


RESEARCH ARTICLE

WILEY

Joint uplink and downlink delay-aware resource allocation in C-RAN

Masumehsadat Tohidi¹ | Hamidreza Bakhshi¹  | Saeedeh Parsaeefard²

¹Department of Electrical Engineering,
Shahed University, Tehran, Iran

²Iran Telecommunications Research
Center, Tehran, Iran

Correspondence

Hamidreza Bakhshi, Department of
Electrical Engineering, Shahed University,
Tehran 3319118651, Iran.
Email: bakhshi@shahed.ac.ir

Abstract

This work considers two-way communication between each pair of users with highly delay-aware applications. We formulate a joint uplink and downlink resource allocation problem in a cloud radio access network. Assuming average end-to-end (E2E) delay of each user pair and practical limitation such as maximum transmit power, we maximize the total throughput of all pair of users in the cloud radio access network. In this setup, we consider that each user can be connected to at most one remote radio head and a limited capacity fronthaul link between each remote radio head and baseband unit. To present the resource allocation problem in a more tractable manner, we replace the E2E delay limitation with its equivalent throughput-based formulation. Due to inherent NP-hard and nonconvex nature of the proposed problem, we apply successive convex approximation to reach a two-step iterative algorithm where, in each step, a specific set of optimization variable derived while other variables are fixed. The problem of each step is transformed into the standard geometric programming via the arithmetic-geometric mean approximation. Simulation results reveal that our proposed joint uplink-downlink resource allocation algorithm outperforms a case that uplink and downlink resources are allocated separately in terms of total throughput and outage probability of E2E delay, ie, a chance that E2E delay does not hold.

1 | INTRODUCTION

Next generation of cellular networks should be able to face the increasing number of network users and their requisites for the high quality of service (QoS) such as lower delay requirements.¹ Many applications that are uniquely contemplated for the fifth generation (5G) of cellular networks such as autonomous vehicles, factory automation, tactile internet, remote control, and healthcare have strict requirements end-to-end (E2E) or round trip delay (say around 1 millisecond).^{2,3} Furthermore, for popular multimedia services such as seamless lip-synchronized video conferencing and interactive gaming, providing low E2E delay is essential.⁴ For implementing these applications in 5G, cloud-based architecture is identified.

Cloud-radio access network (C-RAN) is a new radio access network architecture, which can be responded to the increasing demand for traffic requirements with high spectral efficiency, lower energy consumption, and reduced cost.⁵ Typical C-RAN consists of multiple low-cost remote radio heads (RRHs), which are distributed over geographical locations and controlled by a centralized baseband unit (BBU). The RRHs are connected to BBU via fronthaul links.⁶ The fronthaul links usually suffer from the limited capacity issue, leading to a strict constraint on the maximum number of served users over the coverage of interest.⁷ By considering this constraint, an important issue is associating RRHs and subcarriers to users such that the best performance is achieved. Resource allocation problem for C-RAN has been studied in some related works.⁸⁻¹⁰

Due to the importance of delay QoS requirement, a considerable amount of literature has been investigated resource allocation problems to meet this essential requirement in the C-RAN architecture.¹¹⁻¹³ One well-known approach for delay QoS guarantees is to impose average delay constraint.¹⁴ By exploiting the queuing theory, this restriction can be expressed in terms of the minimum transmission rate constraint.^{15,16} This approach was extended for dealing with delay-aware power minimization in C-RAN by considering a double-layer queuing network structure (in each user equipment data processing in BBU and data transmission in RRH).¹⁷ Tang et al.¹⁸ extended this approach for jointly optimizing the virtual machine activation in the BBU pool and sparse beamforming in the coordinated RRH cluster, which is constrained by limited fronthaul capacity, to minimize the system cost of C-RAN. In the same vein for conducting delay constraint, Li et al.¹⁹ investigated the tradeoff between energy efficiency performance and capacity in C-RAN of high-speed railways. The problem of delay-aware downlink beamforming with discrete rate adaptation to minimize the power consumption of C-RANs was considered in the work of Zhi et al.²⁰

In the aforementioned works that consider average delay constraint in C-RAN,¹⁷⁻²⁰ only downlink delay was investigated. Note that for ensuring a bounded average delay requirement for critical technologies beyond 5G (ie, ultra reliable low latency communication), a global control loop for each pair of transmit and receive users must be considered during resource allocation.²¹ Considering E2E delay quality of service leads us to couple both uplink and downlink resource management and dynamically adjust uplink and downlink delay based on channel state information (CSI) for each pair of user, in contrast to a fixed delay requirement for each uplink and downlink session. To the best of our knowledge, there are no works that investigated average E2E delay constraint in joint uplink and downlink resource allocation problems.

To fill this gap, in this work, we focus on the joint uplink and downlink resource allocation problem in orthogonal frequency-division multiple-access (OFDMA) C-RAN, with a constraint on E2E average delay requirement. In summary, the main contributions of the paper are as follows.

1. We first formulate the problem of resource allocation for any pair of delay-aware users, where the E2E delay includes delays of uplink transmission, BBU process, and downlink transmission. This constraint ensures flexibility to the pair of uplink and downlink user to adjust their throughputs, according to the CSI, such that the delay constraint meet. However, it imposes high complexity to the resource allocation problem. By exploiting the concept of queuing theory, we replace delay requirement by a constraint in terms of uplink and downlink throughput, which leads to a more tractable resource allocation problem.
2. We consider joint uplink and downlink resource allocation. Because QoS requirement and available resource of uplink and downlink are coupled, this is important to providing high QoS and efficient resource management since it tracks the performance between pair of transmit and receive users.
3. We consider OFDMA and jointly associate subcarrier and RRH for each uplink and downlink user pair by defining a binary variable as the user association factor (UAF). Due to the UAFs and nonlinearity nature of constraints, the resource allocation problem is nonconvex and NP-hard. To propose an efficient algorithm, we apply successive convex approximation (SCA)²² and diverse techniques of relaxation on convexification such as arithmetic-geometric mean approximation (AGMA).²³ Following a two-step iterative algorithm for UAFs assignment and power allocation, we reshape the problem of each step into geometric programming (GP).²⁴ Therefore, our problem can be solved efficiently by software optimization package, same as CVX.
4. By simulation results, we demonstrate how optimizing joint uplink and downlink transmission parameters can improve the performance of the network in contrast to the traditional approach where the one-way delay was considered and disjointly allocated resources for uplink and downlink session.

The remainder of this paper is organized as follows. In Section 2, the system model for an E2E delay-aware C-RAN is presented, and the resource allocation problem is formulated. The proposed solution for our algorithm is offered in Section 3. In Section 4, we investigate the simulation results, and the conclusions are presented in Section 5.

2 | SYSTEM MODEL AND PROBLEM FORMULATION

Consider a multicell OFDMA C-RAN, consisting of $\mathcal{M} = \{1, 2, \dots, M\}$ single antenna RRHs, connected to the BBU pool through a fiber fronthaul link with limited capacity C_{\max} bps. We study a certain case of applications such as tactile internet, which places E2E delay constraint on the order of milliseconds between a master user and a controller user. The pair of master and controller users are considered as a tactile user pair.²⁵ In this area, there are $\mathcal{K} = \{1, 2, \dots, K\}$ user pairs that tend to communicate to its own peer via a C-RAN environment with the average E2E delay QoS requirement.

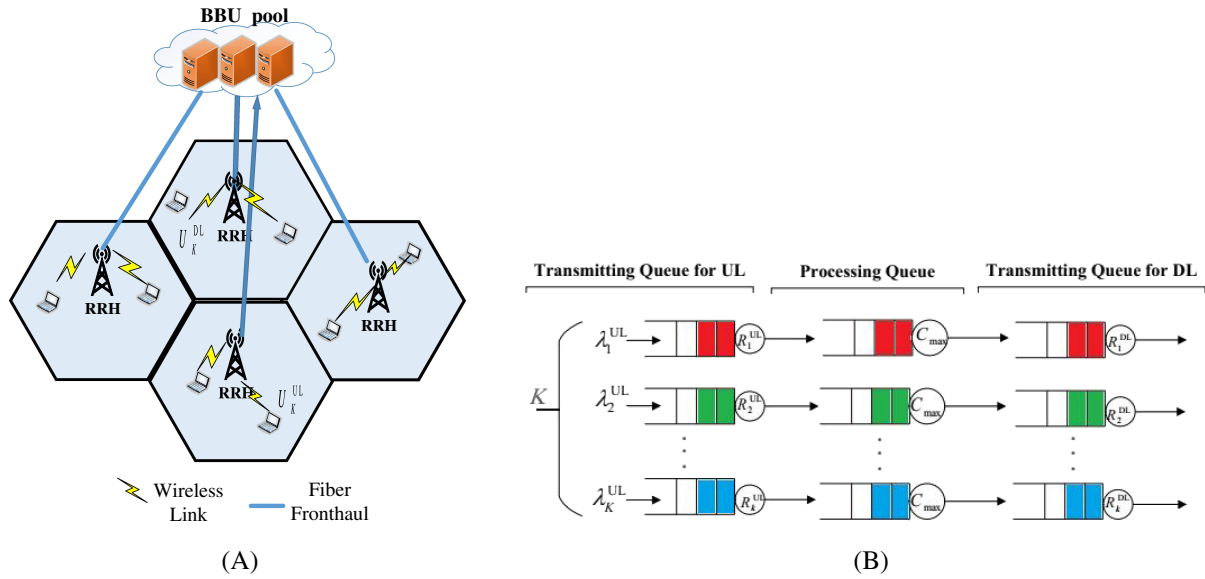


FIGURE 1 A, System model; B, End-to-end queuing model of our system. BBU, baseband unit; DL, downlink; RRH, remote radio head; UL, uplink

In Figure 1A, the overall scheme of our setup is plotted. For instance, as depicted, the data of U_k^{UL} is sent to the closest RRH and then via a fronthaul link to the BBU in uplink session. After processing in BBU, at downlink transmission mode, this information is dispatched to the corresponding RRH and is forwarded to the paired user, ie, U_k^{DL} . In this setup, we consider E2E transmission for each pair of users, including joint uplink and downlink sessions. For the sake of simplicity, we utilize an upper case index $\mathcal{F} = \{UL, DL\}$ to demonstrate parameters of each uplink/downlink sessions.

We assume frequency division duplexing technique where W^{UL} and W^{DL} are two nonoverlapping bands, dedicated for the uplink and downlink, respectively. We further consider OFDMA-based transmission for users in both downlink and uplink sessions where the frequency band of each subcarrier is assumed to be W_S and the set of subcarriers is $\mathcal{N}^f = \{1, \dots, n^f, \dots, N^f\}$, where $N^f = W^f/W_S$. The instantaneous channel power gain $h_{k,n^f,m}$ from user k to RRH m on subcarrier n^f is assumed to be flat and available in BBU pool at each time slot. $P_{k,n^f,m}$ is defined as the transmit power of user k and RRH m on subcarrier n^f . Let σ^2 be the noise power in each subcarrier. The maximum achievable throughput of user k at the RRH m on subcarrier n^f is expressed as

$$R_{k,n^f,m}(\mathbf{P}^f) = \frac{W^f}{N^f} \log_2 \left(1 + \frac{P_{k,n^f,m} h_{k,n^f,m}}{\sigma^2 + I_{k,n^f,m}} \right), \quad \forall k \in \mathcal{K}, \forall n^f \in \mathcal{N}^f, \forall m \in \mathcal{M}, f \in \mathcal{F}, \quad (1)$$

where $\mathbf{P}^f = [P_{k,n^f,m}]$ and $\alpha^f = [\alpha_{k,n^f,m}]$ are the vectors of allocated power and UAF of users, respectively, and

$$I_{k,n^f,m} = \sum_{\substack{m' \in \mathcal{M}, \\ m' \neq m}} \sum_{\substack{k' \in \mathcal{K}, \\ k' \neq k}} p_{k',n^f,m'} h_{k',n^f,m'} \quad (2)$$

is the interference of user k on subcarrier n^f in RRH m . Moreover, $\alpha_{k,n^f,m} = \{0, 1\}$ is defined as the UAF, where

$$\alpha_{k,n^f,m} = \begin{cases} 1, & \text{if the subcarrier } n^f \text{ and RRH } m \text{ is assigned to user } k. \\ 0, & \text{else.} \end{cases} \quad (3)$$

We assume that each user can be connected to one RRH at a time slot; hence,²⁴

$$C1 : \left[\sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m} \right] \left[\sum_{\substack{n^f \in \mathcal{N}^f \\ \forall m' \neq m}} \alpha_{k,n^f,m'} \right] = 0, \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, f \in \{\mathcal{F}\}.$$

C1 ensures that when user k is assigned to RRH m with subcarrier n^f , that user would not be assigned by other RRHs m' with any subcarriers.

Furthermore, because of OFDMA implementation limitation, each subcarrier exclusively can be assigned to at most one user in each cell. Therefore, we have

$$C2 : \sum_{k \in \mathcal{K}} \alpha_{k,n^f,m} \leq 1, \quad \forall m \in \mathcal{M}, \forall n^f \in \mathcal{N}^f, f \in \mathcal{F}.$$

In the uplink and downlink sessions, the maximum total available transmit power for each user and each RRH for every time slot is limited, respectively, which can be formulated as

$$C3 : \sum_{m \in \mathcal{M}} \sum_{n^{\text{UL}} \in \mathcal{N}^{\text{UL}}} \alpha_{k,n^{\text{UL}},m} p_{k,n^{\text{UL}},m} \leq P_k^{\text{max}}, \quad \forall k \in \mathcal{K}.$$

$$C4 : \sum_{k \in \mathcal{K}} \sum_{n^{\text{DL}} \in \mathcal{N}^{\text{DL}}} \alpha_{k,n^{\text{DL}},m} p_{k,n^{\text{DL}},m} \leq P_m^{\text{max}}, \quad \forall m \in \mathcal{M}.$$

Another practical limitation is coming from the limited capacity fronthaul link between RRHs and BBU, where we have

$$C5 : \sum_{k \in \mathcal{K}} \sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m} R_{k,n^f,m}(\mathbf{P}^f) \leq C_{\text{max}}, \quad \forall m \in \mathcal{M}, f \in \mathcal{F}.$$

In general, E2E delay of system consists of four components: propagation delay, processing delay, queueing delay, and transmission delay. Propagation delay depending on the physical distance between the transmitter and receiver.²⁶ Although the distance between RRHs and BBU are usually far, for fixed RRHs, propagation delay between RRH and BBU is constant and does not affect resource allocation. Transmission delay depends on the actual bit rate that transfers in the channel. In C-RAN, data transfer is done in a two-hop case, ie, from user to RRH and then from RRH to BBU for the uplink session and vice versa for downlink. Therefore, we consider a three-step queueing model¹⁷ for each user pair data, ie, a transmission queue in uplink RRH, a processing queue in BBU pool, and a transmission queue in downlink RRH. We assume that each user's information at the BBU and RRH sit in a separate queue; in fact, we have \mathcal{K} parallel queues for the user's data. Figure 1B represents considered queueing model. Arrival data for uplink users is considered to follow a Poisson distribution with a mean rate of λ_k . The service time of each queue data packet in uplink transmission, processing, and downlink transmission queues have exponential distribution with means $\frac{1}{R_k^{\text{UL}}}$, $\frac{1}{C_{\text{max}}}$, and $\frac{1}{R_k^{\text{DL}}}$, respectively. Thus, each queue can be represented as an M/M/I one.¹⁷ Using Little's formula,²⁷ the average waiting time in transmission queue (queueing delay) plus transmission time is derived by the following:

$$D_k^t = \frac{1}{R_k^{\text{UL}} - \lambda_k} + \frac{1}{R_k^{\text{DL}} - \lambda_k}, \quad (4)$$

where $R_k^f > \lambda_k$ is the throughput of user k and defined as

$$R_k^f = \sum_{n^f \in \mathcal{N}^f} \sum_{m \in \mathcal{M}} \alpha_{k,n^f,m} R_{k,n^f,m}(\mathbf{P}^f), \quad \forall k \in \mathcal{K} \text{ and } f \in \mathcal{F}. \quad (5)$$

We suppose that in RRH, only RF functions performed, and other processes are done in BBU. In there, the transmit data of each user stay on a processing queue and then transmitted on the fronthaul link. The expected delay of the processing queue (ie, the expected delay in the BBU pool) is $\frac{1}{C_{\text{max}} - \lambda_k}$, and for our assumption, it is constant.

By considering the variable delay terms, for delay-aware service like tactile internet, for satisfying delay constraint of system, we should have

$$C6 : \frac{1}{R_k^{\text{UL}} - \lambda_k} + \frac{1}{R_k^{\text{DL}} - \lambda_k} \leq D_t^{\text{max}},$$

where D_t^{max} is the maximum tolerable delay of transmission and queueing.

According to (1), the overall throughput of the system defines as

$$R_{\text{total}}(\mathbf{P}^{\text{UL}}, \mathbf{P}^{\text{DL}}, \boldsymbol{\alpha}^{\text{UL}}, \boldsymbol{\alpha}^{\text{DL}}) = \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} \left(\sum_{n^{\text{UL}} \in \mathcal{N}^{\text{UL}}} \alpha_{k,n^{\text{UL}},m} R_{k,n^{\text{UL}},m}(\mathbf{P}^{\text{UL}}) + \sum_{n^{\text{DL}} \in \mathcal{N}^{\text{DL}}} \alpha_{k,n^{\text{DL}},m} R_{k,n^{\text{DL}},m}(\mathbf{P}^{\text{DL}}) \right). \quad (6)$$

In this context, the radio resource allocation problem to be tackled is the maximization of (6) via joint user association and power allocation under the mentioned constraints. Therefore, the optimization problem can be formulated as

$$\begin{aligned} & \max_{\mathbf{P}^{\text{UL}}, \mathbf{P}^{\text{DL}}, \boldsymbol{\alpha}^{\text{UL}}, \boldsymbol{\alpha}^{\text{DL}}} R_{\text{total}}(\mathbf{P}^{\text{UL}}, \mathbf{P}^{\text{DL}}, \boldsymbol{\alpha}^{\text{UL}}, \boldsymbol{\alpha}^{\text{DL}}), \\ & \text{subject to :} \quad \text{C1, C2, C5, C6.} \end{aligned} \quad (7)$$

This problem is a nonconvex and NP-hard optimization due to the binary nature of variable $\boldsymbol{\alpha}^f$ and nonlinearity of C1 and C6. To overcome these challenges, we propose an efficient iterative algorithm, which is explained in the following sections.

3 | PROPOSED TWO-STEP ITERATIVE ALGORITHM

In this section, we provide a detailed solution for resource allocation problem introduced in (7). The constraints C1 and C6 are nonlinear and nonconvex, which yield (7) to be an NP-hard problem. To cope with the computational complexity of (7), we utilize an SCA approach and decompose the problem into two separate subproblems for finding $\boldsymbol{\alpha}^f$ and \mathbf{P}^f as presented in Algorithm 1. We start with an initialized feasible solution for $\boldsymbol{\alpha}^f[0]$ and $\mathbf{P}^f[0]$. In each iteration t , in the first step, for a given power allocation vector, the best UAF ($\boldsymbol{\alpha}^f$) is computed. With this derived UAF, in the second step, power allocation problem is solved. This sequence can be represented as

$$\underbrace{\boldsymbol{\alpha}^f(0) \rightarrow \mathbf{P}^f(0)}_{\text{Initialization}} \rightarrow \underbrace{\dots \boldsymbol{\alpha}^{*f}(t) \rightarrow \mathbf{P}^{*f}(t)}_{\text{Iteration } t} \rightarrow \underbrace{\boldsymbol{\alpha}^{*f} \rightarrow \mathbf{P}^{*f}}_{\text{Optimal solution}}. \quad (8)$$

The solution is improved in each iteration, and the algorithm stops when $\|\boldsymbol{\alpha}^f(t) - \boldsymbol{\alpha}^f(t-1)\| \leq \epsilon_1$ and $\|\mathbf{P}^f(t) - \mathbf{P}^f(t-1)\| \leq \epsilon_2$, where $0 < \epsilon_1, \epsilon_2 \ll 1$. These steps are explained in details as follows.

3.1 | Step 1: User association problem

In step 1, at iteration t for a given $\mathbf{P}^f(t)$, user association problem from (7) is formulated as

$$\begin{aligned} & \max_{\boldsymbol{\alpha}^{\text{UL}}, \boldsymbol{\alpha}^{\text{DL}}} R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \boldsymbol{\alpha}(t_1)^{\text{UL}}, \boldsymbol{\alpha}(t_1)^{\text{DL}}), \\ & \text{subject to :} \quad \text{C1} - \text{C6.} \end{aligned} \quad (9)$$

This problem is also inherently nonconvex due to the integer nature of $\boldsymbol{\alpha}^f$ and nonlinear constraints in C1-C2 and C6. To solve efficiently, first, we relax variables $\alpha_{k,n^{\text{UL}},m}$ and $\alpha_{k,n^{\text{DL}},m}$ to be continuous on the interval $[0, 1]$ and then use AGMA to convert the problem into GP. Although GP in standard form is a nonconvex optimization problem, it can be readily turned into a convex optimization problem.²⁸ In other words, according to SCA,²⁹ the nonconvex optimization problem can be approximated as a convex problem in each iteration. In standard GP formulation, objective function and inequality constraint should be posynomial functions, ie, products of monomial terms and equality constraint should only involve monomial functions. For obtaining the monomial approximation of C1, Proposition 1 is presented.

Proposition 1. At iteration t_1 , by defining

$$x_{k,m}^f(t_1) = \sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m}(t_1), \quad y_k^f(t_1) = \sum_{m \in \mathcal{M}} \sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m}^f(t_1), \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, f \in \{\mathcal{F}\}.$$

C1 can be approximated as following constraints²⁴:

$$\text{C1.1 : } \left(s_{k,m}^f(t_1) \right)^{-1} + x_{k,m}^f(t_1) y_k^f(t_1) \left(s_{k,m}^f(t_1) \right)^{-1} \leq 1, \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, f \in \{\mathcal{F}\},$$

$$\text{C1.2 : } \left[\frac{1}{\theta_{k,m}^f(t_1)} \right]^{-\theta_{k,m}^f(t_1)} s_{k,m}^f(t_1) \left[\frac{\left(s_{k,m}^f(t_1) \right)^2}{\beta_{k,m}^f(t_1)} \right]^{-\beta_{k,m}^f(t_1)} \leq 1, \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, f \in \{\mathcal{F}\},$$

$$\text{C1.3 : } x_{k,m}^f(t_1) \prod_{n^f \in \mathcal{N}^f} \left[\frac{\alpha_{k,n^f,m}(t_1)}{\nu_{k,n^f,m}(t_1)} \right]^{-\nu_{k,n^f,m}(t_1)} = 1, \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, f \in \{\mathcal{F}\},$$

$$\text{C1.4 : } y_k^f(t_1) \prod_{n^f \in \mathcal{N}^f, m \in \mathcal{M}} \left[\frac{\alpha_{k,n^f,m}(t_1)}{\eta_{k,n^f,m}(t_1)} \right]^{-\eta_{k,n^f,m}(t_1)} = 1, \quad \forall k \in \mathcal{K}, f \in \{\mathcal{F}\},$$

where $s_{k,m}^f(t_1)$ is an auxiliary variable and

$$\theta_{k,m}^f(t_1) = \frac{1}{1 + \left(s_{k,m}^f(t_1 - 1) \right)^2}, \quad (10)$$

$$\beta_{k,m}^f(t_1) = \frac{\left(s_{k,m}^f(t_1 - 1) \right)^2}{1 + \left(s_{k,m}^f(t_1 - 1) \right)^2}, \quad (11)$$

$$\nu_{k,n^f,m}(t_1) = \frac{\alpha_{k,n^f,m}(t_1 - 1)}{\sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m}(t_1 - 1)}, \quad (12)$$

$$\eta_{k,n^f,m}(t_1) = \frac{\alpha_{k,n^f,m}(t_1 - 1)}{\sum_{n^f \in \mathcal{N}^f} \sum_{m \in \mathcal{M}} \alpha_{k,n^f,m}(t_1 - 1)}, \quad (13)$$

for all $k \in \mathcal{K}, m \in \mathcal{M}, n^f \in \mathcal{N}^f$, and $f \in \mathcal{F}$.

Proof. See Appendix 1. □

Now, we can replace C1 with C1.1-C1.4, which contain monomial equalities and posynomial inequalities.

Proposition 2. Constraint C6 can be transformed as

$$\text{C}\tilde{6}.1 = \left(R_k^{\text{DL}}(t_1) + R_k^{\text{DL}}(t_1) \right) \left(1 + D_t^{\text{max}} \lambda_k \right) \left(\frac{D_t^{\text{max}} \lambda_k^2 + 2\lambda_k}{\delta_k(t_1)} \right)^{-\delta_k(t_1)} \left(\frac{D_t^{\text{max}} R_k^{\text{UL}}(t_1) R_k^{\text{DL}}(t_1)}{\pi_k(t_1)} \right)^{-\pi_k(t_1)} \leq 1, \quad \forall k \in \mathcal{K},$$

$$\text{C}\tilde{6}.2 : R_k^f(t_1) \prod_{\substack{n^f \in \mathcal{N}^f \\ m \in \mathcal{M}}} \left[\frac{\alpha_{k,n^f,m}(t_1) R_{k,n^f,m}(\mathbf{P}^f)}{\tau_{k,n^f,m}(t_1)} \right]^{\tau_{k,n^f,m}(t_1)} = 1, \quad \forall k \in \mathcal{K}, f \in \mathcal{F},$$

where

$$\tau_{k,n^f,m}(t_1) = \frac{\alpha_{k,n^f,m}(t_1) R_{k,n^f,m}(\mathbf{P}^f)}{\sum_{n^f \in \mathcal{N}^f} \sum_{m \in \mathcal{M}} \alpha_{k,n^f,m}(t_1) R_{k,n^f,m}(\mathbf{P}^f)}, \quad (14)$$

$$\delta_k(t_1) = \frac{D_t^{\text{max}} \lambda_k^2 + 2\lambda_k}{D_t^{\text{max}} R_k^{\text{UL}} R_k^{\text{DL}} + D_t^{\text{max}} \lambda_k^2 + 2\lambda_k}, \quad (15)$$

$$\pi_k(t_1) = \frac{D_t^{\text{max}} R_k^{\text{UL}}(t_1) R_k^{\text{DL}}(t_1)}{D_t^{\text{max}} R_k^{\text{UL}}(t_1) R_k^{\text{DL}}(t_1) + D_t^{\text{max}} \lambda_k^2 + 2\lambda_k}. \quad (16)$$

Proof. See Appendix 2. □

Algorithm 1 Iterative resource allocation algorithm

Initialization: Set $t = 0$, $\mathbf{P}^{\text{UL}}(t = 0) = P_k^{\text{max}}/N^{\text{UL}}$ and $\mathbf{P}^{\text{DL}}(t = 0) = P_m^{\text{max}}/K$.

Repeat: Set $t = t + 1$.

Step1: User Association

Initialization for Step 1: Set $t_1 = 0$, $\alpha^f(t_1) = \alpha^f(t)$, $\mathbf{P}^f(t_1) = \mathbf{P}^f(t)$ and set arbitrary initial for x_0 , $s_{k,m}^f$, $c_0(t_1)$ and $c_{k,n^f,m}(t_1)$.

Repeat: Set $t_1 = t_1 + 1$.

Step 1.1: Update $v_{k,n^f,m}(t_1)$, $\eta_{k,n^f,m}(t_1)$, $\theta_{k,m}^f(t_1)$, $\beta_{k,m}^f(t_1)$, $\tau_{k,n^f,m}(t_1)$, $\pi_k(t_1)$, $\delta_k(t_1)$, $c_0(t_1)$, $c_{k,n^f,m}(t_1)$ using (10)-(16) and (18)-(20).

Step 1.2: Find UAF according to (17) by CVX.

Until: $\|\alpha^f(t) - \alpha^f(t-1)\| \leq \epsilon_1$.

Step 2: Power Allocation

Initialization for Step 2: Set $t_2 = 0$, $\alpha^f(t_2) = \alpha^f(t)$ and set arbitrary initial for v_k^f .

Repeat: Set $t_2 = t_2 + 1$,

Step 2.1: Update $\kappa_{k,n^f,m}(t_2)$, $\chi_{k,n^f,m}(t_2)$, $\vartheta_{k,n^{\text{DL}},m}(t_2)$, $\delta_{k,n^{\text{DL}},m}(t_2)$ using (22)-(25).

Step 2.2: Allocate power according to (26) by CVX.

Until: $\|\mathbf{P}^f(t_2) - \mathbf{P}^f(t_2-1)\| \leq \epsilon_2$.

Until: $\|\alpha^f(t) - \alpha^f(t-1)\| \leq \epsilon_1$ and $\|\mathbf{P}^f(t_2) - \mathbf{P}^f(t_2-1)\| \leq \epsilon_2$.

Proposition 3. *The objective function should be transformed into the monomial function in order to reach the standard GP-based formulation. By defining $\Xi \gg 1$ as a sufficiently large constant and $x_0 > 0$ as an auxiliary variable, we have*

$$\begin{aligned} & \min_{\alpha^{\text{UL}}, \alpha^{\text{DL}}, x_0, s_{k,m}^f, y_k^f, s_{k,m}^f} x_0(t_1), \\ & \text{subject to : C0, C1.1 – C1.4, C2 – C5, C6.1, C6.2,} \end{aligned} \quad (17)$$

where

$$\begin{aligned} \text{C0 : } & \prod_{\substack{k \in \mathcal{K} \\ m \in \mathcal{M}}} \prod_{n^{\text{UL}} \in \mathcal{N}^{\text{UL}}} \left(\frac{\alpha_{k,n^{\text{UL}},m}(t_1) R_{k,n^{\text{UL}},m}(\mathbf{P}^{\text{UL}}(t))}{c_{k,n^{\text{UL}},m}(t_1)} \right)^{-c_{k,n^{\text{UL}},m}(t_1)} \\ & \prod_{n^{\text{DL}} \in \mathcal{N}^{\text{DL}}} \left(\frac{\alpha_{k,n^{\text{DL}},m}(t_1) R_{k,n^{\text{DL}},m}(\mathbf{P}^{\text{DL}}(t))}{c_{k,n^{\text{DL}},m}(t_1)} \right)^{-c_{k,n^{\text{DL}},m}(t_1)} \Xi \left[\frac{x_0(t_1)}{c_0(t_1)} \right]^{-c_0(t_1)} \leq 1 \end{aligned}$$

and

$$c_0(t_1) = \frac{x_0(t_1 - 1)}{x_0(t_1 - 1) + R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1))}, \quad (18)$$

$$c_{k,n^{\text{UL}},m}(t_1) = \frac{\alpha_{k,n^{\text{UL}},m}(t_1) R_{k,n^{\text{UL}},m}(\mathbf{P}(t))}{x_0(t_1 - 1) + R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1))}, \quad (19)$$

$$c_{k,n^{\text{DL}},m}(t_1) = \frac{\alpha_{k,n^{\text{DL}},m}(t_1) R_{k,n^{\text{DL}},m}(\mathbf{P}(t))}{x_0(t_1 - 1) + R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1))}. \quad (20)$$

Proof. See Appendix 3. □

Thus, UAF can be iteratively derived by solving GP approximated problem in (17) with CVX.

3.2 | Step 2: Power allocation

With a fixed UAF, the optimization problem in (7) can be transformed into the following:

$$\begin{aligned} \max_{\mathbf{p}^{\text{UL}}(t_2), \mathbf{p}^{\text{DL}}(t_2)} & \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} \sum_{n^{\text{UL}} \in \mathcal{N}^{\text{UL}}} \alpha_{k,n^{\text{UL}},m}(t) \log_2 \left(1 + \frac{p_{k,n^{\text{UL}},m}(t_2) h_{k,n^{\text{UL}},m}}{\sigma^2 + I_{k,n^{\text{UL}},m}(t_2)} \right) \\ & + \sum_{n^{\text{DL}} \in \mathcal{N}^{\text{DL}}} \alpha_{k,n^{\text{DL}},m}(t) \log_2 \left(1 + \frac{p_{k,n^{\text{DL}},m}(t_2) h_{k,n^{\text{DL}},m}}{\sigma^2 + I_{k,n^{\text{DL}},m}(t_2)} \right), \end{aligned} \quad (21)$$

subject to : C3 – C6,

where t_2 is the power allocation iteration index. This problem is nonconvex due to the C6 and interference term. To simplify this, we consider predefined the threshold for tolerate interference and add a new constraint for the optimization problem as

$$\text{C7 : } \sum_{\substack{m' \in \mathcal{M}, \\ m' \neq m}} \sum_{\substack{k' \in \mathcal{K}, \\ k' \neq k}} p_{k,n^f,m} h_{k,n^f,m} \leq I_{\text{th}}, \quad \forall n^f \in \mathcal{N}^f \text{ and } f \in \mathcal{F},$$

where I_{th} is the maximum tolerable interference level of all users. By substituting the interference term with this threshold, due to the smaller feasible set, the lower bound performance of optimization problem is achieved, but the computational complexity significantly is reduced.^{30,31} Then, we employ AGMA to convert this problem into standard GP. C5 can be represented as polynomial function as

$$\text{C5 : } \prod_{k \in \mathcal{K}, n^f \in \mathcal{N}^f} \left(1 + \frac{h_{k,n^f,m} p_{k,n^f,m}(t_2)}{\sigma^2} \right)^{\frac{w^f}{N^f}} \leq 2^{C_{\text{max}}}, \quad \forall m \in \mathcal{M}, f \in \mathcal{F}.$$

Proposition 4. To handle the non-linear function in C6, we use the approximation $\ln u \approx a(u^{1/a} - 1)$. Afterwards, we deploy AGMA. Consequently, C6 can be replaced by

$$\text{C6.1 : } \left((v_k^{\text{UL}})^{1/a} + (v_k^{\text{DL}})^{1/a} \right) \left(1 + \frac{aD_t^{\text{max}}}{\ln 2} \right)^{-\zeta_k(t_2)} \left(\frac{\frac{aD_t^{\text{max}}}{\ln 2} (v_k^{\text{UL}})^{1/a} (v_k^{\text{DL}})^{1/a}}{\rho_k(t_2)} \right)^{-\rho_k(t_2)} \leq 1, \quad \forall k \in \mathcal{K},$$

$$\text{C6.2 : } \prod_{\substack{n^f \in \mathcal{N}^f \\ m \in \mathcal{M}}} 2^\lambda \sigma^2 v_{k,n,m}^f \left(\frac{\sigma^2}{\kappa_{k,n^f,m}(t_2)} \right)^{-\kappa_{k,n^f,m}(t_2)} \left(\frac{p_{k,n^f,m}(t_2) h_{k,n^f,m}}{\chi_{k,n^f,m}(t_2)} \right)^{-\chi_{k,n^f,m}(t_2)} = 1, \quad \forall k \in \mathcal{K}, f \in \mathcal{F},$$

where $v_{k,n,m}^f$ is an auxiliary variable and

$$\zeta_k(t_2) = \frac{\frac{aD_t^{\text{max}}}{\ln 2} + 2}{\frac{aD_t^{\text{max}}}{\ln 2} (v_k^{\text{UL}})^{1/a} (v_k^{\text{DL}})^{1/a} + \left(\frac{aD_t^{\text{max}}}{\ln 2} + 2 \right)}, \quad (22)$$

$$\rho_k(t_2) = \frac{\frac{aD_t^{\text{max}}}{\ln 2} (v_k^{\text{UL}})^{1/a} (v_k^{\text{DL}})^{1/a}}{\frac{aD_t^{\text{max}}}{\ln 2} (v_k^{\text{UL}})^{1/a} (v_k^{\text{DL}})^{1/a} + \left(\frac{aD_t^{\text{max}}}{\ln 2} + 2 \right)}, \quad (23)$$

$$\kappa_{k,n^f,m}(t_2) = \frac{\sigma^2}{\sigma^2 + h_{k,n^f,m} p_{k,n^f,m}(t_2 - 1)}, \quad f \in \mathcal{F}, \quad (24)$$

$$\chi_{k,n^f,m}(t_2) = \frac{h_{k,n^f,m} p_{k,n^f,m}(t_2 - 1)}{\sigma^2 + h_{k,n^f,m} p_{k,n^f,m}(t_2 - 1)}, \quad f \in \mathcal{F}. \quad (25)$$

Proof. See Appendix 4. □

The objective function in (21) does not belong to the posynomial function. By applying AGMA, we will transform it into (26)

$$\begin{aligned} \min_{\mathbf{P}^{\text{UL}}(t_2), \mathbf{P}^{\text{DL}}(t_2)} \prod_{k \in \mathcal{K}} \prod_{m \in \mathcal{M}} \prod_{n^{\text{UL}} \in \mathcal{N}^{\text{UL}}} \left(\frac{\sigma^2}{\kappa_{k,n^{\text{UL}},m}(t_2)} \right)^{-\kappa_{k,n^{\text{UL}},m}(t_2)} \left(\frac{p_{k,n^{\text{UL}},m}(t_2) h_{k,n^{\text{UL}},m}}{\chi_{k,n^{\text{UL}},m}(t_2)} \right)^{-\chi_{k,n^{\text{UL}},m}(t_2)} \\ \left(\frac{\sigma^2}{\kappa_{k,n^{\text{DL}},m}(t_2)} \right)^{-\kappa_{k,n^{\text{DL}},m}(t_2)} \left(\frac{p_{k,n^{\text{DL}},m}(t_2) h_{k,n^{\text{DL}},m}}{\chi_{k,n^{\text{DL}},m}(t_2)} \right)^{-\chi_{k,n^{\text{DL}},m}(t_2)} \end{aligned} \quad (26)$$

subject to : C3 – C3, C6.1, C6.2, C7.

The power allocation problem in (26) is iteratively solved until the power vector converges, ie, $\|\mathbf{P}^f(t_2) - \mathbf{P}^f(t_2 - 1)\| \leq \epsilon_2$, where $0 < \epsilon_2 \ll 1$.

4 | COMPUTATIONAL COMPLEXITY AND CONVERGENCE ANALYSIS

In this section, we investigate the computational complexity and convergence of Algorithm 1 analytically. Since we apply CVX to solve two subproblems with interior point method in steps 1 and 2, the number of required iteration is $\frac{\log(c/t^0 \rho)}{\log \xi} 32$, where c is the total number of constraints, t^0 is the initial point to approximate the accuracy of interior point method, $0 < \rho \ll 1$ is the stopping criterion for interior point method, and ξ is used for updating the accuracy of interior point method.³² For the user association and power allocation problems, the number of constraints are $C_{\text{UA}} = N^{\text{UL}} + N^{\text{DL}} + 6KM + 6K + M + 1$ and $C_{\text{PA}} = 3M + 4K$, respectively. In steps 1 and 2, for each additional iteration, the number of computations that are required to convert the nonconvex problems using AGMA into the GP approximations are $i_{\text{UA}} = (MK + 3M + 1)(N^{\text{UL}} + N^{\text{DL}})$ and $i_{\text{PA}} = MN^{\text{UL}}(K + 1) + MN^{\text{DL}}(K + 1) + N^{\text{UL}} + N^{\text{DL}}$, respectively. Therefore, the order of computational complexity for each step is

$$\begin{aligned} \text{–For step 1 (User association problem) : } i_{\text{UA}} &\times \frac{\log(c_{\text{UA}}/t_{\text{UA}}^0 \rho_{\text{UA}})}{\log \xi_{\text{UA}}} \\ \text{–For step 2 (Power allocation problem) : } i_{\text{PA}} &\times \frac{\log(c_{\text{PA}}/t_{\text{PA}}^0 \rho_{\text{PA}})}{\log \xi_{\text{PA}}}. \end{aligned} \quad (27)$$

The proposed two-step iterative algorithm in this paper belongs to the block coordinate descent methods where, at each iteration, a single block of variables is optimized, while the remaining variables are fixed. If the subproblems are exactly solved to its unique optimal solution, the convergence of the block coordinate descent method is guaranteed.³³ Moreover, in the work of Razaviyayn et al,³⁴ the general convergence analysis of SCA method is established. In this paper, we utilize AGMA to transform the problem into GP based on the SCA approach. Therefore, the convergence of the proposed algorithm is guaranteed.

5 | SIMULATION RESULTS

In this section, we illustrate the numerical results to evaluate the performance of our proposed algorithm. We consider a C-RAN environment consisting of $K = 6$ pairs of users, $N = 10$ subcarriers for each uplink and downlink sessions, and $M = 4$ RRHs. The RRHs are connected to BBU with fiber fronthaul links and are located at a square area in coordinates: (0.5, 0.5), (0.5, 1.5), (1.5, 0.5), and (1.5, 1.5) wherein users are randomly scattered with uniform distribution. Channel power gains are derived based on the path loss and Rayleigh fading model for both uplink and downlink, ie, $h_{k,n^f,m} = a_{k,n^f,m} d_{k,m}^{-b}$, where $a_{k,n^f,m} \sim \text{Exp}\{1\}$, $b = 3$ is the path loss exponent, and $d_{k,m} > 0$ is normalized distance between user k and RRH m .²⁴ The parameters used for the simulations are listed in Table 1. The simulation results are derived by averaging over 100 channel realizations. When there is no feasible solution for the system, ie, C1-C6 do not hold simultaneously, the total throughput is set to zero.

In our simulation, two scenarios are considered and compared for delay constraint. In scenario I, we seek a joint uplink and downlink resource allocation with E2E delay constraint, presented as C6. In scenario II, we assume an equal but disjoint delay constraint for each uplink and downlink part of one E2E transmission between each pair of users. Such scenario leads to disjoint resource allocation for each pair of users separately in uplink and downlink sessions with two separate delay constraints for uplink and downlink transmission.³⁵

TABLE 1 System simulation parameters

Parameter	Symbol	Value
Number of users	K	6
Number of subcarriers	N	10
Number of remote radio heads	M	4
Path loss exponent	b	3
Mean arrival rate	λ_k	4 kHz
Fronthaul capacity	C_{\max}	100 kbit/sec
Bandwidth of each subcarrier	$\frac{W}{N}$	10 kHz
Accuracy of convergence	ϵ_1, ϵ_2	10^{-3}
Positive large constant	Ξ	10^7
Noise power	σ^2	1 Watt
Maximum tolerable interference level	I_{th}	8 Watt

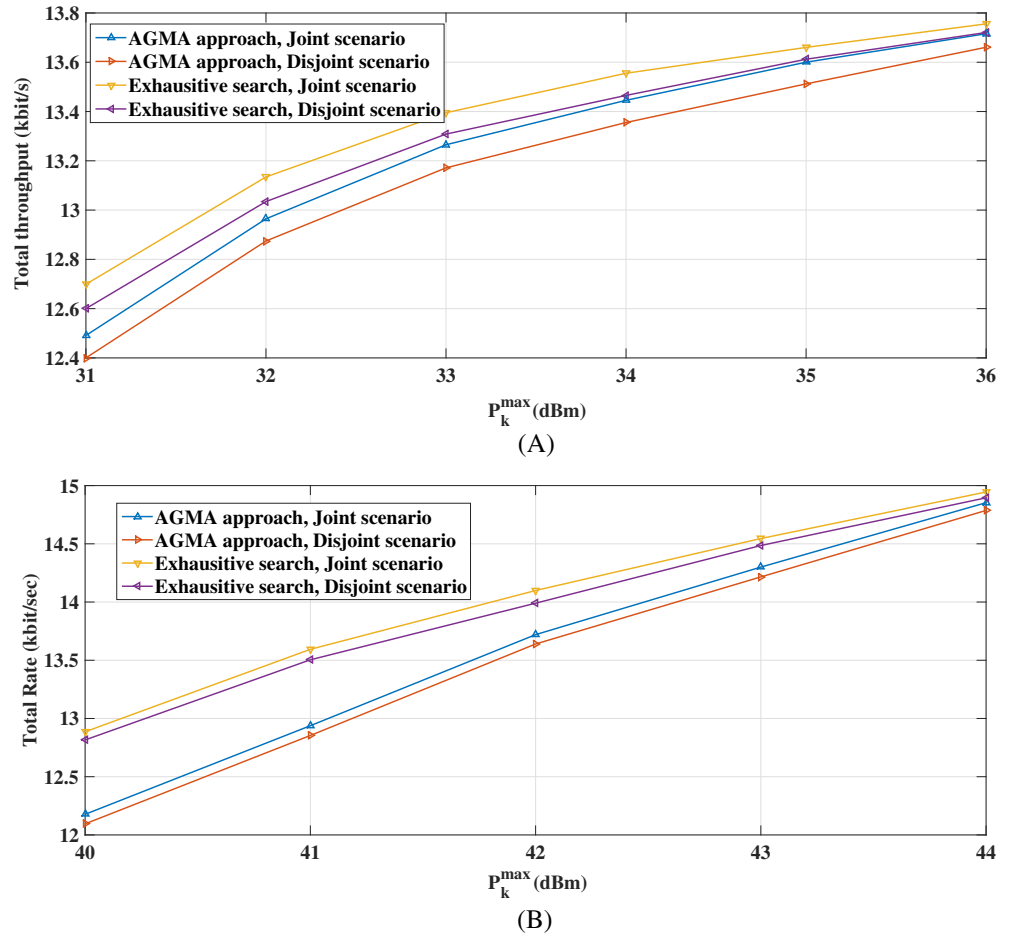


FIGURE 2 A, Total throughput versus maximum transmit power of each user, P_k^{\max} (in dBm); B, Total throughput versus maximum transmit power of each remote radio head, P_m^{\max} (in dBm). AGMA, arithmetic-geometric mean approximation

Figures 2A and 2B demonstrate the total throughputs of uplink and downlink versus maximum transmit power of each user and each RRH, respectively, for $D^{\max} = 2\text{ms}$. Here, we set $P_m^{\max} = 37\text{dBm}$ in Figure 2A and $P_k^{\max} = 33\text{dBm}$ in Figure 2B. For a fixed value of D^{\max} , as we expect, by increasing P_k^{\max} and P_m^{\max} , total uplink and downlink throughputs are growing. On the other hand, for fixed uplink and downlink throughputs, as the delay threshold decreases (the delay is more stringent), the maximum transmit power for each RRH and user increase. More importantly, the proposed joint resource allocation outperforms the disjoint scenario since, in the proposed approach, the delay of uplink and downlink transmission can dynamically be adjusted based on the CSI of each users' pair. Therefore, the total throughputs of users in Figures 2A and 2B are increased for joint scenario compared to the disjoint one. From these figures, it can be observed that by increasing P_m^{\max} and P_k^{\max} , the utility function of Algorithm 1 is approaching the exhaustive search solution. This is due to the fact that for the large SINR scenario, the optimal solution obtained by AGMA approach is the best-fit approximation.

In Figure 3, we investigate the effect of delay threshold D^{\max} on the total throughput of system, for $P_k^{\max} = 33\text{dBm}$ and $P_m^{\max} = 37\text{dBm}$. This figure indicates that the total throughput increases with increasing in D^{\max} . In other words, via

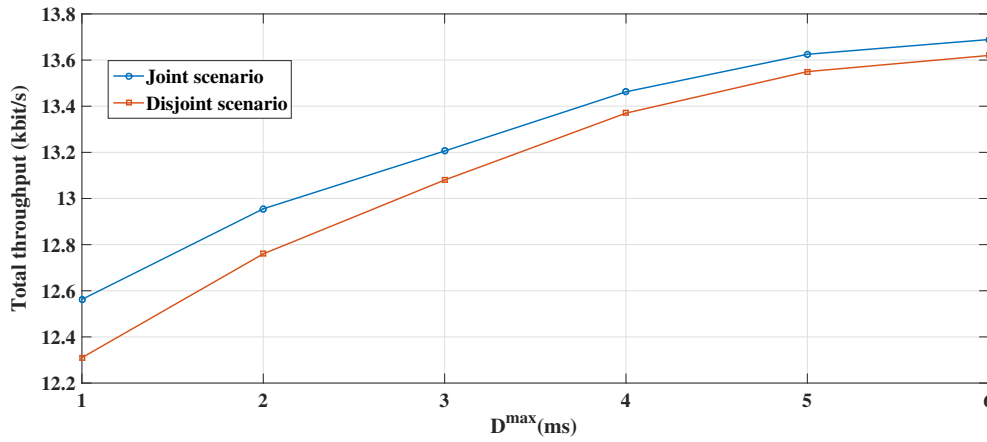


FIGURE 3 Total throughputs versus D^{\max} (ms)

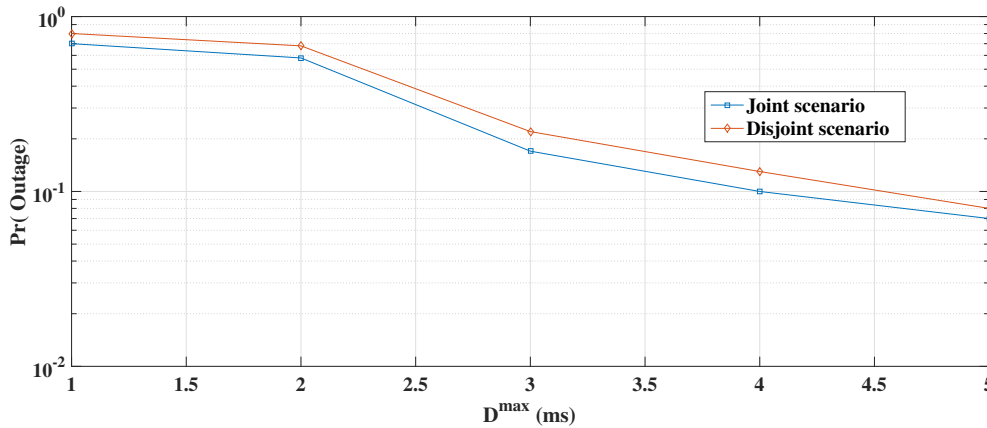


FIGURE 4 Delay-outage probability versus D^{\max} (ms)

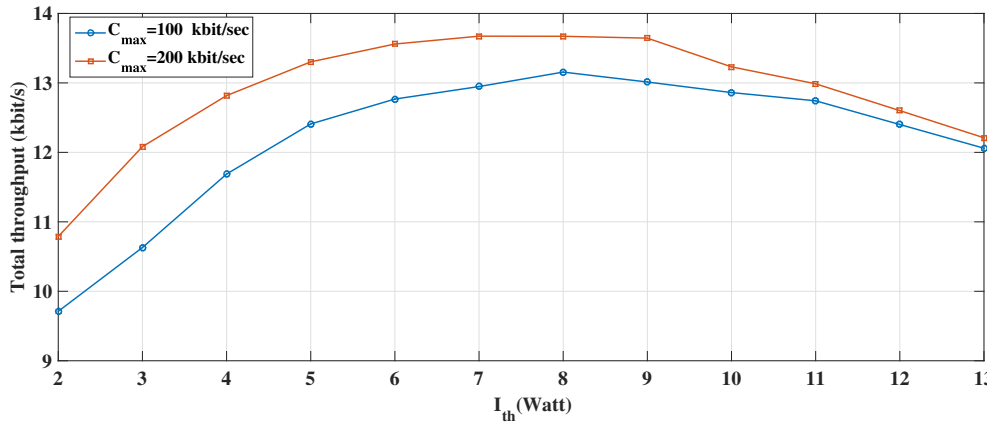


FIGURE 5 Total throughputs versus interference threshold for different I^{th}

decreasing D^{\max} , the feasibility region of our considered problem is shrinking, which leads to less total throughput. In a joint scenario, since there exists flexibility to allocate the resource according to C_6 based on the uplink and downlink channel gains of each pair of users, throughput increases in contrast to the disjoint scenario. This improvement of our proposed approach has more effect in low delay threshold which is related to the ultrareliable low latency communication applications in 5G.

To further study about the chance of feasibility in different delay threshold, we define the delay outage probability as

$$\Pr(\text{delay-outage}) = \Pr\{D_k > D^{\max}\}. \quad (28)$$

Figure 4 depicts $\Pr(\text{delay-outage})$ versus the delay permissive threshold (D^{\max}) via Mont Carlo simulation. As the D^{\max} decreases, ie, more strict constraint on delay, the delay-outage probability increases, which imply that the feasibility region of resource allocation in (7) is shrinking. From Figure 4, it is interesting to note that our proposed joint uplink and downlink

resource allocation has better capability to match the delay constraint. Therefore, there is a higher chance to have a feasible solution for (7) compared with the disjoint scenario, ie, when C6.1 and C6.2 are considered.

We evaluate the performance of the proposed joint uplink and downlink resource allocation problem under different value of interference level threshold I_{th} and different fronthaul capacity C_{max} . We fix $D_{max} = 2$ ms, $P_k^{max} = 37$ dBm, and $P_m^{max} = 37$ dBm. Figure 5 demonstrates that total throughput of system increases with I_{th} and then decreases when I_{th} is larger than 8. This effect occurs because when I_{th} is small, few channels can be reused among different users to satisfy the interference constraint. As I_{th} increases, frequency reuse among users can be admitted, which increases the system throughput. In practice, I_{th} can be selected based on the simulation results under different network settings.

6 | CONCLUSIONS

In this paper, we investigate joint uplink and downlink resource allocation problem under E2E average delay constraint in C-RAN. Taking into account the queuing model of E2E transmission mode for each user pair as an M/M/I queue, we derive a tractable formulation to transform average delay requirement into equivalent uplink and downlink throughput constraints for each user. Via this approach, our cross-layer resource allocation problem is transformed into the case that the proposed resource allocation problem only involves physical-layer parameters. However, the derived formulation still suffers from high computational complexity due to its inherent nonconvex nature. We employ a threshold-based policy to limit aggregate interference and simplify the problem. We also introduce a two-step iterative algorithm based on SCA for joint user association and power allocation where, in each step, by applying AGMA, the corresponding problem is converted into a geometric one and can be solved via CVX. Simulation results reveal that considering E2E delay and joint resource allocation outperforms the disjoint scenario where the delay of uplink and downlink transmissions for each pair of users are fixed, and this proposed approach increases the total throughput and the feasibility probability of optimization problem considerably.

ORCID

Hamidreza Bakhshi  <https://orcid.org/0000-0001-6758-0431>

REFERENCES

- Andrews J, Buzzi S, Choi W. What will 5G be? *IEEE J Sel Areas Commun*. 2014;32(6):1065-1082.
- Parvez I, Rahmati A, Guvenc I, Sarwat H. A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Commun Surv Tutor*. 2018;20(4):3098-3130.
- She C, Yang C, Quek T. Radio resource management for ultra-reliable and low-latency communications. *IEEE Commun Mag*. 2017;55(6):72-78.
- Ren H, Liu N, Pan C, et al. Low-latency C-RAN: a next-generation wireless approach. *IEEE Veh Technol Mag*. 2018;13(2):48-56.
- Wu J, Zhang Z, Hong Y, Wen Y. Cloud radio access network (C-RAN): a primer. *IEEE Netw*. 2015;29:35-41.
- Wang X, Huang Y, Cui C, Chen K, Chen M. C-RAN: The road towards green RAN. *China Commun*. 2010;7(3):107-112.
- Peng M, Wang C, Lau V, Poor H. Fronthaul-constrained cloud radio access networks: insights and challenges. *IEEE Wirel Commun*. 2015;22(2):125-160.
- Liu L, Bi S, Zhang R. Joint power control and fronthaul rate allocation for throughput maximization in OFDMA-based cloud radio access network. *IEEE Trans Commun*. 2015;63(11):4097-4110.
- Lyazidi M, Aitsaadi N, Langar R. Resource allocation and admission control in OFDMA-based Cloud-RAN. Paper presented at: 2016 IEEE Global Communications Conference; 2016; Washington, DC.
- Lin Z, Liu Y. Joint uplink-downlink resource allocation in OFDMA cloud radio access networks. Paper presented at: IEEE International Conference on Communications (ICC); 2018; Kansas, MO.
- Cui Y, Lau V, Wang R, Huang H, Zhang S. A survey on delay-aware resource control for wireless systems-large deviation theory, stochastic Lyapunov drift, and distributed stochastic learning. *IEEE Wirel Commun Lett*. 2012;58(3):1677-1701.
- Chen M, Mozaffari M, Saad W, Yin C, Debbah M, Seon Hong C. Caching in the sky: proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience. *IEEE J Sel Areas Commun*. 2017;3(5):1046-1061.
- Chen M, Saad W, Yin C, Debbah M. Echo state networks for proactive caching in cloud-based radio access networks with mobile users. *IEEE Trans Wirel Commun*. 2017;16(2):3520-3535.
- Collins B, Cruz R. Transmission policies for time varying channels with average delay constraints. Paper presented at: Allerton Conference on Communication, Control, and Computing; 1999; Monticello, IL.
- Sh D, Hui W, Lau V, Lam W. Cross-layer design for OFDMA wireless systems with heterogeneous delay requirements. *IEEE Trans Wirel Commun*. 2007;6(8):2872-2880.

16. Zarakovitis C, Ni Q, Skordoulis D, Hadjinicolaou M. Power-efficient cross-layer design for OFDMA systems with heterogeneous QoS, imperfect CSI, and outage considerations. *IEEE Trans Veh Tech.* 2012;61(2):781-796.
17. Tang J, Peng W, Quek T. Cross-layer resource allocation with elastic service scaling in cloud radio access network. *IEEE Trans Wirel Commun.* 2015;14(9):5068-5081.
18. Tang J, Peng W, Quek T, Liang B. System cost minimization in Cloud RAN with limited fronthaul capacity. *IEEE Trans Wirel Commun.* 2017;16(5):3371-3384.
19. Li S, Zhu G, Lin S, et al. Energy efficiency and capacity tradeoff in cloud radio access network of high-speed railways. *Mobile Inf Syst.* 2017;2017:1-12.
20. Zhi Y, Ke W, Hong J. Delay-aware downlink beamforming with discrete rate adaptation for green cloud radio access network. *J China Univ Posts Telecommun.* 2017;24(1):26-34.
21. Aijaz A. Towards 5G-enabled tactile internet: Radio resource allocation for haptic communications. Paper presented at: IEEE Wireless Communications and Networking Conference; 2016; Doha, Qatar.
22. Wang T, Vandendorpe L. Iterative resource allocation for maximizing weighted sum min-rate in downlink cellular OFDMA systems. *IEEE Trans Signal Process.* 2011;59(1):223-234.
23. Ngo D, Khakurel S, Le-ngoc T. Joint subchannel assignment and power allocation for OFDMA femtocell networks. *IEEE Trans Wirel Commun.* 2014;13(1):125-160.
24. Parsaeefard S, Dawadiy R, Derakhshaniz M, Le-Ngoc T. Joint user-association and resource-allocation in virtualized wireless networks. *IEEE Access.* 2016;4:2738-2750.
25. Simsek M, Aijaz A, Dohler M, Sachs J, Fettweis G. 5G-enabled tactile internet. *IEEE J Sel Areas Commun.* 2016;34(3):460-473.
26. Parvez I, Rahmati A, Guvenc I, Sarwat A, Huaiyu D. A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Commun Surv Tutor.* 2018;20(4):3098-3130.
27. Bertsekas D, Gallager R. Delay models in data networks. In: *Data Networks*. 2nd ed.. Englewood Cliffs, NJ: Prentice-Hall; 1992.
28. Chiang M. Geometric programming for communication systems. *Found Trends Commun Inf Theory.* 2005;2(1-2):1-154.
29. Marks B, Wright G. A general inner approximation algorithm for nonconvex mathematical programs. *Oper Res.* 1987;26(4):681-683.
30. Abdelnasser A, Hossain E, Kim A. Tier-aware resource allocation in OFDMA, macrocell-small cell networks. *IEEE Trans Commun.* 2015;63(3):695-710.
31. Ng D, Lo E, Schober R. Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas. *IEEE Trans Wirel Commun.* 2015;11(2):3292-3303.
32. Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge, UK: Cambridge University Press; 2009.
33. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optim Theory Appl.* 2001;109:475-494.
34. Razaviyayn M, Hong M, Lue Z. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J Optim.* 2013;23(2):1126-1153.
35. Amani N, Pedram H, Taheri H, Parsaeefard S. Energy-efficient resource allocation in heterogeneous cloud radio access networks via BBU-offloading. *IEEE Trans Veh Tech.* 2019;68(2):1365-1377.

How to cite this article: Tohidi M, Bakhshi H, Parsaeefard S. Joint uplink and downlink delay-aware resource allocation in C-RAN. *Trans Emerging Tel Tech.* 2019;e3778. <https://doi.org/10.1002/ett.3778>

APPENDIX

A.1 | Proof of Proposition 1

From the definitions of $x_{k,m}^f(t_1)$ and $y_k^f(t_1)$, C1 can be rewritten as

$$x_{k,m}^f(t_1) \left[y_k^f(t_1) - x_{k,m}^f(t_1) \right] = 0 \quad \forall k \in \mathcal{K}, m \in \mathcal{M}, f \in \mathcal{F}. \quad (\text{A1})$$

By adding 1 to both sides of equality and defining $s_{k,m}^f \geq 0$ as an auxiliary variable, $\forall k \in \mathcal{K}, m \in \mathcal{M}$, we have,

$$x_{k,m}^f(t_1)y_k^f(t_1) + 1 = \left(x_{k,m}^f(t_1) \right)^2 + 1, \quad \forall k \in \mathcal{K}, m \in \mathcal{M}, f \in \mathcal{F}, \quad (\text{A2})$$

$$x_{k,m}^f(t_1)y_k^f(t_1) + 1 \leq s_{k,m}^f(t_1) \leq \left(x_{k,m}^f(t_1)\right)^2 + 1, \quad \forall k \in \mathcal{K}, m \in \mathcal{M}, f \in \mathcal{F}. \quad (\text{A3})$$

The above inequalities can be rewritten as the two following inequalities:

$$\frac{x_{k,m}^f(t_1)y_k^f(t_1) + 1}{s_{k,m}^f(t_1)} \leq 1, \quad \frac{s_{k,m}^f(t_1)}{\left(x_{k,m}^f(t_1)\right)^2 + 1} \leq 1 \quad \forall k \in \mathcal{K}, m \in \mathcal{M}, f \in \mathcal{F}. \quad (\text{A4})$$

Via AGMA, $\forall k \in \mathcal{K}, m \in \mathcal{M}, f \in \mathcal{F}$, (A4) can be rewritten as

$$\text{C1.1} : \left(s_{k,m}^f(t_1)\right)^{-1} + x_{k,m}^f(t_1)y_k^f(t_1)\left(s_{k,m}^f(t_1)\right)^{-1} \leq 1,$$

$$\text{C1.2} : \left[\frac{1}{\theta_{k,m}^f(t_1)}\right]^{-\theta_{k,m}^f(t_1)} s_{k,m}^f(t_1) \left[\frac{(s_{k,m}^f(t_1))^2}{\beta_{k,m}^f(t_1)}\right]^{-\beta_{k,m}^f(t_1)} \leq 1,$$

$$\text{C1.3} : x_{k,m}^f(t_1) = \sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m}(t_1),$$

$$\text{C1.4} : y_k^f(t_1) = \sum_{m \in \mathcal{M}} \sum_{n^f \in \mathcal{N}^f} \alpha_{k,n^f,m}(t_1),$$

where $\theta_{k,m}^f(t_1)$, $\beta_{k,m}^f(t_1)$ are defined in (10) and (11).

Moreover, from AGMA, C1.3 and C1.4 can be converted into monomial functions as

$$\text{C1.3} : x_{k,m}^f(t_1) \prod_{n^f \in \mathcal{N}^f} \left[\frac{\alpha_{k,n^f,m}(t_1)}{v_{k,n^f,m}(t_1)}\right]^{-v_{k,n^f,m}(t_1)} = 1,$$

$$\text{C1.4} : y_k^f(t_1) \prod_{n^f \in \mathcal{N}^f, m \in \mathcal{M}} \left[\frac{\alpha_{k,n^f,m}(t_1)}{\eta_{k,n^f,m}}\right]^{-\eta_{k,n^f,m}(t_1)} = 1.$$

$v_{k,n^f,m}(t_1)$ and $\eta_{k,n^f,m}(t_1)$ are determined in (12) and (13).

A.2 | Proof of Proposition 2

We can rewrite C6 as

$$\text{C6} : \frac{R_k^{\text{DL}}(t_1) - \lambda_k + R_k^{\text{UL}}(t_1) - \lambda_k}{(R_k^{\text{UL}}(t_1) - \lambda_k)(R_k^{\text{DL}}(t_1) - \lambda_k)} \leq D_t^{\text{max}}$$

or equivalently

$$\text{C6.1} : \frac{(R_k^{\text{DL}} + R_k^{\text{UL}})(1 + D_t^{\text{max}}\lambda_k)}{D_t^{\text{max}}R_k^{\text{UL}}R_k^{\text{DL}} + D_t^{\text{max}}\lambda_k^2 + 2\lambda_k} \leq 1.$$

By applying AGMA to the denominator of previous relation, we have C6.1, which is a posynomial function of R_k^f and

$$\text{C6.2} : R_k^f(t_1) = \sum_{n^f \in \mathcal{N}^f} \sum_{m \in \mathcal{M}} \alpha_{k,n^f,m}(t_1) R_{k,n^f,m}(\mathbf{P}^f). \quad (\text{A5})$$

To convert C6.2 to a monomial function, we again deploy AGMA and reach to C6.2.

A.3 | Proof of Proposition 3

In the standard GP-based optimization problems, the objective is minimizing a positive posynomial function. Therefore, to reach the standard format, we first represent the problem in a minimization form as

$$\min_{\alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1)} -R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1)). \quad (\text{A6})$$

To reform the objective as a positive function, we add $\Xi \gg 1$ and rewrite it as

$$\min_{\alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1)} \Xi - R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1)), \quad (\text{A7})$$

which is always positive. By considering an auxiliary variable x_0 and rewriting the objective function, we have

$$\frac{\Xi}{x_0 + R_{\text{total}}(\mathbf{P}^{\text{UL}}(t), \mathbf{P}^{\text{DL}}(t), \alpha^{\text{UL}}(t_1), \alpha^{\text{DL}}(t_1))} \leq 1. \quad (\text{A8})$$

By using AGMA to reform the objective function as a monomial function, the total optimization problem is reformulated as (17).

A.4 | Proof of Proposition 4

Constraint C6 can be rewritten as following in terms of $P_{k,n^{\text{UL}},m}$ and $P_{k,n^{\text{DL}},m}$:

$$\text{C6} : \frac{1}{\ln \prod_{\substack{n^{\text{UL}} \in \mathcal{N}^{\text{UL}} \\ m \in \mathcal{M}}} \frac{1 + \frac{P_{k,n^{\text{UL}},m} h_{k,n^{\text{UL}},m}}{\sigma^2}}{2^{\lambda_k}}} + \frac{1}{\ln \prod_{\substack{n^{\text{DL}} \in \mathcal{N}^{\text{DL}} \\ m \in \mathcal{M}}} \frac{1 + \frac{P_{k,n^{\text{DL}},m} h_{k,n^{\text{DL}},m}}{\sigma^2}}{2^{\lambda_k}}} \leq \frac{D_t^{\max}}{\ln 2}.$$

By considering the approximation $\ln u \approx a(u^{1/a} - 1)$, we have

$$\text{C}\check{6}.1 : \frac{1}{(v_k^{\text{UL}}(t_2))^{1/a} - 1} + \frac{1}{(v_k^{\text{DL}}(t_2))^{1/a} - 1} \leq \frac{aD_t^{\max}}{\ln 2},$$

where v_k^f is defined as

$$\text{C}\check{6}.2 : v_k^f(t_2) = \prod_{\substack{n^f \in \mathcal{N}^f \\ m \in \mathcal{M}}} \frac{1 + \frac{P_{k,n^f,m} h_{k,n^f,m}}{\sigma^2}}{2^{\lambda_k}}.$$

C6.1 can be also reformulated as

$$\text{C}\check{6}.1 : \frac{(v_k^{\text{UL}}(t_2) + v_k^{\text{DL}}(t_2)) \left(1 + \frac{aD_t^{\max}}{\ln 2}\right)}{\frac{aD_t^{\max}}{\ln 2} (v_k^{\text{UL}}(t_2))^{1/a} (v_k^{\text{DL}}(t_2))^{1/a} + \left(\frac{aD_t^{\max}}{\ln 2} + 2\right)} \leq 1.$$

The above constraints can be reformed to GP standard format by applying AGMA as

$$\text{C}\check{6}.1 : \left((v_k^{\text{UL}}(t_2))^{1/a} + (v_k^{\text{DL}}(t_2))^{1/a} \right) \left(1 + \frac{aD_t^{\max}}{\ln 2} \right) \left(\frac{\frac{aD_t^{\max}}{\ln 2} + 2}{-\zeta_k(t_2)} \right)^{-\zeta_k(t_2)} \left(\frac{\frac{aD_t^{\max}}{\ln 2} (v_k^{\text{UL}}(t_2))^{1/a} (v_k^{\text{DL}}(t_2))^{1/a}}{\rho_k(t_2)} \right)^{-\rho_k(t_2)} \leq 1,$$

$$\text{C}\check{6}.2 : \prod_{\substack{n^f \in \mathcal{N}^f \\ m \in \mathcal{M}}} 2^{\lambda} \sigma^2 v_k^f(t_2) \left(\frac{\sigma^2}{\kappa_{k,n^f,m}(t_2)} \right)^{-\kappa_{k,n^f,m}(t_2)} \left(\frac{p_{k,n^f,m}(t_2) h_{k,n^f,m}}{\chi_{k,n^f,m}(t_2)} \right)^{-\chi_{k,n^f,m}(t_2)} = 1,$$

where ζ_k , ρ_k , $\kappa_{k,n^f,m}$, and $\chi_{k,n^f,m}$ are defined in (22)- (25).