{S}$, where $c_{s,k} = 1$ states that the packets of the kth user are processed by a VM in the sth PS and $c_{s,k} = 0$, otherwise. In addition, let binary variable $d_s \in \{0,1\}$ and $\mathbf{d} = \{d_s, \forall s \in \mathscr{S}\}$ denote the operation mode of the sth PS, where $d_s = 0$ means the sth PS is turned off and $d_s = 1$ otherwise.

5.2.3 Processing Queue Model

The packet arrival of the kth UE is assumed to follow a Poisson process with arrival rate Λ_k. For simplicity, we assume each packet has identical length. As illustrated in Fig. 5.1, packets of the kth UE first arrive at the dispatcher and are subsequently split into smaller fragments that are then routed to VMs in different PSs for parallel processing. It is worth mentioning that each small fragment from the kth UE's packets assigned to the VM in the sth PS also follows a Poisson process with arrival rate $\lambda_{s,k}$, where we have

$$\begin{cases} \sum_{s \in \mathscr{S}} \lambda_{s,k} = \Lambda_k \\ \lambda_{s,k} \leq c_{s,k} \Lambda_k \end{cases} , \forall k \in \mathscr{K}, s \in \mathscr{S} \tag{C2}$$

We assume that the baseband processing of each VM on each UE packets can be described as a $M/M/1$ processing queue, where the service time at the VM of the sth PS follows an exponential distribution with mean $1/\mu_{s,k}$, where $\mu_{s,k}$ represents the computing capacity that the VM of sth PS can process the kth UE's packets. Note that since each PS has a maximum computing capacity $C_s^{\mathrm{PS}}, \forall s \in \mathscr{S}$, we have the following constraints

$$\begin{cases} \sum_{k \in \mathscr{K}} \mu_{s,k} \leq d_s C_s^{\mathrm{PS}} \\ \mu_{s,k} \leq c_{s,k} C_s^{\mathrm{PS}} \\ c_{s,k} \leq d_s \end{cases} , \forall k \in \mathscr{K}, s \in \mathscr{S} \tag{C3}$$

Based on these, the average response time to process each packet for the kth UE at the VM of the sth PS is computed as $\frac{c_{s,k}}{\mu_{s,k}-\lambda_{s,k}}$, where $\lambda_{s,k} < \mu_{s,k}, \forall s \in \mathscr{S}, k \in \mathscr{K}$. Since the packets for the kth UE can be processed by multiple VMs of different computing capacities, the effective response time τ_k to process all packets of the kth UE in the BBU pool should be larger than the worst average response time among its serving VMs, leading to the following constraint

$$\begin{cases} \tau_k \geq \frac{c_{s,k}}{\mu_{s,k}-\lambda_{s,k}} \\ \mu_{s,k} \geq \lambda_{s,k} \end{cases}, \forall k \in \mathscr{K}, s \in \mathscr{S} \tag{C4}$$

5.2.4 Transmission Queue Model

After being processed by the VMs, the outcome packets from the processing queue are aggregated at a virtual switching node. Then, they are transported via the corresponding fronthaul links to the RRHs and eventually transmitted to the UEs. For simplicity, we neglect the transportation delay. By Burke's Theorem (Burke, 1956), the arrival process of transmission queue for the kth UE, (i.e., the departure process of processing queue for the kth UE) is still Poisson with rate Λ_k. Therefore, the data transmission to the kth UE from its serving RRHs can be modeled as a $M/M/1$ transmission queue service time $1/R_k(\mathbf{w})$ (Zhuang *et al.*, 2016) (cf. Fig. 5.1). Therefore, the average response time in the wireless transmission queue for the kth UE is simply given by

$$t_k(\mathbf{w}) = \frac{1}{R_k(\mathbf{w}) - \Lambda_k}, \forall k \in \mathscr{K} \tag{5.4}$$

where $R_k(\mathbf{w}) > \Lambda_k$ should be guaranteed for the queue stability. In this paper, we restrict the total response time of the processing and transmission queue by a delay value D_k to ensure a low-latency transmission for each UE, which is expressed as

$$\tau_k + \frac{1}{R_k(\mathbf{w}) - \Lambda_k} \leq D_k, \forall k \in \mathscr{K} \tag{C5}$$

It is noteworthy that virtual computing constraints are coupled with the physical constraints via QoS delay constraint in (C5), motivating the cross-layer joint design considered in this paper.

5.2.5 Power Consumption Model

5.2.5.1 RRH power consumption

According to (Luo *et al.*, 2015; Shi *et al.*, 2016b; Guo *et al.*, 2016b), the power consumption at each RRH is categorized into two types: data-dependent power, which is related to the transmitted signal, and data-independent power. The data-independent power can be further sub-categorized into two types: power to keep each ith RRH active, denoted as P_i^{ra}, and power to keep each ith RRH idle, denoted as P_i^{ri}. To formulate the design problem, we introduce a binary variable $b_i = \{0,1\}, \forall i \in \mathscr{I}$ to represent the operation mode of each ith RRH, where $b_i = 0$ indicates that the ith RRH is in sleep mode and $b_i = 1$ otherwise. The total power consumption at the ith RRH is written as

$$P_i^{\text{RRH}}\big(\mathbf{w}, b_i\big) = \frac{1}{\eta_i}\sum\nolimits_{k\in\mathscr{K}} \big\|\mathbf{w}_{i,k}\big\|_2^2 + b_i P_i^{\text{ra}} + \big(1-b_i\big)P_i^{\text{ri}} \tag{5.5}$$

where $\eta_i \in [0,1]$ is the power amplifier efficiency.

5.2.5.2 Fronthaul power consumption

We adopt the model in (Dai & Yu, 2016) where the fronthaul link power consumption directly depends on transmission rates served by the corresponding RRH. Specifically, the power consumption of the ith fronthaul link for forwarding information data and beamformers is written as

$$P_i^{\text{FH}}(\mathbf{w}, \mathbf{a}_i) = \rho_i \sum\nolimits_{k\in\mathscr{K}} a_{i,k} R_k(\mathbf{w}) \tag{5.6}$$

where $\mathbf{a}_i = [a_{i,1} \ldots, a_{i,K}]^T$ and $\rho_i = P_{i,\max}^{\text{FH}}/C_i^{\text{FH}}$ is the constant scaling factor associated to the ith fronthaul with $P_{i,\max}^{\text{FH}}$ is the power dissipation of ith fronthaul.

5.2.5.3 BBU power consumption

Let us define P_s^{PS} and $\kappa_s \mu_{s,k}^{\alpha_s}, \forall s \in \mathscr{S}, k \in \mathscr{K}$ as the power spent by the sth PS and the associated VMs for processing the kth UE's traffic, respectively, where in the polynomial approximation $\kappa_s \mu_{s,k}^{\alpha_s}$ (c.f.,(Tang *et al.*, 2015, 2014)), $\kappa_s > 0$, $\alpha_s > 1$ are the positive multiplication and exponent factors.Thus, by denoting $\mu = \{\mu_{s,k}, \forall s \in \mathscr{S}, \forall k \in \mathscr{K}\}$, the overall power consumption in the BBU pool is

$$P^{\mathrm{BBU}}(\mathbf{d}, \mu) = \sum_{s \in \mathscr{S}} d_s P_s^{\mathrm{PS}} + \sum_{s \in \mathscr{S}} \sum_{\forall k \in \mathscr{K}} \kappa_s \mu_{s,k}^{\alpha_s} \tag{5.7}$$

5.2.5.4 Total power consumption

Finally, the entire network power consumption in the considered system model is formulated as

$$P(\mathbf{w}, \mu, \mathbf{a}, \mathbf{b}, \mathbf{d}) = \sum_{i \in \mathscr{I}} \left(P_i^{\mathrm{RRH}}(\mathbf{w}, b_i) + P_i^{\mathrm{FH}}(\mathbf{w}, \mathbf{a}_i) \right) + \phi P^{\mathrm{BBU}}(\mathbf{d}, \mu) \tag{5.8}$$

where $\phi > 0$ is a parameter to strike a balance between the power consumption of RRHs, fronthaul and BBU pool, $\mathbf{b} = [b_1, \ldots, b_I]^T$ and $\mathbf{a} = \left[\mathbf{a}_1^T, \ldots, \mathbf{a}_I^T\right]^T$.

5.3 Problem Formulation

We aim at jointly optimizing the virtual computing resource allocation with beamforming, RRH selection and RRH-UE association to maximize the global network energy efficiency. To guarantee the stability of the transmission queue as shown in (5.4) and the minimum QoS UE rate requirement $R_k^{\min}$ for each UE k, we impose the following constraint

$$R_k(\mathbf{w}) \geq \max\left\{R_k^{\min}, \Lambda_k\right\} \tag{C6}$$

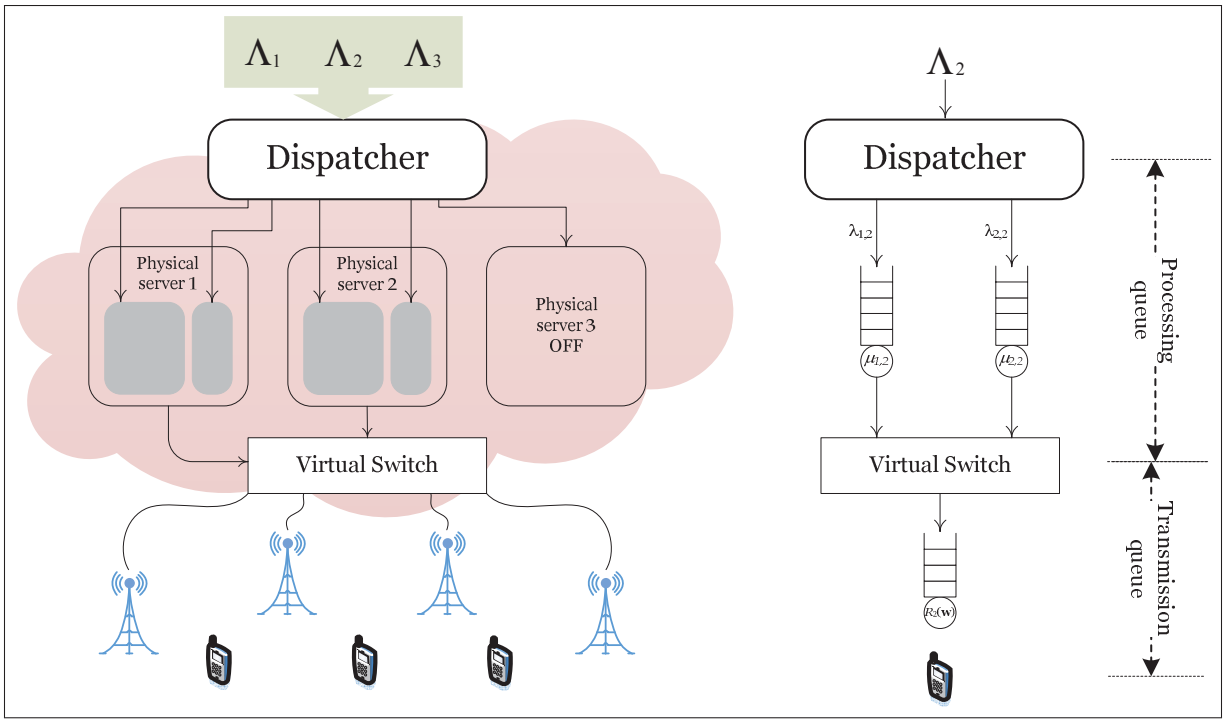


Figure 5.1 (a) Limited fronthaul C-RANs with VCRA scheme, (b) Queuing model, i.e., for UE 2

Moreover, the total transmit power at each RRH is limited by a power budget $P^{\max}$, which is expressed as

$$\sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq b_i P^{\max};\ \left\|\mathbf{w}_{i,k}\right\|_2^2 \leq a_{i,k} P^{\max};\ a_{i,k} \leq b_i \tag{C7}$$

where $\mathbf{a}$ and $\mathbf{b}$ are defined in (C1) and (5.5). The above constraint implies that when the ith RRH is in sleep mode, e.g., $b_i = 0$, no power will be transmitted from it. Similarly, we also guarantee that the transmit power $\left\|\mathbf{w}_{i,k}\right\|_2^2$ from the ith RRH to the kth UE is zero if $a_{i,k} = 0$. Also, whereas $b_i = 0$, then $a_{i,k} = 0$ for all $k \in \mathscr{K}$ and $\sum_{k\in\mathscr{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 = 0$. Now the considered problem is formulated as

$$(\mathscr{P}_0): \underset{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},\lambda,\tau,\mathbf{w},\mu}{\text{maximize}} \quad \frac{\sum_{k\in\mathscr{K}} R_k(\mathbf{w})}{P(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d})} \tag{5.9a}$$

$$\text{subject to} \quad (\text{C1});(\text{C2});(\text{C3});(\text{C4});(\text{C5});(\text{C6});(\text{C7}) \tag{5.9b}$$

where **a**, **b**, **c**, **d** are implicitly understood to be binary. To solve (5.9), we first customize a branch-and-reduce-and-bound (BnRnB) based algorithm which is presented in the next section.

5.4 Proposed Global Optimization Method

We present an algorithm to solve (5.9) optimally. Before proceeding further, we provide some comments on the complexity of (5.9). First, problem (5.9) is generally NP-hard due to the presence of binary variables **a**, **b**, **c** and **d**. Moreover, even when these binary variables are relaxed to be continuous, the obtained problem is still non-convex because of the non-convexity of the objective function (5.9a) and the constraints in (C1) and (C5). In mathematical programming, (5.9) is categorized as a mixed-integer non-convex program for which such a method in (Tang *et al.*, 2015; Guo *et al.*, 2016b; Cheng *et al.*, 2013) is not applicable to find a globally optimal solution. To the best of our knowledge, there is no off-the-shelf solver for (5.9). In what follows, we present an equivalent formulation of (5.9), based on which a BnRnB algorithm using monotonic optimization (MO) is customized to solve it optimally. We note that there are also other global optimization techniques such as inner and outer approximation, cutting-plane methods, etc. These optimal algorithms will yield the same optimal objective value. In this paper, we adopt the branch-and-reduce-and-bound (BnRnB) method to find a globally optimal solution for the considered problem since it lends itself to the considered problem, especially with a novel reformulation presented next.

5.4.1 Equivalent Formulation

Let us introduce the slack variables $\nu = \{\nu_i \geq 0, \forall i \in \{0 \cup \mathscr{K}\}\}$ and $\zeta = \{\zeta_k \geq 0, \forall k \in \mathscr{K}\}$ and rewrite (5.9) as the following problem

$$\begin{aligned}
\underset{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},\lambda,\tau,\mathbf{w},\mu,\nu,\zeta}{\text{maximize}} \quad & f(\nu) = \nu_0 \sum\nolimits_{k\in\mathscr{K}} \nu_k && \text{(5.10a)}\\
\text{subject to} \quad & R_k(\mathbf{w}) \geq \nu_k && \text{(5.10b)}\\
& \nu_k \geq \max\left\{R_k^{\min}, \Lambda_k\right\} && \text{(5.10c)}\\
& \hat{P}(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d},\nu) \leq 1/\nu_0 && \text{(5.10d)}\\
& \sum\nolimits_{k\in\mathscr{K}} a_{i,k}\nu_k \leq C_i^{\text{FH}} && \text{(5.10e)}\\
& \tau_k \geq c_{s,k}^2/(\mu_{s,k} - \lambda_{s,k}) && \text{(5.10f)}\\
& \zeta_k \geq 1/(\nu_k - \Lambda_k) && \text{(5.10g)}\\
& \tau_k + \zeta_k \leq D_k && \text{(5.10h)}\\
& \text{(C2)}, \text{(C3)}, \text{(C7)} && \text{(5.10i)}
\end{aligned}$$

where $\hat{P}(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d},\nu) = \sum_{i\in\mathscr{I}} \rho_i \sum_{k\in\mathscr{K}} a_{i,k}\nu_k + \sum_{i\in\mathscr{I}} P_i^{\text{RRH}}(\mathbf{w}, b_i) + \phi P^{\text{BBU}}(\mathbf{d},\mu)$. To arrive at (5.10), several slack variables have been introduced and the idea behind this step is justified as follows. First, we have replaced $c_{s,k}$ in (C4) by $c_{s,k}^2$, resulting in (5.10f). However, this maneuver still maintains the equivalence between (C4) and (5.10f) since $c_{s,k} = c_{s,k}^2$ for $c_{s,k} \in \{0,1\}$. The benefit of considering (5.10f) is that it is a convex constraint and particularly can be recast as a second order cone constraint, while (C4) is a non-convex one. This property will be exploited to develop a global optimization algorithm to (5.10). Second and more importantly, $R_k(\mathbf{w})$ has been replaced by ν_k or Λ_k at various places in (5.9). We remark that this move does not follow the standard reformulation technique based on epigraph form and thus the equivalence between (5.9) and (5.10) is not guaranteed in general. In this regard one of our main contributions is stated in the following lemma.

Lemma 3. *The formulations in* (5.9) *and* (5.10) *are equivalent in the sense that they have the same optimal solution set and objective.*

Proof. The proof is presented in Appendix 5. □

5.4.2 Optimal Solution based BnRnB Algorithm

The key benefits of the reformulation given in (5.10) are two-fold: first, it facilitates the customization of the BnRnB algorithm based on the MO framework to solve (5.10); and secondly in this regard, the search space for an optimal solution is reduced from $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \lambda, \tau, \mathbf{w}, \mu, \nu, \zeta)$ in (5.9) to only ν in (5.10), which then results in less computational complexity. Application of MO to solve (5.10) is possible due to the following two important observations.

- The objective in (5.10a) monotonically increases with respect to each entry of ν, which is obvious from the expression of $f(\nu)$ in (5.10a).

- For a given ν, the following feasibility problem

$$\text{find } \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \lambda, \tau, \mathbf{w}, \mu, \zeta \tag{5.11a}$$

$$\text{s.t. } (5.10\text{b}) - (5.10\text{i}). \tag{5.11b}$$

is a mixed-integer second order cone (MISOC) feasibility problem, which can be solved optimally by dedicated MISOCP solvers such as MOSEK.[2] Note that for a given ν_k, (5.10b) can be reformulated as a SOC constraint as

$$c' \Re(\mathbf{h}_k^H \mathbf{w}_k) \geq \left\| \mathbf{h}_k^H \mathbf{w}_1, \ldots, \mathbf{h}_k^H \mathbf{w}_K, \sigma_0 \right\|_2 \tag{5.12}$$

where $c' = \sqrt{\frac{1}{2^{\nu_k}-1} + 1}$.

From the above two facts, we can develop a BnRnB method to solve (5.10) optimally as done in (Tervo *et al.*, 2015). The detailed content and description of the BnRnB algorithm is similar to the presentation in (Tervo *et al.*, 2015, Algorithm 1), which is skipped here due to space constraint. Herein, we briefly present some important steps and definitions required to solve

[2] https://www.mosek.com/

the considered problem (5.10). Specifically, we first define the compact normal set $\mathscr{Q} = \{\nu \in \mathbb{R}_+^{K+1} | (5.10b) - (5.10i)\}$, i.e., $\forall \nu \in \mathscr{Q}$ such that problem (5.11) is feasible. We also define $\mathscr{V} = [\underline{\nu}; \overline{\nu}]$ to be the box that contains all ν feasible to (5.10). Note that $Q \subset \mathscr{V}$. The calculation of $\underline{\nu}$ and $\overline{\nu}$ is presented in Appendix 6. Problem (5.10) can now be abstractly expressed as $\max\{f(\nu) | \nu \in \mathscr{Q} \subset \mathscr{V}\}$.

The main idea to solve problem (5.10) optimally using MO framework is to check if a given ν belongs to $\mathscr{Q}$ or not, which amounts to solving the MISOC feasibility problem in (5.11). At the beginning of the proposed algorithm, we check whether $\underline{\nu}$ (i.e., the lower corner of $\mathscr{V}$) is feasible or not. If so, we apply the BnRnB method to find a globally optimal solution to (5.10). The proposed method recursively branches a box $\mathscr{B}$, which has the largest upper bound compared to others, into two smaller boxes, checks the feasibility of each new box, updates the current upper and lower bounds by the box reduction and bound computation process, and removes the boxes that do not contain an optimal solution. The details of these operations can be found in (Tervo *et al.*, 2015), and thus omitted here for the sake of brevity. As mentioned earlier, these steps are performed over ν, not over all optimization variables. This dimension reduction significantly reduce the overall complexity. Moreover, due to the monotonicity of the objective, the upper and lower bound of a box $\mathscr{B} = [\underline{\nu}, \bar{\nu}]$ can be quickly found as $U(\mathscr{B}) = f(\bar{\nu})$ and $L(\mathscr{B}) = f(\underline{\nu})$, respectively.According to Tervo *et al.* (2015), the proposed algorithm is bound improving and terminates after finitely many iterations for a given desired accuracy level ε.

To conclude this section, we remark that the proposed optimal Algorithm BnRnB presented in this section requires extremely high computational complexity for two apparent reasons. First, the MISOC feasibility problem in (5.11) is NP-hard in general and thus the complexity can increase exponentially with the problem size in the worst case. Second, the BnRnB algorithm (even when all binary variables are relaxed to be continuous) generally requires a large number of iterations to terminate. As a result, Algorithm BnRnB is practically useful for networks of relatively small size and is mainly used for benchmarking purpose in this paper. For a more

practically appealing solution we propose a low-complexity method based on the framework of DC programming in the next section.

5.5 Low-complexity Method

Given the inherent non-convexity and combinatorial nature of $(\mathscr{P}_0)$, a pragmatic goal is to find a sufficiently good feasible solution in a reasonable amount of time. To this end we will present a low-complexity algorithm in this section based on the following steps:

- Binary variables are relaxed to be continuous to obtain the continuous relaxation problem of $(\mathscr{P}_0)$, denoted as $(\mathscr{P}_1)$. This step is routine to handle the discreteness of the considered problem.
- As mentioned above, $(\mathscr{P}_1)$ is still non-convex and solving it is difficult. Although finding a globally optimal solution to $(\mathscr{P}_1)$ is possible by slightly modifying Algorithm BnRnB, the run time will be prohibitively high which is not suitable for real-time applications. The main idea of the proposed low-complexity method is to solve $(\mathscr{P}_1)$ using a local optimization method to compute a high-quality estimate of $(\mathscr{P}_0)$. Thus, we resort to the DC programming framework.
- The last step is to devise a post-processing procedure to map the solution produced by solving $(\mathscr{P}_1)$ which is not binary in general into a binary one.

In the next subsections we will present the details of the steps listed above.

5.5.1 DC Decomposition

The main target of the proposed low-complexity algorithm is to solve $(\mathscr{P}_1)$ efficiently. We recall that the non-convexity of $(\mathscr{P}_1)$ is due to that of function $R_k(\mathbf{w})$ and also the term $a_{i,k}R_k(\mathbf{w})$. Based on the concept of DC programming, we will express each of the non-convex functions as a difference of two convex ones. To illustrate this point let us first consider the rate function

$R_k(\mathbf{w})$. Note that the following decomposition holds

$$R_k(\mathbf{w}) = \underbrace{R_k(\mathbf{w}) + \xi_k \|\mathbf{w}\|_2^2}_{f_k(\mathbf{w})} - \xi_k \|\mathbf{w}\|_2^2 \tag{5.13}$$

for any ξ_k. Intuitively if ξ_k is sufficiently large, the quadratic term $\xi_k \|\mathbf{w}\|_2^2$ will dominate $R_k(\mathbf{w})$ and thus $f_k(\mathbf{w})$ becomes convex eventually. We remark that this kind of DC decomposition is not entirely new. But the problem is that finding a proper value for ξ_k to make (5.13) a DC expression is very challenging and problem-specific. In this regard our contribution is the following lemma.

Lemma 4. *For $\xi_k > \bar{\xi}_k$, where $\bar{\xi}_k$ is given in* (A I-20) *in Appendix 7, $f_k(\mathbf{w})$ is strongly convex .*

Proof. The proof and the derivation of $\bar{\xi}_k$ in Lemma 4 are involved and all the detailed algebra is presented in Appendix 7. The idea is to show that $R_k(\mathbf{w})$ is $\bar{\xi}_k$-smooth, i.e.,

$$\|\nabla R_k(\mathbf{x}) - \nabla R_k(\mathbf{y})\|_2 \leq \bar{\xi}_k \|\mathbf{x} - \mathbf{y}\|_2 \tag{5.14}$$

where $\nabla f(\mathbf{x})$ is the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$. Equivalently, $\nabla R_k(\mathbf{x})$ is Lipschitz continuous with a constant $\bar{\xi}_k$. □

We now turn the attention to a DC decomposition of the term $a_{i,k}R_k(\mathbf{w})$. By the same way, we consider the following DC decomposition

$$a_{i,k}R_k(\mathbf{w}) = \gamma_k\left(\|\mathbf{w}\|_2^2 + a_{i,k}^2\right) - \underbrace{\left(\gamma_k\left(\|\mathbf{w}\|_2^2 + a_{i,k}^2\right) - a_{i,k}R_k(\mathbf{w})\right)}_{u_k(\mathbf{w},a_{i,k})} \tag{5.15}$$

and the following lemma is in order.

Lemma 5. *For $\gamma_k > \bar{\gamma}_k$ where $\bar{\gamma}_k$ is given in* (A I-36) *in Appendix 8, $u_k(\mathbf{w}, a_{i,k})$ is strongly convex.*

Proof. The proof of Lemma 5 follows the same steps as those for that of Lemma 4 and is provided in Appendix 8. □

Based on the above DC decomposition, we are now in a position to describe the proposed algorithm to solve $(\mathscr{P}_1)$ efficiently. First $(\mathscr{P}_1)$ can be equivalently rewritten as

$$(\mathscr{P}_2): \max_{\substack{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}\\ \lambda,\tau,\mathbf{w},\mu}} \frac{\sum_{k\in\mathscr{K}} f_k(\mathbf{w}) - \xi_k \|\mathbf{w}\|_2^2}{\tilde{P}(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d})} \tag{5.16a}$$

$$\text{s.t. } f_k(\mathbf{w}) - \xi_k \|\mathbf{w}\|_2^2 \geq \max\left\{R_k^{\min}, \Lambda_k\right\} \tag{5.16b}$$

$$\sum_{k\in\mathscr{K}} \left(\gamma_k\left(\|\mathbf{w}\|_2^2 + a_{i,k}^2\right) - u_k(\mathbf{w}, a_{i,k})\right) \leq C_i^{\text{FH}} \tag{5.16c}$$

$$\frac{\left(\tau_k + \mu_{s,k} - \lambda_{s,k}\right)^2}{4} \geq c_{s,k} + \frac{\left(\tau_k - \mu_{s,k} + \lambda_{s,k}\right)^2}{4} \tag{5.16d}$$

$$f_k(\mathbf{w}) - \xi_k \|\mathbf{w}\|_2^2 - \Lambda_k \geq \frac{1}{D_k - \tau_k} \tag{5.16e}$$

$$a_{i,k}, b_i, c_{s,k}, d_s \in [0,1] \tag{5.16f}$$

$$\mu_{s,k} \geq \lambda_{s,k} \tag{5.16g}$$

$$\text{(C2); (C3); (C7)} \tag{5.16h}$$

where $\tilde{P}(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d}) = \sum_{i\in\mathscr{I}} P_i^{\text{RRH}}(\mathbf{w}, b_i) + \phi P^{\text{BBU}}(\mathbf{d},\mu) + \sum_{i\in\mathscr{I}} \rho_i \sum_{k\in\mathscr{K}} (\gamma_k\left(\|\mathbf{w}\|_2^2 + a_{i,k}^2\right) - u_k(\mathbf{w}, a_{i,k}))$. Note that we have equivalently rewritten constraints (C4) and (C5) as (5.16d) and (5.16e), respectively. The purpose of these reformulations is to express $(\mathscr{P}_2)$ as a DC program that is amenable to application of DCA, which is presented next subsection.

Before proceeding further, we remark that to deal with the non-convexity of such a problem as $(\mathscr{P}_1)$, a class of existing methods introduce some slack variables to expose the hidden convexity of non-convex objective and/or constraints, and then apply successive convex approximation to solve the resulting problem (Nguyen *et al.*, 2014b; Luong *et al.*, 2017a). The drawback of such a method is that the eventual number of optimization variables increases (quickly in many cases) with the problem size. In this regard, the DC form in $(\mathscr{P}_2)$ does not introduce any new

auxiliary variable, which certainly achieves more favorable scalability property and thus makes it more suitable for C-RANs.

5.5.2 DCA-based Method

In this section we apply DCA to solve $(\mathscr{P}_2)$. The main idea of DCA can be briefly explained as follows. Let us consider the following general DC constraint

$$p(\mathbf{x}) - q(\mathbf{x}) \leq 0 \tag{5.17}$$

where $p(\mathbf{x})$ and $q(\mathbf{x})$ are convex with respect to $\mathbf{x}$. It is obvious that the non-convex part in the above constraint is $-q(\mathbf{x})$ which is concave. Assuming $q(\mathbf{x})$ is differentiable (which is true for all constraints in $(\mathscr{P}_2)$), DCA linearizes $q(\mathbf{x})$ around the current iteration $\mathbf{x}^{(n)}$ to arrive at the following constraint

$$p(\mathbf{x}) - q(\mathbf{x}^{(n)}) - \left\langle \nabla q(\mathbf{x}^{(n)}), \mathbf{x} - \mathbf{x}^{(n)} \right\rangle \leq 0 \tag{5.18}$$

Note that (5.18) implies (5.17) as a concave function (i.e., $-q(\mathbf{x})$ as mentioned above) is upper bounded by its linearization. In other words, DCA arrives at an inner approximation of the feasible set of the considered nonconvex program and updates the point $\mathbf{x}^{(n)}$ until convergence.

Let us deal with the DC constraints in $(\mathscr{P}_2)$ first. According to the philosophy of DCA, we can approximate (5.16b) as

$$F_k(\mathbf{w}; \mathbf{w}^{(n)}) - \xi_k \|\mathbf{w}\|_2^2 \geq \max\left\{R_k^{\min}, \Lambda_k\right\} \tag{5.19}$$

where $F_k(\mathbf{w}; \mathbf{w}^{(n)})$ is given as

$$\begin{aligned} F_k(\mathbf{w}; \mathbf{w}^{(n)}) = f_k(\mathbf{w}^{(n)}) + \frac{\sum_{j \in \mathscr{K}} \left(2\Re\left(\mathbf{w}_j^{(n)H} \mathbf{H}_k \mathbf{w}_j\right) - 2\mathbf{w}_j^{(n)H} \mathbf{H}_k \mathbf{w}_j^{(n)}\right)}{\sum_{j \in \mathscr{K}} |\mathbf{h}_k^H \mathbf{w}_j^{(n)}|^2 + \sigma_0^2} \\ - \frac{\sum_{j \in \mathscr{K} \setminus k} \left(2\Re\left(\mathbf{w}_j^{(n)H} \mathbf{H}_k \mathbf{w}_j\right) - 2\mathbf{w}_j^{(n)H} \mathbf{H}_k \mathbf{w}_j^{(n)}\right)}{\sum_{j \in \mathscr{K} \setminus k} |\mathbf{h}_k^H \mathbf{w}_j^{(n)}|^2 + \sigma_0^2} + 2\xi_k \Re\left(\mathbf{w}^{(n)H} \mathbf{w}\right) - 2\xi_k \left\|\mathbf{w}^{(n)}\right\|_2^2 \end{aligned} \tag{5.20}$$

Note that $F_k(\mathbf{w};\mathbf{w}^{(n)})$ is simply a linearization of $f_k(\mathbf{w})$ around $\mathbf{w}^{(n)}$ and its derivation is in fact a by-product of Appendix 7. Thus, constraint (5.16e) can be approximated by the following constraint

$$F_k(\mathbf{w};\mathbf{w}^{(n)}) - \xi_k \|\mathbf{w}\|_2^2 - \Lambda_k \geq \frac{1}{D_k - \tau_k} \tag{5.21}$$

which is a convex constraint since $1/(D_k-\tau_k)$ is convex and $F_k(\mathbf{w};\mathbf{w}^{(n)}) - \xi_k \|\mathbf{w}\|_2^2 - \Lambda_k$ is concave with respect to all feasible variables $\mathbf{w}, \tau_k$. In the same way, we can also derive the upper bound convex approximation of the non–convex DC function $\gamma_k\left(\|\mathbf{w}\|_2^2 + a_{i,k}^2\right) - u_k(\mathbf{w}, a_{i,k})$ by deriving the lower bound concave approximation of $u_k(\mathbf{w}, a_{i,k})$ as follow

$$u_k(\mathbf{w}, a_{i,k}) \geq \tilde{F}_k(\mathbf{w}, a_{i,k}; \mathbf{w}^{(n)}, a_{i,k}^{(n)}) \tag{5.22}$$

where $\tilde{F}_k(\mathbf{w}, a_{i,k}; \mathbf{w}^{(n)}, a_{i,k}^{(n)})$ is given as

$$\begin{aligned}\tilde{F}_k(\mathbf{w}, a_{i,k}; \mathbf{w}^{(n)}, a_{i,k}^{(n)}) = \tilde{f}_k(\mathbf{w}^{(n)}, a_{i,k}^{(n)}) + \left[\log(\sum_{j\in\mathscr{K}} |\mathbf{h}_k^H \mathbf{w}_j^{(n)}|^2 + \sigma_0^2) - \log(\sum_{j\in\mathscr{K}\setminus k} |\mathbf{h}_k^H \mathbf{w}_j^{(n)}|^2 + \sigma_0^2)\right](a_{i,k} - a_{i,k}^{(n)}) \\ + a_{i,k}^{(n)}\left[\frac{\sum_{j\in\mathscr{K}}\left(2\Re\left(\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j\right) - 2\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j^{(n)}\right)}{\sum_{j\in\mathscr{K}}|\mathbf{h}_k^H\mathbf{w}_j^{(n)}|^2 + \sigma_0^2} - \frac{\sum_{j\in\mathscr{K}\setminus k}\left(2\Re\left(\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j\right) - 2\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j^{(n)}\right)}{\sum_{j\in\mathscr{K}\setminus k}|\mathbf{h}_k^H\mathbf{w}_j^{(n)}|^2 + \sigma_0^2}\right] \\ + 2\gamma_k\Re\left(\mathbf{w}^{(n)H}\mathbf{w} + a_{i,k}^{(n)}a_{i,k}\right) - 2\gamma_k\left(\left\|\mathbf{w}^{(n)}\right\|_2^2 + a_{i,k}^{(n)2}\right)\end{aligned} \tag{5.23}$$

Thus, constraint in (5.16c) can be approximated by its concave upper bound as

$$\sum_{k\in\mathscr{K}}\left(\gamma_k\left(\|\mathbf{w}\|_2^2 + a_{i,k}^2\right) - \tilde{F}_k(\mathbf{w}, a_{i,k}; \mathbf{w}^{(n)}, a_{i,k}^{(n)})\right) \leq C_i^{\mathrm{FH}} \tag{5.24}$$

By applying the above approximations, we can formulate the approximation of problem $(\mathscr{P}_2)$ at iteration $n+1$ as

$$\max_{\substack{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}\\ \boldsymbol{\lambda},\boldsymbol{\tau},\mathbf{w},\mu}} \quad \frac{\sum_{k\in\mathscr{K}}\left(F_k(\mathbf{w};\mathbf{w}^{(n)})-\xi_k\|\mathbf{w}\|_2^2\right)}{\hat{P}\left(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d};\mathbf{w}^{(n)},\mathbf{a}^{(n)}\right)} \tag{5.25a}$$

$$\text{s.t. } F_k(\mathbf{w};\mathbf{w}^{(n)})-\xi_k\|\mathbf{w}\|_2^2 \geq \max\left\{R_k^{\min},\Lambda_k\right\} \tag{5.25b}$$

$$\sum_{k\in\mathscr{K}}\gamma_k\left(\|\mathbf{w}\|_2^2+a_{i,k}^2\right)-\sum_{k\in\mathscr{K}}\tilde{F}_k(\mathbf{w},a_{i,k};\mathbf{w}^{(n)},a_{i,k}^{(n)}) \leq C_i^{\mathrm{FH}} \tag{5.25c}$$

$$c_{s,k}+\frac{(\tau_k-\mu_{s,k}+\lambda_{s,k})^2}{4} \leq \frac{\tau_k^{(n)}+\mu_{s,k}^{(n)}-\lambda_{s,k}^{(n)}}{2}(\tau_k+\mu_{s,k}-\lambda_{s,k}) - \frac{(\tau_k^{(n)}+\mu_{s,k}^{(n)}-\lambda_{s,k}^{(n)})^2}{4} \tag{5.25d}$$

$$F_k(\mathbf{w};\mathbf{w}^{(n)})-\xi_k\|\mathbf{w}\|_2^2-\Lambda_k \geq \frac{1}{D_k-\tau_k} \tag{5.25e}$$

$$a_{i,k},b_i,c_{s,k},d_s \in [0,1] \tag{5.25f}$$

$$\text{(C2);(C3);(C7);(5.16g)} \tag{5.25g}$$

where

$$\hat{P}\left(\mathbf{w},\mu,\mathbf{a},\mathbf{b},\mathbf{d};\mathbf{w}^{(n)},a_{i,k}^{(n)}\right)=\sum_{i\in\mathscr{I}}P_i^{\mathrm{RRH}}\left(\mathbf{w},b_i\right)+\phi P^{\mathrm{BBU}}\left(\mathbf{d},\mu\right)+\sum_{i\in\mathscr{I}}\rho_i\sum_{k\in\mathscr{K}}\gamma_k\times \left(\|\mathbf{w}\|_2^2+a_{i,k}^2\right)-\sum_{i\in\mathscr{I}}\rho_i\sum_{k\in\mathscr{K}}\tilde{F}_k(\mathbf{w},a_{i,k};\mathbf{w}^{(n)},a_{i,k}^{(n)}). \tag{5.26}$$

Note that the fractional objective (5.25a) can be easily transformed into a linear subtractive form using Dinkelback approach (Ng *et al.*, 2012). This subsequently makes (5.25a) convex and can be solved by the DCA-based algorithm, which is outlined in Algorithm 5.1.

Convergence analysis: We now prove that Algorithm 5.1 is guaranteed to converge. This can be established by showing that the sequence of objectives returned by Algorithm 5.1 is monotonically convergent. Towards this end, let $\theta^{(n)}$ and $\Theta^{(n)}$ denote the optimal objective value and the achieved optimal solution at the nth iteration of Algorithm 5.1, respectively. We will show that $\Theta^{(n)}$ is also feasible to problem (5.25) at the $(n+1)$th iteration. To see this let

Algorithm 5.1: DCA-based Algorithm.

1: Set $n := 0$ and initialize starting points of $\mathbf{w}^{(n)}, a^{(n)}, \tau^{(n)}, \mu^{(n)}, \lambda^{(n)}$;
2: **repeat**
3: Solve the approximated problem (5.25) at $\mathbf{w}^{(n)}, a^{(n)}, \tau^{(n)}, \mu^{(n)}, \lambda^{(n)}$ to achieve the optimal solution $\mathbf{a}^\star, \mathbf{b}^\star, \mathbf{c}^\star, \mathbf{d}^\star, \lambda^\star, \tau^\star, \mathbf{w}^\star, \mu^\star$;
4: Set $n := n+1$;
5: Update $\mathbf{w}^{(n)} = \mathbf{w}^\star, \mathbf{a}^{(n)} = \mathbf{a}^\star, \tau^{(n)} = \tau^\star, \mu^{(n)} = \mu^\star, \lambda^{(n)} = \lambda^\star$
6: **until** Convergence of the objective (5.25a);

us focus on the general DC constraint in (5.17). Due to the concavity of the term $-q(\mathbf{x})$, the following inequality holds

$$p(\mathbf{x}) - q(\mathbf{x}) \leq p(\mathbf{x}) - q(\mathbf{x}^{(n)}) - \left\langle \nabla q(\mathbf{x}^{(n)}), \mathbf{x} - \mathbf{x}^{(n)} \right\rangle \tag{5.27}$$

for all $\mathbf{x}$. Note that the right hand side of the above inequality stands for the resulting approximate constraint in the problem considered at the $(n+1)$th iteration of Algorithm 5.1. The inequality in (5.27) means that if $\mathbf{x}$ satisfies the approximate constraint in (5.18), then it also satisfies the DC constraint in (5.17). Thus, Algorithm 5.1 produces a sequence of iterates $\{\mathbf{x}^{(k)}\}$ that are feasible to the original problem, i.e., $p(\mathbf{x}^{(k)}) - q(\mathbf{x}^{(k)}) \leq 0$. Substituting $\mathbf{x}$ by $\mathbf{x}^{(n)}$ in (5.18) we have

$$p(\mathbf{x}^{(n)}) - q(\mathbf{x}^{(n)}) - \left\langle \nabla q(\mathbf{x}^{(n)}), \mathbf{x}^{(n)} - \mathbf{x}^{(n)} \right\rangle = p(\mathbf{x}^{(n)}) - q(\mathbf{x}^{(n)}) \leq 0. \tag{5.28}$$

The above inequality holds because $\mathbf{x}^{(n)}$ is feasible to the original problem. Thus, the solution of the nth iteration is feasible to the problem at iteration $(n+1)$. This leads to $\theta^{(n+1)} \geq \theta^{(n)}$, meaning that Algorithm 5.1 generates a non-decreasing sequence of objective values. Due to the power budget constraint (C7), the sequence of objectives $\{\theta^{\{n\}}\}$ is upper bounded and thus, is convergent.

5.5.3 An Accelerated Version of Algorithm 5.1: Practical Choices of ξ_k and γ_k

As being shown in the previous section, the sufficient conditions of ξ_k and γ_k to ensure the DC forms of functions (5.13) and (5.15) (and thus the convergence of Algorithm 5.1) are that $\xi_k \geq \bar{\xi}_k$ and $\gamma_k \geq \bar{\gamma}_k$, where $\bar{\xi}_k$ and $\bar{\gamma}_k$ are analytically computed in Appendix 7 and 8. However, smaller values of ξ_k and γ_k may significantly increase the convergence rate of Algorithm 5.1 in practice since they can lead to tighter approximation in each iteration of Algorithm 5.1. This will be easily seen from (5.13) where $f_k(\mathbf{w})$ is close to $R_k(\mathbf{w})$ for small ξ_k. Based on this observation, we set ξ_k and γ_k to a small value (elaborated in the numerical results section) in each iteration of Algorithm 5.1. If monotonic increase of the objective is not achieved, we then set ξ_k and γ_k to $\bar{\xi}_k$ and $\bar{\gamma}_k$, respectively. This variant is numerically shown to remarkably improve the convergence of Algorithm 5.1 and thus referred to as an accelerated version of Algorithm 5.1.

5.5.4 Post-Processing Procedure

A post-processing step is proposed to map the relaxed variables from solving $(\mathscr{P}_2)$ to the binary values, which is required due to the continuous relaxation. The process starts by assuming that all the RRHs and PSs are OFF and there is no association between RRHs, VMs and UEs. In each iteration, $(\mathscr{P}_2)$ is optimally solved given a set of active RRHs, RRH-UE association, active PSs and VM assignment that is connected. Let us denote optimal solution of $(\mathscr{P}_2)$ in the mth iteration of post-processing algorithm by $\{\mathbf{a}^\star, \mathbf{b}^\star, \mathbf{c}^\star, \mathbf{d}^\star, \lambda^\star, \tau^\star, \mathbf{w}^\star, \mu^\star\}$. The RRH-UE association and VM assignment are then gradually updated by fixing untreated relaxed variables to be 1. Apparently, which unfixed relaxed variables are prior to be picked up is a critical decision. Intuitively, the connection between the ith RRH and the kth UE is more likely if the channel link condition is good and the power consumed to transmit fronthaul data is smaller than the others. Similarly, the kth UE is preferred to be processed by VM in the sth PS if the power expended for switching on the sth PS is the smaller than the others and the total signal processing power consumed in the sth PS is larger. Based on the above intuitive observations, we define a virtual energy efficiency for assigning the ith RRH to serve the kth UE (i.e., incen-

tive measure to set $a_{i,k} = 1$) as $E_{i,k}$ and a normalized importance index for assigning the VM in the sth PS to process for the kth UE's traffic (i.e., incentive measure to set $c_{s,k} = 1$) as $I_{s,k}$, which are given by

$$E_{i,k} = \frac{R_k(\mathbf{w}^\star)}{\frac{1}{\eta_i}\left\|w_{i,k}^\star\right\|^2 + \rho_i R_k(\mathbf{w}^\star)} \tag{5.29}$$

$$I_{s,k} = \frac{\sum_{k\in\mathscr{K}} \kappa_s \left(\mu_{s,k}^\star\right)^{\alpha_s}}{P_s^{\mathrm{PS}} + \kappa_s \left(\mu_{s,k}^\star\right)^{\alpha_s}} \tag{5.30}$$

Hence, we propose to fix the untreated relaxed variables to be 1 in the mth iteration of post-processing algorithm, i.e., $a_{i',k'} = 1$, $c_{s',k'} = 1$ whose indices $(i',k'), (s',k')$ are selected by

$$(i',k') = \underset{(i,k)\in\mathscr{R}_{\mathrm{off}}^{(m-1)}}{\arg\max}\ E_{i,k} \quad \vee \quad (s',k') = \underset{(s,k)\in\mathscr{R}_{\mathrm{off}}^{(m-1)}}{\arg\max}\ \mathrm{I}_{s,k} \tag{5.31}$$

where $\mathscr{R}_{\mathrm{off}}^{(m-1)} = \{(i,k),(s,k)|\forall(i,k)\in(\mathscr{I},\mathscr{K}),\forall(s,k)\in(\mathscr{S},\mathscr{K}), a_{i,k}^{(m-1)} = 0, c_{s,k}^{(m-1)} = 0\}$ denotes the set of unfixed RRH-UE association and VM assignment in the $(m-1)$th iteration of the post-processing algorithm. This selection rule means that the unfixed relaxed binary variable that contributes mostly to the entire energy efficiency is set to be 1. According to constraints in (C3) and (C7), the variables b_i and d_s, $\forall i,s$ are fixed with respect to its associated variables $\mathbf{a}_i = \left\{a_{i,k}, \forall k \in \mathscr{K}\right\}$ and $\mathbf{c}_s = \left\{c_{s,k}, \forall k \in \mathscr{K}\right\}$, respectively. In particular, we need to set $b_i = 1$ or $d_s = 1$ if we fix any $a_{i,k} = 1$ or $c_{s,k} = 1$. Moreover, if variables $\mathbf{a}_i$ and $\mathbf{c}_s$ are fixed to $\mathbf{0}$, then we need to set $b_i = 0$ or $d_s = 0$. The RRH-UE association and VM assignment with the largest incentive measures (5.31) will be made connected and the resulting RRH and PS will be set active, following the relationship in $c_{s,k} \le d_s$ in (C3) and $a_{i,k} \le b_i$ in (C7). The overall algorithm is presented in Algorithm 5.2.

Convergence analysis: Algorithm 5.2 is provably convergent due to two facts. First, the DCA-based algorithm to solve $(\mathscr{P}_2)$ is guaranteed to converge as proved in the previous section. Second, the post-processing procedure is executed $\max\{(I-1)K,(S-1)K\}$ times in the worst

Algorithm 5.2: Post-processing algorithm.

1: Set $m := 0$, $\pi^{(m)}$ is significantly small, and initialize $a_{i,k}^{(m)} = 0, c_{s,k}^{(m)} = 0, b_i^{(m)} = 0, d_s^{(m)} = 0$, $\forall i \in \mathscr{I}, \forall k \in \mathscr{K}, \forall s \in \mathscr{S}$ and the set
$\mathscr{R}_{\text{off}}^{(m)} = \left\{ (i,k),(s,k) | \forall (i,k) \in (\mathscr{I},\mathscr{K}), \forall (s,k) \in (\mathscr{S},\mathscr{K}), a_{i,k}^{(m)} = 0,\ c_{s,k}^{(m)} = 0 \right\}$.
2: **repeat**
3: Set $m := m+1$;
4: Solve $(\mathscr{P}_2)$ given $a_{i',k'} = b_{i'} = 1$, $c_{s',k'} = d_{s'} = 1, \forall (i',k'), (s',k') \notin \mathscr{R}_{\text{off}}^{(m-1)}$ until convergence ;
5: Update $\mathscr{R}_{\text{off}}^{(m)} = \mathscr{R}_{\text{off}}^{(m-1)} \setminus \{ (i',k'),(s',k') | (5.31) \}$;
6: Solve $(\mathscr{P}_2)$ until convergence given all binary values:
$a_{i',k'} = b_{i'} = 1, c_{s',k'} = d_{s'} = 1, \forall (i',k'),(s',k') \notin \mathscr{R}_{\text{off}}^{(m)}$ and
$a_{i,k} = 0, c_{s,k} = 0, \forall (i,k),(s,k) \in \mathscr{R}_{\text{off}}^{(m)}$. If $(\mathscr{P}_2)$ is feasible, set $\pi^{(m)}$ as the value of objective function achieved at the convergence. If not, set $\pi^{(m)} = \pi^{(0)}$.
7: **until** $(\mathscr{P}_2)$ starts to be infeasible or it is feasible and $\pi^{(m)} < \pi^{(m-1)}$;

case. Hence, the post-processing algorithm mainly consists in solving finite times the problem $(\mathscr{P}_2)$ and it converges in finite iterations with a polynomial time computational complexity.

5.5.5 Complexity Analysis

We now discuss the worst-case per-iteration computational complexity of Algorithms 5.1 and 5.2. For Algorithm 5.1, (5.25) can be easily rewritten as a second order cone program (SOCP), whose total number of variables is $KM + 3SK + IK + K + S + I$ and total number of constraints is $4SK + 2IK + 3K + 2I$. Thus, the worst-case per-iteration computational complexity of Algorithm 5.1 and accelerated version of Algorithm 5.1 (ignoring the small orders) can be written as $\mathscr{O}(K^4(M^3+S^3+I^3)(S+I))$. Next, we analyze the worst-case per-iteration complexity of Algorithm 5.2. First we remark that in the worst case, Algorithm 5.2 must iteratively solve and update the resulting parameters for the problem $(\mathscr{P}_2)$ for $\max\{(I-1)K,(S-1)K\}$ times. In each step, the worst-case per-iteration complexity of solving $(\mathscr{P}_2)$ is $\mathscr{O}(K^4(M^3+S^3+I^3)(S+I))$. Therefore, the overall worst-case per-iteration computational complexity of Algorithm 5.2 $\mathscr{O}(\max\{(I-1)K,(S-1)K\}K^4(M^3+S^3+I^3)(S+I))$.

5.6 Numerical Results

We carry out extensive numerical experiments to evaluate the performance of the proposed algorithms. Unless mentioned otherwise, we employ the parameters in Table 5.1 in our simulations, which are taken from (Cheng *et al.*, 2013; Dai & Yu, 2016; Kansal *et al.*, 2010). We set the number of PSs S to be equal to the number of UEs K to ensure that all UEs's packets are always served in order to satisfy the worst case of schemes B and C in Section 5.6.2. We consider a network consisting of I RRHs which are uniformly located and K UEs are randomly scattered across the considered network coverage. The path-loss is modelled as $(d_{ik}/d_0)^{-3}$ where d_{ik} is the distance between the ith RRH and the kth UE and $d_0 = 100$ m is the reference distance. For the VM power consumption, we set $\kappa_s = 10^{-26}$ and $\mu_{s,k}$ is in cycle/s (Guo *et al.*, 2016b). In addition, we consider the conversion calculation $\mu_{s,k}$ b/s $= (8/1900) \times \mu_{s,k}$ cycle/s to compute the processing response time for UEs (Guo *et al.*, 2016b). Throughout our simulations, the accelerated variant of Algorithm 5.1 is used where ξ_k and γ_k are both first set to 0.1 in each iteration. Algorithm 5.1 is terminated when the increase in the objective between two consecutive iterations is less than 10^{-5}.

Table 5.1 Simulation parameters in Chapter 5

Notation	Value	Notation	Value
P_s	17 dBW	$P_{i,k}^{\mathrm{FH}}$	3 dBW
η_i	0.35	$C_s^{\mathrm{PS}} = C^{\mathrm{PS}}, \forall s$	2.5×10^3 cycle/s
P_i^{ra}	12.5 dBW	$\rho_i, \forall i$	1
M_i	2	$C_i^{\mathrm{FH}} = C^{\mathrm{FH}}, \forall i$	15 b/s/Hz
$P^{\max}$	10 dBW	$D_k = D, \forall k$	0.5 s
P_i^{ri}	2.5 dBW	α_s	3

5.6.1 Convergence Speed and Performance Gains by Proposed Algorithms

In Fig. 5.2, we show the convergence of the lower and upper bounds returned by Algorithm BnRnB for $I = 4$ and $K = 3$. As can be seen, Algorithm BnRnB requires about 10^3 iterations to compute an optimal solution. Fig. 5.3 compares the convergence behavior and gains be-

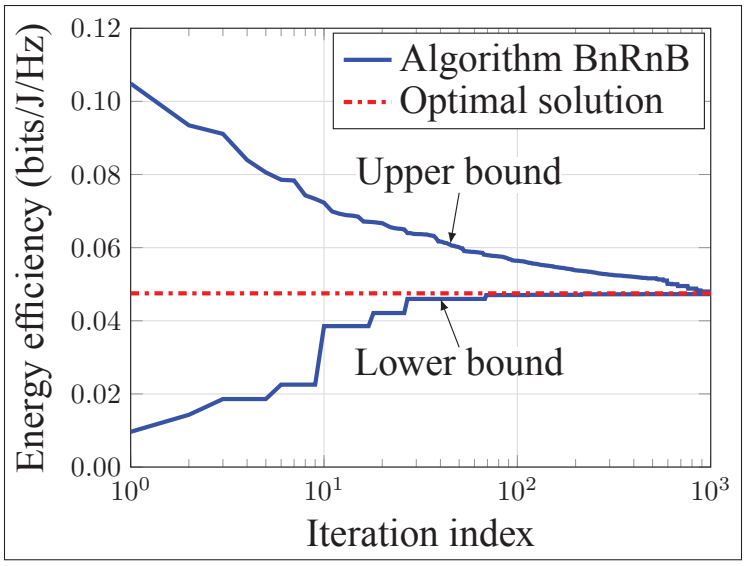


Figure 5.2 Convergence of the optimal BnRnB algorithm.

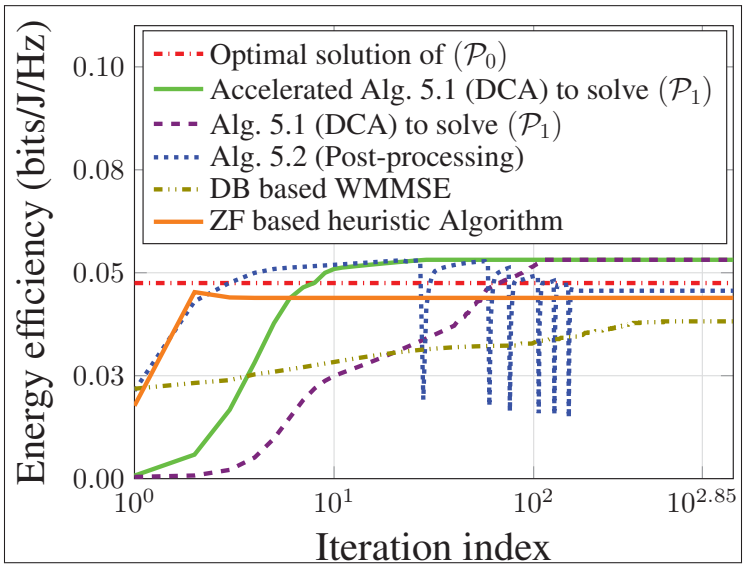


Figure 5.3 Convergence of different low complexity algorithms.

tween our proposed low-complexity algorithm (i.e., Algorithm 5.1), the DB-based WMMSE algorithm in (Peng *et al.*, 2016), and the zero-forcing (ZF) based heuristic algorithm. Algorithm 5.1 needs a much smaller number of iterations to converge, compared to the DB-based WMMSE algorithm. As expected, the accelerated version of Algorithm 5.1 achieves an improved convergence rate due to the reason explained in 5.5.3. Moreover, the convergence of

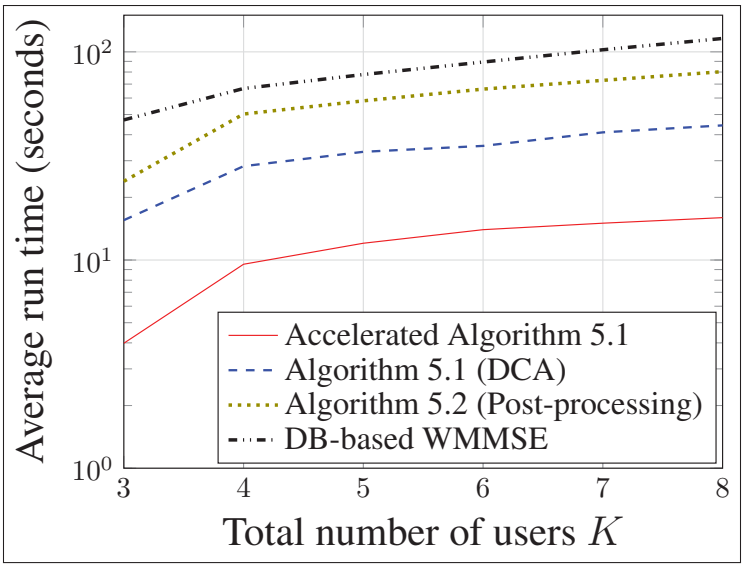
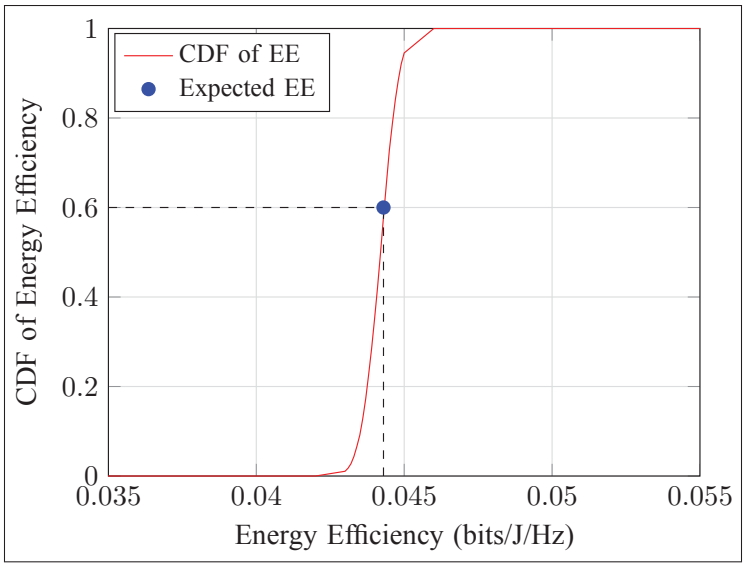


Figure 5.4 Average run time of low complexity algorithms versus K

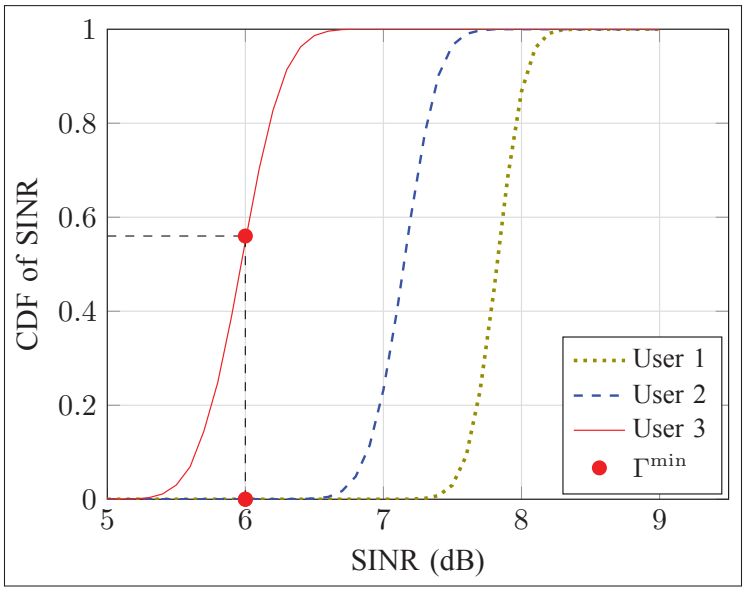
Algorithm 5.1 combined with the post-processing (i.e., Algorithm 5.2) is also presented Fig. 5.3. It is clear that the objective value achieved by Algorithm 5.1 combined with Algorithm 5.2 at convergence is very close to the optimal value returned by Algorithm BnRnB. On the other hand, the DB-based WMMSE and ZF based heuristic algorithms converges to a smaller objective compared to that achieved by our proposed algorithms. This demonstrates the superiority of our proposed low-complexity algorithms.

Fig. 5.4 plots the average run time required for the proposed algorithms to obtain the final solution versus K. We observe that the average run time increases with K which is expected from the complexity analysis presented in Section 5.5.5. Noticeably, Accelerated Algorithm 5.1 achieves lowest run time to return a solution which is consistent with the results shown in Fig. 5.3. We also observe that Algorithm 5.2 is much faster than the DB-based WMMSE algorithm. This again illustrates the effectiveness of our proposed algorithms compared to other existing ones.

To evaluate the time sensitivity of the computed solution from our proposed algorithm, in Figs. 5.5a and 5.5b, we plot the cumulative distribution function (CDF) of the EE and each UE's SINR. The empirical CDFs are obtained as follows. First (5.16) is solved for a given set



a)



b)

Figure 5.5 (a) CDF of the EE; (b) CDF of each UE's SINR.

of channel realizations using Algorithm 5.2. The resulting EE and minimum SINRs are denoted as the expected EE and $\Gamma^{\min}$. Then new channel realizations are generated by adding errors to the given channel realizations. The channel errors are drawn from a zero-mean Gaussian distribution with a variance of 0.01. The obtained solution is used to compute the EE and SINRs for these new channel realizations. As shown from Fig. 5.5a, 60% of the cases, the

achieved EE is equal and larger than the expected EE. Fig. 5.5b also shows that UE 1 and UE 2 achieve the minimum SINR requirement $\Gamma^{\min} = 6$ dB for 100% of simulated cases. That number for UE 3 is reduced to 50%. These results are quite positive in terms of the time sensitivity of the solution.

In Fig. 5.6, we compare the energy efficiency performance of our proposed algorithms with the DB-based WMMSE in (Peng *et al.*, 2016), SCA-based algorithms in (Tran & Pompili, 2017b), and the ZF based heuristic algorithm, with respect to the maximum fronthaul capacity C_i^{FH}. Here, we choose $C_i^{\text{FH}} = C^{\text{FH}}, \forall i \in \mathscr{I}$. We observe that when C^{FH} increases, energy efficiency of all methods in comparison increases accordingly. This is because larger fronthaul capacities allows more data to be transported, which requires a smaller number of activated RRHs to serve the demanding UEs and subsequently leads to reduced total power consumption. It is worth mentioning that larger fronthaul capacity promotes more coordination among RRHs so that inter-RRH interference is more effectively managed, and thus improves the overall achievable sum rate. These two factors collectively increase the energy efficiency. However, all the energy efficiency curves saturate at the high fronthaul capacity regime. This can be explained as the multi-user interference always exists despite more cooperation among the RRHs. For this situation, there is an upper bound on the achievable sum rate determined by the wireless interface of the network. Thus increasing more fronthaul capacity provides no benefit to the system performance. It is also shown that our proposed algorithms outperforms the DB-based WMMSE, SCA-based, and ZF based heuristic algorithms in terms of achieved energy efficiency, which again justifies the effectiveness of our proposed methods.

5.6.2 Advantages of Proposed Computing Model

Next we evaluate the performance of our proposed VCRA scheme and the rate-dependent fronthaul power consumption (RDFP) model. The following schemes are compared:

- Proposed Scheme: the proposed VCRA scheme and RDFP model are considered, where our proposed method in Algorithm 5.1 and Algorithm 5.2 is employed.

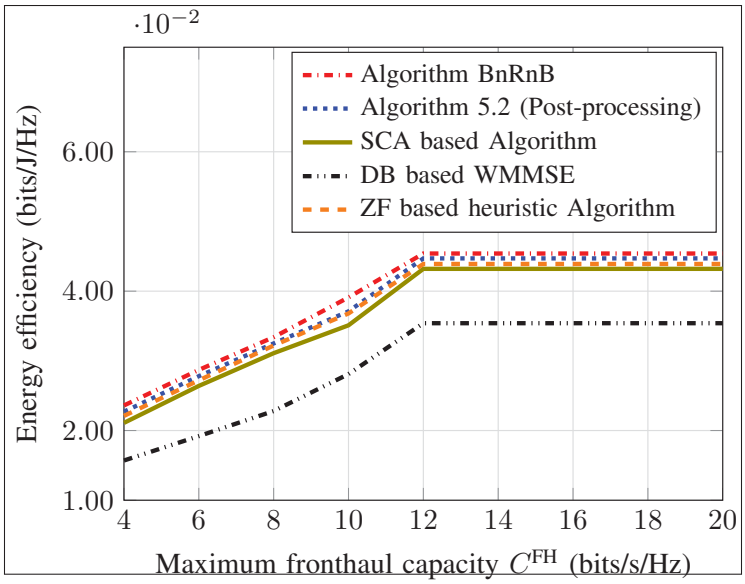


Figure 5.6 EE performance of different algorithms.

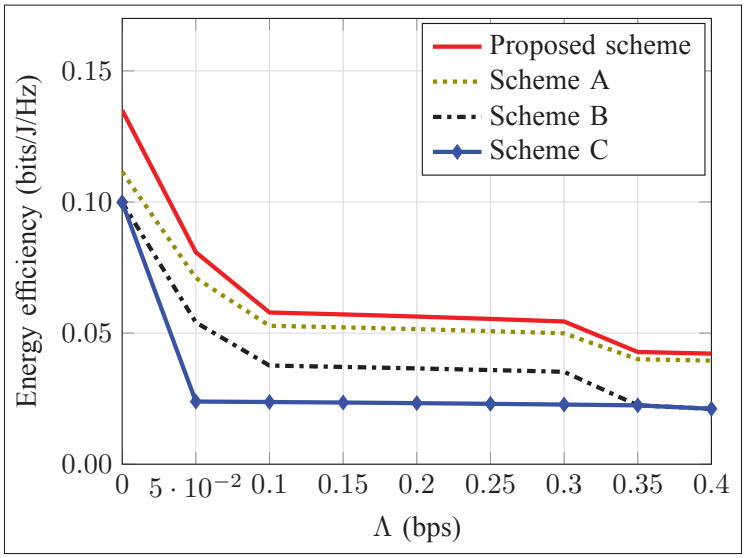


Figure 5.7 EE comparison of different algorithms versus Λ

- Scheme A: the proposed VCRA scheme without the RDFP model is considered, where the SCA-based algorithm in (Tran & Pompili, 2017b) is employed. Instead of using the RDFP model, a linear fronthaul power consumption model in (Luo *et al.*, 2015; Shi *et al.*, 2014) is applied by setting $P_i^{\mathrm{FH}} = \sum_{k \in \mathcal{K}} a_{i,k} P_{i,k}^{\mathrm{FH}}$ with $P_{i,k}^{\mathrm{FH}}$ being a fixed power consumption, i.e.,

$P_{i,k}^{\mathrm{FH}} = 2$ Watts (Shi *et al.*, 2014), for each data transmission between the ith fronthaul link and the kth UE.

- Scheme B: This scheme considers that the UE's workload is not split and thus the entire workload of one UE is served by only one VM (c.f., (Tang *et al.*, 2015; Guo *et al.*, 2016b)). Additionally, the linear fronthaul power consumption model in Scheme A is applied to this scheme. Here, the DB-based WMMSE combined with reweighted ℓ_1-norm technique in (Peng *et al.*, 2016; Dai & Yu, 2014) is employed and greedy algorithm in (Guo *et al.*, 2016b) is used to determine the active RRHs and RRH-UE associations.

- Scheme C: no PS switching ON/OFF is considered (which is similar to (Tang *et al.*, 2015; Wang *et al.*, 2016a)) and DB-based WMMSE combined with reweighted ℓ_1-norm technique used in Scheme B is applied.

Fig. 5.7 plots the energy efficiency performance of the above listed schemes as a function of the workload arrival rates when $I = 6, K = 4$. Here, we set the UE's workload arrival rates Λ_k equally to Λ, $\forall k \in \mathscr{K}$. As can be seen from Fig. 5.7, the energy efficiency attained by all schemes decreases when Λ increases, which can be explained as follows. As the traffic arrival rate grows, more computing resources and active PSs are needed to process the data. This results in the increase of power consumption in the BBU pool which then reduces the overall energy efficiency of the C-RAN system. In addition, we observe that our Proposed Scheme outperforms Scheme A, which verifying the benefit of considering the RDFH model in the formulated C-RAN optimization problem. Moreover, there is a remarkably large performance gap between our Proposed Scheme and Schemes B and C. This is because when Λ becomes larger, splitting the UE's workload into smaller fractions allows for more flexibility in assigning multiple VMs for different tasks and thus consolidate the existence of VMs to active PSs. As a result, more PSs can be switched OFF and more system power consumption can be saved, thereby enhancing the overall system energy efficiency. This again validates the advantages of our Proposed Scheme over the others in comparison.

In Fig. 5.8, the energy efficiency performance of different schemes is studied with respect to the total number of UEs K where $I = 6$. In this figure, we set the UE's workload arrival rate $\Lambda_k = \Lambda = 0.3$ bps and the delay $D_k = D = 0.5$ seconds, $\forall k \in \mathcal{K}$. It is obvious that when K increases, the energy efficiency first increases and then slightly decreases. The fact is that when the UE number increases, the total achievable rate first increases due to the multiuser diversity gain, which leads to an increase in the energy efficiency. However, when K becomes sufficiently large, a large number of RRHs and the PSs need to be activated to coordinate the induced interference. This in turn produces a huge amount of power consumption which subsequently decreases the achieved energy efficiency. Again, the energy efficiency achieved by our Proposed Scheme is much higher than that by Schemes A, B, and C.

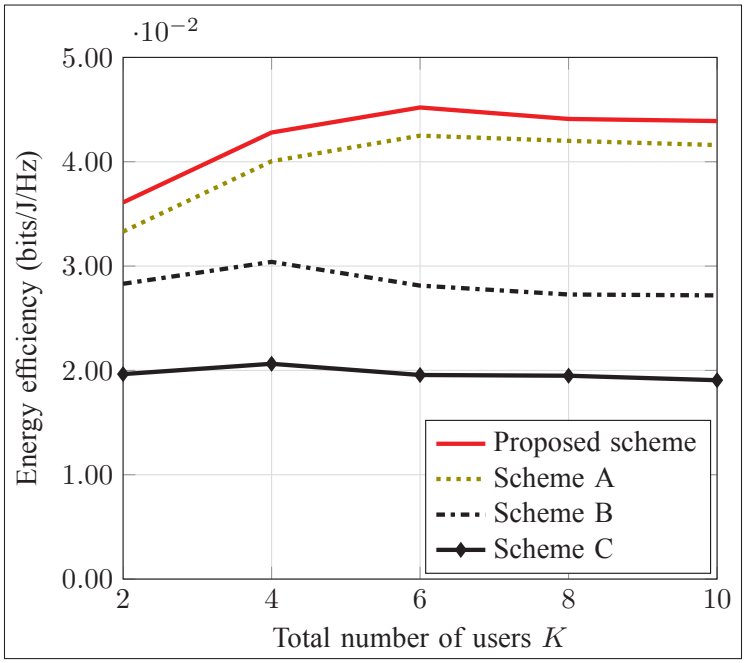


Figure 5.8 EE comparison of different algorithms versus K.

Fig. 5.9 demonstrates the impact of the joint optimization of virtual computing resource in the cloud and radio resource allocation in the RAN by comparing it with the decoupled optimization problem. Note that we can integrate the coupled and decoupled problem on all the above schemes. It is worth mentioning that the decoupled problem separately optimize

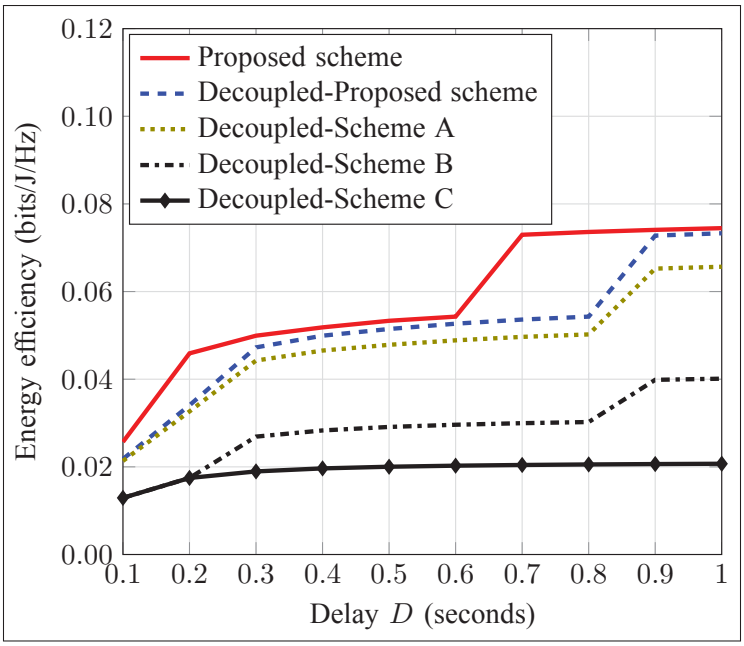


Figure 5.9 EE comparison of different schemes versus D.

each component. For the decoupled problem, to enable the separation between the the virtual computing resource and the radio resource allocation, the cross-layer delay constraint in (C5) are divided into two separated constraints, namely, the processing delay constraint $\tau_k \leq D_k/2$ and the transmission delay constraint $\frac{1}{R_k(\mathbf{w})-\Lambda_k} \leq D_k/2$, $\forall k \in \mathscr{K}$. Then, similar transformations and approximation techniques presented in Section 5.5 can be straightforwardly applied to solve this decoupled problem. The numerical results in Fig. 5.9 are obtained by setting the UEs' delay requirement to be $D_k = D\ \forall k \in \mathscr{K}$. From Fig. 5.9, we can see that the energy efficiency for both the coupled and decoupled schemes increases when D grows. This is because that large delay requirement offers less computing resources which leads to more idle PSs and more power savings. Moreover, it is obvious from Fig. 5.9 that the coupled design outperforms the decoupled design for each scheme. Furthermore, the joint design applied to our Proposed Scheme achieves better performance than the decoupled one for Schemes A, B, and C, which again verifies the advantages of our Proposed Scheme compared to other known methods in (Peng *et al.*, 2016; Tang *et al.*, 2015; Dai & Yu, 2014; Tran & Pompili, 2017b).

5.7 Concluding Remarks

We have considered the joint design of VCRA and radio resource allocation in a limited fronthaul C-RAN under a RDFP model for maximizing the network EE. To solve the formulated problem, we have customized a BnRnB algorithm to search for a globally optimal solution. We have also developed a novel low-complexity and more appealing algorithm which is provably convergent. The proposed method is inspired by the DCA method combined with the Lipschitz continuity concept, approximating the non-convex problem into a sequence of convex quadratic ones, which can be efficiently solved by dedicated convex conic solvers. After solving the continuous relaxation, a post-processing routine is executed to find a high-performance solution which is feasible to the original problem. Numerical results showed that our proposed algorithms converge rapidly and achieve a near-optimal performance as well as outperform the other existing methods. Additionally, we have also numerically demonstrated that the proposed VCRA scheme not only efficiently allocates virtual computing in the BBU pool to process the user's workload in parallel but also significantly reduces the total power consumption.

CONCLUSION AND RECOMMENDATIONS

6.1 Summary

By decoupling BBU pools from RRH that allows the computing resource cloudification and virtualization in the BBU pool and easy radio transmission coordination between RRHs, C-RAN enables more flexible, scalable and efficient resource designs that can not be achieved in the other traditional wireless networks. In particular, centralized baseband signal processing in the BBU pool appropriately allocates the resources from multiple coordinated RRHs to serve users at higher achievable rate and lower power consumption. Through virtualization technology, computation resources can be flexibly adjusted as on-demand request to meet the fluctuation of user traffic. With the proven significant benefits of SE and EE performance gains, C-RAN again illustrates its role as a key technology for the 5G system. However, such a distinct network structure and functions in C-RAN pose several major challenges to seek an optimal resource allocation strategy in order to attain the best system performance. Specifically, C-RAN performance strictly depends on the transport network, e.g., fronthaul, that connects RRHs and the BBU pool. Thus, considering the stringent requirements in the fronthaul transmission such as capacity, latency, power consumption, etc., is necessary when designing the resource allocation in C-RAN. Likely, the cooperative RRHs clusters for user-centric strategy must be carefully designed in such a way to improve C-RAN throughput performance without violating the fronthaul constraints. Moreover, to cope with the very dense RRH and huge BBU pool deployment, the power saving management needs to be intensively scrutinized, especially the computing power consumption, fronthaul power consumption and RRH on/off strategy.

In this dissertation, we have presented the technical solutions to deal with these challenges in C-RAN. In detail, we have developed the energy efficient resource allocation algorithms to maximize the system performance of limited fronthaul capacity constrained C-RAN. In Chapter 3, we have proposed the joint design of transmit beamforming, user association and RRH

selection to optimize the trade-off between sum achievable rate and total power consumption, considering the limited fronthaul capacity constraints. Our algorithms based on the SCA method proved the significant enhancement of achieved performance compared to other works.

Additionally, we have investigated the energy efficient resource allocation design to maximize the different EE metrics, consisting of global energy efficiency (GEE), weighted sum energy efficiency (WSEE) and fairness of energy efficiency (FEE), taking into account the rate dependent fronthaul power consumption model in Chapter 4. We have proposed the SCA based framework to solve the different EE maximization problems. The performance evaluation of GEE, WSEE and FEE was provided. Moreover, the numerical results demonstrated that the rate dependent power consumption model had the largest impact on the EE performance, and thus highlights our contributions.

In the final contribution, we have proposed the joint design of virtual computing, beamforming, user association and RRH selection for C-RAN, which aims at maximizing the system energy efficiency considering the end-to-end latency constraints. We have developed the difference of convex algorithms to tackle the nonconvex optimization problem without introducing any slack variables. The numerical results were provided to evaluate the better improvement of our proposed strategy and algorithms than other works.

6.2 Future Research

Our dissertation focused on the resource allocation to improve the performance of energy efficiency in C-RAN considering fronthaul capacity limitation. However, there are still other important research directions that should be further investigated as the following.

6.2.1 Resource allocation design with imperfect CSI

In our dissertation, we assume that perfect CSI between all RRHs and the users are available in the BBU pool. In fact, the CSI between RRHs and users obtained in the BBU pool for the resource allocation design is not perfect due to the delayed CSI feedback, quantization error, channel estimation error. Since perfect CSI is hardly achieved in practice, there is also a class of studies dealing with imperfect CSI which are based on worst-case design (Shi *et al.*, 2015) or probabilistic approaches (Pan *et al.*, 2018; Lau *et al.*, 2013). However, the stochastic CSI uncertainty approaches will produce the probabilistic QoS constraints, resulting in chance constrained programming problems that are highly intractable in general. Thus, this deserves a thorough separate study of robust resource allocation algorithms which may use the newly introduced constraints to capture the imperfect CSI should be considered in our future work.

6.2.2 Cache allocation optimization for C-RAN

To decrease the data information exchanged through fronthauls and alleviate radio signal processing computation burden in the BBU pool, there are promising approaches that distribute a part of resources such as storage capacity, radio resources and signal processing capacity to the base stations and edge devices. In particular, each RRH can be equipped a cache with finite size to store a mount of file contents so that the BBU pool needs to deliver only the rest contents of the requested file to that RRH. By doing this way, the fronthaul congestion is further reduced and the end-to-end downloading time from the BBU pool to the users is much enhanced. It can be seen that C-RAN with RRH caching has been investigated in the recent literature (Ugur *et al.*, 2016; Zhou *et al.*, 2015; Sun *et al.*, 2016; Chen *et al.*, 2016), however most of the works focus on analyzing the effect of RRH caching to fronthaul capacity consumption and the performance of C-RAN utilities under the fixed cache allocation among the RRHs. In addition, there have been some works that aim at optimizing the caches at the RRHs (Dai *et al.*, 2018;

Bidokhti *et al.*, 2018, 2017; Nguyen *et al.*, 2018a, 2017b), but the channel conditions between the BBU pool and RRHs in the wireless fronthaul network coupling with different multicast channels have not been deeply considered. Consequently, this can be a potential direction of our research in the future.

6.2.3 Mobile Edge Computing in C-RAN

Mobile Edge Computing (MEC) has recently been seen as a potential technology that significantly reduces mobile device energy by offloading heavy-loaded application task to the edge server equipping with powerful computing and storage resources. In the context of C-RAN, MEC servers can be installed at the RRHs, thus enables the operation of low-latency and context-aware applications of 5G networks to be performed closely to the side of end users. Specifically, the applications which are more sensitive to delay and demand high computing tasks can be executed at the network edge through the design of resource allocation between mobile RAN and cloud computing center. In addition, partial amount of data can be pre-processed before being sent to the cloud, thus not all data is transmitted between cloud center and RAN. This results in the alleviation of fronthaul congestion and core workload (Tran *et al.*, 2017). Therefore, optimizing the resource allocation and task offloading in the MEC and C-RAN integration can further improve the overall network performance and should be investigated in our future research.

6.2.4 Machine learning for C-RAN

Recently, it has been shown that the integration of artificial intelligence (AI) and machine learning into the wireless infrastructure and edge devices has brought more reliable, predictable, better self-organized and self-optimized features and applications in the future wireless networks, especially the IoT systems (Chen *et al.*, 2017). One of important aspect in wireless networks is that the amount of CSI obtained at the BBU pool is limited due to the fronthaul requirements

and must be updated very quickly to adapt the real-time resource allocation. Furthermore, with the advantages of cloud computing techniques in the BBU pool that enables huge scalable computation resources to be massively deployed and enormous amount of unstructured data to be stored cost effectively, machine learning techniques can be easily employed in the BBU pool that provide the unprecedented perception on channel variation and user behaviors. Specially, using machine learning techniques to analyze the vast amount of collected data related to file requested from users is able to generate the precise predictive cache allocation strategies and intelligent cached content recommendations. Hence, the development of machine learning algorithms with predictive capabilities for resource allocation in C-RAN to improve user scheduling, cache allocation, and channel estimation are mandatory and an important research direction for our future research.

APPENDIX I

APPENDICES

1. Proof of Lemma 1

We prove that the constraints in (3.8b) and (3.8d) of problem (3.8) are active at optimality by contradiction. Let $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \nu^\star, \mathbf{u}^\star)$ denote an optimal solution of (3.8). By contradiction, suppose that (3.8d) is inactive, i.e., $P^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star) < 1/u_0^\star$. Then there exists u_0' such that $u_0' > u_0^\star$ and $P^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star) < 1/u_0'$. That is, u_0' is feasible to (3.8) but yields a strictly larger objective, which contradicts with the fact that $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \nu^\star, \mathbf{u}^\star)$ is an optimal solution. Similarly, assume that $R_k(\mathbf{w}^\star) > u_k^\star$ for some k. We then create a new set of beamformers as $\mathbf{w}' = [\mathbf{w}_1'^T, \mathbf{w}_2'^T, \ldots, \mathbf{w}_k'^T]^T$ where

$$\mathbf{w}_i' = \begin{cases} \mathbf{w}_i^\star & i \neq k \\ \zeta \mathbf{w}_k^\star & i = k \end{cases} \tag{A I-1}$$

for some $0 < \zeta < 1$. Intuitively, the beamforming vector of user k is scaled down by a factor of ζ and the beamforming vectors of other users remain the same. From (**??**), it is easy to see that there exists $\zeta \in (0,1)$ such that $R_k(\mathbf{w}') > u_k^\star$ for all k. Note that $\|\mathbf{w}'\|_2 < \|\mathbf{w}\|_2$ and thus $P^{\text{tot}}(\mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star) < P^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star) \leq 1/u_0^\star$. Consequently, we find u_0' such that $u_0' > u_0^\star$ and $P^{\text{tot}}(\mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star) \leq 1/u_0'$, meaning that a strictly larger objective can be obtained. Again, this contradicts with the fact that $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \nu^\star, \mathbf{u}^\star)$ is an optimal solution, and thus completes the proof.

2. Proof of (3.25) and (4.42)

We first show that the gradient of the function $g(x) = -\log(1+x)$ for $x \geq 0$ is Lipschitz continuous with parameter $L = 1$. This can be easily proved since

$$\|\nabla g(x_1) - \nabla g(x_2)\|_2 = \left| -\frac{1}{1+x_1} + \frac{1}{1+x_2} \right| = \left| \frac{x_1 - x_2}{(1+x_1)(1+x_2)} \right| \overset{(a)}{\leq} |x_1 - x_2| \tag{A I-2}$$

where (a) is due to $(1+x_1)(1+x_2) > 1$ for $x_1, x_2 > 0$. Due to the Lipschitz continuity of $\nabla g(x)$, it holds that Parikh & Boyd (2014)

$$g(\gamma_k) \leq g(\gamma_k) + \nabla g(\gamma_k^{(n)}) \left(\gamma_k - \gamma_k^{(n)}\right) + \frac{1}{2\lambda} \left(\gamma_k - \gamma_k^{(n)}\right)^2 \tag{A I-3}$$

for $\lambda \in (0,1]$, and thus completes the proof by noting that (A I-3) is actually (4.42) when $\lambda = 1$.

3. Proof of the equivalence of (4.12) and (4.14)

We prove that the constraints in (4.14b) and (4.14d) of problem (4.14) are hold with equalities at optimality by contradiction. Let $\Theta^\star = \{\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \nu^\star, \tau^\star, \mathbf{t}^\star\}$ denote an optimal solution of (4.14). By contradiction suppose that (4.14d) is inactive, i.e., $\tilde{P}^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star) < 1/t_0^\star$ for GEE maximization problem and $\tilde{P}_i(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star) < 1/t_i^\star$ for WSEE and FEE maximization problems. Then, for GEE, there exists t_0' such that $t_0' > t_0^\star$ and $\tilde{P}^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star) \leq 1/t_0'$. That is, t_0' is feasible to (4.14) but yields a strictly larger objective, which contradicts with the fact that $\Theta^\star$ is an optimal solution. Similarly, assume that $R_k(\mathbf{w}^\star) > \tau_k^\star$ for some k. We then create a new set of beamformers as $\mathbf{w}' = [\mathbf{w}_1'^T, \mathbf{w}_2'^T, \ldots, \mathbf{w}_k'^T]^T$ where

$$\mathbf{w}_i' = \begin{cases} \mathbf{w}_i^\star & i \neq k \\ \xi \mathbf{w}_k^\star & i = k \end{cases} \tag{A I-4}$$

for some $0 < \xi < 1$. Intuitively, the beamforming vector of user k is scaled down by a factor of ξ and the beamforming vectors of other users are remain the same. It is easy to see that there exists $\xi \in (0,1)$ such that $R_k(\mathbf{w}') > \tau_k^\star$ for all k. Note that $\|\mathbf{w}'\|_2 < \|\mathbf{w}\|_2$ and thus $\tilde{P}^{\text{tot}}(\mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star) < \tilde{P}^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star) \leq 1/t_0{}^\star$. Consequently, we can find t_0' such that $t_0' > t_0^\star$ and $\tilde{P}^{\text{tot}}(\mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star) \leq 1/t_0'$, meaning that a strictly larger objective can be obtained. Again, this contradicts with the fact that $\Theta^\star$ is an optimal solution, and thus proves that the constraints in (4.14b) and (4.14d) of problem (4.14) are hold with equalities at optimality. As a result, for given optimal solution $\Theta^\star$ to (4.14) we can simply obtain optimal solution $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \nu^\star)$ to (4.12) and the same objective value as that of (4.14) and vice versa for given optimal solution

$(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, v^\star)$ to (4.12), we can easily find the corresponding optimal solution $\Theta^\star$ to (4.14) by calculating $\tau_k^\star = R_k(\mathbf{w}^\star), \forall k \in \mathscr{K}$ and $1/t_0^\star = \tilde{P}^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \tau^\star)$, and also achieve the same objective value as that of (4.12). Thus, Lemma 3 is proved. Similar proof steps can be applied for WSEE and FEE maximization problems.

4. Calculation of box $\mathscr{U}$

The values of $\underline{s} = [\underline{\mathbf{a}}^T, \underline{\tau}^T, \underline{\mathbf{t}}^T]^T$ and $\overline{s} = [\overline{\mathbf{a}}^T, \overline{\tau}^T, \overline{\mathbf{t}}^T]^T$ can be computed as follows. It is easily seen that $\underline{\mathbf{a}} = \mathbf{0}_N$ and $\overline{\mathbf{a}} = \mathbf{1}_N$ where $N = IK$ since $a_{i,k} \in \{\underline{a}_{i,k}, \overline{a}_{i,k}\} = \{0,1\}$, where $\mathbf{0}_N$ and $\mathbf{1}_N$ denote vectors of all zeros and ones with length of vector is N. Additionally, from (4.14b), it holds that $\tau_k \geq \log(1+\Gamma_k^{\min}) = \underline{\tau}_k, \forall k \in \mathscr{K}$. Moreover, we have

$$\tau_k \overset{(a)}{\leq} \log(1+\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sigma_0^2}) \overset{(b)}{\leq} \log\left(1+\frac{\left\|\mathbf{h}_k\right\|_2^2 \left\|\mathbf{w}_k\right\|_2^2}{\sigma_0^2}\right) \overset{(c)}{\leq} \log\left(1+I\times P^{\max}\frac{\left\|\mathbf{h}_k\right\|_2^2}{\sigma_0^2}\right) = \overline{\tau}_k, \forall k \in \mathscr{K}. \tag{A I-5}$$

where (a) is due to omitting the inter-user interference, (b) is the result of applying Cauchy–Schwarz inequality, and (c) is obvious from the power constraint for each $\mathbf{w}_{i,k}$. Similarly, an upper bound and lower bound of t_i can be given by

$$t_i \geq \frac{1}{\overline{p}_i} = \underline{t}_i; t_i \leq \frac{1}{\underline{p}_i} = \overline{t}_i \tag{A I-6}$$

where

$$\begin{cases} \overline{p}_0 = \sum_{i\in\mathscr{I}} \rho_i \sum_{k\in\mathscr{K}} \overline{\tau}_k + \sum_{i\in\mathscr{I}} (\frac{1}{\eta_i} P^{\max} + P_i^{\text{ra}}) & \text{if X is GEE} \\ \overline{p}_i = \rho_i \sum_{k\in\mathscr{K}} \overline{\tau}_k + \frac{1}{\eta_i} P^{\max} + P_i^{\text{ra}} & \text{otherwise} \end{cases} \tag{A I-7}$$

and

$$\begin{cases} \underline{p}_0 = \sum_{i\in\mathscr{I}} P_i^{\text{ri}} & \text{if X is GEE} \\ \underline{p}_i = P_i^{\text{ri}}, \forall i \in \mathscr{I} & \text{otherwise} \end{cases} \tag{A I-8}$$

5. Proof of the equivalence of (5.9) and (5.10)

We prove that the constraints in (5.10b) and (5.10d) of problem (5.10) are hold with equalities at optimality by contradiction. Let $\Theta^\star = \{\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{c}^\star, \mathbf{d}^\star, \lambda^\star, \tau^\star, \mathbf{w}^\star, \mu^\star, \nu^\star, \zeta^\star\}$ denote an optimal solution of (5.10). By contradiction suppose that (5.10d) is inactive, i.e.,

$$\hat{P}(\mathbf{d}^\star, \mu^\star, \mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \nu^\star) < \frac{1}{\nu_0^\star}$$

Then there exists ν_0' such that $\nu_0' > \nu_0^\star$ and

$$\hat{P}(\mathbf{d}^\star, \mu^\star, \mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \nu^\star) \leq \frac{1}{\nu_0'}$$

ν_0' is feasible to (5.10) but yields a strictly larger objective, which contradicts with the fact that $\Theta^\star$ is an optimal solution. Similarly, assume that $R_k(\mathbf{w}^\star) > \nu_k^\star$ for some k. We then create a new set of beamformers as $\mathbf{w}' = [\mathbf{w}_1'^T, \mathbf{w}_2'^T, \ldots, \mathbf{w}_k'^T]^T$ where $\mathbf{w}_i' = \mathbf{w}_i^\star$ if $i \neq k$ and $\mathbf{w}_i' = \xi \mathbf{w}_k^\star$ otherwise, for some $0 < \xi < 1$. Intuitively, the beamforming vector of UE k is scaled down by a factor of ξ and the beamforming vectors of other UEs remain the same. It is easy to see that there exists $\xi \in (0,1)$ such that $R_k(\mathbf{w}') > \nu_k^\star$ for all k. Note that $\|\mathbf{w}'\|_2 < \|\mathbf{w}\|_2$ and thus $\hat{P}(\mathbf{d}^\star, \mu^\star, \mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star, \nu^\star) < \hat{P}(\mathbf{d}^\star, \mu^\star, \mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \nu^\star) \leq 1/\nu_0{}^\star$. Consequently, we can find ν_0' such that $\nu_0' > \nu_0^\star$ and $\hat{P}(\mathbf{y}^\star, \mu^\star, \mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star, \nu^\star) \leq 1/\nu_0'$, meaning that a strictly larger objective can be obtained. Again, this contradicts with the fact that $\Theta^\star$ is an optimal solution, and thus proves that the constraints in (5.10b) and (5.10d) of problem (5.10) hold with equalities at optimality. As a result, for given optimal solution $\Theta^\star$ to (5.10) we can simply obtain optimal solution $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{c}^\star, \mathbf{d}^\star, \lambda^\star, \tau^\star, \mathbf{w}^\star, \mu^\star)$ to (5.9) and the same objective value as that of (5.10). Similarly, for given optimal solution $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{c}^\star, \mathbf{d}^\star, \lambda^\star, \tau^\star, \mathbf{w}^\star, \mu^\star)$ to (5.9), we can easily find a feasible solution $\Theta^\star$ to (5.10) by setting $\nu_k^\star = R_k(\mathbf{w}^\star)$, $1/\nu_0^\star = \hat{P}(\mathbf{d}^\star, \mu^\star, \mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star, \nu^\star)$ and $D_k - \tau_k^\star \geq \zeta_k^\star \geq 1/(\nu_k^\star - \Lambda_k)$, $\forall k \in \mathscr{K}$ that achieves the same objective value as that of (5.9).

6. Calculation of box $\mathscr{V}$

The values of $\underline{\nu} = \{\underline{\nu}_0, \underline{\nu}_1, \ldots, \underline{\nu}_K\}$ and $\overline{\nu} = \{\overline{\nu}_0, \overline{\nu}_1, \ldots, \overline{\nu}_K\}$ can be computed as follows. From (5.10c), it holds that $\nu_k \geq \log(1+\Gamma_k^{\min}) = \underline{\nu_k}$, $\forall k \in \mathscr{K}$. Moreover, we have

$$\nu_k \overset{(a)}{\leq} \log\Big(1+\frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sigma_0^2}\Big) \overset{(b)}{\leq} \log\Big(1+I\times P^{\max}\frac{\big\|\mathbf{h}_k\big\|_2^2}{\sigma_0^2}\Big) = \overline{\nu}_k, \forall k \in \mathscr{K}. \tag{A I-9}$$

where (a) is due to omitting the inter-user interference, (b) is by applying the result of Cauchy-Schwarz inequality to have $|\mathbf{h}_k^H \mathbf{w}_k|^2 \leq \big\|\mathbf{h}_k\big\|_2^2 \big\|\mathbf{w}_k\big\|_2^2 \leq IP^{\max}\big\|\mathbf{h}_k\big\|_2^2$. Similarly, an upper bound and lower bound of ν_0 can be given by $\nu_0 \geq \frac{1}{\overline{P}} = \underline{\nu}_0; \nu_0 \leq \frac{1}{\sum_{i\in\mathscr{I}} P_i^{\mathrm{ri}}} = \overline{\nu}_0$, where

$$\overline{P} = \sum_{i\in\mathscr{I}} \rho_i \sum_{k\in\mathscr{K}} \overline{\nu}_k + \sum_{i\in\mathscr{I}} \big(\frac{1}{\eta_i} P^{\max} + P_i^{\mathrm{ra}}\big) + \sum_{s\in\mathscr{S}} \big(P_s^{\mathrm{PS}} + \kappa_s C_s^{\alpha_s}\big)\big) \tag{A I-10}$$

7. Proof of Lemma 4

In this section, we show that $\bar{\xi}_k$ in Lemma 4 is a Lipschitz constant of $\nabla R_k(\mathbf{w})$, which is then used to prove Lemma 4. Since there is a mapping rule from a complex-valued vector into a real domain and also due to the space limitation, we will treat $\mathbf{w}$ as a real-valued vector in the following. For ease of mathematical presentation, let us rewrite $R_k(\mathbf{w})$ as

$$R_k(\mathbf{w}) = \log\big(\sigma_0^2 + \mathbf{w}^H \mathbf{H}_k \mathbf{w}\big) - \log\big(\sigma_0^2 + \mathbf{w}^H \mathbf{G}_k \mathbf{w}\big) \tag{A I-11}$$

where $\tilde{\mathbf{H}}_k = \mathbf{h}_k \mathbf{h}_k^H$, $\mathbf{H}_k = \mathrm{blkdiag}(\underbrace{\tilde{\mathbf{H}}_k, \ldots, \tilde{\mathbf{H}}_k}_{K \text{ elements}})$, and $\mathbf{G}_k = \mathrm{blkdiag}(\tilde{\mathbf{H}}_k, \ldots, \underbrace{0}_{k\text{th position}}, \ldots, \tilde{\mathbf{H}}_k)$. Then the gradient of $R_k(\mathbf{w})$ is given by

$$\nabla R_k(\mathbf{w}) = \frac{2\mathbf{w}^H \mathbf{H}_k}{\mathbf{w}^H \mathbf{H}_k \mathbf{w} + \sigma_0^2} - \frac{2\mathbf{w}^H \mathbf{G}_k}{\mathbf{w}^H \mathbf{G}_k \mathbf{w} + \sigma_0^2} = h_1(\mathbf{w}) h_2(\mathbf{w}) - g_1(\mathbf{w}) g_2(\mathbf{w}) \tag{A I-12}$$

where $h_1(\mathbf{w}) = 2\mathbf{w}^H \mathbf{H}_k$, $h_2(\mathbf{w}) = \frac{1}{\mathbf{w}^H \mathbf{H}_k \mathbf{w} + \sigma_0^2}$, $g_1(\mathbf{w}) = 2\mathbf{w}^H \mathbf{G}_k$, and $g_2(\mathbf{w}) = \frac{1}{\mathbf{w}^H \mathbf{G}_k \mathbf{w} + \sigma_0^2}$. Next we will find a Lipschitz constant of each term in (A I-12). From the definition of $h_1(\mathbf{w})$, the

following inequality holds

$$\|h_1(\mathbf{w}) - h_1(\bar{\mathbf{w}})\|_2 = \left\|\left(\mathbf{w}^H - \bar{\mathbf{w}}^H\right)\mathbf{H}_k\right\|_2 \leq \|\mathbf{H}_k\|_F \|(\mathbf{w} - \bar{\mathbf{w}})\|_2 \tag{A I-13}$$

In words, a Lipschitz constant of $h_1(\mathbf{w})$ is $\|\mathbf{H}_k\|_F$. Next we have

$$\begin{aligned}
\|h_2(\mathbf{w}) - h_2(\bar{\mathbf{w}})\|_2 &= \left\| \frac{\bar{\mathbf{w}}^H \mathbf{H}_k \bar{\mathbf{w}} - \mathbf{w}^H \mathbf{H}_k \mathbf{w}}{\left(\mathbf{w}^H \mathbf{H}_k \mathbf{w} + \sigma_0^2\right)\left(\bar{\mathbf{w}}^H \mathbf{H}_k \bar{\mathbf{w}} + \sigma_0^2\right)} \right\|_2 \\
&\leq \left\|\bar{\mathbf{w}}^H \mathbf{H}_k \bar{\mathbf{w}} - \mathbf{w}^H \mathbf{H}_k \mathbf{w}\right\|_2 \\
&\leq \left\|\bar{\mathbf{w}}^H \mathbf{H}_k \bar{\mathbf{w}} - \bar{\mathbf{w}}^H \mathbf{H}_k \mathbf{w} + \bar{\mathbf{w}}^H \mathbf{H}_k \mathbf{w} - \mathbf{w}^H \mathbf{H}_k \mathbf{w}\right\|_2 \\
&\leq \left\|\bar{\mathbf{w}}^H \mathbf{H}_k (\bar{\mathbf{w}} - \mathbf{w})\right\|_2 + \left\|\left(\bar{\mathbf{w}}^H - \mathbf{w}^H\right)\mathbf{H}_k \mathbf{w}\right\|_2 \\
&\leq ||\bar{\mathbf{w}}||_2 \|\mathbf{H}_k\|_F \|\mathbf{w} - \bar{\mathbf{w}}\|_2 + ||\mathbf{w}||_2 \|\mathbf{H}_k\|_F \|\mathbf{w} - \bar{\mathbf{w}}\|_2 \\
&\leq 2P \|\mathbf{H}_k\|_F \|\mathbf{w} - \bar{\mathbf{w}}\|_2
\end{aligned} \tag{A I-14}$$

where $P = \sqrt{IP^{\max}}$. Note that the last inequality occurs due to the sum power constraint. Using (A I-13) and (A I-14) we can find a Lipschitz constant of the product $h_1(\mathbf{w}) h_2(\mathbf{w})$ as

$$\begin{aligned}
&\|h_1(\mathbf{w}) h_2(\mathbf{w}) - h_1(\bar{\mathbf{w}}) h_2(\bar{\mathbf{w}})\|_2 \\
&\leq \|h_2(\mathbf{w})\|_2 \|h_1(\mathbf{w}) - h_1(\bar{\mathbf{w}})\|_2 + \|h_1(\bar{\mathbf{w}})\|_2 \|h_2(\mathbf{w}) - h_2(\bar{\mathbf{w}})\|_2 \\
&\leq \left(\|\mathbf{H}_k\|_F \|h_2(\mathbf{w})\|_2 + 2P \|\mathbf{H}_k\|_F \|h_1(\bar{\mathbf{w}})\|_2\right) \|(\mathbf{w} - \bar{\mathbf{w}})\|_2 \\
&\leq \left(\|\mathbf{H}_k\|_F + \left(2P \|\mathbf{H}_k\|_F\right)^2\right) \|(\mathbf{w} - \bar{\mathbf{w}})\|_2
\end{aligned} \tag{A I-15}$$

We now study the Lipschitz continuity of the term $g_1(\mathbf{w}) g_2(\mathbf{w})$ in (A I-12). Following the same algebraic manipulations the following inequalities are obtained

$$\|g_1(\mathbf{w}) - g_1(\bar{\mathbf{w}})\|_2 = \left\|\left(\mathbf{w}^H - \bar{\mathbf{w}}^H\right)\mathbf{G}_k\right\|_2 \leq \|\mathbf{G}_k\|_F \|(\mathbf{w} - \bar{\mathbf{w}})\|_2 \tag{A I-16}$$

$$\|g_2(\mathbf{w}) - g_2(\bar{\mathbf{w}})\|_2 \leq 2P \|\mathbf{G}_k\|_F \|\mathbf{w} - \bar{\mathbf{w}}\|_2 \tag{A I-17}$$

Thus a Lipschitz constant of $g_1(\mathbf{w})\, g_2(\mathbf{w})$ is simply given by

$$\|g_1(\mathbf{w})\, g_2(\mathbf{w}) - g_1(\bar{\mathbf{w}})\, g_2(\bar{\mathbf{w}})\| \leq \left(\|\mathbf{G}_k\|_F + 4P^2 \|\mathbf{G}_k\|_F^2\right) \|(\mathbf{w}-\bar{\mathbf{w}})\|_2 \tag{A I-18}$$

Combining (A I-15) and (A I-18) results in

$$\begin{aligned}\|\nabla R_k(\mathbf{w}) - \nabla R_k(\bar{\mathbf{w}})\|_2 \leq \|h_1(\mathbf{w})\, h_2(\mathbf{w}) - h_1(\bar{\mathbf{w}})\, h_2(\bar{\mathbf{w}})\|_2 \\ + \|g_1(\mathbf{w})\, g_2(\mathbf{w}) - g_1(\bar{\mathbf{w}})\, g_2(\bar{\mathbf{w}})\|_2 \leq \bar{\xi}_k \|(\mathbf{w}-\bar{\mathbf{w}})\|_2\end{aligned} \tag{A I-19}$$

where

$$\bar{\xi}_k = \|\mathbf{H}_k\|_F + (2P\|\mathbf{H}_k\|_F)^2 + \|\mathbf{G}_k\|_F + (2P\|\mathbf{G}_k\|_F)^2. \tag{A I-20}$$

In other words $R_k(\mathbf{w})$ is a $\bar{\xi}_k$-smooth function.

We now show that $f_k(\mathbf{w}) = R_k(\mathbf{w}) + \xi_k \|\mathbf{w}\|_2^2$ is strongly convex. Since $R_k(\mathbf{w})$ is $\bar{\xi}_k$-smooth, the following inequality holds

$$\left|R_k(\mathbf{w}) - R_k(\bar{\mathbf{w}}) - \nabla R_k(\bar{\mathbf{w}})^T (\mathbf{w}-\bar{\mathbf{w}})\right| \leq \frac{\bar{\xi}_k}{2} \|\mathbf{w}-\bar{\mathbf{w}}\|^2 \tag{A I-21}$$

which implies

$$R_k(\mathbf{w}) \geq -\frac{\bar{\xi}_k}{2} \|\mathbf{w}-\bar{\mathbf{w}}\|^2 + R_k(\bar{\mathbf{w}}) + \nabla R_k(\bar{\mathbf{w}})^T (\mathbf{w}-\bar{\mathbf{w}}) \tag{A I-22}$$

Due to the strong convexity of $\xi_k \|\mathbf{w}\|_2^2$ we have

$$\xi_k \|\mathbf{w}\|_2^2 \geq \xi_k \|\bar{\mathbf{w}}\|_2^2 + 2\xi_k \bar{\mathbf{w}}^T (\mathbf{w}-\bar{\mathbf{w}}) + \frac{\xi_k}{2} \|\mathbf{w}-\bar{\mathbf{w}}\|^2 \tag{A I-23}$$

Combining (A I-22) and (A I-23) we obtain

$$\begin{aligned}R_k(\mathbf{w}) + \xi_k \|\mathbf{w}\|_2^2 \geq \frac{\xi_k - \bar{\xi}_k}{2} \|\mathbf{w}-\bar{\mathbf{w}}\|^2 + R_k(\bar{\mathbf{w}}) + \\ \xi_k \|\bar{\mathbf{w}}\|_2^2 + \Re\left\{(\nabla R_k(\bar{\mathbf{w}}) + 2\xi_k \bar{\mathbf{w}})^T (\mathbf{w}-\bar{\mathbf{w}})\right\}\end{aligned} \tag{A I-24}$$

which is equivalent to

$$f_k(\mathbf{w}) \geq \frac{\xi_k - \bar{\xi}_k}{2} \|\mathbf{w} - \bar{\mathbf{w}}\|^2 + f_k(\bar{\mathbf{w}}) + \nabla f_k(\bar{\mathbf{w}})^T (\mathbf{w} - \bar{\mathbf{w}}) \tag{A I-25}$$

(A I-25) implies that $f_k(\mathbf{w})$ is $(\xi_k - \bar{\xi}_k)$-strongly convex, $\forall \xi_k > \bar{\xi}_k$ which completes the proof.

8. Proof of Lemma 5

Similar to the proof of Lemma 4, $u_k(\mathbf{w}, a_{i,k})$ is strongly convex if $\nabla\big(a_{i,k}R_k(\mathbf{w})\big)$ has a Lipschitz constant of γ_k. First we have $\nabla\big(a_{i,k}R_k(\mathbf{w})\big) = [R_k(\mathbf{w}); a_{i,k}\nabla R_k(\mathbf{w})]$ which leads to

$$\begin{aligned}\big\|\nabla\big(a_{i,k}R_k(\mathbf{w})\big) - \nabla\big(\bar{a}_{i,k}R_k(\bar{\mathbf{w}})\big)\big\|_2^2 &= \big(R_k(\mathbf{w}) - R_k(\bar{\mathbf{w}})\big)^2 \\ &\quad + \big\|a_{i,k}\nabla R_k(\mathbf{w}) - \bar{a}_{i,k}\nabla R_k(\bar{\mathbf{w}})\big\|_2^2 \end{aligned} \tag{A I-26}$$

From (A I-12) we can write

$$\begin{aligned}&\big\|a_{i,k}\nabla R_k(\mathbf{w}) - \bar{a}_{i,k}\nabla R_k(\bar{\mathbf{w}})\big\|_2^2 \\ &= \|a_{i,k}\nabla R_k(\mathbf{w}) - a_{i,k}\nabla R_k(\bar{\mathbf{w}}) + a_{i,k}\nabla R_k(\bar{\mathbf{w}}) - \bar{a}_{i,k}\nabla R_k(\bar{\mathbf{w}})\|_2^2 && \text{(A I-27)} \\ &\leq 2\big(\big\|a_{i,k}\nabla R_k(\mathbf{w}) - a_{i,k}\nabla R_k(\bar{\mathbf{w}})\|_2^2 + \|a_{i,k}\nabla R_k(\bar{\mathbf{w}}) - \bar{a}_{i,k}\nabla R_k(\bar{\mathbf{w}})\big\|_2^2\big) && \text{(A I-28)} \\ &\leq 2\big(|a_{i,k}|^2\big\|\nabla R_k(\mathbf{w}) - \nabla R_k(\bar{\mathbf{w}})\|_2^2 + |a_{i,k} - \bar{a}_{i,k}|^2\|\nabla R_k(\bar{\mathbf{w}})\big\|_2^2\big) && \text{(A I-29)} \\ &\leq 2\big(\big\|\nabla R_k(\mathbf{w}) - \nabla R_k(\bar{\mathbf{w}})\|_2^2 + |a_{i,k} - \bar{a}_{i,k}|^2\|\nabla R_k(\bar{\mathbf{w}})\big\|_2^2\big) && \text{(A I-30)}\end{aligned}$$

Also from (A I-12) the following inequality holds

$$\begin{aligned}\|\nabla R_k(\mathbf{w})\|_2^2 &= \Big\|\frac{2\mathbf{w}^H\mathbf{H}_k}{\mathbf{w}^H\mathbf{H}_k\mathbf{w} + \sigma_0^2} - \frac{2\mathbf{w}^H\mathbf{G}_k}{\mathbf{w}^H\mathbf{G}_k\mathbf{w} + \sigma_0^2}\Big\|_2^2 && \text{(A I-31a)} \\ &\leq 2\Big(\Big\|\frac{2\mathbf{w}^H\mathbf{H}_k}{\mathbf{w}^H\mathbf{H}_k\mathbf{w} + \sigma_0^2}\Big\|_2^2 + \Big\|\frac{2\mathbf{w}^H\mathbf{G}_k}{\mathbf{w}^H\mathbf{G}_k\mathbf{w} + \sigma_0^2}\Big\|_2^2\Big) && \text{(A I-31b)} \\ &\leq 4\left(\|\mathbf{w}^H\mathbf{H}_k\|_2^2 + \|\mathbf{w}^H\mathbf{G}_k\|_2^2\right) && \text{(A I-31c)} \\ &\leq 4\left(\|\mathbf{H}_k\|_F^2 + \|\mathbf{G}_k\|_F^2\right)\|\mathbf{w}\|_2^2 && \text{(A I-31d)}\end{aligned}$$

We now study the Lipschitz continuity of $R_k(\mathbf{w})$. To this end we will show that the function $\log(1+x)$ for $x \geq 0$ is Lipschitz continuous with a Lipschitz constant of 1. That is, for $u \geq 0$ and $v \geq 0$, $|\log(1+u) - \log(1+v)| \leq |u-v|$. Obviously, we only need to prove this for the case $u > v \geq 0$. Since $\log(1+x)$ is continuous and differentiable in the interval $[v,u]$, by the mean-value theorem **?**, there exists $v < x_0 < u$ such that $\frac{1}{1+x_0} = \frac{\log(1+u)-\log(1+v)}{u-v}$, and thus

$$\log(1+u) - \log(1+v) = (u-v)/(1+x_0) \leq u-v \tag{A I-32}$$

Now we have

$$\begin{aligned}
&\left(R_k(\mathbf{w}) - R_k(\bar{\mathbf{w}})\right)^2 \\
&= \left(\log\left(1+\mathbf{w}^H\mathbf{H}_k\mathbf{w}\right) - \log\left(1+\bar{\mathbf{w}}^H\mathbf{H}_k\bar{\mathbf{w}}\right) + \log\left(1+\bar{\mathbf{w}}^H\mathbf{G}_k\bar{\mathbf{w}}\right) - \log\left(1+\mathbf{w}^H\mathbf{G}_k\mathbf{w}\right)\right)^2 && \text{(A I-33a)} \\
&\leq 2\left(\log\left(1+\mathbf{w}^H\mathbf{H}_k\mathbf{w}\right) - \log\left(1+\bar{\mathbf{w}}^H\mathbf{H}_k\bar{\mathbf{w}}\right)\right)^2 + 2\left(\log\left(1+\bar{\mathbf{w}}^H\mathbf{G}_k\bar{\mathbf{w}}\right) - \log\left(1+\mathbf{w}^H\mathbf{G}_k\mathbf{w}\right)\right)^2 && \text{(A I-33b)}
\end{aligned}$$

Applying (A I-32) results in

$$\begin{aligned}
&\left(R_k(\mathbf{w}) - R_k(\bar{\mathbf{w}})\right)^2 \leq 2|\mathbf{w}^H\mathbf{H}\mathbf{w} - \mathbf{w}^H\mathbf{H}\bar{\mathbf{w}} + \mathbf{w}^H\mathbf{H}\bar{\mathbf{w}} - \bar{\mathbf{w}}^H\mathbf{H}\bar{\mathbf{w}}|^2 + \\
&2|\mathbf{w}^H\mathbf{G}_k\mathbf{w} - \mathbf{w}^H\mathbf{G}_k\bar{\mathbf{w}} + \mathbf{w}^H\mathbf{G}_k\bar{\mathbf{w}} - \bar{\mathbf{w}}^H\mathbf{G}_k\bar{\mathbf{w}}|^2 && \text{(A I-34a)} \\
&\leq 4|\mathbf{w}^H\mathbf{H}(\mathbf{w}-\bar{\mathbf{w}})|^2 + 4|(\mathbf{w}^H-\bar{\mathbf{w}})\mathbf{H}\bar{\mathbf{w}}|^2 + 4|\mathbf{w}^H\mathbf{G}_k(\mathbf{w}-\bar{\mathbf{w}})|^2 + 4|(\mathbf{w}^H-\bar{\mathbf{w}})\mathbf{G}_k\bar{\mathbf{w}}|^2 && \text{(A I-34b)} \\
&\leq 8P^2(||\mathbf{H}_k||_F^2 + ||\mathbf{G}_k||_F^2)||\mathbf{w}-\bar{\mathbf{w}}||_2^2 && \text{(A I-34c)}
\end{aligned}$$

Combining (A I-19),(A I-30), (A I-31d) and (A I-34c) we obtain

$$\left\|\nabla(a_{i,k}R_k(\mathbf{w})) - \nabla(\bar{a}_{i,k}R_k(\bar{\mathbf{w}}))\right\|_2 \leq \bar{\gamma}_k\sqrt{|a_{i,k}-\bar{a}_{i,k}|^2 + ||\mathbf{w}-\bar{\mathbf{w}}||_2^2} \tag{A I-35}$$

where

$$\bar{\gamma}_k = \sqrt{2\bar{\xi}_k^2 + 8\left(||\mathbf{H}_k||_F^2 + ||\mathbf{G}_k||_F^2\right)P^2} \tag{A I-36}$$

This completes the proof.

BIBLIOGRAPHY

(2006). *Open Base Station Architecture Initiative (OBSAI) BTS System Reference Document Version 2.0.*

(2011). *ETSI GS ORI 002-1 V1.1.1 (2011-10). Open Radio equipment Interface (ORI); ORI Interface Specification; Part 1: Low Layers (Release 1).*

(2013). *Common Public Radio Interface (CPRI); Interface Specification V6.0.*

(2014). Mosek ApS. Consulted at http://www.mosek.com.

Abid, H., Luong, P., Wang, J., Lee, S. & Qaisar, S. (2011). V-Cloud: vehicular cyber-physical systems and cloud computing. *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL'11)*.

Andrews *et al.*, J. G. (2014). What Will 5G Be? *IEEE J. Sel. Areas Commun.*, 32(6), 1065–1082.

Ariffin, W. N. S. F. W., Zhang, X. & Nakhai, M. R. (2017). Sparse Beamforming for Real-Time Resource Management and Energy Trading in Green C-RAN. *IEEE Trans. Smart Grid*, 8(4), 2022–2031.

Auer *et al.*, G. (2011). How much energy is needed to run a wireless network? *IEEE Wireless Commun.*, 18(5), 40-49.

Ayach, O. E., Rajagopal, S., Abu-Surra, S., Pi, Z. & Heath, R. W. (2014). Spatially Sparse Precoding in Millimeter Wave MIMO Systems. *IEEE Trans. Wireless Commun.*, 13(3), 1499–1513.

Beck, A., Ben-Tal, A. & Tetruashvili, L. (2010). A sequential parametric convex approximation method with applications to nonconvex truss topology design problems. *J. Global Opti.*, 47(1), 29-51.

Bernardos, C. J., Domenico, A. D., Ortin, J., Rost, P. & Wubben, D. (2013). Challenges of Designing Jointly the Backhaul and Radio Access Network in a Cloud-based Mobile Network. *Proc. IEEE Future Netw. and Mobile Summit*, pp. 1–10.

Bhushan *et al.*, N. (2014). Network densification: the dominant theme for wireless evolution into 5G. *IEEE Commun. Mag.*, 52(2), 82-89.

Bidokhti, S. S., Wigger, M. & Yener, A. (2017, Jun.). Benefits of cache assignment on degraded broadcast channels. *IEEE Int. Symp. on Inf. Theory (ISIT 2017)*, pp. 1222–1226.

Bidokhti, S. S., Wigger, M. & Timo, R. (2018). Noisy Broadcast Networks with Receiver Caching. *IEEE Trans. Inf. Theory*, Early Access.

Biermann, T., Scalia, L., Choi, C., Kellerer, W. & Karl, H. (2013). How Backhaul Networks Influence the Feasibility of Coordinated Multipoint in Cellular Networks. *IEEE Commun. Mag.*, 51(8), 168–176.

Boccardi, F., Heath, R. W., Lozano, A., Marzetta, T. L. & Popovski, P. (2014). Five disruptive technology directions for 5G. *IEEE Commun. Mag.*, 52(2), 74–80.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge, UK: Cambridge University Press.

Burke, P. J. (1956). The output of a queuing system. *Oper. Res.*, 4(6), 699-704.

Buzzi *et al.*, S. (2016). A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead. *IEEE J. Sel. Areas Commun.*, 34(4), 697–709.

Candes, E., Wakin, M. & Boyd, S. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5), 877-905.

Chandrasekhar, V., Andrews, J. G. & Gatherer, A. (2008). Femtocell networks: A survey. *IEEE Commun. Mag.*, 46(9), 59-67.

Chen, D., Schedler, S. & Kuehn, V. (2016, Jul.). Backhaul Traffic Balancing and Dynamic Content-Centric Clustering for the Downlink of Fog Radio Access Network. *Proc. IEEE Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC 2016)*, pp. 1–6.

Chen, M., Challita, U., Saad, W., Yin, C. & Debbah, M. (2017). Machine Learning for Wireless Networks with Artificial Intelligence: A Tutorial on Neural Networks. *http://arxiv.org. DOI: arXiv:1606.08950*.

Cheng, Y., Pesavento, M. & Phillip, A. (2013). Joint Network Optimization and Downlink Beamforming for CoMP Transmission Using Mixed Integer Conic Programming. *IEEE Trans. Signal Process.*, 61(16), 3972-3987.

Chin, W. H., Fan, Z. & Haines, R. (2014). Emerging technologies and research challenges for 5G wireless networks. *IEEE Wireless Commun.*, 21(2), 106-112.

Dai, B. & Yu, W. (2014). Sparse Beamforming and User-Centric Clustering for Downlink Cloud Radio Access Network. *IEEE Access*, 31(2), 1326-1339.

Dai, B. & Yu, W. (2016). Energy Efficiency of Downlink Transmission Strategies for Cloud Radio Access Networks. *IEEE J. Sel. Areas Commun.*, PP(99), 1-14.

Dai, B., Liu, Y.-F. & Yu, W. (2018). Optimized Base-Station Cache Allocation for Cloud Radio Access Network with Multicast Backhaul. *IEEE J. Sel. Areas Commun.*, To Appear.

Dhifallah, O., Dahrouj, H., Al-Naffouri, T. Y. & Alouini, M.-S. (2015, Dec). Decentralized Group Sparse Beamforming for Multi-Cloud Radio Access Networks. *Proc. IEEE Global Telecom. Conf. (GLOBECOM 2015)*, pp. 1–6.

Ding, Z. & Poor, H. V. (2013). The use of spatially random base stations in Cloud Radio Access Networks. *IEEE Signal Process. Lett.*, 20(11), 1138–1141.

Dinkelbach, W. (1967). On linear fractional programming. *Bulletin of the Australian Mathematical Society*, 13, 492-498.

Fan, C., Zhang, Y. J. & Yuan, X. (2016). Dynamic nested clustering for parallel PHY-layer processing in cloud-RANs. *IEEE Trans. Wireless Commun.*, 15(3), 1881-1894.

Ge, X., Tu, S., Mao, G., Wang, C.-X. & Han, T. (2016). 5G Ultra-Dense Cellular Networks. *IEEE Wireless Commun. Mag.*, 23(1), 72–79.

Ghanbari, A., Laya, A., A.-Zarate, J. & Markendahl, J. (2017). Business Development in the Internet of Things: A Matter of Vertical Cooperation. *IEEE Commun. Mag.*, 55(2), 135–141.

Guo, K., Sheng, M., Tang, J., Quek, T. Q. & Qiu, Z. (2016a). Exploiting Hybrid Clustering and Computation Provisioning for Green C-RAN. *IEEE J. Sel. Areas Commun.*, 34(12), 4063-4076.

Guo, K., Sheng, M., Tang, J., Quek, T. Q. S. & Qiu, Z. (2016b). Exploiting Hybrid Clustering and Computation Provisioning for Green C-RAN. *IEEE J. Sel. Areas Commun.*, 34(12), 4063–4076.

Gurobi Optimization, I. (2015). Gurobi Optimizer Reference Manual. Consulted at http://www.gurobi.com.

Ha, V. N., Le, L. B. & Đào, N.-D. (2014, Mar.). Energy-Efficient Coordinated Transmission for Cloud-RANs: Algorithm Design and Trade-off. *Conf. on Inf. Sciences and Systems (CISS)*, pp. 1-6.

Ha, V. N., Nguyen, D. H. N. & Le, L. B. (2015, Mar.). Sparse Precoding Design for Cloud-RANs Sum-Rate Maximization. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC)*, pp. 1–6.

Ha, V. N., Le, L. B. & Đào, N.-D. (2016). Coordinated Multipoint (CoMP) Transmission Design for Cloud-RANs with Limited Fronthaul Capacity Constraints. *IEEE Trans. Veh. Technol.*, 65(9), 7432–7447.

Hawilo, H., Shami, A., Mirahmadi, M. & Asal, R. (2014). NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC). *IEEE Network*, 28(6), 18–26.

Hoydis, J., ten Brink, S. & Debbah, M. (2013). Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need? *IEEE J. Sel. Areas Commun.*, 31(2), 160–171.

Hur, S., Kim, T., Love, D. J., Krogmeier, J. V., Thomas, T. A. & Ghosh, A. (2013). Millimeter Wave Beamforming for Wireless Backhaul and Access in Small Cell Networks. *IEEE Trans. Commun.*, 61(10), 4391–4403.

Intelligence, G. (2014). *Understanding 5G: Perspectives on future technological advancements in mobile*. White paper.

Irmer *et al.*, R. (2011). Coordinated Multipoint: Concepts, Performance, and Field Trial Results. *IEEE Commun. Mag.*, 49(2), 102–111.

Jing, S., Tse, D. N. C., Soriaga, J. B., J.Hou, Smee, J. E. & Padovani, R. (2008). Multicell Downlink Capacity with Coordinated Processing. *EURASIP J. Wireless Commun. and Netw.*, 2008, 1-19.

Jungnickel *et al.*, V. (2014). The Role of Small Cells, Coordinated Multipoint, and Massive MIMO in 5G. *IEEE Commun. Mag.*, 52(5), 44–51.

Kansal, A., Zhao, F., Kothari, N. & Bhattacharya, A. A. (2010). Virtual machine power metering and provisioning. *Proc. ACM Symp. on Cloud Computing (SoCC'10)*, pp. 39–50.

Larsson, E. G., Edfors, O., Tufvesson, F. & Marzetta, T. L. (2014). Massive MIMO for next generation wireless systems. *IEEE Commun. Mag.*, 52(2), 186–195.

Lau, V. K. N., Zhang, F. & CuiLow, Y. (2013). Complexity Delay-Constrained Beamforming for Multi-User MIMO Systems With Imperfect CSIT. *IEEE Trans. Signal Process.*, 61(16), 4090-4099.

Lee, N., Heath, R. W., M.-Jimenez, D. & Lozano, A. (2013, Dec.). Base Station Cooperation with Dynamic Clustering in Super-Dense Cloud-RAN. *Proc. IEEE Globecom 2013 Workshop*, pp. 784–788.

Lee *et al.*, D. (2012). Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges. *IEEE Commun. Mag.*, 50(2), 148–155.

Li, J., Wu, J., Peng, M., Wang, W. & Lau, V. K. N. (2015a, Dec.). Queue-aware joint remote radio head activation and beamforming for green cloud radio access networks. *IEEE Global Commun. Conf. (Globecom)*, pp. 1-6.

Li, J., Wu, J., Peng, M. & Zhang, P. (2016). Queue-Aware Energy-Efficient Joint Remote Radio Head Activation and Beamforming in Cloud Radio Access Networks. *IEEE Trans. Wireless Commun.*, 15(6), 3880–3894.

Li, Y., Sheng, M., Wang, C.-X., Wang, X., Shi, Y. & Li, J. (2015b). Throughput-Delay Tradeoff in Interference-Free Wireless Networks With Guaranteed Energy Efficiency. *IEEE Trans. Wireless Commun.*, 14(3), 1608–1621.

Lia, W.-C., Hong, M., Liu, Y.-F. & Luo, Z.-Q. (2014). Base Station Activation and Linear Transceiver Design for Optimal Resource Management in Heterogeneous Networks. *IEEE Trans. Signal Process.*, 62(15), 3939-3952.

Liang, C. & Yu, F. R. (2015). Wireless Network Virtualization: A Survey, Some Research Issues and Challenges. *IEEE Commun. Surveys Tuts.*, 17(1), 358–380.

Li *et al.*, J. (2016). Energy-Efficient Joint Congestion Control and Resource Optimization in Heterogeneous Cloud Radio Access Networks. *IEEE Trans. Veh. Technol.*, 65(12), 9873–9887.

Lin, C.-C., Liu, P. & Wu, J.-J. (2011). Energy-Aware Virtual Machine Dynamic Provision and Scheduling for Cloud Computing. *Proc. IEEE Inter. Conf. Cloud Computing*, pp. 736–737.

Lin, J.-Y., Lee, C.-H. & Tsao, H.-W. (2016, May). On the Optimization of User-Centric Energy-Efficient C-RAN. *IEEE Intern. Conf. Commun. (ICC)*, pp. 1962-1967.

Lipp, T. & Boyd, S. (2016). Variations and extension of the convex–concave procedure. *Opt. and Engineering*, 17(2), 263–287.

Liu, D. & Yang, C. (2016). Energy efficiency of downlink networks with caching at base stations. *IEEE J. Sel. Areas Commun.*, 34(4), 907–922.

Lozano, A., Heath, R. W. & Andrews, J. G. (2013). Fundamental Limits of Cooperation. *IEEE Trans. Inf. Theory*, 59(9), 5213–5226.

Luo, S., Zhang, R. & Lim, T. J. (2015). Downlink and Uplink Energy Minization Through User Association and beamforming in C-RAN. *IEEE Trans. Wireless Commun.*, 14(1), 494-508.

Luong, P., Tran, L.-N., Despins, C. & Gagnon, F. (2016a). Joint Beamforming and Remote Radio Head Selection in Limited Fronthaul C-RAN. *Proc. IEEE Veh. Technol. Conf. (VTC'16 Fall)*, pp. 1-5.

Luong, P., Nguyen, H., Wang, J., Lee, S. & Lee, Y.-K. (2011). Energy Efficiency Based on Quality of Data for Cyber Physical Systems. *International Conference on Internet of Things and International Conference on Cyber, Physical and Social Computing (2011)*, pp. 232–241.

Luong, P., Lee, S. & Lee, Y.-K. (2012). Distributed queuing based-TDMA for real-time service in vehicle to roadside communications. *Proc. 12th International Conference on ITS Telecommunications*, pp. 232–241.

Luong, P., Nguyen, T. M. & Le, L. B. (2014a). Throughput Analysis and Design for Coexisting WLAN and ZigBee Network. *Proc. IEEE Veh. Technol. Conf. (VTC'14 Fall)*, pp. 1-5.

Luong, P., Nguyen, T. M., Le, L. B. & Đào, N.-D. (2014b). Admission control design for integrated WLAN and OFDMA-based cellular networks. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC'14)*, pp. 1–6.

Luong, P., Nguyen, T. M. & Le, L. B. (2016b). Throughput analysis for coexisting IEEE 802.15.4 and 802.11 networks under unsaturated traffic. *EURASIP Journal on Wireless Communications and Networking*, 2016(1), 127–141.

Luong, P., Nguyen, T. M., Le, L. B., Đào, N.-D. & Hossain, E. (2016c). Energy-efficient WiFi offloading and network management in heterogeneous wireless networks. *IEEE Access*, 4, 10210–10227.

Luong, P., Despins, C., Gagnon, F. & Tran, L.-N. (2017a, May). Designing Green C-RAN with Limited Fronthaul via Mixed-Integer Second Order Cone Programming. *Proc. IEEE Int. Conf. Communications (ICC'17)*, pp. 1–6.

Luong, P., Despins, C., Gagnon, F. & Tran, L.-N. (2017b, May). A Fast Converging Algorithm for Limited Fronthaul C-RANs Design: Power and Throughput Trade-off. *Proc. IEEE Int. Conf. Communications (ICC'17)*, pp. 1–6.

Luong, P., Gagnon, F., Despins, C. & Tran, L.-N. (2017c). Optimal Joint Remote Radio Head Selection and Beamforming Design for Limited Fronthaul C-RANs. *IEEE Trans. Signal Process.*, 65(21), 5605–5620.

Luong, P., Despins, C., Gagnon, F. & Tran, L.-N. (2018a). A Novel Energy-Efficient Resource Allocation Approach in Limited Fronthaul Virtualized C-RANs. *Proc. IEEE Veh. Technol. Conf. (VTC'18 Spring)*, pp. 1-6.

Luong, P., Gagnon, F., Despins, C. & Tran, L.-N. (2018b). Joint Virtual Computing and Radio Resource Allocation in Limited Fronthaul Green C-RANs. *IEEE Trans. Wireless Commun.*, 17(4), 2602-2617.

Luong, P., Gagnon, F., Despins, C. & Tran, L.-N. (submitted). Energy-Efficient Beamforming Strategies for Cloud-RANs with Rate-Dependent Power Consumption. *IEEE Trans. Green Commun. and Netw.*

Manosha, K. B. S., Codreanu, M., Rajatheva, N. & Latva-aho, M. (2014). Power–Throughput Tradeoff in MIMO Heterogeneous Networks. *IEEE Trans. Wireless Commun.*, 13(8), 4309-4322.

Marks, B. & Wright, G. (1977). A general inner approximation algorithm for nonconvex mathematical programs. *Oper. Res.*, 26(4), 681–683.

Marler, R. T. & Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optim.*, 26(6), 369–395.

Marsch, P. & Fettweis, G. (2011). Uplink CoMP under a Constrained Backhaul and Imperfect Channel Knowledge. *IEEE Trans. Wireless Commun.*, 10(6), 1730–1742.

Nejad, M. M., Mashayekhy, L. & Grosu, D. (2015). Truthful Greedy Mechanisms for Dynamic Virtual Machine Provisioning and Allocation in Clouds. *IEEE Trans. Parallel Distrib. Syst.*, 26(2), 594-603.

Ng, D. W. K. & Schober, R. (2015). Secure and green SWIPT in distributed antenna networks with limited backhaul capacity. *IEEE Trans. Wireless Commun.*, 14(9), 5082–5097.

Ng, D. W. K., Lo, E. S. & Schober, R. (2016). Multiobjective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer. *IEEE Trans. Veh. Technol.*, 65(5), 3166–3184.

Ng, D. W. K., Lo, E. S. & Schober, R. (2012). Energy-Efficient Resource Allocation in Multi-Cell OFDMA Systems with Limited Backhaul Capacity. *IEEE Trans. Wireless Commun.*, 11(10), 3618–3631.

Nguyen, D., Tran, L.-N., Pirinen, P. & L.-aho, M. (2014a). On the Spectral Efficiency of Full-Duplex Small Cell Wireless Systems. *IEEE Trans. Wireless Commun.*, 13(9), 4896–4910.

Nguyen, D., Tran, L.-N., Pirinen, P. & Latva-aho, M. (2014b). On the spectral efficiency of full-duplex small cell wireless systems. *IEEE Trans. Wireless Commun.*, 13(39), 4896–4910.

Nguyen, T. M., Ajib, W. & Assi, C. A Novel Cooperative NOMA For Designing Unmanned Aerial Vehicle (UAV)–Assisted Wireless Backhaul Networks. *IEEE J. Sel. Areas Commun.*

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2016a). Achieving Energy-Efficiency in Two-Tier Wireless Backhaul HetNets. *Proc. IEEE Int. Conf. Communications (ICC'16)*, pp. 1-6.

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2016b). Resource Allocation in Two-Tier Wireless Backhaul Heterogeneous Networks. *IEEE Trans. Wireless Commun.*, 15(10), 6690-6704.

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2017a). Centralized and Distributed Energy Efficiency Designs in Wireless Backhaul HetNets. *IEEE Trans. Wireless Commun.*, 16(7), 4711-4726.

Nguyen, T. M., Ajib, W. & Assi, C. (2018a). Designing Wireless Backhaul Heterogeneous Networks with Small Cell Buffering. *IEEE Trans. Commun.*

Nguyen, T. M., Ajib, W. & Assi, C. (2018b). A Novel Cooperative Non-Orthogonal Multiple Access (NOMA) in Wireless Backhaul Two-Tier HetNets. *IEEE Trans. Wireless Commun.*, 17(7), 4873–4887.

Nguyen, T. M. & Le, L. B. (2014a). Cognitive spectrum access in femtocell networks exploiting nearest interferer information. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC' 14)*, pp. 1–6.

Nguyen, T. M. & Le, L. B. (2014b). Opportunistic spectrum sharing in Poisson femtocell networks. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC' 14)*, pp. 1–6.

Nguyen, T. M. & Le, L. B. (2015). Joint pilot assignment and resource allocation in multicell massive MIMO network: Throughput and energy efficiency maximization. *Proc. IEEE Wireless Commun. and Net. Conf. (WCNC' 15)*, pp. 1–6.

Nguyen, T. M., Quek, T. Q. S. & Shin, H. (2012a). Opportunistic interference alignment in MIMO femtocell networks. *Proc. IEEE Int. Symp. on Inf. Theory (ISIT' 2012)*, pp. 1–6.

Nguyen, T. M., Shin, H. & Quek, T. Q. S. (2012b). Network throughput and energy efficiency in MIMO femtocells. *Proc.18th European Wireless (EW' 2012)*, pp. 1–6.

Nguyen, T. M., Jeong, Y., Quek, T. Q. S., Tay, W. P. & Shin, H. (2013). Interference alignment in a Poisson field of MIMO femtocells. *IEEE Trans. Wireless Commun.*, 12(6), 2633–2645.

Nguyen, T. M., Ha, V. N. & Le, L. B. (2015). Resource allocation optimization in multi-user multi-cell massive MIMO networks considering pilot contamination. 3, 1272-1287.

Nguyen, T. M., Yadav, A., Ajib, W. & Assi, C. (2016c). Energy Efficiency with Adaptive Decoding Power and Wireless Backhaul Small Cell Selection. *Proc. IEEE Global Telecom. Conf. (GLOBECOM 2016)*.

Nguyen, T. M., Ajib, W. & Assi, C. (2017b). Online Algorithm for Wireless Backhaul HetNets with Advanced Small Cell Buffering. *Proc.26th Int. Conf. on Comp. Comm. and Netw. (ICCCN'17)*, pp. 1–8.

Niu, B., Zhou, Y., Shah-Mansouri, H. & Wong, V. W. S. (2016). A dynamic resource sharing mechanism for cloud radio access networks. *IEEE Trans. Wireless Commun.*, 15(12), 8325–8338.

Pan, C., Zhu, H., Gomes, N. J. & Wang, J. (2017). Joint Precoding and RRH Selection for User-Centric Green MIMO C-RAN. *IEEE Trans. Wireless Commun.*, 16(5), 2891-2906.

Pan, C., Xu, W., Wang, J., Ren, H., Zhang, W., Huang, N. & Chen, M. (2016). Pricing-Based Distributed Energy-Efficient Beamforming for MISO Interference Channels. *IEEE Trans. Veh. Technol.*, 34(4), 710–722.

Pan, C., Mehrpouyan, H., Liu, Y., Elkashlan, M. & Nallanathan, A. (2018). Joint Pilot Allocation and Robust Transmission Design for Ultra-Dense User-Centric TDD C-RAN With Imperfect CSI. *IEEE Trans. Wireless Commun.*, 17(3), 2038–2053.

Parikh, N. & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3).

Park, S.-H., Simeone, O., Sahin, O. & Shamai, S. (2013a). Joint Decompression and Decoding for Cloud Radio Access Networks. *IEEE Signal Process. Lett.*, 20(5), 503–506.

Park, S.-H., Simeone, O., Sahin, O. & Shamai, S. (2013b). Joint Precoding and Multivariate Backhaul Compression for the Downlink of Cloud Radio Access Networks. *IEEE Trans. Signal Process.*, 61(22), 5646–5658.

Park, S.-H., Simeone, O., Sahin, O. & Shamai, S. (2013c). Robust and Efficient Distributed Compression for Cloud Radio Access Networks. *IEEE Trans. Veh. Technol.*, 62(2), 692–703.

Park, S.-H., Lee, K.-J., Song, C. & Lee, I. (2016). Joint Design of Fronthaul and Access Links for C-RAN With Wireless Fronthauling. *IEEE Signal Process. Lett.*, 23(11), 1657–1661.

Parsaeefard, S., R. Dawadi, M. D., L.-Ngoc, T. & Baghani, M. (2017). Dynamic Resource Allocation for Virtualized Wireless Networks in Massive-MIMO-Aided and Fronthaul-Limited C-RAN. *IEEE Trans. Veh. Technol.*, 66(10), 9512-9520.

Peng, M., Yan, S. & Poor, H. V. (2014a). Ergodic Capacity Analysis of Remote Radio Head Associations in Cloud Radio Access Networks. *IEEE Wireless Commun. Lett.*, 3(4), 365–368.

Peng, M., Wang, C., Lau, V. & Poor, H. V. (2015). Fronthaul-Constrained Cloud Radio Access Networks: Insights And Challenges. *IEEE Wireless Commun. Mag.*, 22(2), 152–160.

Peng, M., Yu, Y., Xiang, H. & Poor, H. V. (2016a). Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks. *IEEE Trans. Multimedia*, 8(5), 879–892.

Peng, M., Li, Y., Jiang, J., Li, J. & Wang, C. (2014b). Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies. *IEEE Wireless Commun. Mag.*, 21(6), 126–135.

Peng, M., Sun, Y., Li, X., Mao, Z. & Wang, C. (2016b). Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues. *IEEE Commun. Surveys Tuts.*, 18(3), 2282–2308.

Peng, M., Yu, Y., Xiang, H. & Poor, H. V. (2016c). Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks. *IEEE Trans. Multimedia*, 18(5), 879–892.

Peng *et al.*, M. (2015a). Energy-Efficient resource assignment and power allocation in Heterogeneous Cloud radio access networks. *IEEE Trans. Veh. Technol.*, 64(11), 5275–5287.

Peng *et al.*, M. (2015b). Contract-Based Interference Coordination in Heterogeneous Cloud Radio Access Networks. *IEEE J. Sel. Areas Commun.*, 33(6), 1140–1153.

Peng *et al.*, M. (2016). Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks. *IEEE Trans. Multimedia*, 18(5), 879–892.

Pham, T. D. & Thi, H. A. L. (2014). *Recent Advances in DC Programming and DCA*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Pompili, D., Hajisami, A. & Tran, T. X. (2016). Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN. *IEEE Commun. Mag.*, 54(1), 26–32.

Qualcomm. (2016). *Making 5G NR a Reality*.

Ramamonjison, R., Haghnegahdar, A. & Bhargava, V. K. (2014). Joint Optimization of Clustering and Cooperative Beamforming in Green Cognitive Wireless Networks. *IEEE Trans. Wireless Commun.*, 13(2), 982–997.

Rost *et al.*, P. (2014). Cloud Technologies for Flexible 5G Radio Access Networks. *IEEE Commun. Mag.*, 52(5), 68–76.

Sanderovich, A., Somekh, O., Poor, H. V. & S.Shamai. (2009). Uplink Macro Diversity of Limited Backhaul Cellular Network. *IEEE Trans. Inf. Theory*, 55(8), 3457–3478.

Saxena, N., Roy, A. & Kim, H. (2016). Traffic-Aware Cloud RAN: A Key for Green 5G Networks. *IEEE J. Sel. Areas Commun.*, 34(4), 1010–1021.

Shi, Q., Peng, C., Xu, W., Hong, M. & Cai, Y. (2016a). Energy Efficiency Optimization for MISO SWIPT Systems With Zero-Forcing Beamforming. *IEEE Trans. Signal Process.*, 64(4), 842–854.

Shi, Y., Zhang, J. & Letaief, K. B. (2014). Group Sparse Beamforming for Green Cloud-RAN. *IEEE Trans. Wireless Commun.*, 13(5), 2809-2823.

Shi, Y., Cheng, J., Zhang, J., Bai, B., W.Chen & Letaief, K. B. (2016b). Smoothed ℓ_p Minimization for Green Cloud-RAN With User Admission Control. *IEEE J. Sel. Areas Commun.*, 34(4), 1022–1036.

Shi, Y., Zhang, J. & Letaief, K. B. (2015). Robust Group Sparse Beamforming for Multicast Green Cloud-RAN With Imperfect CSI. *IEEE Trans. Signal Process.*, 63(17), 4647–4659.

Simeone, O., Somekh, O., Poor, H. V. & Shamai, S. (2009). Local Base Station Cooperation Via Finite-Capacity Links for the Uplink of Linear Cellular Networks. *IEEE Wireless Commun.*, 55(1), 190–204.

Simeone, O., Maeder, A., Peng, M., Sahin, O. & Yu, W. (2016). Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems. IEEE J. Commun. Net., 18(2), 135–149.

Sun, Y., Dang, T. & Zhou, J. (2016). User scheduling and cluster formation in Fog Computing based Radio access networks. *Proceedings of ICUWB*, pp. 1-4.

Tang, J., Tay, W. P. & Wen, Y. (2014). Dynamic request redirection and elastic service scaling in cloud-centric media networks. *IEEE Trans. Multimedia*, 16(5), 1434–1445.

Tang, J., Tay, W. P. & Quek, T. Q. S. (2015). Cross-layer resource allocation with elastic service scaling in cloud radio access network. *IEEE Trans. Wireless Commun.*, 14(9), 5068–5081.

Tang, J., Tay, W. P. & Quek, T. Q. S. (2017). System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity. *IEEE Trans. Wireless Commun.*, PP(99), 1–15.

Tervo, O., Tran, L.-N. & Juntti, M. (2015). Optimal Energy-Efficient Transmit Beamforming for Multi-User MISO Downlink. *IEEE Trans. Signal Process.*, 63(20), 5574-5587.

Tolli, A., Pennanen, H. & Komulainen, P. (2011). Decentralized Minimum Power Multi-Cell Beamforming with Limited Backhaul Signaling. *IEEE Trans. Wireless Commun.*, 10(2), 570–580.

Tran, T. X. & Pompili, D. (2017a). Dynamic Radio Cooperation for User-Centric Cloud-RAN With Computing Resource Sharing. *IEEE Trans. Wireless Commun.*, 16(4), 2379–2393.

Tran, T. X., Hajisami, A., Pandey, P. & Pompili, D. (2017). Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges. *IEEE Commun. Mag.*, 55(4), 54-61.

Tran, T. X. & Pompili, D. (2017b). Dynamic Radio Cooperation for User-Centric Cloud-RAN with Computing Resource Sharing. *IEEE Trans. Wireless Commun.*, 16(4), 2379–2393.

Tuy, H. & khayya-and P. Thach, F. A. (2005). Monotonic Optimization: Branch and Cut Methods. *Essays and Surveys in Global Optimization*, 39-78.

Ugur, Y., Awan, Z. H. & Sezgin, A. (2016, Mar.). Cloud Radio Access Networks with Coded Caching. *Int. ITG Workshop on Smart Antennas (WSA 2016)*, pp. 1–6.

Vu, Q.-D., Tran, L.-N., Farrell, R. & Hong, E.-K. (2016). Energy-Efficient Zero-Forcing Precoding Design for Small-Cell Networks. *IEEE Trans. Commun.*, 64(2), 790-804.

Wang, K., Yang, K. & Magurawalage, C. S. (2016a). Joint Energy Minimization and Resource Allocation in C-RAN with Mobile Cloud. *IEEE Trans. Cloud Computing*, pp(99), 1–11.

Wang, X., Thota, S., Tornatore, M., Chung, H. S., Lee, H. H., Park, S. & Mukherjee, B. (2016b). Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN. *IEEE J. Sel. Areas Commun.*, 34(5), 1130–1139.

Wang, Z., Ng, D. W. K., Wong, V. W. S. & Schober, R. (2017). Robust Beamforming Design in C-RAN with Sigmoidal Utility and Capacity-Limited Backhaul. *IEEE Trans. Wireless Commun.*, 16(9), 5583–5598.

Wübben *et al.*, D. (2014). Benefits and Impact of Cloud Computing on 5G Signal Processing:. *IEEE Signal Process. Mag.*, 31(6), 35–44.

Wen, C.-K., Chen, J.-C., Wong, K.-K. & Ting, P. (2014). Message Passing Algorithm for Distributed Downlink Regularized Zero-Forcing Beamforming with Cooperative Base Stations. *IEEE Trans. Wireless Commun.*, 13(5), 2920–2930.

Wiesel, A., Eldar, Y. C. & Shamai, S. (2006). Linear Precoding via Conic Optimization for Fixed MIMO Receivers. *IEEE Trans. Signal Process.*, 54(1), 161–176.

Wong, V. W. S., Schober, R., Ng, D. W. K. & Wang, L.-C. (2017). *Key Technologies for 5G Wireless Systems*. Cambridge, UK: Cambridge University Press.

Wu, J., Zhang, Z., Hong, Y. & Wen, Y. (2015). Cloud Radio Access Network (C-RAN): A Primer. *IEEE Network*, 29(1), 35–41.

Wu, Q. & Liang, Q. (2015). Increasing Capacity of Multi-Cell Cooperative Cellular Networks with Nested Deployment. *Proc. IEEE Int. Conf. Communications (ICC'15)*, pp. 4647-4652.

Wu *et al.*, Y. (2014). Green Transmission Technologies for Balancing the Energy Efficiency and Spectrum Efficiency Trade-off. *IEEE Commun. Mag.*, 52(11), 112-120.

Xiong, K., Fan, P., Lu, Y. & Letaief, K. B. (2016). Energy Efficiency With Proportional Rate Fairness in Multirelay OFDM Networks. *IEEE J. Sel. Areas Commun.*, 34(5), 1431–1447.

Xu, J. & Fortes, J. A. B. (2010). Multi-objective Virtual Machine Placementin Virtualized Data Center Environments. *Proc. IEEE/ACM GREENCOM-CPSCOM*, pp. 179–188.

Yadav, A., Nguyen, T. M. & Ajib, W. (2016a). Optimal Energy Management in Hybrid Energy Small Cell Access Points. *IEEE Trans. Commun.*, 64(12), 5334–5348.

Yadav, A., Nguyen, T. M. & Ajib, W. (2016b). Joint grid energy-throughput optimization for hybrid energy small cell access points. *Proc. IEEE 12th Int. Conf. on Wireless and Mobile Comp., Netw. and Comm. (WiMob) (2016)*, pp. 1-6.

Zakhour, R. & Gesbert, D. (2011). Optimized Data Sharing in Multicell MIMO With Finite Backhaul Capacity. *IEEE Trans. Signal Process.*, 59(12), 6102–6111.

Zhang, Q., Yang, C. & Molisch, A. F. (2013). Downlink Base Station Cooperative Transmission Under Limited-Capacity Backhaul. *IEEE Trans. Wireless Commun.*, 12(8), 3746-3759.

Zhao, J., Quek, T. Q. & Lei, Z. (2013). Coordinated Multipoint Transmission with Limited Backhaul Data Transfer. *IEEE Trans. Wireless Commun.*, 12(6), 2762-2775.

Zhao, W. & Wang, S. (2016). Traffic Density-Based RRH Selection for Power Saving in C-RAN. *IEEE J. Sel. Areas Commun.*, 34(12), 3157–3167.

Zhao, Z., Peng, M., Ding, Z., Wang, C. & Poor, H. V. (2015, June). Cluster Formation in Cloud-Radio Access Networks: Performance Analysis and Algorithms Design. *Proc. IEEE ICC*, pp. 3903–3908.

Zhao *et al.*, Z. (2016). Joint Design of Iterative Training-Based Channel Estimation and Cluster Formation in Cloud-Radio Access Networks. IEEE Access, 4, 9643—9658.

Zhou, H., Tao, M., Chen, E. & Yu, W. (2015). Content-Centric Multicast Beamforming in Cache-Enabled Cloud Radio Access Networks. *Proc. IEEE Global Telecom. Conf. (GLOBECOM 2015)*, pp. 1–6.

Zhou, Y. & Yu, W. (2014). Optimized Backhaul Compression for Uplink Cloud Radio Access Networkk. *IEEE J. Sel. Areas Commun.*, 32(6), 1295–1307.

Zhou, Z., Dong, M., Ota, K., Wang, G. & Yang, L. T. (2016). Energy-Efficient Resource Allocation for D2D Communications Underlaying Cloud-RAN-Based LTE-A Networks. *IEEE Internet of Things Jour.*, 3(3), 428–438.

Zhou *et al.*, L. (2013). Uplink Multicell Processing with Limited Backhaul via Per-Base-Station Successive Interference Cancellation. *IEEE J. Sel. Areas Commun.*, 31(10), 1981–1993.

Zhuang, B., Guo, D. & Honig, M. L. (2016). Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks. *IEEE J. Sel. Areas Commun.*, 34(4), 823–831.