Resource Allocation for Content Delivery in Cache-enabled OFDMA Small Cell Networks

Xiuhua Li[1], Xiaofei Wang[2], Keqiu Li[2], Hongjun Chi[3], and Victor C. M. Leung[1]
[1]Dept. Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada
[2]Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin, China.
[3]Shandong Academy of Sciences, Shangdong, China.
Email: {lixiuhua, vleung}@ece.ubc.ca, xiaofeiwang@tju.edu.cn, {likeqiu, hjchi2016}@gmail.com

***Abstract*—To deal with explosively growing demands for multimedia contents from mobile users, content caching in base stations has been considered as an effective solution to improve the network performance by, e.g., offloading network traffic and improving users' Quality of Service (QoS). Moreover, the proactive caching policy in a cache-enabled system needs to be optimized taking into account of content delivery by wireless transmissions. Thus, in this paper, we investigate and propose an efficient resource allocation scheme for min-rate guaranteed content delivery in the downlink multiuser cache-enabled orthogonal frequency division multiple access small cell networks (OFDMA-SCNs). Our aim is to maximize the weighted sum of data rates in an OFDMA-SCN based on the constraints of the caching method, users' QoS, subcarrier reuse and small base stations' transmit power. We employ the alternating direction multiplier method to decompose the formulated complex nonconvex optimization problem into a series of simpler subproblems for which optimal solutions can be easily obtained, and propose corresponding low-complexity methods to solve the subproblems and then the whole problem. Numerical results demonstrate the effectiveness of the proposed resource allocation scheme.**

I. Introduction

With modern lifestyle, especially in the increasing popularity of online social communities, demands for multimedia contents (e.g., video, photos and audio) from mobile users are sustaining explosive growth [1], [2]. Satisfying these demands cost effectively is becoming a big challenge for mobile network operators (MNOs), which is heightened by the scarcity of resources especially in the radio access networks (RANs) and backhaul networks. To overcome the limitations of current generation mobile networking technologies [3]–[8], innovations in mobile networking technologies including new network architectures and advanced data transmission techniques [1]–[8] would be needed to effectively support the increasing network traffic load for content delivery without degrading mobile users' Quality of Service (QoS).

Deploying caches at the edges of mobile networks has recently emerged as an effective technique to offload network traffic and improve users' QoS by satisfying users' content requests locally [1]–[7]. There have been many studies focusing on designing content caching strategies in mobile networks. For instance, collaborative multi-cell caching in [1], [5], [6] and FemtoCaching in [9] were proposed to cache popular contents at base stations (BSs) in small cells to offload network traffic and increase the number of served users. Besides, the studies in [3], [4] proposed the concept of Caching-as-a-Service (CaaS), a caching virtualization framework along with the development of Cloud-based RANs (C-RANs) and the virtualization of Evolved Packet Core, aiming to offload network traffic. The studies in [2], [7], [10] proposed collaborative BS caching schemes to enhance users' QoS especially on access delay. However, most of these studies do not take into account the last mile of content delivery via wireless transmissions from BSs to users.

In this paper, we aim to explore the joint design of content delivery by wireless transmissions. Generally, contents can be stored in BS caches for a long period since their popularity changes slowly, while scheduling wireless transmissions of contents requires knowledge of mobile users' instantaneous channel state information (CSI) and is inherently a short-time process. Thus, when designing wireless content delivery schemes to achieve the potential performance gains of content caching and improve network capacity, we can assume that the states of the caches are static during the transmissions. Moreover, the corresponding resource allocation plays an essential role in the scheme design. There are only a few studies on designing resource allocation schemes in cache-enabled systems. For instance, the pricing and resource allocation scheme in [11] used stochastic geometry optimization to maximize the profit of a video caching system for small cells. Resource allocation schemes for software defined networking, caching and computing were proposed in [12], [13] to minimize the system costs. However, wireless content delivery has not been taken into account in [11]–[13]. Multicast beamforming schemes in [14], [15] were proposed for content delivery from BSs to users through wireless links with given caching methods. However, the study in [14] focused on the theoretical analysis of system performance without considering the detailed design of resource allocation schemes for real-time content delivery satisfying users' QoS requirements. On the other hand, the study in [15] did not consider resource allocation based on orthogonal frequency-division multiple access (OFDMA), which is widely used in contemporary wireless access networks. A resource allocation

978-1-5090-5935-5/17/$31.00 ©2017 IEEE

scheme was proposed in [16] for cache-enabled OFDMA C-RANs to minimize the total transmit power, but this work does not consider the limit of maximum transmit power of each BS. Thus resource allocation for wireless content delivery in OFDMA cache-enabled mobile networks is still not well explored.

To fill the gap, this paper focusses on optimal resource allocation for content delivery in the downlink multiuser cache-enabled OFDMA small cell networks (OFDMA-SCNs), where the objective is maximizing the weighted sum of data rates in an OFDMA-SCN based on the constraints of caching policy, users' QoS, subcarrier reuse and small base stations' (SBSs') transmit power. To solve the formulated complex nonconvex optimization problem, we employ alternating direction method of multipliers (ADMM) [17]–[19] to decompose it into a series of simpler subproblems for which optimal solutions can be easily obtained, and propose low-complexity methods to solve the subproblems as well as the whole problem. Numerical results demonstrate the effectiveness of the proposed scheme.

The rest of this paper is organized as follows. In Section II, we describe the system model and then formulate the optimal resource allocation problem. In Section III, we decompose the nonconvex optimization problem into a series of subproblems and solve these subproblems. We also propose low-complexity solutions for the subproblems and whole problem. Numerical results are shown to evaluate the performance of the proposed resource allocation scheme in Section IV. Finally, Section V concludes this paper.

II. System Model and Problem Formulation

A. Downlink OFDMA-SCN Model

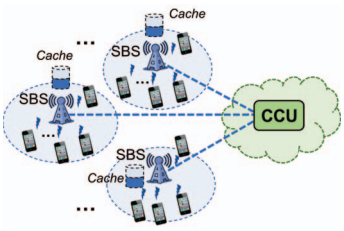


Fig. 1. An illustration of a cache-enabled OFDMA-SCN model.

As illustrated in Fig. 1, we consider the downlink transmissions of a cache-enabled OFDMA-SCN with a cloud central unit (CCU), M single-antenna SBSs (denoted by $\mathcal{M} = \{1, 2, \ldots, M\}$), N subcarriers (denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$), and K active single-antenna mobile users (denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$) accessing contents. All SBSs are connected to the CCU via backhaul links with large but limited capacity, while the CCU can perform all computations in the network. There are F contents (denoted by $\mathcal{F} = \{1, 2, \ldots, F\}$) in the network, and each SBS can only cache a limited number of contents. Each user requests a content based on the content popularity and can be served by multiple SBSs. The N subcarriers are orthogonal and have an identical bandwidth of B_s. We assume interference between adjacent cells can be avoided even with the maximum subcarrier reuse factor of 1. Besides, we assume that the CSI and users' content requests are available at the CCU before content delivery. The downlink channel is slotted, and resource allocation decisions are made on a slot-by-slot basis over much shorter time intervals than those of state changes in content caching, which are relatively static and known to the CCU.

Denote $x_m^f \in \{0, 1\}$ for whether content f is cached at SBS$_m$ or not. Denote $y_k^f \in \{0, 1\}$ for whether content f is requested by user k or not, and each user accesses only one content in a time slot, i.e., $\sum_{f=1}^{F} y_k^f = 1, \forall k \in \mathcal{K}$. Denote $\mathcal{S}_k = \{m | x_m^f = y_k^f = 1, m \in \mathcal{M}, f \in \mathcal{F}\}, k \in \mathcal{K}$ for the sets of SBSs that cache the requested contents of users, $\mathcal{K}_1 = \{k | \mathcal{S}_k \neq \emptyset, k \in \mathcal{K}\}$ for the set of users whose requested contents are locally available, and $\mathcal{K}_0 = \mathcal{K} \setminus \mathcal{K}_1$ for the set of users whose requested contents are not cached and need to be downloaded via backhaul links. To offload backhaul traffic and reduce network costs, each user in the set $\mathcal{K}_1$ is required to be associated with at least one the SBSs that cache the requested content while each user in the set $\mathcal{K}_0$ can be associated with any SBS in the network. Denote $h_{k,m,n}$ and $p_{k,m,n}$ for the complex channel gain (consisting of large-scale fading and small-scale fading) and transmit power from SBS$_m$ to user k on subcarrier n. Denote $\delta_{k,n} \in \{0, 1\}$ and $\delta_{k,m,n} \in \{0, 1\}$ for whether subcarrier n is allocated to user k in the network and from SBS$_m$ or not, respectively. Here, $\{\delta_{k,n}\}$ satisfies $\sum_{k=1}^{K} \delta_{k,n} \leq 1, \forall n \in \mathcal{N}$, while $\delta_{k,m,n} = 0, \forall k \in \mathcal{K}_1, \forall m \in \mathcal{M} \setminus \mathcal{S}_k, \forall n \in \mathcal{N}$ holds. Based on our previous work [19], the mathematical relationship between $\delta_{k,m,n}$ and $p_{k,m,n}$ satisfies

$$\delta_{k,m,n} = \text{sign}(p_{k,m,n}) \text{ and } \delta_{k,m,n} p_{k,m,n} = p_{k,m,n} \tag{1}$$

where $\text{sign}(x) \triangleq \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \end{cases} (x \geq 0)$ is the step function. Besides, the relationship between $\delta_{k,n}$ and $\delta_{k,m,n}$ satisfies

$$\delta_{k,n} = \max_{m \in \mathcal{M}} \{\delta_{k,m,n}\} = \text{sign}\Big(\sum_{m=1}^{M} p_{k,m,n} \Big). \tag{2}$$

Thus, (1) and (2) indicate that the joint user association and subcarrier-power allocation can be equally transformed into the power allocation.

By considering (1) and allowing multiple SBSs to transmit to one user in a coordinated fashion, e.g., using the technique of maximum ratio transmission (MRT) [19], the signal-to-noise ratio (SNR) of user k on subcarrier n is expressed as

$$\rho_{k,n} = \frac{\sum_{m=1}^{M} p_{k,m,n} \mid h_{k,m,n} \mid^2}{\sigma_N^2}, \ \forall k \in \mathcal{K}, \forall n \in \mathcal{N} \tag{3}$$

where σ_N^2 denotes the power of the zero-mean additive white Gaussian noise (AWGN) at the receiver input. Thus, we can get the channel capacity of user k on subcarrier n as

$$r_{k,n} = B_s \log_2(1 + \rho_{k,n}), \forall k \in \mathcal{K}, \forall n \in \mathcal{N}. \tag{4}$$

Then we can get the overall data rate of user k as

$$R_k = \sum_{n=1}^{N} r_{k,n}, \quad \forall k \in \mathcal{K}. \tag{5}$$

B. Problem Formulation

In this paper, our objective is to maximize the weighted sum of data rates in an OFDMA-SCN based on joint user association and subcarrier-power allocation for min-rate guaranteed content delivery. The overall optimization problem is formulated as

$$\max_{\mathbf{P}\in\mathbb{R}^{K\times M\times N}} \ \lambda \sum_{k\in\mathcal{K}_1} \omega_k R_k + \sum_{k\in\mathcal{K}_0} \omega_k R_k \tag{6a}$$

$$s.t.\ p_{k,m,n} \geq 0, \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{6b}$$

$$\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,m,n} \leq p_m^{\max}, \quad \forall m \in \mathcal{M}, \tag{6c}$$

$$R_k \geq C_k^{\min}, \forall k \in \mathcal{K}, \tag{6d}$$

$$\delta_{k,m,n} = \text{sign}(p_{k,m,n}), \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{6e}$$

$$\delta_{k,m,n} = 0, p_{k,m,n} = 0, \ \forall k \in \mathcal{K}_1, \forall m \in \mathcal{M} \setminus \mathcal{S}_k, \forall n \in \mathcal{N}, \tag{6f}$$

$$\delta_{k,n} = \max_{m\in\mathcal{M}} \{\delta_{k,m,n}\}, \quad \forall\, k \in \mathcal{K}, \forall\, n \in \mathcal{N}, \tag{6g}$$

$$\sum_{k=1}^{K} \delta_{k,n} \leq 1, \quad \forall\, n \in \mathcal{N} \tag{6h}$$

where λ and ω_k are weighting factors denoting the network priority of the users whose requested contents are locally cached and the individual priority of user k, respectively; $\mathbf{P} = \{p_{k,m,n}\}^{K\times M\times N}$; $p_m^{\max}$ and $C_k^{\min}$ denote the maximum transmit power of SBS m and the required minimum data rate of user k, respectively. Define $F(\mathbf{P}) = -\lambda \sum_{k\in\mathcal{K}_1} \omega_k R_k - \sum_{k\in\mathcal{K}_0} \omega_k R_k$. Besides, the power constraints in (6b) and (6c) imply that

$$0 \leq p_{k,m,n} \leq p_m^{\max}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{7}$$

which is useful in the algorithm design [20]. Clearly, the problem in (6) is a mixed 0-1 nonconvex optimization problem and thus is NP-hard. Denote $\mathcal{P}$ as the feasible solution set of the problem in (6).

III. Solutions to the Optimization Problem

A. ADMM-based Decomposition

To solve the complex optimization problem in (6), we aim to provide a suboptimal solution and also employ the ADMM used in [17]–[19] to decompose the complex problem into subproblems that are simpler to solve. Thus, based on the idea of ADMM, we divide the constraints in (6) into two groups and define two sets as

$$\mathcal{S}_P = \left\{\mathbf{P} \in \mathbb{R}^{K\times M\times N} \middle|\ (6c), (6d), (6f) \text{ and } (7)\right\}, \tag{8}$$

$$\mathcal{S}_Q = \left\{\mathbf{P} \in \mathbb{R}^{K\times M\times N} \middle|\ (6e) - (6h) \text{ and } (7)\right\}. \tag{9}$$

Clearly, the set $\mathcal{S}_P$, which satisfies the constraints of users' required minimum data rates and SBSs' maximum transmit power, is convex. $\mathcal{S}_Q$, which aims at satisfying the constraints of user association and subcarrier allocation, is discrete. As a result, we have the feasible solution set $\mathcal{P} = \mathcal{S}_P \cap \mathcal{S}_Q$, i.e., $\mathbf{P} \in \mathcal{S}_P$ and $\mathbf{P} \in \mathcal{S}_Q$. Then the problem in (6) becomes

$$\min_{\mathbf{P}\in\mathcal{S}_\mathbf{P}, \mathbf{Q}\in\mathcal{S}_\mathbf{Q}} \ F(\mathbf{P}) \tag{10a}$$

$$s.t.\ \mathbf{P} = \mathbf{Q} \tag{10b}$$

where $\mathbf{Q} \in \mathbb{R}^{K\times M\times N}$ and $\mathbf{Z} \in \mathbb{R}^{K\times M\times N}$ are introduced variable matrices. Thus, the problem in (6) is equally transformed to the problem as shown in (10) with one equality constraint.

Then the problem in (10) can be turned into a minimization problem by introducing the augmented Lagrangian function as

$$\mathcal{L}(\mathbf{P}, \mathbf{Q}, \mathbf{L}, \theta) = F(\mathbf{P}) + \langle \mathbf{P} - \mathbf{Q}, \mathbf{L}\rangle + \frac{\theta}{2}(\|\mathbf{P} - \mathbf{Q}\|_2^2) \tag{11}$$

where $\mathbf{L} \in \mathbb{R}^{K\times M\times N}$ is the Lagrange multiplier matrix for constraints (10b); $\theta > 0$ is a quadratic penalty scalar; $\langle \mathbf{x}, \mathbf{y}\rangle$ denotes the sum of all the elements of $\mathbf{x} \circ \mathbf{y}$ and $\circ$ denotes the Hadamard product.

By using ADMM-based decomposition, the joint optimization problem w.r.t. the augmented Lagrangian function in (11) can be decomposed into the following three subproblems:

1) Subproblem 1: Optimization of $\mathbf{P}$ under fixed $\mathbf{Q}$, $\mathbf{L}$ and θ, which can be formulated as

$$\min_{\mathbf{P}\in\mathcal{S}_\mathbf{P}} \ F(\mathbf{P}) + \frac{\theta}{2}\|\mathbf{P} - \mathbf{C}_P\|_2^2 \tag{12}$$

where $\mathbf{C}_P = \mathbf{Q} - \frac{\mathbf{L}}{\theta}$ is a constant matrix w.r.t. $\mathbf{P}$.

2) Subproblem 2: Optimization of $\mathbf{Q}$ under fixed $\mathbf{L}$, and θ, which can be formulated as

$$\min_{\mathbf{Q}\in\mathcal{S}_\mathbf{Q}} \|\mathbf{Q} - \mathbf{C}_Q\|_2^2 \tag{13}$$

where $\mathbf{C}_Q \triangleq \mathbf{P}^* + \frac{\mathbf{L}}{\theta}$ is a constant matrix w.r.t. $\mathbf{Q}$, and $\mathbf{P}^*$ is the optimal solution of (12).

3) Subproblem 3: Updating of $\mathbf{L}$ and θ with given $(\mathbf{P}^*, \mathbf{Q}^*)$, where $\mathbf{Q}^*$ is the optimal solution of (13).

B. Solutions to Subproblems

After the ADMM-based decomposition, Subproblem 1 is convex and its optimal solution can be obtained using standard optimization techniques (e.g., subgradient method or interior point method) [21].

Subproblem 2 is nonconvex with binary variables, and can be further divided into N subproblems and solved in parallel using the proposed distributed search method as shown in Algorithm 1 to arrive at the optimal solution.

With $\mathbf{P}$ and $\mathbf{Q}$ achieved by respectively solving Subproblem 1 and Subproblem 2, Subproblem 3 aims to update the multiplier $\mathbf{L}$ and the quadratic penalty scalar θ. With ADMM, they can be updated in each step as

$$\mathbf{L}^{(\tau+1)} = \mathbf{L}^{(\tau)} + \theta^{(\tau)}\big(\mathbf{P}^{(\tau+1)} - \mathbf{Q}^{(\tau+1)}\big), \tag{14}$$

$$\theta^{(\tau+1)} = \min\big\{\theta^{\max}, \Delta \cdot \theta^{(\tau)}\big\} \tag{15}$$

where τ denotes the iteration index, $\theta^{\max}$ is a relatively large scalar and $\Delta > 1$ is a scalar. The ADMM process for solving the whole problem in (10) is described in Algorithm

Algorithm 1 Distributed Search Algorithm for Solving Subproblem 2 w.r.t. $\mathbf{Q}$.

Input: $\mathbf{P}$, $\mathbf{L}$, θ, $(p_m^{\max})_{M\times 1}$.
Initialize $\mathbf{Q} = \{q_{k,m,n}\}^{K\times M\times N} = \mathbf{0}_{K\times M\times N}$, $\mathbf{T} = \{t_{k,m,n}\}^{K\times M\times N} = \mathbf{0}_{K\times M\times N}$.
Calculate $\mathbf{C}_Q$.
for $n = 1$ to N **do**
  for $k = 1$ to K **do**
    if $k \in \mathcal{K}_1$ **then**
      Set $t_{k,m,n} = \min\{\left[\left(\mathbf{C}_Q\right)_{k,m,n}\right]^+, p_m^{\max}\}$ for $\forall m \in \mathcal{S}_k$.
    else
      Set $t_{k,m,n} = \min\{\left[\left(\mathbf{C}_Q\right)_{k,m,n}\right]^+, p_m^{\max}\}$ for $\forall m \in \mathcal{M}$.
    end if
  end for
  Given n, find $k_n^* = \arg\min_{k\in\mathcal{K}}\{\sum_{m=1}^{M}\left(t_{k,m,n} - \mathbf{C}_Q\right)_{k,m,n}\right)^2\}$.
  if k_n^* not unique **then**
    Select one randomly.
  end if
  Set $q_{k_n^*,m,n} \leftarrow t_{k_n^*,m,n}$ for $\forall m \in \mathcal{M}$.
end for
Output: $\mathbf{Q}$.

2. Given the multiplier $\mathbf{L}$ and the quadratic penalty scalar θ, Algorithm 2 updates $\mathbf{P}$ and $\mathbf{Q}$ by solving Subproblem 1 and Subproblem 2, respectively. Besides, ADMM converges to the corresponding suboptimal solution of the problem in (10) [17], [22]. Let ε be the convergence precision of Algorithm 2.

Algorithm 2 ADMM for Solving the whole Problem in (10).

Input: σ_N^2, $(h_{k,m,n})_{K\times M\times N}$, $(p_m^{\max})_{M\times 1}$, B_s, $(C_k^{\min})_{K\times 1}$.
Initialize $\tau = 0$, $\mathbf{P}^{(0)} = \mathbf{0}_{K\times M\times N}$, $\mathbf{Q}^{(0)} = \mathbf{0}_{K\times M\times N}$, $\mathbf{L}^{(0)} = 0.02 \times \mathbf{1}_{K\times M\times N}$, $\theta^{(0)} = 10^{-3}$, $\theta^{\max} = 10^6$, $\Delta = 1.2$, $\varepsilon = 10^{-4}$.
while not converge **do**
  Update $\mathbf{P}^{(\tau)}$ by solving Subproblem 1.
  Update $\mathbf{Q}^{(\tau)}$ by solving Subproblem 2.
  Set $\tau \leftarrow \tau + 1$.
  Update $\mathbf{L}^{(\tau)}$ and $\theta^{(\tau)}$ according to (14) and (15), respectively.
  Check the convergence condition: $\|\mathbf{P}^{(\tau)} - \mathbf{Q}^{(\tau)}\|_\infty \leq \varepsilon$.
end while
Output: $\mathbf{P}$.

Moreover, for the initialization of the multiplier $\mathbf{L}$, it can be set randomly. For the scalars θ, $\theta^{\max}$ and Δ, θ is usually initialized as a small value, e.g., $\theta^{(0)} = 10^{-4}$; $\theta^{\max}$ is set as a relatively large value, e.g., $\theta^{\max} = 10^4$, which is usually not reached when the algorithm converges; Δ is set neither too small nor too large based on the algorithm's convergence rate, e.g., $\Delta = 1.2$. Most importantly, using different feasible initializations, ADMM always converges but requires various numbers of iterations to converge, the complexity and convergence analysis of which can be referred to [17], [22].

IV. Numerical Results

In this section, we evaluate the performance of our proposed joint user association and subcarrier-power allocation scheme for content delivery in an OFDMA-SCN via Monte-Carlo simulations. The whole service area is a circle with a radius of 500 meters, and fully covered by five small cells, i.e., $M = 5$. The five SBSs, each with a radius of 250 meters and the same maximum transmit power (i.e., $p_m^{\max} \equiv p^{\max}, \forall\, m \in \mathcal{M}$), are uniformly distributed in a circle, while the users are uniformly distributed in the whole area. Based on [19], we set the system bandwidth B to 2.5 MHz, subcarrier number N to 128, carrier center frequency to 2.5 GHz, path loss exponent to 3.7, lognormal shadowing standard deviation to 8 dB, and noise power density to -174 dBm/Hz, respectively. The random channel fluctuations are modeled as Rayleigh fading with unit average power. For illustration purpose, we set the content number F to 2,000 and each SBS can only cache 10% of the contents, i.e., 200 contents. Each active user is set to have the same individual priority (i.e., $\omega_k = 1, \forall k \in \mathcal{K}$) and randomly requests only one of the available contents at a time slot. Besides, we set the network priority $\lambda \in \{1, 5, 10\}$, and assume that around 60% of the users requested contents are locally cached. The required minimum data rate (i.e., $\{C_k^{\min}\}$) for delivering a content to a user is set as 128 Kbps. All the simulation results are averaged over 100 random channel realizations. Note that based on the above settings, when the network priority λ is chosen as 1, the proposed scheme is reduced to the general scheme that maximizes the total data rate based on the same considered constraints.

Fig. 2 illustrates the convergence performance evaluation of the proposed ADMM in Algorithm 2. For illustration purpose, we set the user number and the maximum transmit power $(K, p^{\max}) = (10, 25\text{ dBm}), (20, 30\text{ dBm})$. As shown in Fig. 2, in all the settings, Algorithm 2 takes at most 120 iterations to converge. Specifically, all the weighted sum of data rates (the considered optimization objective) decreases significantly in [20, 90] iterations and then gradually converges to satisfy the given convergence precision. Besides, we can observe that at the beginning of the iterative procedure of Algorithm 2, the weighted sum of data rates may be relatively large since the achieved transmit power matrix $\mathbf{P}$ only satisfies some of the considered constraints, but Algorithm 2 converges to a local optimum at the end to satisfy all the considered constraints.

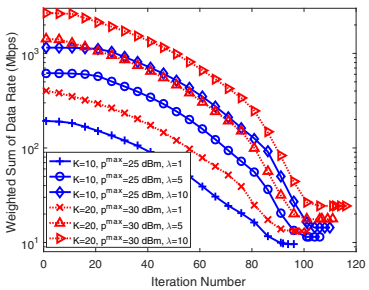
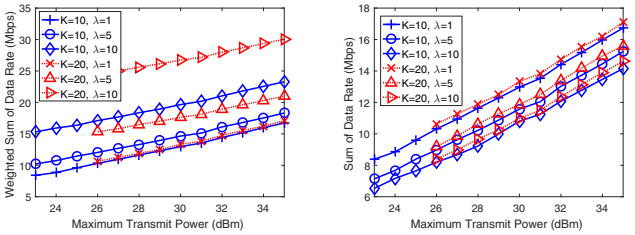


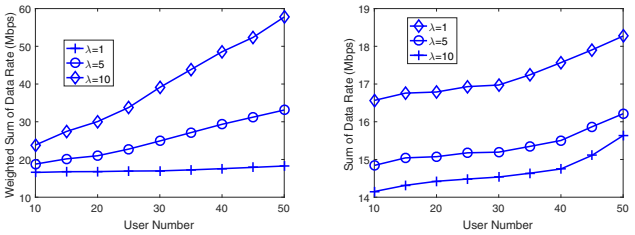
Fig. 2. Weighted sum of data rates versus iteration number.

Fig. 3 compares the weighted sum of data rates and sum of data rates versus maximum transmit power in different settings. From Fig. 3, as the maximum transmit power increases, all the weighted sums of data rates and sums of data rates also go up. Specifically, a greater value of the chosen λ leads to



(a) Weighted sum of data rate (b) Sum of data rate

Fig. 3. Data rate versus maximum transmit power.



(a) Weighted sum of data rate (b) Sum of data rate

Fig. 4. Data rate versus number of users when $p^{\max} = 35$ dBm.

a greater weighted sum of data rates but smaller sum of data rate, which can be explained by the fact that, mathematically, maximizing the objective $\lambda \sum_{k\in\mathcal{K}_1} \omega_k R_k + \sum_{k\in\mathcal{K}_0} \omega_k R_k$ is equivalent to maximizing $\sum_{k\in\mathcal{K}_1} \omega_k R_k + \frac{1}{\lambda} \sum_{k\in\mathcal{K}_0} \omega_k R_k$, and that more resources need to be allocated to the users whose requested contents are locally cached at SBSs. Besides, a greater number of users leads to increases in weighted sum of data rates and sum of data rates due to the multiuser diversity gain.

Fig. 4 compares the weighted sum of data rates and sum of data rates versus number of users in different settings. From Fig. 4, all the weighted sum of data rates and sum of data rates increase with increasing number of users because of the multiuser diversity gain. Moreover, a greater value of the chosen λ also leads to a greater weighted sum of data rates but smaller sum of data rate.

V. Conclusions

In this paper, we have proposed a joint user association and subcarrier-power allocation scheme for min-rate guaranteed content delivery in the downlink of a multiuser cache-enabled OFDMA-SCN with some practical considerations, employed ADMM to split the complex nonconvex optimization problem into a series of simpler subproblems for which optimal solutions can be easily achieved, and proposed the corresponding methods to solve the subproblems as well as the whole problem with low complexity to realize a design that is attractive for practical implementation. Numerical results have shown that the proposed algorithm has good convergence performance and can effectively achieve the performance gain in weighted sum of data rates as well as sum of data rates, which is important in designing cache-enabled wireless communication systems.

Acknowledgment

This work was supported in part by National NSFC through Grants No. 61671088 and No. 61271182, a CSC Four Year Doctoral Fellowship, and grants from Canadian NSERC.

References

[1] X. Wang, X. Li, V. C. M. Leung, and P. Nasiopoulos, "A framework of cooperative cell caching for the future mobile networks," *in Proc. HICSS*, pp. 5404-5413, Jan. 2015.

[2] X. Li, X. Wang, S. Xiao, and V. C. M. Leung, "Delay performance analysis of cooperative cell caching in future mobile networks," *in Proc. IEEE ICC*, pp. 5652-5657, Jun. 2015.

[3] X. Li, X. Wang, C. Zhu, W. Cai, and V. C. M. Leung, "Caching-as-a-Service: virtual caching framework in the cloud-based mobile networks," *in Proc. IEEE INFOCOM Workshops*, pp. 372-377, May 2015.

[4] X. Li, X. Wang, K. Li, and V. C. M. Leung, "CaaS: caching as a service for 5G networks," *IEEE Access*, vol. 5, pp. 5982-5993, May 2017.

[5] X. Li, X. Wang, and V. C. M. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," *in Proc. IEEE ICC*, pp. 1-6, May 2016.

[6] X. Li, X. Wang, K. Li, and V. C. M. Leung, "Collaborative hierarchical caching for traffic offloading in heterogeneous networks," *in Proc. IEEE ICC*, May 2017.

[7] X. Li, P. Wu, X. Wang, K. Li, Z. Han, and V. C. M. Leung, "Collaborative hierarchical caching in cloud radio access networks," *in Proc. IEEE INFOCOM Workshops*, May 2017.

[8] X. Ge, X. Li, H. Jin, J. Cheng, and V. C. M. Leung, "Joint User Association and Scheduling for Load Balancing in Heterogeneous Networks," *in Proc. IEEE GLOBECOM*, pp. 1-6, Dec. 2016.

[9] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: wireless video content delivery through distributed caching helpers," *in Proc. IEEE INFOCOM*, Mar. 2012.

[10] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: an energy-efficient approach to improve Quality of Service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207-1221, May 2016.

[11] J. Li, H. Chen, Y. Chen, Z, et al., "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115-2129, Aug. 2016.

[12] Q. Chen, F.R. Yu, T. Huang, R. Xie, J. Liu, and Y. Liu, "Joint resource allocation for software defined networking, caching and computing," *in Proc. IEEE GLOBECOM*, pp. 1-6, Dec. 2016.

[13] Y. Jin, Y. Wen, and C. Westphal, "Towards joint resource allocation and routing to optimize video distribution over future Internet," *in Proc. IFIP Networking*, pp. 1-9, May 2015.

[14] A. Liu and V. K. N. Liu, "Exploiting base station caching in MIMO cellular networks: opportunistic cooperation for video streaming," *IEEE Trans. Signal Processing*, vol. 63, no. 1, pp. 57-69, Jan. 2015.

[15] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118-6131, Sept. 2016.

[16] R.G. Stephen, and R. Zhang, "Green OFDMA resource allocation in cache-enabled CRAN," *in Proc. IEEE OnlineGreenComm*, Dec. 2016.

[17] S. Boyd, N. Parikh, et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp.1-122, 2011.

[18] C. Shen, T. H. Chang, K. Y. Wang, Z. Qiu, and C. Y. Chi, "Distributed robust multicell coordinated beamforming with imperfect csi: An admm approach," *IEEE Trans. on Signal Processing*, vol. 60, no. 6, pp. 2988-3003, Jun. 2012.

[19] X. Li, X. Ge, X. Wang, J. Cheng, and V. C. M. Leung, "Energy efficiency optimization: joint antenna-subcarrier-power allocation in OFDM-DASs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7470-7483, Nov. 2016.

[20] X. Li and V. C. M. Leung, "Optimizing power allocation in wireless networks: are the implicit constraints really redundant?" *in Proc. Ad Hoc Networks*, Sept. 2016.

[21] S. Boyd and L. Vandenbergh. *Convex Optimization*. Cambridge University Press, 2004.

[22] M. Hong, Z. Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," Oct. 2014. http://arxiv.org/abs/1410.1390