Joint Caching and Radio Resource Allocation for the Downlink of Multi-Cell OFDMA Systems

Sepehr Rezvani and Nader Mokari

***Abstract*—The unprecedented growth of internet contents, specially social media, makes a challenge to the load of cellular networks. Latency is one of the most important metrics at end-users. To this end, we propose a resource allocation (RA) algorithm to design both caching and delivery policies with the aim of minimizing total latency of mobile users (MUs), where in the caching phase, the content placement is investigated and in the delivery phase, we allocate radio resources (i.e., transmit powers and subcarriers) in a multi-cell orthogonal frequency division multiple access (OFDMA)-based network communicating with a data center via backhaul links. In order to achieve an efficient caching policy, we propose an optimization problem to minimize the latency of MUs subject to maximum delivery deadline, maximum allowable transmit power of each base station (BS) and data center, and exclusive subcarrier assignment constraints. Hence, we devise an iterative algorithm to solve the main optimization problem and prove that the proposed approach converges to a near-optimal solution, when the number of iterations increases. Moreover, simulation results illustrate that devising the transmission-aware caching policy can significantly improve the performance, compared to the conventional proactive caching policies which are only based on the popularity of contents and the storage capacity of BSs.**

Index Terms– Caching policy, delivery policy, OFDMA, resource allocation, latency.

I. Introduction

In the past decade, with introducing smart phones, tablets and other advanced mobile devices, wireless cellular networks have been developed. Moreover, the global data traffic is predicted to increase 11-fold from 2013 to 2018 and reach 15.9 exabytes per month. Consequently, increasing the mobile data traffic creates challenges for cellular network operators (CNOs) [1].

Social media has already been exceeded 50 percent of the global mobile data traffic in content transportation of internet in 2012 [2]. In order to satisfy the tremendous growing mobile data traffic demand, developers have to increase the cellular network capacity and backhaul bandwidth, simultaneously. Moreover, developing small cells is a potential solution to achieve the Shannon limit in long term evolution (LTE) standard [3]. In LTE systems, it needs to be considered more control signals like transmit power control and interference alignment, where significant amount of available bandwidth is considered by the mentioned control signals [4]. Latency is one of the most important metrics at end-users. According to the confined available radio spectrum and more expensive bandwidth of backhaul links, increasing wireless capacity is not a sufficient solution to improve the latency [5].

Orthogonal frequency division multiplexing (OFDM) is a modulation technique which is utilized in many communication networks, such as LTE cellular networks to increase the data rate of mobile users (MUs). In OFDM, the available bandwidth is divided into some smaller bandwidths, called subcarriers [6]. For more flexibility in resource allocation (RA), orthogonal frequency division multiple access (OFDMA) technique is developed in which separated sets of subcarriers are assigned to different MUs [6], [7].

In the current LTE system, social media should travel from data centers to base stations (BSs) and finally to the end-users. Travelling data over the network takes a time called latency. Since other solution for reducing latency is closing data to the end-users, we should not solely concentrate our attentions to radio resources (i.e., subcarriers and transmit powers), but also to the storages (i.e., caching) in cellular networks [3]. The storage capacity of BSs are other resources which can be simply considered in LTE system to improve latency [3]. Since the latency at MUs mainly comes from traveling contents through access and backhaul links, decreasing the data traffic of radio access and backhaul links, simultaneously, is an effective approach to reduce latency [3].

A. Related Works

Recently, some joint transmit power and subcarrier allocation algorithms have been designed for the downlink of cellular OFDMA-based networks [8]–[11]. A joint subcarrier and transmit power allocation algorithm for the downlink of an OFDMA mixed femtocell/macrocell network is considered in [12]. Moreover, several fractional programming algorithms are proposed to solve the joint transmit power and subcarrier allocation in the downlink of OFDMA systems [13]–[15]

Content caching in cellular wireless networks is classified into two main categories. The first category is considering a scenario, where each MU is assigned to at most one storage capacity. The second category is considering a multi-coverage scenario in which each MU is covered by several caches. In the second scenario, the authors study on the optimal cache placement with a larger distributed cache. In [16], the authors investigate the design of a distributed caching policy for the downlink of the heterogeneous cellular networks (HCNs) based on the network channel conditions. Hence, they designed distributed caching optimization algorithms via the

Manuscript received June 20, 2016 ; accepted September 20, 2017.

S. Rezvani is with the Electrical and Computer Engineering Department, Tarbiat Modares University, Tehran, Iran (e-mail: s.rezvani@modares.ac.ir).

N. Mokari (corresponding author) is with the Electrical and Computer Engineering Department, Tarbiat Modares University, Tehran, Iran (e-mail: nader.mokari@modares.ac.ir).

belief propagation (BP) algorithm for minimizing the latency. To alleviate the complexity of the proposed BP algorithm, the authors proposed a heuristic BP algorithm and evaluate the average latency of the network. They also show that designing an efficient content placement algorithm is more better than storing popular contents everywhere [17]. Moreover, in [18], the authors propose an in-network cache assisted eNodeB caching scheme for LTE networks, where content objects are cached at eNodeBs with the aim of minimizing latency. They also showed that utilizing the caching technique in the networks can save bandwidth for mobile operators and reduce the latency for end-users. In [3], the authors study on a single-cell scenario in LTE system with considering device-to-device (D2D) communication and design a caching policy based on the achieved fixed data rates at MUs, popularity of contents, storage capacity of BSs and MUs, and maximum delivery deadline at each MU. They show that finding the optimal content placement based on the network conditions improves the latency, in contrast to the conventional caching policies which are only based on the storage capacity of caches and popularity of requesting contents.

B. Our Contributions

Our contributions are presented as follows:

- We consider a cache-enabled OFDMA multi-cell network in which BSs are able to store the social media in the network. We focus on the downlink of the wireless backhaul and access links.
- In the considered model, BSs serve the requests of MUs. Moreover, in each cell, MUs request the social media only from the associated BS, and the requests of MUs could be served in two cases based on the existence of the social media in the cache of the BSs[1]. In order to reduce the latency of MUs, we can decrease the data traffic of both the access and backhaul links and increase the data rate of MUs, simultaneously. In contrast to previous works, this paper considers a new viewpoint in which both the caching and delivery policies are jointly devised.
- In this system, we minimize the total latency of MUs subject to the maximum transmit power, delivery deadline and storage capacity constraints.
- We propose an iterative algorithm to solve the proposed optimization problem. Specifically, in each iteration, we divide the main optimization problem into two subproblems, where the first subproblem is content placement problem and the second subproblem is joint transmit power and subcarrier allocation problem. We solve the binary linear programming (BLP) content placement problem via the available optimization softwares, such as CVX with the internal solver MOSEK. The joint transmit power and subcarrier allocation problem can be classified as general linear fractional programming (GLFP) problem. Accordingly, we use the fractional programming algorithm to solve it. We repeat these iterations until more improvement is not made. We mathematically prove that the proposed iterative algorithm converges to a near-optimal solution.
- In simulation results, we investigate the the performance of the proposed algorithm.

[1]In this paper, we consider a set of social media which are frequently requested. However, the proposed algorithm can be applied for all types of contents.

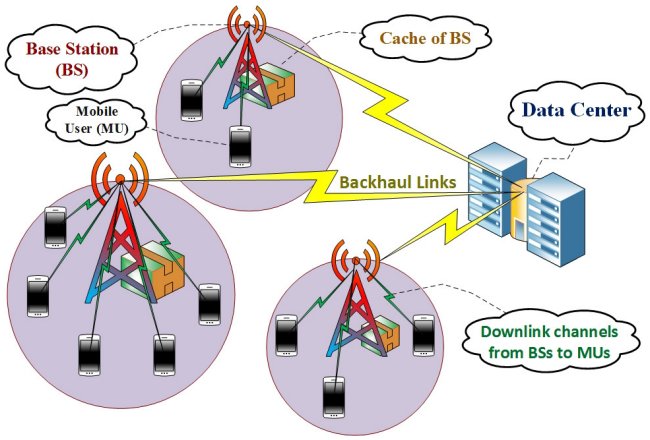


Fig. 1: The network model of the considered system.

C. Paper Organization

The rest of this paper is organised as follows: We describe the network model in Section II. The main optimization problem is presented in Section III. Moreover, Section IV solves the proposed optimization problem. The computational complexity of the proposed algorithm is presented in Section V. The numerical results and simulations are presented in Section VI. We also present the conclusion of this paper in Section VII.

II. System Model

Consider an OFDMA-based multi-cell radio access network (RAN) consisting of B BSs and U MUs, communicates with a data center via wireless backhaul links. We assume that in each cell, there is only one BS which is servicing the set of MUs located in that cell. We focus on the downlink of the network which is shown in Fig. 1.

A. Network Model

In this model, we have B cells which are not spanning together. Denoted by $\mathcal{B} = \{1, 2, \ldots, B\}$, the set of BSs, where $b \in \mathcal{B}$ represents the b^{th} BS in the network. We also denote the data center by 0. The set of MUs in the network is expressed by $\mathcal{U} = \bigcup_{b=1}^{B} \mathcal{U}_b$, where $\mathcal{U}_b$ represents the set of MUs located in cell b and subsequently are associated with BS b. We also indicate the number of MUs located in cell b by $U_b = |\mathcal{U}_b|$. Hence, the total number of MUs in the network is obtained by $U = \sum_{b=1}^{B} U_b$. The set of social media is denoted by $\mathcal{C} = \{1, 2, \ldots, C\}$, where $c \in \mathcal{C}$ represents the c^{th} social media in the network. Moreover, the size of social media c is denoted by s_c which is assumed to be Log-normal distributed, i.e., $s_c \sim \ln\mathcal{N}(\mu_s, \sigma_s^2)$ [3]. In the considered model, we assume that each MU only requests the social media

from the associated BS, and then, the BS serves the requests. The popularity of requesting the c^{th} social media, i.e., rank c, is non-homogeneous in a period of time. Moreover, we characterize it by the Zipf distribution as follows: [3], [19], [20]

$$p_c = \frac{1/c^{\zeta_1}}{\sum_{c=1}^{C} 1/c_{\zeta_1}}, \forall c \in \mathcal{C}, \tag{1}$$

where $0 \leq \zeta_1 \leq 1$ is a real number which tunes the considered popularity model of the social media. Furthermore, we note that the term MU is used instead of 'active MU' for brevity, which means each MU requests at least one social media in the network. We also define a binary indicator $\delta_{b,u,c} \in \{0,1\}$ for requesting the social media c from the MU $u \in \mathcal{U}_b$ as:

$$\delta_{b,u,c} = \begin{cases} 1, & \text{if MU } u \in \mathcal{U}_b \text{ requests social media } c \\ & \text{from BS } b; \\ 0, & \text{otherwise.} \end{cases}$$

According to (1), $\frac{\sum_{b=1}^{B}\sum_{u\in\mathcal{U}_b}\delta_{b,u,c}}{U} \approx p_c, \forall c \in \mathcal{C}$ should be satisfied [3]. We assume that each BS has a storage capacity and is able to cache social media in the network. The set of cache size of the BSs is denoted by $\mathcal{M} = \{m_1, \ldots, m_B\}$, where m_b represents the cache size of the b^{th} BS in the network. Note that the data center has all social media.

In this model, we assume that all channels are composed of Rayleigh fading, path loss model and additive white Gaussian noise (AWGN) with the power spectral density (PSD) of N_0. We also assume a fixed available radio spectrum with the non-spanning frequency bandwidth W_{BH} and W_{Ac} for wireless backhaul and access links, respectively. We divide the bandwidth of wireless backhaul and access links into N_{BH} and N_{Ac} orthogonal subcarriers, respectively, such that each subcarrier has W_s frequency bandwidth which is much less than the available coherence bandwidth of the network. Hence, the transmitted signal over each subcarrier may experience flat fading [10]. Denoted by $\mathcal{N}_{\text{BH}} = \{1, 2, \ldots, N_{\text{BH}}\}$ the set of subcarriers of backhaul links, where $n_{\text{BH}} \in \mathcal{N}_{\text{BH}}$ represents the n^{th} subcarrier of the backhaul channels in the network. Moreover, the set of subcarriers of the access links is denoted by $\mathcal{N}_{\text{Ac}} = \{1, 2, \ldots, N_{\text{Ac}}\}$, where $n_{\text{Ac}} \in \mathcal{N}_{\text{Ac}}$ represents the n^{th} subcarrier of the access channels in the network.

B. Serving Model and Latency Formulation

In this paper, we assume that each social media can be stored at the cache of BSs. Hence, we introduce the following binary content placement indicator as:

$$\rho_{b,c} = \begin{cases} 1, & \text{if social media } c \text{ is cached at BS } b; \\ 0, & \text{otherwise.} \end{cases}$$

The serving request model of the considered network is described in the following. In each time duration, BSs receive the requests of MUs located in their cells. If the requested social media is stored in the cache of the BS, the BS disseminates it via access link, immediately. If the social media is not stored at the BS, the BS receives the requested social media from the data center via the backhaul link and then, sends it to the requested MU via access link. According to the above, when MU $u \in \mathcal{U}_b$ requests social media c, the requested social media will be served based on the following cases:

1. ***First case***: If the requested social media c is stored at BS b, i.e., $\rho_{b,c} = 1$, the requested social media c will be served by BS b via the access link between BS b and MU $u \in \mathcal{U}_b$. Moreover, we obtain the data traffic of the access link of MU $u \in \mathcal{U}_b$ as [3]:

$$f_{b,u}^{\text{Ac},1} = \sum_{c=1}^{C} \delta_{b,u,c}\rho_{b,c}s_c. \tag{2}$$

In addition, we assume a limitation for the latency of each MU $u \in \mathcal{U}_b$ as $\tau_{b,u}^{\text{Ac}}$, which means that the latency of each MU $u \in \mathcal{U}_b$ should not exceed $\tau_{b,u}^{\text{Ac}}$. Hence, we should consider a sufficient data rate for the access link of MUs to serve all requests. Moreover, the channel capacity of the access link for MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac} is given by [21]

$$r_{b,u}^{\text{Ac},1,n_{\text{Ac}}} = \log_2(1 + \frac{p_{b,u}^{\text{Ac},1,n_{\text{Ac}}} h_{b,u}^{n_{\text{Ac}}}}{\sigma^2}), \tag{3}$$

where $p_{b,u}^{\text{Ac},1,n_{\text{Ac}}}$ is the transmit power of BS b to MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac} and $h_{b,u}^{n_{\text{Ac}}}$ is the channel power gain of it. We also note that σ^2 is the received AWGN noise power at MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac}. For simplicity, we do not consider any inter-cell interference in our model[2]. The data rate of the access link for MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac} is obtained by

$$w_{b,u}^{\text{Ac},1,n_{\text{Ac}}} = W_s r_{b,u}^{\text{Ac},1,n_{\text{Ac}}}. \tag{4}$$

Moreover, we introduce the following binary subcarrier assignment indicator for the access link of MU $u \in \mathcal{U}_b$ as:

$$\gamma_{b,u}^{\text{Ac},1,n_{\text{Ac}}} = \begin{cases} 1, & \text{if subcarrier } n_{\text{Ac}} \text{ is assigned to the access link} \\ & \text{of MU } u \in \mathcal{U}_b \text{ for the } \textit{first case}; \\ 0, & \text{otherwise.} \end{cases}$$

The data rate of the access link for MU $u \in \mathcal{U}_b$ is formulated as follows:

$$w_{b,u}^{\text{Ac},1,\text{tot}} = \sum_{n_{\text{Ac}}=1}^{N_{\text{Ac}}} \gamma_{b,u}^{\text{Ac},1,n_{\text{Ac}}} w_{b,u}^{\text{Ac},1,n_{\text{Ac}}}. \tag{5}$$

To obtain the latencies, we first formulate the latency of the data packet transmission for the access link of MU $u \in \mathcal{U}_b$ as:

$$t_{b,u}^{\text{Ac},1} = \frac{1}{w_{b,u}^{\text{Ac},1,\text{tot}}}. \tag{6}$$

Hence, the access latency of MU $u \in \mathcal{U}_b$ is given by

$$D_{b,u}^{\text{Ac},1} = f_{b,u}^{\text{Ac},1} t_{b,u}^{\text{Ac},1}. \tag{7}$$

2. ***Second case***: If the requested social media c is not stored at BS b, i.e., $\rho_{b,c} = 0$, the data center disseminates social media c to BS b via backhaul link b and then, BS b serves social media c via the access link between BS b and MU $u \in \mathcal{U}_b$. At first, we note that BS b receives all social media disseminated from the data center, simultaneously, and then,

[2]The inter-cell interference management will be considered as a future work.

the BS serves the received social media from the data center to the MUs located in cell b. Hence, the transmission of social media in the *second case* takes occur in two hops: In the *first hop*, the data center disseminates the total social media which is requested from all MUs located in cell b and is not stored at BS b. We can formulate the data traffic of the backhaul link b as:

$$f_{0,b}^{\text{BH}} = \sum_{c=1}^{C} \sum_{u \in \mathcal{U}_b} \frac{\delta_{b,u,c}(1-\rho_{b,c})s_c}{\max\{\sum\limits_{u \in \mathcal{U}_b} \delta_{b,u,c}, 1\}}. \tag{8}$$

Note that for all requests of social media c from MUs in $\mathcal{U}_b$, if social media c is not stored at BS b, the requested social media c is disseminated to BS b from the data center just once. According to the above, we assume a limitation for the latency of each backhaul link b as $\tau_{0,b}^{\text{BH}}$, which means the latency of each backhaul link b should not exceed $\tau_{0,b}^{\text{BH}}$. Therefore, we should consider a sufficient data rate for the backhaul links to serve all requests. Hence, the channel capacity of backhaul link b on subcarrier n_{BH} is given by

$$r_{0,b}^{\text{BH},n_{\text{BH}}} = \log_2(1 + \frac{p_{0,b}^{n_{\text{BH}}} h_{0,b}^{n_{\text{BH}}}}{\sigma^2}), \tag{9}$$

where $p_{0,b}^{n_{\text{BH}}}$ is the transmit power of the data center to BS b on subcarrier n_{BH} and $h_{0,b}^{n_{\text{BH}}}$ is the channel power gain of it. The data rate of the backhaul link b on subcarrier n_{BH} is obtained as:

$$w_{0,b}^{\text{BH},n_{\text{BH}}} = W_s r_{0,b}^{\text{BH},n_{\text{BH}}}. \tag{10}$$

We introduce the following binary subcarrier assignment indicator for backhaul link b as $\gamma_{0,b}^{n_{\text{BH}}}$, where if subcarrier n_{BH} is assigned to backhaul link b, $\gamma_{0,b}^{n_{\text{BH}}} = 1$, and otherwise $\gamma_{0,b}^{n_{\text{BH}}} = 0$. Hence, we formulate the data rate of the backhaul link b as:

$$\gamma_{0,b}^{n_{\text{BH}}} = \begin{cases} 1, & \text{if subcarrier } n_{\text{BH}} \text{ is assigned to} \\ & \text{backhaul link } b; \\ 0, & \text{otherwise.} \end{cases}$$

The latency of the data packet transmission for backhaul link b is given by

$$t_{0,b}^{\text{BH}} = \frac{1}{w_{0,b}^{\text{BH,tot}}}. \tag{11}$$

According to the above, we formulate the latency of the *first hop* of BS b as follows:

$$D_{0,b}^{\text{BH}} = f_{0,b}^{\text{BH}} t_{0,b}^{\text{BH}}. \tag{12}$$

When BS b receives the total social media disseminated from the data center, sends them to MUs located in cell b in the *second hop* via access links which are considered for the *second case*. In the *second case*, the data traffic of the access link for MU $u \in \mathcal{U}_b$ can be obtained by

$$f_{b,u}^{\text{Ac},2} = \sum_{c=1}^{C} \delta_{b,u,c}(1-\rho_{b,c})s_c. \tag{13}$$

According to the latency threshold $\tau_{b,u}^{\text{Ac}}$, we should consider a sufficient data rate for the access links in the *second case*. We note that a fraction of the access subcarriers $\mathcal{N}_{\text{Ac}}$ belongs to the access links of the *first case* and the remaining fraction of it, is assigned to access links of the *second case*. Hence, we introduce a binary subcarrier assignment indicator for the access link of MU $u \in \mathcal{U}_b$ in the *second case* as:

$$\gamma_{b,u}^{\text{Ac},2,n_{\text{Ac}}} = \begin{cases} 1, & \text{if subcarrier } n_{\text{Ac}} \text{ is assigned to the access link} \\ & \text{of MU } u \in \mathcal{U}_b \text{ for the } \textit{second case}; \\ 0, & \text{otherwise.} \end{cases}$$

The channel capacity of access link for MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac} for the *second case* is given by

$$r_{b,u}^{\text{Ac},2,n_{\text{Ac}}} = \log_2(1 + \frac{p_{b,u}^{\text{Ac},2,n_{\text{Ac}}} h_{b,u}^{n_{\text{Ac}}}}{\sigma^2}), \tag{14}$$

where $p_{b,u}^{\text{Ac},2,n_{\text{Ac}}}$ is the transmit power of BS b to MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac} for the *second case*. Accordingly, the data rate of MU $u \in \mathcal{U}_b$ on subcarrier n_{Ac} in the *second case* is given by

$$w_{b,u}^{\text{Ac},2,n_{\text{Ac}}} = W_s r_{b,u}^{\text{Ac},2,n_{\text{Ac}}}. \tag{15}$$

Therefore, the data rate of MU $u \in \mathcal{U}_b$ in the *second case* is given by

$$w_{b,u}^{\text{Ac},2,\text{tot}} = \sum_{n_{\text{Ac}}=1}^{N_{\text{Ac}}} \gamma_{b,u}^{\text{Ac},2,n_{\text{Ac}}} w_{b,u}^{\text{Ac},2,n_{\text{Ac}}}, \tag{16}$$

and subsequently, the latency of the data packet for the access links in the *second case* for each MU $u \in \mathcal{U}_b$ can be formulated as follows:

$$t_{b,u}^{\text{Ac},2} = \frac{1}{w_{b,u}^{\text{Ac},2,\text{tot}}}. \tag{17}$$

Furthermore, the latency of the *second hop* for MU $u \in \mathcal{U}_b$ which means the access latency of MU $u \in \mathcal{U}_b$ in the *second case*, can be formulated as:

$$D_{b,u}^{\text{Ac},2} = f_{b,u}^{\text{Ac},2} t_{b,u}^{\text{Ac},2}. \tag{18}$$

As shown in Fig. 1, in the *second case*, each social media should travel through the backhaul link (yellow line) and also the access link (green line) to achieve the requested MU. Besides, in the *first case*, each social media only be transferred through the access link between MU and BS, i.e., green line.

III. The Main Optimization Problem

In this paper, our goal is to determine how to allocate the radio resources and store social media in the cache of the BSs. In doing so, we formulate an optimization problem with the aim of minimizing total latencies in the network as follows:

$$D_{\text{tot}} = \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} \left(D_{b,u}^{\text{Ac},1} + D_{b,u}^{\text{Ac},2}\right) + \sum_{b=1}^{B} D_{0,b}^{\text{BH}}, \tag{19}$$

subject to the delivery deadline, storage capacity and transmit power constraints. Hence, we formulate an optimization problem as follows:

$$\min_{\boldsymbol{p},\boldsymbol{\gamma},\boldsymbol{\rho}} \quad D_{\text{tot}} \tag{20a}$$

$$\text{s.t.} \quad D_{0,b}^{\text{BH}} \leq \tau_{0,b}^{\text{BH}}, \forall b \in \mathcal{B}, \tag{20b}$$

$$D_{b,u}^{\text{Ac},1} \leq \tau_{b,u}^{\text{Ac}}, \forall b \in \mathcal{B}, u \in \mathcal{U}_b, \quad (20c)$$

$$D_{b,u}^{\text{Ac},2} \leq \tau_{b,u}^{\text{Ac}}, \forall b \in \mathcal{B}, u \in \mathcal{U}_b, \quad (20d)$$

$$\sum_{u \in \mathcal{U}_b} \sum_{n_{\text{Ac}}=1}^{N_{\text{Ac}}} \gamma_{b,u}^{\text{Ac},1,n_{\text{Ac}}} p_{b,u}^{\text{Ac},1,n_{\text{Ac}}} + \gamma_{b,u}^{\text{Ac},2,n_{\text{Ac}}} p_{b,u}^{\text{Ac},2,n_{\text{Ac}}} \leq P_b^{\max}, \forall b \in \mathcal{B}, \quad (20e)$$

$$\sum_{b=1}^{B} \sum_{n_{\text{BH}}=1}^{N_{\text{BH}}} \gamma_{0,b}^{n_{\text{BH}}} p_{0,b}^{n_{\text{BH}}} \leq P_0^{\max}, \quad (20f)$$

$$\sum_{c=1}^{C} \rho_{b,c} s_c \leq m_b, \forall b \in \mathcal{B}, \quad (20g)$$

$$\sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} (\gamma_{b,u}^{\text{Ac},1,n_{\text{Ac}}} + \gamma_{b,u}^{\text{Ac},2,n_{\text{Ac}}}) \leq 1, \forall n_{\text{Ac}} \in \mathcal{N}_{\text{Ac}}, \quad (20h)$$

$$\sum_{b=1}^{B} \gamma_{0,b}^{n_{\text{BH}}} \leq 1, \forall n_{\text{BH}} \in \mathcal{N}_{\text{BH}}, \quad (20i)$$

$$\rho_{b,c} \in \{0,1\}, \forall b \in \mathcal{B}, \forall c \in \mathcal{C}, \quad (20j)$$

$$\gamma_{0,b}^{n_{\text{BH}}} \in \{0,1\}, \forall b \in \mathcal{B}, \forall n_{\text{BH}} \in \mathcal{N}_{\text{BH}}, \quad (20k)$$

$$\gamma_{b,u}^{\text{Ac},1,n_{\text{Ac}}}, \gamma_{b,u}^{\text{Ac},2,n_{\text{Ac}}} \in \{0,1\}, \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_b, \forall n_{\text{Ac}} \in \mathcal{N}_{\text{Ac}}, \quad (20l)$$

$$p_{0,b}^{n_{\text{BH}}} \geq 0, \forall b \in \mathcal{B}, \forall n_{\text{BH}} \in \mathcal{N}_{\text{BH}}, \quad (20m)$$

$$p_{b,u}^{\text{Ac},1,n_{\text{Ac}}}, p_{b,u}^{\text{Ac},2,n_{\text{Ac}}} \geq 0, \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_b, \forall n_{\text{Ac}} \in \mathcal{N}_{\text{Ac}}, \quad (20n)$$

where $\boldsymbol{p} = [\boldsymbol{p}^{\text{Ac}}, \boldsymbol{p}^{\text{BH}}]$, $\boldsymbol{p}^{\text{Ac}} = [\boldsymbol{p}^{\text{Ac},1}, \boldsymbol{p}^{\text{Ac},2}]$, $\boldsymbol{p}^{\text{Ac},1} = [p_{b,u}^{\text{Ac},1,n_{\text{Ac}}}]$, $\boldsymbol{p}^{\text{Ac},2} = [p_{b,u}^{\text{Ac},2,n_{\text{Ac}}}]$, $\boldsymbol{p}^{\text{BH}} = [p_{0,b}^{n_{\text{BH}}}]$, $\boldsymbol{\gamma} = [\boldsymbol{\gamma}^{\text{Ac}}, \boldsymbol{\gamma}^{\text{BH}}]$, $\boldsymbol{\gamma}^{\text{Ac}} = [\gamma_{b,u}^{\text{Ac},1,n_{\text{Ac}}}, \gamma_{b,u}^{\text{Ac},2,n_{\text{Ac}}}]$, $\boldsymbol{\gamma}^{\text{BH}} = [\gamma_{0,b}^{n_{\text{BH}}}]$ and $\boldsymbol{\rho} = [\rho_{b,c}]$. Constraint (20b) represents that the latency of each backhaul link should not exceed the considered latency limitation of that backhaul link. Moreover, the latency limitation of each MU in RAN is mentioned in (20c) and (20d) for the *first case* and the *second case*, respectively. The transmit power limitation for each BS and the data center are presented in (20e) and (20f), where $P_b^{\max}$ and $P_0^{\max}$ represent the maximum transmit power of BS b and the data center, respectively. The limitation of the cache size of each BS is mentioned in constraint (20g). Moreover, constraints (20h) and (20l) guarantee the OFDMA assumption for the access links, and (20i) and (20k) guarantee the OFDMA assumption for the backhaul links.

IV. Solution

The optimization problem (20) is a mixed-integer nonlinear programming (MINLP) problem. Hence, finding an optimal solution for (20) is very difficult. To this end, we propose an iterative algorithm to solve it. In the proposed algorithm, the main optimization problem (20) is divided into two sub-problems as: 1) content placement problem; 2) joint transmit power and subcarrier allocation problem. We repeat these iterations until the proposed algorithm converges to a sub-optimal solution.

A. The Proposed Iterative Algorithm

Algorithm 1 The proposed iterative algorithm

1: Initialize $\boldsymbol{p}_0, \boldsymbol{\gamma}_0, \boldsymbol{\rho}_0$ according to the Subsection IV-A2.
repeat
2: **for** $t_1 = 1$ to T_1 **do**
3: Find $\boldsymbol{\rho}_{t_1}$ by solving (26).
4: For a fixed $\boldsymbol{\rho}_{t_1}$, find $\boldsymbol{p}_{t_1}$ and $\boldsymbol{\gamma}_{t_1}$ by solving (27).
5: **Until** $||\boldsymbol{p}_{t_1} - \boldsymbol{p}_{t_1-1}|| \leq \omega_1$ or $t_1 = T_1$.
6: Set $t_1 = t_1 + 1$
7: **end for**
8: $\boldsymbol{p}_{t_1}$, $\boldsymbol{\gamma}_{t_1}$ and $\boldsymbol{\rho}_{t_1}$ are adapted for the network.

1) Algorithm overview: The proposed iterative algorithm is summarized in Alg. 1. At first, we initialize the content placement indicator $\boldsymbol{\rho}_0$, transmit power $\boldsymbol{p}_0$ and subcarrier assignment indicator $\boldsymbol{\gamma}_0$ to feasible values. In each iteration t_1, we first find $\boldsymbol{\rho}_{t_1}$ for a fixed $(\boldsymbol{p}_{t_1-1}, \boldsymbol{\gamma}_{t_1-1})$ and then, give it to the joint transmit power and subcarrier allocation problem (27) for a fixed $\boldsymbol{\rho} = \boldsymbol{\rho}_{t_1}$ and find $(\boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1})$. After that, we investigate the stopping criterion which is presented in Alg. 1. If $||\boldsymbol{p}_{t_1} - \boldsymbol{p}_{t_1-1}|| \leq \omega_1$ or the iteration number exceeds a predefined value T_1, the last output is considered for our system as a sub-optimal solution. The function $||.||$ represents the vector Euclidean norm.

Proposition 1: In the proposed iterative Alg. 1, the objective function (20a) is improved or remains constant, in each iteration. Hence, the proposed Alg. 1 converges to a sub-optimal solution.

Proof. Please refer to Appendix A. □

2) Initialization method: To find a feasible solution $(\boldsymbol{p}, \boldsymbol{\gamma}, \boldsymbol{\rho})$ for the optimization problem (20), the difficulty is how to satisfy constraints (20b), (20c) and (20d). To initialize the transmit power $\boldsymbol{p}$ and subcarrier allocation indicator $\boldsymbol{\gamma}$, we should consider the worst states for both the data traffic of the *first case* and the *second case*. In this state, we assume that all social media is cached at the BSs and the data traffic of the access links in the *first case* are in their maximum values. Moreover, we suppose that all social media are not cached at the cache of BSs and all requests of MUs will be forwarded to the data canter in the *second case*. Hence, we formulate the data traffic of the access link of MU $u \in \mathcal{U}_b$ for the *first case* and the *second case* to initialize $\boldsymbol{p}$ and $\boldsymbol{\gamma}$, respectively, as: $\hat{f}_{b,u}^{\text{Ac},1} = \hat{f}_{b,u}^{\text{Ac},2} = \sum_{c=1}^{C} \delta_{b,u,c} s_c$. The data traffic of the backhaul link b in the *second case* is thus given by

$$\hat{f}_{0,b}^{\text{BH}} = \sum_{c=1}^{C} \sum_{u \in \mathcal{U}_b} \frac{\delta_{b,u,c} s_c}{\max\{\sum_{u \in \mathcal{U}_b} \delta_{b,u,c}, 1\}}. \quad (21)$$

According to the above, we solve the following optimization problem to allocate transmit power $\boldsymbol{p}_0$ as:

$$\min_{\boldsymbol{p}_0} \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} \hat{f}_{b,u}^{\text{Ac},1} t_{b,u}^{\text{Ac},1} + \hat{f}_{b,u}^{\text{Ac},2} t_{b,u}^{\text{Ac},2} + \sum_{b=1}^{B} \hat{f}_{0,b}^{\text{BH}} t_{0,b}^{\text{BH}} \quad (22a)$$

$$\text{s.t.} \quad \hat{f}_{0,b}^{\text{BH}} t_{0,b}^{\text{BH}} \leq \tau_{0,b}^{\text{BH}}, \forall b \in \mathcal{B}, \quad (22b)$$

$$\hat{f}_{b,u}^{\text{Ac},1} t_{b,u}^{\text{Ac},1} \leq \tau_{b,u}^{\text{Ac}}, \forall b \in \mathcal{B}, u \in \mathcal{U}_b, \quad (22c)$$

$$\hat{f}_{b,u}^{\mathrm{Ac},2} t_{b,u}^{\mathrm{Ac},2} \leq \tau_{b,u}^{\mathrm{Ac}}, \forall b \in \mathcal{B}, u \in \mathcal{U}_b, \quad (22d)$$
(20e),(20f),(20m),(20n).

Subsequently, we solve (22) by the proposed Alg. 2. Therefore, the access transmit powers are obtained by

$$p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}(0) = \left[\frac{(\mu_{b,u}(0)\tau_{b,u}^{\mathrm{Ac}} - q_{b,u}^{\mathrm{Ac},1})W_s}{\ln 2.\alpha_b(0)} - \frac{\sigma^2}{h_{b,u}^{n_{\mathrm{Ac}}}}\right]^+, \quad \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_b, \forall n_{\mathrm{Ac}} \in \mathcal{N}_{\mathrm{Ac}}, \quad (23)$$

$$p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}(0) = \left[\frac{(\nu_{b,u}(0)\tau_{b,u}^{\mathrm{Ac}} - q_{b,u}^{\mathrm{Ac},2})W_s}{\ln 2.\alpha_b(0)} - \frac{\sigma^2}{h_{b,u}^{n_{\mathrm{Ac}}}}\right]^+, \quad \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_b, \forall n_{\mathrm{Ac}} \in \mathcal{N}_{\mathrm{Ac}}, \quad (24)$$

where $[.]^+ = \max\{.,0\}$. We update the Lagrangian multipliers $\boldsymbol{\mu}_0$, $\boldsymbol{\nu}_0$ and $\boldsymbol{\alpha}_0$ with the subgradient method corresponding to constraints (22c), (22d) and (20e), respectively. By using the Karush-Kuhn-Tucker (KKT) conditions, the transmit power $\boldsymbol{p}_0^{\mathrm{BH}}$ is obtained by

$$p_{0,b}^{n_{\mathrm{BH}}}(0) = \left[\frac{(\lambda_b(0)\tau_{0,b}^{\mathrm{BH}} - q_{0,b}^{\mathrm{BH}})W_s}{\ln 2.\beta(0)} - \frac{\sigma^2}{h_{0,b}^{n_{\mathrm{BH}}}}\right]^+, \quad \forall b \in \mathcal{B}, \forall n_{\mathrm{BH}} \in \mathcal{N}_{\mathrm{BH}}. \quad (25)$$

We update the Lagrangian multipliers $\boldsymbol{\lambda}_0$ and $\boldsymbol{\beta}_0$ with the subgradient method corresponding to constraints (20b) and (20f), respectively. After finding $\boldsymbol{p}_0$, we find $\boldsymbol{\gamma}_0$ using the CVX software with MOSEK solver.

3) Content placement optimization: In this subsection, we solve the following content placement problem for a given $(\boldsymbol{p}, \boldsymbol{\gamma})$ with respect to the variable $\boldsymbol{\rho}$ as:

$$\min_{\boldsymbol{\rho}} \ D_{\mathrm{tot}} \quad (26a)$$
$$\text{s.t.} \ \ (20b)\text{-}(20d),(20g),(20j).$$

The optimization problem (26) is a BLP problem which can be solve by the BLP optimization toolboxes, such as CVX with the internal solver MOSEK [22].

4) Joint transmit power and subcarrier allocation optimization: For a given $\boldsymbol{\rho}$, we solve the following optimization problem to find $\boldsymbol{p}$ and $\boldsymbol{\gamma}$ as:

$$\min_{\boldsymbol{p},\boldsymbol{\gamma}} \ D_{\mathrm{tot}} \quad (27a)$$
$$\text{s.t.} \ \ (20b)\text{-}(20f),(20h),(20i),\ (20k)\text{-}(20n).$$

The optimization problem (27) can be divided into two separated subproblems, where the *first subproblem* is formulated as:

$$\min_{\boldsymbol{p}^{\mathrm{Ac}},\boldsymbol{\gamma}^{\mathrm{Ac}}} \ \sum_{b=1}^{B}\sum_{u\in\mathcal{U}_b} D_{b,u}^{\mathrm{Ac},1} + D_{b,u}^{\mathrm{Ac},2} \quad (28a)$$
$$\text{s.t.} \ \ (20c)\text{-}(20e),(20h),(20l),(20n)\ ,$$

and the *second subproblem* is given by

$$\min_{\boldsymbol{p}^{\mathrm{BH}},\boldsymbol{\gamma}^{\mathrm{BH}}} \ \sum_{b=1}^{B} D_{0,b}^{\mathrm{BH}} \quad (29a)$$
$$\text{s.t.} \ \ (20b),(20f),(20i),(20k),(20m).$$

The objective functions (28a) and (29a) are the ratio of two functions which result in non-convex functions. Hence, (28) and (29) can be classified as GLFP problems [23], [24]. Without loss of generality, we define the minimum latency of each MU $u \in \mathcal{U}_b$ for the *first case* as $q_{b,u}^{\mathrm{Ac},1^*}$, for the *first hop* of the *second case* as $q_{0,b}^{\mathrm{BH}^*}$ and for the *second hop* of the *second case* as $q_{b,u}^{\mathrm{Ac},2^*}$ which are formulated, respectively, as follows:

$$q_{b,u}^{\mathrm{Ac},1^*} = \frac{f_{b,u}^{\mathrm{Ac},1}}{w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}{}^*}, \gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}{}^*})}, \quad (30)$$

$$q_{b,u}^{\mathrm{Ac},2^*} = \frac{f_{b,u}^{\mathrm{Ac},2}}{w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}{}^*}, \gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}{}^*})}, \quad (31)$$

and

$$q_{0,b}^{\mathrm{BH}^*} = \frac{f_{0,b}^{\mathrm{BH}}}{w_{0,b}^{\mathrm{BH},\mathrm{tot}}(p_{0,b}^{n_{\mathrm{BH}}{}^*}, \gamma_{0,b}^{n_{\mathrm{BH}}{}^*})}. \quad (32)$$

According to Theorem 1 in [13] which is proved in [14], the minimum $q_{b,u}^{\mathrm{Ac},1^*}$ can be achieved for each MU $u \in \mathcal{U}_b$ *if and only if*

$$\min_{p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}} f_{b,u}^{\mathrm{Ac},1} - q_{b,u}^{\mathrm{Ac},1^*} w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}) = f_{b,u}^{\mathrm{Ac},1} - q_{b,u}^{\mathrm{Ac},1^*} w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}{}^*}, \gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}{}^*}) = 0, \quad (33)$$

for $f_{b,u}^{\mathrm{Ac},1} \geq 0$ and $w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}) > 0$. In addition, the minimum $q_{b,u}^{\mathrm{Ac},2^*}$ can be achieved for each MU $u \in \mathcal{U}_b$ *if and only if*

$$\min_{p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}} f_{b,u}^{\mathrm{Ac},2} - q_{b,u}^{\mathrm{Ac},2^*} w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}) = f_{b,u}^{\mathrm{Ac},2} - q_{b,u}^{\mathrm{Ac},2^*} w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}{}^*}, \gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}{}^*}) = 0, \quad (34)$$

for $f_{b,u}^{\mathrm{Ac},2} \geq 0$ and $w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}) > 0$. The minimum $q_{0,b}^{\mathrm{BH}^*}$ is also achieved for each BS $b \in \mathcal{B}$ *if and only if*

$$\min_{p_{0,b}^{n_{\mathrm{BH}}}, \gamma_{0,b}^{n_{\mathrm{BH}}}} f_{0,b}^{\mathrm{BH}} - q_{0,b}^{\mathrm{BH}^*} w_{0,b}^{\mathrm{BH},\mathrm{tot}}(p_{0,b}^{n_{\mathrm{BH}}}, \gamma_{0,b}^{n_{\mathrm{BH}}}) = f_{0,b}^{\mathrm{BH}} - q_{0,b}^{\mathrm{BH}^*} w_{0,b}^{\mathrm{BH},\mathrm{tot}}(p_{0,b}^{n_{\mathrm{BH}}{}^*}, \gamma_{0,b}^{n_{\mathrm{BH}}{}^*}) = 0, \quad (35)$$

for $f_{0,b}^{\mathrm{BH}} \geq 0$ and $w_{0,b}^{\mathrm{BH},\mathrm{tot}}(p_{0,b}^{n_{\mathrm{BH}}}, \gamma_{0,b}^{n_{\mathrm{BH}}}) > 0$. Consequently, there exist equivalent objective functions in subtractive forms for the objective functions (28a) and (29a) as $f_{b,u}^{\mathrm{Ac},1} - q_{b,u}^{\mathrm{Ac},1^*} w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}) + f_{b,u}^{\mathrm{Ac},2} - q_{b,u}^{\mathrm{Ac},2^*} w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}, \gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}})$ and $f_{0,b}^{\mathrm{BH}} - q_{0,b}^{\mathrm{BH}^*} w_{0,b}^{\mathrm{BH},\mathrm{tot}}(p_{0,b}^{n_{\mathrm{BH}}}, \gamma_{0,b}^{n_{\mathrm{BH}}})$, respectively. In the following, we propose an iterative approach which is known as the Dinkelbach method [24] to solve (28) and (29) with the equivalent objective functions. In each iteration, for given parameters $q_{b,u}^{\mathrm{Ac},1}$, $q_{b,u}^{\mathrm{Ac},2}$ and $q_{b,u}^{\mathrm{BH}}$, we should solve the *first inner subproblem* and *second inner subproblem*, respectively as:

$$\min_{\boldsymbol{p}^{\mathrm{Ac}},\boldsymbol{\gamma}^{\mathrm{Ac}}} \ \sum_{b=1}^{B}\sum_{u\in\mathcal{U}_b} f_{b,u}^{\mathrm{Ac},1} - q_{b,u}^{\mathrm{Ac},1} w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}} + f_{b,u}^{\mathrm{Ac},2} - q_{b,u}^{\mathrm{Ac},2} w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}} \quad (36a)$$

Algorithm 2 The Dinkelbach algorithm for solving the joint transmit power and subcarrier allocation optimization problem (27)

Initialize $q_{b,u}^{\mathrm{Ac},1} = q_{b,u}^{\mathrm{Ac},2} = 0, \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_b$ and $q_{0,b}^{\mathrm{BH}} = 0, \forall b \in \mathcal{B}$.
repeat
for $t_2 = 1$ to T_2 **do**
  Solve the *first inner subproblem* (36) for given $q_{b,u}^{\mathrm{Ac},1}$ and $q_{b,u}^{\mathrm{Ac},2}$ using Alg. 3 and obtain the optimal ${p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}}^*$, ${p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}}^*$, ${\gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}}^*$ and ${\gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}}^*$.
  if $f_{b,u}^{\mathrm{Ac},1} - q_{b,u}^{\mathrm{Ac},1} w_{b,u}^{\mathrm{Ac},1}(p'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u}) < \epsilon_{q_{\mathrm{Ac},1}}$ & $f_{b,u}^{\mathrm{Ac},2} - q_{b,u}^{\mathrm{Ac},2} w_{b,u}^{\mathrm{Ac},2}(p'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u}) < \epsilon_{q_{\mathrm{Ac},2}}$, **then**
    Convergence1 =**true**
    return $({p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}}^*, {\gamma_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}}^*) = (p'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u})$, ${q_{b,u}^{\mathrm{Ac},1}}^* = \frac{f_{b,u}^{\mathrm{Ac},1}}{w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u})}$, and $({p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}}^*, {\gamma_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}}^*) = (p'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u})$, ${q_{b,u}^{\mathrm{Ac},2}}^* = \frac{f_{b,u}^{\mathrm{Ac},2}}{w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u})}$
  else
    Set $q_{b,u}^{\mathrm{Ac},1} = \frac{f_{b,u}^{\mathrm{Ac},1}}{w_{b,u}^{\mathrm{Ac},1,\mathrm{tot}}(p'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},1,n_{\mathrm{Ac}}}_{b,u})}$ and $q_{b,u}^{\mathrm{Ac},2} = \frac{f_{b,u}^{\mathrm{Ac},2}}{w_{b,u}^{\mathrm{Ac},2,\mathrm{tot}}(p'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u}, \gamma'^{\mathrm{Ac},2,n_{\mathrm{Ac}}}_{b,u})}$
    Convergence1 =**false**
  end if
  Set $t_2 = t_2 + 1$
  Until Convergence1 =**true** or $t_2 = T_2$.
end for
for $t_3 = 1$ to T_3 **do**
  Solve the *second inner subproblem* (37) for a given $q_{0,b}^{\mathrm{BH}}$ using Alg. 4 and obtain the optimal ${p_{0,b}^{n_{\mathrm{BH}}}}^*$ and ${\gamma_{0,b}^{n_{\mathrm{BH}}}}^*$.
  if $f_{0,b}^{\mathrm{BH}} - q_{0,b}^{\mathrm{BH}} w_{0,b}^{\mathrm{BH,tot}}(p'^{n_{\mathrm{BH}}}_{0,b}, \gamma'^{n_{\mathrm{BH}}}_{0,b}) < \epsilon_{q_{\mathrm{BH}}}$ **then**
    Convergence2 =**true**
    return $({p_{0,b}^{n_{\mathrm{BH}}}}^*, {\gamma_{0,b}^{n_{\mathrm{BH}}}}^*) = (p'^{n_{\mathrm{BH}}}_{0,b}, \gamma'^{n_{\mathrm{BH}}}_{0,b})$ and ${q_{0,b}^{\mathrm{BH}}}^* = \frac{f_{0,b}^{\mathrm{BH}}}{w_{0,b}^{\mathrm{BH,tot}}(p'^{n_{\mathrm{BH}}}_{0,b}, \gamma'^{n_{\mathrm{BH}}}_{0,b})}$
  else
    Set $q_{0,b}^{\mathrm{BH}} = \frac{f_{0,b}^{\mathrm{BH}}}{w_{0,b}^{\mathrm{BH,tot}}(p'^{n_{\mathrm{BH}}}_{0,b}, \gamma'^{n_{\mathrm{BH}}}_{0,b})}$
    Convergence2 =**false**
  end if
  Set $t_3 = t_3 + 1$
  Until Convergence2 =**true** or $t_3 = T_3$.
end for

s.t. (20c)-(20e),(20h),(20l),(20n),

and

$$\min_{\boldsymbol{p}^{\mathrm{BH}}, \boldsymbol{\gamma}^{\mathrm{BH}}} \quad \sum_{b=1}^{B} f_{0,b}^{\mathrm{BH}} - q_{0,b}^{\mathrm{BH}} w_{0,b}^{\mathrm{BH,tot}} \tag{37a}$$
$$\text{s.t.} \quad \text{(20b),(20f),(20i),(20k),(20m).}$$

The pseudo code of the proposed Dinkelbach method is summarized in Alg. 2. Accordingly, for given $q_{b,u}^{\mathrm{Ac},1}$ and $q_{b,u}^{\mathrm{Ac},2}$, the *first inner subproblem* (36) and for a given $q_{0,b}^{\mathrm{BH}}$, the *second inner subproblem* (37) can be solved by utilizing the Lagrange dual decomposition approach with the subgradient method as follows [28]: The Lagrangian function of the optimization problem (36) is given by

$$\begin{aligned} L(\boldsymbol{p}^{\mathrm{Ac}}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = &\sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} \big(f_{b,u}^{\mathrm{Ac},1} - q_{b,u}^{\mathrm{Ac},1} \sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} w_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}} \\ &+ f_{b,u}^{\mathrm{Ac},2} - q_{b,u}^{\mathrm{Ac},2} \sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} w_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}} \big) + \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} \mu_{b,u} \big(\sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} \\ &w_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}} \tau_{b,u}^{\mathrm{Ac}} - f_{b,u}^{\mathrm{Ac},1} \big) + \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} \nu_{b,u} \big(\sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} w_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}} \\ &\tau_{b,u}^{\mathrm{Ac}} - f_{b,u}^{\mathrm{Ac},2} \big) + \sum_{b=1}^{B} \alpha_b \big(P_b^{\max} - \sum_{u \in \mathcal{U}_b} \sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}} + \\ &p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}} \big), \end{aligned} \tag{38}$$

where $\boldsymbol{\mu} = [\mu_{b,u}]$, $\boldsymbol{\nu} = [\nu_{b,u}]$ and $\boldsymbol{\alpha} = [\alpha_b]$ are the vectors of Lagrangian multipliers corresponding to constraints (20c), (20d) and (20e), respectively. Moreover, the Lagrangian dual function of (36) is formulated by

$$g(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = \min_{\boldsymbol{p}^{\mathrm{Ac}}} L(\boldsymbol{p}^{\mathrm{Ac}}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}). \tag{39}$$

According to (39), the dual problem of (36) is formulated as:

$$\max_{\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}} \quad g(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) \tag{40a}$$
$$\text{s.t.} \quad \alpha_b \geq 0, \forall b \in \mathcal{B}, \tag{40b}$$
$$\mu_{b,u} \geq 0, \nu_{b,u} \geq 0, \forall b \in \mathcal{B}, \forall u \in \mathcal{U}_b. \tag{40c}$$

Moreover, the Lagrangian dual function (39) is decomposed as follows:

$$\begin{aligned} g(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = &\sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} g_{n_{\mathrm{Ac}}}^{1}(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) + \sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}} g_{n_{\mathrm{Ac}}}^{2}(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) \\ &+ \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} (1 - \mu_{b,u}) f_{b,u}^{\mathrm{Ac},1} + (1 - \nu_{b,u}) f_{b,u}^{\mathrm{Ac},2} + \sum_{b=1}^{B} \alpha_b P_b^{\max}, \end{aligned} \tag{41}$$

where,

$$\begin{aligned} g_{n_{\mathrm{Ac}}}^{1}(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = \min_{\boldsymbol{p}^{\mathrm{Ac}}} \min_{u \in \mathcal{U}_b} \Big\{ &(\mu_{b,u} \tau_{b,u}^{\mathrm{Ac}} - q_{b,u}^{\mathrm{Ac},1}) w_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}} \\ &- \alpha_b p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}} \Big\}, \forall n_{\mathrm{Ac}} \in \mathcal{N}_{\mathrm{Ac}}, \end{aligned} \tag{42}$$

and

$$\begin{aligned} g_{n_{\mathrm{Ac}}}^{2}(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = \min_{\boldsymbol{p}^{\mathrm{Ac}}} \min_{u \in \mathcal{U}_b} \Big\{ &(\nu_{b,u} \tau_{b,u}^{\mathrm{Ac}} - q_{b,u}^{\mathrm{Ac},2}) w_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}} \\ &- \alpha_b p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}} \Big\}, \forall n_{\mathrm{Ac}} \in \mathcal{N}_{\mathrm{Ac}}. \end{aligned} \tag{43}$$

The optimization problem (40) is convex, because the Lagrangian dual function (40a) is a linear function of $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$. To maximize the objective function (40a), since the Lagrangian function (38) is a differentiable function of $\boldsymbol{p}^{\mathrm{Ac}}$,

we should use KKT conditions to obtain the transmit power $\boldsymbol{p}^{\mathrm{Ac}}$. By using KKT conditions, the transmit power $\boldsymbol{p}^{\mathrm{Ac}}$ and subcarrier assignment indicator $\boldsymbol{\gamma}^{\mathrm{Ac}}$ can be obtained in the following steps:

1. We should calculate the gradient of the Lagrangian function (38) with respect to $\boldsymbol{p}^{\mathrm{Ac}}$. Therefore, we have

$$\frac{\partial L(\boldsymbol{p}^{\mathrm{Ac}},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\alpha})}{\boldsymbol{p}^{\mathrm{Ac},1}}=\frac{(\mu_{b,u}\tau_{b,u}^{\mathrm{Ac}}-q_{b,u}^{\mathrm{Ac},1})W_s h_{b,u}^{n_{\mathrm{Ac}}}/\sigma^2}{\ln 2.(1+p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}h_{b,u}^{n_{\mathrm{Ac}}}/\sigma^2)}-\alpha_b. \tag{44}$$

$$\frac{\partial L(\boldsymbol{p}^{\mathrm{Ac}},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\alpha})}{\boldsymbol{p}^{\mathrm{Ac},2}}=\frac{(\nu_{b,u}\tau_{b,u}^{\mathrm{Ac}}-q_{b,u}^{\mathrm{Ac},2})W_s h_{b,u}^{n_{\mathrm{Ac}}}/\sigma^2}{\ln 2.(1+p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}h_{b,u}^{n_{\mathrm{Ac}}}/\sigma^2)}-\alpha_b. \tag{45}$$

Then, the transmit power $\boldsymbol{p}^{\mathrm{Ac}}$ is given by

$$p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}=\left[\frac{(\mu_{b,u}\tau_{b,u}^{\mathrm{Ac}}-q_{b,u}^{\mathrm{Ac},1})W_s}{\ln 2.\alpha_b}-\frac{\sigma^2}{h_{b,u}^{n_{\mathrm{Ac}}}}\right]^+, \quad \forall b\in\mathcal{B},\forall u\in\mathcal{U}_b,\forall n_{\mathrm{Ac}}\in\mathcal{N}_{\mathrm{Ac}}, \tag{46}$$

$$p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}=\left[\frac{(\nu_{b,u}\tau_{b,u}^{\mathrm{Ac}}-q_{b,u}^{\mathrm{Ac},2})W_s}{\ln 2.\alpha_b}-\frac{\sigma^2}{h_{b,u}^{n_{\mathrm{Ac}}}}\right]^+, \quad \forall b\in\mathcal{B},\forall u\in\mathcal{U}_b,\forall n_{\mathrm{Ac}}\in\mathcal{N}_{\mathrm{Ac}}, \tag{47}$$

where $[.]^+=\max\{.,0\}$.

2. According to (42) and (43), with respect to the optimal transmit power $\boldsymbol{p}^{\mathrm{Ac}}$, the optimal MU $\tilde{u}\in\mathcal{U}_{\tilde{b}}$ for assigning subcarrier n_{Ac} for the *first case* is obtained as:

$$\tilde{u}\in\mathcal{U}_{\tilde{b}}=\arg\min_{u\in\mathcal{U}_b}\varphi_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}, \tag{48}$$

where,

$$\varphi_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}=(\mu_{b,u}\tau_{b,u}^{\mathrm{Ac}}-q_{b,u}^{\mathrm{Ac},1})w_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}-\alpha_b p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}, \quad \forall b\in\mathcal{B},\forall u\in\mathcal{U}_b,\forall n_{\mathrm{Ac}}\in\mathcal{N}_{\mathrm{Ac}}, \tag{49}$$

and the optimal MU $\bar{u}\in\mathcal{U}_{\bar{b}}$ for assigning subcarrier n_{Ac} for the *second case* is given by

$$\bar{u}\in\mathcal{U}_{\bar{b}}=\arg\min_{u\in\mathcal{U}_b}\varphi_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}, \tag{50}$$

where,

$$\varphi_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}=(\nu_{b,u}\tau_{b,u}^{\mathrm{Ac}}-q_{b,u}^{\mathrm{Ac},2})w_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}-\alpha_b p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}, \quad \forall b\in\mathcal{B},\forall u\in\mathcal{U}_b,\forall n_{\mathrm{Ac}}\in\mathcal{N}_{\mathrm{Ac}}, \tag{51}$$

and finally, the subcarrier allocation for access links is based on the following condition which is given by

$$\begin{cases}\gamma_{\tilde{b},\tilde{u}}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}=1\ \&\ \gamma_{\bar{b},\bar{u}}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}=0, & \text{if } \varphi_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}<\varphi_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}};\\ \gamma_{\tilde{b},\tilde{u}}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}=0\ \&\ \gamma_{\bar{b},\bar{u}}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}=1, & \text{if } \varphi_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}>\varphi_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}.\end{cases} \tag{52}$$

3. We update the Lagrangian multipliers $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ with the subgradient method as follows:

$$\mu_{b,u}^{(i+1)}=\left[\mu_{b,u}^{(i)}-\epsilon_\mu\Big(\sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}}w_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}\tau_{b,u}^{\mathrm{Ac}}-f_{b,u}^{\mathrm{Ac},1}\Big)\right]^+, \quad \forall b\in\mathcal{B},\forall u\in\mathcal{U}_b, \tag{53}$$

Algorithm 3 The Lagrangian dual decomposition algorithm for solving the *first inner subproblem* (36)

1: Initialize $\boldsymbol{\mu}_0,\boldsymbol{\nu}_0,\boldsymbol{\alpha}_0$.
repeat
2: **for** $t_4=1$ to T_4 **do**
3: Find $\boldsymbol{p}^{\mathrm{Ac}}$ using (46) and (47).
4: Find $\boldsymbol{\gamma}^{\mathrm{Ac}}$ using (48), (50) and (52).
5: Update the Lagrangian multipliers $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ using the subgradient approach.
6: Set $t_4=t_4+1$
7: **Until** $||g(\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\alpha})-g^*(\boldsymbol{\mu}^*,\boldsymbol{\nu}^*,\boldsymbol{\alpha}^*)||\le\omega_2$ or $t_4=T_4$.
8: **end for**

$$\nu_{b,u}^{(i+1)}=\left[\nu_{b,u}^{(i)}-\epsilon_\nu\Big(\sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}}w_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}\tau_{b,u}^{\mathrm{Ac}}-f_{b,u}^{\mathrm{Ac},2}\Big)\right]^+, \quad \forall b\in\mathcal{B},\forall u\in\mathcal{U}_b, \tag{54}$$

$$\alpha_b^{(i+1)}=\left[\alpha_b^{(i)}-\epsilon_\alpha\Big(P_b^{\max}-\sum_{u\in\mathcal{U}_b}\sum_{n_{\mathrm{Ac}}=1}^{N_{\mathrm{Ac}}}p_{b,u}^{\mathrm{Ac},1,n_{\mathrm{Ac}}}+p_{b,u}^{\mathrm{Ac},2,n_{\mathrm{Ac}}}\Big),\right]^+, \forall b\in\mathcal{B}, \tag{55}$$

where $[.]^+=\max\{.,0\}$ and i is the iteration index and ϵ_μ, ϵ_ν and ϵ_α are the positive step-size sequences specified to (53), (54) and (55), respectively.

The pseudo code of the Lagrangian dual decomposition approach for solving the *first inner subproblem* (36) is presented in Alg. 3. Moreover, the Lagrangian function of the *second inner subproblem* (37) is formulated as follows:

$$L(\boldsymbol{p}^{\mathrm{BH}},\boldsymbol{\lambda},\boldsymbol{\beta})=\sum_{b=1}^{B}f_{0,b}^{\mathrm{BH}}-q_{0,b}^{\mathrm{BH}}\sum_{n_{\mathrm{BH}}=1}^{N_{\mathrm{BH}}}w_{0,b}^{\mathrm{BH},n_{\mathrm{BH}}}+\sum_{b=1}^{B}\lambda_b\Big(\sum_{n_{\mathrm{BH}}=1}^{N_{\mathrm{BH}}}w_{0,b}^{\mathrm{BH},n_{\mathrm{BH}}}\tau_{0,b}^{\mathrm{BH}}-f_{0,b}^{\mathrm{BH}}\Big)+\beta\Big(P_0^{\max}-\sum_{b=1}^{B}\sum_{n_{\mathrm{BH}}=1}^{N_{\mathrm{BH}}}p_{0,b}^{n_{\mathrm{BH}}}\Big), \tag{56}$$

where $\boldsymbol{\lambda}=[\lambda_b]$ and $\boldsymbol{\beta}=[\beta]$ are the vectors of Lagrangian multipliers corresponding to constraints (20b) and (20f), respectively. In addition, the dual problem of (37) is formulated as:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\beta}}\ g(\boldsymbol{\lambda},\boldsymbol{\beta}) \tag{57a}$$

$$\text{s.t.}\ \ \lambda_b\ge 0,\forall b\in\mathcal{B}, \tag{57b}$$

$$\beta\ge 0, \tag{57c}$$

where,

$$g(\boldsymbol{\lambda},\boldsymbol{\beta})=\min_{\boldsymbol{p}^{\mathrm{BH}}}L(\boldsymbol{p}^{\mathrm{BH}},\boldsymbol{\lambda},\boldsymbol{\beta})=\min_{\boldsymbol{p}^{\mathrm{BH}}}\Bigg\{\sum_{b=1}^{B}f_{0,b}^{\mathrm{BH}}-q_{0,b}^{\mathrm{BH}}\sum_{n_{\mathrm{BH}}=1}^{N_{\mathrm{BH}}}w_{0,b}^{\mathrm{BH},n_{\mathrm{BH}}}+\sum_{b=1}^{B}\lambda_b\Big(\sum_{n_{\mathrm{BH}}=1}^{N_{\mathrm{BH}}}w_{0,b}^{\mathrm{BH},n_{\mathrm{BH}}}\tau_{0,b}^{\mathrm{BH}}-f_{0,b}^{\mathrm{BH}}\Big)+$$

$$\beta\big(P_0^{\max}-\sum_{b=1}^{B}\sum_{n_{\rm BH}=1}^{N_{\rm BH}} p_{0,b}^{n_{\rm BH}}\big)\Big\}, \quad (58)$$

is the Lagrangian dual function of (37). Moreover, the decomposition of the Lagrangian dual function (58) is given by

$$g(\boldsymbol{\lambda},\boldsymbol{\beta})=\sum_{n_{\rm BH}=1}^{N_{\rm BH}} g_{n_{\rm BH}}(\boldsymbol{\lambda},\boldsymbol{\beta})+\sum_{b=1}^{B}(1-\lambda_b)f_{0,b}^{\rm BH}+\beta P_0^{\max}, \quad (59)$$

where,

$$g_{n_{\rm BH}}(\boldsymbol{\lambda},\boldsymbol{\beta})=\min_{\boldsymbol{p}^{\rm BH}}\min_{b\in\mathcal{B}}\Big\{(\lambda_b\tau_{0,b}^{\rm BH}-q_{0,b}^{\rm BH})w_{0,b}^{{\rm BH},n_{\rm BH}}- \beta p_{0,b}^{n_{\rm BH}}\Big\}, \forall n_{\rm BH}\in\mathcal{N}_{\rm BH}. \quad (60)$$

By exploiting KKT conditions, the transmit power $\boldsymbol{p}^{\rm BH}$ can be obtained as:

$$p_{0,b}^{n_{\rm BH}}=\Big[\frac{(\lambda_b\tau_{0,b}^{\rm BH}-q_{0,b}^{\rm BH})W_s}{\ln 2.\beta}-\frac{\sigma^2}{h_{0,b}^{n_{\rm BH}}}\Big]^+, \forall b\in\mathcal{B}, \forall n_{\rm BH}\in\mathcal{N}_{\rm BH}. \quad (61)$$

Moreover, the updating of the Lagrangian multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ is obtained as follows:

$${\lambda_b}^{(i+1)}=\Big[{\lambda_b}^{(i)}-\epsilon_\lambda\big(\sum_{n_{\rm BH}=1}^{N_{\rm BH}} w_{0,b}^{{\rm BH},n_{\rm BH}}\tau_{0,b}^{\rm BH}-f_{0,b}^{\rm BH}\big)\Big]^+, \forall b\in\mathcal{B}, \quad (62)$$

$$\beta^{(i+1)}=\Big[\beta^{(i)}-\epsilon_\beta\big(P_0^{\max}-\sum_{b=1}^{B}\sum_{n_{\rm BH}=1}^{N_{\rm BH}} p_{0,b}^{n_{\rm BH}}\big)\Big]^+. \quad (63)$$

Accordingly, the optimal BS b for assigning subcarrier $n_{\rm BH}$ with respect to the transmit power $\boldsymbol{p}^{\rm BH}$ is obtained as:

$$\hat{b}\in\mathcal{B}=\arg\min_{b\in\mathcal{B}}\Big\{(\lambda_b\tau_{0,b}^{\rm BH}-q_{0,b}^{\rm BH})w_{0,b}^{{\rm BH},n_{\rm BH}}-\beta p_{0,b}^{n_{\rm BH}}\Big\}. \quad (64)$$

Then, we have

$$\begin{cases}\gamma_{0,\hat{b}}^{n_{\rm BH}}=1, & \forall b=\hat{b};\\ \gamma_{0,b}^{n_{\rm BH}}=0, & \text{otherwise.}\end{cases} \quad (65)$$

The pseudo code of the Lagrangian dual decomposition approach for solving the *second inner subproblem* (37) is presented in Alg. 4.

Proposition 2: The convergence of the proposed algorithms for solving (28) and (29) to optimal points are guaranteed if we are able to solve the inner problems (36) and (37) in each iteration, respectively. Therefore, the proposed Dinkelbach Alg. 2 improves the objective functions (28a) and (29a) and converges to an optimum solution, if we are able to solve the inner problems (36) and (37), in each iteration.

Proof. Please refer to [24] and [14] for the proof of convergence for the proposed Dinkelbach Alg. 2. □

Proposition 3: The inner problems (36) and (37) can satisfy the time-sharing condition and the duality gaps of the considered Lagrange dual decomposition methods tend to zero as the total number of subcarriers tends to infinity for both the inner problems (36) and (37).

Proof. Please refer to Appendix B. □

Algorithm 4 The Lagrangian dual decomposition algorithm for solving the *second inner subproblem* (37)

1: Initialize $\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0$.
repeat
2: **for** $t_5=1$ to T_5 **do**
3: Find $\boldsymbol{p}^{\rm BH}$ using (61).
4: Find $\boldsymbol{\gamma}^{\rm BH}$ using (64).
5: Update the Lagrangian multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ using (62) and (63), respectively.
6: Set $t_5=t_5+1$
7: **Until** $||g(\boldsymbol{\lambda},\boldsymbol{\beta})-g^*(\boldsymbol{\lambda}^*,\boldsymbol{\beta}^*)||\le\omega_3$ or $t_5=T_5$.
8: **end for**

Algorithm 5 The exhaustive-search method

1: Select Z to establish $\Delta_p=\frac{1}{Z}$
2: Find the possible solutions vectors $\boldsymbol{p}^{\rm pos}$, $\boldsymbol{\gamma}^{\rm pos}$ and $\boldsymbol{\rho}^{\rm pos}$, where $p_{b,u}^{{\rm Ac},1,n_{\rm Ac}}=\{0,\Delta_p,2\Delta_p,\dots,1\},\forall b\in\mathcal{B},u\in\mathcal{U}_b,\forall n_{\rm Ac}\in\mathcal{N}_{\rm Ac}$, $p_{b,u}^{{\rm Ac},2,n_{\rm Ac}}=\{0,\Delta_p,2\Delta_p,\dots,1\},\forall b\in\mathcal{B},u\in\mathcal{U}_b,\forall n_{\rm Ac}\in\mathcal{N}_{\rm Ac}$, $p_{0,b}^{n_{\rm BH}}=\{0,\Delta_p,2\Delta_p,\dots,1\},\forall b\in\mathcal{B},\forall n_{\rm BH}\in\mathcal{N}_{\rm BH}$, $\gamma_{b,u}^{{\rm Ac},1,n_{\rm Ac}},\gamma_{b,u}^{{\rm Ac},2,n_{\rm Ac}}\in\{0,1\},\forall b\in\mathcal{B},u\in\mathcal{U}_b,\forall n_{\rm Ac}\in\mathcal{N}_{\rm Ac}$, $\gamma_{0,b}^{n_{\rm BH}}\in\{0,1\},\forall b\in\mathcal{B},\forall n_{\rm BH}\in\mathcal{N}_{\rm BH}$, $\rho_{b,c}\in\{0,1\},\forall b\in\mathcal{B},\forall c\in\mathcal{C}$.
3: Find the feasible solutions from $\boldsymbol{p}^{\rm pos}$, $\boldsymbol{\gamma}^{\rm pos}$ and $\boldsymbol{\rho}^{\rm pos}$ that satisfy the constraints of the considered optimization problem.
4: From the solutions which are found in Step 3, select the optimum solution which minimizes the objective function (20a)

B. Exhaustive-Search Method

The performance of the proposed iterative Alg. 1 is compared with the exhaustive-search method which gives us the global optimum solution for the optimization problem (20). The pseudo code of the exhaustive-search method is presented in Alg. 5. Moreover, in this method, all possible solutions for the transmit power and subcarrier and content placement indicators are investigated to obtain the global optimum solution for a given specific step Z [25].

V. Computational Complexity

Now, we investigate the computational complexity of the proposed iterative Alg. 1. The mentioned Alg. 1 consists of two main steps: 1) solving the content placement problem (26) via CVX with the internal solver MOSEK; 2) solving the joint transmit power and subcarrier allocation problem (27) using fractional programming. We also note that the CVX solver employs GP with the interior point method (IPM) [25]. In addition, the total number of constraints in (26) is $(2B+2U)$.

Hence, the total number of required iterations for solving (26) via CVX is given by [25]

$$\Omega_1^{\text{CVX}} = \frac{\log\big((2B+2U)/t^0\varrho\big)}{\log\zeta_2}, \tag{66}$$

where ζ_2 is the accuracy updating parameter of IPM, $0 \leq \varrho \ll 1$ is the stopping criterion for IPM and t^0 is the initial point for approximated the accuracy of IPM. MTo solve (27) using Alg. 2, we note that the proposed Dinkelbach method consists of two nested loops. The outer loop can be proved to have a linear time complexity as Φ_1 [15]. Moreover, the inner loop optimization problem is divided into two separated subproblems such that both the *first inner subproblem* (36) and the *second inner subproblem* (37) are solved via Lagrangian dual decomposition algorithm with the subgradient method in Alg. 3 and 4, respectively. The computational complexity of solving the *first inner subproblem* (36) consists of three main steps: obtaining the transmit power $\boldsymbol{p}^{\text{Ac}}$, subcarrier allocation $\boldsymbol{\gamma}^{\text{Ac}}$ and updating the Lagrangian multipliers corresponding to (53), (54), and (55). Since the dual function (39) is convex and non-differentiable, we use the subgradient method. Let Ψ^i be a diminishing step size, where i is the number of iterations. According to [26], we have $\sum_{i=1}^{\infty}(\Psi_i)_2 < \infty$ and $\sum_{i=1}^{\infty}\Psi^i \rightarrow \infty$. Hence, the iteration number required to achieve ω_2-optimality, i.e., $||g(\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\alpha}) - g^*(\boldsymbol{\mu}^*,\boldsymbol{\nu}^*,\boldsymbol{\alpha}^*)|| \leq \omega_2$ is on the order of $O\big(\frac{1}{\omega_2^2}\big)$ [26]. Therefore, the required iteration number does not depend on the number of dual variables [27]. In each iteration, the computational complexity of obtaining the transmit power $\boldsymbol{p}^{\text{Ac}}$ using (46) and (47) is on the order of $O\big(2UN_{\text{Ac}}\big)$. Moreover, it is required to find (42) and (43) in each iteration and find the subcarrier indicator $\boldsymbol{\gamma}^{\text{Ac}}$ using (52). The computational complexity of obtaining (42) and (43) are on the order of $O\big(N_{\text{Ac}}\big)$ and computing (52) has a linear time complexity as Υ_1. Moreover, corresponding to (53), (54), and (55), the computational complexity of updating Lagrangian multipliers $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ is on the order of $O\big(2U+B\big)$. Hence, the computational complexity of solving the *first inner subproblem* (36) is given by

$$\Omega_1^{\text{Ac}} = O\big(\frac{2UN_{\text{Ac}} + (2N_{\text{Ac}}+\Upsilon_1) + (2U+B)}{\omega_2^2}\big). \tag{67}$$

According to the above, we obtain the computational complexity of solving the *second inner subproblem* (37) to achieve ω_3-optimality solution via Alg. 4 as:

$$\Omega_1^{\text{BH}} = O\big(\frac{BN_{\text{BH}} + N_{\text{BH}} + (B+1)}{\omega_3^2}\big), \tag{68}$$

where the computational complexity of obtaining $\boldsymbol{p}^{\text{BH}}$, $\boldsymbol{\gamma}^{\text{BH}}$ and updating Lagrangian multipliers corresponding to (62) and (63) are on the order of $O\big(BN_{\text{BH}}\big)$, $O\big(N_{\text{BH}}\big)$ and $O\big(B+1\big)$, respectively. Finally, the total computational complexity of solving (20) is given by

$$\Omega_1^{\text{tot}} = O\big(\Omega_1^{\text{CVX}} + \Phi_1(\Omega_1^{\text{Ac}} + \Omega_1^{\text{BH}})\big). \tag{69}$$

The computational complexity of the proposed algorithm is summarized in Table I.

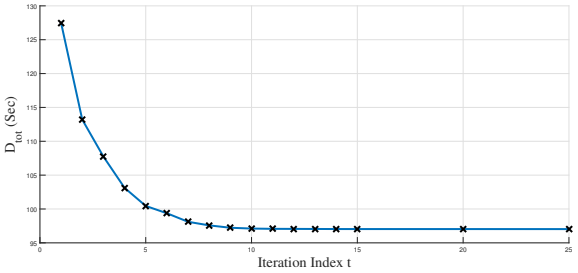


Fig. 2: The convergence of the proposed iterative algorithm in terms of total latency of the network versus the number of iterations t.

VI. Simulation Results

The simulation results are given to evaluate the performance of the proposed algorithm. The data center is positioned at the center of the area and there are 4 BSs located with the coordinates (+3,+3), (-3,+3), (-3,-3) and (+3,-3) Kilometers. Moreover, we suppose that 6 MUs are uniformly spread in the coverage area of each cell which are 200×200 m^2 square areas. The mean value of the Rayleigh distribution for all channel fadings are set to be 0.8 which are modeled as independent and identically distributed (i.i.d.) exponentially distribution. In addition, the pathloss exponent of the pathloss model is set to be 2. We also assume that the bandwidth of the access and backhaul links are set to be $W_{\text{Ac}} = 5$ MHz and $W_{\text{BH}} = 2.5$ MHz, respectively. The bandwidth of each subcarrier is also assumed to be $W_s = 19.53125$ KHz. Hence, the number of access and backhaul subcarriers are given by $N_{\text{Ac}} = 256$ and $N_{\text{BH}} = 128$, respectively. The noise power is assemed to be $N_0 = -174$ dBm/Hz. We assume that $C = 50$ social media exist in the library of the data center which are requested from MUs in the network. The size of social media are modeled as Log-normal distribution, where $\mu_s = 0.7$ and $\sigma_s^2 = 0.5$, and the social media sizes are in terms of Mbit (Mb). The Zipf parameter is also set to be $\zeta_1 = 0.6$ for the popularity distribution of social media. We assume that each BS b has $m_b = 1.875$ Mbyte (MB) storage capacity. Moreover, the maximum transmit power of each BS and the data center are assumed to be $P_b^{\max} = 15$ Watts and $P_0^{\max} = 25$ Watts, respectively. We also set $\tau_{b,u}^{\text{Ac}} = \tau_{0,b}^{\text{BH}} = 300$ Seconds (Sec) for all access and backhaul links. The system parameters are summarized in Table II.

Fig. 2 investigates the convergence of the proposed iterative Alg. 1. As shown, the proposed iterative algorithm converges to a fixed point after 8 iterations. This result ensures us that the proposed algorithm can be applied in a multi-cell scenario.

We also investigate the total latency in the network versus different storage capacities of the BSs which varies from $m_b = 0$ to $m_b = 7$ MB in Fig. 3(a). Note that $m_b = 0$ represents no caching scheme is considered for the network. We also show that when the storage capacity of the BSs increases, the total latency of the network decreases, because more social media can be stored and therefore, more social media are closed to the end-users and consequently, D_{tot} decreases. Specifically, we first formulate the total access traffic and sum data rate of

TABLE I: COMPUTATIONAL COMPLEXITY OF THE PROPOSED ITERATIVE ALGORITHMS

Iterative Algorithm	Content Placement	Joint Subcarrier and Power Allocation
Alg. 1	$\frac{\log\left((2B+2U)/t^0_\varrho\right)}{\log \zeta_2}$	$O\left(\Phi_1 \frac{2UN_{\text{Ac}}+(2N_{\text{Ac}}+\Upsilon_1)+(2U+B)}{\omega_2^2} + \Phi_1 \frac{BN_{\text{BH}}+N_{\text{BH}}+(B+1)}{\omega_3^2}\right)$

TABLE II: SYSTEM PARAMETERS

Parameter	Value
Number of BSs	$B = 4$
Number of MUs in each cell	$U_b = 6$
Distance between data center and each BS	$3\sqrt{2}$ Km
coverage area of each BS	200×200 m^2
Mean of Rayleigh distributions	0.8
Path-loss exponent	2
PSD of AWGN noise	$N_0 = -174$ dBm/Hz
Bandwidth of access links	$W_{\text{Ac}} = 5$ MHz
Bandwidth of backhaul links	$W_{\text{BH}} = 2.5$ MHz
Bandwidth of each subcarrier	$W_{\text{S}} = 19.53125$ KHz
Number of access subcarriers	$N_{\text{Ac}} = 256$
Number of backhaul subcarriers	$N_{\text{BH}} = 128$
Number of social media	$C = 50$
Size of contents	$\mu_s = 0.7$, $\sigma_s^2 = 0.5$
Zipf parameter	$\zeta_1 = 0.6$
Storage capacity of each BS	$m_b = 1.875$ MB
Maximum transmit power of each BS	$P_b^{\max} = 15$ Watts
Maximum transmit power of data center	$P_0^{\max} = 25$ Watts
Maximum deadline of access links	$\tau_{b,u}^{\text{Ac}} = 300$ Sec
Maximum deadline of backhaul links	$\tau_{0,b}^{\text{BH}} = 300$ Sec

access links in the *first case*, respectively as:

$$f_{\text{Ac},1} = \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} f_{b,u}^{\text{Ac},1}, \tag{70}$$

and

$$r_{\text{Ac},1} = \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} w_{b,u}^{\text{Ac},1,\text{tot}}, \tag{71}$$

and subsequently for the *second case*, we have

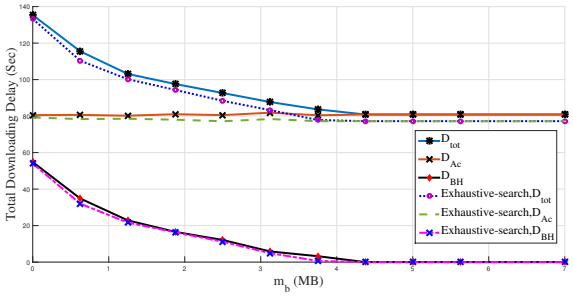
$$f_{\text{Ac},2} = \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} f_{b,u}^{\text{Ac},2}, \tag{72}$$

and

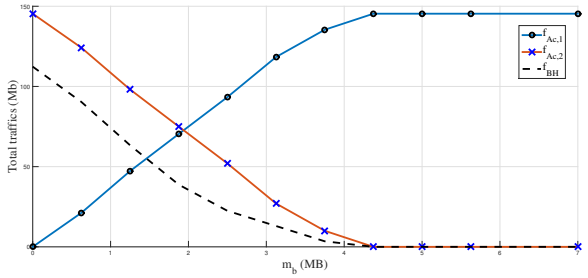
$$r_{\text{Ac},2} = \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} w_{b,u}^{\text{Ac},2,\text{tot}}. \tag{73}$$

According to Figures 3(b) and 3(c), with increasing m_b, both the total access traffic $f_{\text{Ac},1}$ and sum data rate $r_{\text{Ac},1}$ increase, while for the *second case*, both the access traffic $f_{\text{Ac},2}$ and sum data rate $r_{\text{Ac},2}$ decrease, because more social media lead to be stored in the *first case* and accordingly, more radio resources should be allocated to the access links in the first case which increases the sum data rate $r_{\text{Ac},1}$. In addition, we show there is a trade-off between the first case and the second case for allocating both storage capacity and radio resources. Moreover, we formulate the total backhaul data traffic and sum data rate of backhaul links, respectively, as:

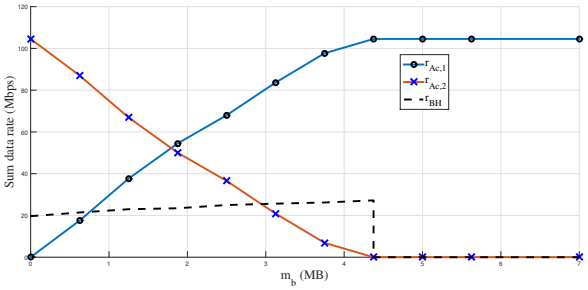
$$f_{\text{BH}} = \sum_{b=1}^{B} f_{0,b}^{\text{BH}}, \tag{74}$$



(a) The total latency over the storage capacity.



(b) The total traffic over the storage capacity.



(c) The sum data rate over the storage capacity.

Fig. 3: The total latency of the network versus storage capacity. We show that increasing the storage capacity of the BSs, decreases the latency of the network.

and

$$r_{\text{BH}} = \sum_{b=1}^{B} r_{0,b}^{\text{BH}}. \tag{75}$$

We also show that the total data traffic of the backhaul links decreases which causes a decreasing in D_{tot}. It can be seen that increasing m_b does not significantly affect the total access latency which is formulated as $D_{\text{Ac}} = \sum_{b=1}^{B} \sum_{u \in \mathcal{U}_b} \left(D_{b,u}^{\text{Ac},1} + D_{b,u}^{\text{Ac},2}\right)$, but also affects the total backhaul latency $D_{\text{BH}} = \sum_{b=1}^{B} D_{0,b}^{\text{BH}}$, seriously, in which the total backhaul traffic f_{BH} decreases and subsequently, decreasing f_{BH} extends the feasible area of the constraint (20b) which causes increasing the sum data rate of the backhaul links r_{BH}. Note that increasing $f_{\text{Ac},1}$ shrinks the feasible area of the constraint (20c) and subsequently,

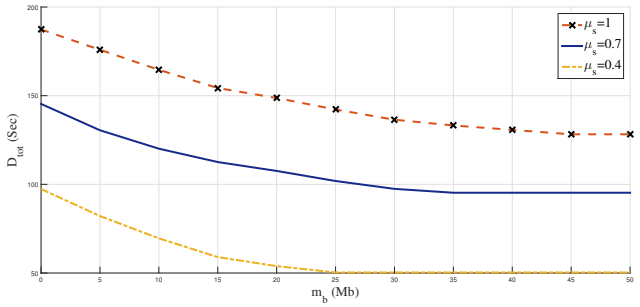


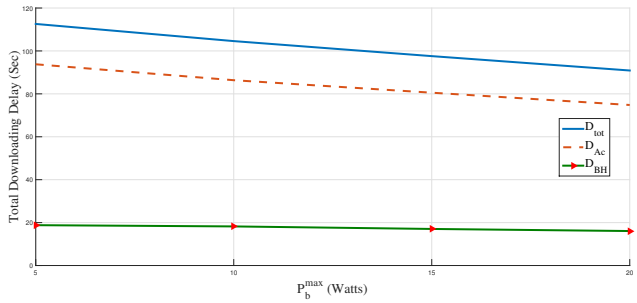
Fig. 4: The total latency over the storage capacity of BSs for different values of μ_s as 0.4, 0.7 and 1.

decreasing $f_{\text{Ac},2}$ extends the feasible area of the constraint (20d) for allocating radio resources. On the other hand, the trade-off of the radio resources between the first and second cases leads us to allocate more radio resources to the *first case* and subsequently less radio resources to the *second case*. Moreover, it can be seen that when the backhaul traffic tends to zero, D_{tot} tends to D_{Ac}. In Fig. 3(a), it is shown that the proposed iterative algorithm has a near-optimal solution.

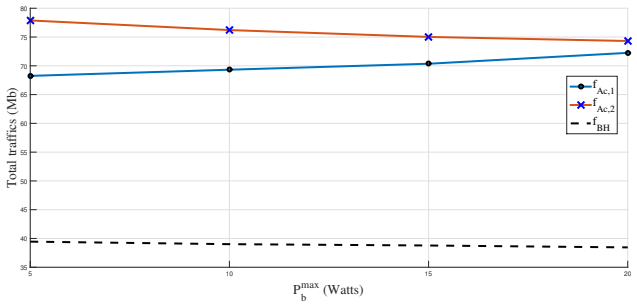
We also evaluate the total latency of the network for different size distributions of social media, in Fig. 4. We note that increasing μ_s from 0.4 to 1 increases the size of social media, inherently. The data traffics are related to requests of MUs, size of social media, and the place of them in the caches. Hence, increasing μ_s increases the data traffics and consequently, D_{tot} increases. Note that more increasing μ_s shrinks the feasible area of (20b), (20c), (20d) and (20g) and makes (20) infeasible.

We investigate the effect of the maximum allowable transmit power of BSs on the total latency D_{tot} by increasing $P_b^{\max}$ from 5 to 20 Watts and obtaining D_{tot} for each scenario, in Fig. 5. As we see, increasing $P_b^{\max}$ increases the sum data rate $r_{\text{Ac},1}$ and $r_{\text{Ac},2}$, and then, more social media can be stored at the cache of BSs, because of extending the feasible area of constraint (20c) in (26). Hence, the total data traffic of the *first case*, i.e., $f_{\text{Ac},1}$, increases and because of existing of the storage resources trade-off, $f_{\text{Ac},2}$ decreases and subsequently, f_{BH} decreases. Since more transmit powers are allocated to the data rates of both the first case and the second case, D_{Ac} decreases. Moreover, the backhaul sum data rate r_{BH} is also increased, smoothly, because of decreasing f_{BH} and subsequently, the feasible area of (20b) in (27) is extended.

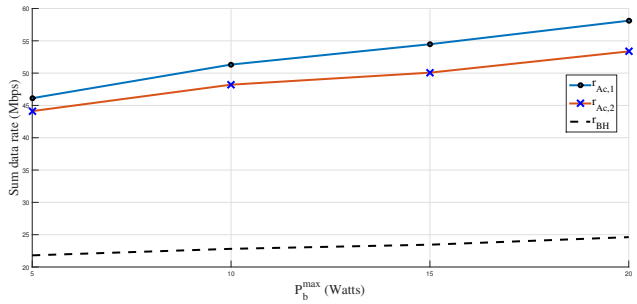
The effect of maximum transmit power of the data center on D_{tot} is also investigated in Fig. 6, by increasing the maximum allowable transmit power of the data center from $P_0^{\max} = 20$ to $P_0^{\max} = 35$ Watts. We note that increasing $P_0^{\max}$ decreases D_{BH}, because of increasing r_{BH}. Moreover, increasing $P_0^{\max}$ does not affect D_{Ac}, seriously.



(a) The total latency over the maximum transmit power of each BS $P_b^{\max}$.



(b) The total traffic over the maximum transmit power of each BS $P_b^{\max}$.



(c) The sum data rate over the maximum transmit power of each BS $P_b^{\max}$.

Fig. 5: The total latency of the network versus $P_b^{\max}$.

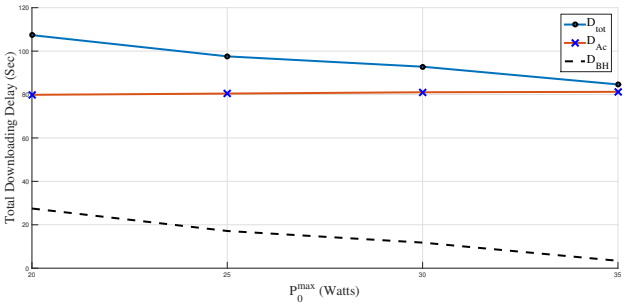


Fig. 6: The total latency over the maximum transmit power of the data center.

VII. Conclusion

In this paper, we considered a cache-enabled multiuser OFDMA-based cellular network consisting of several BSs communicate with a data center. We aimed to design a RA algorithm to minimize total latency of MUs under delivery deadline constraints. Hence, we obtained a near-optimal caching policy, as well as a delivery policy, where joint access and backhaul radio resources are optimized. In the proposed algorithm, we proposed an iterative algorithm, where the main optimization problem is divided into two subproblems. The first subproblem is a content placement problem and the second subproblem is a radio resource allocation problem. Via simulating, we evaluated the performance of the proposed low-complex iterative algorithm and compared it with the

exhaustive-search method.

APPENDIX A
PROOF OF PROPOSITION 1

After finding the solution $\boldsymbol{\rho} = \boldsymbol{\rho}_{t_1+1}$ for a given $(\boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1})$, we have $D_{\text{tot}}(\boldsymbol{\rho}_{t_1+1}, \boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1}) \leq D_{\text{tot}}(\boldsymbol{\rho}_{t_1}, \boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1})$ due to the fact that for a given feasible solution $(\boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1})$, solving the BLP problem (26) by optimization toolboxes, minimizes the objective function (26a) after each iteration and after that, we have $D_{\text{tot}}(\boldsymbol{\rho}_{t_1+1}, \boldsymbol{p}_{t_1+1}, \boldsymbol{\gamma}_{t_1+1}) \leq D_{\text{tot}}(\boldsymbol{\rho}_{t_1+1}, \boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1})$ which is proved in *Proposition 2*. Accordingly, we can show that

$$0 \leq D_{\text{tot}}(\boldsymbol{\rho}_{\text{opt}}, \boldsymbol{p}_{\text{opt}}, \boldsymbol{\gamma}_{\text{opt}}) \leq \cdots \leq D_{\text{tot}}(\boldsymbol{\rho}_{t_1+1}, \boldsymbol{p}_{t_1+1}, \boldsymbol{\gamma}_{t_1+1}) \leq D_{\text{tot}}(\boldsymbol{\rho}_{t_1+1}, \boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1}) \leq D_{\text{tot}}(\boldsymbol{\rho}_{t_1}, \boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1}) \leq \ldots, \quad (76)$$

where, $\boldsymbol{\rho}_{\text{opt}}$, $\boldsymbol{p}_{\text{opt}}$ and $\boldsymbol{\gamma}_{\text{opt}}$ are the final solution which is considered as a suboptimal solution. According to the above, after each iteration, the duality gap $D_{\text{Gap}} = D_{\text{tot}}(\boldsymbol{\rho}_{t_1}, \boldsymbol{p}_{t_1}, \boldsymbol{\gamma}_{t_1}) - D_{\text{tot}}(\boldsymbol{\rho}_{\text{opt}}, \boldsymbol{p}_{\text{opt}}, \boldsymbol{\gamma}_{\text{opt}})$ decreases and guarantees the convergence of the proposed iterative approach in Alg. 1. The optimization problem (20) is also lower-bounded by zero due to constraints (20e), (20f) and (20g).

APPENDIX B
PROOF OF PROPOSITION 3

In order to solve the inner problems (36) and (37) using the dual method, we should investigate the duality gaps exist between sub-optimal and optimal solutions. Accordingly, we investigate a condition which is called time-sharing property [28] for both subproblems (36) and (37). According to [28] and [29], as the number of subcarriers tends to infinity, the solution of the dual method tends to global-optimal solution. Hence, the time sharing property holds in the inner problems (36) and (37) and the duality gaps tend to zero as the number of access and backhaul subcarriers go to infinity, respectively in (36) and (37).

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," San Jose, CA, USA, White Paper, Feb. 2014.

[2] Ericsson. (2015). *Ericsson Mobility Report* [Online]. Available: http://www.ericsson.com/mobility-report

[3] H. Hsu and K. Chen, "A Resource Allocation Perspective on Caching to Achieve Low Latency," *IEEE Commun. Letters*, vol. 20, no. 1, pp. 145–148, Jan. 2016.

[4] 3GPP. (2015). *3GPP Specification Series* [Online]. Available: http://www.3gpp.org/dynareport/36-series.htm

[5] I. V. Loumiotis, E. F. Adamopoulou, K. P. Demestichas, T. A. Stamatiadi, and M. E. Theologou, "Dynamic backhaul resource allocation: An evolutionary game theoretic approach," *IEEE Trans. on Commun.*, vol. 62, no. 2, pp. 691-698, Feb. 2014.

[6] A. Biagioni, R. Fantacci, D. Marabissi, and D. Tarchi, "Adaptive Subcarrier Allocation Schemes for Wireless OFDMA Systems in WiMAX Networks," *IEEE J. on Sel. Areas in Commun.*, vol. 27, no. 2, pp. 217–225, Feb. 2009.

[7] R. Kwan, C. Leung, and J. Zhang. 2009. "Resource allocation in an LTE cellular communication system," *In Proceedings of the 2009 IEEE international conference on Communications* (ICC'09). IEEE Press, Piscataway, NJ, USA, 3915–3919.

[8] M. Moretti, A. Todini and A. Baiocchi, "Distributed radio resource allocation for the downlink of multi-cell OFDMA radio systems," *Teletraffic Congress (ITC), 2010 22nd International.* IEEE, 2010

[9] H. Shahrokh Shahraki, K. Mohamed-pour, L. Vangelista, "Sum capacity maximization for MIMO-OFDMA based cognitive radio networks," *Physical Commun.*, Vol. 10, March 2014, Pages 106–115, ISSN 1874–4907, http://dx.doi.org/10.1016/j.phycom.2012.10.002.

[10] T. Wang and L. Vandendorpe, "Iterative Resource Allocation for Maximizing Weighted Sum Min-Rate in Downlink Cellular OFDMA Systems," *IEEE Trans. on Signal Processing*, vol. 59, no. 1, pp. 223–234, Jan. 2011.

[11] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Y. D. Kim, "Fairness-aware radio resource management in downlink OFDMA cellular relay networks," *IEEE Trans. on Wireless Commun.*, vol. 9, no. 5, pp. 1628–1639, May 2010.

[12] D. Trong Ngo, S. Khakurel, T. Le-Ngoc, "Joint Subchannel Assignment and Power Allocation for OFDMA Femtocell Networks," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 1, pp. 342–355, Jan. 2014.

[13] D. W. K. Ng, E. S. Lo, R. Schober, "Energy-Efficient Resource Allocation in Multi-Cell OFDMA Systems with Limited Backhaul Capacity." *IEEE Trans. on Wireless Commun.*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.

[14] D. W. K. Ng, E. S. Lo, R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas." *IEEE Trans. on Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.

[15] D. W. K. Ng, E. S. Lo, R. Schober, "Energy-efficient resource allocation for secure OFDMA systems." *IEEE Trans. on Veh. Tech.*, vol. 61, no. 6, pp. 2572–2585, 2012.

[16] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic and L. Hanzo, "Distributed Caching for Data Dissemination in the Downlink of Heterogeneous Networks," *IEEE Trans. on Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[17] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," arXiv preprint arXiv:1601.07322 (2016).

[18] Z. Ming, M. Xu, D. Wang, "InCan: In-network cache assisted eNodeB caching mechanism in 4G LTE networks," *Computer Networks*, vol. 75, Part A, Dec. 2014, pp. 367–380.

[19] Y. Cui, D. Jiang and Y. Wu, "Analysis and Optimization of Caching and Multicasting in Large-Scale Cache-Enabled Wireless Networks," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.

[20] S. Krishnan, M. Afshang, and H. S. Dhillon, "Effect of Retransmissions on Optimal Caching in Cache-enabled Small Cell Networks," available online: http://arxiv.org/abs/1606.03971 (2016).

[21] C. E. Shannon. 2001. "A mathematical theory of communication.," *ACM SIGMOBILE Mobile Computing and Commun. Review* vol. 5, no. 1, pp. 3–55, Jan. 2001.

[22] "CVX Research: "CVX: Matlab software for disciplined convex programming, version 2.1". Available at http://cvxr.com/cvx, accessed 4 Oct. 2015.

[23] Y. J. Zhang, L. Qian, and J. Huang, "Monotonic optimization in communication and networking systems.," *Found Trends ®Networking*, vol. 7, no. 1, pp. 1-75, 2013.

[24] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, pp. 492-498, Mar. 1967.

[25] N. Mokari, F. Alavi, S. Parsaeefard, and T. Le-Ngoc, "Limited-Feedback Resource Allocation in Heterogeneous Cellular Networks," *IEEE Trans. on Veh. Tech.*, vol. 65, no. 4, pp. 2509–2521, Apr. 2016.

[26] S. Boyd, and L. Vandenberghe, *Convex optimization* Cambridge, U.K.: Cambridge university press, 2004.

[27] N. Mokari, M. R. Javan, and K. Navaie, "Cross-Layer Resource Allocation in OFDMA Systems for Heterogeneous Traffic With Imperfect CSI," *IEEE Trans. on Veh. Tech.*, vol. 59, no. 2, pp. 1011–1017, Feb. 2010.

[28] W. Yu, L. Raymond, "Dual methods for nonconvex spectrum optimization of multicarrier systems." *IEEE Trans. on Commun.*, vol. 54, no. 7, pp. 1310-1322, Jul. 2006.

[29] M. R. Abedi, N. Mokari, M. R. Javan, and H. Yanikomeroglu, "Secure Communication in OFDMA Based Cognitive Radio Networks: An Incentivized Secondary Network Coexistence Approach," *IEEE Trans. on Veh. Tech.*, vol. 66, no. 2, pp. 1171–1185, Feb. 2017.