Performance Analysis of Non-Orthogonal Multiple Access (NOMA) enabled Cloud Radio Access Networks

Rupesh Rai, *Student Member, IEEE,* Huiling Zhu, *Senior Member, IEEE,* and Jiangzhou Wang, *Fellow, IEEE*

Abstract

In this paper, 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 obtain a feasible solution to the remote radio head (RRH) muting problem for given bandwidth and transmit power. The second subproblem is then reformulated for tractibility purpose and a low-complexity algorithm is proposed to bandwidth and power allocation subject to users data rate constraints. Moreover, for given user association and muting states, the near optimal power allocation is derived in a closed-form expression. 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.

I. Introduction

Cloud radio access network (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 [1]-[4]. The pool of baseband unit (BBU) and remote radio heads (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

Rupesh Rai, Huiling Zhu and Jiangzhou Wang are with the School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NT, UK (email:{rr304, H.Zhu, J.Z.Wang}@kent.ac.uk).

respective base station (BS) with various services such as inter-cell interference management and increased network capacity [5]. A C-RAN provides a new cost-effective way to achieve network densification where the conventional base stations (BSs) are replaced by low-power and low-complexity remote radio heads (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 (ICI) 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 [6,7]. In [8], two optimization models were proposed for i) resource allocation and power minimization ii) BBU-RRH assignment problem in C-RAN. In [9], the authors proposed a quality of service (QoS)-aware radio resource optimization solution for maximizing downlink system utility in C-RAN. In [10], a joint scheduling strategy for resource allocation in C-RAN was proposed where the time/frequency resources of multiple RRHs were 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 spectrum efficiency in wireless networks, non-orthogonal multiple access (NOMA) was proposed in [11,12]. In NOMA, multiple users are multiplexed by superposition coding in the power domain on the transmitter side and employ successive interference cancellation (SIC) to seperate multi-user signal on the receiver side. Compared to orthogonal multiple access (OMA), NOMA allows BSs to serve multiple users simultaneously in the same frequency band 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) were investigated in [13,14]. In order to enhance user fairness in cellular downlink, the proportional fair (PF) based scheduling was introduced in NOMA [15]. The problem of effective capacity of NOMA systems subject to QoS was investigated in [16], in which a sub-optimal power control approach was proposed to maximize the sum capacity. In [17], the resource allocation problem in the NOMA system was divided into two cases and an algorithm based on dynamic programming and Lagrangian dual optimization was proposed. A joint power control and subcarrier allocation problem was studied in [18] to minimize the overall transmit power. Although the authors in [19] studied user association and power control in single-cell NOMA networks, the effect of inter-cell interference (ICI) in practical multi-cell scenerios was not considered.

A. Related Work and Motivation

Although some outstanding works have been researched on C-RAN and NOMA, these two areas were addressed seperately. Most of the existing work on resource allocation in C-RAN considered orthogonal frequency division multiple access (OFDMA) based multi-user transmission [20]-[22]. However NOMA is more appealing for the high throughput demanding wireless network such as 5G. To acheive higher spectral efficiency of NOMA in more realistic setting, it is necessary to consider multi-cell network. Some recent work on NOMA has been extended to multi-cell systems in [23,24]. Applying NOMA technology into the C-RAN network may bring significant benefits.

In dense wireless networks, ICI becomes the major obstacle, which degrades the QoS of cell-edge users (CEUs). In order to tackle these problems, some recent work has applied NOMA with other techniques such as coordinated multi-point (CoMP) [25,26] and cooperative communication [27]. [28] proposed a NOMA scheme for wireless downlink C-RAN to improve spectral efficiency as well as to support number of connections in C-RAN. In [29], the authors analyzed the outage probability of the NOMA-enabled C-RAN. Similarly, the authors in [30] derived the expressions in terms of outage probability for both cell-edge and cell-center users (CCUs). A lot of work has focused NOMA in different network scenerios [31-32] in order to improve resource utilization. The motivation behind considering C-RAN architecture is that the traditional distributed networks lacks in global network information and with local network information, scheduling can be performed sub-optimally. Therefore many tasks of scheduling can benefit from centralized global perspective. 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 of work has 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.

B. Contributions and Organization

An important aspect of C-RAN design is that there exists a large set of trade-off parameters and objectives. Therefore, a framework of multi-objective problem is investigated in this paper. In a dense C-RAN, supporting all users by higher throughput and lower power consumption is critical. Moreover, the accompanying inter-RRH interference has become the fundamental challenge for an effective resource management. Therfore, this paper conducts the joint research on downlink NOMA-enabled C-RAN which aims at minimizing the number of active RRHs and assigning feasible power and bandwidth while satisfying the data rate requirements of all users. Specifically, we investigate the joint optimization of RRH-UE association, muting to minimize the power consumption and bandwidth-power allocation (BPA) subject to all UEs rate requirements and per-RRH bandwidth and power constraints. To tackle the joint optimization problem we decouple the main problem into two subproblems as joint RRH and user equipment association and muting, and BPA. The main contributions of this paper are summarized as follows:

1) 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 BPA with the aim of maximizing UEs sum rate and network utility while considering users' fairness issues.
2) The problem belongs to mixed-integer non-linear programming (MINLIP) 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 centralized muting algorithm which updates the RRHs muting states by subgradient method. We also propose heuristic algorithm to find the muting states which improves the CEUs performance and overall system performance using the Jain's fairness index.
3) We propose an adaptive resource allocation strategy that minimizes the total transmit power by following two strategies: a) reduce the number of active RRHs by employing key idea of coordinated silencing (RRH muting). b) minimize 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 bandwidth and power allocation (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 utility. For a given bandwidth allowance, power allocation for RRHs is formulated as non-convex problem which is solved by transforming it into convex problem and applying Karush-Kuhn-Tucker (KKT) conditions. Based on the transformed Lagrangian function, near optimal power allocation is derived in closed form.

4) The performance of our proposed framework is evaluated for joint optimization problem for NOMA-enabled C-RAN systems via simulation to validate that our proposed algorithms can obtain near 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 RRHs transmission power considering the bandwidth constraint in comparison with the conventional OMA-enabled C-RAN transmission scheme as well as fixed BPA scheme.

The rest of the paper is organized as follows: In Section II we describe the system description and channel model for NOMA-enabled C-RAN. The optimization problem is formulated and decomposed into subproblems in Section III. Section IV discusses the proposed iterative user association and muting problem. Section V investigates the optimal bandwidth and power allocation methods. Simulation results are presented in Section VI, which is followed by conclusions in Section VII.

II. System Description and Channel Model

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

A. System model of NOMA-enabled C-RAN

Consider a downlink of a cloud based radio access network architecture. Fig.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 CCUs and 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 CEUs from the neighbouring cell. It is assumed that CCUs do not suffer from any ICI. Single transmit and receive antenna is assumed to be equipped with all RRHs and users respectively. The set of RRHs which are dominant interfering RRHs that interfere with UE_n is expressed as $I_r = \{r|b_{rn} = 0, \forall r \in \mathcal{R}\}$ where b_{rn} is the user-RRH indicator and $b_{rn} = 1$ indicates the n-th user is served by the r-th RRH, $b_{rn} = 0$ otherwise. $\alpha_{kr} = 1$ or 0 determines whether the r-th RRH exploits the k-th subchannel. Fig. 1 includes the table for dominant interfering RRHs

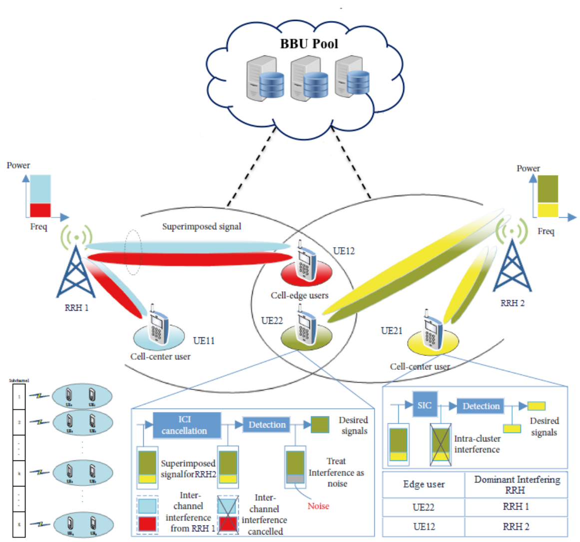


Figure 1: System Model for NOMA-enabled C-RAN

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 on k-th subchannel is P_k^r, and the total available bandwidth is B Hz and the bandwidth allocation factor of the k-th subchannel is B_{ij}^k where B_{ij}^k is the portion of the entire downlink bandwidth allocated to the NOMA pair (i, j) and $0 \leq B_{ij}^k \leq 1$, $\sum_{k=1}^{K_n} B_{ij}^k \leq 1$ where K_n represents the NOMA pairs. The NOMA pair is allowed to utilize $B_{ij}^k \in [0, 1]$ portion of the entire downlink bandwidth B, $\forall r, K_n$. The subchannels are indexed by $\mathcal{K} = \{k|1 \leq n \leq K\}$. 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 $|h_{nk}^{ri}|^2 \geq |h_{nk}^{rj}|^2$, where h_{nk}^{ri} and h_{nk}^{rj} represents the channel gain coefficients between i-th user and RRH r and between j-th user and RRH r.

The superposition coded symbol transmitted by rth RRH is:

Table I: Notation Summary of System Model

Notation	Description
$\mathcal{R}$	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}^k	Portion of the entire downlink bandwidth allocated to the NOMA pair (i, j)
γ_{nk}^i	Received signal-to-interference-plus-noise ratio (SINR) for the i-th decoded user on subchannel k
β_{rk}^m	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}	User-RRH association indicator
α_{kr}	Indicator whether or not the r-th RRH exploits k-th subchannel
I_{nk}^r	Average ICI from other RRHs

$$x_k^r = \sum_{n=1}^{N} b_{rn}\alpha_{kr}\sqrt{P_k^r}x_{kn}^r \tag{1}$$

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

$$y_{nk}^r = h_{nk}^r x_k^r + I_{nk}^r + \zeta_{rk}^n \tag{2}$$

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 cumulative interference to UE n from other RRHs except RRH r with unit bandwidth given by:

$$I_{nk}^r = \sum_{m=1, m\neq r}^{R} h_{nk}^m x_k^m \tag{3}$$

where P_k^r and B_{max} are the maximum power and bandwidth respectively and P_k^{rn} is the allocated power of UE n associated with RRH r on subchannel k.

$$P_k^r = \sum_{n=1}^{N} b_{rn}\alpha_{kr}P_k^{rn} \tag{4}$$

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

$$f_{rk}^n = \frac{b_{rn}\alpha_{kn}|h_{nk}^r|^2}{\left(\sum_{m=1,m\neq r}^{R}|h_{nk}^m|^2 P_k^r + N_0 B_{max}\right)} \tag{5}$$

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

$$\gamma_{nk}^i = \frac{b_{rn}\alpha_{kr}|h_{nk}^r|^2 P_{ki}^r}{\sum_{j=i+1}^{n_i} b_{rn}\alpha_{kr}|h_{nk}^r|^2 P_{kj}^r + B_{max}N_0} \tag{6}$$

where P_{ki}^r is the allocated power of UE i associated with RRH r on subchannel k. 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 paper, we assume that $N_k = 2$ for $k = 1, 2, \dots K$ and $N = 2K$. In each subchannel, the signals transmitted for NOMA users are ordered based on their channel quality i.e.$|h^i_{nk}|^2 \geq |h^j_{nk}|^2$. In each subchannel NOMA protocol is implemented. We consider a pair of two users i and j served by RRH r and the UE_i wants to decode and remove UE_j's signal by SIC on RB k, then the following inequality holds:

$$\frac{b_{rn}\alpha_{kr}a^j_{rk}|h^{ri}_{nk}|^2}{\sum_{m=1,m\neq r}^{R}|h^{mi}_{nk}|^2 a^i_{rk}P^r_k + N_0 B_{max}} \geq \frac{b_{rn}\alpha_{kr}a^i_{rk}|h^{rj}_{nk}|^2}{\sum_{m=1,m\neq r}^{R}|h^{mj}_{nk}|^2 a^j_{rk}P^r_k + N_0 B_{max}} \tag{7}$$

$I^i_{nk} = |h^{mi}_{nk}|^2 a^i_{rk}P^r_k$ and $I^j_{nk} = |h^{mj}_{nk}|^2 a^j_{rk}P^r_k$ are the superposed interference to decode the signal of UE j at UE i and j respectively. RRH r sends messages to RRH users (RUEs) n_1 and n_2 on subchannel k by superposition i.e. RRH r sends $x_n = a^i_{rk}x^{n_1}_{rk} + a^j_{rk}x^{n_2}_{rk}$, where a^i_{rk} and a^j_{rk} are the power sharing coefficients.

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:

$$\gamma^i_{nk} = \frac{b_{rn}\alpha_{kr}|h^{ri}_{nk}|^2 P^{ri}_k}{B_{max}N_0} \tag{8}$$

$$\gamma^j_{nk} = \frac{b_{rn}\alpha_{kn}|h^{rj}_{nk}|^2 P^{rj}_k}{b_{rn}\alpha_{kn}|h^{ri}_{nk}|^2 P^{ri}_k + B_{max}N_0} \tag{9}$$

In terms of vector β^m_{rk}, the average ICI $I^r_{nk}(\beta^m_{rk})$ from other RRHs defined as:

$$I^r_{nk}(\beta^m_{rk}) = \sum_{m=1,m\neq r}^{R}(1-\beta^m_{rk})h^m_{nk}P^{rm}_k x^m_k \tag{10}$$

where β^m_{rk} is the muting arrangement matrix of dimensions $R \times K$ with elements $\beta^m_{rk} = 0$, the RRH $r \in \mathcal{R}$ is muted on resource block (RB) or called chunk [33,34] $k \in \mathcal{K}$. The muting arrangement is determined so as to minimize the interference among concurrent transmissions which 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. From (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.

III. 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 resorce 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:

$$O(\beta_{rk}^m, b_{rn}, B_{ij}^k, P_k^r) = \max_{\beta_{rk}^m, b_{rn}, B_{ij}^k, P_k^r} \sum_{n=1}^{N} (\sum_{r=1}^{R} R_{CS}) \tag{11a}$$

$$s.t. \sum_{n\in N} \sum_{k\in K_N} B_{ij}^k \leq \beta_{rk}^m B_{max} \; \forall r \in R \tag{11b}$$

$$\sum_{r\in R} \sum_{n\in N} \sum_{k\in K} b_{rn} \alpha_{kr} P_k^r \leq P^m \tag{11c}$$

$$\sum_{r\in R} \sum_{k\in K} b_{rn} \alpha_{kr} R_{ij}(\beta_{rk}^m, b_{rn}) \geq r_{min} \quad \forall n \in N \tag{11d}$$

$$b_{rn} \leq \alpha_{kr} \beta_{rk}^m \quad \forall n \in N, \; \forall r \in R \tag{11e}$$

$$a_{rk}^i + a_{rk}^j \leq 1 \; \forall r \in R \tag{11f}$$

$$a_{rk}^i \geq 0, a_{rk}^j \geq 0 \; \forall r \in R \tag{11g}$$

$$\beta_{rk}^m, b_{rn} \in \{0, 1\} \tag{11h}$$

where $R_{CS} = B(1 + \gamma_{nk}^i)$ denotes the throughput that user n can obtain with coordinated scheduling when associated with RRH r, $R_{CS}(t+1) = (1 - \frac{1}{t_c})R_{CS}(t) + \frac{1}{t_c}(\sum_{k\in K_s} R_i(t) + \sum_{k\in K_w} R_j(t))$ is the long-term averaged rate and $R_{ij}(\beta_{rk}^m, b_{rn}) = \sum_{k\in K_s} R_i(t) + \sum_{k\in K_w} R_j(t)$ denotes the achievable rate on RB k and its dependence on muting indicator β_{rk}^m and user-RRH association, $R_i = B(1+\gamma_{nk}^i)$ and $R_j = B(1+\gamma_{nk}^j)$ are the data rates of UEs i and j respectively, K_s and K_w are the RB indices in which the user is scheduled as the strong user or the weak user respectively, constraint 11(b) accounts for the bandwidth budget, constraint 11(c) means that the RRHs total transmit power cannot exceed maximum transmission power capacity. Constraint 11(d) guarantees the quality of service (QoS) requirement of UEs by keeping the rate above or equal to the minimum rate requirements. In constraint 11(e) $\alpha_{kr}\beta_{rk}^m = 1$ indicates that UE can connect to RRH and allocated to subchannel k and $b_{rn} \leq \alpha_{kr}\beta_{rk}^m$ is always satisfied; when $\beta_{rk}^m = 0$, b_{rn} must be zero for all k. Thus the RRHs that are muted can be excluded from the problem formulation. Constraint 11(f) gives the upper bound of the transmit power of the RRHs and constraint 11(g) gives the non-negative transmit power constraint for the RRHs. The joint problem (11) is mixed combinatorial non-convex problem due to binary constraints for user association as well as the muting indicator and non-convex objective function $R_{CS}(\beta, p)$.

Fig. 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

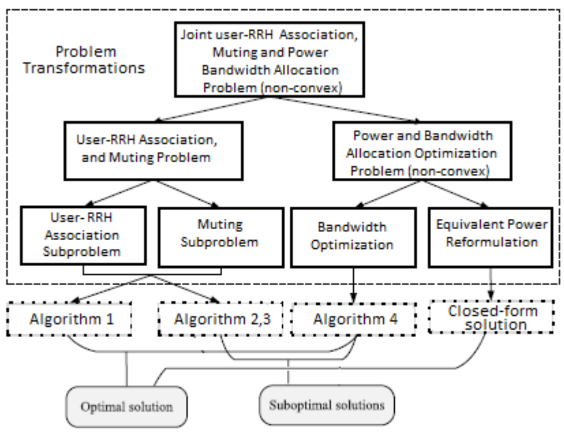


Figure 2: Overview of the proposed approach to solve the joint optimization problem

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 4(B). We develop Algorithm 2 to determine the actual number of muting states. We obtain the optimal solutions with problem transformations and Lagrangian dual analysis. In addition, a low-complexity Algorithm 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 a closed-form expression subject to QoS constraints.

IV. 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}^{m}, 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 $\hat{\beta_{rk}^{m}} = log_2(\beta_{rk}^{m})$, we will have a convex optimization problem with respect to $\hat{\beta_{rk}^{m}}$ which can be expressed as:

$$O_1(\hat{\beta_{rk}^{m}}, b_{rn}) = \max_{\hat{\beta_{rk}^{m}}, b_{rn}} \left[log \left(\sum_{n=1}^{N} \big(\sum_{r=1}^{R} R_{CS}(e^{\hat{\beta_{rk}^{m}}}, b_{rn}) \big) \right) \right] \tag{12a}$$

s.t.

$$\sum_{n \in N} \sum_{k \in K_N} B_{ij}^{k} \leq e^{\hat{\beta_{rk}^{m}}} B_{max} \quad \forall r \in R \tag{12b}$$

$$\sum_{r \in R} \sum_{u \in N} \sum_{k \in K} b_{rn} \alpha_{kr} P_k^r \leq P^m \tag{12c}$$

$$\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kr} R_{ij}(\hat{\beta_{rk}^{m}}, b_{rn}) > r_{min} \quad \forall n \in N \tag{12d}$$

$$b_{rn} \leq \alpha_{kr} e^{\hat{\beta_{rk}^{m}}} \quad \forall n \in N, \ \forall r \in R \tag{12e}$$

$$0 \leq b_{rn} \leq 1 \quad \forall n \in N, \ \forall r \in R \tag{12f}$$

$$\beta_{rk}^{m} \leq 0 \quad \forall r \in R \tag{12g}$$

In order to develop an efficient algorithm, we relax the integer constraints β_{rk}^{m} and b_{rn} to take values in [0,1]. Problem (12) is concave optimization problem. The objective function of (12) contain sum-rate expression inside the log function. Since log(.) is a concave function, the objective of (12) is concave [35]. In (12) the decision variables β_{rk}^{m} and b_{rn} are coupled in the constraints. Moreover, the objective function include the sum of quadratic expressions, which is difficult to tackle directly. We introduce the auxillary variable $\mathcal{S}_r = \sum_{n \in N} b_{rn}$ in the problem $O(\hat{\beta_{rk}^{m}}, b_{rn})$. We decompose the problem into two subproblems in order to decouple the variables.

Firstly, given the values of $\mathcal{S}_r$ and $\hat{\beta^m_{rk}}$, we find the optimal user association b_{rn} and consequently we find the optimal values of $\mathcal{S}_r$ and $\hat{\beta^m_{rk}}$. We provide the NP-hardness analysis as below:

Theorem 1: (11) is NP-hard.

Proof: Firstly we conclude that if N=1, problem (11) is NP-hard according to [38] where the problem reduces to OFDMA subchannel and power allocation. For multi-carrier NOMA, with $M > 2$, we consider an instance of (11) with N users, K RBs and M=2. The total power is given by NKP^r_k. The power limit $P^r_k = 1$ 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 $\frac{g_{\bar{kn}}}{g_{kn}}$ 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 $w_{\bar{k}}R_{\bar{KN}} = e^{KN}log(1+\frac{p}{\epsilon})$ is greater than $KNe^{-KN}log(1+\frac{e^{-KN}p}{\epsilon})$. Thus allocating power to user $\bar{n}$ is preferable for maximizing 11(a). Consequently, a special case of (11) with $M > 1$ is equivalent to the problem in [38] and the result follows.

A. User Association subproblem

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

$$O_{1a}(b_{rn}) = \max_{b_{rn}} \left[log\left(\sum_{n=1}^{N} \sum_{r=1}^{R} R_{CS}(b_{rn}) \right) \right] \tag{13a}$$

s.t. 12(c) - 12(g)

$$\mathcal{S}_r = \sum_{n \in N} b_{rn} \tag{13b}$$

Based on (13), the Lagrangian function can be written as:

$$\begin{aligned} \mathcal{L}(b, \lambda, \mu, \theta, \rho) = & \left[log\left(\sum_{n=1}^{N} \sum_{r=1}^{R} R_{CS}(b_{rn}) \right) \right] - \lambda_r \left(\sum_{r \in R} \sum_{n \in N} \sum_{k \in K} b_{rn}\alpha_{kr}P^r_k - P^m \right) \\ & - \sum_{n \in N} \mu_n \left(r_{min} - \sum_{r \in R} \sum_{k \in K} b_{rn}\alpha_{kr}R_{ij}(\hat{\beta^m_{rk}}, b_{rn}) \right) + \sum_{r \in R} \sum_{n \in N} \theta_{rn}(\alpha_{kr}e^{\hat{\beta^m_{rk}}} - b_{rn}) \\ & + \rho_r(\mathcal{S}_r - \sum_{n \in N} b_{rn}) \end{aligned} \tag{14}$$

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 (12e) and (13b). The operator $\succeq 0$ indicates that all the elements of the vector are nonnegative. The dual problem is given by:

$$\min_{\lambda,\mu,\theta,\rho} g(\lambda, \mu, \theta, \rho) \tag{15}$$

$$s.t. \quad \lambda_r \geq 0, \mu_n \succeq 0, \theta_{rn} \succeq 0 \; and \; \rho_r \succeq 0 \tag{16}$$

$$g(\lambda, \mu, \theta, \rho) = \begin{cases} \max_{b_{rn}} \mathcal{L}(b, \lambda, \mu, \theta, \rho) \\ s.t. \quad (12c), (12d), (12e) \end{cases} \tag{17}$$

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

$$b_{rn^*} = \begin{cases} 1, \text{ if } r = r^* \\ 0, \text{otherwise} \end{cases} \tag{18}$$

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

$$\zeta = R_{CS} b_{rn} - \lambda_r P_k^r + R_{ij} - \theta_{rn} + \rho_r \tag{19}$$

Lemma 1: Problem (13) holds a strong duality.

Proof: In Problem (13), there exists a feasible **x**, such that the linear constraints are satisfied and the problem is feasible while inequalities hold (12e). The constraint qualification (Slater's condition) is satisfied and strong duality holds.

The subgradient method [36] can be utilized to obtain the optimal solution of given dual problem.

$$\lambda_r(t+1) = \left[\lambda_r(t) - \xi_1 \times \left(P^m - \sum_{r\in R}\sum_{n\in N}\sum_{k\in K} b_{rn}\alpha_{kr}P_k^r\right)\right]^+ \tag{20}$$

$$\mu_n(t+1) = \left[\mu_n(t) - \xi_2 \times \left(\sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kr}R_{ij}(\hat{\beta_{rk}^m}, b_{rn} - r_{min}\right)\right]^+ \tag{21}$$

$$\theta_{rn}(t+1) = \left[\theta_{rn}(t) - \xi_3 \times \left(\alpha_{kr}e^{\hat{\beta_{rk}^m}} - b_{rn}\right)\right]^+ \tag{22}$$

$$\rho_r(t+1) = \left[\rho_r(t) - \xi_4 \times \left(\mathcal{S}_r - \sum_{n\in N} b_{rn}\right)\right]^+ \tag{23}$$

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

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 C-RAN 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 for the updated values.

conditions for RUE. The RUEs receives pilot signal to calculate the RSRP (received signal received power) and reports back to the C-RAN 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 $r^* = \max_r(\zeta)$.

B. 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:

$$O_{1b}(S_r, \hat{\beta^m_{rk}}) = \max_{S_r, \hat{\beta^m_{rk}}} O(b_{rn}(S_r, \hat{\beta^m_{rk}})) \tag{24}$$

s.t. (12b), (12c), (12d), (12e)

We find the optimal S_r and $\hat{\beta^m_{rk}}$ by solving the above problem. Let $b^*_{rn}(S^{'}_r)$ be the optimal solution for given problem (13) and $O^*(S^{'}_r)$ be the objective function and we find the optimal value by:

$$O^*(S_r) = \max_{S_r} log[f^*(S^{'}_r)] \tag{25}$$

s.t. (12b), (12c), (12d), (12e) We consider solution b_{rn} for S_r. The following inequalities hold:

$$\begin{aligned}
f^*(S^{'}_r) &= \mathcal{L}(S^*(b^*_{rn}), \lambda^*(b^*_{rn}), \mu^*(b^*_{rn}), \theta^*(b^*_{rn}), \rho^*(b^*_{rn})) \\
&\geq \mathcal{L}(S^*(b_{rn}), \lambda^*(b^*_{rn}), \mu^*(b^*_{rn}), \theta^*(b^*_{rn}), \rho^*(b^*_{rn})) \\
= log[f^*(S^{'}_r) - \sum_{n=1}^{N}\sum_{r=1}^{R} & R_{CS}(S^{'}_r - S_r)] + \rho^*_r(b_{rn})\Big(\sum_{n\in N} b_{rn} - S^*_r(b_{rn})\Big) \\
&+\lambda_r(b^*_{rn}))\Big(\sum_{r\in R}\sum_{n\in N}\sum_{k\in K} b_{rn}\alpha_{kr}P^r_k - P^m\Big) \\
-\mu_n(b^*_{rn}))\Big(r_{min} - \sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kr}R_{ij}(\hat{\beta^m_{rk}}, b_{rn})\Big) &+ \sum_{r\in R}\sum_{n\in N}\theta_{rn}(b^*_{rn})(\alpha_{kr}e^{\hat{\beta^m_{rk}}} - b_{rn})
\end{aligned} \tag{26}$$

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:

$$S(t+1) = S(t) + \frac{f(S_r(t), \hat{\beta_{rk}^m}(t) - f(S_r^*, \hat{\beta_{rk}^m}(t)}{|\rho_r^*(t) - \sum_{n=1}^{N}\sum_{r=1}^{R} R_{CS}|}(\rho_r^*(t) - \sum_{n=1}^{N}\sum_{r=1}^{R} R_{CS}) \tag{27}$$

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

$$\begin{aligned} h^*(\hat{\beta_{rk}^{m'}}) &= \mathcal{L}(\hat{\beta_{rk}^{m'}}((b_{rn}^*)), \lambda^*(b_{rn}^*), \mu^*(b_{rn}^*), \theta^*(b_{rn}^*)) \\ &\geq \mathcal{L}(\hat{\beta_{rk}^{m'}}(b_{rn}), \lambda^*(b_{rn}^*), \mu^*(b_{rn}^*), \theta^*(b_{rn}^*)) \\ &= h^*(\hat{\beta_{rk}^m} + \sum_{n\in N} \theta_{rn}^*(b_{rn}^*)(\alpha_{kr}e^{\hat{\beta_{rk}^m}} - b_{rn}) + \nabla \end{aligned} \tag{28}$$

where λ_r, μ_n are the Lagrangian multipliers corresponding to constraints (12c) and (12d). ∇ is given by:

$$\nabla = \lambda_r(b_{rn}^*)\left(\sum_{r\in R}\sum_{n\in N}\sum_{k\in K} b_{rn}\alpha_{kr}P_k^r - P^m\right) - \mu_n(b_{rn}^*))\left(r_{min} - \sum_{r\in R}\sum_{k\in K} b_{rn}\alpha_{kr}R_{ij}(\hat{\beta_{rk}^m}, b_{rn}\right) \tag{29}$$

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

$$\begin{aligned} \hat{\beta_{rk}^m}(t+1) &= \hat{\beta_{rk}^m}(t) + \frac{f(S_r(t), \hat{\beta_{rk}^m}(t) - f(S_r, \hat{\beta_{rk}^m}^*(t)}{||\theta(t)||^2}\theta(t), \\ where \quad \theta(t) &= (\sum_{n\in N}\theta_{1n}(t), \sum_{n\in N}\theta_{2n}(t) + \ldots + \sum_{n\in N}\theta_{rn}(t)) \end{aligned} \tag{30}$$

The update depicts the performance gain acheived in terms of data rate with the increase in number of muting dominant interfering RRHs. S_r needs to be an integer no greater than $\hat{\beta_{rk}^m}$ in accordance with the constraint (11e). After the convergence of S_r, the final value of S_r is rounded up to a greater value of objective function and less than $\hat{\beta_{rk}^m}$. The update given by (30) is a projection to the feasible regions of S_r and $\hat{\beta_{rk}^m}$ and terminates whenever the boundary values are achieved. The equalities and inequalities hold in (28) where the equality in (28) is due to the strong duality and inequality is due to the near optimality of $\hat{\beta_{rk}^{m'}}$. 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.

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

1. Inputs $S_r, \hat{\beta_{rk}^m}, \rho, \lambda, \theta, \mu$
2. Initialize $S_r, \hat{\beta_{rk}^m}, \theta_{rn}, \rho_r$
3. (Repeat)
4. Solve the problem (14) by Lagrangian dual method.
5. Update θ_{rn}, ρ_r
6. Until θ_{rn}, ρ_r do not converge;
7. Update $S_r, \hat{\beta_{rk}^m}$ from (27) and (30)
8. Until $S_r, \hat{\beta_{rk}^m}$ do not converge;

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

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

$$I_{nk}^r(\beta_{rk}^m) = \sum_{m=1, m \neq r}^{R} (1 - \beta_{rk}^m) h_{nk}^m P_k^{rm} x_k^m \tag{31}$$

The ICI experienced by CEU considering dominant interferers is given by:

$$I_c = \sum_{j=2}^{I_r} |h_{ij}|^2 P_k^r \tag{32}$$

where power transmitted by RRH is $P_k^r = E|x|^2$, $x = \sqrt{P_r^{ki}} x_1 + \sqrt{P_r^{kj}} x_2$. 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:

$$I_{total} = \sum_{j=2}^{I_r} F[r', c] \tag{33}$$

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 CEUs are identified based on normalized channel gains derived in Appendix B. If we assume that a CEU is selected that is liable to suffer from ICI from the neighbouring cell, the following constraint will apply:

$$R(P_k^{r1j}) - R(P_k^{r2j'}) \leq \eta \tag{34}$$

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 (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 shows the signalling sequences for the proposed approach. The central processor is in charge of

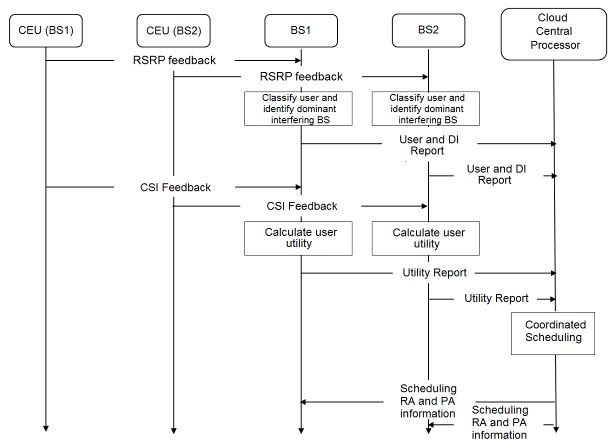


Figure 3: Signalling sequences for the proposed approach

collecting 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 CEUs 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., $\beta_{rk}^{m} = 1$. We initialize the scheduling matrix S, which is the scheduling set of all UEs. Users are scheduled based on the categorization of CEUs and CCUs. 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. Each RRH performs

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 (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:

$$I_{r_m} = \bigcup_{r_m=\{r_1,r_2...r_m\}} \binom{I_{r_m}}{b_r} \tag{35}$$

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

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:

$$w_k(t) = \sum_{n \in N} b_{rn}(t) \left(\frac{r_{nk}(t)}{R_u(t)} \right) \tag{36}$$

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

$$R_n(t+1) = \left(1 - \frac{1}{t_c}\right) R_n(t) + \frac{1}{t_c} \sum_{k \in K} s_{nk}(t) R_{ij}(t) \tag{37}$$

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. $r_{nk}(t)$ is the instantaneous data rate.

$$R_n(t+1) = \left(1 - \frac{1}{t_c}\right) R_n(t) + \frac{1}{t_c} \left(\sum_{k \in K_s} r_{sk}(t) + \sum_{k \in K_w} r_{wk}(t) \right) \tag{38}$$

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 (34). When RRH is muted, the maximum PF metrices are calculated among all UEs n on RBs k. The muting is stopped when the additional muting of RRHs does not improve the sum of PF metrices. The set of muting decisions are defined in (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 (36) that maximizes the weghted sum rate. The loop runs until all users are scheduled. Flag is set to 2 for a maximum number of multiplexed users on the same RB. We get the scheduling set S with scheduled UEs.

V. 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 following minimization problem is formulated as:

$$O_2(B^k_{ij}, P^r_k) = \min_{B^k_{ij}, P^r_k} B_{ij} \sum_{k=1}^{K} P^r_k \tag{39a}$$

s.t.

$$\sum_{n \in N} \sum_{k \in K_N} B^k_{ij} \leq \beta^m_{rk} B_{max} \; \forall r \in R \tag{39b}$$

$$\sum_{r \in R} \sum_{n \in N} \sum_{k \in K} b_{rn} \alpha_{kn} P^r_k \leq P^m \tag{39c}$$

$$\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kr} R_{ij}(\hat{\beta^m_{rk}}, b_{rn}) > r_{min} \quad \forall n \in N \tag{39d}$$

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.

A. 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 $P^{r^*}_k$. 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:

$$\sum_{n \in N} \sum_{k \in K_N} B^k_{ij} \leq B \; \forall r \in R \tag{40}$$

We obtain the following optimization problem:

$$O_{2a}(B^k_{ij}, P^{r*}_k) = \min_{B^k_{ij}, P^{r^*}_k} \sum_{n \in N} \sum_{k \in K} B^k_{ij} \tag{41a}$$

s.t.

$$\sum_{m=1}^{M} B^k_{ij} \leq B \; \forall r \in R \tag{41b}$$

$$\sum_{n\in N}\sum_{k\in K_N} B_{ij}^{k} \leq \beta_{rk} B_{max}\ \forall r \in R \tag{41c}$$

$$R_{ij}(B_{ij}^{k}, P_k^{r^*}) > r_{min} \quad \forall n \in N,\ \ k \in K \tag{41d}$$

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:

$$\begin{aligned}\mathcal{L}(B_{ij}^{k}, \lambda, \mu, \theta) = \sum_{n\in N}\sum_{k\in K} B_{ij}^{k} - \lambda \sum_{r\in R}(B - \sum_{m=1}^{M} B_{ij}^{k} + \mu \sum_{r\in R}(\sum_{n\in N}\sum_{k\in K_N} B_{ij}^{k} - \beta_{rk}^{m} B_{max}) \\ + \sum_{n\in N} \theta \left(r_{min} - R_{ij} B_{ij}^{k}, P_k^{r^*}) \right)\end{aligned} \tag{42}$$

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

$$C_{i,j} = \sum_{n\in N} \mu \beta_{rk}^{m} \tag{43}$$

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

$$\sum_{n\in N}\sum_{k\in K} B_{ij}^{k} + \lambda + \mu = C_{ij} \tag{44}$$

$$-\lambda(B - \sum_{m=1}^{M} B_{ij}^{k}) = 0 \tag{45}$$

$$\mu(\sum_{n\in N}\sum_{k\in K_N} B_{ij}^{k} - \beta_{rk}^{m} B_{max}) = 0 \tag{46}$$

$$\theta \left(r_{min} - R_{ij}(B_{ij}^{k}, P_k^{r^*}) \right) \tag{47}$$

$$\lambda, \mu, \theta \geq 0 \tag{48}$$

From (43) and (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.

$$B^* = \left[\frac{C - \theta}{b_{rn}}\right]^{B_{max}} \tag{49}$$

We update the bandwidth allocation for each subchannel as:

$$B(t+1) = \left[\frac{C(t) - \theta(t)}{b_{rn}}\right]^{B_{max}} \tag{50}$$

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

$$B_{ij}^{k}(t+1) = [B_{ij}^{k} - \delta(C(t) - C_{ij}(t))]^{+} \tag{51}$$

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:
   $B^*(t+1) = \left[\frac{C(t)-\theta(t)}{b_{rn}}\right]^{B_{max}}$ and $B_{ij}^{k}(t+1) = [B_{ij}^{k} - \delta(C(t) - C_{ij}(t))]^{+}$
5. The bandwidth allocated with minimum cost is given by: $B^*(t+1) - B_{ij}^{k}(t+1)$
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 (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.

B. Power Allocation

The problem given in (39) can be reformulated into an equivalent form. Given that 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:

$$O_{2b}(B_{ij}^{k*}, P_k^{r*}) = \min_{P_k^{r*}} P_k^r \tag{52a}$$

s.t.

$$\sum_{n \in N} B_{ij}^{k} = \beta_{rk}^{m} B_{max} \; \forall r \in R \tag{52b}$$

$$\sum_{r \in R} \sum_{n \in N} \sum_{k \in K} b_{rn} \alpha_{kr} P_k^r \leq P^m \tag{52c}$$

$$\sum_{r \in R} \sum_{k \in K} b_{rn} \alpha_{kr} R_{ij}(\hat{\beta_{rk}^{m}}, b_{rn}) > r_{min} \quad \forall n \in N \tag{52d}$$

$$a_{rk}^{i} + a_{rk}^{j} \leq 1 \; \forall r \in R \tag{52e}$$

To transform the given problem into concave function, we set $a_{rk}^{i} = 2^{x_{rk}^{i}}$, $a_{rk}^{j} = 2^{x_{rk}^{j}}$ and define $P_k^{r*} = [x_{rk}^{i}, x_{rk}^{j}]$. 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:

$$\frac{b_{rn}\alpha_{kn}|h^r_{nk}|^2 P^r_{kl}}{\sum_{j=l+1}^{u_l} b_{rn}\alpha_{kr}|h^r_{nl}|^2 P^r_{kj} + B_{max}N_0} \geq \gamma^i_{nk} \tag{53}$$

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

$$P^{rj}_k = \frac{B_{max}N_0}{b_{rn}\alpha_{kn}|h^{ri}_{nk}|^2}\gamma^i_{nk} \tag{54}$$

$$P^{ri}_k = \frac{b_{rn}\alpha_{kn}|h^{ri}_{nk}|^2 + B_{max}N_0}{b_{rn}\alpha_{kn}|h^{rj}_{nk}|^2}\gamma^j_{nk} \tag{55}$$

where $\gamma^i_{nk} = (2^{\frac{R_i}{B^k_{ij}}} - 1)$ and $\gamma^j_{nk} = (2^{\frac{R_j}{B^k_{ij}}} - 1)$ 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 $g^{ri}_{nk} = |h^{ri}_{nk}|^2$, $g^{ri*}_{nk} = |h^{ri*}_{nk}|^2$, $g^{rj}_{nk} = |h^{rj}_{nk}|^2$ and $g^{rj*}_{nk} = |h^{rj*}_{nk}|^2$ and the power levels are indicated by $g^{ri}_{nk}P^{ri}_k$, $g^{ri*}_{nk}P^{ri*}_k$, $g^{rj}_{nk}P^{rj}_k$ and $g^{rj*}_{nk}P^{rj*}_k$. The optimization problem is equivalent to solving (52).

$$\min_{P^{r*}_k}(P^{ri*}_k + P^{rj*}_k) \tag{56a}$$

$$P^{ri}_k + P^{rj}_k \leq P^r \tag{56b}$$

$$f^i_{rk} > f^j_{rk} \tag{56c}$$

$$R_{i*} \geq R_{min} \tag{56d}$$

$$R_{j*} \geq R_{min} \tag{56e}$$

where $P^r_k \triangleq [P^{ri}_k, P^{rj}_k]$ 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:

$$\delta R_{i*} = B^k_{ij} log_2\left(\frac{B_{max}N_0 + b_{rn}\alpha_{kr}|h^{ri}_{nk}|^2 P^{ri}_k}{B_{max}N_0 + b_{rn}\alpha_{kn}|h^{ri}_{nk}|^2 P^{ri*}_k}\right) \tag{57}$$

$$\delta R_{j*} = B^k_{ij} log_2\left(\frac{b_{rn}\alpha_{kr}|h^{ri}_{nk}|^2 P^{ri}_k + B_{max}N_0 + b_{rn}\alpha_{kr}|h^{rj}_{nk}|^2 a^j_{rk} P^{rj}_k}{b_{rn}\alpha_{kr}|h^{ri}_{nk}|^2 P^{ri*}_k + B_{max}N_0 + b_{rn}\alpha_{kr}|h^{rj}_{nk}|^2 P^{rj*}_k}\right) \tag{58}$$

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

$$R_{i*} = B^k_{ij} log_2\left[1 + min\left(\frac{b_{rn}\alpha_{kr}g^{ri}_{nk}P^{ri}_k}{b_{rn}\alpha_{kr}g^{ri*}_{nk}P^{ri*}_k + B_{max}N_0}, \frac{b_{rn}\alpha_{kr}g^{rj}_{nk}P^{rj}_k}{b_{rn}\alpha_{kr}g^{rj*}_{nk}P^{rj*}_k + B_{max}N_0}\right)\right] \tag{59}$$

$$R_{j*} = B_{ij}^{k} log_2\left(1 + \frac{b_{rn}\alpha_{kr}\lambda_c P_c^T}{b_{rn}\alpha_{kr}\lambda_c P_{c'}^T + B_{max}N_0}\right) \tag{60}$$

where $\lambda_c P_c^T = P_k^{rj} g_{nk}^{rj} + P_k^{rj*} g_{nk}^{rj*}$ represents the desired signal from user j jointly transmitted 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. (56d) and (56e) can be rewritten using (59) and (60) as follows:

$$\frac{b_{rn}\alpha_{kr} g_{nk}^{ri} P_k^{ri}}{b_{rn}\alpha_{kr} g_{nk}^{ri*} P_k^{ri*} + B_{max}N_0} \geq \gamma_{min}^* \tag{61}$$

$$\frac{b_{rn}\alpha_{kr} g_{nk}^{rj} P_k^{rj}}{b_{rn}\alpha_{kr} g_{nk}^{rj*} P_k^{rj*} + B_{max}N_0} \geq \gamma_{min}^* \tag{62}$$

$$\frac{b_{rn}\alpha_{kn}\lambda_c P_c^T}{b_{rn}\alpha_{kr}\lambda_c P_{c'}^T + B_{max}N_0} \geq \gamma_{min}^* \tag{63}$$

The optimal solutions to problem (56) with given γ_{min}^* are derived as:

$$P_k^{ri} = \frac{(\gamma_{min}^*)^2 P_k^r (g_{nk}^{rj} - \gamma_{min}^* \lambda_c) + B_{max}N_0 g_{nk}^{ri*} \gamma_{min}^*}{g_{nk}^{ri} g_{nk}^{ri*}} \tag{64}$$

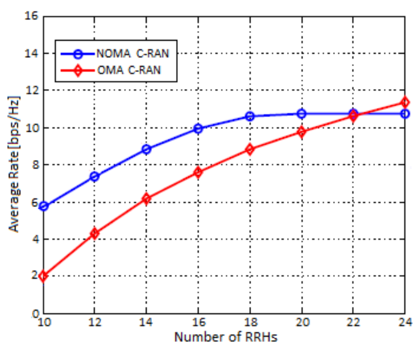
$$P_k^{rj} = \frac{\gamma_{min}^* [B_{max}N_0 + b_{rn}\alpha_{kn}\lambda_c P_k^r}{b_{rn}\alpha_{kr} g_{nk}^{rj}} \tag{65}$$

$$P_k^{ri*} = \frac{\gamma_{min}^* (P_k^r g_{nk}^{rj} - \gamma_{min}^* \lambda_c P_k^r)}{g_{nk}^{ri*}} \tag{66}$$

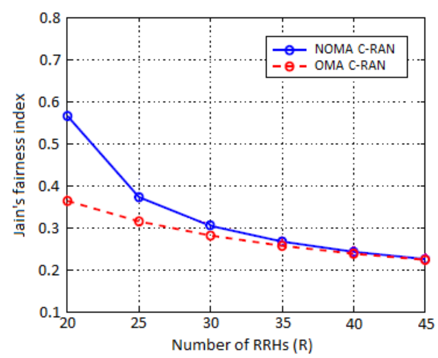
$$P_k^{rj*} = \frac{\gamma_{min}^* \lambda_c P_k^r}{1 + \gamma_{min}^* g_{nk}^{rj*}} \tag{67}$$

VI. 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 covering 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 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. 4(a) compares the average rates for coordinated scheduling scheme for NOMA C-RAN with OMA C-RAN. It can be seen 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 is because RRH coordination contributes to increase in average achieveble rate for both schemes when there are low to medium numbers of RRHs. However the average rates of both schemes have an intersection at some specific threshold. This indicates that NOMA C-RAN 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 users which decreases with increasing number of RRHs. Fig. 4(b) shows the relationship between the Jain's fairness index and number of RRHs



(a) Average rates for NOMA and OMA C-RANs



(b) Jain's fairness index vs number of RRHs

Figure 4: Performance comparison of different transmission schemes in downlink C-RAN networks

for fixed number of RBs. To provide measurement for fairness, Jain's fairness index is used. The fairness index [37] in C-RAN is defined as:

Table II: Simulation parameters

Parameter	Values
Distance dependent path-loss from RRH to RUE	$148.1 + 37.6log_{10}(d)$, 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

$$J_{fi} = \frac{\left[\sum_{n=1}^{N}(R_i + R_j) + \sum_{b=1}^{B} \beta(R_{i'} + R_{j'}) \right]^2}{N_n \left[\left(\sum_{n=1}^{N}(R_i + R_j)^2 + \sum_{b=1}^{B} \beta^2(R_{i'} + R_{j'})^2 \right) \right]} \tag{68}$$

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

$$\beta = \frac{\frac{1}{N}\sum_{n=1}^{N}(R_i + R_j)}{\frac{1}{B}\sum_{b=1}^{B}(R_{i'} + R_{j'})} \tag{69}$$

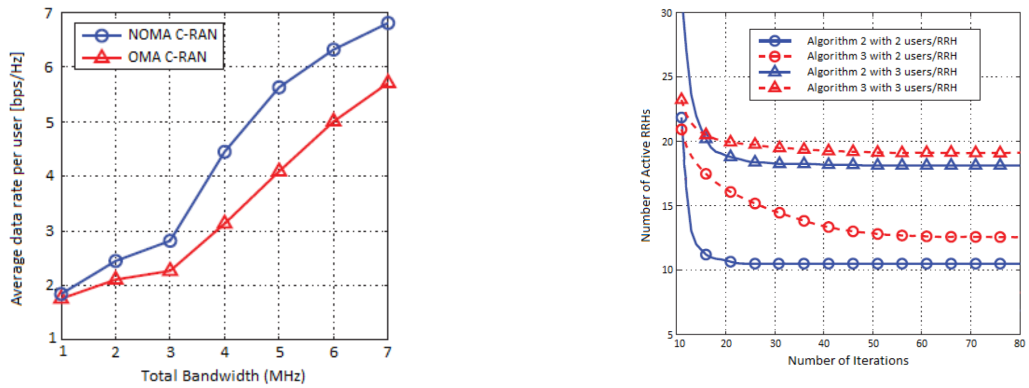
The value of Jain's fairness index is between 0 and 1. The rate allocation is perfectly fair if $J_{fi} = 1$.

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 can be seen that the fairness level is significantly improved with the proposed NOMA-enabled C-RAN compared to OMA C-RAN especially when the numbers of RRHs are in low to medium range.

Fig. 5(a) 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 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 to 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 obtain significant capacity gains of the wireless access links.

We evaluate the effectiveness of the proposed technique in muting RRHs. In order to show the convergence behaviour of the proposed algorithm, Fig.5(b) plots the number of active RRHs in each iteration for different numbers 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 within 35 iterations while the Algorithm 3 requires 55 iterations to converge.

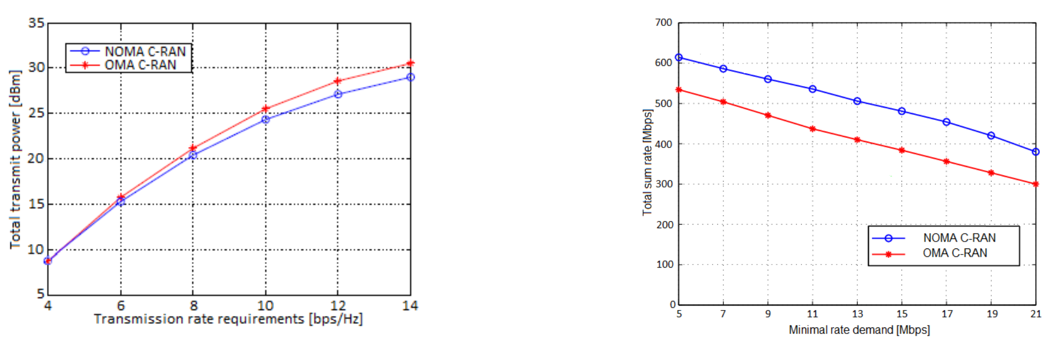
Fig. 6(a) illustrates the transmit power with target data rate for k=6 and n=12. All users have an identical data rate requirements with rates varying from 1 to 14 bps/Hz. It can be



(a) Average data rate versus bandwidth (b) Convergence behaviour of proposed algorithm

Figure 5: Performance comparison and convergence behaviour of the proposed algorithm

seen that the transmit power consumption 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 power consumption (about 4 dB) compared to proposed scheme. Fig. 6(b) shows that the sum rate of NOMA C-RAN outperforms OMA C-RAN, which demonstrates the benefit of NOMA in improving the overall throughput of the system. It can be seen that the sum rate monotonically decreases with the minimum rate requirement for both NOMA C-RAN and OMA C-RAN system. This implies that high rate requirement requires that large transmission power needs to be allocated to users having poor channel conditions, resulting in small transmission power for users having good channel conditions and subsequently low throughput of the overall system.



(a) Transmit power v.s. minimum rate requirements (b) Total sum rate v.s. minimum rate requirements

Figure 6: Performance comparison of transmit power and sum rate vs minimal rate requirements

VII. Conclusions

In this paper 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 optimisation problem in a significantly reduced computational time and showed that NOMA-enabled C-RAN achieves improved network performance in both data rate and network utility with proportional fairness consideration in comparison with the conventional OMA-enabled C-RAN transmission scheme.

APPENDIX A

The Lagrangian to the problem (56) is :

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

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

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

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

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

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

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

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

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

$$\left(\gamma_{min}^* - \frac{b_{rn}\alpha_{kn}g_{nk}^{ri}P_k^{ri}}{b_{rn}\alpha_{kn}g_{nk}^{ri*}P_k^{ri*} + B_{max}N_0}\right) \leq 0 \tag{A.9}$$

$$\left(\gamma_{min}^* - \frac{b_{rn}\alpha_{kn}g_{nk}^{rj}P_k^{rj}}{b_{rn}\alpha_{kn}g_{nk}^{rj*}P_k^{rj*} + B_{max}N_0}\right) \leq 0 \tag{A.10}$$

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

$$\frac{\partial \mathcal{L}}{\partial P_k^{ri}} = \mu_1 - \mu_3 \frac{b_{rn}\alpha_{kn}g_{nk}^{ri}}{b_{rn}\alpha_{kn}g_{nk}^{ri*}P_k^{ri*} + B_{max}N_0} - \lambda_1 = 0 \tag{A.12}$$

$$\frac{\partial \mathcal{L}}{\partial P_k^{ri*}} = \mu_1 - \mu_2 \frac{b_{rn}\alpha_{kn}g_{nk}^{rj}}{b_{ru}\alpha_{kn}\lambda_c P_{c'}^T + B_{max}N_0} - \mu_4 \frac{b_{rn}\alpha_{kn}g_{nk}^{rj}}{b_{rn}\alpha_{kn}g_{nk}^{rj*}P_k^{rj*} + B_{max}N_0} - \lambda_2 = 0 \tag{A.13}$$

$$\frac{\partial \mathcal{L}}{\partial P_k^{rj}} = 1 - \mu_3 \frac{{b_{rn}}^2{\alpha_{kn}}^2(g_{nk}^{ri})^2P_k^{ri}}{(b_{rn}\alpha_{kn}g_{nk}^{ri*}P_k^{ri*} + B_{max}N_0)^2} - \lambda_3 = 0 \tag{A.14}$$

$$\frac{\partial \mathcal{L}}{\partial P_k^{rj*}} = 1 - \mu_2 \frac{b_{rn}\alpha_{kn}g_{nk}^{rj*}}{b_{rn}\alpha_{kn}\lambda_c P_{c'}^T + B_{max}N_0} - \lambda_4 = 0 \tag{A.15}$$

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.

$$P_k^{ri} = \frac{(\gamma_{min}^*)^2 P_k^r(g_{nk}^{rj} - \gamma_{min}^*\lambda_c) + B_{max}N_0 g_{nk}^{ri*}\gamma_{min}^*}{g_{nk}^{ri}g_{nk}^{ri*}} \tag{A.16}$$

$$P_k^{rj} = \frac{\gamma_{min}^*[B_{max}N_0 + b_{rn}\alpha_{kn}\lambda_c P_k^r}{b_{rn}\alpha_{kn}g_{nk}^{rj}} \tag{A.17}$$

$$P_k^{ri*} = \frac{\gamma_{min}^*(P_k^r g_{nk}^{rj} - \gamma_{min}^*\lambda_c P_k^r)}{g_{nk}^{ri*}} \tag{A.18}$$

$$P_k^{rj*} = \frac{\gamma_{min}^*\lambda_c P_k^r}{1 + \gamma_{min}^* g_{nk}^{rj*}} \tag{A.19}$$

(A.16)-(A.19) are the optimal solutions of problem (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 (56). As it can be observed from (A.16) and (A.18) P_k^{ri} and P_k^{ri*} can be negative values. Following are the constraints to make the problem feasible.

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

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

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

APPENDIX B

Initially, the CEUs 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 CEU:

$$P_1^k \leq \frac{\sqrt{1+\frac{d_2^{-\alpha}P^k}{I_i+\sigma^2}}-1}{\frac{d_2^{-\alpha}}{I_i+\sigma^2}} \tag{B.4}$$

Since $\frac{d_2^{-\alpha}}{I_i+\sigma^2} < L_2$ above equation always hold when $P_2^k < \frac{\sqrt{1+P_kL_2}-1}{L_2}$. (B.3) and (B.5) are true simultaneously when following inequality is satisfied:

$$\frac{\sqrt{1+P_kL_1}-1}{L_1} < P_1^k < \frac{\sqrt{1+P_kL_2}-1}{L_2} \tag{B.5}$$

i.e.

$$\frac{\sqrt{1+P_kL_1}-1}{L_1} < \epsilon < \frac{\sqrt{1+P_kL_2}-1}{L_2} \tag{B.6}$$

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

$$D = \left(\frac{1-2\epsilon}{P_b\epsilon^2}\right) \tag{B.7}$$

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:

$$R^k(\epsilon) = w_1R_1^k(\epsilon) + w_2R_2^k(\epsilon) \tag{B.8}$$

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:

$$w_n(t) = \frac{1}{R_n(t-1)}, \ \forall n \in N \tag{B.9}$$

REFERENCES

[1] M. Vaezi, Y. Zhang, " Cloud Mobile Networks: From RAN to EPC", in *Springer*, 2017.

[2] 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.

[3] 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.

[4] 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, 2016.

[5] 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.

[6] 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.

[7] 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.

[8] 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 International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-6.

[9] 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.

[10] 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.

[11] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, Anxin Li, K. Higuchi, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," in *IEEE 77th Vehicular Technology Conference (VTC Spring), 2013*, pp.1, 5, 2-5 June 2013.

[12] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)*, 2013.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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

[18] 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.

[19] 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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," in *IEEE Communications Letters*.

[26] 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.

[27] Z. Ding, M. Peng, and V. Poor, "Cooperative non-orthogonal multiple access in 5g systems," in *IEEE Communications Letters*, 2015.

[28] 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.

[29] 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.

[30] 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

[31] 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.

[32] 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

[33] 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.

[34] 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.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[36] S. Boyd and A. Mutapcic, *Subgradient Methods*. Stanford, CA, USA: Stanford Univ. Press, 2008

[37] 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.

[38] 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.