Performance Analysis of Non-Orthogonal Multiple Access (NOMA) in C-RAN, H-CRAN and F-RAN for 5G Systems

A Thesis Submitted to The University of Kent

For The Degree of Doctor of Philosophy

In Electronic Engineering

By

Rupesh Rai

October, 2019

Supervisor(s)

Professor Jiangzhou Wang, Dr. Huiling Zhu

Dedication

I dedicate this work to .... My family and loved ones

Acknowledgements

I would like to thank Prof Jiangzhou Wang for his invaluable contributions to my scientific and personal development. He always encouraged me to move forward, develop myself and take the further step. Without his comments and contributions the work of this thesis could not be achieved.

I would like to thank Dr Huiling Zhu for her insightful and constructive comments and contributions. Her suggestions were always of great help and improved the work of this thesis greatly. Finally, I would like to thank the colleagues in the lab and school for their support and friendly environment.

List of Publications and Submissions

1. Rupesh Rai, Huiling Zhu and Jiangzhou Wang, "Performance Analysis of Non-Orthogonal Multiple Access (NOMA) enabled Cloud Radio Access Networks", submitted to *IEEE Transactions on Wireless Communications.*

2. Rupesh Rai, Huiling Zhu and Jiangzhou Wang, "Coordinated Scheduling and Power Control for Non-Orthogonal Multiple Access (NOMA) enabled Heterogeneous Cloud Radio Access Networks", submitted to *IEEE Transactions on Wireless Communications.*

3. Rupesh Rai, Huiling Zhu and Jiangzhou Wang, "Performance Analysis of Non-Orthogonal Multiple Access (NOMA) enabled Fog Radio Access Networks", submitted to *IEEE Transactions on Wireless Communications.*

4. R. Singh, H. Zhu and J. Wang, "Performance of Non-Orthogonal Multiple Access (NOMA) in a C-RAN System," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, ON, 2017

5. R. S. Rai, "Coordinated Scheduling for Non-Orthogonal Multiple Access (NOMA) in a Cloud-RAN System," in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018, pp. 1-6.

6. R. Singh, "Sub-channel assignment and resource scheduling for non-orthogonal mul-

tiple access (NOMA) in downlink coordinated multi-point systems," in *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, Paris, 2017, pp. 17-22.

7. R. Rai, H. Zhu and J. Wang, "Resource scheduling in non-orthogonal multiple access (NOMA) based cloud-RAN systems," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, New York.

8. R. Rai, "Uplink Power Control for Ultra-Dense C-RAN with Non-Orthogonal Multiple Access," in *2017 Advances in Wireless and Optical Communications (RTUWO)*, Riga, 2017,

Abstract

The world of telecommunication is witnessing a swift transformation towards fifth generation (5G) cellular networks. The future networks present requisite needs in ubiquitous throughput, low latency, and high reliability. They are also envisioned to provide diversified services such as enhanced Mobile BroadBand (eMBB) and ultra-reliable low-latency communication (URLLC) as well as improved quality of user experience. More interestingly, a novel mobile network architecture allowing centralized processing and cloud computing has been proposed as one of the best candidates for fifth generation. It is denoted as Cloud Radio Access Network (C-RAN) and Heterogeneous Cloud Radio Access Network (H-CRAN). Furthermore, the 5G architecture will be fog-like, namely fog radio access networks (F-RAN) enabling a functional split of network functionalities between cloud and edge nodes with caching and fog computing capabilities.

Meanwhile non-orthogonal multiple access (NOMA) has been proposed as an promising multiple access (MA) technology for future radio access networks (RANs) to meet the heterogeneous demands for high throughput, low latency and massive connectivity. One of the main challenges of NOMA is that how well it is to be compatible with other emerging techniques for meeting the requirements of 5G. However, comprehensive performance analysis on NOMA and practical resource allocation designs in co-existence with other emerging networks have not been fully studied and investigated in the literature. This thesis focuses on potential performance enhancement brought by NOMA for the C-RAN, H-CRAN and F-RAN and is expected to address some of the aforementioned key challenges of 5G. The research work of this thesis can be divided into three parts.

In the first part of our research, we focus on investigating the performance analysis of NOMA in a C-RAN. The problem of jointly optimizing user association, muting and power-bandwidth allocation is formulated for NOMA-enabled C-RANs. To solve the mixed integer programming problem, the joint problem is decomposed into two subproblems as 1) user association and muting 2) power-bandwidth allocation optimization. To deal with the first subproblem, we propose a centralized and heuristic algorithm to provide the optimal and suboptimal solutions to the remote radio head (RRH) muting problem for given bandwidth and transmit power, respectively. The second subproblem is then reformulated and we propose an optimal solution to bandwidth and power allocation subject to users data rate constraints. Moreover, for given user association and muting states, the optimal power allocation is derived in a closed-form. Simulation results show that the proposed NOMA-enabled C-RAN outperforms orthogonal multiple access (OMA)-based C-RANs in terms of total achievable rate, interference mitigation and can achieve significant fairness improvement.

Our second work investigates the performance of NOMA in H-CRAN, where coordination of macro base station (MBS) and remote radio heads (RRHs) for H-CRAN with NOMA is introduced to improve network performance. We formulate the problem of jointly optimizing user association, coordinated scheduling and power allocation for NOMA-enabled H-CRANs. To efficiently solve this problem, we decompose the joint optimization problem into two subproblems as 1) user association and scheduling 2) power allocation optimization. Firstly the users are divided based on different interference they suffer. This interference-aware NOMA approach account for the inter-tier interference. Proportional fairness (PF) scheduling for NOMA is

utilized to schedule users with a two-loop optimization method to enhance throughput and fairness. Based on the user scheduling scheme, optimal power allocation optimization is performed by the hierarchical decomposition approach. It is then followed by algorithm for joint scheduling and power allocation. Simulation results show that the proposed NOMA-enabled H-CRAN outperforms OMA-based H-CRANs in terms of total achievable rate and can achieve significant fairness improvement.

In the third part of our research, we propose a NOMA-enabled fog-cloud structure in a novel density-aware F-RAN to tackle different aspects such as throughput and latency requirements of high and low user-density regions, in order to meet the heterogeneous requirements of eMBB and URLLC traffic. A framework of the multi-objective problem is formulated to cater the high throughput and low-latency requirements in a high and low user-density mode respectively. In the first problem, we study the joint caching placement and association strategy aiming at minimizing the average delay. To deal with the first problem, we apply McCormick envelopes and Lagrange partial relaxation method to transform it into three convex sub-problems, which is then solved by proposed distributed algorithm. The second problem is to jointly optimize transmission mode selection, subchannel assignment and power allocation to maximize the sum data rate of all fog user equipments (F-UEs) while satisfying fronthaul capacity and fog-computing access point (F-AP) power constraints. Moreover, for given transmission mode selection and subchannel assignment, the optimal power allocation is derived in a closed-form. Simulation results are provided for the proposed NOMA-enabled F-RAN framework and reveal that the ultra-low latency and high throughput can be achieved by properly utilizing

the available resources.

Contents

List of Figures

List of Tables

Abbreviations

3GPP ..... Third Generation Partnership Project

5G ..... Fifth Generation

AP ..... Access Point

AWGN ..... Additive White Gaussian Noise

BBU ..... Baseband Unit

CCU ..... Cell Center User

CDF ..... Cumulative Distribution Function

CEU ..... Cell Edge User

CoMP ..... Coordinated Multipoint

CPU ..... Central Processing Unit

C-RAN ..... Cloud Radio Access Network

CSI ..... Channel State Information

D2D ..... Device-to-Device

DAS ..... Distributed Antennas System

EB ..... Exabytes

eMBB ..... Enhanced Mobile Broad Band

F-AP ..... Fog-computing Access Point

F-RAN ..... Fog Radio Access Network

H-CRAN ... Heterogeneous Cloud Radio Access Network

ICI ..... Inter-Cell Interference

i.i.d ..... Independent and Identically Distributed

IoT ..... Internet of Things

KKT ..... Karush-Kuhn-Tucker

LTE ..... Long Term Evolution

M2M ..... Machine to Machine

MBS ..... Macro Base Station

MIMO ..... Multiple-Input Multiple-Output

MMSE ..... Minimum Mean Squared Error

MUD ..... Multi-User Detection

NOMA ..... Non-Orthogonal Multiple Access

OFDMA .... Orthogonal Frequency Division Multiple Access

QoS ..... Quality of Service

RB ..... Resource Block

RRH ..... Remote Radio Head

SC ..... Superposition Coding

SE ..... Spectral Efficiency

SIC ..... Successive Interference Cancelation

SINR ..... Signal-to-interference-plus-noise Ratio

TDD ..... Time-division Duplexing

UE ..... User-equipment

URLLC ..... Ultra-Reliable Low-Latency Communication

List of Notations

$\mathcal{R}$	Set of all RRHs in the network
$\mathcal{N}$	Set of all users in the network
P_k^r	Maximum transmit power of RRH r
B_{ij}	Portion of the entire downlink bandwidth allocated to the NOMA pair (i, j)
γ_{nk}^u	Received signal-to-interference-plus-noise ratio (SINR) for the i-th decoded user on subchannel k
β_{rk}	Muting arrangement matrix of dimensions $R \times K$
R_{CS}	Throughput that the user n can obtain with coordinated scheduling when associated with RRH r
b_{rn}	Achievable rate on RB k and its dependence on muting indicator
β_{rk}	User-RRH association
$\mathcal{R}_m$	Total number of MBS and RRHs
p^m, p^r	Transmit power from MBS and RRH to nth MUE and bth RUE respectively
$x_{r,k}^{b_1}$	The symbols transmitted from rth RRH to its serving RUE b_1

$\gamma_{r,k}^{b}$	The received SINR at RUE b served by RRH r on RB k
$\gamma_{m,k}^{n}$	The received SINR of MUE n considering interference from RRH
γ_{cs}	Received SINR of UEu in the MBS+RRH range
$U(R(t), \alpha)$	The network-level system throughput
$U_R(C_i)$	Utility function of MBS or RRH
$\mathcal{N}$	Index of F-UEs directly served by MBS
$\mathcal{F}$	Set of F-APs in a macrocell
S_i^{max}	Cache capacity
p^m, p^f	Transmit power from MBS and F-AP to mth UE and nth F-UE respectively
$\gamma_{f,k}^{n}$	The received SINR at F-UE n served by F-AP considering cross-tier interference
R_n^k	The delivery rate of user n served by F-AP f for non-cooperative transmission
R_n^{ck}	The delivery rate of user n served by F-AP f for cooperative transmission
R_n^k	Achievable rate of user n
R_{ac}	Achievable data rate between controlling F-AP and the coordinated F-APs
α^i	Binary association variable between UE and cache content c
$c_{f'f}^{n}$	Coordinated F-APs and subchannel assignment for UE
G_f	Total number of computing tasks
L_jS	Amount of requested contents served by single F-AP to UE n

x_{fn} Binary association variable between UE and F-AP

P_t Detection threshold required to distinguish between the signal at the SIC receiver

Chapter 1

Introduction

1.1 Motivation

The demands for high data rates and efficient use of limited radio-frequency spectrum resources and radio network infrastructure have been motivating the rapid evolution in mobile networks. According to the Ericsson Mobility report, for the near future, the total mobile data traffic will experience a 30 percent compound annual growth rate (CAGR) between 2018 and 2024 to reach 131 exabytes (EB) per month by the end of 2024 as shown in Figure 1.1 [1]. Meeting this growing amount of data traffic is a critical issue for wireless systems. While the fourth generation (4G) mobile systems have approached theoretical capacity, the fifth generation (5G) mobile networks is to be launched in 2020 to deal with growing expectations of quality of services (QoS) [2]. Another significant requirement that customers anticipate is the availability of equally good services anywhere [3]. The demand for wireless data connectivity is an additional challenge that will continue to increase in the near future. The fully networked society will be a feature of the 5G mobile networks, where all electronic devices are connected to the Internet as well as communications, such

as machine-to-machine (M2M), Internet of things (IoT), etc, need to be supported besides personal communications. Moreover, on the basis of the Ericsson Mobility Report, the number of connected devices including the IoT, and other devices such as mobile phones, is expected to reach 31.4 billion in 2024. Therefore, it is essential to deal with all these intense demands through designing new revolutionary wireless network technologies.

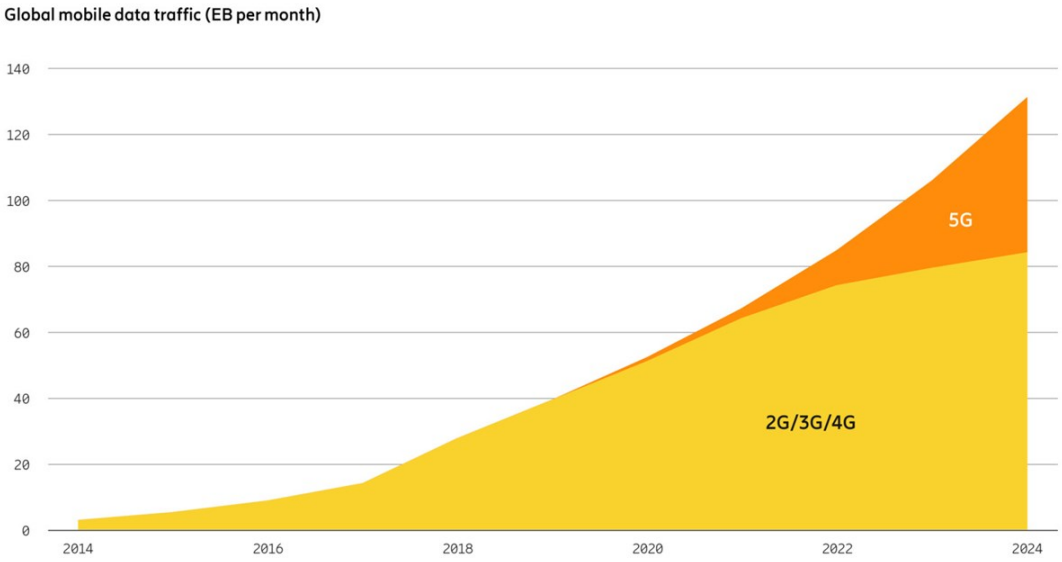

Figure 1.1: Global mobile data traffic (EB per month)

To deal with the above mentioned challenges in the forthcoming 5G wireless networks, compelling new technologies, such as massive multiple-input multiple-output (MIMO) [4], non-orthogonal multiple access (NOMA) [5-7], millimeter wave (mmWave) communications [8], cloud radio access networks (C-RAN) [9], heterogeneous cloud radio access networks (H-CRAN) [10] and fog radio access networks (F-RAN) [11] etc. have been proposed. In particular, wireless access technologies for 5G and beyond 5G networks, will be more flexible, reliable, and efficient in terms of energy and spectrum [12-16]. In fact, multiple access technology is the most fundamental

aspect in physical layer and it significantly affects the whole system performance in each generation of wireless networks. As such, this thesis focuses on the potential performance enhancement brought by NOMA for the C-RAN, H-CRAN and F-RAN that is expected to address some of the aforementioned key challenges of 5G.

The motivation behind considering C-RAN architecture is that the traditional distributed networks lacks in global network information while with only local network information, scheduling can be performed sub-optimally. Therefore many tasks of scheduling can benefit from a centralized global perspective. In the considered C-RAN architecture, a central processor (cloud) is responsible for scheduling users. The centralized scheduling approach offers the highest spectral efficiency (SE). The SE with a decentralized approach is lower than what is achieved with a centralized approach, due to the lack of global information.

In contrast to orthogonal multiple access (OMA), NOMA allows allocating one frequency channel to multiple users at the same time within the same cell. NOMA offers number of advantages, including improved SE, higher cell-edge user throughput, massive connectivity, low latency, and higher user fairness [17].

Inspired by the aforementioned potential benefits of NOMA and C-RAN, we therefore explore the potential performance enhancement brought by NOMA for the C-RANs. Although a substantial prior work in the available literature has investigated C-RAN and NOMA, as discussed above, multiple access techniques and in particular NOMA, which are of great importance in C-RANs for interference mitigation and SE improvement has not been fully explored.

Moreover, H-CRANs is considered as a promising solution to meet the huge traffic demands of fifth generation (5G) networks. To further enhance the C-RAN concept,

H-CRAN [10] has been proposed to decouple the control plane and the user plane, in which functions of control plane are implemented in macro base stations (MBSs) or high power nodes (HPNs) whereas remote radio heads (RRHs) are deployed to provide high data rates for users with diverse quality of service (QoS) requirements. The capacity and time delay constraints on the fronthaul is alleviated by shifting the delivery of control and broadcast signalling from RRH to MBS.

Centralized baseband processing facilitate better coordination among MBS and RRHs as the cell-site information like channel conditions, user requirements and traffic loads are available across the network. Such information can be used for optimization of radio resource allocation, manage inter-cell interference and improve coverage.

To meet high traffic requirements in dense user deployments, 5G will have to provide enhanced Mobile BroadBand (eMBB) services that will share radio and computational resources with Ultra-Reliable Low-Latency Communication (URLLC). Moreover the network architecture will evolve from a traditional base station-centric architecture to a F-RAN architecture, which enhances the C-RAN architecture by allowing fog access points (F-APs) to be equipped with storage capacity and signal processing functionalities [11,18]. F-RAN will enable network functionalities to be distributed among F-APs and cloud depending on their latency and reliability requirements. Compared with the pure C-RAN mode, the adaptive mode selection in F-RANs can significantly improve SE and decrease latency. In order to achieve high SE and low latency, a UE may be served by the local F-APs, which cache the desired content, instead of being served through the fronthaul connection.

In a dense F-RAN, providing all users with higher throughput and lower power

consumption is critical whereas in low user-density region, latency is an important parameter to be considered as an objective function. In low user-density region F-APs are deployed in a sparse manner, which leads to high transmission delay between F-APs and cloud. Moreover, the accompanying inter-F-AP interference becomes the fundamental challenge for an effective resource management scheme. We assume that the users in low-density region to have URLLC transmissions only from users with high average channel gain to the F-APs. This condition ensures the high reliability requirement of URLLC traffic. It is assumed that in the high-density region eMBB users do not guarantee this condition. The users are assumed to have non-negligible channel gains to all F-APs. In this work we evaluate the performance of both eMBB and URLLC traffic considering high user-density and low user-density respectively in NOMA-enabled F-RAN system. Due to the limited storage and computation capacity in the fog node, the implementation of NOMA scheme will have great impact on high-throughput and low-latency requirements in F-RAN. In contrast to orthogonal multiple access (OMA), NOMA not only improves the reliability of content delivery by pushing multiple files to F-APs simultaneously but also ensures more user file requests can be served concurrently by the F-APs. While in OMA, only most popular file can be pushed during a single time slot, in NOMA, multiple files can be pushed to F-APs and users simultaneously, which leads to the efficient use of limited resources reserved for content pushing. We consider that the request files are cooperatively processed in the fog and cloud before being conveyed to the user terminals. If the request files are stored in both cloud and F-AP, one part will be processed in the F-AP, and the remaining part will be processed in the cloud. If the request files only existed in the cloud, the files will be processed totally in the cloud resulting in increased latency. The NOMA principle enables the

MBS to perform content delivery and content pushing simultaneously, i.e., it can push new content to the F-APs while serving UEs directly in case the requested files are not found at the F-APs.

1.2 Contribution of the Thesis

The key contributions of this thesis are summarized as follows:

(a) We propose a NOMA-enabled C-RAN model in which NOMA technology is utilized for spectrum efficiency enhancement and user access improvement. Based on the proposed model, we formulate a joint user association, muting and bandwidth power allocation (BPA) with the aim of maximizing UE's sum rate and network utility while considering users' fairness issues. The formulation problem from (a) belongs to the mixed-integer non-linear programming (MINLIP) class of optimisation problems with high complexity. We relax the integer constraints and then decompose the joint optimization problem into two subproblems. We first solve the user association and muting problem under fixed BPA. We first study the user association (UA) strategy under given number of active RRHs. A semi-distributed algorithm is proposed to find an efficient user association solution based on the Lagrangian dual analysis. Based on the given UA solution, we propose a centralized muting algorithm which updates the RRHs muting states using the subgradient method. We also propose a heuristic algorithm to find the muting states which improves the cell-edge users' performance and overall system performance using the Jain's fairness index. We propose an adaptive resource allocation strategy that minimizes the total transmit power by following two strategies: a) reducing the num-

ber of active RRHs by employing the key idea of coordinated silencing (RRH muting). b) minimizing total transmit power of all RRHs while satisfying the data rate requirements of all users. Under the proposed user association and muting schemes, we propose a BPA problem which aims at assigning feasible bandwidth and minimizing the required power. Based on the hierarchical decomposition method the BPA approach iteratively updates the bandwidth allocation to maximize the network utilisation. For a given bandwidth allowance, the optimal power allocation for RRH is formulated as a non-convex problem which is solved by transforming it into a convex problem and applying the Karush-Kuhn-Tucker (KKT) conditions. Based on the transformed Lagrangian function, the optimal power allocation is derived in closed form subject to QoS constraints. Finally, we evaluate the performance of our proposed framework for joint optimization problem for NOMA-enabled C-RAN systems via simulation to validate that our proposed algorithms can obtain the optimal solution of the joint optimization problem in a significantly reduced computational time and show that NOMA can greatly improve network performance in both data rate and network utility with proportional fairness consideration. Additionally, we present numerical results to show that our proposed joint channel bandwidth and power allocations for NOMA-enabled C-RAN transmission can significantly minimize the total RRH transmission power considering the bandwidth constraint in comparison with the conventional OMA-enabled C-RAN transmission scheme as well as the corresponding fixed BPA scheme.

(b) We propose a NOMA-enabled H-CRAN model in which NOMA technology

is utilized for spectrum efficiency enhancement and user access improvement. Based on the proposed model, we formulate a joint user association, resource allocation and scheduling, and power allocation with the aim of maximizing UEs sum rate while considering users' fairness issues. The problem belongs to the mixed-integer non-linear programming (MINLIP) class of optimisation problems with high complexity. We first propose a two loop optimization algorithm to jointly solve the user association and scheduling problem under fixed transmit powers. Based on the proposed joint user association and coordinated scheduling scheme, iterative power allocation scheme is proposed for the NOMA H-CRAN and the optimal power allocation for each user is derived by hierarchical decomposition method. We demonstrate that the proposed iterative algorithm to optimize the schedule and power iteratively increases the average data rate and network utility with proportional fairness consideration for NOMA-enabled H-CRAN as compared to OMA H-CRAN.

(c) We propose a NOMA-enabled fog-cloud structure in a F-RAN system to tackle different aspects such as throughput and latency of high and low user-density regions, in order to meet the heterogeneous requirements of eMBB and URLLC traffic. A framework of the multi-objective problem is formulated to cater the high throughput and low-latency requirements in a high and low user-density mode respectively. In the first problem, we study the joint caching placement and association strategy aiming at minimizing the average delay. The average delay function is formulated for cooperative and non-cooperative transmission modes, cache placement subject to F-AP storage capacity and maximum number of cooperating F-APs constraints. Since the formulated problem belongs to

integer non-linear optimization problem, we apply McCormick envelopes and the Lagrange partial relaxation method to transform it into three convex subproblems, which are then solved by the proposed distributed algorithm. The second problem is to maximize the sum-rate of the macro-cell based transmission and F-AP based transmission. We consider that the request files of the UEs are processed collaboratively in both cloud and F-APs. Moreover by using NOMA principle, additional files are pushed to the F-APs simultaneously serving the users' requests by making efficient use of the available resources. We tackle the problem of jointly optimizing transmission mode selection, subchannel assignment and power allocation to maximize the sum data rate of all F-UEs while satisfying fronthaul rate and F-AP power constraints. We evaluate the performance of our proposed framework for multi-objective optimization problem for NOMA-enabled F-RAN systems via simulation to validate that our proposed algorithms has a low complexity and can obtain the near-optimal solution and show that NOMA can greatly improve network performance in both throughput and latency.

1.3 Thesis Outline

The structure of this thesis is based on six chapters and appendices as follows:

In Chapter 1, the motivation and the main contributions of this thesis are summarised, and the thesis contents are outlined.

In Chapter 2, the background theory of the thesis is presented, including the fundamental concepts of NOMA, the current state of the art on NOMA research and its coexistence with C-RAN, H-CRAN and F-RAN.

In Chapter 3, we present the system description and channel model for NOMA-enabled C-RAN. The optimization problem is formulated and decomposed into sub-problems. Thereafter we discuss the proposed iterative user association and muting problem. Following that, the optimal bandwidth and power allocation methods are investigated. Simulation results to validate that our proposed algorithms can obtain the optimal solution are presented.

In Chapter 4, the signal transmission model of NOMA enabled H-CRAN is presented. Following that we investigate the user association methods. Thereafter, we describe the proportional fairness (PF) scheduling policy followed by investigation into the power control problem. Simulation results to demonstrate that the proposed algorithm to optimize the scheduling and power iteratively with proportional fairness consideration are presented.

In Chapter 5, we present the system description and channel model for NOMA-enabled F-RAN. The optimization problems are then formulated. Following that the transmission modes and interference-aware NOMA-enabled F-RAN are discussed. Thereafter we investigate the solution to sum-rate maximization problem. Finally, simulation results are presented to assess the proposed algorithms and to show the impact of various parameters on the performance of the proposed schemes.

In Chapter 6, the conclusions of this thesis are drawn, and future research directions are discussed.

The Appendices provided at the end of the document include parameter definitions and specific formula derivations.

A list of the related publications to this work is provided on page v.

Chapter 2

Theoretical Background and Literature Review

In this chapter, the background theory relating to the technical content in this Thesis is provided. The basic principles of non-orthogonal multiple access (NOMA) are introduced, especially the coexistence of NOMA with other technologies, which lays a solid foundation for the technical works in Chapters 3, 4 and 5. The basic concept of NOMA, C-RAN, H-CRAN and F-RAN are introduced and literature review of NOMA and its co-existence with other 5G key enabling technologies are presented. Moreover, the technical challenges when applying NOMA to C-RAN, H-CRAN and F-RAN are discussed. Finally, the present standardization status of NOMA in 5G is discussed.

2.1 Fundamental Concepts of NOMA

In this section, we first introduce two key enabling technologies for NOMA [18,19], superposition coding (SC) and successive interference cancellation (SIC). Then, the

basic concept and system model for downlink NOMA are presented. Finally, the multi-user capacity regions for downlink NOMA is discussed.

2.1.1 Superposition Coding

In this subsection, we introduce the concepts of SC in the context of downlink communications. The SC was first proposed in [21] as a technique in which a single source is responsible for simultaneous communicating information to several receivers. SC allows the transmitter to transmit multiple users' information at the same time. SC is an effective technique to increase capacity in the NOMA system [22]. SC is a coding technique in which multiplexing of symbols is performed in power domain, where a symbol is an output from a multi-level symbol mapper such as quadrature amplitude multiplexing (QAM). In order to illustrate how SC is performed, a schematic diagram is given in Fig. 2.1, where QAM constellation of user 1 and user 2 with low transmit power. [23] proposed a design technique for SC by using single user coding and decoding blocks. During the superposition coding

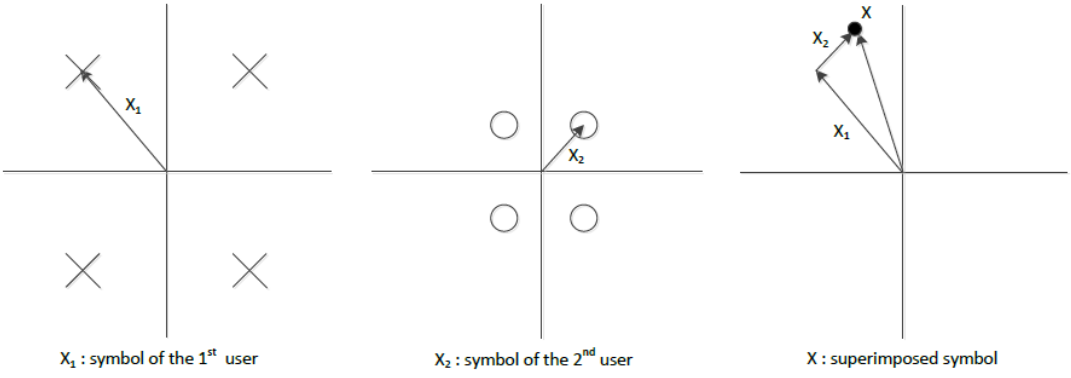

Figure 2.1: SC of two 4-QAM signals [24]

phase, two point-to-point encoders, $f_1 : \{0,1\}^{2^{TR_1^k}} \rightarrow C^T$ and $f_2 : \{0,1\}^{2^{TR_2^k}} \rightarrow C^T$

first map the input bits to output bit sequences x_1 and x_2, respectively. R_1^k and R_2^k represents the transmission rates and T denotes the block length. C is the code library and [.] denotes the floor operator. The output sequence is given by:

$$X(n) = \sqrt{P^k \beta_1} x_1 + \sqrt{P^k \beta_2} x_2 \tag{2.1}$$

where β_i represents the fraction of the total power P assigned to user i, subject to the constraint $\beta_1 + \beta_2 = 1$. At the transmitter side the total power allocated to all U users is limited to P, and the base station (BS) transmits the signal x_u to the uth user subjected to the power-scaling coefficient P_k where k is the resource block (RB). In other words, the signals intended for different users are weighted by different power-scaling coefficients and then they are superimposed at the BS with different power levels. Without loss of generality, the channel gains of users are assumed to be w.r.t. a particular ordering. The user with a better channel gain is usually called a strong user, while the user with a worse channel gain is called a weak user [25]. The transmit powers for the strong and weak users are allocated in accordance with their channel gain order.

2.1.2 Successive Interference Cancellation

In order to provide fairness and to perform the SIC decoding, the transmitter usually allocates more power to the weak user with a poor channel condition. At the receiver, SIC decoding is employed to exploit the disparity in channel gains and transmit powers. In SIC user signals are successively decoded. After one users' signal is decoded it is subtracted from the combined signal before the next user's signal is decoded. When SIC is applied, one of the user signals is decoded, treating the signal of other user as interferer. Prior to SIC, users are ordered in accordance with their

signal strenghts, so that the receiver can decode the user having the strongest signal first, subtract it from the composite signal, and isolate the user with weak signal from the residue. As a result, the interference can be successively removed and the achievable data rate is improved. In order to mitigate the inter-user interference for downlink communications, users with better channel conditions can perform SIC. It has been demonstrated that SIC is capable of reaching the region boundaries of Shannon capacity, both in terms of the broadcast channel and multiple access networks.

Fig. 2.2 illustrates the SC-SIC detection where the constellation point of user 1 is decoded first, removed from the received superposed signal and the constellation point of user 2 is decoded. The concept of SC-SIC can be extended to more users by multiplexing more symbols. SC-SIC can achieve data rate to multiple users simultaneously without the need to expand the bandwidth which is favourable to increase the channel capacity.

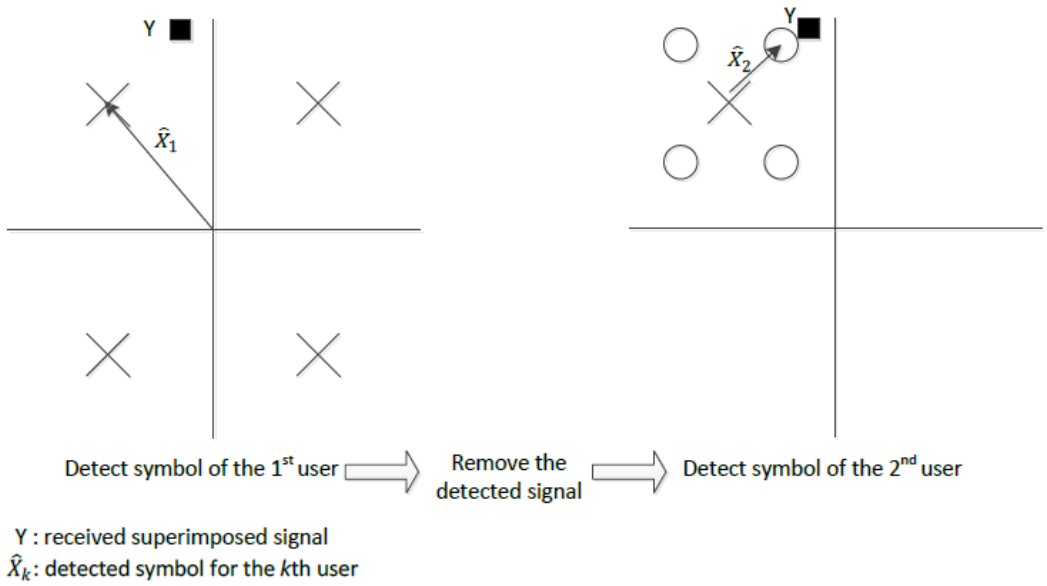

Figure 2.2: SC-SIC detection of two 4-QAM signals [24]

In brief the particular process involved in decoding the superposed messages can be mathematically expressed as follows [23]:

1) At user 1, a single user decoder $g_1 : C^T \rightarrow \{0,1\}^{2^{TR_1^k}}$ decodes the message x_1 by treating x_2 as noise.

2) User 2 performs the following steps to successively recover its message from its received signal $Y_2(n)$:

a) Decode user 1 message x_1 by using a single user decoder $g_1 : C_T \rightarrow \{0,1\}^{2^{TR_1^k}}$

b) Subtract $\sqrt{P^k\beta_1}h_2x_1$ from the received signal $Y_2(n)$

$$Y_2^{'}(n) = Y_2(n) - \sqrt{P^k\beta_1}h_2x_1 \tag{2.2}$$

where h_2 is the channel gain for user 2.

c) Decode user 2 message x_2 by applying another single-user decoder $g_2 : C_T \rightarrow \{0,1\}^{2^{TR_2^k}}$ on $\mathrm{Y}_2^{'}(n)$.

2.1.3 Downlink NOMA

In this chapter, to facilitate the presentation for the basic concepts of NOMA, we consider a simple single-carrier, two-user downlink NOMA system. The generalization to the case of multi-carrier and multi-user communications will be presented in the following chapters when necessary. The generic system model for downlink NOMA is shown in Figure 2.1. with one BS and two users.

Since the system employs multiplexing of users through superposition coding by using NOMA technique [20,25], a single RB can be assigned to multiple users. In this thesis NOMA study assumes a group size of two. Grouping more than two users provides better performance but increases processing complexity for successive

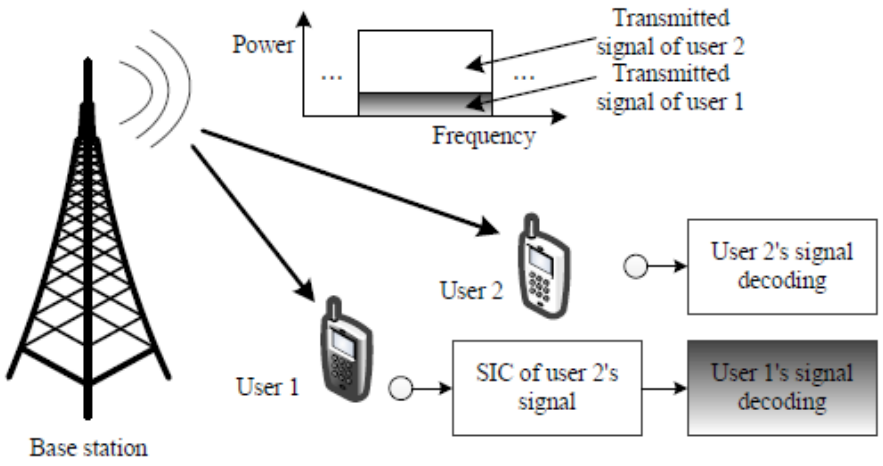

Figure 2.3: The downlink NOMA system model with one BS and two users.

interference cancellation (SIC) receivers [26]. Intra-group interference is mitigated by the NOMA principle [27].

In NOMA principle, higher power is allocated to far users. We assume that the base station transmits the messages of both UE 1 and UE 2, i.e., x_1 and x_2 on the k-th resource block with a total transmission power P_k^b. The powers allocated to users UE_1 and UE_2 are expressed as $P_1^k = \epsilon P_k^b$ and $P_2^k = (1-\epsilon)P_k^b$ respectively, where ϵ is a power allocation variable. The corresponding transmitted signal is given by

$$x = \sqrt{P_1^k} x_1 + \sqrt{P_2^k x_2} \tag{2.3}$$

The received signal at user u is given by

$$y_u = h_u x + n_u, \ u = \{1, 2\} \tag{2.4}$$

where h_u denotes the channel coefficient between user u and BS including joint effects of small and large-scale fading. n_u denotes the additive white Gaussian noise (AWGN) at user u with a noise power of σ_u^2, i.e., $n_u \sim \mathcal{CN}(0, \sigma_n^2)$.

Then the signals transmitted for NOMA users on the k-th RB are ordered based on their channel gain as $\frac{|h_1^k|^2}{\sigma_1^2} \geq \frac{|h_2^k|^2}{\sigma_2^2}$. According to this order for each RB, one user can successfully decode the signal of the other user whose decoding order is lower. The achievable rates of UE_1 and UE_2 on k-th RB can be expressed as:

$$R_1^k = log_2\left(1 + \frac{|h_1|^2 P_1^k}{\sigma_1^2}\right) \tag{2.5}$$

and

$$R_2^k = log_2\left(1 + \frac{|h_2|^2 P_2^k}{P_1^k |h_2|^2 + \sigma_2^2}\right) \tag{2.6}$$

Thus, by tuning power allocation coefficients, the BS can adjust the data rate of each user. The NOMA principle makes a full use of the channel gain differences among the users, which implies that the near-far effect is effectively harnessed to achieve higher SE. As a result, both the attainable sum capacity and the cell-edge user data rate can be improved [25].

Note that SIC is unable to eliminate the interference caused by user u_1 for user u_2. Fortunately, if larger power is allocated to user u_2 as compared to user u_1 in the aggregate received signal y_2, it does not introduce much performance degradation compared to allocating user u_2 on this RB exclusively. The achievable rate region of downlink NOMA is shown in Fig. 2.2 in comparison with that of orthogonal multiple access (OMA).

It can be observed from Fig. 2.2 that the achievable rate region of OMA is only a subset of that of NOMA. Consequently, NOMA provides a higher flexibility in resource allocation in order to improve the system SE, especially considering the diverse quality of service (QoS) requirements of users.

Table 2.1: Architecture and qualitative comparison between C-RAN, H-CRAN, and F-RAN in 5G mobile networks

	C-RAN	H-CRAN	F-RAN
Storage	Centralized	Centralized	Centralized and Distributed
Caching	Centralized	Centralized	Centralized and Distributed
Control	Centralized	Centralized	Centralized and Distributed
Communication	Centralized	Centralized	Centralized and Distributed
CRRM	Centralized	Centralized and Distributed	Centralized and Distributed
Complexity in Fronthaul	Highest	High	Lowest
Burden on the Fronthaul	Highest	High	Lowest
Decouple of Control/User Planes	No	Yes	Yes
Complexity in BBU Pool	Highest	High	Lowest
Burden on BBU Pool	Highest	High	Lowest
Transmission Delay	Long	Long	Low
Latency	High	High	Low
Social and Local Awareness Services	No	No	Yes

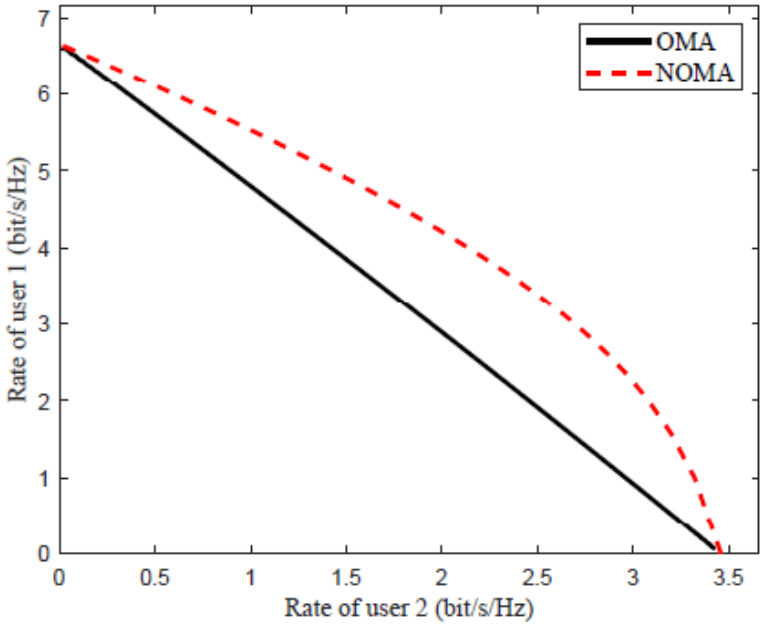

Figure 2.4: The multi-user achievable rate region for downlink NOMA with one BS and two users.u_1 is the strong user with $\frac{|h_1|^2}{\sigma_1^2} = 100$, while u_2 is the weak user with $\frac{|h_2|^2}{\sigma_2^2} = 10$.

2.1.4 C-RAN, H-CRAN and F-RAN for 5G Systems

In C-RAN, the traditional BS functions are decoupled into two parts: distributed low-power and low-complexity remote radio heads (RRHs) and the BBUs clustered as BBU pool in a centralized cloud server [28]. RRHs are deployed to support seamless coverage and provide high data rates in hot spots, while BBUs performs centralized signal processing, provides collaborative transmission and real time cloud computing provides network coordination and flexible spectrum management. RRHs work as soft relays with the capability of compressing and forwarding the signals received from UEs to the BBU pool as shown in Fig. 2.5. BBU pool and RRHs are interconnected via a high bandwidth connectivity interface (i.e. wired/wireless network). It is called fronthaul which ensures the transportation of both data and control signalling. Fronthaul can be realized via optical fiber communications, mil-

limeter wave communications and cellular communications. This type of C-RAN

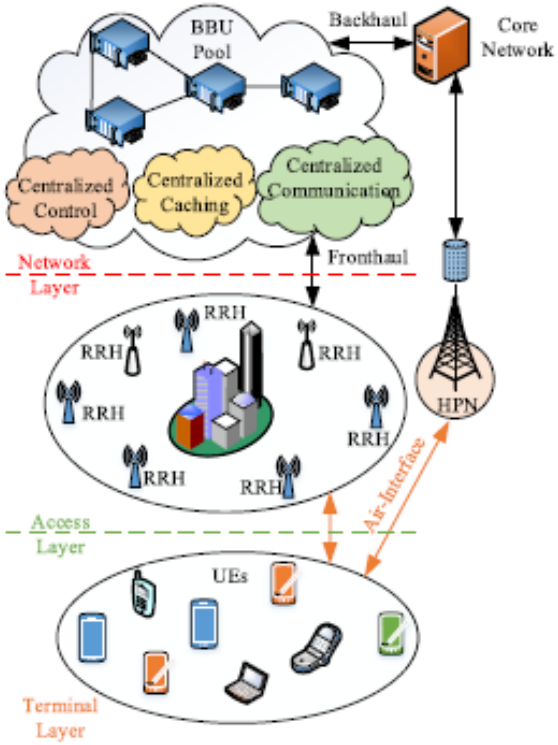

Figure 2.5: CRAN Architecture

structure is denoted full centralized. Although there are various possibilities for C-RAN structures, according to the constraints on fronthaul and the distribution of functionalities between BBUs and RRHs, another prominent structure of C-RAN, namely, partially centralized introduces some RF related baseband processing to the functionalities of RRHs which alleviates the constraints on fronthaul as well as reduces the RRH-BBU overhead.

H-CRAN has been proposed to decouple both control and user planes to enhance the performance and functionalities of C-RAN. In H-CRAN full advantages of Het-Nets and C-RAN have been taken for improving SE and energy efficiency (EE) through suppressing inter-tier interference and enhancing the cooperative process-

ing capabilities [10] as shown in Fig. 2.6. BBU pool efficiently performs baseband signal processing to take advantage of cloud computing capabilities and virtualization techniques. RRHs remotely conduct radio transmission/reception processing and are aimed to guarantee improved network capacity and fulfill the diverse QoS requirements of users. High power nodes (HPNs) are deployed to improve network coverage and control signalling.

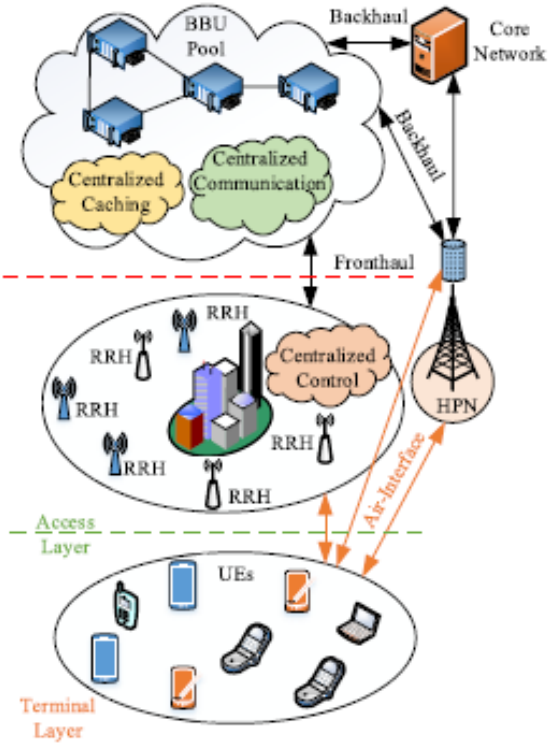

Figure 2.6: HRAN Architecture

F-RAN takes full advantages of the cloud computing, heterogeneous networking and fog computing. Fog computing [29] is a model for storing, managing, processing and analyzing network data. The term "Fog Computing" was initiated by Cisco [30], used to extend the cloud computing to the edge of the network. Some distributed and storage functions exist in the fog layer as shown in Fig. 2.7. Four different

kinds of clouds [11] are defined: 1) global centralized communication and storage cloud 2) centralized control cloud 3) distributed logical communication cloud and 4) distributed logical storage cloud. The global centralized communication and storage cloud is same as the centralized cloud in C-RANs, and the centralized control cloud is used to complete functions of control plane and it is located in HPNs. The distributed logical communication cloud located in fog access points and F-UEs are responsible for the local collaboration radio signal processing (CRSP) and cooperative radio resource management (CRRM) functions, whereas the distributed storage cloud represents the sharing and local caching in edge devices. The burden on the fronthaul and BBU pool are alleviated because a large number of CRSP and CRRM functions are shifted from F-APs and F-UEs.

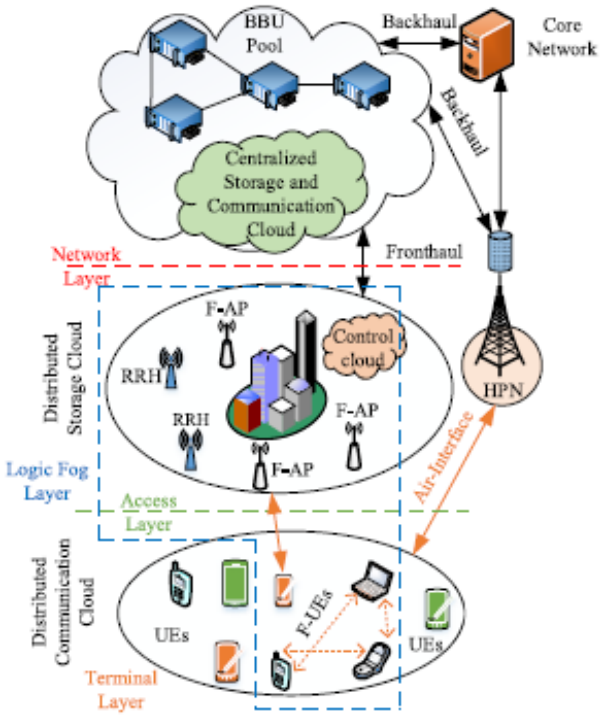

Figure 2.7: FRAN Architecture

2.1.5 Literature Review

As one promising technique in future 5G networks, one of the main challenges of NOMA is whether it is be compatible with other emerging techniques for meeting the requirements of 5G. In this section, we survey the existing research contributions considering the co-existence of NOMA with other 5G key enabling technologies. In this thesis, performance analysis of NOMA in the C-RAN and its two evolutions, namely, H-CRAN and F-RAN is studied.While the first is adopted to support the very dense heterogeneous networks, the second expands the cloud capabilities to the edge through fog computing. The CRAN, H-CRAN, and F-RAN are considered as emerging technical enablers that are capable to fulfill the challenging goals of 5G system.

C-RAN is an emerging network architecture capable of supporting the high data rate services for the fifth generation (5G) mobile networks. C-RAN has been proposed as a novel mobile network architecture allowing centralized processing [31-33]. The pool of BBU and RRHs are connected through high bandwidth optical fronthaul links. BBU pool performs centralized signal processing, provides collaborative transmission and real time cloud computing. In C-RAN, central processor provides support to respective base station (BS) with various services such as inter-cell interference management and increased network capacity [34]. Mobile operators can benefit from various advantages of C-RAN such as reduced Capital Expenditure (CAPEX) and Operational Expenditure (OPEX), decreased power consumption, low latency, improved efficiency in resource allocation, increased adaptability to non uniform traffic, and more flexibility during network upgrading [108]. In C-RAN, the conventional base stations (BSs) are replaced by low-power and low-complexity

RRHs or remote antenna units that are coordinated by a central processor. Therefore the investigation of the resource allocation and scheduling algorithms in C-RAN networks has become the sole impetus of many researchers.

Centralizing baseband processing facilitates better coordination across the RRHs as the cell site information like channel conditions, user requirements and traffic loads are available across the network. Such information can be used for optimization of radio resource allocation, manage inter-cell interference and improve coverage. Therefore, based on the global perspective of the network condition and the information available at the BBU pool, dynamic provisioning and radio resource allocation can improve network performance [35,36]. In [37], two optimization models has been proposed for the 1) resource allocation and power minimization 2) BBU-RRH assignment problem in C-RAN. In [38], the authors propose a QoS-aware radio resource optimization solution for maximizing downlink system utility in C-RAN. In [39], a joint scheduling strategy for resource allocation in C-RAN has been proposed where the time/frequency resources of multiple RRHs are jointly optimized to schedule network users for network throughput improvement.

In cellular wireless networks, the multiple access (MA) technology is one of the important aspects in improving system capacity. In order to enhance SE in wireless networks, NOMA has been proposed in recent work [19,20]. In NOMA, multiple users are multiplexed by superposition coding in the power domain on the transmitter side and employ SIC to separate multi-user signal on the receiver side. Compared to orthogonal multiple access (OMA), NOMA allows multiple users to share time and frequency resources in the same spatial layer via power domain or code domain multiplexing and can enhance the spectral efficiency significantly. The system-level

performance of downlink NOMA and potential issues (i.e. candidate user set selection, power allocation, error propagation for SIC) are investigated in [40,41]. In order to enhance user fairness in cellular downlink, the proportional fair (PF) based scheduling is introduced in NOMA [42]. The problem of effective capacity of NOMA systems subject to quality of service (QoS) is investigated in [43], in which a sub-optimal power control approach is proposed to maximize the sum capacity. In [44], the resource allocation problem in the NOMA system is divided into two cases and an algorithm based on dynamic programming and Lagrangian dual optimization is proposed. A joint power control and subcarrier allocation problem was studied in [45] to minimize the overall transmit power. Although the authors in [46] studied user association and power control in single-cell NOMA networks, the effect of inter-cell interference (ICI) in practical multi-cell scenerios is not considered.

Although some significant contributions have been made on C-RAN and NOMA, these two areas are addressed in seperation. Most of the existing work on resource allocation in C-RAN considered orthogonal frequency division multiple access (OFDMA) based multi-user transmission [47-49]. However NOMA is more appealing for the high throughput demanding wireless network such as 5G. To verify the benefits of NOMA in more realistic setting, it is necessary to consider multi-cell network [50]. Some recent work on NOMA is extended to multi-cell systems in [51]. Applying NOMA technology into the C-RAN network may bring significant benefits.

In dense wireless networks, ICI becomes the major obstacle, which degrades the quality of service (QoS) of cell-edge users. In order to avoid this problem, some recent work applies NOMA with other techniques such as coordinated multi-point

(CoMP) [52,53] and cooperative communication [54]. Recent work in [55] proposes a NOMA scheme for wireless downlink C-RAN to improve SE as well as to support a number of connections in C-RAN. In [56], the authors analyzed the outage probability of the NOMA-enabled C-RAN. Similarly, the authors in [57] derived the expressions in terms of outage probability for both cell-edge and cell-center users. A lot of work have focused NOMA in different network scenerios such as heterogeneous, C-RAN and ultra-dense [58-61] in order to improve resource utilization. The motivation behind considering C-RAN architecture is that the traditional distributed networks lack in global network information and with local network information, scheduling can be performed sub-optimally. Therefore many task of scheduling can benefit from centralized global perspective. In the considered C-RAN architecture, central processor (cloud) is responsible for scheduling users. Centralized scheduling approach offers the highest SE. The spectral efficiency of decentralized approach is lower than that of the centralized approach, due to the lack of global information.

Inspired by the aforementioned potential benefits of NOMA and C-RAN, we therefore explore the potential performance enhancement brought by NOMA for the C-RANs. Although a lot work have exploited C-RAN and NOMA extensively as discussed above, investigation of multiple access techniques particularly NOMA which are of great importance in C-RANs for interference mitigation and spectral efficiency improvement has not been fully explored.

Regarding NOMA systems related to HetNets in [61], authors proposed a cooperative NOMA scheme in HetNets and the inter-user interference is minimized with the aid of dirty paper coding (DPC) precoding in order to intelligently cope with the interference from multi-layers. The distinct power disparity between the MBS

(macro BS) and PBS (pico BSs) was investigated.

H-CRAN is considered as a promising solution to meet the huge traffic demand in future 5G systems. In the H-CRAN [10], various BS types, such as MBS, RRH based BSs, PBS (pico BSs), FBS (femto BSs), are incorporated via a cloud to cooperatively assist the mobile users. Such high density of BSs may cause severe interferences.

To further enhance the C-RAN concept, H-CRAN [10] has been proposed to decouple the control plane and the user plane, in which functions of control plane are implemented in macro base stations (MBSs) or high power nodes (HPNs) whereas RRHs are deployed to provide high data rates for users with diverse quality of service (QoS) requirements. The capacity and time delay constraints on the fronthaul is alleviated by shifting the delivery of control and broadcast signalling from RRH to MBS.

Centralized baseband processing facilitate better coordination among MBS and RRHs as the cell-site information like channel conditions, user requirements and traffic loads are available across the network. Such information can be used for optimization of radio resource allocation, manage inter-cell interference and improve coverage.

In [62] the energy efficiency (EE) of the practical H-CRAN is analysed utilising NOMA by taking into account practical channel modelling with power consumptions at BSs of different cell types (e.g. macro-cell, micro-cell, etc.) and backhauling power.

With the growing popularity of smart devices and various applications, our daily life witnesses a significant increase in the demands for mobile data rate and compu-

tational abilities for running sophisticated applications (social networking, business etc.). However due to scarcity of network resources in the radio access networks (RANs), nowadays mobile devices suffer from constrained computational capability which degrades users' quality of experience. To meet high traffic requirements of crowded users, 5G is obliged to provide enhanced Mobile BroadBand (eMBB) services that will share radio and computational resources with Ultra-Reliable Low-Latency Communication (URLLC). Moreover the network architecture will evolve from a traditional base station-centric architecture to a F-RAN architecture, which enhances the C-RAN architecture by allowing F-APs to be equipped with storage capacity and signal processing functionalities [11,18]. F-RAN will enable network functionalities to be distributed among F-APs and cloud depending on their latency and reliability requirements. Compared with the pure C-RAN mode, the adaptive mode selection in F-RANs can significantly improve SE and decrease latency. In order to achieve high SE and low latency, UE prefers to be served by accessing the local F-APs which cache the desired content and not being served via fronthaul.

A promising approach to reduce the time latency is to relieve unnecessary traffic load by jointly optimizing the transmission schemes (cooperative or non-cooperative) and caching strategies in the F-RAN [63]. For the sake of achieving ultra-low latency, authors in [64] proposed joint distributed computing to enhance the computational capacity of the edge nodes (ENs). In [65], a cooperative communication model is considered with the aim of choosing appropriate number of fog nodes, given the computation and communication resource constraints. The authors in [66] proposed the computing and communication tradeoff factors affecting the objective of minimizing service latency in F-RAN network. Caching in the F-APs can alleviate backhaul

load by pre-storing parts of files in the F-APs at off-peak periods, which greatly improves the quality of servise (QoS) of users. During the peak traffic periods, the cached files can be processed in the fog node and then conveyed to the user terminals. In [67], the fronthaul-aware design of the pre-fetching policy was studied with the aim of achieving ultra-low latency given the cache memory constraints.

On the other hand, the eMBB services aim to provide high data rate guarantee with minimum transmission power consumption.The authors in [68] studied the joint optimization of cloud and edge processing for the downlink of F-RANs, where popular content caching strategies among F-APs were designed to maximize the delivery rate under fronthaul capacity and per-FAP power constraints. In [69], the authors propose a downlink sum-rate optimization scheme in F-RAN, which utilizes the Hungarian method and greedy algorithm combination to balance the resource allocation scheduling among RRHs in order to achieve an optimal downlink sum-rate. In [70], authors proposed a two-level transmission scheme including a cache-level and network-level transmission aiming at maximizing the delivery rate under the constraints of fronthaul capacity, maximum transmit power and size of files.

The authors in [71] addressed the resource allocation for NOMA-enabled F-RAN where each user maximizes the utility function, with power constraints and interference-aware pricing function. The heterogeneous services such as eMBB and URLLC has been studied by assuming orthogonal resource allocation. In [72], the authors proposed a hierarchical radio resource allocation has been modeled as Stackelberg game in order to help the global radio resource manager (GRRM) and local radio resource manager (LRRM) slices achieve the spectral efficiency interactively. The coexistence of heterogeneous services has been studied in [73], where the authors investigated

the joint scheduling of eMBB and URLLC with the goal of maximizing the utility of eMBB traffic while satisfying QoS requirements of URLLC traffic. The authors in [74] propose different methods for optimizing joint scheduling and power adaptation in the downlink of a NOMA-based F-RAN which maximizes a network-wide rate based utility function subject to fronthaul capacity constraints. In [75], the cooperation between fog and cloud computing is investigated by jointly optimizing the offloading decisions and the allocation of computation resource, transmit power, and radio bandwidth while guaranteeing user fairness and maximum tolerable delay.

Although some significant contributions have been made on F-RAN and NOMA, these two areas are addressed seperately. Most of the existing work on resource allocation in F-RAN considered orthogonal frequency division multiple access (OFDMA) based multi-user transmission. However NOMA is more appealing to improve the performance of heterogeneous eMBB and URLLC services. In [76], the resource allocation problem in the NOMA system is divided into two cases and an algorithm based on dynamic programming and Lagrangian dual optimization is proposed. A joint power control and subcarrier allocation problem was studied in [77] to minimize the overall transmit power. Although the authors in [78] studied user association and power control in single-cell NOMA networks, the effect of inter-cell interference (ICI) in practical multi-cell scenerios is not considered.

2.1.6 Key Technologies and Challenges

NOMA is an important enabling technology for achieving the 5G key performance requirements including high system throughput, low latency and massive connectivity. Aiming to achieve higher SE and to satisfy 5G requirements NOMA techniques

has been considered for downlink [20] and the uplink [79]. NOMA is also helpful in the multi-cell scenario to manage ICI [50].

In order to realize low-latency data exchange for multi-cell environment in NOMA, a part of the physical layer processing of the cooperating cells, C-RAN architecture is needed. Moreover C-RAN is also required to accomodate large bandwidth and to enable to serve a higher number of users per cell.

NOMA is anticipated to be used as the key technology in the physical layer of downlink C-RANs. In C-RANs, RRHs are used to provide high data-rate for the users with basic signal coverage. However, such high density of RRHs may cause inefficient usage of communication resources and a considerably degraded throughput at distant cells at the cloud edge. Therefore, NOMA can be applied in C-RANs to support the multiple users to achieve the better SE, high capacity and low latency.

An H-CRAN with NOMA poses great challenges for ICI management due to severe ICI from both intra-tier and inter-tier BSs and intra-cell or inter-user interference caused by NOMA. Although conventional ICIC and eICIC [80] are still applicable and effective, but more sophisticated techniques are needed to further eliminate ICI in H-CRANs with NOMA. In NOMA multiple users are scheduled on same RB by power domain multiplexing. Therefore CoMP scheme should be investigated in order to satisfy the multi-user scenario. Moreover, combining CoMP and ICI management is also important to further improve the performance of H-CRANs with NOMA.

Although some significant works has been done on fog computing and NOMA, how to integrate these techniques and efficient performance of the computation offloading are still open issues. F-APs usually do not have enough storage and

computing resources and may not meet large-scale users' requests. Fog and cloud must be carefully managed in order to operate effectively to improve the overall performance and give an efficient computation offloading. How to split the physical layer network functionalities between edge and cloud so that the coexistence of eMBB and URLLC traffic can be improved is a challenging problem. One more important issue is that: Can NOMA improve the performance of the coexistence between eMBB and URLLC services? Wireless caching is also an important enabling technology for 5G communication networks [81]. How to apply NOMA principle to wireless caching is a challenging problem.

2.1.7 Status of NOMA in 5G

There are some existing evidence of performance improvement when NOMA is integrated with various effective wireless communications techniques, such as cooperative communications [54], multiple-input multiple-output (MIMO), beamforming [51], network coding, etc. With all advancements and experimental outcomes, standardization of NOMA has been established for the next-generation American digital TV standard (ATSC 3.0) [82] under the term layered-division multiplexing (LDM), and was initiated for the third generation partnership project (3GPP) under the name multi-user superposition transmission (MUST) [83]. Below is the summary of the current status of ongoing standardization of NOMA. The 3rd Generation Partnership Project (3GPP) LTE Release 13 approved the study item (SI) of downlink MUST. The primary objective of MUST was to identify the enhancements of downlink multi-user transmission schemes within one cell. In order to achieve this objective, the SI focused on the evaluation of system level gain and

complexity-performance trade-off under practical deployment scenarios and traffic models [84–[89]. The outcome of the study is that NOMA can increase system capacity and improve user experiences. A new work item (WI) of downlink MUST for LTE has also been approved by 3GPP LTE Release 14 [90]. The core objective of this WI was to identify necessary techniques to enable LTE to support downlink intra-cell MUST for the physical downlink shared channel.

The Rel-16 SI [91] has not led to a dedicated NOMA WI. Instead, it is expected that different aspects of NOMA will be continued in more specialized 3GPP studies, e.g. on random access, or URLLC/eMBB multiplexing, in Rel-17 and beyond. Candidate features for future 3GPP NR Releases such as 17 and 18 [92] has developed enhanced NOMA techniques for increasing the number of supported devices per cell, which is particularly important for mMTC. The approach includes regular spreading matrices, spatial preamble reuse, and reinforcement learning for preamble selection. Moreover, NOMA has been proposed for service coexistence, particularily for sharing resources between different service types, e.g. eMBB and URLLC.

Chapter 3

Performance Analysis of NOMA in Cloud Radio Access Networks (C-RAN)

3.1 Introduction

In this chapter, the performance analysis of non-orthogonal multiple access (NOMA) in a cloud radio access networks (C-RAN) is carried out. The problem of jointly optimizing user association, muting and power-bandwidth allocation is formulated for NOMA-enabled C-RANs. To solve the mixed integer programming problem, the joint problem is decomposed into two subproblems as 1) user association and muting 2) power-bandwidth allocation optimization. To deal with the first subproblem, we propose a centralized and heuristic algorithm to provide the optimal and suboptimal solutions to the remote radio head (RRH) muting problem for given bandwidth and transmit power, respectively. The second subproblem is then reformulated and we propose an optimal solution to bandwidth and power allocation subject to users data rate constraints. Moreover, for given user association and muting states,

the optimal power allocation is derived in a closed-form. Simulation results show that the proposed NOMA-enabled C-RAN outperforms orthogonal multiple access (OMA)-based C-RANs in terms of total achievable rate, interference mitigation and can achieve significant fairness improvement.

3.2 System Description and Channel Model

In this section we present a system model of NOMA-enabled C-RAN.

3.2.1 System model of NOMA-enabled C-RAN

Consider downlink direction of a cloud based radio access network architecture. Fig. 3.1 shows a system model of multi-cell downlink of C-RAN architecture with central cloud connected to R remote radio heads (RRHs) via transport networks such as optical transport network and the signalling is assumed to be perfectly synchronized. Let $\mathcal{R} = \{r|1 \leq b \leq R\}$ and $\mathcal{N} = \{n|1 \leq n \leq N\}$ denote the set of all RRHs and users respectively in the network. Users are classified as cell-center users (CCUs) and cell-edge users (CEUs). In a multi-cell network, cell-edge users suffer from interference. In NOMA principle, higher power is allocated to far users which results in more severe inter-cell interference of cell-edge users from the neighbouring cell. We assume that CCUs do not suffer from any ICI. RRHs and users are equipped with single transmit and receive antenna. The set of RRHs which are dominant interfering RRHs that interfere with UE_i is expressed as $I_r = \{r|b_{ru} = 0, \forall r \in \mathcal{R}\}$ where b_{ru} is the user-RRH indicator and $b_{ru} = 1$ indicates the nth user is served by the rth RRH, $b_{ru} = 0$ otherwise. $\alpha_{kn} = 1$ or 0 determines whether the rth RRH exploits the k-th subchannel.

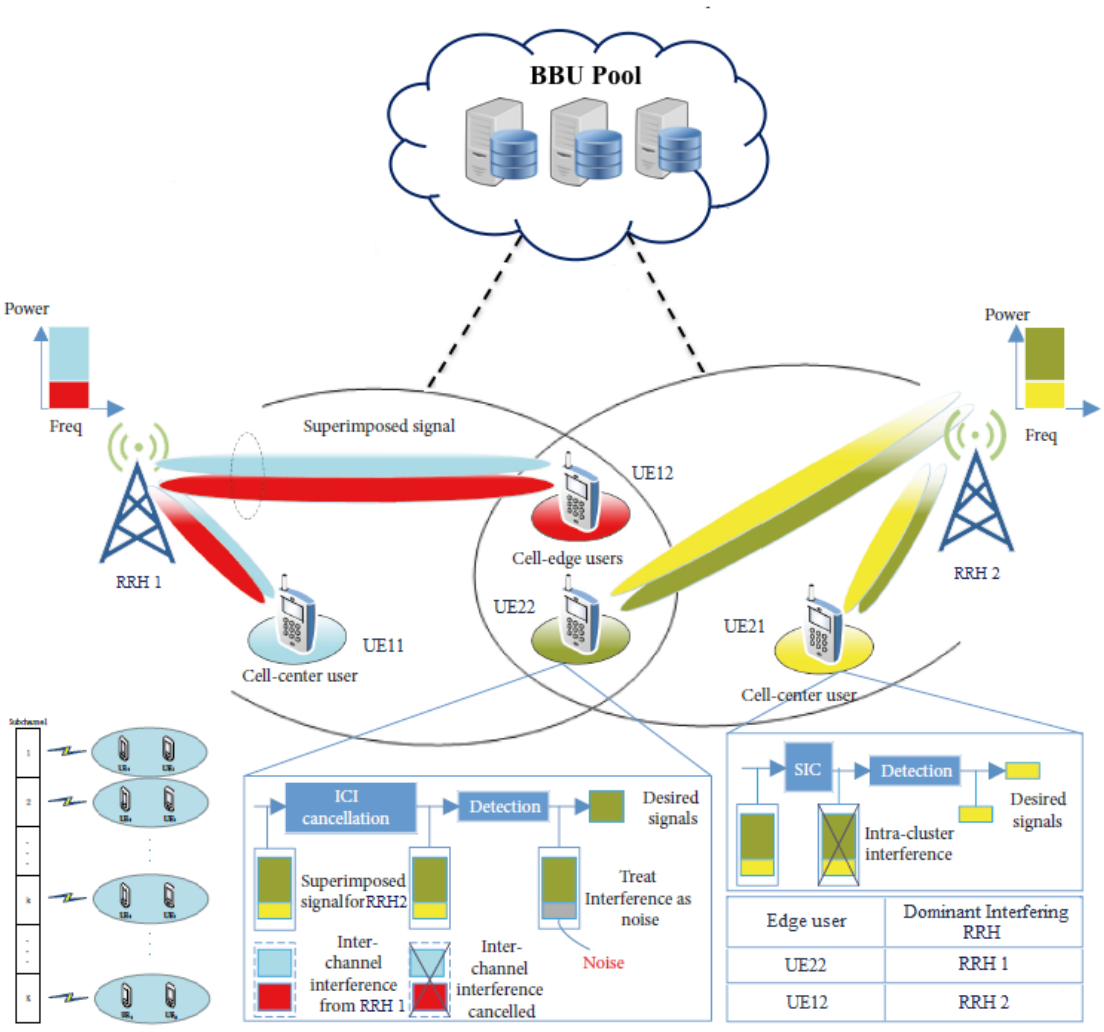

Figure 3.1: System Model for NOMA-enabled C-RAN

Fig. 3.1 includes the table for dominant interfering RRHs shown for the scenerio in which the CEUs UE12 and UE22 are within the range of RRH2 and RRH1 respectively. Here the central controller recognises that RRH1 is the dominant interfering RRH for UE22 and RRH2 is the dominant interfereing RRH for UE12. The maximum transmit power of RRH r is P_k^r and the total available bandwidth is B Hz and the bandwidth allocation factor of the kth subchannel is B_{ij} where B_{ij} is the portion of the entire downlink bandwidth allocated to the NOMA pair (i, j) and

$0 \leq B_{ij} \leq 1$, $\sum_{k=1}^{k} B_{ij} \leq 1$. In NOMA, SIC is performed at the users with stronger channel conditions. It is assumed that the user channel gains on subchannel k are sorted as $|h_{nk}^{u_1}|^2 \geq |h_{nk}^{u_l}|^2 \ldots |h_{nk}^{u_k}|^2$, where h_{nk}^{i} is the channel gain between UE u and RRH r. u_l is the number of UEs that RRH r can serve at the same time.

The superposition coded symbol transmitted by RRH is given by:

$$x_k^r = \sum_{n=1}^{N} b_{ru} \alpha_{kn} \sqrt{P_k^r} x_{ku}^r \tag{3.1}$$

The received signal of UE u associated with RRH r on subchannel k is given by:

$$y_{nk}^r = h_{nk}^r x_k^r + I_{nk}^r + \zeta_{rk}^u \tag{3.2}$$

where h_{nk}^r is the channel gain between RRH r and UE n on subchannel k. ζ_{rk} is the additive white Gaussian noise with power spectral density N_0 and I_{nk}^r is the interference to UE n from other RRHs with unit bandwidth given by:

$$I_{nk}^r = \sum_{m=1, m \neq r}^{R} h_{nk}^m \sqrt{P_k^{rm}} x_k^m / B_{max} \tag{3.3}$$

where P_k^{rm} and B_k^{max} are the maximum power and bandwidth respectively.

$$P_k^{rm} = \sum_{n=1}^{N} b_{rn} \alpha_{kn} P_k^r \tag{3.4}$$

We introduce the following auxillary variable f_{rk}^u given by:

$$f_{rk}^u = \frac{b_{rn} \alpha_{kn} |h_{nk}^r|^2}{\left(\sum_{m=1, m \neq r}^{R} |h_{nk}^m|^2 P_k^{rm} + N_0 B_{max} \right)} \tag{3.5}$$

The received signal-to-interference-plus-noise ratio (SINR) for the i-th decoded user on subchannel k is given by:

$$\gamma_{nk}^u = \frac{b_{ru} \alpha_{kn} |h_{nk}^r|^2 P_{kl}^r}{\sum_{j=l+1}^{u_l} b_{rn} \alpha_{kn} |h_{nl}^r|^2 P_{kj}^r + B_{max}(I_{nk}^r + N_0)} \tag{3.6}$$

where index l denotes that the corresponding user has the l-th highest channel gain among the UEs served by RRH r. Therefore UE l first decode the messages in the i-th order and then successively subtracts the messages of $(l-1)$ UEs to decode its own information.

In practice the maximum number of UEs that can be multiplexed over a channel is often restricted to two to reduce the receiver complexity. In this thesis, we assume that $n_k = 2$ for $k = 1, 2, \ldots K$ and $n = 2K$.

We consider a pair of users (i,j) served by RRH r in which UE i can successfully decode and receive UE j's signal by successive interference cancellation (SIC). In each subchannel, the signals transmitted for NOMA users are ordered based on their channel quality i.e.$|h_{nk}^{u_1}|^2 \geq |h_{nk}^{u_2}|^2$. The condition of SIC decoding order is given by:

$$\frac{b_{rn}\alpha_{kn}|h_{nk}^{ri}|^2}{\sum_{m=1,m\neq r}^{R}|h_{nk}^{mi}|^2 a_{rk}^{i} P_k^{rm} + N_0 B_{max}} \geq \frac{b_{rn}\alpha_{kn}|h_{nk}^{rj}|^2}{\sum_{m=1,m\neq r}^{R}|h_{nk}^{mj}|^2 a_{rk}^{j} P_k^{rm} + N_0 B_{max}} \tag{3.7}$$

RRH r sends messages to RUEs u_1 and u_2 on subchannel k by superposition i.e. RRH r sends $x_n = a_{rk}^{i} x_{rk}^{u_1} + a_{rk}^{j} x_{rk}^{u_2}$, where a_{rk}^{i} and a_{rk}^{j} are the power sharing coefficients.

In terms of vector β_{rk} which is equivalent to the k-th column of muting arrangement matrix, the signal-to-interference-plus-noise ratio (SINR) of UE_i and UE_j served by RRH, $r \in \mathcal{R}$ on subchannel, $k \in \mathcal{K}$ is expressed as:

$$\gamma_{nk}^{i} = \frac{b_{ru}\alpha_{kn}|h_{nk}^{ri}|^2 P_k^{ri}}{B_{max}(I_{nk}^{r}(\beta_{rk}) + N_0)} \tag{3.8}$$

$$\gamma_{nk}^{j} = \frac{b_{ru}\alpha_{kn}|h_{nk}^{rj}|^2 P_k^{rj}}{b_{rn}\alpha_{kn}|h_{nk}^{ri}|^2 P_k^{ri} + B_{max}(I_{nk}^{r}(\beta_{rk} + N_0)} \tag{3.9}$$

where β_{rk} is the muting arrangement matrix of dimensions $R \times K$. If $\beta_{rk} = 0$, the RRH $r \in \mathcal{R}$ is muted on RB $k \in \mathcal{K}$. The muting arrangement is determined so

as to minimize the interference among concurrent transmissions. It indicates the dependence of acheivable rates of NOMA users on the muting decisions β_k on RB k of the dominant interfering RRHs I_r.

$I^r_{nk}(\beta_{rk})$ is the average ICI from other BSs defined as:

$$I^r_{nk}(\beta_{rk}) = \sum_{m=1, m\neq r}^{R} (1-\beta_{rk}) h^m_{nk} \sqrt{P^{rm}_k} x^m_k / B^{max}_k \tag{3.10}$$

From (3.10) it is observed that as the number of muting RRHs increases, the $SINR$ of users on RB k increases which results in increased achievable data rate of users.

3.3 Problem Formulation

The main objective of the work is to optimize the service fairness and network spectral efficiency. A transmission mechanism is proposed with the following requirements

1) to decide the association for each UE

2) to dynamically mute the RRHs and minimize transmit power to mitigate inter-RRH interference.

3) to adjust the bandwidth and power allocation in order to maximize the performance gain

Our aim is to optimize the resource allocation based on designing user-RRH association, user scheduling and bandwidth-power allocation. The joint problem of user association, muting and power-bandwidth allocation for C-RAN is a combinatorial problem. The joint problem is expressed as:

$$O(\beta_{rk}, b_{rn}, B_{ij}, P^r_k) = \max_{\beta_{rk}, b_{rn}, B_{ij}, P^r_k} \sum_{n=1}^{N} \big(\sum_{r=1}^{R} R_{CS} \big) \tag{3.11a}$$

s.t.

$$\sum_{n\in N} B_{ij} \leq \beta_{rk} B_{max} \ \forall r \in R \tag{3.11b}$$

$$\sum_{r\in R}\sum_{u\in N}\sum_{k\in K} b_{rn}\alpha_{kn}P_k^r \leq P^r \tag{3.11c}$$

$$\sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kn}R_{ij}(\hat{\beta_{rk}}, b_{rn}) \geq r_{min} \ \ \forall n \in N \tag{3.11d}$$

$$b_{rn} \leq \beta_{rk} \ \ \forall n \in N, \ \forall r \in R \tag{3.11e}$$

$$\beta_{rk}, b_{rn} \in \{0, 1\} \tag{3.11f}$$

$$\sum_{k\in K} n_k \leq M \tag{3.11g}$$

where R_{CS} denotes the throughput that the user n can obtain with coordinated scheduling when associated with RRH r, $R_{ij}(\beta_{rk}, b_{rn})$ denotes the achievable rate on RB k and its dependence on muting indicator β_{rk} and user-RRH association, constraint 3.11(b) accounts for the bandwidth budget, constraint 3.11(c) means that the RRHs total transmit power cannot exceed RRH maximum power capacity. Constraint 3.11(d) guarantees the quality of service (QoS) requirement of UEs by keeping the rate above or equal to the minimum rate requirements. Constraint 3.11(e) means that the user can connect to RRH only when it is active. Constraint 3.11(g) ensures that the maximum number of users assigned to a particular RB is M. We assume M=2. The joint problem 3.11 is mixed combinatorial non-convex NP-hard optimization problem due to binary constraints for user association as well as the muting indicator and non-convex objective function $R_{CS}(\beta, p)$. We provide the NP-hardness analysis as below:

Theorem 1: (3.11) is NP-hard.

Proof: Firstly we conclude that if N=1 in 3.11(g), problem (3.11) is NP-hard according to [93-94] where the problem reduces to OFDMA subchannel and power

allocation. For multi-carrier NOMA, with $M > 2$, we consider an instance of 3.11 with N users, K RBs and M=2. The total power is given by NKP_k^r. The power limit $P_k^r = 1$ is uniform for users $n \in N$. We select an arbitary user $n \in N$ and assign a dominating weight $w_n = e^{KN}$ and channel gain $g_{kn} = 1$ on all RBs, where other users' dominant weight is $w_k = \epsilon$ and channel gain is $g_{kn} \leq \frac{1}{e}^{KN}$, where ϵ denotes a small value with $0 < \epsilon < \frac{1}{e}^{KN}$. The ratios $\frac{w_{\bar{k}}}{w_k}$ and $\frac{g_{\bar{k}n}}{g_{kn}}$ are large such that the allocation of power $p \leq p_{\bar{k}}$ to user n on any RB k, the function $w_{\bar{k}}R_{CS} > max\big(\sum_{n=1}^{N}\big(\sum_{r=1}^{R} w_{kn}R_{CS}\big)\big)$ is bounded by $KNe^{-KN}log(1+\frac{e^{-KN}p}{\epsilon})$ and $w_{\bar{k}}R_{\bar{K}N} = e^{KN}log(1+\frac{p}{\epsilon})$ is greater than $KNe^{-KN}log(1+\frac{e^{-KN}p}{\epsilon})$. Thus allocating power to user $\bar{n}$ is preferable for maximizing 3.11(a). Consequently, a special case of 3.11 with $M > 1$ is equivalent to the problem in [94] and the result follows.

Fig. 3.2 gives an overview of the proposed approach to solve the joint optimization problem. The key problem transformations, algorithms and generated solutions are shown in different boxes. The boxes with solid and dotted boundaries show the problem reformulations and the proposed algorithms respectively. The optimal and suboptimal solutions generated as a result are shown in rounded rectangular boxes.

In order to solve the combinatorial problem, we decompose the problem into subproblems. We first study the user-RRH association subproblem with given number of active RRHs and fixed power. A semi-distributed Algorithm 1 is developed to find an efficient user-RRH association based on Lagrangian dual method. Then we update the RRH muting states based on user-RRH association strategy using subgradient method as shown in Section 3.4.2. We develop Algorithm 2 determine the actual number of muting states. We obtain the optimal solutions with problem transformations and Lagrangian dual analysis. In addition, a low-complexity Algo-

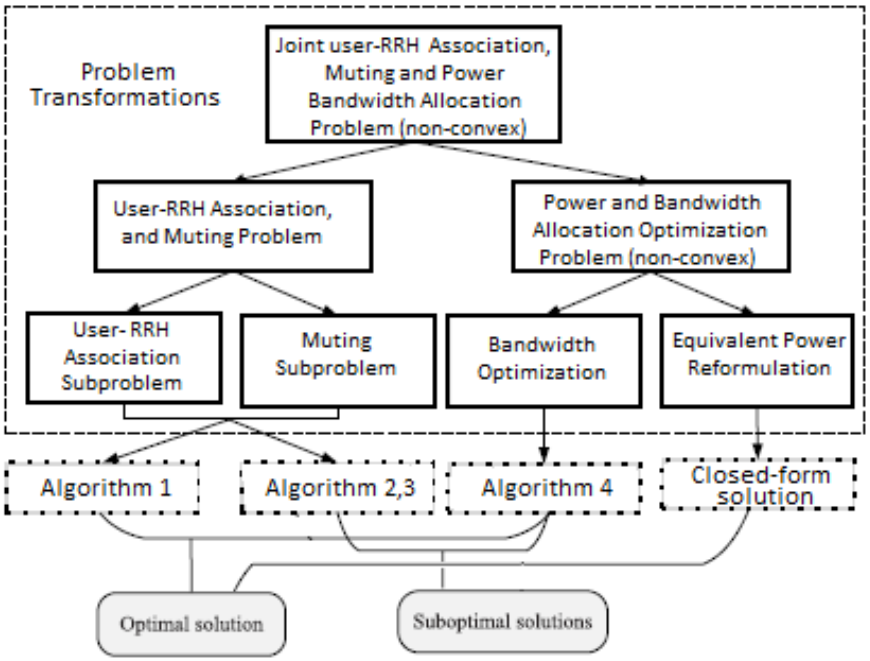

Figure 3.2: Overview of the proposed approach to solve the joint optimization problem

rithm 3 is also developed taking into account the effects of ICI. Then we estimate the total bandwidth of k subchannels to support all users taking into account the target data rate requirements. Then we determine the subchannel assignment based on the bandwidth budget. We develop an iterative bandwidth allocation Algorithm 4 that minimizes the consumed bandwidth which is bounded by the data rate constraint. For a given bandwidth allowance, optimal power allocation is derived in closed form subject to QoS constraints.

3.4 Optimal User Association under fixed bandwidth and Transmit Power

In this section we propose iterative method to solve the formulated problem which is non-convex NP-hard optimization problem. To solve the joint optimization problem we propose two-stage iterative method that decomposes the problem into two stages and solve them iteratively. We first assume fixed bandwidth and transmit power and consider the muting and user association problem. We solve the muting problem with subgradient approach and obtain the optimal user association with given muting indicator. We relax the integer constraints β_{rk}, b_{rn} from $\{0, 1\}$ to $[0, 1]$. However the problem is still non-convex, since the objective function is not concave. By utilising the new variable $\hat{\beta_{rk}} = log_2(\beta_{rk})$, we will have a convex optimization problem with respect to $\hat{\beta_{rk}}$ which can be expressed as:

$$O(\hat{\beta_{rk}}, b_{rn}) = \max_{\hat{\beta_{rk}}, b_{rn}} \left[log\left(\sum_{n=1}^{N} \left(\sum_{r=1}^{R} R_{CS}(e^{\hat{\beta_{rk}}}, b_{rn}) \right) \right) \right] \tag{3.12a}$$

s.t.

$$\sum_{n \in N} B_{ij} \leq e^{\hat{\beta_{rk}}} B_{max} \quad \forall r \in R \tag{3.12b}$$

$$\sum_{r \in R} \sum_{u \in N} \sum_{k \in K} b_{rn} \alpha_{kn} P_k^r \leq P^r \tag{3.12c}$$

$$\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kn} R_{ij}(\hat{\beta_{rk}}, b_{rn}) > r_{min} \quad \forall n \in N \tag{3.12d}$$

$$b_{rn} \leq e^{\hat{\beta_{rk}}} \quad \forall n \in N, \ \forall r \in R \tag{3.12e}$$

$$0 \leq b_{rn} \leq 1 \quad \forall n \in N, \ \forall r \in R \tag{3.12f}$$

$$\hat{\beta_{rk}} \leq 0 \quad \forall r \in R \tag{3.12g}$$

Since log-sum-exp is convex [95], then problem (3.12) is a standard concave maximisation problem. In the problem $O(\hat{\beta_{rk}}, b_{rn})$, due to coupled variables $\hat{\beta_{rk}}$ and b_{rn} in

the constraints, we utilise new variable $\mathcal{S}_r = \sum_{n \in N} b_{rn}$. We decompose the problem into two subproblems in order to decouple the variables. Firstly, given the values of $\mathcal{S}_r$ and $\hat{\beta_{rk}}$, we find the optimal user association b_{rn} and consequently we find the optimal values of $\mathcal{S}_r$ and $\hat{\beta_{rk}}$.

3.4.1 Optimal User Association subproblem

The subproblem of given optimization problem with given values of $\mathcal{S}_r$ and $\hat{\beta_{rk}}$ can be rewritten as:

$$O(b_{rn}) = \max_{b_{rn}} \left[log\left(\sum_{u=1}^{U} \sum_{r=1}^{R} R_{CS}(b_{rn}) \right) \right] \tag{3.13a}$$

s.t. 3.12(c) - 3.12(g)

$$\mathcal{S}_r = \sum_{n \in N} b_{rn} \tag{3.13b}$$

Based on (3.13), the Lagrangian function can be written as:

$$\begin{aligned} \mathcal{L}(b, \lambda, \mu, \theta, \rho) = \left[log\left(\sum_{n=1}^{N} \sum_{r=1}^{R} R_{CS}(b_{rn}) \right) \right] - \lambda_r \left(\sum_{r \in R} \sum_{n \in N} \sum_{k \in K} b_{rn} \alpha_{kn} P_k^r - P^r \right) \\ - \sum_{n \in N} \mu_n \left(r_{min} - \sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kn} R_{ij}(\hat{\beta_{rk}}, b_{rn}) \right) + \sum_{r \in R} \sum_{n \in N} \theta_{rn} (e^{\hat{\beta_{rk}}} - b_{rn}) \\ + \rho_r (\mathcal{S}_r - \sum_{n \in N} b_{rn}) \end{aligned} \tag{3.14}$$

where $\lambda_r \geq 0$ is the Lagrange multiplier for total transmit power constraint, $\mu_n \succeq 0$ is the Lagrange multiplier associated with the required minimum data rate constraint, $\theta_{rn} \succeq 0$ and $\rho \succeq 0$ are the Lagrange multipliers corresponding to the constraints (3.12e) and (3.13b). The operator $\succeq 0$ indicates that all the elements of the vector are nonnegative.

The dual problem is given by:

$$\min_{\lambda, \mu, \theta, \rho} g(\lambda, \mu, \theta, \rho) \tag{3.15}$$

$$s.t. \;\; \lambda_r \geq 0, \mu_n \succeq 0, \theta_{rn} \succeq 0 \; and \; \rho_r \succeq 0 \tag{3.16}$$

$$g(\lambda, \mu, \theta, \rho) = \begin{cases} \max_{b_{ru}} \mathcal{L}(b, \lambda, \mu, \theta, \rho) \\ s.t. \;\; (3.12c), (3.12d), (3.12e) \end{cases} \tag{3.17}$$

Given the dual variables $\lambda, \mu, \theta, \rho$, the optimal solution obtained by maximizing the Lagrangian w.r.t. b_{ru} is:

$$b_{rn^*} = \begin{cases} 1, \text{ if } r = r^* \\ 0, \text{otherwise} \end{cases} \tag{3.18}$$

where $r^* = \max_r(\zeta)$ with

$$\zeta = R_{CS} b_{rn} - \lambda_r P_k^r + R_{ij} - \theta_{rn} + \rho_r \tag{3.19}$$

The subgradient method [96] can be utilized to obtain the optimal solution of given dual problem.

$$\lambda_r(t+1) = \left[\lambda_r(t) - \xi_1 \times \left(P^r - \sum_{r \in R} \sum_{n \in N} \sum_{k \in K} b_{rn} \alpha_{kn} P_k^r \right) \right]^+ \tag{3.20}$$

$$\mu_n(t+1) = \left[\mu_n(t) - \xi_2 \times \left(\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kn} R_{ij}(\hat{\beta_{rk}}, b_{rn} - r_{min} \right) \right]^+ \tag{3.21}$$

$$\theta_{rn}(t+1) = \left[\theta_{rn}(t) - \xi_3 \times \left(e^{\hat{\beta_{rk}}} - b_{rn} \right) \right]^+ \tag{3.22}$$

$$\rho_r(t+1) = \left[\rho_r(t) - \xi_4 \times \left(\mathcal{S}_r - \sum_{n \in N} b_{rn} \right) \right]^+ \tag{3.23}$$

where t is the iteration index. ξ_1, ξ_2, ξ_3 and ξ_4 are the positive step sizes. After obtaining the optimal $\lambda^*, \mu^*, \theta^*$ and ρ^* the corresponding b_{ru} is the solution to the primal problem.

The proposed user association Algorithm 1 is described with initial values of μ_n, $\forall r \in R$ calculated based on the initial user association. The C-RAN centre (BBU pool) collects the channel conditions for RUEs. The RUEs receives pilot signal to

Algorithm 1: Proposed User Association Algorithm

1. Initialize μ_n, $\forall r \in R$ equals to some non-negative value, Set i=1
2. Each user measures its received inter-RRH interference according to the pilot signal from BS and calculates average SINR by accounting pilot signal from BS. They are reported to the CRAN centre.
3. If Avg SINR is greater than the threshold
4. UE selects its BS according to the Avg SINR value.
5. else
6. User receives ζ and R_{cs} values from the BSs.
7. User determines the serving BS according to the maximum $r^* = \max_r(\zeta)$
8. Update μ_n
9. end if
10. Set i=i+1.
11. Each user feedbacks the user association request to the chosen BS broadcast the updated values.

calculate the RSRP (received signal received power) and reports back to the CRAN centre via serving RRH. After collecting the measurements and averaging SINR for each RUE, the centre compare it with the threshold values. If the SINR is greater than threshold it is associated with the RRH else it means that the user can't cope with high interference and it is associated based on the the function $r^* = \max_r(\zeta)$.

3.4.2 Optimal Muting subproblem

We consider the muting problem and develop algorithm. As discussed in the previous section, we first determine the user association indicators given S_r and $\hat{\beta_{rk}}$. Then under fixed user association b_{rn} the problem of optimizing $(S_r, \hat{\beta_{rk}})$ is written as:

$$O(S_r, \hat{\beta_{rk}}) = \max_{S_r, \hat{\beta_{rk}}} O(b_{rn}(S_r, \hat{\beta_{rk}})) \tag{3.24}$$

s.t. (3.12b), (3.12c), (3.12d), (3.12e)

We find the optimal S_r and $\hat{\beta_{rk}}$ by solving the above problem. Let $b^*_{rn}(S'_r)$ be the optimal solution for given problem (3.13) and $O^*(S'_r)$ be the objective function and

we find the optimal value by:

$$O^*(S_r) = \max_{S_r} log[f^*(S_r^{'})] \tag{3.25}$$

s.t. (3.12b), (3.12c), (3.12d), (3.12e)

We consider another solution b_{rn} for S_r. The following inequalities hold:

$$\begin{aligned} f^*(S_r^{'}) &= \mathcal{L}(S^*(b_{rn}^*), \lambda^*(b_{rn}^*), \mu^*(b_{rn}^*), \theta^*(b_{rn}^*), \rho^*(b_{rn}^*)) \\ &\geq \mathcal{L}(S^*(b_{rn}), \lambda^*(b_{rn}^*), \mu^*(b_{rn}^*), \theta^*(b_{rn}^*), \rho^*(b_{rn}^*)) \\ &= log[f^*(S_r^{'}) - \sum_{n=1}^{N}\sum_{r=1}^{R} R_{CS}(S_r^{'} - S_r)] + \rho_r^*(b_{rn})\Big(\sum_{n\in N} b_{rn} - S_r^*(b_{rn})\Big) \\ &\quad +\lambda_r(b_{rn}^*))\Big(\sum_{r\in R}\sum_{n\in N}\sum_{k\in K} b_{rn}\alpha_{kn}P_k^r - P^r\Big) \\ &\quad -\mu_n(b_{rn}^*))\Big(r_{min} - \sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kn}R_{ij}(\hat{\beta_{rk}}, b_{rn})\Big) + \sum_{r\in R}\sum_{n\in N}\theta_{rn}(b_{rn}^*)(e^{\hat{\beta_{rk}}} - b_{rn}) \end{aligned} \tag{3.26}$$

The inequalities above are due to strong duality and optimality of $S_r^{'}(b_{rn}^*)$. Therefore the problem can be updated with the following subgradient method:

$$S(t+1) = S(t) + \frac{f(S_r(t), \hat{\beta_{rk}}(t) - f(S_r^*, \hat{\beta_{rk}}(t)}{|\rho_r^*(t) - \sum_{n=1}^{N}\sum_{r=1}^{R} R_{CS}|}(\rho_r^*(t) - \sum_{n=1}^{N}\sum_{r=1}^{R} R_{CS}) \tag{3.27}$$

To update $\hat{\beta_{rk}}$ we denote $h^*(\hat{\beta_{rk}})$ as the optimal value. We consider another solution b_{rn} for $\hat{\beta_{rk}}$. The following inequalities hold:

$$\begin{aligned} h^*(\hat{\beta_{rk}^{'}}) &= \mathcal{L}(\hat{\beta_{rk}^{'}}(b_{rn}^*), \lambda^*(b_{rn}^*), \mu^*(b_{rn}^*), \theta^*(b_{rn}^*)) \\ &\geq \mathcal{L}(\hat{\beta_{rk}^{'}}(b_{rn}), \lambda^*(b_{rn}^*), \mu^*(b_{rn}^*), \theta^*(b_{rn}^*)) \\ &= h(\hat{\beta_{rk}} + \sum_{n\in N}\theta_{rn}^*(b_{rn}^*)(e^{\hat{\beta_{rk}}} - b_{rn}) + \nabla \end{aligned} \tag{3.28}$$

where λ_r, μ_n are the Lagrangian multipliers corresponding to constraints (3.12c) and

(3.12d). ∇ is given as:

$$\begin{aligned}\nabla = \lambda_r(b^*_{rn})\Bigg(\sum_{r\in R}\sum_{n\in N}\sum_{k\in K} b_{rn}\alpha_{kn}P^r_k - P^r\Bigg)\\ -\mu_n(b^*_{rn})\Bigg(r_{min} - \sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kn}R_{ij}(\hat{\beta_{rk}}, b_{rn})\Bigg)\end{aligned} \tag{3.29}$$

Thus the update for $\hat{\beta_{rk}}$ is given by:

$$\begin{aligned}\hat{\beta_{rk}}(t+1) = \hat{\beta_{rk}}(t) + \frac{f(S_r(t), \hat{\beta_{rk}}(t) - f(S_r, \hat{\beta_{rk}}^*(t)}{|\sum_{n\in N}\theta_{1n}(t), \sum_{n\in N}\theta_{2n}(t) + \ldots + \sum_{n\in N}\theta_{rn}(t)|}\\ (\sum_{n\in N}\theta_{1n}(t), \sum_{n\in N}\theta_{2n}(t) + \ldots + \sum_{n\in N}\theta_{rn}(t))\end{aligned} \tag{3.30}$$

The update depicts the performance gain acheived in terms of data rate with the increase in number of muting dominant interfering RRHs. Conducting the above process iteratively, the convergence of the Algorithm 2 is achieved. Once the convergence condition meets, the optimal solution is achieved.

Algorithm 2: Proposed RRH-muting Algorithm for NOMA based C-RAN Systems

1. **Inputs**
2. **Initialize** $S_r, \hat{\beta_{rk}}, \theta_{rn}, \rho_r$
3. (Repeat)
4. Solve the problem (3.14) by Lagrangian dual method.
5. Update θ_{rn}, ρ_r
6. **Until** θ_{rn}, ρ_r converge;
7. Update $S_r, \hat{\beta_{rk}}$ from (3.27) and (3.30)
8. **Until** $S_r, \hat{\beta_{rk}}$ converge;

We now propose a low complexity algorithm. We develop greedy heuristic search algorithm to solve the problem for muting. The objective is to assign the radio resources in such a way as to mitigate the inter-cell interference by using the concept of coordinated silencing, whilst still improving the downlink user throughput. To guarantee the required data rates of CEUs the RRH may decide not to transmit (coordinated silencing) a superposed message to a set of NOMA users but a dedicated message to CCU.

Initially, the dominant neighbouring interfering RRHs I_b are identified. We derive the average ICI power experienced by cell-edge user assuming no ICI is experienced by cell-center users. The set of RRHs which are dominant interfering RRHs that interfere with UE_i is expressed as $I_r = \{r|b_{ru} = 0, \forall r \in \mathcal{R}\}$

We denote $I^r_{nk}(\beta_{rk})$ as the average ICI from other RRHs defined as:

$$I^r_{nk}(\beta_{rk}) = \sum_{m=1,m\neq r}^{R} (1-\beta_{rk}) h^m_{nk} \sqrt{P^{rm}_k} x^m_k / B^{max}_k \tag{3.31}$$

The ICI experienced by cell-edge user considering dominant interferers is given by:

$$I_c = \sum_{j=2}^{I_r} |h_{ij}|^2 P^r_k \tag{3.32}$$

where power transmitted by RRH is $P^r_k = E|x|^2$, $x = \sqrt{P^{ki}_r} x_1 + \sqrt{P^{kj}_r} x_2$. At the beginning of each scheduling instance, the instantaneous ICI is unknown. Therefore the average power is computed by simple summation of the product of number of users in dominant interfering RRHs I_r and their respective per-user interference factor defined by:

$$I_{total} = \sum_{j=2}^{I_r} F[r^{'}, c] \tag{3.33}$$

where $F[r^{'}, c]$ is a matrix with $r^{'}$ as the number of dominant interfering RRHs and c is the per-user interference exerted by RRHs I_b on RRH r.

Considering the two-cell scenerio in which central processor determines that one RRH is dominant interferer for the neighbouring RRH. The cell-edge users are identified based on normalized channel gains derived in Appendix B. If we assume that a cell-edge user is selected that is liable to suffer from ICI from the neighbouring cell, the following constraint will apply:

$$R(P^{r1j}_k) - R(P^{r2j'}_k) \leq \eta \tag{3.34}$$

where $R(P_k^{r1j})$ is the power received from the user's serving RRH $r1$ and $R(P_k^{r2j'})$ is the power received from the neighbouring RRH $r2$ and η is the pre-defined ICI threshold value and is set equal to noise power in simulations. The condition in (3.34) is checked for both cells by considering only the CEUs. If the condition is met then the RRH $r2$ is one of the dominant interfering RRH for CEU j. In order to improve the cell-edge throughput RRH $r2$ will remain silent (coordinated silencing) in that slot. RRH $r1$ will form a pair having highest PF metric. Fig. 3.3 shows the signalling sequences for the proposed approach. $R^k_{i^*j^*}$ is the highest user utility which a CCU i or CEU j can achieve by assigning a RB k and is defined as $R^k_{i^*j^*} = max(L^*)$, where L^* is the matrix generated for a list of values which defines the utility of an CCU i or CEU j. The central processor is in charge of collecting

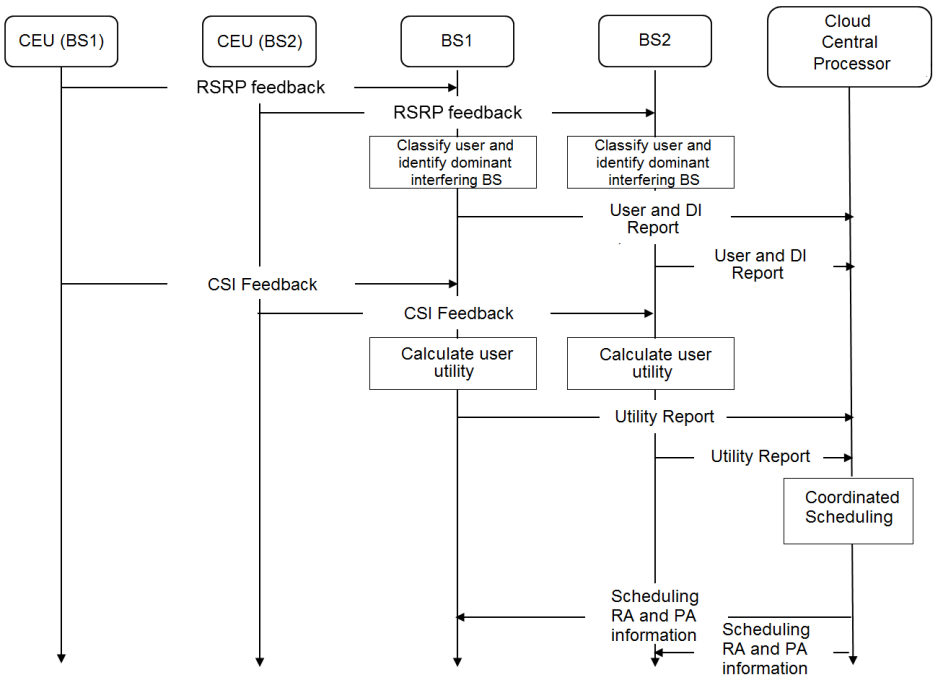

Figure 3.3: Signalling sequences for the proposed approach

and using channel state information (CSI) to make a coordinated scheduling decision among the connected RRHs via the fronthaul links. For R RRHs a total of $J = 2^R - 1$

muting combinations are possible per RB. Considering the dominant interfering RRHs I_r, it is assumed that UEs generate a total of $J = 2^{I_r}$ CSI reports per RB. Initially the cell-edge users are identified based on the received signal power. Then the dominant interferers to that user are identified. The UE and dominant interfering RRHs report are collected at the central processor. RRHs calculate the utility function which is the scheduling matrix that maximizes the sum-rate of the NOMA users. Finally the coordinated scheduling alongwith muting operation is performed at the central processor. Muting decisions are imposed by the centralized processor to the dominant interfering RRHs. The inputs to the Algorithm 3 are UEs, $n \in \mathcal{N}$, RRHs, $r \in \mathcal{R}$, RBs, $k \in \mathcal{K}$. All RRHs are assumed to be activated, i.e. $\beta_{rk} = 1$. We initialize the scheduling matrix S, which is the scheduling set of all UEs. Users are scheduled based on the categorization of cell-edge and cell-center users. Two UEs of different channel conditions are paired on the same RB. UEs with normalized channel gain above L_1 are classified as CCUs and UEs with channel gain below L_2 are CEUs. L_1 and L_2 are the pre-defined threshold values defined in Appendix B.

The proportional fairness (PF) scheduling metric [97] is defined as:

$$w(t) = \sum_{n \in N} \left(\frac{r_{uk}(t)}{R_u(t)} \right)$$

Each RRH performs scheduling by picking UE that has the maximum PF metric. The PF metric is calculated using the instantaneous user data rate and long-term average rate. The scheduling factor is defined as:

$$w_k(t) = \sum_{n \in N} b_{ru}(t) \left(\frac{r_{uk}(t)}{R_u(t)} \right) \tag{3.36}$$

Algorithm 3: Heuristic Muting Algorithm for NOMA based C-RAN Systems

1. **Inputs** $\beta_{rk} = 1$,$\mathcal{N}$,$\mathcal{R}$,$\mathcal{K}$
2. Initialize $S \leftarrow 0$, where S is the scheduling set
3. Set **flag** $= [flag_1, ..., flag_K] \leftarrow \mathbf{0}$;
4. **for** all S **do**
5. **for** $i < 2K$ **do**
6. Classify the users into U_{ceu}(cell edge users) and U_{ccu}(cell center users) based on channel gains.
   If channel gain$>L_1$ then user is U_{ccu}
   If channel gain$<L_2$ then user is U_{ceu}
7. Each BS performs scheduling by picking UE that has maximum PF metric. Find the serving BS for CEU. Find the dominant interfering BS based on the condition in (3.34). The set of dominant interference BSs I_b silences (coordinated
   silencing) on RB k. The set of muting indicators are defined by the set:

$$I_{r_m} = \bigcup_{r_m=\{r_1,r_2...r_m\}} \binom{I_{r_m}}{b_r} \tag{3.35}$$

8. Calculate the sum PF metric $PF(i, j)$ for UEs
9. Compute the metric $T_{i,j}, j \in K$
10. Compute $T_i = max_{j\in K} T_{i,j}$
11. **If** $T_i > T_{i-1}$ then schedule the user as:
    $\bar{j} = argmax_{j\in K} T_{i,j}$
12. Update S with $\bar{j}$
13. **Repeat** till $flag_k$=2.
14. $i \leftarrow i + 1$;
15. **end while**
16. Output S contains set of scheduled UEs

The long-term average rate is updated by the following:

$$R_u(t+1) = \left(1 - \frac{1}{t_c}\right) R_u(t) + \frac{1}{t_c} \sum_{k\in K} s_{uk}(t) r_{uk}(t) \tag{3.37}$$

where $\alpha_{kn}(t)$ is the scheduling index which is equal to one if user n is scheduled in k-th RB, otherwise 0. t_c is the time-window length. $r_{uk}(t)$ is the instantaneous data rate.

$$R_u(t+1) = \left(1 - \frac{1}{t_c}\right) R_u(t) + \frac{1}{t_c} \left(\sum_{k\in K_s} r_{sk}(t) + \sum_{k\in K_w} r_{wk}(t) \right) \tag{3.38}$$

where K_s and K_w are the RB indices in which the user is scheduled as the strong user or the weak user respectively.

At each iteration one RRH is muted $b \in I_b$ by checking the condition in (3.34).When RRH is muted, the maximum PF metrices is calculated among all UEs u on RBs k. The muting is stopped when the additional muting os RRHs does not improve the sum of PF metrices. The set of muting decisions are defined in (3.35). The bionomial coefficients of set I_r are evaluated by taking r_m BSs at a time so as to reduce the complexity of muting decisions for each RB k.. Then we calculate the metric in (3.36) that maximizes the weghted sum rate. The loop runs until all users are scheduled. Flag is set to 2 for maximum number of multiplexed users on the same RB. We get the scheduling set S with scheduled UEs.

3.5 Optimal Power and Bandwidth Allocation with given RRH muting states and User-Association

This section solves the power and bandwidth problem for UE NOMA pair with individual QoS constraints assuming $f^i_{rk} > f^j_{rk}$. We formulate the feasibility problem with the given bandwidth and power budgets to satisfy the QoS constraints for each RRH r. The aim of the subproblem (3.39) is to assign the power and bandwidth budgets of RRHs so that the rate requirements of all users are met. The following minimization problem is formulated as:

$$O(B_{ij}, P^r_k) = \min_{B_{ij}, P^r_k} B_{ij} \sum_{k=1}^{K} P^r_k \tag{3.39a}$$

s.t.

$$\sum_{n \in N} B_{ij} \leq \beta_{rk} B_{max} \ \forall r \in R \tag{3.39b}$$

$$\sum_{r \in R} \sum_{u \in N} \sum_{k \in K} b_{rn} \alpha_{kn} P^r_k \leq P^r \tag{3.39c}$$

$$\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kn} R_{ij}(\hat{\beta_{rk}}, b_{rn}) > r_{min} \ \ \forall n \in N \tag{3.39d}$$

The formulated problem is equivalent to the task of finding minimum bandwidth and power of the RRH r while satisfying the rate requirements of users. We solve the problem in two phases.

3.5.1 Bandwidth Allocation

First we estimate the total bandwidth of k allocated subchannels required to satisfy the rate requirements of users. The power allocation is then determined. We fix the transmission power to a feasible value $P_k^{r^*}$. The bandwidth which is sum of the bandwidth resource allocated to the mth NOMA user pair (i,j) must be minimized. In order to decompose the joint problem new variable which is the total bandwidth of all subchannels is defined as:

$$\sum_{m=1}^{M} B_{ij} \leq B \ \forall r \in R \tag{3.40}$$

We obtain the following optimization problem:

$$\min_{B_{ij}, P_k^{r^*}} \sum_{n \in N} \sum_{k \in K} B_{ij} \tag{3.41a}$$

s.t.

$$\sum_{m=1}^{M} B_{ij} \leq B \ \forall r \in R \tag{3.41b}$$

$$\sum_{n \in N} B_{ij} \leq \beta_{rk} B_{max} \ \forall r \in R \tag{3.41c}$$

$$\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kn} R_{ij}(B_{ij}, P_k^{r^*}) > r_{min} \ \ \forall n \in N \tag{3.41d}$$

Firstly the Lagrange function of the problem is formulated. Sub-gradient approach is then utilized to allocate bandwidth to subchannels. The Lagrange function of the

problem is:

$$\mathcal{L}(B_{ij},\lambda,\mu,\theta)=\sum_{n\in N}\sum_{k\in K}B_{ij}+\lambda\sum_{r\in R}(B-\sum_{m=1}^{M}B_{ij})-\mu\sum_{r\in R}(\sum_{n\in N}B_{ij}-\beta_{rk}B_{max}) \\ -\sum_{n\in N}\theta\left(r_{min}-\sum_{r\in R}\sum_{k\in K}b_{rn}\alpha_{kn}R_{ij}(B_{ij},P_k^{r^*})\right) \tag{3.42}$$

μ can be viewed as cost of assigning subchannel to user pair (i,j) defined as:

$$C_{i,j}=\sum_{n\in N}\mu\beta_{rk} \tag{3.43}$$

The optimal solution B_{ij} must satisfy the Karush-Kuhn-Tucker (KKT) conditions as below:

$$\sum_{n\in N}\sum_{k\in K}B_{ij}-\lambda+\theta=C_{ij} \tag{3.44}$$

$$\lambda(B-\sum_{m=1}^{M}B_{ij})=0 \tag{3.45}$$

$$\mu(\sum_{n\in N}B_{ij}-\beta_{rk}B_{max})=0 \tag{3.46}$$

$$\theta\left(r_{min}-\sum_{r\in R}\sum_{k\in K}b_{rn}\alpha_{kn}R_{ij}(B_{ij},P_k^{r^*})\right) \tag{3.47}$$

$$\lambda,\mu,\theta\geq 0 \tag{3.48}$$

From (3.43) and (3.44) it can be observed that the subchannel k with low cost can be used to assign user pair (i,j). The cost associated with each subchannel is based on the gain value observed by user pair at that subchannel. The minimum cost of the subchannel assigned to user pair can be defined as $C_{ij}=minC$

The dual decomposition results for each subchannel are also the optimal bandwidth allocated given C.

$$B^*=\left[\frac{C+\theta}{b_{ru}}\right]^{B_{max}} \tag{3.49}$$

We update the bandwidth allocation for each subchannel as:

$$B(t+1) = \left[\frac{C(t)+\theta(t)}{b_{ru}}\right]^{B_{max}} \tag{3.50}$$

where t is the iteration index. The split in bandwidth among different pairs can be expressed as:

$$B_{ij}(t+1) = [B_{ij} - \delta(C(t) - C_{ij}(t))]^{+} \tag{3.51}$$

Algorithm 4: Iterative Bandwidth Allocation

1. **Inputs** K_n, K, N
2. **for** $K^* \leftarrow K_n$ **do**
3. Compute the cost associated with each subchannel C_{ij} by using μ
4. Update the bandwidth of each subchannel by:
   $B^*(t+1) = \left[\frac{C(t)+\theta(t)}{b_{ru}}\right]^{B_{max}}$ and $B_{ij}(t+1) = [B_{ij} - \delta(C(t) - C_{ij}(t))]^{+}$
5. The bandwidth allocated with minimum cost is given by: $B^*(t+1) - B_{ij}(t+1)$
6. **end for**

We define the total number of subchannels for all users as bandwidth budget K_n. The assignment table is formed with K subchannels and $|2N|$ users. We then construct a cost matrix given by (3.43). After computing the cost matrix we obtain the best subchannels for each user pair sorted according to the cost. The bandwidth of the subchannel is converged as the cost converges in this algorithm.

3.5.2 Power Allocation

The problem given in (3.39) can be reformulated into an equivalent form. Given UEs consume all bandwidth B, we aim to find the minimum power consumption of RRHs while satisfying the rate requirements of users. Thus the optimization problem can be represented as:

$$O(B^*_{ij}, P^r_k) = \min_{P^r_k} P^r_k \tag{3.52a}$$

s.t.

$$\sum_{n\in N} B_{ij} = \beta_{rk} B_{max} \ \forall r \in R \tag{3.52b}$$

$$\sum_{r\in R}\sum_{u\in N}\sum_{k\in K} b_{rn}\alpha_{kn} P_k^r \leq P^r \tag{3.52c}$$

$$\sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kn} R_{ij}(\hat{\beta_{rk}}, b_{rn}) > r_{min} \quad \forall n \in N \tag{3.52d}$$

We propose centralized power control optimization for fixed $b_{rn}\alpha_{kn}$, i.e. fixed user-RRH and subchannel-RRH indicator. Power allocated to users i and j on the same subchannel is adjusted. For each subchannel the best user pair and its required transmit power is selected in a way that the ICI can be minimized while maintaining QoS. Suppose each user has a minimum SINR level as the QoS then the transmit power needs to satisfy the following equation:

$$\frac{b_{ru}\alpha_{kn}|h_{nk}^r|^2 P_{kl}^r}{\sum_{j=l+1}^{u_l} b_{rn}\alpha_{kn}|h_{nl}^r|^2 P_{kj}^r + B_{max}(I_{nk}^r + N_0)} \geq \gamma_{nk}^u \tag{3.53}$$

The transmit power of user i and j is given by:

$$P_k^{rj} = \frac{B_{max}(I_{nk}^r(\beta_{rk}) + N_0)}{b_{ru}\alpha_{kn}|h_{nk}^{ri}|^2}\gamma_{nk}^i \tag{3.54}$$

$$P_k^{ri} = \frac{b_{rn}\alpha_{kn}|h_{nk}^{ri}|^2 + B_{max}(I_{nk}^r(\beta_{rk} + N_0)}{b_{ru}\alpha_{kn}|h_{nk}^{rj}|^2}\gamma_{nk}^j \tag{3.55}$$

where $\gamma_{nk}^i = (2^{\frac{R_i}{B_{ij}}} - 1)$ and $\gamma_{nk}^j = (2^{\frac{R_j}{B_{ij}}} - 1)$ are the SINR of users i and j respectively. We solve the optimization problem in the case where two different RRHs transmit powers on the subcarrier and four user-RRH wireless links are involved. The channel gains are $g_{nk}^{ri} = |h_{nk}^{ri}|^2$, $g_{nk}^{ri*} = |h_{nk}^{ri*}|^2$, $g_{nk}^{rj} = |h_{nk}^{rj}|^2$ and $g_{nk}^{rj*} = |h_{nk}^{rj*}|^2$ and the power levels are indicated by $g_{nk}^{ri}P_k^{ri}$, $g_{nk}^{ri*}P_k^{ri*}$, $g_{nk}^{rj}P_k^{rj}$ and $g_{nk}^{rj*}P_k^{rj*}$. The following optimization problem is equivalent to solving the problem in (3.52).

$$\min_{P_k^r}(P_k^{ri*} + P_k^{rj*}) \tag{3.56a}$$

$$P_k^{ri} + P_k^{rj} \leq P^r \tag{3.56b}$$

$$f_{rk}^i > f_{rk}^j \tag{3.56c}$$

$$R_{i*} \geq R_{min} \tag{3.56d}$$

$$R_{j*} \geq R_{min} \tag{3.56e}$$

$$P_k^r \succeq 0 \tag{3.56f}$$

where $P_k^r \triangleq [P_k^{ri}, P_k^{rj}]$ is the transmit power vector and P^r is the maximum power constraint on each subchannel. R_{i*} and R_{j*} are the rate variations due to power minimization expressed as:

$$\delta R_{i*} = B_{ij} log_2 \left(\frac{B_{max}(I_{nk}^r(\beta_{rk}) + N_0) + b_{ru}\alpha_{kn}|h_{nk}^{ri}|^2 P_k^{ri}}{B_{max}(I_{nk}^r(\beta_{rk}) + N_0) + b_{ru}\alpha_{kn}|h_{nk}^{ri}|^2 P_k^{ri*}} \right) \tag{3.57}$$

$$\delta R_{j*} = B_{ij} log_2 \left(\frac{b_{rn}\alpha_{kn}|h_{nk}^{ri}|^2 P_k^{ri} + B_{max}(I_{nk}^r(\beta_{rk} + N_0) + b_{ru}\alpha_{kn}|h_{nk}^{rj}|^2 a_{rk}^j P_k^{rj}}{b_{rn}\alpha_{kn}|h_{nk}^{ri}|^2 P_k^{ri*} + B_{max}(I_{nk}^r(\beta_{rk} + N_0) + b_{ru}\alpha_{kn}|h_{nk}^{rj}|^2 P_k^{rj*}} \right) \tag{3.58}$$

The achivable rates of users i and j with two powering RRHs is given by:

$$R_{i*} = B_{ij} log_2 \left[1 + min \left(\frac{b_{ru}\alpha_{kn} g_{nk}^{ri} P_k^{ri}}{b_{rn}\alpha_{kn} g_{nk}^{ri*} P_k^{ri*} + B_{max}(I_{nk}^r + N_0)}, \frac{b_{ru}\alpha_{kn} g_{nk}^{rj} P_k^{rj}}{b_{rn}\alpha_{kn} g_{nk}^{rj*} P_k^{rj*} + B_{max}(I_{nk}^r + N_0)} \right) \right] \tag{3.59}$$

$$R_{j*} = B_{ij} log_2 \left(1 + \frac{b_{ru}\alpha_{kn}\lambda_c P_c^T}{b_{ru}\alpha_{kn}\lambda_c P_{c'}^T + B_{max}(I_{nk}^{r'} + N_0} \right) \tag{3.60}$$

where $\lambda_c P_c^T = P_k^{rj} g_{nk}^{rj} + P_k^{rj*} g_{nk}^{rj*}$ represents the desired signal from for user j jointly ttansmitted from both RRHs and $\lambda_c P_{c'}^T$ represents interference from other NOMA pairs. $P_c = [P_k^{rj} P_k^{rj*}]^T$ and P_c^T is the transpose of P_c

The constraints in (3.56d) and (3.56e) can be rewritten using (3.59) and (3.60) as following:

$$\frac{b_{ru}\alpha_{kn} g_{nk}^{ri} P_k^{ri}}{b_{rn}\alpha_{kn} g_{nk}^{ri*} P_k^{ri*} + B_{max}(I_{nk}^r + N_0)} \geq \gamma_{min}^* \tag{3.61}$$

$$\frac{b_{ru}\alpha_{kn}g_{nk}^{rj}P_k^{rj}}{b_{rn}\alpha_{kn}g_{nk}^{rj*}P_k^{rj*}+B_{max}(I_{nk}^{r}+N_0)} \geq \gamma_{min}^{*} \tag{3.62}$$

$$\frac{b_{ru}\alpha_{kn}\lambda_c P_c^T}{b_{ru}\alpha_{kn}\lambda_c P_{c'}^T+B_{max}(I_{nk}^{r'}+N_0} \geq \gamma_{min}^{*} \tag{3.63}$$

The optimal solutions of problem (3.56) with given γ_{min}^{*} are derived as:

$$P_k^{ri} = \frac{(\gamma_{min}^{*})^2 P_k^r(g_{nk}^{rj}-\gamma_{min}^{*}\lambda_c)+B_{max}(I_{nk}^{r}+N_0)g_{nk}^{ri*}\gamma_{min}^{*}}{g_{nk}^{ri}g_{nk}^{ri*}} \tag{3.64}$$

$$P_k^{rj} = \frac{\gamma_{min}^{*}[B_{max}(I_{nk}^{r'}+N_0)+b_{rn}\alpha_{kn}\lambda_c P_k^r}{b_{rn}\alpha_{kn}g_{nk}^{rj}} \tag{3.65}$$

$$P_k^{ri*} = \frac{\gamma_{min}^{*}(P_k^r g_{nk}^{rj}-\gamma_{min}^{*}\lambda_c P_k^r)}{g_{nk}^{ri*}} \tag{3.66}$$

$$P_k^{rj*} = \frac{\gamma_{min}^{*}\lambda_c P_k^r}{1+\gamma_{min}^{*}g_{nk}^{rj*}} \tag{3.67}$$

For proof please refer Appendix A.

3.6 Simulation Results

In this section performance of our proposed scheme for NOMA-based C-RAN systems is evaluated with system level simulations. We consider multi-cell NOMA based C-RAN system consisting of RRHs and users are uniformly and independently placed within the RRHs' circular coverage area of radius 500m and whose centre is located at a distance of 2 km from the cloud. Each RRH has a coverage radius of D_R. The active mode and the sleep mode power for each RRH is 84W and 56W of power. We assume that all fronthaul links are identical, therefore $R_{fh}^{r} = R^{r}$, where R_{fh}^{r} is the maximum traffic load that can be carried by the fronthaul link associated with RRH r. The maximum number of users that can be multiplexed on the same RB is 2. Moreover we assume that the power sharing coefficients of NOMA users are $a_r^i = 1/4$ and $a_r^j = 3/4$. The simulation parameters are listed in Table 3.1.

The performance of OMA based C-RAN is illustrated as benchmark to demonstrate the effectiveness of our proposed NOMA-enabled C-RAN system. Fig. 3.4 compares the average rates for coordinated scheduling scheme for NOMA C-RAN with OMA C-RAN. We observe that when the number of RRHs increases, the sum rates for NOMA C-RAN and OMA C-RAN first increases and then begin to decrease if the number of RRHs exceeds certain threshold. This crossover effect is due to the fact that overcoverage generates severe aggregated interference. Moreover the performance of NOMA depends on the channel gain differences among users. When the number of RRHs increases, the channel gain differences gradually disappear. To utilize inter-RRH interference, coordinated multipoint joint transmission (CoMP-JT) is considered, which enables multiple RRHs to transmit the same data on radio resources. The CEUs can can combine multiple signals to enhance the performance. We observe that RRH coordination contribute to increase in average achieveble rate for both schemes when there are low to medium number of RRHs. However the average rates of both schemes have an intersection at some specific threhold. This indicates that NOMA CRAN outperforms OMA C-RAN only when the number of RRHs are below some threshold. After certain threshold OMA C-RAN can be better choice than NOMA C-RAN to improve average rate. This is due to the fact that performance of NOMA depends on the channel gain difference between the

Table 3.1: Simulation parameters

Parameter	Values
Distance dependent path-loss from RRH to UE	$148.1 + 37.6log_{10}(d)$, d in km
Number of antenna at RRH/UE	1
Scheduler	Proportional Fairness
Maximum RRH Transmit Power P_k^r	24 dBm
Noise power spectral density	-174 dBm/Hz
Noise Figure	9 dB
Throughput Calculation	Based in Shannon's Formula

users which becomes less with increasing number of RRHs.

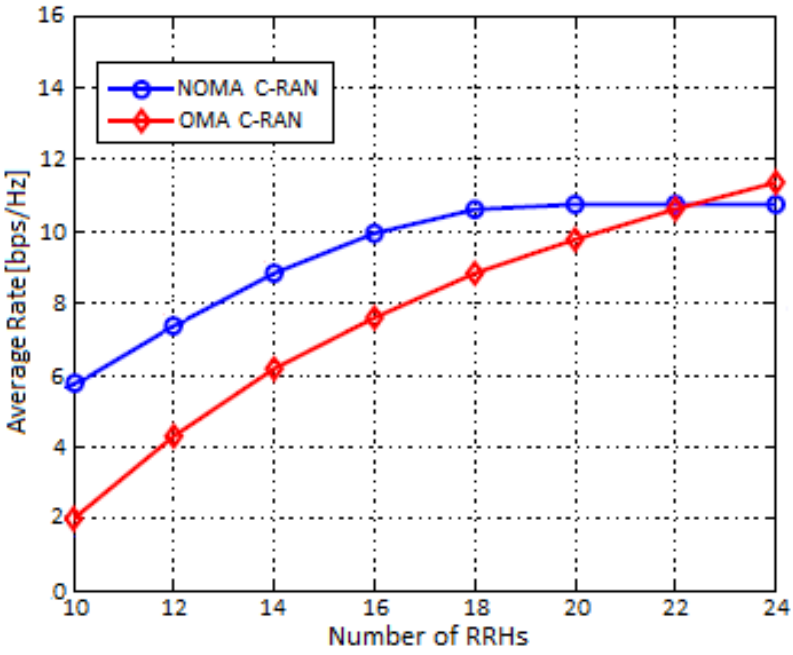

Figure 3.4: Average rates for NOMA and OMA C-RANs

Fig. 3.5 shows the relationship between the Jain's fairness index and number of RRHs for fixed number of RBs. To provide measurement for fairness, Jain's fairness index is used. The fairness index [98] in C-RAN is defined as:

$$J_{fi} = \frac{\left[\sum_{n=1}^{N}(R_i + R_j) + \sum_{b=1}^{B}\beta(R_{i'} + R_{j'})\right]^2}{N_u\left[\left(\sum_{n=1}^{N}(R_i + R_j)^2 + \sum_{b=1}^{B}\beta^2(R_{i'} + R_{j'})^2\right)\right]} \tag{3.68}$$

where β is used to measure the relative throughput of two-RRHs and is defined as:

$$\beta = \frac{\frac{1}{N}\sum_{n=1}^{N}(R_i + R_j)}{\frac{1}{B}\sum_{b=1}^{B}(R_{i'} + R_{j'})} \tag{3.69}$$

The value of Jain's fairness index is between 0 and 1. The rate allocation is perfectly fair if $J_{fi} = 1$.

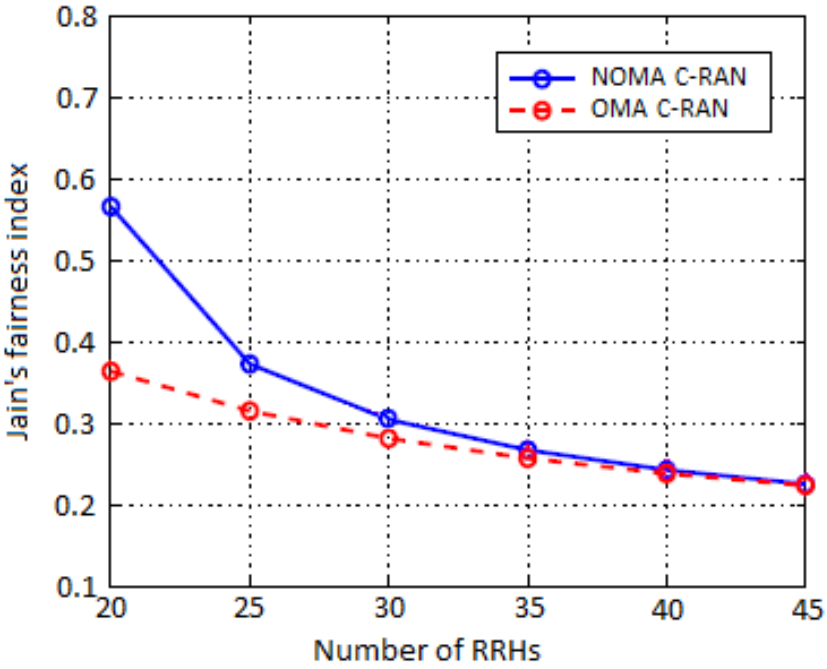

Figure 3.5: Jain's fairness index vs number of RRHs

For a given number of RBs, we observe that the Jain's fairness index decreases with the increasing number of RRHs. This occurs because the aggregated interference experienced by users due to overcoverage is more complicated. Moreover, the network cannot be accessed by the users with poor channel conditions due to more competitiveness for limited resources. We can observe that the fairness level is significantly improved with the proposed NOMA-enabled C-RAN compared to OMA C-RAN especially when the number of RRHs are in low to medium range.

Fig. 3.6 shows that the average data rate increases with the increase in bandwidth. We observe that the proposed approach outperforms OMA scheme for fixed power. The cell-edge users (CEUs) experience less interference due to optimal bandwidth allocation. Although there is possibility for RRHs to mute, it is also possible to serve users on all subchannels to increase network capacity. The NOMA technique

enables the multiple users share a whole frequency band which is occupied by the same RRH to transmit data by proper power allocation. This leads to the feasible bandwidth allocation. Therefore the proposed approach can enhance the spectral efficiency.

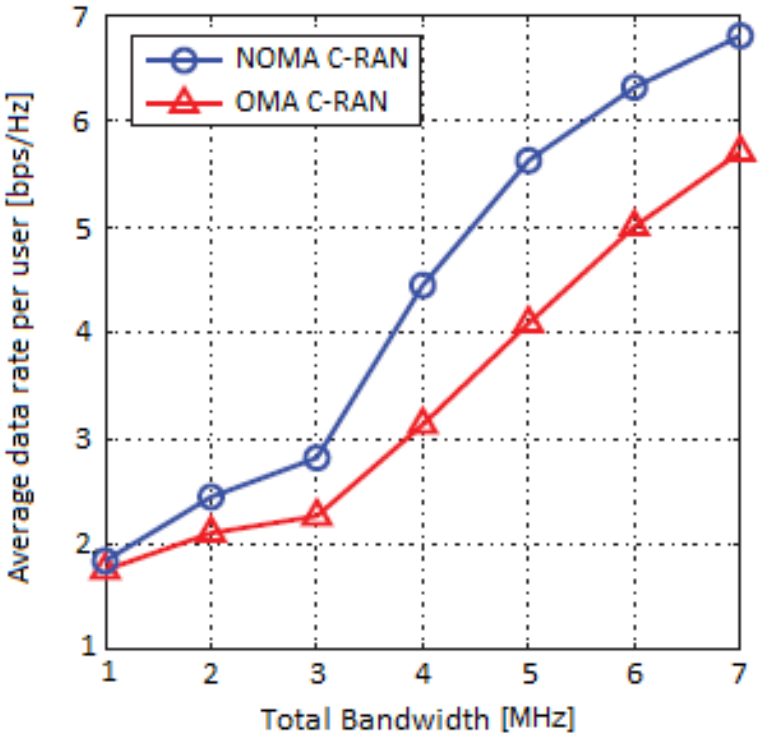

Figure 3.6: Average data rate versus bandwidth

We evaluate the effectiveness of the proposed technique in muting RRHs. We plot the number of active RRHs remaining in each iteration for Algorithm 2 and Algorithm 3 in Fig. 3.7. The convergence behaviour of the proposed algorithm is shown. Fig.3.7 plots the number of active RRHs in each iteration for different number of users/RRH. It can be observed that all the RRHs are initially active. However as the number of iteration increases the number of active RRHs decreases. This implies that when more users are served, more RRHs need to remain active.We observe that Algorithm 2 has better convergence speed than Algorithm 3. Algorithm 2 converges

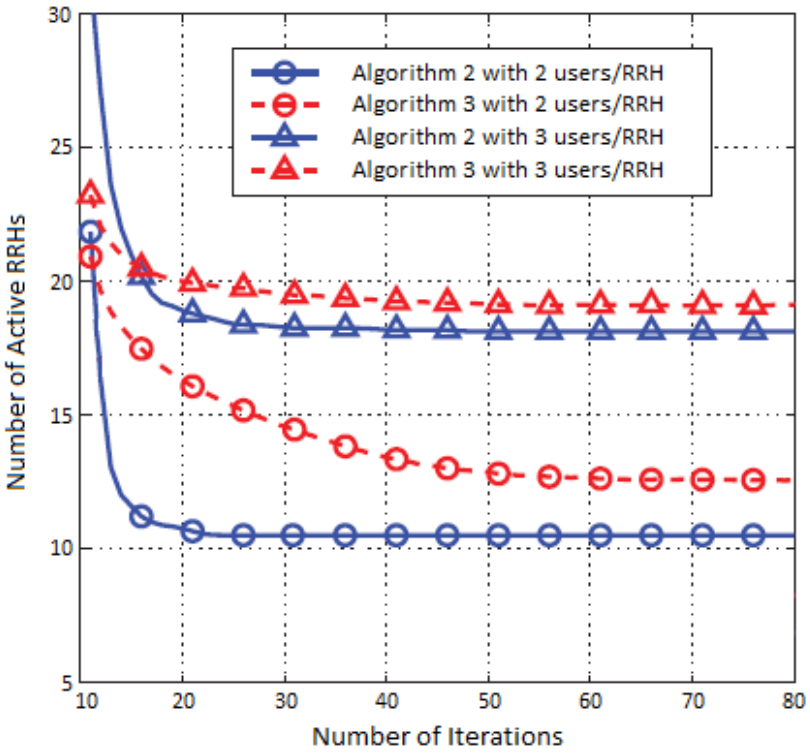

Figure 3.7: Convergence behaviour of proposed algorithm

within 35 iterations while the Algorithm 3 requires 55 iterations to converge.

Fig. 3.8 illustrates the transmit power with target data rate for k=6 and n=12. All users have an identical data rate requirements with rates varying from 1 to 14 bps/Hz. It can be seen that the transmit power increases with the target rate requirement for both schemes. The RRH needs to transmit with a higher power in order to support a more stringent data rate requirement. The proposed optimal power allocation approach provides a significant power reduction as compared to the conventional OMA scheme. Specifically, benchmark scheme requires a higher transmit power (about 2 dB) compared to proposed scheme.

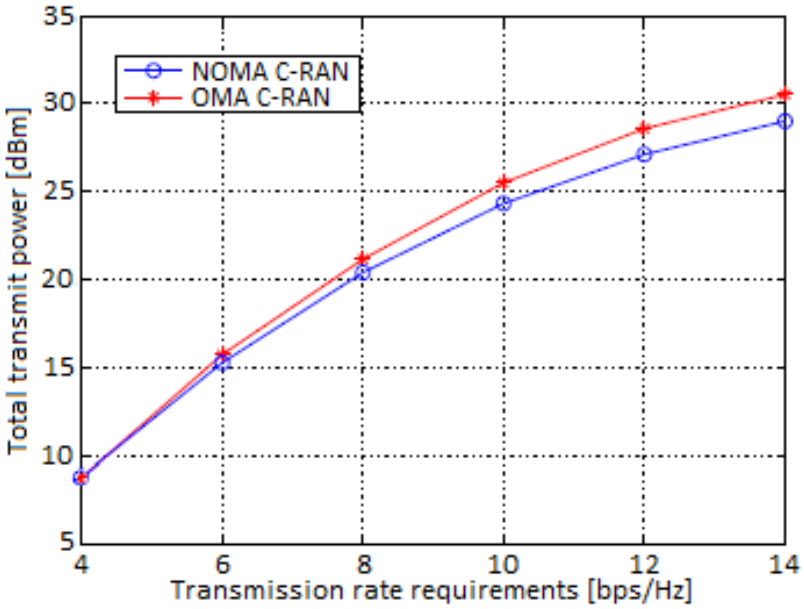

Figure 3.8: Transmit power v.s. minimum transmission rate requirements

3.7 Conclusions

In this chapter we have studied joint user association, muting and power-bandwidth optimization in multi-cell NOMA-enabled C-RAN system. The problem has been formulated as a combinatorial non-convex optimization problem. By formulating joint user association and muting problem, we have proposed a centralized algorithm to provide the optimal solution to the RRH muting problem for fixed bandwidth and transmit power. Besides, a suboptimal algorithm considering ICI has also been proposed to achieve a trade-off between performance and computational complexity. The bandwidth-power allocation problem has been reformulated and an efficient algorithm has been proposed to solve the problem. Moreover the optimal power allocations have been given in closed-form expressions. Specifically, our NOMA-enabled C-RAN framework can find the best RB allocation, number of active RRHs and transmission BPA strategy, while satisfying users' data rate con-

straints and per-RRH bandwidth and power constraints. Simulation results have revealed that our proposed algorithms can obtain the optimal solution of the joint optimisation problem in a significantly reduced computational time and show that NOMA-enabled C-RAN achieves improved network performance in both data rate and network utilisation with proportional fairness consideration. Moreover, numerical results have showed that our proposed joint channel bandwidth and power allocations for NOMA-enabled C-RAN transmission can significantly minimize the total RRHs transmission power considering the bandwidth constraint in comparison with the conventional OMA-enabled C-RAN transmission scheme as well as the corresponding fixed BPA scheme.

Chapter 4

Coordinated Scheduling and Power Control for Non-Orthogonal Multiple Access (NOMA) enabled Heterogeneous Cloud Radio Access Networks(H-CRAN)

4.1 Introduction

In this chapter we investigate the performance of non-orthogonal multiple access (NOMA) in heterogeneous cloud radio access networks (H-CRAN), where coordination of macro base station (MBS) and remote radio heads (RRHs) for H-CRAN with NOMA is introduced to improve network performance. We formulate the problem of jointly optimizing user association, coordinated scheduling and power allocation for NOMA-enabled H-CRANs. To efficiently solve this problem, we decompose the joint optimization problem into two subproblems as 1) user association and scheduling 2) power allocation optimization. Firstly the users are divided based on different interference they suffer. This interference-aware NOMA approach account for the inter-tier interference. Proportional fairness (PF) scheduling for NOMA is utilized to schedule users with a two-loop optimization method to enhance throughput and fairness. Based on the user scheduling scheme, optimal power allocation optimization is performed by the hierarchical decomposition approach. It is then followed

by algorithm for joint scheduling and power allocation. Simulation results show that the proposed NOMA-enabled H-CRAN outperforms OMA-based H-CRANs in terms of total achievable rate and can achieve significant fairness improvement. To the best of our knowledge none of the existing works on H-CRAN investigated multiple access techniques particularly NOMA which are of great importance in H-CRANs for interference mitigation and spectral efficiency improvement.

4.2 System Description and Channel Model

In this section we present a system model of NOMA enabled H-CRAN. We propose a NOMA-enabled H-CRAN model and consider the downlink scenario where one macro base station (MBS) and multiple remote radio heads (RRHs) communicates with multiple UE's via the NOMA protocol. Therefore we formulate the problem where user association, resource allocation and scheduling, and power allocation are jointly considered for downlink H-CRAN to optimize the network utility. To tackle the joint optimization problem we decouple the main problem into two subproblems as joint user association and scheduling, and power allocation.

The system categorizes users into three regions according to the access via MBS, RRH or both. Under this NOMA-enabled H-CRAN model, we first consider MBS and each RRH adopts NOMA scheme where no BSs in the network are coordinated to jointly transmit the NOMA signals. To mitigate the effect of interference the coordination scheme is invoked where the BSs are coordinated to do the joint transmission to the farthest user and thus improve UEs signal-to-interference-plus-noise (SINR) ratio especially around cell-edges. We categorise this region as CS-NOMA region which not only helps near users to perform successive interference cancella-

tion (SIC) but also helps farthest user decode its own signal. To further analyze the performance we categorise users according to different interference they experience from MBS, RRH or both i.e. macro users (MUEs), RRH users (RUEs) suffering interference from MBS and RUEs suffering interference from both MBS and RRHs.

4.2.1 System model of NOMA-enabled H-CRAN

We consider a downlink of NOMA enabled Heterogeneous C-RAN shown in Fig. 4.1 under which multiple RRHs are underlaid within the coverage of one MBS. We assume M macro UEs (MUEs) are uniformly distributed within the MBS. Similar to MBS and RRHs, the MUEs and RRH UEs (RUEs) are each equipped with single antenna. During the optimization process channel remains unchanged. This assumption is justified for networks with very low degree of mobility and/or very high throughput. A cloud center is employed to collect all the channel state information and perform the network optimization.

We denote the set of RRHs in a macrocell as $\mathcal{R} = \{1, 2, \ldots, R\}$. $\mathcal{R}_m = R + 1$ is the total number of MBS and RRHs denoted by set $\mathcal{R}_m = \{0, 1, 2, \ldots, \mathcal{R}\}$ where 0 is the index of the MBS. The total bandwidth B is divided into K resource blocks (RBs) indexed as $\mathcal{K} = \{1, 2, \ldots, K\}$ and each RB occupies a bandwidth of $B_k = B/K$. In order to improve the spectrum efficiency we assume MUEs and RUEs reuse the same set of RBs and we refer RRHs as the underlay tier.

4.2.2 Channel Model

According to the NOMA-based transmission, multiplexing of users through superposition coding (SC) at the MBS and successive interference cancellation (SIC) technique is implemented at the UEs. A single RB can be assigned to multiple

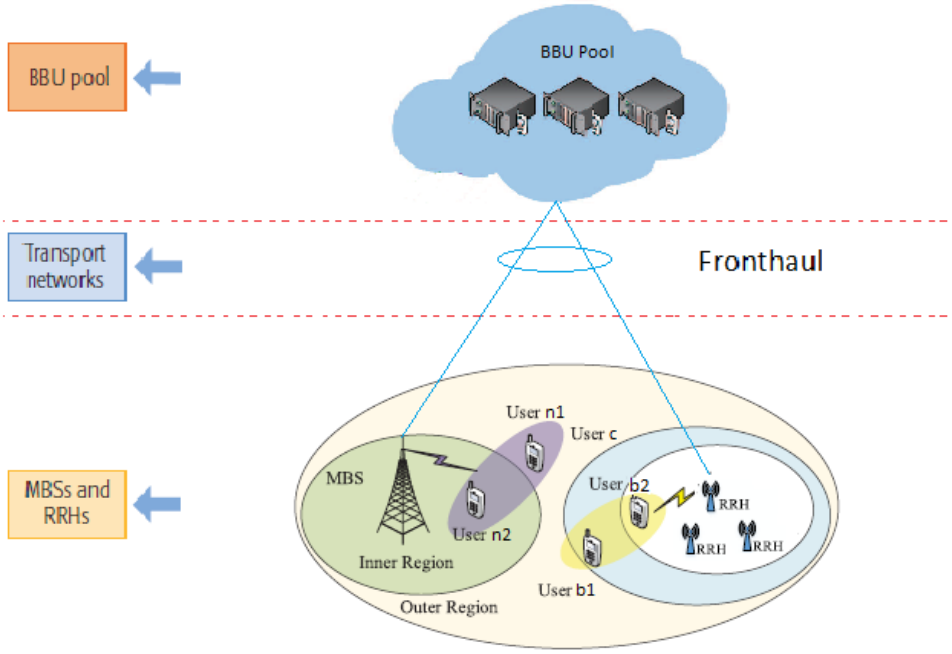

Figure 4.1: System Model of NOMA-enabled H-CRAN

number of users. We denote p^m and p^r as transmit power from MBS and RRH to nth MUE and bth RUE respectively. $f_{r,k}^{b_1} = |\tilde{f}_{r,k}^{b_1}|^2 d_{b_1}^{-\alpha}$ with $\tilde{f}_{r,k}^{b_1} \sim \mathcal{CN}(0,1)$, where $f_{r,k}^{b_1}$ is the Rayleigh fading channel coefficient, α represents the path loss exponent and d_{b_1} is the distance between RRH r and RUE b. $g_{r,k}^{mb_1} = |\tilde{g}_{r,k}^{mb_1}|^2 d_{mb_1}^{-\alpha}$ and $g_{r,k}^{mb_2} = |\tilde{g}_{r,k}^{mb_2}|^2 d_{mb_2}^{-\alpha}$ are the channel coefficients between MBS m and RUE b_1 and b_2, d_{mb_1} and d_{mb_2} are the distances between MBS m and RUEs b_1 and b_2 respectively. β_{bk}^r represents the RB indicator for RRHs i.e. if RB k is assigned to RRH r then $\beta_{bk}^r = 1$ and 0 otherwise. β_{nk}^m represents the RB indicator for MUEs. We define $\mathcal{B} = \{b|1 \leq b \leq b\}$ and $\mathcal{N} = \{n|1 \leq n \leq N\}$ as the index of RUEs and MUEs respectively.

RRH sends messages to RUEs b_1 and b_2 on RB k by superposition i.e. RRH r sends $a_{r,k}^{b_1} x_{r,k}^{b_1} + a_{r,k}^{b_2} x_{r,k}^{b_2}$ where $a_{r,k}^{b_1}$ and $a_{r,k}^{b_2}$ are the power sharing coefficients. $x_{r,k}^{b_1}$ is the symbols transmitted from rth RRH to its serving RUE b_1. The received signal by

RUE b_1 on RB k is given by:

$$y_{r,k}^{b_1} = f_{r,k}^{b_1}\sqrt{p^r a_{r,k}^{b_1}}x_{r,k}^{b_1} + f_{r,k}^{b_1}\sqrt{p^r a_{r,k}^{b_2}}x_{r,k}^{b_2} + \sum_{n=1}^{N}\beta_{n,k}g_{r,k}^{mb_1}\sqrt{p_{n,k}^{m}}x_{n,k}^{m} + \xi_{r,k}^{b_1} \tag{4.1}$$

The channel gains are sorted as $|f_{b,k}^r| \geq \ldots \geq |f_{1,k}^r|$

The received signal by RUE on RB k is

$$y_{b,k}^r = f_{b,k}^r\sqrt{p_{b,k}^r}x_{b,k}^r + f_{r,k}^b\sum_{l=b+1}^{U}\sqrt{p_{l,k}^r}x_{l,k}^r + \sum_{m=1}^{M}\beta_m, kg_{n,k}^m\sqrt{p_{n,k}^m}x_{n,k}^m + \rho_{b,k}^r \tag{4.2}$$

Th first term in the above equation is desired received signal, the second term is the interference from the neighbouring RUEs, the third term is the cross-tier interference and $\xi_{b,k}^r$ is the additive white gaussian noise (AWGN) at RUE b_1 with variance σ^2.

The condition for successive interference cancellation (SIC) decoding order is given by:

$$\frac{|f_{r,k}^{b_2}|^2 p^r a_{r,k}^{b_1}}{|f_{r,k}^{b_1}|^2 p^r a_{r,k}^{b_1} + \sum_{n=1}^{N}\beta_{nk}|g_{r,k}^{mb_1}|^2 p^m + \sigma^2} \geq \frac{|f_{r,k}^{b_1}|^2 p^r a_{r,k}^{b_1}}{|f_{r,k}^{b_2}|^2 p^r a_{r,k}^{b_2} + \sum_{n=1}^{N}\beta_{nk}|g_{r,k}^{mb_1}|^2 p^m + \sigma^2} \tag{4.3}$$

The received SINR at RUE b_1 served by RRH r on RB k is given by

$$\gamma_{r,k}^{b_1} = \frac{|f_{r,k}^{b_1}|^2 p^r a_{r,k}^{b_1}}{|f_{r,k}^{b_2}|^2 p^r a_{r,k}^{b_2} + \sum_{n=1}^{N}\beta_{nk}|g_{r,k}^{mb_1}|^2 p^m + \sigma^2} \tag{4.4}$$

After decoding SINR of RUE b_2 is given by:

$$\gamma_{r,k}^{b_2} = \frac{|f_{r,k}^{b_2}|^2 p^r a_{r,k}^{b_2}}{\sum_{n=1}^{N}\beta_{nk}|g_{r,k}^{mb_2}|^2 p^m + \sigma^2} \tag{4.5}$$

Given the SINR values, the achievable data rate in terms of bit/s/Hz for RUEs b_1 and b_2 can be calculated using Shannon formula as:

$$R_{b_1} = B_k log_2(1 + \gamma_{r,k}^{b_1}) \tag{4.6}$$

and

$$R_{b_2} = B_k log_2(1 + \gamma_{r,k}^{b_2}) \tag{4.7}$$

In the macrocell the SINR of MUE n_1 considering interference from RRH is given by:

$$\gamma_{m,k}^{n_1} = \frac{|f_{m,k}^{n_1}|^2 p^m a_{r,k}^{n_1}}{|f_{r,k}^{n_2}|^2 p^m a_{r,k}^{n_2} + \sum_{k=1}^{K}\sum_{b=1}^{B} \phi_{b,k}|g_{m,k}^{rn_1}|^2 p^r + \sigma^2} \tag{4.8}$$

After decoding, SINR of MUE n_2 is given by:

$$\gamma_{m,k}^{n_2} = \frac{|f_{m,k}^{n_2}|^2 p^m a_{m,k}^{n_2}}{\sum_{k=1}^{K}\sum_{b=1}^{B} \phi_{b,k}|g_{m,k}^{rn_2}|^2 p^r + \sigma^2} \tag{4.9}$$

The data rate for MUEs n_1 and n_2 is given by:

$$R_{n_1} = B_k log_2(1 + \gamma_{m,k}^{n_1}) \tag{4.10}$$

and

$$R_{n_2} = B_k log_2(1 + \gamma_{m,k}^{n_2}) \tag{4.11}$$

The cross-tier interference from MBS to RUEs is given by:

$$\sum_{n=1}^{N} \beta_{nk}|g_{r,k}^{mb_1}|^2 p^m \tag{4.12}$$

The cross-tier interference from RRH to MUEs which are multiplexed in RB k is given by:

$$\sum_{k=1}^{K} \phi_{b,k} \sum_{b=1}^{B} |g_{m,k}^{rn_1}|^2 p^r \tag{4.13}$$

where $g_{m,k}^{rn_1}$ is the channel gain beween RRH r to MUE n_1 on RB k. SINR of UE u in the MBS+RRH range is given as

$$\gamma_{cs} = \frac{\lambda_c p_c^T}{\sum_{c'=c+1}^{\phi_{cs}} p_{c'}^T \lambda_c + \sum_{n=1}^{N} \beta_{nk}|g_{m,k}^{u}|^2 p^m + \sum_{r=1}^{R}\sum_{b=1}^{B} \phi_{b,k} p^r |f_{r,k}^{u}|^2 + 1} \tag{4.14}$$

The term $p_{c'}^T \lambda_c = p_c^r f_{r,k}^u + p_c^m g_{m,k}^u$ represents the desired signal from joint transmission from MBS m and RRH r, $p_c^T = [p_c^r, p_c^m]^T$, $\lambda_c = [f_{r,k}^u, g_{m,k}^u]^T$ and the

term $\sum_{c'=c+1}^{\phi_{cs}} p_{c'}^T \lambda_c$ represents the inter-user interference due to other users in the MBS+RRH range.

4.3 Problem Formulation for Joint User Association, Scheduling and Power Control

We categorize overall service area into three regions: MBS NOMA region, RRH NOMA region and MBS+RRH CS NOMA region based on the association schemes and the received powers from MBS and RRH. The main objective of the work is to optimize the service fairness and network spectral efficiency. A transmission mechanism is proposed with the following requirements

1) to decide the association for each UE

2) to allocate RBs to user pairs at the end of each scheduling cycle.

3) to adjust the power allocation in order to maximize the performance gain

We define the variables to indicate the association status between UE and the MBS or RRH. We define the vectors C_i^m, C_i^r and C_i^{cs} to identify the regions of MUEs, RUEs and CS-CoMP UEs respectively. $\beta_{ij}^m(t)$, $\beta_{ij}^r(t)$ and $\beta_i^{cs}(t)$ represents whether RB k is assigned to RUE, MUE or CS-CoMP UE. $j = 1, 2$ represents index for cell-edge and cell-center RUEs or MUEs. User association and scheduling variables are defined as follows:

$$C_i^m = \begin{cases} 1, \text{ if UE is associated with MBS } m \\ 0, \text{otherwise} \end{cases} \tag{4.15}$$

$$C_i^r = \begin{cases} 1, \text{ if UE is associated with RRH } r \\ 0, \text{otherwise} \end{cases} \tag{4.16}$$

$$C_i^{cs} = \begin{cases} 1, \text{ if UE is in the CS-CoMP NOMA range} \\ 0, \text{otherwise} \end{cases} \tag{4.17}$$

In each time slot each RB can be assigned to only one pair of UEs. Scheduling variables are defined as:

$$\beta_{i1}^m(t) = \begin{cases} 1, \text{ if UE k1 is served by MBS on k-th RB as first UE forming a} \\ \text{NOMA pair at time t} \\ 0, \text{otherwise} \end{cases} \tag{4.18}$$

$$\beta_{i2}^m(t) = \begin{cases} 1, \text{ if UE k1 is served by MBS on k-th RB as second UE forming a} \\ \text{NOMA pair at time t} \\ 0, \text{otherwise} \end{cases} \tag{4.19}$$

$$\beta_{i1}^r(t) = \begin{cases} 1, \text{ if UE k1 is served by RRH on k-th RB as first UE forming a} \\ \text{NOMA pair at time t} \\ 0, \text{otherwise} \end{cases} \tag{4.20}$$

$$\beta_{i2}^r(t) = \begin{cases} 1, \text{ if UE k2 is served by RRH on k-th RB as second UE forming a} \\ \text{NOMA pair at time t} \\ 0, \text{otherwise} \end{cases} \tag{4.21}$$

$$\beta_i^{cs}(t) = \begin{cases} 1, \text{ if UE u is served by MBS and RRH on k-th RB at time t} \\ 0, \text{otherwise} \end{cases} \tag{4.22}$$

The network-wide optimization of joint scheduling and power control in NOMA-enabled H-CRAN with a long-term proportional fair resource allocation can be

formulated as follows:

$$\max_{\beta(t),p(t)} \sum_{u\in U} \sum_{r\in\mathcal{R}_m} U(R(t),\alpha) \tag{4.23a}$$

s.t.

$$\sum_{b=1}^{\mathcal{B}} C_i^r \beta_{ij}^r(t) + \sum_{b=1}^{\mathcal{B}} \sum_{c=1}^{\phi_c} C_i^r C_i^{cs} \beta_i^{cs} \leq 1 \tag{4.23b}$$

$$\sum_{n=1}^{\mathcal{N}} C_i^m \beta_{ij}^m(t) + \sum_{n=1}^{\mathcal{N}} \sum_{c=1}^{\phi_c} C_i^m C_i^{cs} \beta_i^{cs}(t) \leq 1 \tag{4.23c}$$

$$\sum_{k=1}^{K} \left(\sum_{b=1}^{\mathcal{B}} p_{r,k}^b + \sum_{c=1}^{\phi_c} p_c^r \right) \leq P_{thr}^r \tag{4.23d}$$

$$\sum_{k=1}^{K} \left(\sum_{n=1}^{\mathcal{N}} p_{m,k}^n + \sum_{c=1}^{\phi_c} p_c^m \right) \leq P_{thr}^m \tag{4.23e}$$

$$\beta_{ij}^r(t),\ \beta_{ij}^m(t) \textit{ and } \beta_i^{cs}(t)\ \in \{0,1\} \tag{4.23f}$$

where $U(R(t),\alpha)$ is the network-level system throughput defined in (4.37). Constraints (4.23b), (4.23c) and (4.23f) are imposed to ensure that at each time slot RB can be occupied by only one pair of UEs. P_{thr}^r and P_{thr}^m are the maximum MBS and RRH power respectively. Constraints (4.23d) and (4.23e) limit the peak transmitted power of RRHs and MBS. The formulated problem is a mixed combinatorial non-convex problem due to binary variables β and real variable $p(t)$ as well as non-convex objective function. There is no systematic approach to solve this problem optimally. But the problem becomes more tractable if we seperate it into subproblems. To solve the joint optimization problem we propose two-stage iterative method that decomposes the problem into two stages and solve them iteratively.

1) In each iteration of our joint algorithm we associate each user to MBS or RRH based on the association schemes discussed in section 4.4.

2) Transforming into iterative scheduling. The objective of $U(t)$ can be optimized by maximizing $\frac{\sum r_{ij}}{R_i(t-1)}$ for fixed power.

3) Lastly we develop an algorithm for joint user scheduling and power control.

4.4 User Association and Interference Aware NOMA

We define MBS or RRH as the interfering BS for a UE if the average received signal strength from MBS or RRH i.e $p^m d_{mb}^{-\alpha_m}$ OR $p^r d_{rn}^{-\alpha_r}$ satisfies $p_t^{m(r)} d_{mb(rn)}^{-\alpha_{m(r)}} \geq \frac{p_m ax}{\delta}$ where $\delta > 1$ is a user-defined constant for setting the level of interference.

Furthermore we divide the users as cell-edge users and cell-center users based on the boundary distance D as:

$$D = \left(\frac{1 - 2a}{P_b a^2} \right) \tag{4.24}$$

where a is the portion of the BS transmit power allocated to strong NOMA user. For derivation please refer Appendix B. We classify users into five types according to the user access methods. CEU and CCU associated with the MBS or RRH and the UE in the CS-CoMP NOMA range. For any cell-center MUE u the serving BS is MBS and the received power from MBS and RRH satisfies the inequality $p^m d_{mu}^{-\alpha_m} > p^r d_{ru}^{-\alpha_r} \delta$ or $d_{ru} > \delta^{\frac{1}{r}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_r}}$ where $\theta = \left(\frac{p^r}{p^m}\right)^{\frac{1}{\alpha_r}}$.

For the UE u located in the cell-edge region of MBS $p^r d_{ru}^{-\alpha_r} < p^m d_{mu}^{-\alpha_m} \leq p^r d_{ru}^{-\alpha_r} \delta$ or equivalently $\theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} < d_{ru} \leq \delta^{\frac{1}{r}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_r}}$. Therefore this UE suffers main interference from RRH r.

Similarly the cell-center RUE does not experience interference from MBS and its distance to the RRH must satisfy $d_{ru} < \delta^{\frac{-1}{r}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_r}}$.

The cell-edge RUE suffers main interference from MBS i.e $p^r d_{ru}^{-\alpha_r} < p^m d_{mu}^{-\alpha_m}\delta$. Therefore its distance from the RRH must satisfy $\delta^{\frac{1}{r}}\theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} \leq d_{ru} < \theta d_{mu}^{\frac{\alpha_m}{\alpha_r}}$.

For CS-CoMP UE in the cell-range extension region (MBS+RRH region) the difference in the measured signal strength is small (less distictive). Therefore this UE is jointly served by MBS and RRH.

The user association scheme for different users is summarized as below:

$$U^i = \begin{cases} n_1, \text{if } \theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} < d_{ru} \leq \delta^{\frac{1}{r}}\theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} \text{ and } d_{mu} > D \\ n_2, \text{if } d_{ru} > \delta^{\frac{1}{r}}\theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} \text{ and } d_{mu} \leq D \\ b_1, \text{if } \delta^{\frac{1}{r}}\theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} \leq d_{ru} < \theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} \text{ and } d_{ru} > D \\ b_2, \text{if } d_{ru} < \delta^{\frac{-1}{r}}\theta d_{mu}^{\frac{\alpha_m}{\alpha_r}} \text{ and } d_{ru} \leq D \\ c, \text{if } p^m \approx p^r \text{ and } d > D \end{cases} \tag{4.25}$$

Obviously users n_1 and n_2 are associated with MBS and paired to form NOMA group. Similarly users b_1 and b_2 are associated with RRH and paired to form NOMA group. The payoff of MBS or RRH is defined as the sum utility for all users associated with it and is given by:

$$U_R(C_i) = \sum_{b\in B}\sum_{n\in N} C_i \nu_r \tag{4.26}$$

The utility of UE if associated with MBS or RRH is defined as:

$$\nu_r = \begin{cases} a\ log(R_{b(n)}^{d_{r(m)}^{-\alpha_{r(m)}}})\forall U \\ -exp\left(\frac{-bR_{b(n)}^{d_{r(m)}^{-\alpha_{r(m)}}}}{R_{b(n)}^{min}}\right), \forall U \end{cases} \tag{4.27}$$

where a and b are the coefficients with $0 < a < 1$ and $b > 0$, $d_r^{-\alpha_r}$ and $d_m^{-\alpha_m}$ is the distance from UE to RRH and MBS respectively. R_{min} is the minimum data rate requirement.

In order to ensure fairness among users, the utility function [99] is a logarithmic function and is concave, which ensure fairness by allocating more resources to users with lower rate.

The fundamental of NBS (Nash Bargaining Solution) has been widely used for fairly distributing resources among competing players [99]. From eqs (4.26) and (4.27) the utility function of MBS or RRH is:

$$U_R(C_i) = \sum_{b=1}^{B}\sum_{n=1}^{N} C_i a\ log(R_{b(n)}^{d_{r(m)}^{-\alpha_{r(m)}}}) + \sum_{b=1}^{B}\sum_{n=1}^{N} C_i - exp\left(\frac{-bR_{b(n)}^{d_{r(m)}^{-\alpha_{r(m)}}}}{R_{b(n)}^{min}}\right) \tag{4.28}$$

The user association problem in H-CRAN is formulated as:

$$\max_{C_i} U(C_i) = \prod_{r=0}^{R_m}(U_R(C_i) - U_R(C_i)^{min}) \tag{4.29}$$

In (4.29) the optimization goal is to determine which user should be associated with MBS or RRH, so as to maximize the NBS utility function, where $U_R(C_i)^{min}$ is the minimal payoff of MBS or RRH. The main aim of the optimization is to determine C_i which maximizes all $U_R(C_i)$ simultaneously with the constraints $U_R(C_i) \geq U_R^{min}(C_i)$ and $C_i = \{0, 1\}\ \forall\ R, b, n$. Each MBS or RRH has $U_R(C_i)$ as payoff function and is concave since its Hessian matrix is negative semidefinite.

The user association problem is integer programming problem. We adopt continuous relaxation approach to solve the problem by relaxing the constraints $C_i = \{0, 1\}$ to $0 \leq C_i \leq 1$ where C_i is the user association probability.

For one MBS and one RRH the utility is defined as

$U(C_i) = (U_0(C_i) - U_0(C_i)^{min})(U_1(C_i) - U_1(C_i)^{min})$

The Lagrangian function of the optimization problem is given by:

$$\mathcal{L} = \prod_{r=0}^{1}(U_R(C_i) - U_R(C_i)^{min}) + \sum_{b=1}^{B}\sum_{n=1}^{N}\lambda\left(\sum_{r=0}^{1} C_i - 1\right) \tag{4.30}$$

Taking derivative of the above function w.r.t. C_i

$$\frac{\nu_0 + \sum_{b=1}^{B}\sum_{n=1}^{N} C_i^m \frac{d\nu_0}{dC_i^m}}{U_0(C_i) - U_0(C_i)^{min}} = \frac{\nu_1 + \sum_{b=1}^{B}\sum_{n=1}^{N} C_i^r \frac{d\nu_1}{dC_i^r}}{U_1(C_i) - U_1(C_i)^{min}} \tag{4.31}$$

where C_i^m and C_i^r are the association indexes for user to be associated with MBS or RRH.

We define a difference function $f(\nu_0, \nu_1)$ which decides whether UE should be associated with MBS or RRH as:

$$f(\nu_0, \nu_1) = \frac{\nu_0 + X_0 + Y_0}{U_0(C_i) - U_0(C_i)^{min}} - \frac{\nu_1 + X_1 + Y_1}{U_1(C_i) - U_1(C_i)^{min}} \tag{4.32}$$

where

$$X_r = \sum_{b=1}^{B}\sum_{n=1}^{N} \left(\frac{-bR_{b(n)}^{d_{r(m)}} C_i}{\sum_{b=1}^{B}\sum_{n=1}^{N} C_i R_{b(n)}^{min}} \right) exp\left(\frac{-bR_{b(n)}^{d_{r(m)}^{-\alpha_{r(m)}}}}{R_{b(n)}^{min}} \right),\ r \in \{0,1\} \tag{4.33}$$

and

$$Y_r = \sum_{b=1}^{B}\sum_{n=1}^{N} C_i \left(\frac{-a}{\sum_{b=1}^{B}\sum_{n=1}^{N} C_i} \right),\ r \in \{0,1\} \tag{4.34}$$

If the function $f(\nu_0, \nu_1)$ is greater than zero then the user is associated with MBS and if its less than zero then user is associated with RRH. The users will be associated with MBS and RRH simultaneously if the function is equal to zero.

We focus on the simple user association algorithm for one MBS and one RRH. The proposed user association Algorithm 5 is described with initial values of X_r, Y_r and $U_R(0)$ calculated based on the initial user association. The H-CRAN centre (BBU pool) collects the channel conditions for RUEs and MUEs. The RUEs receives pilot signal from MBS and other RRH to calculate the RSRP (received signal received power) and reports back to the H-CRAN centre via serving BS. After collecting the measurements and averaging SINR for each RUE or MUE, the centre compare it with the threshold values. If the SINR is greater than threshold it is associated

Algorithm 5: Proposed User Association Algorithm

1. Initialize δ, λ, X_r, Y_r $U_R(0)$ $R \in \{0, 1\}$ Set i=1
2. Each user measures its received inter-tier interference according to the pilot signal from BS and calculates average SINR by accounting pilot signal from BS. They are reported to the H-CRAN centre via MBS.
3. If Avg SINR is greater than the threshold
4. UE selects its BS according to the Avg SINR value.
5. else
6. User receives $U_R(C_i)$ value from the BSs.
7. User determines the serving BS according to the maximum $f(\nu_0, \nu_1)$
8. Update λ, X_r, Y_r, $U_R(C_i)$
9. end if
9. Set i=i+1.
10. Each user feedbacks the user association request to the chosen BS broadcast the updated values.

with the BS else it means that the user can't cope with high interference and it is associated based on the the function $f(\nu_0, \nu_1)$.

4.5 Proportional Fairness Scheduling

The optimization problem in (4.23a) is solved by transforming into a more amenable form. Given fixed P_m and P_r, the scheduling problem can be expresssed as:

$$\begin{gathered} F_1(\beta) = \max_{\beta(t)} \sum_{u \in \mathcal{U}} U_\alpha(t) \\ s.t\ (4.23b),\ (4.23c)\ and\ (4.23f) \end{gathered} \tag{4.35}$$

Outer problem maximizing $U(t)$ by varying $\beta(t)$ for a given $p(t)$ The achievable rate at discrete time t on RB k can be written as:

$$R(t) = \beta_{ij}^m(t)(1 - \beta_{ij}^r(t))C_i^m R_{nj} + \beta_{ij}^r(t)(1 - \beta_{ij}^r(t))C_i^r R_{bj} + \beta_i^{cs}(t)C_i^{cs} R_{cs} \tag{4.36}$$

Proportional fairness (PF) is used in order to maintain fairness among users. The RB k is allocated to user pair having the highest PF metric. To quantify the fairness among users we adopt the α-fairness utility function [100], where $U(t)$ is

the network-level system throughput defined as:

$$U(R(t),\alpha) = \begin{cases} \frac{R(t)^{1-\alpha}}{1-\alpha} \ for \ \alpha \geq 0, \alpha \neq 1 \\ \\ log \ R(t), for \ \alpha = 1 \end{cases} \tag{4.37}$$

where $R(t)$ is the average throughput of user at discrete time t. α is a non-negative parameter standing for the tradeoff between user fairness and spectrum efficiency. The fairness among users will be enhanced at the cost of reduced spectrum efficiency as α increases. The average throughput of user at time t is defined as:

$$\begin{aligned} R^k(t) = R^k(t-1) + \frac{1}{R_{avg}} \Bigg[\sum_{k \in K} \sum_{r \in \mathcal{R}_m} \beta_{ij}^m(t)(1-\beta_{ij}^r(t)) C_i^m R_{nj} + \\ \beta_{ij}^r(t)(1-\beta_{ij}^r(t)) C_i^r R_{bj} + \beta_i^{cs}(t) C_i^{cs} R_{cs} - R(t-1) \Bigg] \end{aligned} \tag{4.38}$$

PF metric is calculated from the instantaneous rate and long-term average rates given by:

$$w(t) = \frac{R(t)}{R^k(t-1)} \tag{4.39}$$

where $r(t)$ denotes the estimated instantaneous rate UE receives in scheduling interval t, which is updated in each iteration and $R^k(t-1)$ is the long-term average rate. The long-term average rate is updated by the following:

$$\begin{aligned} R^k(t+1) = \left(1 - \frac{1}{t_c}\right) R^k(t) + \frac{1}{t_c} \sum_{k \in K} \sum_{r \in \mathcal{R}_m} \beta_{ij}^m(t)(1-\beta_{ij}^r(t)) C_i^m R_{nj} + \\ \beta_{ij}^r(t)(1-\beta_{ij}^r(t)) C_i^r R_b + \beta_i^{cs}(t) C_i^{cs} R_{cs} \end{aligned} \tag{4.40}$$

where $s_{uk}(t)$ is the scheduling index which is equal to one if user u is scheduled in k-th RB, otherwise 0. t_c is the time-window length. $r_{uk}(t)$ is the instantaneous data rate.

The scheduling problem is solved based on gradient descent method for each RB k:

$$\max_{\beta(t)} \sum_{u\in U} \sum_{r\in\mathcal{R}_m} \frac{\beta_{ij}^m(t)(1-\beta_{ij}^r(t))C_i^m R_{nj} + \beta_{ij}^r(t)(1-\beta_{ij}^r(t))C_i^r R_{bj} + \beta_i^{cs}(t)C_i^{cs}R_{cs}}{R_u(t-1)} \tag{4.41}$$

The gradient problem above is the only way to exploit the multiuser diversity whilst keeping full control of fairness among users through the flexible tuning knob α.

Algorithm 6: Proposed Two-Loop Optimization for User Association and Scheduling

Outer Loop: For fixed δ from Algorithm 5 we categorize UEs based on their access via MBS, RRH or both.
Inner Loop: Initialize Lagrangian multipliers ρ_i^m, ρ_i^r
We obtain the optimal indexes for MUE, RUE and CS-UE as $x_{ij}^r, x_{ij}^m, x_{ij}^{cs}$

We propose two-loop optimization for user association and scheduling in Algorithm 6. In the outer loop, we initialize a fixed bias value δ from Algorithm 5 and determine the association status for each user. The interference power from neighbouring BS is calculated and users are categorized into three groups based on access via MBS, RRH or both. In the inner loop Lagrange multipliers ρ_i^m, ρ_i^r are initialized and updated. Finally, the optimal indexes are obtained from (4,43), (4.44) and (4.45). The Lagrangian function of the problem can be defined as follows:

$$\begin{aligned}
&\mathcal{L}(\beta_{ij}^m(t), \beta_{ij}^r(t), \beta_i^{cs}(t), \rho_i^m, \rho_i^r, x_{ij}^r, x_{ij}^m, x_{ij}^{cs}) = \\
&\sum_{u\in U}\sum_{r\in\mathcal{R}} \frac{\beta_{ij}^m(t)(1-\beta_{ij}^r(t))C_i^m R_{nj} + \beta_{ij}^r(t)(1-\beta_{ij}^r(t))C_i^r R_b + \beta_i^{cs}(t)C_i^{cs}R_{cs}}{R_u(t-1)} + \\
&\rho_i^r\left(\sum_{b=1}^{B} C_i^r\beta_{ij}^r(t) + \sum_{b=1}^{B}\sum_{r=1}^{R}\beta_{ij}^r(t)C_i^{cs}\beta_i^{cs} - 1\right) + \\
&\rho_i^m\left(\sum_{n=1}^{N} C_i^m\beta_{ij}^m(t) + \sum_{n=1}^{N}\beta_{ij}^m(t)C_i^{cs}\beta_i^{cs} - 1\right) - \\
&\sum_{r=1}^{\mathcal{R}_m}\sum_{b=1}^{B}\xi_r\beta_{ij}^r(t) - \sum_{n=1}^{N}\xi_m\beta_{ij}^m(t) - \sum_{c=1}^{\phi_c}\xi_c\beta_i^{cs}(t)
\end{aligned} \tag{4.42}$$

For each RB k we need to find the optimal index for MUEs, RUEs and CS-UEs at

time t. The optimal RRH-user index is derived as:

$$x_{ij}^{r} = arg \max_{x_{ij}^{r} \in B} \frac{R_{bj}(t)}{R(t-1)} \tag{4.43}$$

The optimal Macro-user index is:

$$x_{ij}^{m} = arg \max_{x_{ij}^{m} \in N} \frac{R_{nj}(t)}{R(t-1)} \tag{4.44}$$

and the optimal CS-UE index is:

$$x_{ij}^{cs} = arg \max_{x_{ij}^{cs} \in \phi_{cs}} \frac{R_{cs}(t)}{R(t-1)} \tag{4.45}$$

For the derivations please refer to Appendix C

4.6 Power Control Problem

The original problem reduces to power control problem for optimal $x_{ij}^{r}, x_{ij}^{m}, x_{ij}^{cs}$ as:

$$\begin{gathered} F_2(\mathbf{p}) = \max_{p(t)} \sum_{u \in \mathcal{U}} U_\alpha(R(p)) \\ s.t.\ (4.23d)\ and\ (4.23e) \end{gathered} \tag{4.46}$$

Due to inter-tier interference the above optimization problem is non-convex. In order to obtain optimal solution to problem F_2 we adopt heirarchical decomposition method [101] For $\alpha > 0$ we define a vector $x = [x_1, x_2, \ldots, x_n]$ and rewrite the problem F_2 as:

$$F_2(\mathbf{p}) = \max_{p,x} x \tag{4.47}$$

s.t (4.23d) and (4.23e)

$$x < R(p) \tag{4.48}$$

We define the Lagrange function of problem by relaxing the above constraint (4.48) as:

$$\mathcal{L}(p, x, \lambda) = \max_{p,x} x + \lambda(R(p) - x) \tag{4.49}$$

where $\lambda \geq 0$ is the Lagrangian multiplier corresponding to the constraint (4.48). The dual function is given as:

$$G(\lambda) = \begin{cases} \max_{p,x} \mathcal{L}(p, x, \lambda), \\ s.t\ (4.23d)\ and\ (4.23e) \end{cases} \tag{4.50}$$

The above dual function can be seperated into two maximization subproblems:

$$G_1(\lambda)) = \begin{cases} \max_{p} \lambda_n\ R(p), \\ s.t\ (4.23d)\ and\ (4.23e) \end{cases} \tag{4.51}$$

$$G_2(\lambda) = \max_{x} f(x)\ = \max_{x} U_\alpha(x) - \lambda_n \tag{4.52}$$

The subproblem G_1 can be solved with Lagrangian dual decomposition method. By relaxing the constraints the LDD function becomes:

$$\begin{aligned} H(p, \lambda_r, \lambda_m) = & \sum_{b\in B}\sum_{n\in N}\sum_{r\in\mathcal{R}_m}\sum_{c\in\phi_c} \lambda_n R(p) + \\ & \lambda_r\left(P^r_{thr} - \sum_{k=1}^{K}\left(\sum_{b=1}^{B} p^b_{r,k} + \sum_{c=1}^{\phi_c} p^r_c\right)\right) + \\ & \lambda_m\left(P^m_{thr} - \sum_{k=1}^{K}\left(\sum_{n=1}^{N} p^n_{m,k} + \sum_{c=1}^{\phi_c} p^m_c\right)\right) \end{aligned} \tag{4.53}$$

where λ_r and λ_m are the dual vectors corresponding to the constraints (4.23d) and (4.23e).

$$\begin{aligned} H(p, \lambda_r, \lambda_m) = & \sum_{k\in K}\sum_{b\in B}\sum_{n\in N}\sum_{c\in\phi_c}\sum_{r\in\mathcal{R}_m} \lambda_n R(k) + \\ & \lambda_r\left(P^r_{thr} - \sum_{k=1}^{K}\left(\sum_{b=1}^{B} p^b_{r,k} + \sum_{c=1}^{\phi_c} p^r_c\right)\right) + \\ & \lambda_m\left(P^m_{thr} - \sum_{k=1}^{K}\left(\sum_{n=1}^{N} p^n_{m,k} + \sum_{c=1}^{\phi_c} p^m_c\right)\right) \end{aligned} \tag{4.54}$$

$$H(p,\lambda_r,\lambda_m) = \Bigg(\sum_{k\in K}\sum_{b\in B}\sum_{r=1}^{\mathcal{R}_m} R_{bj} + \sum_{k\in K}\sum_{n\in N} R_{nj} + \sum_{k\in K}\sum_{c\in\phi_c} R_{cs}\Bigg)\lambda_n - \lambda_r \sum_{k\in K}\sum_{b=1}^{B} p_{r,k}^{b} - \lambda_m \sum_{k=1}^{K}\sum_{n=1}^{N} p_{m,k}^{n} - \lambda_r \sum_{k\in K}\sum_{c=1}^{\phi_c} p_c^r - \lambda_m \sum_{k\in K}\sum_{c=1}^{\phi_c} p_c^m + \lambda_r P_{thr}^r + \lambda_m P_{thr}^m \tag{4.55}$$

To obtain the optimal power allocation for the given RB assignment, we solve the following:

$$\begin{aligned}\widetilde{H}(p,\lambda_r,\lambda_m) = \max_p \Bigg(\sum_{k\in K}\sum_{b\in B}\sum_{r=1}^{\mathcal{R}_m} R_{bj} + \sum_{k\in K}\sum_{n\in N} R_{nj} + \sum_{k\in K}\sum_{c\in\phi_c} R_{cs}\Bigg)\lambda_n - \\ \lambda_r \sum_{k\in K}\sum_{b=1}^{B} p_{r,k}^{b} - \lambda_m \sum_{k=1}^{K}\sum_{n=1}^{N} p_{m,k}^{n} - \lambda_r \sum_{k\in K}\sum_{c=1}^{\phi_c} p_c^r - \lambda_m \sum_{k\in K}\sum_{c=1}^{\phi_c} p_c^m\end{aligned} \tag{4.56}$$

Consequently the function can be decoupled by dual decomposition method [101]. We seperate the problem into three parts:

$$\widetilde{H}_1 = \sum_{b\in B} R_{bj} - \lambda_r \sum_{b=1}^{B} p_{r,k}^{b} \tag{4.57}$$

$$\widetilde{H}_2 = \sum_{n\in N} R_{nj} - \lambda_m \sum_{n=1}^{N} p_{m,k}^{n} \tag{4.58}$$

$$\widetilde{H}_3 = \sum_{c\in\phi_c} R_{cs} - \lambda_r \sum_{c=1}^{\phi_c} p_c^r - \lambda_m \sum_{c=1}^{\phi_c} p_c^m \tag{4.59}$$

$\widetilde{H}_1$ is a concave in $p_{r,k}^b$. Let $\frac{d\widetilde{H}_1}{dp_{r,k}^b} = 0$ and we can derive the optimal power allocation for RUEs maximizing $\widetilde{H}_1$ as:

$$p_{b1}^r = \frac{B_k(1+\hat{\gamma}_{rk}^{b1})}{ln\ 2\lambda_r} \tag{4.60}$$

$$p_{b2}^r = \frac{B_k P_{b1}^r (1+\hat{\gamma}_{rk}^{b2})}{\lambda_r ln\ 2P_{b1}^r + B_k(1+\hat{\gamma}_r k^b 1)(-\hat{\gamma}_{rk}^{b1})} \tag{4.61}$$

where

$$\gamma_{r,k}^{b_1} = \frac{\left|f_{r,k}^{b_1}\right|^2 p_{b1}^r}{\left|f_{r,k}^{b_2}\right|^2 p_{b2}^r + \sum_{n=1}^{N}\left|g_{r,k}^{mb_1}\right|^2 p^m + \sigma^2} \tag{4.62}$$

The optimal power allocation for MUEs is derived as:

$$p_{n1}^{m} = \frac{B_k(1+\hat{\gamma}_{mk}^{n1})}{ln\ 2\lambda_m} \tag{4.63}$$

$$p_{n2}^{r} = \frac{B_k P_{n1}^{r}(1+\hat{\gamma}_{mk}^{b2)}}{\lambda_m ln\ 2P_{n1}^{m} + B_k(1+\hat{\gamma}_{mk}^{n1})(-\hat{\gamma}_{mk}^{n1})} \tag{4.64}$$

where

$$\gamma_{m,k}^{n_1} = \frac{|f_{m,k}^{n_1}|^2 p_{n1}^{m}}{|f_{r,k}^{n_2}|^2 p_{n2}^{m} + \sum_{k=1}^{K}\sum_{b=1}^{B}|g_{m,k}^{rn_1}|^2 p^{r} + \sigma^2} \tag{4.65}$$

Optimal power allocation for CS-UE is derived as:

$$p_{c'}^{T}\lambda_c = p_c^r f_{r,k}^u + p_c^m g_{m,k}^u \geq \left(1 - \frac{1}{\Lambda}\right)(\Lambda^r + \Lambda^m + 1) \tag{4.66}$$

where

$$\Lambda = \left(1 + \frac{a_m^n P_t^m g_{mk}^c}{B_k} + \frac{a_r^b P_t^r f_{rk}^c}{B_k}\right) \tag{4.67}$$

and $\Lambda^m = P_t^m g_{mk}^c$, $\Lambda^r = P_t^r f_{rk}^c$. a_r^b and a_m^n are the power allocation coefficients. For derivations please refer Appendix D.

The subgradient method can be used to update the dual vectors λ_r and λ_m in each iteration.

$$\lambda_r(i+1) = \left[\lambda_r(i) - s_1(i)\left(P_{thr}^r - \sum_{k=1}^{K}\left(\sum_{b=1}^{B} p_{r,k}^b + \sum_{c=1}^{\phi_c} p_c^r\right)\right)\right]^+ \tag{4.68}$$

$$\lambda_m(i+1) = \left[\lambda_m(i) - s_2(i)\left(P_{thr}^m - \sum_{k=1}^{K}\left(\sum_{n=1}^{N} p_{m,k}^n + \sum_{c=1}^{\phi_c} p_c^m\right)\right)\right]^+ \tag{4.69}$$

In the subproblem G_2, $U_\alpha(x)$ is a concave function of x and hence $f(x)$ is also a concave function of x.Therefore the optimal solution x^* can be obtained by taking the derivative of $f(x)$ with respect to x and equating it to zero.

$$x^* = \begin{cases} 1,\ if\lambda_n\ >\ 1 \\ \max_x U_\alpha(x)\ if\ \lambda_n \leq\ 1 \end{cases} \tag{4.70}$$

The proposed joint user association, scheduling and power allocation scheme is summarized in Algorithm 7.

Algorithm 7: Joint user association, scheduling and power control

1. Input P^m, P^r where the vecor $P = [P^m, P_1^r, \ldots, P^R]$
2. While not converge iteratively optimize P^m and P^r do
3. Determine the optimal user association and scheduling from Algorithm 5 and Algorithm 6 i.e $x_{ij}^r, x_{ij}^m, x_{ij}^{cs}$
4. Given $x_{ij}^r, x_{ij}^m, x_{ij}^{cs}$ update the transmit power $P = [P^m, P_1^r, \ldots, P^R]$ with the steps as follows:

a) Obtain P^r and P^m according to equations(4.60, 4.61, 4.63, 4.64)
b) Update λ_n using subgradient method.
c) Until convergence
5. Obtain the optimal solution $U(p^*, \beta^*)$
$i = i + 1$
end if

4.7 Simulation Results

In this section we present simulation results to evaluate the performance of the proposed algorithm. We consider several RRHs located in the coverage of one MBS with 1 km diameter. Obviously the UEs closer to MBS prefers to access the network via MBS. It is assumed that the RRHs are uniformly distributed in a ring area centered around the MBS with the radius of inner ring to be 250 m and the outer radius of the ring is equal to the radius of the cell. The distance of MUEs and RUEs are measured with respect to their serving MBS and RRHs respectively. Each RRH has a coverage radius of D_R. If the distance between UE and RRH is within D_R, the UE chooses to access the network via the RRH. The distance of UEs in the CS-NOMA range are measured with respect to RRH. We assume that all fronthaul links are identical, therefore $R_{fh}^r = R^r$, where R_{fh}^r is the maximum traffic load that can be carried by the fronthaul link associated with RRH r. The simulation parameters are listed in Table 4.1.

The maximum number of users that can be multiplexed on the same RB is 2. Moreover we assume that the power sharing coefficients of NOMA for each tier are same i.e. $a_r^{b1} = a_m^{n1}$ and $a_r^{b2} = a_m^{n2}$.

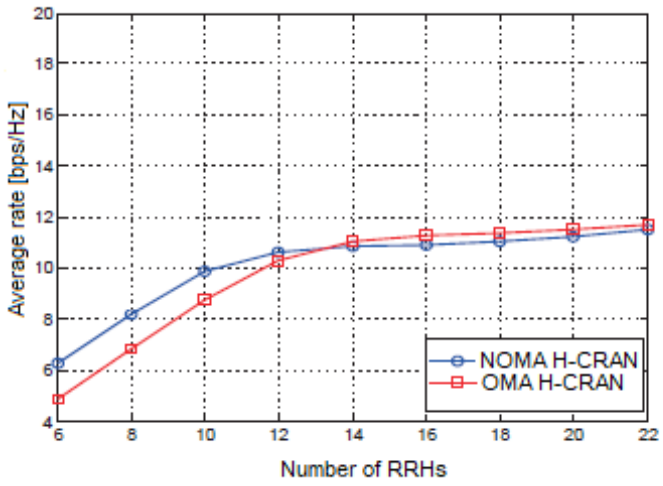

Figure 4.2: Average rates for NOMA and OMA H-CRANs ($\{a_r^{b1}, a_r^{b2}\} = \{a_m^{n1}, a_m^{n2}\} = \{0.3, 0.7\}$)

The performance of OMA based H-CRAN is illustrated as benchmark to demonstrate the effectiveness of our proposed NOMA-enabled H-CRAN system. Fig. 4.2 compares the average rates for coordinated scheduling scheme for NOMA H-CRAN with OMA H-CRAN. We observe that when the number of RRHs increases, the sum rates for NOMA H-CRAN and OMA H-CRAN first increases and then begin to decrease if the number of RRHs exceeds certain threshold. We observe that MBS and RRH coordination contribute to increase in average achieveble rate for both schemes when there are low to medium number of RRHs. However the average rates of both schemes have an intersection at some specific threhold. This indicates that NOMA

Table 4.1: Simulation parameters

Parameter	Values
Distance dependent path-loss from MBS to RUE	$140.7 + 36.7 * log_{10} d_b$, d in km
Distance dependent path-loss from RRH to RUE	$128.1 + 37.6 * log_{10} d_n$, d in km
Number of antenna at BS/UE	1
Scheduler	Proportional Fairness
Scheduling Interval	1ms
Maximum Tx power of MBS	43 dBm
Maximum Tx power of RRH	29 dBm
Throughput Calculation	Based in Shannon's Formula
Fronthaul Capacity C_{max}	variable

H-CRAN outperforms OMA H-CRAN only when the number of RRHs are below some threshold. After certain threshold OMA HetNets can be better choice than NOMA H-CRAN to improve average rate. This is due to the fact that performance of NOMA depends on the channel gain difference between the users which becomes less with increasing number of RRHs.

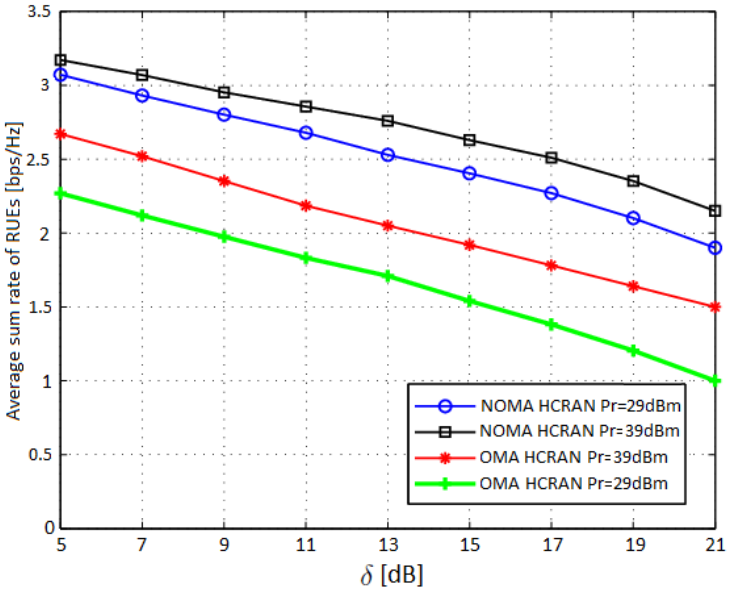

Figure 4.3: Average sum rate of RUEs vs δ $(\{a_r^{b1}, a_r^{b2}\} = \{a_m^{n1}, a_m^{n2}\} = \{0.3, 0.7\})$

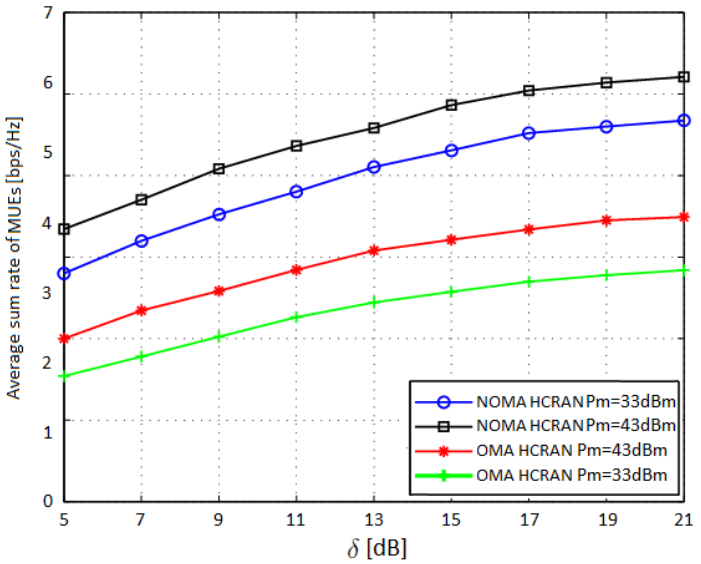

Figure 4.4: Average sum rate of MUEs vs δ $(\{a_r^{b1}, a_r^{b2}\} = \{a_m^{n1}, a_m^{n2}\} = \{0.3, 0.7\})$

Fig. 4.3 shows the average sum rate of RUEs with NOMA HCRAN and OMA H-CRAN versus δ for different transmit powers where δ is the factor for interference as discussed in section 4. We can observe that the average sum-rate of RUEs decreases with the increase in factor δ. This degradation can be explained as follows: larger value of δ means increase in main interference. The proportion of RRHs which use CoMP with large size increases due to increase in main interference. This is because the RRHs are more likely to cooperate. Moreover more users are associated with RRHs with low SINR which in turn degrades the performance. It is shown that the average sum-rate performance of NOMA enabled H-CRAN RUEs outperforms OMA H-CRAN.

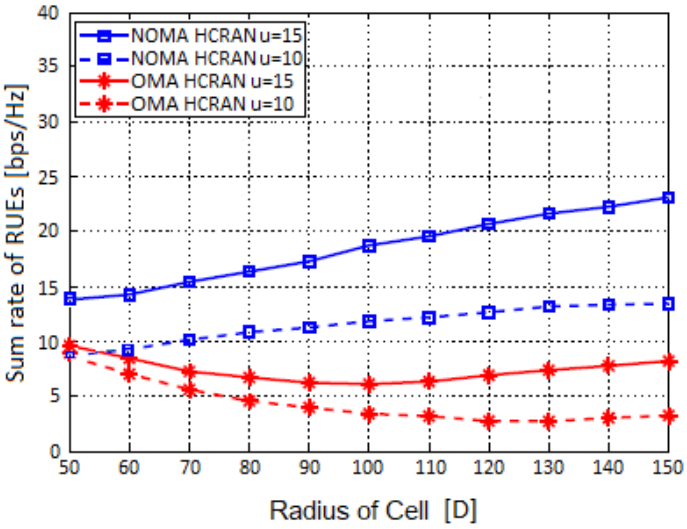

Figure 4.5: Average sum rate for RUEs vs D $(\{a_r^{b1}, a_r^{b2}\} = \{a_m^{n1}, a_m^{n2}\} = \{0.3, 0.7\})$

Fig. 4.4 shows the average sum-rate of MUEs with NOMA-enabled H-CRAN and OMA H-CRAN versus δ for different transmit powers. It is observed that the average sum-rate of MUEs improves with the factor δ. This is because when low SINR users are associated to RRHs, the sum-rate of MUEs is enhanced.

Fig. 4.5 demonstrates the sum-rate of RUEs as a function of radius of RRH coverage for different number of UEs. As the distance D_r increases, the SINR of RUEs is reduced owing to severe interference in case of OMA H-CRAN. Since the proposed coordinated scheduling cancels out interference from the coordinated MBS, the RUEs receive least interference and exihibit performance improvement over the OMA H-CRAN. However when the RRH radius is low the NOMA-enabled H-CRAN do not show much improvement in performance due to inherently remaining inter-user interference from NOMA transmission [102].

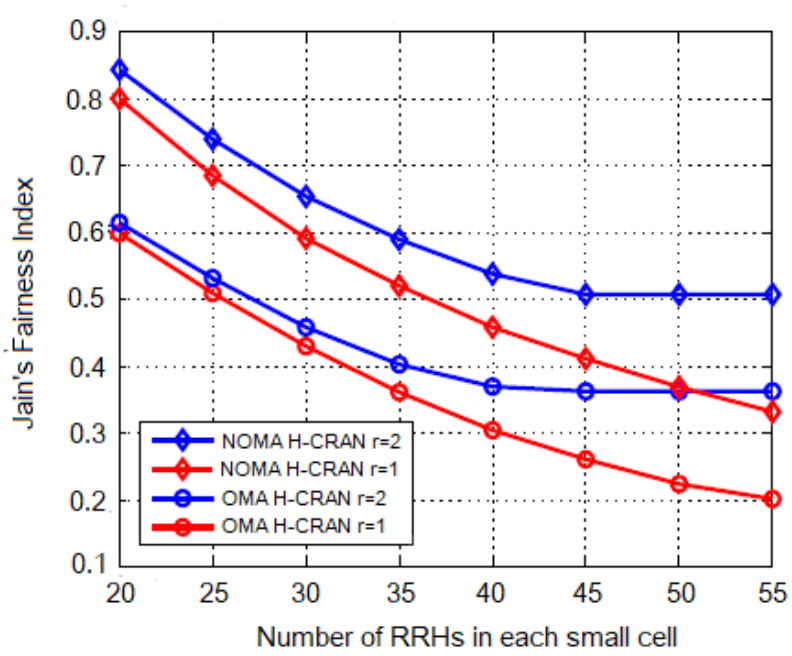

Figure 4.6: Jain's fairness index vs number of RRHs, transmit power at MBS is 43 dBm and the transmit power at RRHs is 29 dBm

Fig. 4.6 shows the relationship between the Jain's fairness index and number of RRHs for fixed number of RBs. To provide measurement for fairness, Jain's fairness index is used. The fairness index [98] considering heterogeneity in H-CRAN is defined as:

$$J_{fi} = \frac{\left[\sum_{n=1}^{N}(R_{n1}+R_{n2}) + \sum_{b=1}^{B}\beta(R_{b1}+R_{b2})\right]^2}{N_u\left[\left(\sum_{n=1}^{N}(R_{n1}+R_{n2})^2 + \sum_{b=1}^{B}\beta^2(R_{b1}+R_{b2})^2\right)\right]} \tag{4.71}$$

where β is used to measure the relative throughput of two-tiers and is defined as:

$$\beta = \frac{\frac{1}{N}\sum_{n=1}^{N}(R_{n1}+R_{n2})}{\frac{1}{B}\sum_{b=1}^{B}(R_{b1}+R_{b2})} \tag{4.72}$$

The value of Jain's fairness index is between 0 and 1. The rate allocation is perfectly fair if $J_{fi} = 1$. For a given number of RBs, we observe that the Jain's fairness index decreases with the increasing number of RRHs. This occurs because the aggregated interference experienced by users due to overcoverage is more complicated. Moreover the users with poor channel conditions may not be accessed by network due to more competitiveness for limited resources. It is worth noting that as r increases, a higher fairness level can be achieved. This is due to the fact that more small cell RRHs can be multiplexed on each RB which increases the multi-user diversity gain. We can observe that the fairness level is significantly improved with the proposed NOMA-enabled H-CRAN compared to OMA H-CRAN especially when the number of RRHs are in low to medium range.

4.8 Conclusions

In this chapter joint coordinated scheduling and power control for NOMA-enabled H-CRAN is proposed. For the coordinated scheduling scheme, iterative power allocation scheme is proposed for the NOMA H-CRAN and the optimal power allocation for each user is derived by hierarchical decomposition method.The simulation results demonstrate that the proposed iterative algorithm to optimize the scheduling and

power iteratively increases the average data rate and network utilisation with proportional fairness consideration for NOMA-enabled H-CRAN as compared to OMA H-CRAN.

Chapter 5

Performance Analysis of NOMA in Fog Radio Access Networks (F-RAN)

5.1 Introduction

In the fifth generation (5G) era, mobile networks will provide diversified services such as enhanced Mobile BroadBand (eMBB) and ultra-reliable low-latency communication (URLLC). Furthermore, the 5G architecture will be fog-like, enabling a functional split of network functionalities between cloud and edge nodes. In this chapter, we propose a non-orthogonal multiple access (NOMA)-enabled fog-cloud structure in a novel density-aware fog-based radio access networks (F-RAN) to tackle different aspects such as high throughput and low-latency requirements of high and low user-density regions, in order to meet the heterogeneous requirements of eMBB and URLLC traffic. A framework of the multi-objective problem is formulated to cater the high throughput and low-latency requirements in a high and low user-density mode respectively. In the first problem, we study the joint caching placement and association strategy aimed at minimizing the average delay. To deal with the first problem, we apply the McCormick envelopes and Lagrange partial relaxation method to transform the problem into three convex sub-problems, which are

then solved by proposed distributed algorithm. The second problem is to jointly optimize transmission mode selection, subchannel assignment and power allocation to maximize the sum data rate of all fog UEs (F-UEs) while satisfying fronthaul capacity and fog-computing access point (F-AP) power constraints. Moreover, for given transmission mode selection and subchannel assignment, the optimal power allocation is derived in a closed-form. Simulation results are provided to validate the effectiveness of the proposed NOMA-enabled F-RAN framework and reveal that the ultra-low latency and high throughput can be achieved by properly utilizing the available resources.

5.2 System model of NOMA-enabled F-RAN

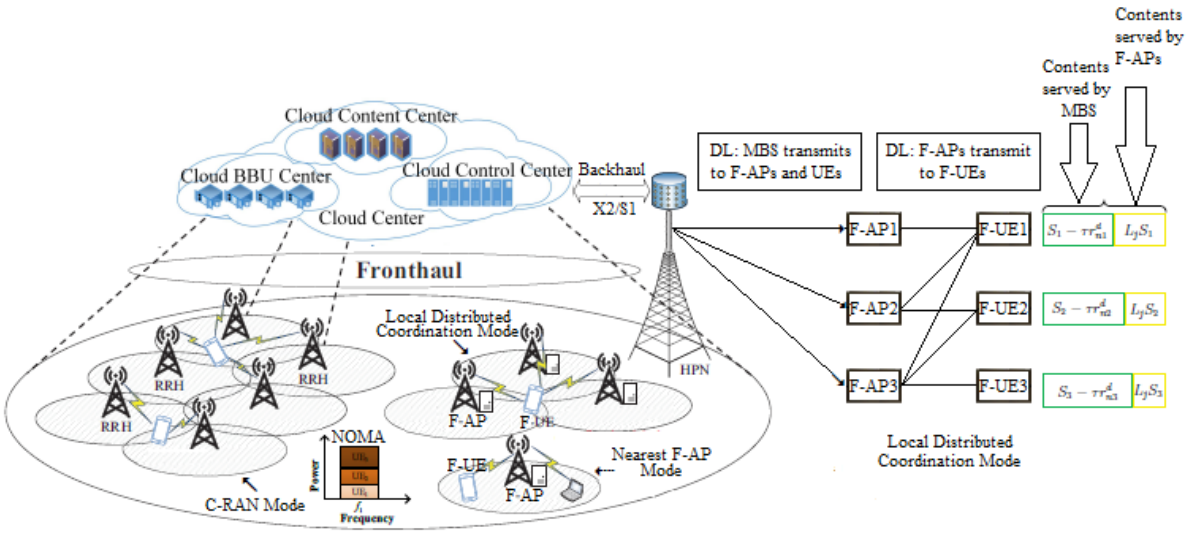

Figure 5.1: System Model for NOMA-enabled F-RAN

We consider a downlink direction of NOMA-enabled F-RAN as shown in Fig. 5.1. UEs, and the fog access points (F-APs) are connected to a macro base station

(MBS), or a cloud server, through a shared wireless fronthaul link. We define $\mathcal{N} = \{1, 2, \ldots, n\}$ and $\mathcal{M} = \{N+1, N+2, \ldots, N+M\}$ as the index of F-UEs and UEs directly served by MBS respectively.

We denote the set of F-APs in a macrocell as $\mathcal{F} = \{1, 2, \ldots, F\}$. $\mathcal{F}_m = F + 1$ is the total number of MBS and F-APs denoted by set $\mathcal{F}_m = \{0, 1, 2, \ldots, \mathcal{F}\}$ where 0 is the index of the MBS. In this work, we assume the wireless fronthaul transmission between MBS and F-APs (to fetch missed cached contents directly) operates concurrently with the wireless access transmission from MBS to all UEs. This means the user will be served directly by MBS if the file requested cannot be found directly at local F-AP. Using the NOMA principle, MBS sends a superposition signal containing the file requested by UE and at the same time push the new content to F-APs. If the file requested by user is cached at F-APs, multiple F-APs can communicate with their associated users to boost the transmission cooperation in the access link. In order to improve the spectrum efficiency we assume UEs served directly by MBS and F-UEs reuse the same set of resource blocks (RBs). The F-AP is equipped with a cache of finite memory storing ($nB_i > 0$) some of the popular contents which might be requested by the UEs. We denote the maximum cache capacity $S_i^{max}, 1 \leq i \leq \mathcal{F}$. Let J denote the number of content files and L_j denote the size of the file j, $1 \leq j \leq J$. Assume that all the files in the library $L = \{1, 2, \ldots, L\}$ stored in the BBU are of same same size of nS bits. The F-AP pre-stores S_i^{max}/S files in its cache. In case the UEs required content is stored in the local cache at its associated F-AP, UE will retrieve the content directly from the F-AP without interacting with the remote server. The total bandwidth B is divided into K resource blocks (RBs) indexed as $\mathcal{K} = \{1, 2, \ldots, K\}$ and each RB

occupies a bandwidth of $B_k = B/K$.

A single RB can be assigned to multiple number of users. We denote p^m and p^f as transmit power from MBS and F-AP to mth UE and nth F-UE respectively. $h_{f,k}^{n_1} = |\tilde{h}_{f,k}^{n_1}|^2 d_{n_1}^{-\alpha}$ with $\tilde{h}_{f,k}^{n_1} \sim \mathcal{CN}(0,1)$, where $h_{f,k}^{n_1}$ is the Rayleigh fading channel coefficient, α represents path loss exponent and d_{n_1} is the distance between F-AP f and F-UE n. $g_{f,k}^{mn_1} = |\tilde{g}_{f,k}^{mn_1}|^2 d_{mn_1}^{-\alpha}$ and $g_{f,k}^{mn_2} = |\tilde{g}_{f,k}^{mn_2}|^2 d_{mn_2}^{-\alpha}$ are the channel coefficients between MBS m and F-UE n_1 and n_2, d_{mn_1} and d_{mn_2} are the distances between MBS m and F-UEs n_1 and n_2 respectively. β_{bk}^f represents the RB indicator for F-APs i.e. if RB k is assigned to F-AP f then $\beta_{bk}^f = 1$ and 0 otherwise. β_{nk}^m represents the RB indicator for UEs served directly by MBS.

According to the NOMA-based transmission, multiplexing of users through superposition coding (SC) at the MBS and successive interference cancellation (SIC) technique is implemented at the UEs. Therefore the most popular files belonging to the same library can be pushed. The MBS superimposes S_i popular files. The F-APs carry out SIC. The SIC decoding order is based on the priorty of the files. The F-AP experiences the interference from other F-APs whose contents are more popular. The F-APs decodes its signal from other F-APs whose channel gains after that F-AP depending on the order of channel gain.

Suppose that the SIC decoding order is determined by the popularity of the files, i.e a more popular file s_i is to be pushed at f_1 F-AP before the less popular content s_j at F-AP f_2, $i < j$. The f_2^{th} F-AP decodes the f_1^{th} F-APs message. Similarly at f_3^{th} F-AP, the signals of f_1 ane f_2 are decoded first.The achievable rate at the F-AP

is given by:

$$\gamma_{m,k}^{n_1} = \frac{|h_{m,k}^{n_1}|^2 p^m a_{f,k}^{n_1}}{|h_{f,k}^{n_2}|^2 p^m a_{f,k}^{n_2} + \sum_{k=1}^{K}\sum_{b=1}^{B} \phi_{b,k} |g_{m,k}^{fn_1}|^2 p^f + \sigma^2} \tag{5.1}$$

F-AP sends messages to FUEs n_1 and n_2 on RB k by superposition i.e. F-AP f sends $a_{f,k}^{n_1} x_{f,k}^{n_1} + a_{f,k}^{n_2} x_{f,k}^{n_2}$ where $a_{f,k}^{n_1}$ and $a_{f,k}^{n_2}$ are the power sharing coefficients. $x_{f,k}^{n_1}$ is the symbols transmitted from fth F-AP to its serving FUE n_1. The received signal by F-UE n_1 on RB k is given by:

$$y_{f,k}^{n_1} = h_{f,k}^{n_1}\sqrt{p^f a_{f,k}^{n_1}} x_{f,k}^{n_1} + h_{f,k}^{n_1}\sqrt{p^f a_{f,k}^{n_2}} x_{f,k}^{n_2} + \sum_{n=1}^{N} \beta_{n,k} g_{f,k}^{mn_1} \sqrt{p_{n,k}^{m}} x_{n,k}^{m} + \xi_{f,k}^{n_1} \tag{5.2}$$

The first term in the above equation is desired received signal, the second term is the interference from the neighbouring F-UEs, the third term is the cross-tier interference and $\xi_{b,k}^{f}$ is the additive white gaussian noise (AWGN) at F-UE n_1 with variance σ^2.

The channel gains are sorted as $|f_{b,k}^{r}| \geq \ldots \geq |f_{1,k}^{r}|$

The received SINR at F-UE n served by F-AP considering cross-tier interference is represented by:

$$\gamma_{f,k}^{n} = \frac{|h_{f,k}^{n}|^2 p^f}{\sum_{l=b+1}^{U} |h_{f,k}^{n}|^2 p^f + \sum_{m=1}^{M} \beta_{nk} |g_{f,k}^{mn}|^2 p^m + \sigma^2} \tag{5.3}$$

The received SINR at FUE n_1 served by F-AP f on RB k is given by

$$\gamma_{f,k}^{n_1} = \frac{|h_{f,k}^{n_1}|^2 p^r a_{f,k}^{n_1}}{|h_{f,k}^{n_2}|^2 p^f a_{f,k}^{n_2} + \sum_{n=1}^{N} \beta_{nk} |g_{f,k}^{mn_1}|^2 p^m + \sigma^2} \tag{5.4}$$

After decoding SINR of FUE n_2 is given by:

$$\gamma_{f,k}^{n_2} = \frac{|h_{f,k}^{n_2}|^2 p^f a_{f,k}^{n_2}}{\sum_{n=1}^{N} \beta_{nk} |g_{f,k}^{mn_2}|^2 p^m + \sigma^2} \tag{5.5}$$

Given the SINR values, the achievable data rate in terms of bit/s/Hz for FUEs n_1 and n_2 can be calculated using Shannon formula as:

$$R_{n_1} = B_k log_2(1 + \gamma_{f,k}^{n_1}) \tag{5.6}$$

and

$$R_{n_2} = B_k log_2(1 + \gamma_{f,k}^{n_2}) \tag{5.7}$$

The SINR of UE directly served by MBS considering interference from F-APs is given by:

$$\gamma_{m,k}^{m1} = \frac{|h_{m,k}^{m_1}|^2 p^m a_{f,k}^{n_1}}{|h_{f,k}^{m_2}|^2 p^m a_{f,k}^{m_2} + \sum_{k=1}^{K} \sum_{b=1}^{B} \phi_{b,k} |g_{m,k}^{fm_1}|^2 p^f + \sigma^2} \tag{5.8}$$

After decoding, SINR at F-AP is given by:

$$\gamma_{m,k}^{m2} = \frac{|f_{m,k}^{m_2}|^2 p^m a_{m,k}^{m_2}}{\sum_{k=1}^{K} \sum_{b=1}^{B} \phi_{b,k} |g_{m,k}^{fm_2}|^2 p^f + \sigma^2} \tag{5.9}$$

Their corresponding data rates are given by:

$$R_{m_1} = B_k log_2(1 + \gamma_{m,k}^{m_1}) \tag{5.10}$$

and

$$R_{m_2} = B_k log_2(1 + \gamma_{m,k}^{m_2}) \tag{5.11}$$

In this thesis, we consider users will access to the F-RAN by three user-centric access modes according to users' communication distance, content caching strategy and the QoS requirements. Three modes are the nearest F-AP mode, local distributed coordinated mode, global C-RAN mode. We divide the user requests at each F-AP into three modes according to their service routes 1, 2 and 3. To serve the request for content f at F-AP f, the associated F-AP checks its own cache and delivers the content directly to the user if this content is cached termed as nearest F-AP

mode. Otherwise in local coordination mode F-AP fetches the file from the chosen coordinated F-AP and send it to the user. In case all the F-APs have not cached this content , then the F-AP receive it from the BBU via fronthaul link and delivers it to the UE denoted as global C-RAN mode.

The serving F-APs jointly decide to transmit data to users, whose requested files are stored in local cache. When the file j, which is the requested file index of user n is cached in one or multiple F-APs, the F-APs transmit this file to the user based on either cooperative or non-cooperative transmission schemes. The user's delivery rate performance depends not only on the cache placement from F-APs or central processor (CP) but also on the transmission scheme adopted to deliver the files. If the F-AP is unable to serve the users' request with cooperative or non- cooperative transmission then the F-AP forwards the requests to the CP.

5.2.1 Non-Cooperative Transmission Scheme

When a user is served by a single F-AP, a non-cooperative scheme is employed to transmit the file to the UE n directly. When the controlling F-AP do not select the coordinating F-AP via the indicator function $c^b_{f'} = 0$, the scheme is employed by the associated F-AP, if the requested file is cached in this F-AP. The delivery rate of user n served by F-AP f, for non-cooperative transmission is given by:

$$R^k_n = \sum_{f\in F}\sum_{k\in K} c^n_{f'f}\left(1 + \frac{|h^n_{f,k}|^2 p^f}{\sum_{l=b+1}^{U}|h^n_{f,k}|^2 p^f + \sum_{m=1}^{M}\beta_{nk}|g^{mn}_{f,k}|^2 p^m + \sigma^2}\right) \tag{5.12}$$

5.2.2 Cooperative Transmission Scheme

When the UE n is served by multiple F-APs and the computing tasks are sent to the grouping F-APs by controlling F-AP, cooperative transmission scheme can be

applied. The delivery rate can be represented by:

$$R_n^{ck} = \sum_{f\in F}\sum_{k\in K} c^n_{f'f}\left(1+\sum_{f_m\in C}\frac{|h^n_{f,k}|^2 p^f}{\sum_{l=b+1}^{U}|h^n_{f,k}|^2 p^f + \sum_{m=1}^{M}\beta_{nk}|g^{mn}_{f,k}|^2 p^m + \sigma^2}\right) \quad (5.13)$$

where $C = \{f_m \in F | m_{nf} = 1\}$ denotes a set of serving F-APs that transmit file to UE n_m via cooperative transmission scheme.

5.3 Problem Formulations

In this work two problems making use of cache and signal processing capability of each F-AP for low and high density modes with different objective functions are proposed. First, we address the problem of minimizing the average latency of FUEs whose requests at each F-AP are divided into three service routes. The second aim of this work is to maximize the total sum rate achieved by considering both macrocell and F-APs.

5.3.1 Low-Density Mode: Latency Problem

For each controlling or master F-AP, there are other F-APs defined as cooperated F-APs, responsible for computing tasks, content fetching and data transmission for users. The number of cooperated F-APs is dynamic according to the requirement of users, network environment, computing and communication resources.

For coordinated task computing, we denote the binary indicator variable $c^n_{f'}$ which decide whether F-AP $f^{'}$ is chosen as controlling F-AP, i.e., $c^n_{f'} = 1$ if selected as controlling F-AP of user b; and 0 otherwise. We denote the binary association variable between UE and cache content c, where

$$\alpha^i = \begin{cases} 1 \text{ if content c is cached by F-AP} \\ 0, \text{ otherwise} \end{cases} \quad (5.14)$$

Let r_n^d denote the file delivery rate $r_n^d < S_l$ of user n served by F-AP f. In a given transmission interval total $nr_n^d \leq nS_l$ bits are transmistted to the UE n. In order to determine the set of F-APs to which the content is fetched from CP via fronthaul. we define the binary variable d_f^i as:

$$d_f^i = \begin{cases} 1, \text{ if content c is transferred to F-AP} \\ 0, \text{ otherwise} \end{cases} \tag{5.15}$$

The fronthaul capacity constraint for each F-AP is represented as:

$$\sum_{l \in L} \sum_{j \in J} d_f^i r_n^d \leq R_f \tag{5.16}$$

where R_f is the rate at F-AP f on subchannel k.

In case of NOMA-enabled F-RAN, the coordinated F-APs transmit signals in non-orthogonal manner to the n-th user on the same subchannel. The signals received from multiple F-APs having distinct channel gain can be ordered in descending order according to their received signal strength. The connection matrix between UEs and F-APs is represented by a $N \times F$ matrix M where each element m_{nf} denotes whether UE n is served by F-AP f i.e. $m_{nf} = 1$ if UE n_m is served by F-AP f: $m_{nf} = 0$, otherwise. The set of serving F-APs of UE n_m is represented as $C = \{f_m \in F | m_{nf} = 1\}$. Similarly the set of users served by F-AP f is represented as $U_m = \{f_m \in F | m_{nf} = 1\}$. We denote $c_{f'f}^n = m_{nf} \times b_n^k$ as coordinated F-APs and subchannel assignment for UE. The chosen controlling F-AP decides whether the F-AP is grouped or not for serving UE via binary association variable $c_{f'f}^n$ i.e. variable $c_{f'f}^n = 1$, if selected as coordinated F-AP; and 0 otherwise. The complexity of SIC receiver depends on the number of F-APs in the same group on the same subchannel. In this work we assume each user can be served by F-APs formed with

upto $\bar{f}$ F-APs on the same subchannel.

$$\sum_{f \in F} c^n_{f'f} \leq \bar{f} \tag{5.17}$$

For each master F-AP f, the sum of allocated RBs for all the controlling F-APs to serve all UEs cannot exceed the total number of RBs.

$$\sum_{b \in B} \sum_{f \in F} c^n_{f'} c^n_{f'f} K_f \leq K \quad \forall f \in F \tag{5.18}$$

where K_f is the number of RBs allocated to F-AP by controlling F-AP.

The cache capacity constraint for F-AP is

$$\sum_{b \in B} c^n_{f'} c^n_{f'f} S \leq S^{max}_i \quad \forall f \in F \tag{5.19}$$

For F-AP computing, the device offloads its tasks to the F-AP via the wireless links. The computation capacity for task execution in terms of CPU cycles per second is defined as G_f which is the total number of computing tasks. If user n sends $L_j S$ contents to edge device to offload network traffic, an additional transmission delay is caused by transmitting the computation data to the edge device and is given by $\frac{L_j S}{R^k_n}$ in (5.20).

We formulate a problem for achieving ultra low latency in NOMA-enabled F-RAN system by satisfying limited fronthaul capacity, caching and power constraints. We assume that user 2 has stringent delay requirements than user 1.

$$\min_{c^n_{f'f}, \alpha_i} \max_{k \in \{1,2\} s \in \{1,2\} \forall n \in N, \forall f' \in F, \forall f' \in \bar{F}} c^n_{f'f} \left[\frac{L_j S_{ks}}{R^{ks}_n} + \frac{(1 - c^n_{f'}) \times G_f}{K_f \times R_c} + \frac{C_{f'f}}{S \times R_{ac}} \right] \tag{5.20a}$$

s.t

$$\sum_{b \in B} \sum_{f \in F} c^n_{f'} c^n_{f'f} K_f \leq K \quad \forall f \in F \tag{5.20b}$$

$$\sum_{b \in B} c^n_{f'} c^n_{f'f} S \leq S^{max}_i \quad \forall f \in F \tag{5.20c}$$

$$\sum_{f \in F} c^n_{f'f} \leq \bar{f} \tag{5.20d}$$

$$\sum_{n \in N} nS_l = \sum_{n \in N} \sum_{F \in f} L_j S + \sum_{n \in N} \sum_{m \in M} (S - \tau r^d_n) \tag{5.20e}$$

where $L_j S$ denotes the amount of requested contents served by single F-AP to UE n and R^k_n is the achievable rate of user n. R_{ac} is the achievable data rate between controlling F-AP and the coordinated F-APs. G_f is the amount of processing data, R_f is the computational rate of F-APs. Constraint (5.20e) states that the contents requested by UE n can be served both by F-AP and CP. If the controlling F-APs and coordinated F-APs are not able to serve the requests, one or more F-APs should fetch the contents from the cloud processor and deliver the content to the user. Let D_f denote the additional transmission delay incurred at the CP, i.e. from the remote server to the F-AP represented by:

$$D_f = (1 - \alpha_i) \frac{S - \tau r^d_n}{R_f} \tag{5.21}$$

where τ is the transmission delay during which cached files are transmitted to the UEs . We propose a caching mechanism which reduces the latency τ and the burden on the fronthaul as per the above constraint. The total latency at the F-APs and CP can be represented as:

$$\begin{aligned} \min_{c^n_{f'f}, \alpha_i} \quad \max_{k \in \{1,2\} s \in \{1,2\} \forall n \in N, \forall f' \in F, \forall f' \in \bar{F}} \quad & c^n_{f'f} \left[\frac{L_j S}{R^k_n} + \frac{(1 - c^n_{f'}) \times G_f}{K_f \times R_c} \right. \\ & \left. + \frac{C_{f'f}}{S \times R_{ac}} \right] + c^n_{f'f} (1 - \alpha_i) \frac{S - \tau r^d_n}{R_f} \end{aligned} \tag{5.22}$$

s.t. 5.20(b), 5.20(c), 5.20(d)

$$r^r_f \leq (1 - \alpha_i) max(S - \tau r^d_n) \tag{5.23}$$

Based on the given subchannel assignment, we consider the delay $\tau = \left[\frac{L_j S}{R^k_n} + \frac{(1 - c^n_{f'}) \times G_f}{K_f \times R_c} + \frac{C_{f'f}}{S \times R_{ac}} \right]$ to be fixed value. In order to decouple the user association

and the caching variable in the optimization problem, we introduce a new variable $u_m = c^n_{f'f}(1-\alpha_i)$. Therefore the optimization problem can be rewritten as:

$$\min_{c^n_{f'f},\alpha_i} \max_{k\in\{1,2\}s\in\{1,2\}\forall n\in N,\forall f^{'}\in F,\forall f^{'}\in\bar{F}} c^n_{f'f}\tau + u_m\frac{S-\tau r^d_n}{R_f} \tag{5.24}$$

s.t. 5.20(b), 5.20(c), 5.20(d) and

$$u_m = c^n_{f'f}(1-\alpha_i)\ \forall n\in N, f\in F, b\in B \tag{5.25}$$

where (5.25) is a non-convex constraint. By relaxing the constraint by using McCormick envelopes [103], it can be represented as:

$$u_m \geq 0\ \forall n\in N, f\in F, b\in B \tag{5.26}$$

$$u_m \geq c^n_{f'f} - \alpha_i\ \forall n\in N, f\in F, b\in B \tag{5.27}$$

$$u_m \leq c^n_{f'f}\ \forall n\in N, f\in F, b\in B \tag{5.28}$$

$$u_m \leq 1-\alpha_i\ \forall n\in N, f\in F, b\in B \tag{5.29}$$

We introduce the Lagrange partial selection method [74] to solve the optimization problem in (5.22). Specifically, we relax the constraints (5.27),(5.28) and (5.29) and define the set of dual Lagrange multipliers as:

$$\mu^{nb}_f \geq 0\ \forall n\in N, f\in F, b\in B \tag{5.30}$$

$$\lambda^{nb}_f \geq 0\ \forall n\in N, f\in F, b\in B \tag{5.31}$$

$$\psi^{nb}_f \geq 0\ \forall n\in N, f\in F, b\in B \tag{5.32}$$

Hence, we obtain the Lagrange function as:

$$\begin{aligned} L(\mu^{nb}_f,\lambda^{nb}_f,\psi^{nb}_f,c^n_{f'f},\alpha_i,u_m) = c^n_{f'f}\tau + u_m\frac{S-\tau r^d_n}{R_f} + \mu^{nb}_f(c^n_{f'f}-\alpha_i-u_m) \\ +\lambda^{nb}_f(u_m - c^n_{f'f}) + \psi^{nb}_f(u_m+\alpha_i-1) \end{aligned} \tag{5.33}$$

The dual problem is represented by:

$$\max_{\mu_f^{nb},\lambda_f^{nb},\psi_f^{nb}} \min_{c^n_{f'f},\alpha_i} L(\mu_f^{nb},\lambda_f^{nb},\psi_f^{nb},c^n_{f'f},\alpha_i,u_m) \tag{5.34}$$

s.t. (5.26), (5.30), (5.31), (5.32) and 5.20(b), 5.20(c), 5.20(d)

Since there is no coupling between user association and caching variables, the optimization problem can be decomposed into three subproblems with independent feasible regions as:

$$SP1 : \min_{c^n_{f'f}} c^n_{f'f}(\tau + \mu_f^{nb} - \lambda_f^{nb}) s.t. 5.21(b,c,d) \tag{5.35}$$

$$SP2 : \min_{\alpha_i} \alpha_i(\psi_f^{nb} - \mu_f^{nb}) s.t. 5.24 \tag{5.36}$$

$$SP3 : \min_{u_m} u_m(\frac{S - \tau r_n^d}{R_f} - \mu_f^{nb} + \lambda_f^{nb} + \psi_f^{nb}) s.t. 5.27 \tag{5.37}$$

It can be shown that the three subproblems are constrained integer programming optimization problems for a given set of Lagrange multipliers. SP1 is a user association problem which can be solved by Hungarian method [104] and SP2 and SP3 can be solved using CVX [105] . After solving the three subproblems, we obtain the locally optimal solution of $c^{n*}_{f'f}$, α_i^* and u_m^*. Therefore the Lagrange multipliers can be updated using subgradient method as follows:

$$\mu_f^{nb}(t+1) = [\mu_f^{nb}(t) + \epsilon_1(c^n_{f'f} - \alpha_i - u_m)]^+ \tag{5.38}$$

$$\lambda_f^{nb}(t+1) = [\lambda_f^{nb}(t) + \epsilon_2(u_m - c^n_{f'f})]^+ \tag{5.39}$$

$$\psi_f^{nb}(t+1) = [\psi_f^{nb}(t) + \epsilon_3(u_m + \alpha_i - 1)]^+ \tag{5.40}$$

where $[x]^+ = max\{0, x\}$ and ϵ_1, ϵ_2 and ϵ_3 are the step sizes with regard to μ_f^{nb}, λ_f^{nb} and ψ_f^{nb} respectively. In order to reduce the complexity of the given problem, we propose a distributed algorithm. The algorithm is guaranteed to converge to the

optimal value by conducting the above process iteratively. Once the convergence is met, we can obtain the global optimal user association and caching placement strategy. The method is summarized as follows:

In order to reduce the complexity of the given problem, we propose a distributed algorithm. The proposed method is summarized in Algorithm 8. To obtain the optimal solution of the subproblems SP1, SP2 and SP3, the lagrange multipliers are solved with the optimization variables. The algorithm is guaranteed to converge to the optimal value by conducting the process iteratively. The first subproblem SP1 involves the UE-F-AP association. The second subproblem SP2 involves the caching variable and third subproblem SP3 only involves adding a new variable.

Algorithm 8: Proposed Decentralized Algorithm for solving the Optimization Problem

1. **Initialization**: Set $t = 0$, $\sigma(t) = 0$ and initial subchannel assignment $b_n^k(t)$
2. Lagrange multipliers initialization: $\mu_f^{nb}(t) = 0$, $\lambda_f^{nb}(t) = 0$ and $\psi_f^{nb}(t)$
3. **Repeat** $t = t + 1$
4. Solve subproblem SP1,SP2 and SP3 to obtain locally solution of $c_{f'f}^n$, α_i and u_m respectively.
5. Set $\sigma(t) = L(\mu_f^{nb}, \lambda_f^{nb}, \psi_f^{nb}, c_{f'f}^n, \alpha_i, u_m)$ and update the dual variable μ_f^{nb}, λ_f^{nb} and ψ_f^{nb} using (5.38), (5.39) and (5.40)
6. **if** $|\sigma(t) - \sigma(t-1)| \leq \epsilon$ **then**
7. Convergence = true
8. **return** $c_{f'f}^{n*} = c_{f'f}^n$, $\alpha_i^* = \alpha_i$, $u_m^* = u_m$
9. **else** $t = t + 1$
10. **end if**
11. **until** Convergence = true

5.3.2 High-Density Mode: Delivery Rate Maximization Problem

When the contents stored in F-APs or F-UEs is requested, the use of edge caching can alleviatie bandwidth requirements and reduce delay. By means of the coefficient (cache revenue), which adds the reward function in (5.41) to the utilization of edge caching, we can compute a certain degree of compensation from the finite edge

caching in NOMA-based F-RANs [71]. As stated in the previous literature [71], [106] and [107] the reward function of F-AP is represented as:

$$CR = \lambda \alpha_i \sum_{n \in N} \sum_{k \in K} G_f r_n^d \tag{5.41}$$

where r_n^d represents the request rate of F-UEs on subchannel k.. The request rate is related to the content requested UEs. All UEs are assumed to have the same request for contents. Therefore r_n^d can be expressed as a random number r_n^R. Therefore (5.41) can be represented as:

$$CR = \lambda \alpha_i \sum_{n \in N} \sum_{k \in K} r_n^R R_n^k \tag{5.42}$$

Let $\mathbf{P}$ and b_n^k are the matrices of dimensions $F \times K$ and $F \times K \times N$ which denotes the tansmit power and subchannel assignment. The access selection vector $\mathbf{d} = [d_f^i]$ which indicates whether content is fetched from CP via fronthaul or not.

In access mode selection phase, a user which requires content can select a communication independently according to the local cache. The total sum data rate of all FUEs can be expressed as:

$$\phi(P, b, \theta) = \sum_{f \in F} \sum_{k \in K} b_n^k (\sum_{n \in N} R_n^k + \sum_{m=N+1}^{M} d_f^i R_f) + \lambda \alpha_i \sum_{f \in F} \sum_{n \in N} \sum_{k \in K} G_f r_n^d \tag{5.43}$$

In NOMA, the power allocation should be configured properly for correct SIC. Besides, the F-APs total transmit power cannot exceed RRH maximum power capacity. The power allocation of F-AP should satisfy the following conditions:

$$b_n^k (\alpha_i p_{nk}^f |h_{fk}^n|^2 - \sum_{j=f+1}^{F} p_{jk}^f |h_{jk}^n|^2) \geq P_t \tag{5.44}$$

$$\sum_{k \in K} (\sum_{n \in N} p_{nk}^f + \sum_{m=N+1}^{M} P_k^m) \leq P_{max} \tag{5.45}$$

$$\sum_{f=1}^{F} x_{fn} p_k^f \leq p_{max} \tag{5.46}$$

where P_t is the detection threshold required to distinguish between the signal at the SIC receiver. Constraint (5.45) limits the maximum MBS and F-AP power. x_{fn} denotes the binary association variable between UE and F-AP i.e. variable $x_{fn} = 1$ indicates UE is associated with F-AP f; and 0 otherwise.

The problem is to maximize the sum-rate of the macro-cell based transmission and F-AP based transmission. In this work, we consider that the request files of the UEs are processed collaboratively in both cloud and fog. We tackle the problem of jointly optimizing transmission mode selection, subchannel assignment and power allocation to maximize the sum data rate of all F-UEs while satisfying fronthaul capacity and FAP power constraints. The optimization problem can then be formulated as:

$$\max_{P,b,\theta} \phi(P, b, \theta) \tag{5.47a}$$

$$b_n^k(\alpha_i p_{nk}^f |h_{fk}^n|^2 - \sum_{j=f+1}^{F} p_{jk}^f |h_{jk}^n|^2) \geq P_t \tag{5.47b}$$

$$\sum_{k\in K}(\sum_{n\in N} p_{nk}^f + \sum_{m=N+1}^{M} P_k^m) \leq P_{max} \tag{5.47c}$$

$$\sum_{f=1}^{F} x_{fn} p_k^f \leq p_{max} \tag{5.47d}$$

$$\sum_{l\in L}\sum_{j\in J} d_f^i r_n^d \leq R_f \tag{5.47e}$$

where the optimization is over the transmit power allocation $\mathbf{P}$, subchannel assignment and access selection matrices $\mathbf{b}$ and $\mathbf{d}$. The formulated problem is a mixed combinatorial non-convex problem due to binary variables β and real variable $p(t)$

as well as non-convex objective function. But the problem becomes more tractable if we seperate it into subproblems.

5.4 User Association and Interference Aware NOMA-enabled F-RAN

If F-APs and MBS share the same frequency band, cross-tier interference is serious in such co-channel deployment. Without effective management of cross-tier interference in the co-channel deployment of F-RAN, both the system throughput would be largely limited. We consider a simple scenario where only one MBS and several F-APs within the coverage area of the MBS are involved. The F-APs share the same spectrum resources with the MBS.

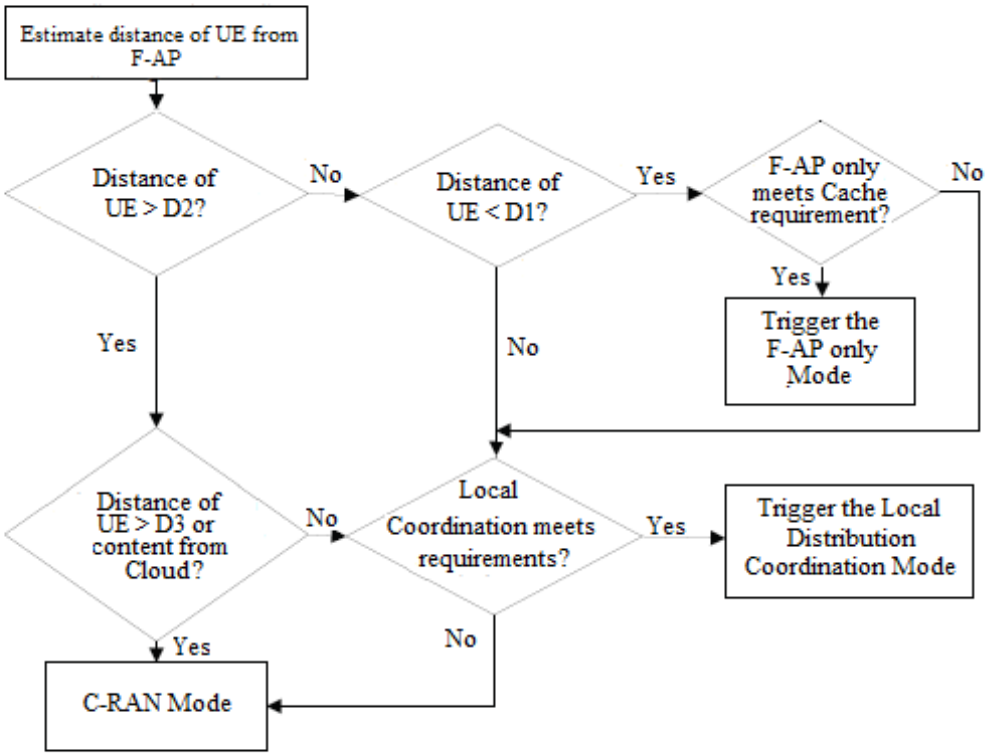

Figure 5.2: Adaptive transmission mode selection in F-RAN

We define MBS or F-AP as the interfering BS for a UE if the average received signal strength from MBS or F-AP i.e $p^m d_{mb}^{-\alpha_m}$ or $p^f d_{fn}^{-\alpha_f}$ satisfies $p_t^{m(f)} d_{mb(fn)}^{-\alpha_{m(f)}} \geq \frac{p_m ax}{\delta}$ where $\delta > 1$ is a constant for deciding interference.

We classify users into three types according to the user access methods. Cell-edge users (CEU) and cell-center users (CCU) served by F-AP and UE served directly by MBS in case cached content is not available at F-APs. For any cell-center UE $\mathcal{M} = \{N+1, N+2, \ldots, N+M\}$, the serving BS is MBS and the received power from MBS and F-AP satisfies the inequality $p^m d_{mu}^{-\alpha_m} > p^f d_{fu}^{-\alpha_r} \delta$ or $d_{fu} > \delta^{\frac{1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}}$ where $\theta = \left(\frac{p^f}{p^m}\right)^{\frac{1}{\alpha_f}}$.

For the UE located in the cell-edge region of MBS $p^f d_{fu}^{-\alpha_f} < p^m d_{mu}^{-\alpha_m} \leq p^f d_{fu}^{-\alpha_f} \delta$ or equivalently $\theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} < d_{fu} \leq \delta^{\frac{1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}}$. Therefore this UE suffers main interference from F-AP.

Similarly the cell-center FUE does not experience interference from MBS and its distance to the F-AP must satisfy $d_{fu} < \delta^{\frac{-1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} (D1)$.

The cell-edge RUE suffers main interference from MBS i.e $p^f d_{ru}^{-\alpha_f} < p^m d_{mu}^{-\alpha_m} \delta$. Therefore its distance from the RRH must satisfy $\delta^{\frac{1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} \leq d_{fu}(D2) < \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} (D3)$.

The user association scheme for different users is summarized as below:

$$U^i = \begin{cases} m_1, \text{if } \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} < d_{fu} \leq \delta^{\frac{1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} \text{ and } d_{mu} > D \\ n_1, \text{if } \delta^{\frac{1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} \leq d_{fu} < \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} \text{ and } d_{fu} > D \\ n_2, \text{if } d_{fu} < \delta^{\frac{-1}{f}} \theta d_{mu}^{\frac{\alpha_m}{\alpha_f}} \text{ and } d_{fu} \leq D \end{cases} \tag{5.48}$$

D is the boundary distance where NOMA outperforms OMA defined in (4.24).

5.5 Solution to Sum-Rate Maximization Problem

The problem in (5.47a) is MINP and its extremely difficult to find the global optimal solution. For high-density F-RANs consisting of even hundreds of F-APs and UEs, low complexity suboptimal solutions are required. We divide the problem by categorising the variables into two groups. The first ones are the discrete variables, including access (mode) section d_f^i, coordinated F-APs and subchannel assignment. The second group consists of continuous transmit power variables p. In the first subproblem, the values d_f^i and b_n^k are optimized given fixed power allocation.

To solve this problem, we propose two suboptimal algorithms. In order to provide the efficient mode selection, we design the utility function as:

$$U = m_{nf}(X_i^k + Y_i^c) \tag{5.49}$$

where Y_i^c represents the required number of computing resources (unit as per CPU instruction) and X_i^k is the required number of radio resources where:

$$\frac{Y_i^c(c_1^u)}{c_1} = \frac{Y_i^c(c_2^u)}{c_2} \quad \forall c_1, c_2 \in \{1, 2, \ldots, c_i^n\} \tag{5.50}$$

$$\frac{X_i^k(k_1^u)}{k_1} > \frac{X_i^k(k_2^u)}{k_2} \quad \forall k_1 \leq k_2 \leq k \tag{5.51}$$

We define the number of computing and communication resources available at each F-AP as c_i^f and k_i^f respectively.

5.5.1 Transmission Mode Selection and Subchannel Assignment Problem

We propose an efficient mode selection algorithm for multiple F-APs and multiple users to deal with the problem in terms of 1) coordinated task computation, 2) mode selection and 3) resource allocation to maximize the sum-rate. It determines which

mode should be selected based on the access priorty based on the distance between UEs and F-AP, number of radio resources, computing resources and processing data. Besides, due to the introduction of caching strategy in the NOMA-enabled F-RAN system, we should add the reward function to the utility to represent the cache revenue earned by the system. The goal is to choose the best mode such that the utility is maximized with sum-rate of users do not exceed the available wireless fronthaul rate. We consider the utility function as an important factor in order to cater the communication and computation resource requirements in NOMA-enabled F-RANs. Moreover, NOMA restricts the number of users served by F-AP.

We define the matching matrix $M(n, f)$ with set of users and F-APs. Firstly, the Algorithm 9 initializes the computation and communication resource requirements of UEs and available resources of F-APs. The F-APs are then sorted by descending order of the utility function defined in (5.49). For F-AP , UEs are sorted in descending order of their distances. Thereafter a 3-dimensional table is created with

Algorithm 9: Proposed Mode Selection Algorithm for NOMA based F-RAN

1. **Input F-APs and UEs**
2. Computation and communication resource requirements of UEs
$U_X^u = \{k_1^u, k_2^u, \ldots, k_n^u\}$ $U_Y^u = \{c_1^u, c_2^u, \ldots, c_n^u\}$
available resources of F-APs
$U_X^f = \{k_1^f, k_2^f, \ldots, k_i^f\}$ $U_Y^f = \{c_1^f, c_2^f, \ldots, c_i^f\}$
3. Utility function of users $X^u + Y^u$
Utility function of F-APs $X^f + Y^f$
4. **for** $f \leftarrow 0$ *to* F **do**
5.**for** $n \leftarrow 0$ *to* N **do**
6. Calculate the distance between UEs and F-APs
7. Sort the UEs by ascening order according to their distance from F-APs
8. $M_{nf} = MCKA\{UEs, U_X^u, U_Y^u, U_X^f, U_Y^f\}$

$U(N \times K \times C)$ and Multiple-Choice Knapsack Algorithm (MCKA) (Algorithm 10) is adopted to resolve the mode selection problem. We consider that each F-AP is a knapsack with rate which needs to be filled with pair of users. Each pair of users is

characterized with a utility value (reward of caching added to the sum-rate utility if the caching is chosen). For each F-AP, the steps are repeated until all the users are served. As a result the output matching matrix M_{nf} and mode selection matrix $\theta = \alpha_i \times d_f^i$ is obtained.

Each user makes a request to the preferred F-AP according to the utility function. If the computation and communication resource requirements are met then the F-AP is not dedicated for coordinated transmission strategy and the user will consume $G(f-1, k_u, c_u)$ resources and $M_{nf} = 1$. For coordinated transmission strategy, the selected F-APs are grouped if the number of F-APs are less than $\bar{f}$. Condition $D_2 \leq d \leq D_3$ corresponds to the local distributed coordination mode. If $d > D_3$, then the CRAN mode is triggered.

The algorithm gets terminated after all the users are served.

Algorithm 10: Multiple-Choice Knapsack Algorithm

1. **Input** the order of UEs, computation and communication resource requirements of
UEs, utility function of UEs derived in (5.49), resource status of F-APs.
2. **for** $i = 1$ **to** $F \times K$, $j = 1$ **to** C^f **do**
3. **If** $K_u \leq K_f$ & $C_n \leq C_f$ & $d \leq D_1$ **then**
$M_{nf} = 1$, $\alpha_c^i = 1$ and $d_f^i = 0$, F-AP only mode and reward of caching added to sum-rate utility function $\phi(P, b, \theta)$ in (5.43).
4.**elseif** $D_2 \leq d \leq D_3$
5. $M_{nf} = 1$, $\alpha_c^i = 1$ and $d_f^i = 1$
6. **else**
7. $M_{nf} = 0$, $\alpha_c^i = 0$ and $d_f^i = 1$
8. $K_i^f \leftarrow K_i^f - 1$
$C_i^f \leftarrow C_i^f - 1$
$F \leftarrow F - 1$
9.**Output** matching matrix M_{nf} and mode selection matrix θ

To perform subchannel assignment efficiently, we transform the problem (5.47a) into linear optimization problem by using auxillary variable z_n^k, where $z_n^k = b_n^k \alpha_i d_f^i$. The auxillary variable depends on subchannel assignment b_n^k, therefore we introduce the

following constraint:

$$b_n^k - z_n^k \geq 0 \tag{5.52}$$

Moreover the following constraint should also be met:

$$d_f^i - z_n^k \geq 0 \tag{5.53}$$

5.5.2 Power Allocation Problem

The problem given in (5.47a) can be reformulated into an equivalent form. Given transmission mode selection and subchannel assignment, we aim to find the power allocation across subchannels. Thus the optimization problem can be represented as:

$$\max_P \sum_{f\in F}\sum_{k\in K} b_n^k \Big(\sum_{n\in N} R_n^k + \sum_{m=N+1}^{N+M} d_f^i R_f\Big) + \lambda\alpha_i \sum_{f\in F}\sum_{k\in K}\sum_{n\in N} G_f r_n^d \tag{5.54}$$

By considering the power allocation dependent constraints, the optimization problem can be represented as:

$$\max_P \sum_{f\in F}\sum_{k\in K}\sum_{n\in N} (b_n^k + \alpha_i\lambda r_n^R) log\left(1 + \frac{|h_{f,k}^n|^2 p^f}{\sum_{l=b+1}^{U} |h_{f,k}^n|^2 p^f + \sum_{m=1}^{M} \beta_{nk} |g_{f,k}^{mn}|^2 p^m + \sigma^2}\right) + \sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M} b_n^k d_f^i log\left(1 + \frac{|f_{m,k}^{b_2}|^2 p^m a_{m,k}^{b_2}}{\sum_{k=1}^{K}\sum_{b=1}^{B} \phi_{b,k} |g_{m,k}^{fb_2}|^2 p^f + \sigma^2}\right) \tag{5.55}$$

s.t

$$b_n^k(\alpha_i p_{nk}^f |h_{fk}^n|^2 - \sum_{j=f+1}^{F} p_{jk}^f |h_{jk}^n|^2) \geq P_t \tag{5.56}$$

$$\sum_{k\in K}\Big(\sum_{n\in N} p_{nk}^f + \sum_{m=N+1}^{M} P_k^m\Big) \leq P_{max} \tag{5.57}$$

$$\sum_{f=1}^{F} x_{fn} p_k^f \leq p_{max} \tag{5.58}$$

$$\sum_{f\in F}\sum_{\gamma B} b_n^k \Big(\sum_{n\in N} R_n^k + \sum_{m=N+1}^{N+M} d_f^i R_f\Big) \geq R_{max} \tag{5.59}$$

After approximation, the negative log terms with affine objective and constraint functions can be represented as:

$$\begin{aligned} U(\mathbf{P}) = \sum_{f\in F}\sum_{n\in N}\sum_{k\in K}(b_n^k + \alpha_i\lambda r_n^R)\Bigg[log(\sum_{l=b+1}^{U}|h_{f,k}^n|^2p^f \\ + \sum_{m=1}^{M}\beta_{nk}|g_{f,k}^{mn}|^2p^m + \sigma^2 + |h_{f,k}^n|^2p^f) \\ -log(\sum_{l=b+1}^{U}|h_{f,k}^n|^2p^f + \sum_{m=1}^{M}\beta_{nk}|g_{f,k}^{mn}|^2p^m + \sigma^2)\Bigg] \\ + \sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M}b_n^k d_f^i\Bigg[log\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f + \sigma^2 + |f_{m,k}^{b_2}|^2p^m a_{m,k}^{b_2} \\ -log(\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f + \sigma^2)\Bigg] \end{aligned} \tag{5.60}$$

Similarly approximation for constraint (5.59) is represented as:

$$\begin{aligned} \sum_{f\in F}\sum_{k\in K}\sum_{n\in N}b_n^k log\Bigg(1 + \frac{|h_{f,k}^n|^2p^f}{\sum_{l=b+1}^{U}|h_{f,k}^n|^2p^f + \sum_{m=1}^{M}\beta_{nk}|g_{f,k}^{mn}|^2p^m + \sigma^2}\Bigg) \\ + \sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M}b_n^k d_f^i log\Bigg(1 + \frac{|f_{m,k}^{b_2}|^2p^m a_{m,k}^{b_2}}{\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f + \sigma^2}\Bigg) \geq R_{max} \end{aligned} \tag{5.61}$$

Equivalently:

$$\begin{aligned} \sum_{f\in F}\sum_{n\in N}\sum_{k\in K}b_n^k\Bigg[log(\sum_{l=b}^{U}|h_{f,k}^n|^2p^f + I_m + \sigma^2) \\ -log(\sum_{l=b+1}^{U}|h_{f,k}^n|^2p_{t-1}^f + I_m + \sigma^2) + \frac{(p^f - p_{t-1}^f)|h_{f,k}^n|^2}{|h_{f,k}^n|^2p_{t-1}^f + I_m + \sigma^2}\Bigg] \\ + \sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M}b_n^k d_f^i\Bigg[log\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f + \sigma^2 + |f_{m,k}^{b_2}|^2p^m a_{m,k}^{b_2} \\ -log(\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}^f + \sigma^2) + \frac{(p^f - p_{t-1}^f)|g_{m,k}^{fb_2}|^2}{\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}f + \sigma^2}\Bigg] \geq R_{max} \end{aligned} \tag{5.62}$$

After transforming the power allocation problem to the convex problem with affine

objective and constraint functions, we get:

$$\begin{aligned} U(\mathbf{P}) = \sum_{f\in F}\sum_{n\in N}\sum_{k\in K}(b_n^k + \alpha_i\lambda r_n^R)\Bigg[log(\sum_{l=b}^{U}|h_{f,k}^n|^2p^f + I_m + \sigma^2) \\ -log(\sum_{l=b+1}^{U}|h_{f,k}^n|^2p_{t-1}^f + I_m + \sigma^2) + \frac{(p^f - p_{t-1}^f)|h_{f,k}^n|^2}{|h_{f,k}^n|^2p_{t-1}^f + I_m + \sigma^2}\Bigg] \\ + \sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M} b_n^k d_f^i\Bigg[log\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f + \sigma^2 + |f_{m,k}^{b_2}|^2p^m a_{m,k}^{b_2} \\ -log(\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}^f + \sigma^2) + \frac{(p^f - p_{t-1}^f)|g_{m,k}^{fb_2}|^2}{\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}f + \sigma^2}\Bigg] \end{aligned} \tag{5.63}$$

We get the following transformed optimization problem:

$$\max_{P} U(\mathbf{p}) \tag{5.64}$$

s.t. (5.62), (5.56), (5.57), (5.58)

The Lagrange function is represented as:

$$
\begin{aligned}
L(p,\lambda,\mu,\gamma,\eta) = \sum_{f\in F}\sum_{n\in N}\sum_{k\in K}(b_n^k+\alpha_i\lambda r_n^R)\Bigg[log(\sum_{l=b}^{U}|h_{f,k}^n|^2p^f+I_m+\sigma^2) \\
-log(\sum_{l=b+1}^{U}|h_{f,k}^n|^2p_{t-1}^f+I_m+\sigma^2)+\frac{(p^f-p_{t-1}^f)|h_{f,k}^n|^2}{|h_{f,k}^n|^2p_{t-1}^f+I_m+\sigma^2}\Bigg] \\
+\sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M}b_n^kd_f^i\Bigg[log\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f+\sigma^2+|f_{m,k}^{b_2}|^2p^ma_{m,k}^{b_2} \\
-log(\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}^f+\sigma^2)+\frac{(p^f-p_{t-1}^f)|g_{m,k}^{fb_2}|^2}{\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}f+\sigma^2}\Bigg] \\
+\lambda\Bigg(\sum_{f\in F}\sum_{n\in N}\sum_{k\in K}b_n^k\Bigg[log(\sum_{l=b}^{U}|h_{f,k}^n|^2p^f+I_m+\sigma^2)-log(\sum_{l=b+1}^{U}|h_{f,k}^n|^2p_{t-1}^f+I_m+\sigma^2) \\
+\frac{(p^f-p_{t-1}^f)|h_{f,k}^n|^2}{|h_{f,k}^n|^2p_{t-1}^f+I_m+\sigma^2}\Bigg] \\
+\sum_{f\in F}\sum_{k\in K}\sum_{m=N+1}^{N+M}b_n^kd_f^i\Bigg[log\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p^f+\sigma^2+|f_{m,k}^{b_2}|^2p^ma_{m,k}^{b_2} \\
-log(\sum_{k=1}^{K}\sum_{b=1}^{B}\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}^f+\sigma^2)+\frac{(p^f-p_{t-1}^f)|g_{m,k}^{fb_2}|^2}{\phi_{b,k}|g_{m,k}^{fb_2}|^2p_{t-1}f+\sigma^2}\Bigg]-R_{max}\Bigg) \\
+\mu\Bigg(b_n^k(\alpha_ip_{nk}^f|h_{fk}^n|^2-\sum_{j=f+1}^{F}p_{jk}^f|h_{jk}^n|^2)-P_t\Bigg)-\gamma\Bigg(\sum_{f=1}^{F}x_{fn}p_k^f-p_{max}\Bigg) \\
-\eta\Bigg(\sum_{k\in K}(\sum_{n\in N}p_{nk}^f+\sum_{m=N+1}^{M}P_k^m)-P_{max}\Bigg)
\end{aligned}
$$

In order to solve the optimization problem, the primal and dual problems are alternately optimized, until differences between variables in each iteration are smaller than the threshold. The optimal power allocation can be derived by taking derivative $dL(p,\lambda,\mu,\gamma,\eta)/dp^f$ for given access selection for subchannel b_n^k as:

If $d_f^i=1$ and $b_n^k=1$

$$
p_{nk}^f=\frac{-(h_{f,k}^n(1+\lambda)+\sigma^2\pm\sqrt{g_{mk}^{fb_2}(1+\lambda+\sigma^2S)^2-4b_n^k\sigma^2Sg_{mk}^{fb_2}}}{2Sg_{mk}^{fb_2}} \tag{5.65}
$$

where

$$S = -(1+\lambda)\frac{-h_{fk}^{n}}{p_{t-1}^{f}h_{fk}^{n}+\sigma^{2}} + \mu\alpha_{i}\frac{h_{fk}^{n}}{\sigma^{2}} - \gamma x_{fn} - \eta \tag{5.66}$$

and

$$p_{nk}^{f} = \left[-\frac{1+\lambda}{T} - \frac{\sigma^{2}}{h_{f,k}^{n}} \right]^{+} \tag{5.67}$$

where

$$T = -d_{f}^{i}\frac{h_{fk}^{n}}{p_{t-1}^{f}h_{f,k}^{n}+\sigma^{2}} + \phi_{bk}\frac{\mu h_{fk}^{n}}{\sigma^{2}} - \gamma - \eta \tag{5.68}$$

If $d_{f}^{i} = 0$ and $b_{n}^{k} = 1$

$$p^{f} = \left[-\frac{1+\lambda}{T} - \frac{\sigma^{2}\phi_{bk}}{g_{nk}^{f_{b2}}} \right]^{2} \tag{5.69}$$

5.6 Simulation Results

In this section we present simulation results to evaluate the performance of the proposed algorithm. We consider several F-APs located in the coverage of one MBS with 1 km diameter. It is assumed that the F-APs are uniformly distributed in a ring area centered around the MBS with the radius of inner ring to be 250 m and the outer radius of the ring is equal to the radius of the cell. The distance of FUEs are measured with respect to their serving MBS or F-APs. Each F-AP has a covering radius of D_R. If the distance between UE and F-AP is within D_R, the UE chooses to access the network via the RRH. We assume that all fronthaul links are identical, therefore $R_f^r = R^r$, where R_f^r is the maximum traffic load that can be carried by the fronthaul link associated with F-AP r. The simulation parameters are listed in Table 5.1.

The maximum number of users that can be multiplexed on the same RB is 2. Moreover we assume that the power sharing coefficients of NOMA for each tier are

same i.e. $a_r^{b1} = a_m^{n1}$ and $a_r^{b2} = a_m^{n2}$.

Fig. 5.3 shows the sum of the utility of UE's considering the heterogeneous resource requirementss $G(f, k_u, c_u)$ by employing different mode selection criterion: one criterion is based on the minimum distance transmission mode selection employed in OMA F-RAN and the other is the proposed adaptive transmission mode selection in NOMA-enabled F-RAN systems. It is observed that the sum of the UEs utility defined in (5.49) increases as the number of UEs increases and the utility in the NOMA-enabled F-RAN is greater than its OMA F-RAN counterpart for different transmission mode selection schemes. It can be seen from the figure that the sum utility is the highest for proposed NOMA F-RAN mode selection scheme followed by minimum distance mode selection scheme in OMA F-RANs.

Fig. 5.4 shows the average sum rate of RUEs with NOMA F-RAN and OMA F-RAN versus δ for different transmit powers. δ is the factor for interference as discussed in section 4. We can observe that the average sum-rate of FUEs decreases with the increase in factor δ. This degradation can be explained as follows: larger value of δ means increase in main interference. Due to increase in main interference more F-APs use CoMP and the proportion of F-APs which use CoMP with large size increases. This is because the F-APs are more likely to cooperate. Moreover

Table 5.1: Simulation parameters

Parameter	Values
Path-loss from F-AP to UE	$140.7 + 36.7 * log_{10} d_b$, d in km
Path-loss from MBS to UE	$128.1 + 37.6 * log_{10} d_n$, d in km
Number of antenna at BS/UE	1
The density of F-APs F λ_f	600, 1000 F-APs $/Km^2$
The density of users N λ_u	400 users$/Km^2$
Maximum Tx power of MBS	43 dBm
Maximum Tx power of RRH	29 dBm
Throughput Calculation	Based in Shannon's Formula
The detection threshold at SIC receiver P_{th}	10dBm

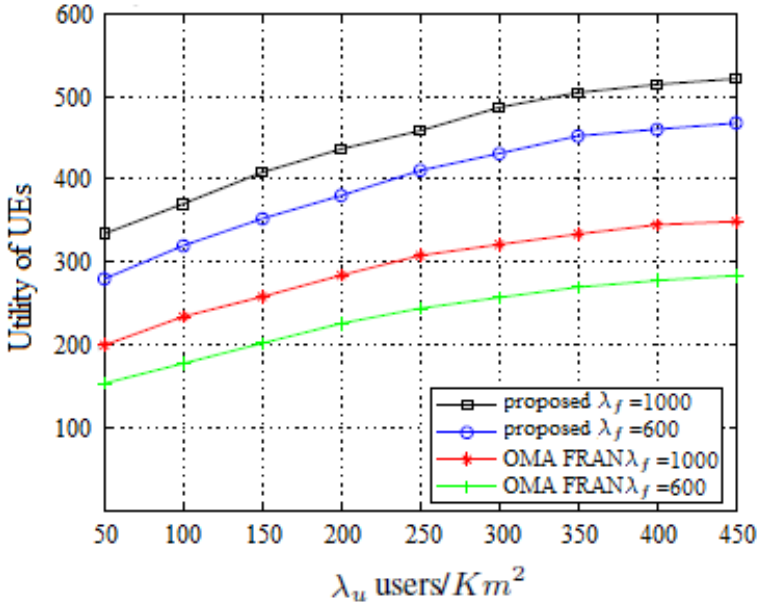

Figure 5.3: The utility of UEs vs density of UEs

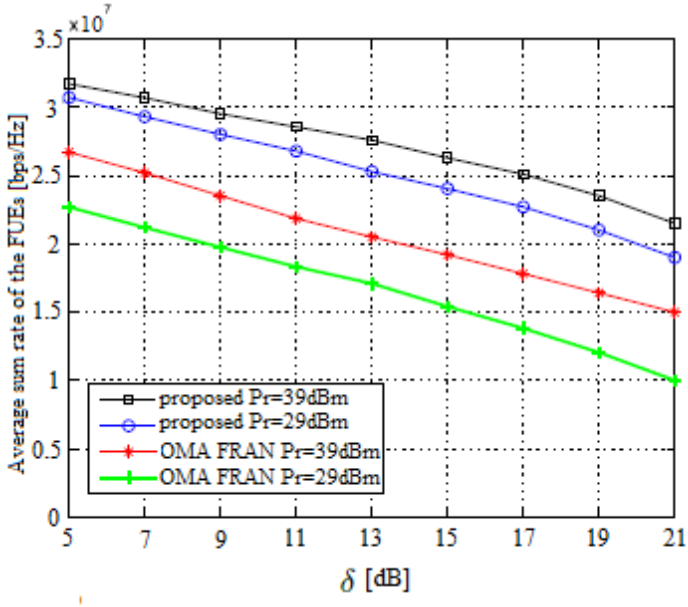

Figure 5.4: Average sum rate of FUEs vs δ $(\{a_r^{b1}, a_r^{b2}\} = \{a_m^{n1}, a_m^{n2}\} = \{0.3, 0.7\})$

more users are associated with F-APs with low SINR which in turn degrades the performance. It is shown that the average sum-rate performance of NOMA enabled F-RAN FUEs outperforms OMA F-RAN.

Fig. 5.5 shows the throughput of the proposed scheme versus the density of F-APs

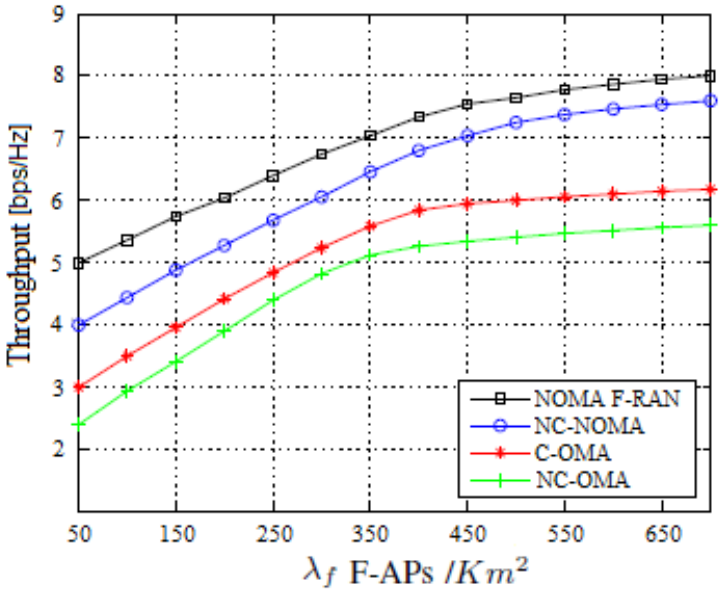

Figure 5.5: The system throughput vs density of the F-APs

for: NOMA-enabled F-RAN and OMA F-RAN with caching and without caching. It is observed that the throughput gradually increases with the density of F-APs. The proposed method has higher throughput than the OMA F-RAN system. The multiple F-APs can simultaneously configure radio resources to a given UE, introducing diversity gain. This diversity gain is achieved due to multiple F-APs coordinating on the same subchannel to serve a UE.

Fig. 5.6 shows the average delay performance of our NOMA-enabled F-RAN scheme compared with the OMA F-RAN scheme versus the computational capacity of the F-APs for both offloading and no offloading tasks. It is observed that with increasing computational capacity of the F-APs, the delay first decreases and then becomes constant. The average delay of the offloading schemes is higher when the computational capability of the F-APs is limited. By utilizing the NOMA principle, multiple F-APs cooperate on the same subchannel. Therefore, it can bring significant capacity gains as compared with OMA F-RAN scheme.

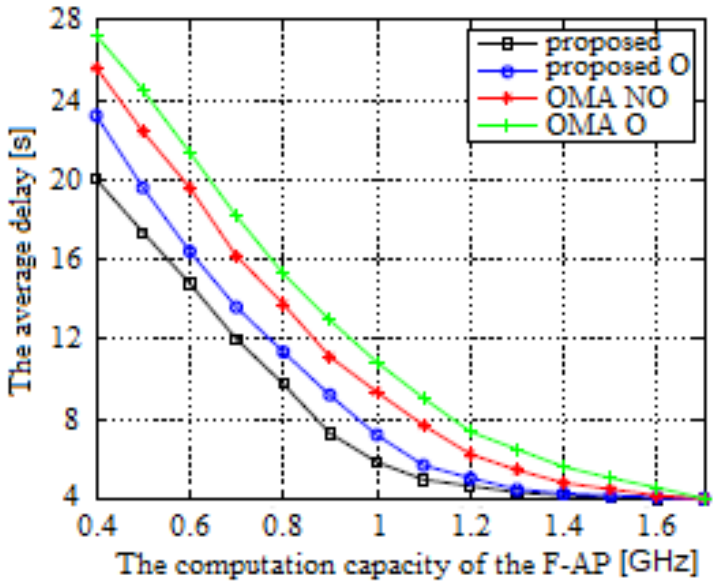

Figure 5.6: Average delay vs the computation capacity of the F-APs

Fig. 5.7 shows the average delay versus the total number of computation tasks G_f, which is the processing data transformed into number of computation tasks. SC is the number of subchannels. It is noted that the average delay for all the computation tasks which are performed locally at the F-APs achieves a lower delay as compared to the mode where the requested tasks are cooperatively processed at the F-APs and cloud. Moreover, Fig. 5.7 shows the abudant subchannels (SC) are beneficial for the average delay performance. With the increase in available subchannels, average delay is decreased. This is because the computation time is reduced greatly when the available spectrum resources are abudant. Moreover the offloading time for tasks is greatly reduced with the increase in the number of resources.

5.7 Conclusions

In this chapter, we have studied the multi-objective resource allocation problem in NOMA-enabled F-RAN system to tackle different aspects such as high throughput

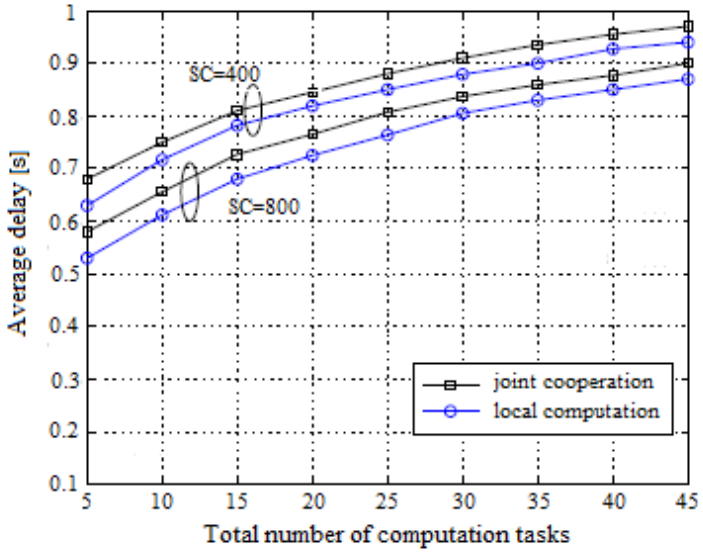

Figure 5.7: Average delay vs the number of computation tasks

and low-latency requirements of high and low user-density regions, in order to meet the heterogeneous requirements of eMBB and URLLC traffic respectively. In low user-density mode we studied the joint caching placement and association strategy aiming at minimizing the average delay. The average delay function has been formulated for both cooperative and non-cooperative transmission modes, cache placement subject to F-AP storage capacity and maximum number of cooperating F-APs constraints. In dense F-RAN mode, sum-rate of the macro-cell based transmission and F-AP based transmission is considered as the objective function. We have proposed sub-optimal algorithms to solve the joint problem of transmission mode selection, subchannel assignment and power allocation to maximize the sum data rate of all F-UEs while satisfying fronthaul rate and F-AP power constraints. Moreover, for a given transmission mode selection and subchannel assignment, the optimal power allocation has been derived in closed-form. Simulation results have shown the effectiveness of the proposed NOMA-enabled F-RAN framework and have

revealed that ultra-low latency and high throughput can be achieved by properly utilizing the available computing and communication resources.

Chapter 6

Conclusions and Future Work

6.1 Overall Conclusion

The main conclusions are drawn as follows:

(a) In the first part of this thesis, we have studied joint user association, muting and power-bandwidth optimization in multi-cell NOMA-enabled C-RAN system. The problem has been formulated as a combinatorial non-convex optimization problem. By formulating joint user association and muting problem, we have proposed a centralized algorithm to provide the optimal solution to the RRH muting problem for fixed bandwidth and transmit power. Besides, a suboptimal algorithm considering ICI has also been proposed to achieve a trade-off between performance and computational complexity. The bandwidth-power allocation problem has been reformulated and an efficient algorithm has been proposed to solve the problem. Moreover the optimal power allocations have been given in closed-form expressions. Specifically, our NOMA-enabled C-RAN framework can find the best RB allocation, number of active RRHs and transmission BPA strategy, while satisfying users' data rate constraints and per-RRH bandwidth and power constraints. Simulation results have revealed that our proposed algorithms can obtain the optimal solution of the joint opti-

misation problem in a significantly reduced computational time and show that NOMA-enabled C-RAN achieves improved network performance in both data rate and network utility with proportional fairness consideration. Moreover, numerical results have showed that our proposed joint channel bandwidth and power allocations for NOMA-enabled C-RAN transmission can significantly minimize the total RRHs transmission power considering the bandwidth constraint in comparison with the conventional OMA-enabled C-RAN transmission scheme as well as the corresponding fixed BPA scheme.

(b) In the second part, we have investigated the performance of NOMA in heterogeneous cloud radio access networks (H-CRAN), where coordination of macro base station (MBS) and RRHs for H-CRAN with NOMA is introduced to improve network performance. The problem of jointly optimizing user association, coordinated scheduling and power allocation for NOMA-enabled H-CRANs has been formulated. To efficiently solve this problem, we have decomposed the joint optimization problem into two subproblems as 1) user association and scheduling 2) power allocation optimization. Firstly the users are divided based on different interference they suffer. This interference-aware NOMA approach account for the inter-tier interference. Proportional fairness (PF) scheduling for NOMA is utilized to schedule users with a two-loop optimization method to enhance throughput and fairness. Based on the user scheduling scheme, optimal power allocation optimization has been performed by the hierarchical decomposition approach. It is then followed by algorithm for joint scheduling and power allocation. Simulation results have showed that the proposed NOMA-enabled H-CRAN outperforms OMA-based H-CRANs in terms of total

achievable rate and can achieve significant fairness improvement.

(c) In the final part of this thesis, we have studied the multi-objective resource allocation problem in NOMA-enabled F-RAN system to tackle different aspects such as high throughput and low-latency requirements of high and low user-density regions, in order to meet the heterogeneous requirements of eMBB and URLLC traffic respectively. In low user-density mode we studied the joint caching placement and association strategy aiming at minimizing the average delay. The average delay function has been formulated for both cooperative and non-cooperative transmission modes, cache placement subject to F-AP storage capacity and maximum number of cooperating F-APs constraints. In dense F-RAN mode, sum-rate of the macro-cell based transmission and F-AP based transmission is considered as the objective function. We have proposed sub-optimal algorithms to solved the joint problem of transmission mode selection, subchannel assignment and power allocation to maximize the sum data rate of all F-UEs while satisfying fronthaul rate and F-AP power constraints. Moreover, for given transmission mode selection and subchannel assignment, the optimal power allocation has been derived in a closed-form. Simulation results have showed the effectiveness of the proposed NOMA-enabled F-RAN framework and have revealed that the ultra-low latency and high throughput can be achieved by properly utilizing the available resources.

6.2 Areas of Future Research

The explosive growth of traffic demand keeps imposing unprecedented challenges for the development in future wireless communication systems, supporting massive

connectivity as well as spectral efficiency improvement, ultra-reliable low-latency communications (URLLC), enhanced mobile broad-band (eMBB) and so on. This thesis has addressed some of these challenges by investigating the concept of NOMA in C-RAN, H-CRAN and F-RAN and improving the system performance. However, there are still many research issues remaining to be addressed to make it feasible to apply these concepts in practice, as listed below:

6.2.1 Beamforming Design for NOMA enabled C-RAN and H-CRAN Systems

The existing research contributions on NOMA enabled C-RAN are still in their infancy. The C-RAN networks are capable of efficient interference management and large-scale data control. This is particularly important for large-scale NOMA enabled C-RAN networks, where sophisticated interference management, e.g., distributed beamforming design, dynamic user association, and efficient power allocation can be jointly considered.

6.2.2 Joint eMBB and URLLC Design for NOMA enabled F-RAN Systems

The future wireless networks are expected to provide services for latency sensitive devices for applications in factory automation, autonomous driving, and remote surgery. Therefore, cross-layer designs and network architecture designs should be taken into account for NOMA systems to support eMBB and URLLC simultaneously in future works. Moreover, optimizing a more generalized NOMA enabled F-RAN architecture empowered by D2D communication, relaying, and caching at the network edge would be the reaearch direction so as to meet the ambitious requirements of the anticipated futuristic wireless communication systems.

Appendix A

The Lagrangian to the problem (3.56) is :

$$\begin{aligned}\mathcal{L}(\mu_1,\mu_2,\mu_3,\mu_4,\lambda_1,\lambda_2,\lambda_3,\lambda_4) = P_k^{ri*} + P_k^{rj*} + \mu_1(P_k^{ri} + P_k^{rj} - P_k^{r}) \\ +\mu_2\left(\gamma_{min}^* - \frac{b_{ru}\alpha_{kn}\lambda_c P_c^T}{b_{ru}\alpha_{kn}\lambda_c P_{c'}^T + B_{max}(I_{nk}^{r'} + N_0)}\right) \\ +\mu_3\left(\gamma_{min}^* - \frac{b_{ru}\alpha_{kn}g_{nk}^{ri}P_k^{ri}}{b_{rn}\alpha_{kn}g_{nk}^{ri*}P_k^{ri*} + B_{max}(I_{nk}^{r} + N_0)}\right) \\ +\mu_4\left(\gamma_{min}^* - \frac{b_{ru}\alpha_{kn}g_{nk}^{rj}P_k^{rj}}{b_{rn}\alpha_{kn}g_{nk}^{rj*}P_k^{rj*} + B_{max}(I_{nk}^{r} + N_0)}\right) \\ -\lambda_1 P_k^{ri} - \lambda_2 P_k^{rj} - \lambda_3 P_k^{ri*} - \lambda_4 P_k^{rj*}\end{aligned} \tag{A.1}$$

where μ and λ are the Lagrange multipliers and γ_{min}^* is the minimum SINR which needs to be maximized with successfull SIC. Applying KKT conditions

$$\mu_1(P_k^{ri} + P_k^{rj} - P_k^{r}) = 0 \tag{A.2}$$

$$\mu_2\left(\gamma_{min}^* - \frac{b_{ru}\alpha_{kn}\lambda_c P_c^T}{b_{ru}\alpha_{kn}\lambda_c P_{c'}^T + B_{max}(I_{nk}^{r'} + N_0)}\right) = 0 \tag{A.3}$$

$$\mu_3\left(\gamma_{min}^* \frac{b_{ru}\alpha_{kn}g_{nk}^{ri}P_k^{ri}}{b_{rn}\alpha_{kn}g_{nk}^{ri*}P_k^{ri*} + B_{max}(I_{nk}^{r} + N_0)}\right) = 0 \tag{A.4}$$

$$\mu_4\left(\gamma_{min}^* - \frac{b_{ru}\alpha_{kn}g_{nk}^{rj}P_k^{rj}}{b_{rn}\alpha_{kn}g_{nk}^{rj*}P_k^{rj*} + B_{max}(I_{nk}^{r} + N_0)}\right) = 0 \tag{A.5}$$

$$\lambda_1 P_k^{ri} = \lambda_2 P_k^{rj} = \lambda_3 P_k^{ri*} = \lambda_4 P_k^{rj*} = 0 \tag{A.6}$$

$$(P_k^{ri} + P_k^{rj} - P_k^{r}) \leq 0 \tag{A.7}$$

$$\left(\gamma_{min}^* - \frac{b_{ru}\alpha_{kn}\lambda_c P_c^T}{b_{ru}\alpha_{kn}\lambda_c P_{c'}^T + B_{max}(I_{nk}^{r'} + N_0)}\right) \leq 0 \tag{A.8}$$

$$\left(\gamma^*_{min} - \frac{b_{ru}\alpha_{kn}g^{ri}_{nk}P^{ri}_k}{b_{rn}\alpha_{kn}g^{ri*}_{nk}P^{ri*}_k + B_{max}(I^r_{nk}+N_0)}\right) \leq 0 \tag{A.9}$$

$$\left(\gamma^*_{min} - \frac{b_{ru}\alpha_{kn}g^{rj}_{nk}P^{rj}_k}{b_{rn}\alpha_{kn}g^{rj*}_{nk}P^{rj*}_k + B_{max}(I^r_{nk}+N_0)}\right) \leq 0 \tag{A.10}$$

$$\mu_1, \mu_2, \mu_3, \mu_4 \geq 0, \lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0 \tag{A.11}$$

$$\frac{\partial \mathcal{L}}{\partial P^{ri}_k} = \mu_1 - \mu_3 \frac{b_{ru}\alpha_{kn}g^{ri}_{nk}}{b_{rn}\alpha_{kn}g^{ri*}_{nk}P^{ri*}_k + B_{max}(I^r_{nk}+N_0)} - \lambda_1 = 0 \tag{A.12}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial P^{ri*}_k} = \mu_1 - \mu_2 \frac{b_{ru}\alpha_{kn}g^{rj}_{nk}}{b_{ru}\alpha_{kn}\lambda_c P^T_{c'} + B_{max}(I^{r'}_{nk}+N_0)} - \mu_4 \frac{b_{ru}\alpha_{kn}g^{rj}_{nk}}{b_{rn}\alpha_{kn}g^{rj*}_{nk}P^{rj*}_k} \\ + B_{max}(I^r_{nk}+N_0) - \lambda_2 = 0\end{aligned} \tag{A.13}$$

$$\frac{\partial \mathcal{L}}{\partial P^{rj}_k} = 1 - \mu_3 \frac{{b_{ru}}^2{\alpha_{kn}}^2(g^{ri}_{nk})^2P^{ri}_k}{(b_{rn}\alpha_{kn}g^{ri*}_{nk}P^{ri*}_k + B_{max}(I^r_{nk}+N_0))^2} - \lambda_3 = 0 \tag{A.14}$$

$$\frac{\partial \mathcal{L}}{\partial P^{rj*}_k} = 1 - \mu_2 \frac{b_{ru}\alpha_{kn}g^{rj*}_{nk}}{b_{ru}\alpha_{kn}\lambda_c P^T_{c'} + B_{max}(I^{r'}_{nk}+N_0)} - \lambda_4 = 0 \tag{A.15}$$

The complementary-slackness conditions in (A.2)-(A.5) are used to obtain optimal equations. From (A.14), (A.15) it can be observed that $\mu_3 > 0$ and $\mu_2 > 0$. From (A.13) we have μ_1 strictly positive.

$$P^{ri}_k = \frac{(\gamma^*_{min})^2P^r_k(g^{rj}_{nk} - \gamma^*_{min}\lambda_c) + B_{max}(I^r_{nk}+N_0)g^{ri*}_{nk}\gamma^*_{min}}{g^{ri}_{nk}g^{ri*}_{nk}} \tag{A.16}$$

$$P^{rj}_k = \frac{\gamma^*_{min}[B_{max}(I^{r'}_{nk}+N_0) + b_{rn}\alpha_{kn}\lambda_c P^r_k}{b_{rn}\alpha_{kn}g^{rj}_{nk}} \tag{A.17}$$

$$P^{ri*}_k = \frac{\gamma^*_{min}(P^r_k g^{rj}_{nk} - \gamma^*_{min}\lambda_c P^r_k)}{g^{ri*}_{nk}} \tag{A.18}$$

$$P^{rj*}_k = \frac{\gamma^*_{min}\lambda_c P^r_k}{1 + \gamma^*_{min}g^{rj*}_{nk}} \tag{A.19}$$

(A.16)-(A.19) are the optimal solutions of problem (3.56) with given γ^*_{min}. In order to find the minimum user SINR which has to be maximized, we need to find the optimal γ^*_{min}. To obtain optimal γ^*_{min}, it needs to be maximized to guarantee the feasibility of problem (3.56). As it can be observed from (A.16) and (A.18) P^{ri}_k

and P_k^{ri*} can be negative values. Following are the constraints to make the problem feasible.

$$P_k^{ri} = \frac{(\gamma_{min}^*)^2 P_k^r (g_{nk}^{rj} - \gamma_{min}^* \lambda_c) + B_{max}(I_{nk}^r + N_0) g_{nk}^{ri*} \gamma_{min}^*}{g_{nk}^{ri} g_{nk}^{ri*}} \geq 0 \tag{A.20}$$

$$P_k^{ri*} = \frac{\gamma_{min}^* (P_k^r g_{nk}^{rj} - \gamma_{min}^* \lambda_c P_k^r)}{g_{nk}^{ri*}} \geq 0 \tag{A.21}$$

$$P_k^{ri*} + P_k^{rj*} = \frac{\gamma_{min}^* (P_k^r g_{nk}^{rj} - \gamma_{min}^* \lambda_c P_k^r)}{g_{nk}^{ri*}} + \frac{\gamma_{min}^* \lambda_c P_k^r}{1 + \gamma_{min}^* g_{nk}^{rj*}} \leq P_k^r \tag{A.22}$$

Appendix B

Initially, the cell-edge users are identified. We denote the users whose channel gain is above the threshold as L_1 as UE_s or CCUs and users with channel gain below threhold L_2 as UE_w or CEUs where $L_2 \leq L_1$. To find the boundary distance in a cell at which NOMA outperforms OMA, the following condition must be met:

$$log_2\left(1+\frac{d_1^{-\alpha}P_1^k}{I_i+d_1^{-\alpha}P_2^k+\sigma^2}\right) \geq \frac{1}{2}log_2\left(1+\frac{d_1^{-\alpha}P^k}{I_i+d_1^{-\alpha}P^k+\sigma^2}\right) \tag{B.1}$$

where $d_1^{-\alpha}$ and $d_2^{-\alpha}$ are the average channel gains and

$$R^{oma} = \frac{1}{2}log_2\left(1+\frac{d_1^{-\alpha}P^k}{I_i+d_1^{-\alpha}P^k+\sigma^2}\right) \tag{B.2}$$

is the OMA rate with equal power allocation to two users. (B.1) is equivaelnt to:

$$P_1^k \geq \frac{\sqrt{1+\frac{d_1^{-\alpha}P^k}{I_i+\sigma^2}}-1}{\frac{d_1^{-\alpha}}{I_i+\sigma^2}} \tag{B.3}$$

If $\frac{d_2^{-\alpha}}{I_i+\sigma^2} > L_1$ above equation holds when $P_1^k > \frac{\sqrt{1+P_kL_1}-1}{L_1}$ Similarly for cell-edge user:

$$log_2\left(1+\frac{d_2^{-\alpha}P_2^k}{I_i+d_2^{-\alpha}P_1^k+\sigma^2}\right) \geq \frac{1}{2}log_2\left(1+\frac{d_2^{-\alpha}P^k}{I_i+d_2^{-\alpha}P^k+\sigma^2}\right) \tag{B.4}$$

which is equivalent to:

$$P_1^k \leq \frac{\sqrt{1+\frac{d_2^{-\alpha}P^k}{I_i+\sigma^2}}-1}{\frac{d_2^{-\alpha}}{I_i+\sigma^2}} \tag{B.5}$$

Since $\frac{d_2^{-\alpha}}{I_i+\sigma^2} < L_2$ above equation always hold when $P_2^k < \frac{\sqrt{1+P_k L_2}-1}{L_2}$. (B.3) and (B.5) are true simultaneously when following inequality is satisfied:

$$\frac{\sqrt{1+P_k L_1}-1}{L_1} < P_1^k < \frac{\sqrt{1+P_k L_2}-1}{L_2} \tag{B.6}$$

i.e.

$$\frac{\sqrt{1+P_k L_1}-1}{L_1} < \epsilon < \frac{\sqrt{1+P_k L_2}-1}{L_2} \tag{B.7}$$

Considering $d_1 \leq D$ and $d_2 \geq D$ and rearranging equations, we derive the distance D approximately as:

$$D = \left(\frac{1-2\epsilon}{P_b \epsilon^2}\right) \tag{B.8}$$

whereas D is the boundary distance in a cell at which the users are classified as CEUs and CCUs. The users are scheduled based on the NOMA weighted sum-rate as:

$$R^k(\epsilon) = w_1 R_1^k(\epsilon) + w_2 R_2^k(\epsilon) \tag{B.9}$$

where ϵ is the optimal power allocation variable. In order to ensure fairness among users the weights for each UE are calculated based on most recent average rates at each scheduling interval as:

$$w_u(t) = \frac{1}{R_u(t-1)}, \ \forall u \in U \tag{B.10}$$

Appendix C

Derivation of optimal MUEs, RUEs and CS-UEs index:

The Karush-Kuhn-Tucker (KKT) conditions [95], necessary for optimality can be obtained by taking the first-order derivation of problem OP_1 with respect to $\beta_{ij}^{m}(t)$, $\beta_{ij}^{r}(t)$ and $\beta_{i}^{cs}(t)$ $j = 1, 2$ and set the resulting equation to zero:

$$\frac{\partial \mathcal{L}}{\partial \beta_{ij}^{m}(t)} = 0, \frac{\partial \mathcal{L}}{\partial \beta_{ij}^{r}(t)} = 0 and \frac{\partial \mathcal{L}}{\partial \beta_{i}^{cs}(t)} = 0 \tag{C.1}$$

are the first-order necessary conditions for optimality.

The complementary slackness conditions are given as follows:

$$\rho_i^r \left(\sum_{b=1}^{B} C_i^r \beta_{ij}^r(t) + \sum_{b=1}^{B} \sum_{r=1}^{R} \beta_{ij}^r(t) C_i^{cs} \beta_i^{cs} - 1 \right) = 0 \tag{C.2}$$

$$\rho_i^m \left(\sum_{n=1}^{N} C_i^m \beta_{ij}^m(t) + \sum_{n=1}^{N} \beta_{ij}^m(t) C_i^{cs} \beta_i^{cs} - 1 \right) = 0 \tag{C.3}$$

$$\xi_r \beta_{ij}^r(t) = 0 \tag{C.4}$$

$$\xi_m \beta_{ij}^m(t) = 0 \tag{C.5}$$

$$\xi_c \beta_i^{cs}(t) = 0 \tag{C.6}$$

From eqs.(4.40),(4.42) we get:

$$-\frac{R_{bj}(t)}{R^k(t-1)} C_i^r + \rho_i^r(t) C_i^r - \xi_r = 0 \tag{C.7}$$

$$-\frac{R_{nj}(t)}{R^k(t-1)} C_i^m + \rho_i^m(t) C_i^m - \xi_m = 0 \tag{C.8}$$

$$-\frac{R_{cs}(t)}{R^k(t-1)}C_i^{cs}+\rho_i^r(t)C_i^{cs}C_i^r+\rho_i^m(t)C_i^{cs}C_i^m-\xi_c=0 \tag{C.9}$$

In order to obtain optimal MUE, RUE and CS-UE indexes we need to obtain optimal values of $\beta_{ij}^r(t)$, $\beta_{ij}^m(t)$ and $\beta_i^{cs}(t)$. We derive the Lagrangian parameters as:

$$\hat{\rho_i^r}=\max_{x_{ij}^r\in B}\frac{R_{bj}(t)}{R(t-1)} \tag{C.10}$$

$$\hat{\rho_i^m}=max(\rho_{iN}^m,\rho_{iF}^{r^*}) \tag{C.11}$$

ρ_{iN}^m and ρ_{iF}^r denotes the gain in proportional fairness function at the MBS where:

$$\rho_{iN}^m=\max_{x_{ij}^r\in B}\frac{R_{bj}(t)}{R(t-1)} \tag{C.12}$$

and

$$\rho_{iF}^{r^*}=\max_{r\in R}\left(\max_{x_{ij}^{cs}\in\phi_{cs}}\frac{R_{cs}(t)}{R(t-1)}-\max_{x_{ij}^r\in B}\frac{R_{bj}(t)}{R(t-1)}\right) \tag{C.13}$$

where ρ_{iN}^m and $\rho_{iF}^{r^*}$ solve for the best users at the MBS and CS-UE in the region of RRH r^* respectively.

$$\hat{\xi_r}=\rho_i^r(t)C_i^r-\frac{R_{bj}(t)}{R^k(t-1)}C_i^r \tag{C.14}$$

$$\hat{\xi_m}=\rho_i^m(t)C_i^m-\frac{R_{nj}(t)}{R^k(t-1)}C_i^m \tag{C.15}$$

$$\hat{\xi_c}=\hat{\rho_i^m}-\left(\frac{R_{cs}(t)}{R(t-1)}-\max_{x_{ij}^r\in B}\frac{R_{bj}(t)}{R(t-1)}\right) \tag{C.16}$$

To find optimal MUE, RUE and CS-UE index for each RB k, we consider the following cases:

Case-1: For $\rho_{iN}^m>\rho_{iF}^r$ This case corresponds to the scenario where the proportional fairness gain achieved by forming pair with user associated in CS-NOMA range is less than the PF gain without coordination. In this scenario we have:

$$\rho_i^r=\max_{x_{ij}^r\in B}\frac{R_{bj}(t)}{R(t-1)} \tag{C.17}$$

Substituting the above equation in equations (C.14) and (C.15) , we get optimal RUEs as follows:

$$x_{ij}^{r} = arg \max_{x_{ij}^{r} \in B} \frac{R_{bj}(t)}{R(t-1)} \quad j = 1, 2 \tag{C.18}$$

and optimal MUEs as:

$$x_{ij}^{m} = arg \max_{x_{ij}^{m} \in N} \frac{R_{nj}(t)}{R(t-1)} \tag{C.19}$$

Case-2: For $\rho_{iN}^{m} < \rho_{iF}^{r}$ In this case the gain achieved by forming CS-NOMA pair is greater than without coordination and the optimal CS-UE is derived as follows:

$$x_{ij}^{cs} = arg \max_{x_{ij}^{cs} \in \phi_{cs}} \frac{R_{cs}(t)}{R(t-1)} \tag{C.20}$$

Appendix D

Derivation of optimal power allocation:

RUEs are sorted according to eq. 4.4 $\frac{d\widetilde{H}_1}{dp^b_{r,k}1} = 0$

$$\frac{B_k(1+\hat{\gamma}^{b1}_{rk})}{ln\ 2P^{b1}_{rk}} - \lambda_r = 0 \tag{D.1}$$

We calculate the following to obtain P^{b1}_{rk}

$$\frac{d\widetilde{H}_1}{dp^{b2}_{r,k}} = \frac{B_k(1+\hat{\gamma}^{b1}_{rk})}{ln\ 2}\left(-\frac{\hat{\gamma}^{b1}_{rk}}{P^{b1}_{rk}}\right) + \frac{B_k(1+\hat{\gamma}^{b2}_{rk})}{P^{b2}_{rk} ln\ 2} - \lambda_r \tag{D.2}$$

Similarly we can obtain power allocated to MUEs.

References

[1] "Ericsson mobility report June 2019," Ericsson, 2019

[2] Jeffrey G. Andrews, Stefano Buzzi, Wan Choi, Stephen V. Hanly, Angel Lozano, Anthony C. K. Soong, Anthony C. K. Soong, Jianzhong Charlie Zhang, " What Will 5G Be?," in *IEEE Journal on selected Areas in Communications*, June 2014.

[3] M. Shafi et al., "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," in *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201-1221, June 2017.

[4] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," in *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[5] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiplelarge-scale underlay access for 5G: solutions, challenges, opportunities, and future research trends," in *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[6] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," in *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[7] Z. Wei, Y. Jinhong, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," in *ZTE Commun.*, vol. 14, no. 4, pp. 17–25, Oct. 2016.

[8] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" in *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[9] M. Vaezi, Y. Zhang, " Cloud Mobile Networks: From RAN to EPC", in *Springer*, 2017.

[10] M. Peng, Y. Li, J. Jiang, J. Li and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," in *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126-135, December 2014.

[11] M. Peng, S. Yan, K. Zhang and C. Wang, "Fog-computing-based radio access networks: issues and challenges," in *IEEE Network*, vol. 30, no. 4, pp. 46-53, July-August 2016.

[12] D.W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," in *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.

[13] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems - Part I: chunk allocation," in *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2734-2744, Sept. 2009.

[14] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems - Part II: joint chunk, power and bit allocation," in *IEEE Transactions on Communications*, vol. 60, no. 2, pp. 499-509, Feb. 2012.

[15] H. Zhu and J. Wang, "Performance analysis of chunk-based resource allocation in multi-cell OFDMA systems," in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 367-375, February 2014.

[16] D. W. K. Ng, E. S. Lo and R. Schober, "Wireless information and power transfer: Energy efficiency optimization in OFDMA systems," in *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, pp. 6352–6370, Dec. 2013.

[17] L. Dai, B. Wang, Y. Yuan, S. Han, C. I and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," in *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74-81, September 2015.

[18] K. Liang, L. Zhao, X. Zhao, Y. Wang and S. Ou, "Joint resource allocation and coordinated computation offloading for fog radio access networks," in *China Communications*, vol. 13, no. 2, pp. 131-139, N/A 2016.

[19] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Personal, Indoor and Mobile Radio Commun. Sympos.*, Sep. 2013, pp. 611–615.

[20] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Techn. Conf.*, Jun. 2013, pp. 1–5.

[21] T. Cover, "Broadcast channels," in *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2-14, January 1972,

[22] R. Zhang and L. Hanzo, "A Unified Treatment of Superposition Coding Aided Communications: Theory and Practice," in *IEEE Communications Surveys and Tutorials*, vol. 13, pp. 503–520, Third 2011.

[23] S. Vanka, S. Srinivasa, Z. Gong, P. Vizi, K. Stamatiou and M. Haenggi, "Superposition Coding Strategies: Design and Experimental Evaluation," in *IEEE Transactions on Wireless Communications*, vol. 11, no. 7, pp. 2628-2639, July 2012.

[24] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[25] Z. Ding, Z. Yang, P. Fan, and H. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," in *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[26] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," in *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294-2323, thirdquarter 2018.

[27] Higuchi, Kenichi Benjebbour, Anass. (2015). Non-orthogonal Multiple Access (NOMA) with Successive Interference Cancellation for Future Radio Access, in *IEICE Transactions on Communications*. E98.B. 403-414. 10.1587/transcom.E98.B.403.

[28] M. Peng, C. Wang, V. Lau and H. V. Poor,"Fronthaul-constrained cloud radio access networks: insights and challenges," in *IEEE Wireless Communications*, vol. 22, no. 2, pp. 152-160, April 2015.

[29] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the art and research challenges," in *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 416-464, 1st Quart., 2018.

[30] M. Mukherjee, L. Shu, and D. Wang, "Survey of fog computing: Fundamental, network applications, and research challenges," in *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1826-1857, 3rd Quart., 2018.

[31] J. Wu, Z. Zhang, Y. Hong and Y. Wen, "Cloud radio access network (C-RAN): a primer," in *IEEE Network*, vol. 29, no. 1, pp. 35-41, Jan.-Feb. 2015.

[32] A. Checko et al., "Cloud RAN for Mobile Networks-A Technology Overview, " in *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405-426, Firstquarter 2015.

[33] M. Peng, Y. Sun, X. Li, Z. Mao and C. Wang, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," in *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2282-2308, third quarter 2016.

[34] Ding, Z., Poor, H.: "The use of spatially random base stations in cloud radio access networks", in *IEEE Signal Processing Letters*, 2013, 20, (11), pp. 1138–1141.

[35] T. X. Tran and D. Pompili, "Dynamic radio cooperation for user-centric cloud-RAN with computing resource sharing," in *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2379–2393, Apr. 2017.

[36] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "Optimal joint remote radio head selection and beamforming design for limited fronthaul C-RAN," in *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5605–5620, Nov. 2017.

[37] M. Y. Lyazidi, N. Aitsaadi and R. Langar, "Dynamic resource allocation for Cloud-RAN in LTE with real-time BBU/RRH assignment," in *2016 IEEE In-*

ternational Conference on Communications (ICC), Kuala Lumpur, 2016, pp. 1-6.

[38] X. Huang, G. Xue, R. Yu, and S. Leng, "Joint scheduling and beamforming coordination in cloud radio access networks with QoS guarantees," in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5449–5460, Jul. 2016.

[39] M. Awais, et al., "Efficient joint user association and resource allocation for cloud radio access networks," in *IEEE Access*, vol. 5, pp. 1439–1448, 2017.

[40] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (noma)," in *Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013.

[41] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink noma for future lte enhancements," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2013.

[42] F. Liu, P. Mähönen and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Hong Kong, 2015, pp. 1127-1131.

[43] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," in *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849–1858, Apr. 2017.

[44] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: tractability and computation," in *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016

[45] Z. Wei, D. W. K. Ng, and J. Yuan, "Power-efficient resource allocation for MC-NOMA with statistical channel state information," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–7.

[46] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016

[47] A. Douik, H. Dahrouj, T. Y. Al-Naffouri and M. S. Alouini, "Coordinated Scheduling and Power Control in Cloud-Radio Access Networks," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2523-2536, April 2016.

[48] W. Xia, J. Zhang, T. Q. S. Quek, S. Jin and H. Zhu, "Power Minimization-Based Joint Task Scheduling and Resource Allocation in Downlink C-RAN," in *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7268-7280, Nov. 2018.

[49] A. Abdelnasser and E. Hossain, "On Resource Allocation for Downlink Power Minimization in OFDMA Small Cells in a Cloud-RAN," in *IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015, pp. 1-6.

[50] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee and H. V. Poor, "Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges," in IEEE Communications Magazine, vol. 55, no. 10, pp. 176-183, October 2017.

[51] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee and H. V. Poor, "Coordinated Beamforming for Multi-Cell MIMO-NOMA," in *IEEE Communications Letters*, vol. 21, no. 1, pp. 84-87, Jan. 2017.

[52] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," in *IEEE Communications Letters*, 2014.

[53] R. Singh, "Sub-channel assignment and resource scheduling for non-orthogonal multiple access (NOMA) in downlink coordinated multi-point systems," in *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, Paris, 2017, pp. 17-22.

[54] Z. Ding, M. Peng, and V. Poor, "Cooperative non-orthogonal multiple access in 5g systems," in *IEEE Communications Letters*, 2015.

[55] Vien, Q.-T., Ogbonna, N., Nguyen, H.X., et al., "Non-orthogonal multiple access for wireless downlink in cloud radio access networks", in *Proc. IEEE EW 2015*, Budapest, Hungary, May 2015, pp. 434–439.

[56] X. Gu, X. Ji, Z. Ding, W. Wu and M. Peng, "Outage Probability Analysis of Non-Orthogonal Multiple Access in Cloud Radio Access Networks," in *IEEE Communications Letters*, vol. PP, no. 99, pp. 1-1.

[57] F. J. Martin-Vega, Y. Liu, G. Gomez, M. C. Aguayo-Torres and M. Elkashlan, "Modeling and analysis of NOMA enabled CRAN with cluster point process," in *IEEE Global Communications Conference*, Singapore, 2017, pp. 1-6

[58] R. S. Rai, "Coordinated Scheduling for Non-Orthogonal Multiple Access (NOMA) in a Cloud-RAN System," in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018, pp. 1-6.

[59] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," in *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, Dec. 2016.

[60] Y. Liu, X. Li, F. R. Yu, H. Ji, H. Zhang, and V. C. M. Leung, "Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access," in *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2295–2311, Oct. 2017

[61] Y. Xu, H. Sun, R. Q. Hu, and Y. Qian, non-orthogonal multiple access in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1-6.

[62] Vien, Quoc-Tuan; Le, Tuan Anh; Barn, Balbir; Phan, Ca V.: 'Optimising energy efficiency of non-orthogonal multiple access for wireless backhaul in heterogeneous cloud radio access network', in *IET Communications*, 2016, 10, (18), p. 2516-2524.

[63] S. Park, O. Simeone and S. Shamai Shitz, "Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7621-7632, Nov. 2016.

[64] J. Liu, B. Bai, J. Zhang and K. B. Letaief, "Cache Placement in Fog-RANs: From Centralized to Distributed Algorithms," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7039-7051, Nov. 2017.

[65] T. Chiu, W. Chung, A. Pang, Y. Yu and P. Yen, "Ultra-low latency service provision in 5G Fog-Radio Access Networks," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, 2016, pp. 1-6.

[66] A. Pang, W. Chung, T. Chiu and J. Zhang, "Latency-Driven Cooperative Task Computing in Multi-user Fog-Radio Access Networks," in *2017 IEEE 37th In-*

ternational Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, 2017, pp. 615-624.

[67] Y. Shih, W. Chung, A. Pang, T. Chiu and H. Wei, "Enabling Low-Latency Applications in Fog-Radio Access Networks," in *IEEE Network*, vol. 31, no. 1, pp. 52-58, January/February 2017.

[68] X. Peng, J. Shen, J. Zhang and K. B. Letaief, "Backhaul-Aware Caching Placement for Wireless Networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015, pp. 1-6.

[69] D. Vu, N. Dao and S. Cho, "Downlink sum-rate optimization leveraging hungarian method in fog radio access networks," in *2018 International Conference on Information Networking (ICOIN)*, Chiang Mai, 2018, pp. 56-60.

[70] S. He, C. Qi, Y. Huang, Q. Hou and A. Nallanathan, "Two-Level Transmission Scheme for Cache-Enabled Fog Radio Access Networks," in *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 445-456, Jan. 2019.

[71] X. Wen, H. Zhang, H. Zhang and F. Fang, "Interference Pricing Resource Allocation and User-Subchannel Matching for NOMA Hierarchy Fog Networks," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 467-479, June 2019.

[72] Y. Sun, M. Peng, S. Mao and S. Yan, "Hierarchical Radio Resource Allocation for Network Slicing in Fog Radio Access Networks," in IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3866-3881, April 2019.

[73] A. Anand, G. De Veciana and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, 2018, pp. 1970-1978.

[74] Randrianantenaina, I., Kaneko, M., Dahrouj, H., Elsawy, H., Alouini, M. (2019), "Interference Management in NOMA-based Fog-Radio Access Networks via Joint Scheduling and Power Adaptation", in *ArXiv*, abs1902.10388.

[75] J. Du, L. Zhao, J. Feng and X. Chu, "Computation Offloading and Resource Allocation in Mixed Fog/Cloud Computing Systems With Min-Max Fairness Guarantee," in *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1594-1608, April 2018.

[76] Y. Wang, B. Ren, S. Sun, S. Kang and X. Yue, "Analysis of non-orthogonal multiple access for 5G," in *China Communications*, vol. 13, no. Supplement2, pp. 52-66, N/A 2016.

[77] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (noma)," in *Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013.

[78] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink noma for future lte enhancements," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2013.

[79] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink nonorthogonal multiple access for 5G wireless networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2014, pp. 781–785.

[80] S. Deb, P. Monogioudis, J. Miernik and J. P. Seymour, "Algorithms for Enhanced Inter-Cell Interference Coordination (eICIC) in LTE HetNets," in *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137-150, Feb. 2014, doi: 10.1109/TNET.2013.2246820.

[81] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," in *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[82] L. Zhang, et al., "Layered Division Multiplexing: Theory and Practice," in *IEEE Trans. Broadcast.*, vol. 62, no. 1, 2016.

[83] MediaTek Inc., "Study on downlink multiuser superposition transmission for LTE," 3GPP TSG RAN Meeting 68, Malmö, Sweden, Tech. Rep. RP-151100, Jun. 2015.

[84] A. Benjebbour et al., "NOMA: From concept to standardization," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Tokyo, Japan, May 2015, pp. 18–23.

[85] NTT DOCOMO, "Deployment scenarios for downlink multiuser superposition transmissions," 3GPP TSG RAN WG1 Meeting 80bis, Belgrade, Serbia, Tech. Rep. R1-152062, Apr. 2015.

[86] NTT DOCOMO, "Evaluation methodologies for downlink multiuser superposition transmissions," 3GPP TSG RAN WG1 Meeting 81, Fukuoka, Japan, Tech. Rep. R1-153332, May 2015.

[87] MediaTek Inc., "Candidate non-orthogonal multiple access," 3GPP TSG RAN WG1 Meeting 81, Fukuoka, Japan, Tech. Rep. R1-153335, May 2015.

[88] NTT DOCOMO, "System-level evaluation results for downlink multiuser superposition transmissions," 3GPP TSG RAN WG1 Meeting 82, Beijing, China, Tech. Rep. R1-154536, Aug. 2015.

[89] NTT DOCOMO, "Link-level evaluation results for downlink multiuser superposition transmissions," 3GPP TSG RAN WG1 Meeting 82, Beijing, China, Tech. Rep. R1-154537, Aug. 2015.

[90] MediaTek Inc. and CMCC, "Downlink multiuser superposition transmissions for LTE," 3GPP TSG RAN Meeting 71, Gothenburg, Sweden, Tech. Rep. RP-160680, Mar. 2016.

[91] 3GPP TR 38.812, "Study on Non-Orthogonal Multiple Access (NOMA) for NR," Release 16, [accessed on 21.03.2019].

[92] ONE5G, *https://one5g.eu/wp-content/uploads/2019/07/ONE5G D1.2 D6.2-1.pdf*, 2019.

[93] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," in *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 583–596, Feb. 2014.

[94] S. Hayashi and Z. Q. Luo, "Spectrum management for interferencelimited multiuser communication systems," in *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1153–1175, Mar. 2009.

[95] S. Boyd L. Vandenberghe, in *Convex Optimization*. Cambridge U.K.:Cambridge Univ. Press 2004.

[96] S. Boyd and A. Mutapcic, *Subgradient Methods*. Stanford, CA, USA: Stanford Univ. Press, 2008

[97] D. M. Andrews, "A survey of scheduling theory in wireless data networks," in *Wireless Communication*, New York, NY, USA: Springer, 2007.

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

[99] Qiaoyang Ye, Beiyu Rong, Yudong Chen, M. Al-Shalash, C. Caramanis, and J.G. Andrews. User association for load balancing in heterogeneous cellular networks.in *IEEE Trans. Wireless Commun.*, 12(6):2706–2716, Jun. 2013.

[100] J.Moand J.Walrand, "Fair end-to-end window-based congestion control," in *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[101] Daniel Perez Palomar and Mung Chiang. "A tutorial on decomposition methods for network utility maximization", in *IEEE J. on Sel. Areas in Commun.*, 24(8):1439–1451, 2006.

[102] Hina Tabassum, Md Shipon Ali, Ekram Hossain, Md. Jahangir Hossain, Dong In Kim, "Non-Orthogonal Multiple Access (NOMA) in Cellular Uplink and Downlink: Challenges and Enabling Techniques" in *eprint arXiv:1608.05783*.

[103] L. Liberti C. C. Pantelides "An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms" in *Journal of Global Optimization* vol. 36 no. 2 pp. 161-189 Oct. 2006.

[104] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[105] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September 2013.

[106] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang and A. Nallanathan, "Resource Allocation in NOMA-Based Fog Radio Access Networks," in *IEEE Wireless Communications*, vol. 25, no. 3, pp. 110-115, June 2018.

[107] C. Liang, F. R. Yu, H. Yao and Z. Han, "Virtual Resource Allocation in Information-Centric Wireless Networks With Virtualization," in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9902-9914, Dec. 2016.

[108] *C-RAN the Road Towards Green RAN-White Paper*, China Mobile Res. Inst., China Mobile, Beijing, China, Oct. 2011.