Cache Placement in Two-Tier HetNets with Limited Storage Capacity: Cache or Buffer?

Zhaohui Yang, Cunhua Pan, Yijin Pan, Yongpeng Wu, *Senior Member, IEEE*, Wei Xu, *Senior Member, IEEE*, Mohammad Shikh-Bahaei, *Senior Member, IEEE*, and Ming Chen

***Abstract*—In this paper, we aim to minimize the average file transmission delay via bandwidth allocation and cache placement in two-tier heterogeneous networks with limited storage capacity, which consists of cache capacity and buffer capacity. For average delay minimization problem with fixed bandwidth allocation, although this problem is nonconvex, the optimal solution is obtained in closed form by comparing all locally optimal solutions calculated from solving the Karush-Kuhn-Tucker conditions. To jointly optimize bandwidth allocation and cache placement, the optimal bandwidth allocation is first derived and then substituted into the original problem. The structure of the optimal caching strategy is presented, which shows that it is optimal to cache the files with high popularity instead of the files with big size. Based on this optimal structure, we propose an iterative algorithm with low complexity to obtain a suboptimal solution, where the closed-from expression is obtained in each step. Numerical results show the superiority of our solution compared to the conventional cache strategy without considering cache and buffer tradeoff in terms of delay.**

***Index Terms*—Caching policy, heterogeneous networks, cache and buffer, bandwidth allocation.**

I. Introduction

To accommodate the growing demand for high data rate transmission and seamless coverage in wireless communications, heterogeneous deployment has been proposed as an effective network architecture [1], [2]. In heterogeneous networks (HetNets), small base stations (BSs) are deployed to offload the traffic in high user density area. To further improve the transmission rate and decrease latency for users, wireless caching is a promising solution by caching popular contents at the network edge [3]–[9]. The recent contributions about cache-aided wireless networks can be classified into two categories: analyzing the content delivery performance and designing cache placement strategies.

This work was supported in part by the Engineering and Physical Science Research Council under grant EP/P003486/1, grant EP/N029666/1 and grant EP/N029720/1, in part by the National Nature Science Foundation of China under Grant 61471114, Grant 61372106 and Grant 61701198, and in part by the Six Talent Peaks project in Jiangsu Province under GDZB-005 (*Corresponding author: Cunhua Pan*).

Z. Yang and M. Shikh-Bahaei are with Centre for Telecommunications Research, King's College London, London, UK, (Emails: {yang.zhaohui, m.sbahaei}@kcl.ac.uk).

Y. Pan, W. Xu, and M. Chen are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 211111, China, (Emails: {panyijin, wxu, chenming}@seu.edu.cn).

C. Pan is with School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK, (Email: c.pan@qmul.ac.uk).

Y. Wu is with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, Minhang 200240, China, (Email: yongpeng.wu@sjtu.edu.cn).

Content delivery performance analysis is crucial in reveling the benefits of distributed cache placement in cache-enabled networks [10]–[15]. The throughput-outage tradeoff was investigated in [10] for one-hop device-to-device (D2D) networks, which showed that the user throughput is proportional to the fraction of cached information. For multi-hop D2D networks, the multi-hop capacity scaling laws were investigated in [11]. It was further shown in [11] that a multi-hop transmission provides a significant throughput gain over one-hop direct transmission for Ad-Hoc networks with cached users. For cache-enabled cellular networks with coordinated D2D communications, cellular and D2D coverage probabilities were derived in [12]. In [13]–[15], multicast beamforming was investigated for cache-enabled content-centric networks.

Cache placement strategies should be properly designed due to the features of link connectivity and channel quality [16]–[18]. There are mainly two issues addressed in cache placement problems: the hit probability maximization [19]–[22] and the average delay minimization [23]–[27]. The hit probability is defined as the probability that a user will find the file he/she is asking for in the cache of the BS he/she is covered from [28]. Considering both small BS caching and cooperation in a downlink HetNet, the authors in [19] first derived a tractable expression for the hit probability by using stochastic geometry, and then optimized the caching distribution to maximize this hit probability. In [20], the structure of the optimal content-placement policies to maximize the hit probability for HetNets was investigated. The impact of file preference and user willingness on hit probability was investigated in [21] by optimizing the cache placement strategy. Instead of maximizing hit probability, the authors in [23] analyzed the optimal way of assigning files to the small BSs to minimize the average transmission delay. Joint caching and user association was considered in [24] to minimize the delay for satisfying the transmission demands in cached-enabled HetNets with wireless backhaul. By exploiting user preference and spatial locality, the authors in [25] investigated the optimal cache policy to minimize the average file download time in HetNets. In user-centric networks, the delay-optimal cooperative edge caching was investigated in [26]. Reference [23]–[25] all assumed equal length for all files, which limited the conclusions for optimal cache placement strategies. Due to limited storage, some files are always not cached in the edge nodes and the delivery of uncached files constitutes a performance bottleneck caused by the buffer, which serves as a short-term memory to temporally store the data. However, the above contributions [19]–[26] all ignored the effect of

buffer. Considering the maximal buffer capacity constraint, the authors in [27] investigated the average delay minimization in cache- and buffer-enabled relaying networks. The maximal buffer capacity is assumed to be fixed in [27], even though considering cache capacity and buffer capacity tradeoff can further improve the system performance.

In this paper, we aim to minimize the average file transmission delay through bandwidth allocation and cache placement in two-tier HetNets with limited storage capacity. Different from [27], we consider the sum cache and buffer capacity constraint to balance cache and buffer. The contributions of this paper are summarized as follows:

1) The average file transmission delay minimization problem is formulated by considering transmission delay, fronthaul delay and buffer delay, which reflect the tradeoff of cache and buffer. Specifically, the average file delay expression is derived by modeling the distribution of users as independent poisson point process (PPP).
2) For cache placement with fixed bandwidth allocation, we have successfully derived the structure of the optimal solution to this nonconvex problem by solving Karush-Kuhn-Tucker (KKT) conditions, which reflects that all locally optimal solutions can be obtained. Based on the optimal structure, the optimal cache strategy can be obtained by comparing finite potentially optimal solutions. In this case, the optimal cache strategy indicates the optimal tradeoff of cache capacity and buffer capacity.
3) For joint bandwidth allocation and cache placement, the optimal bandwidth allocation is first derived in closed form. Then, the original problem is transformed into an equivalent problem with respect to the cache strategy by substituting the optimal bandwidth allocation. To solve the equivalent nonconvex problem, an iterative algorithm with low complexity is proposed to obtain a suboptimal solution.

The rest of the paper is organized as follows. In Section II, we introduce the system model. Problem formulation and analysis are presented in Section III. Optimal cache placement with fixed bandwidth allocation and joint bandwidth and cache optimization are addressed in Section IV and Section V, respectively. Some numerical results are shown in Section VI and conclusions are finally drawn in Section VII.

II. System Model

A. System Model

Consider a cache-enabled HetNet with one macro BS and M pico BSs as shown in Fig. 1. The set of pico BSs is denoted by $\mathcal{M} \triangleq \{1, 2, \cdots, M\}$. Denote the total coverage area of the macro BS by $\mathcal{A}$, while the coverage area of pico m is denoted by $\mathcal{A}_m$. The coverage area $\mathcal{A}_m$ ($\mathcal{A}$) is a circle area centered at pico BS m (macro BS) with radius r_m (r_0). The pico BSs' coverage areas are disjoint, i.e., $\mathcal{A}_m \cap \mathcal{A}_n = \varnothing$ for $m, n \in \mathcal{M}$ and $m \neq n$. Let $\mathcal{A}_0 \triangleq \mathcal{A} \setminus \cup_{m\in\mathcal{M}} \mathcal{A}_m$ denote the set of areas only covered by the macro BS.

The macro BS and all the pico BSs share the same bandwidth W for wireless information transmission to users. The wireless fronthaul links from the macro BS to the pico BSs are assumed to be orthogonal, and denote w_m as the bandwidth allocated to the wireless fronthaul link from the macro BS to pico BS m. As a result, we have

$$\sum_{m\in\mathcal{M}} w_m \leq W. \tag{1}$$

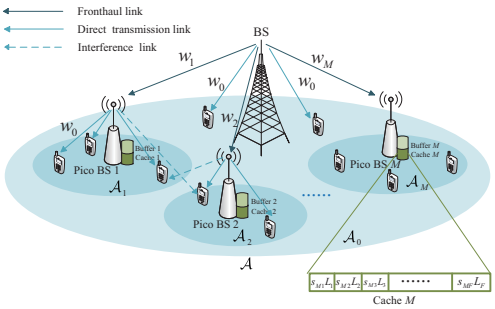

Fig. 1. System model.

Let $\mathcal{F} \triangleq \{1, \cdots, F\}$ denote the set of F files. The length of file f is denoted by $L_f > 0$ (measured in bits). The file popularity distribution $\{q_1, \cdots, q_F\}$ is assumed to be identical for all users, where $q_f \in (0, 1]$ is the popularity of file f and $\sum_{f\in\mathcal{F}} q_f = 1$. Without loss of generality, the files are sorted as $q_1 > q_2 > \cdots > q_F > 0$.

All files are stored at the macro BS, while each pico BS can only cache a subset of the total files due to limited storage capacity. Denote the storage capacity of pico BS m by C_m (measured in bits). Assume that each file is further encoded via rateless maximum distance separable codes [29]. Letting $s_{mf} \in [0, 1]$ denote the fraction of file f cached at pico BS m, we have

$$\sum_{f\in\mathcal{F}} s_{mf} L_f \leq C_m. \tag{2}$$

It is assumed that $C_m < \sum_{f\in\mathcal{F}} L_f$, i.e., each pico BS cannot cache all files. Note that the storage capacity contains both cache capacity and buffer capacity, since cache chip and buffer chip are interchangeable [30]–[33]. Cache is a long-term memory to cache popular files in a long time, while buffer is a short-term memory to temporally store the file. Consequently, the remaining part $C_m - \sum_{f\in\mathcal{F}} s_{mf} L_f$ means the buffer capacity of pico BS m.

The distribution of users in the whole area $\mathcal{A}$ is modeled as independent PPPs with density λ. The users located in $\mathcal{A}_0$ are only served by the macro BS. As shown in Fig. 2, for each file f, user i located in $\mathcal{A}_m$ covered by pico BS m first fetches s_{mf} fraction of cached file f from pico BS m. The remaining $1 - s_{mf}$ fraction of file f is delivered to pico BS m from the macro BS via the wireless fronthaul link and then relayed to user i from pico BS m with the aid of buffer.

B. Delay Model

Users served by the same pico BS are allocated with orthogonal bandwidth, and there does not exist interference

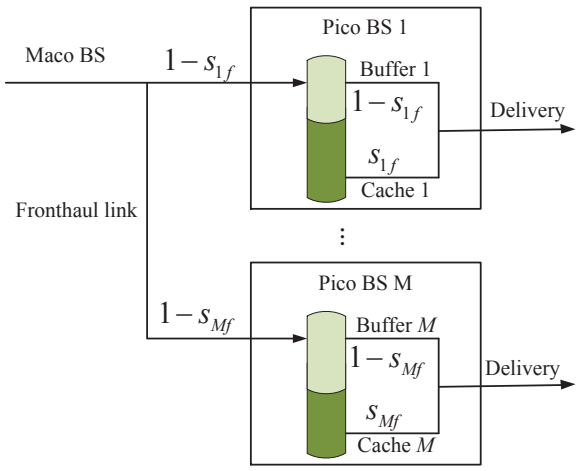

Fig. 2. The structure of cache and buffer in each pico BS.

between the users in the same pico cell. For a user i with location $\xi \in \mathcal{A}_m$ ($\mathcal{A}_0$), the file transmission rate from pico BS m (the macro BS) to user i is given by

$$R_{mi}(\xi) = w_{mi} \log_2 \left(1 + \frac{P_m |h_{mi}|^2 (d_{mi}(\xi))^{-\alpha}}{Z_{mi}}\right), \quad (3)$$

where w_{mi} is the allocated bandwidth for user i by pico BS $m \in \mathcal{M}$ (the macro BS for $m = 0$), $z_{mi} = \sigma^2 + \sum_{n \in \mathcal{M} \cup \{0\} \setminus \{m\}} P_n |h_{ni}|^2 (d_{ni}(\xi))^{-\alpha}$, σ^2 is the noise power, P_m is the transmission power of pico BS m for $m \in \mathcal{M}$ (the macro BS for $m = 0$), $|h_{mi}|^2 \sim \exp(1)$, which is a exponentially distributed random variable with unit mean, denotes the small-scale fading channel gain between user i and pico BS m for $m \in \mathcal{M}$ (the macro BS for $m = 0$), $d_{mi}(\xi)$ is the distance between user i located in ξ and pico BS m for $m \in \mathcal{M}$ (the macro BS for $m = 0$), and α is the pathloss exponent. Since users follow the same PPP distribution, equal bandwidth allocation is adopted in each BS, i.e.,

$$w_{mi} = \frac{W}{U_m}, \quad \forall m \in \mathcal{M} \cup \{0\}, \quad (4)$$

where U_m is the number of users located in $\mathcal{A}_m$. Note that Since equation (4) reflects the allocated bandwidth for the user during direct wireless information transmission, the numerator of equation (4) is W.

For fronthaul link, the file transmission rate from the macro BS to pico BS m is

$$R_m = w_m \log_2 \left(1 + \frac{P_0 |h_m|^2 d_m^{-\alpha}}{\sigma^2}\right), \quad \forall m \in \mathcal{M}, \quad (5)$$

where $|h_m|^2 \sim \exp(1)$ denotes the small-scale fading channel gain between the macro BS and pico BS m, and d_m is the distance between the macro BS and pico BS m.

When $s_{mf} < 1$, the user covered by pico BS m sequently receives separate segments of file f in two stages. In the first stage, the user accesses $s_{mf} L_f$ bits from the cache of pico BS m. In the second stage, pico BS fetches the rest uncached $(1 - s_{mf}) L_f$ bits from the macro BS via fronthaul link, and then delivers these $(1 - s_{mf}) L_f$ bits to the user. As a result, the average delay for a user located in $\mathcal{A}_m$ covered by pico BS m to download file f can be modeled as

$$\begin{aligned} d_{mf} = & \underbrace{\mathbb{E}_{U_m, h_{0i}, \cdots, h_{Mi}, \xi} \frac{s_{mf} L_f}{R_{mi}(\xi)}}_{\text{transmission delay from pico BS } m \text{ in the first stage}} \\ & + \underbrace{\mathbb{E}_{h_m} \frac{(1 - s_{mf}) L_f}{R_m}}_{\text{fronthaul delay from the macro BS in the second stage}} \\ & + \underbrace{\frac{(1 - s_{mf}) L_f}{C_m - \sum_{f \in \mathcal{F}} s_{mf} L_f} D}_{\text{buffer delay from pico BS } m \text{ in the second stage}} \\ & + \underbrace{\mathbb{E}_{U_m, h_{0i}, \cdots, h_{Mi}, \xi} \frac{(1 - s_{mf}) L_f}{R_{mi}(\xi)}}_{\text{transmission delay from pico BS } m \text{ in the second stage}}, \end{aligned} \quad (6)$$

where D is the buffer delay per time. The third term in (6) denotes the buffer time consumed at pico BS m. With capacity $\sum_{f \in \mathcal{F}} s_{mf} L_f$ allocating to cache files, the remaining buffer capacity $C_m - \sum_{f \in \mathcal{F}} s_{mf} L_f$ is only used to relay file from the macro BS to the served user. To deliver file f with capacity $(1 - s_{mf}) L_f$, the average portion of time required for delivering file f in the buffer is $\frac{(1 - s_{mf}) L_f}{C_m - \sum_{f \in \mathcal{F}} s_{mf} L_f}$. When the denominator of the third term in equation (6) equals to 0, which means that the storage capacity of the pico BS is full, it is impossible for the pico BS to fetch uncached files from the maco BS due to the fact that there is no extra capacity to store buffered files. As a result, the case when the third term in equation (6) equals to 0 will never happen in the optimal caching placement strategy. According to Little's law [34] and [35], the average delay for file f in the buffer of pico BS m, i.e., the average time that a packet is stored in the buffer, can be given by $\frac{(1 - s_{mf}) L_f}{C_m - \sum_{f \in \mathcal{F}} s_{mf} L_f} D$. The intuition behind the inverse function $\frac{(1 - s_{mf}) L_f}{C_m - \sum_{f \in \mathcal{F}} s_{mf} L_f} D$ is that small buffer leads to long queue, which results in large delay time.

From (6), it is observed that the fronthaul transmission delay decreases with s_{mf}, while the buffer delay increases with s_{mf} for $L_f < C_m - \sum_{l \in \mathcal{F} \setminus \{f\}} s_{ml} L_l$ and decreases with s_{mf} for $L_f \geq C_m - \sum_{l \in \mathcal{F} \setminus \{f\}} s_{ml} L_l$. Fig. 3 shows the delay given in (6) for the special case of one cached file in pico BS 1 versus the fraction of cached file. It can be seen that the average delay first decreases and then increases with the fraction s_{11} of cached file. This can be explained as follows. For small s_{11} (less than 0.2), more bits will be transferred from the macro BS to pico BS 1, which incurs larger fronthaul delay. When s_{11} becomes large (large than 0.3), the buffer capacity in pico BS is small, which causes large time cost for buffering.

For file f, the average delay for a user located in $\mathcal{A}_0$ served by the macro BS is

$$d_{0f} = \frac{L_f}{\mathbb{E}_{U_0, h_{0i}, \cdots, h_{Mi}, \xi} R_{0i}(\xi)}. \quad (7)$$

As a result, the average file transmission delay is given by

$$D_{\text{avg}} = \sum_{m \in \mathcal{M} \cup 0} \sum_{f \in \mathcal{F}} q_f d_{mf}. \quad (8)$$

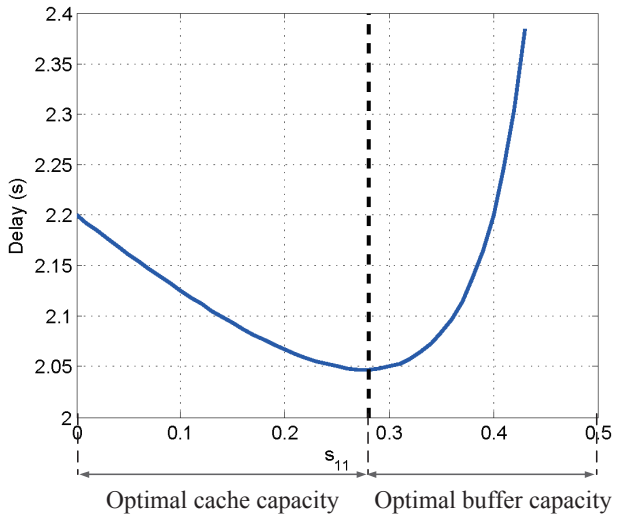

Fig. 3. Delay d_{11} versus cache placement s_{11} for pico BS 1 with $L_1 = 1$ Mbits, $C_1 = 0.5$ Mbits, $\mathbb{E}_{U_1,h_{0i},\cdots,h_{Mi},d_{1i}(\xi)} R_{1i}(\xi) = 1$ Mbps, $\mathbb{E}_{h_1} R_1 = 1$ Mbps, $D = 0.1$ s.

III. Problem Formulation and Analysis

Based on the above system model, we formulate the joint bandwidth allocation and cache placement problem to minimize the average file transmission delay. Then, we provide the expressions for average transmission rate.

A. Problem Formulation

Based on (2)-(8), the average file transmission delay optimization problem is formulated as

$$\min_{\boldsymbol{s},\boldsymbol{w}} \sum_{m\in\mathcal{M}}\sum_{f\in\mathcal{F}} q_f \left(\frac{a_m L_f}{W} + \frac{b_m(1-s_{mf})L_f}{w_m} + \frac{(1-s_{mf})L_f D}{C_m - \sum_{f\in\mathcal{F}} s_{mf} L_f} \right) + \sum_{f\in\mathcal{F}} q_f \frac{a_0 L_f}{W} \tag{9a}$$

$$\text{s.t.} \quad \sum_{f\in\mathcal{F}} s_{mf} L_f \le C_m, \quad \forall m \in \mathcal{M} \tag{9b}$$

$$\sum_{m\in\mathcal{M}} w_m \le W \tag{9c}$$

$$0 \le s_{mf} \le 1, \quad \forall m \in \mathcal{M}, f \in \mathcal{F} \tag{9d}$$

$$w_m \ge 0, \quad \forall m \in \mathcal{M}\cup\{0\}, \tag{9e}$$

where $\boldsymbol{s} = [s_{11},\cdots,s_{1F},\cdots,s_{MF}]^T$, $\boldsymbol{w} = [w_1,\cdots,w_M]^T$, $a_m = \mathbb{E}_{U_m,h_{0i},\cdots,h_{Mi},\xi} \frac{W}{R_{mi}(\xi)}$, $b_m = \mathbb{E}_{h_m} \frac{w_m}{R_m}$, $a_0 = \mathbb{E}_{U_0,h_{0i},\cdots,h_{Mi},\xi} \frac{W}{R_{0i}(\xi)}$, $\forall m \in \mathcal{M}$. From the proof of the following Lemma 1, we find that only the channel distribution information is exploited in the calculation of a_m and a_0 in (10) rather than the instantaneous channel state information. Please note that the distribution of channel varies very slowly. Constraints (9b) reflect the limitation of the storage capacity, and constraint (9c) shows that the bandwidth of the system is constrained. The optimization of cache placement $\boldsymbol{s}$ is to balance the cache capacity and buffer capacity. Increasing cached files can improve the file hit ratio, and the average delay for cached files can be reduced, while the average delay for uncached files is increased due to small buffer capacity. The optimization of bandwidth allocation takes the tradeoff between the resource and traffic demand into consideration, due to the fact that different cache placements lead to different traffic demands among BSs.

There are two difficulties to solve Problem (9). The first one is to obtain the expressions of parameters a_m and b_m, which are determined by the randomness of the number and locations of users as well as channel gains. The second one is that cache placement variable $\boldsymbol{s}$ and bandwidth allocation variable $\boldsymbol{w}$ are coupled in the objective function (9a), which makes Problem (9) a nonconvex problem.

To deal with the first difficulty, we analyze the average delay based on the exponential distribution of channel gains in the following subsection. As for the second difficulty, we obtain the optimal cache placement in closed form with fixed bandwidth allocation, and provide a low-complexity algorithm to solve the joint bandwidth allocation and cache placement problem.

B. Average Delay Analysis

Lemma 1: The average download time multiplied by bandwidth per bit of the user when downloading file from pico BS $m \in \mathcal{M}$ (or the macro BS with $m = 0$) can be obtained as

$$a_m = \lambda A_m \int_0^\infty r \int_{\xi\in\mathcal{A}_m} \exp\left(-\frac{\left(2^{\frac{1}{r}}-1\right)\sigma^2}{P_m (d_{mi}(\xi))^{-\alpha}} \right) \prod_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \frac{P_m (d_{mi}(\xi))^{-\alpha}}{\left(2^{\frac{1}{r}}-1\right) P_n (d_{ni}(\xi))^{-\alpha} + P_m (d_{mi}(\xi))^{-\alpha}}$$
$$\lambda \left(\frac{(\ln 2) 2^{\frac{1}{r}} \sigma^2}{r^2 P_m (d_{mi}(\xi))^{-\alpha}} + \sum_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \frac{(\ln 2) 2^{\frac{1}{r}} P_n (d_{ni}(\xi))^{-\alpha}}{r^2 \left(\left(2^{\frac{1}{r}}-1\right) P_n (d_{ni}(\xi))^{-\alpha} + P_m (d_{mi}(\xi))^{-\alpha} \right)} \right) \mathrm{d}\xi \, \mathrm{d}r, \tag{10}$$

where $A_m = \pi r_m^2$ for $m \in \mathcal{M}$, $A_0 = \pi \left(r_0^2 - \sum_{n\in\mathcal{M}} r_n^2 \right)$, and $\mathrm{Ei}(x) = -\int_{\infty}^{-x} \frac{\mathrm{e}^t}{t} \mathrm{d}t$ is the exponential integral function [36]. The average download time multiplied by the bandwidth per bit of pico BS $m \in \mathcal{M}$ when downloading file from the macro BS is given by

$$b_m = \int_0^\infty \frac{1}{\mathrm{e}^x \log_2\left(1 + \frac{P_0 d_m^{-\alpha}}{\sigma^2} x\right)} \mathrm{d}x \ge \frac{-\ln 2}{\mathrm{e}^{\frac{\sigma^2}{P_0 d_m^{-\alpha}}} \mathrm{Ei}\left(-\frac{\sigma^2}{P_0 d_m^{-\alpha}}\right)}. \tag{11}$$

Proof: Please refer to Appendix A. □

IV. Optimal Cache Placement with Fixed Bandwidth Allocation

Since the objective function (9a) is nonconvex with respect to $(\boldsymbol{s},\boldsymbol{w})$, it is in general hard to obtain the globally optimal solution of Problem (9). In this section, we investigate the optimization of cache placement with fixed bandwidth allocation. Given bandwidth allocation, the original Problem (9) can

be decoupled into multiple cache placement problems for M pico BSs, which fortunately have optimal solutions in closed form.

With given $\boldsymbol{w}$, Problem (9) becomes the following problem:

$$\min_{\boldsymbol{s}} \sum_{m\in\mathcal{M}}\sum_{f\in\mathcal{F}} q_f\left(\frac{b_m(1-s_{mf})L_f}{w_m}+\frac{(1-s_{mf})L_fD}{C_m-\sum_{f\in\mathcal{F}}s_{mf}L_f}\right) \tag{12a}$$

$$\text{s.t.} \sum_{f\in\mathcal{F}} s_{mf}L_f \le C_m, \quad \forall m\in\mathcal{M} \tag{12b}$$

$$0\le s_{mf}\le 1, \quad \forall m\in\mathcal{M}, f\in\mathcal{F}. \tag{12c}$$

Since Problem (12) has a decoupled objective function and decoupled constraints, Problem (12) can be decoupled into M subproblems. Subproblem m for cache optimization in pico BS m is formulated as

$$\min_{\boldsymbol{s}_m} \left(\frac{b_m}{w_m}+\frac{D}{C_m-\sum_{f\in\mathcal{F}}s_{mf}L_f}\right)\sum_{f\in\mathcal{F}} q_fL_f(1-s_{mf}) \tag{13a}$$

$$\text{s.t.} \sum_{f\in\mathcal{F}} s_{mf}L_f \le C_m \tag{13b}$$

$$0\le s_{mf}\le 1, \quad \forall f\in\mathcal{F}, \tag{13c}$$

where $\boldsymbol{s}_m=[s_{m1},\cdots,s_{mF}]^T$.

Problem (13) is nonconvex. To show this, we define function $g(x,y)=\frac{x}{C_m-L_f+x-y}$, $x,y\ge 0$, $y-x<C_m-L_f$. The Hessian matrix of $g(x,y)$ is given by

$$\begin{aligned}\bigtriangledown^2 g(x,y) &= \begin{pmatrix}\frac{\partial^2 g(x,y)}{\partial x^2} & \frac{\partial^2 g(x,y)}{\partial x\partial y}\\ \frac{\partial^2 g(x,y)}{\partial x\partial y} & \frac{\partial^2 g(x,y)}{\partial y^2}\end{pmatrix}\\ &= \frac{1}{(C_m-L_f+x-y)^3}\\ &\quad\begin{pmatrix}2(y-C_m+L_f) & C_m-L_f-x-y\\ C_m-L_f-x-y & 2x\end{pmatrix}.\end{aligned} \tag{14}$$

Since

$$|\bigtriangledown^2 g(x,y)| = -\frac{1}{(1-x-y)^3}(C_m-L_f+x-y)^2<0,$$

Hessian matrix $\bigtriangledown^2 g(x,y)$ is not positive semi-definite, i.e., function $g(x,y)$ is not a convex function with respect to (x,y). Denoting $x=L_f(1-s_{mf})$ and $y=\sum_{l\in\mathcal{F}\setminus\{f\}}s_{ml}L_l$, we can show that the objective function (13a) is not convex, i.e., Problem (13) is a nonconvex problem.

Although Problem (13) is nonconvex, the optimal cache placement in each pico BS can be obtained in closed form. First, a special structure of the optimal solution is revealed by Theorem 1. Second, this special structure shows that the optimal solution has a finite solution space, where each possibly optimal solution can be obtained according to Theorem 1. Finally, the optimal cache placement is one of finite candidate solutions with the best objective value as summarized in Theorem 2.

Theorem 1: In the optimal cache placement $\boldsymbol{s}_m^*$ of Problem (13), at most one file f has the optimal cache probability with the range $s_{mf}^*\in(0,1)$, and for all the other files, i.e., $s_{ml}^*\in\{0,1\}$, $\forall l\in\mathcal{F}\setminus\{f\}$. Besides, the optimal cache placement $\boldsymbol{s}_m^*$ satisfies $s_{m1}^*\ge s_{m2}^*\ge\cdots\ge s_{mF}^*$.

Proof: Please refer to Appendix B. □

According to Theorem 1, at most one variable can choose continuous value and other variables have binary value space due to the following two reasons. One reason is that it is optimal to store the high-popularity file with high priority independent of the file length. This is reasonable since caching files with high popularity can improve file hit probability, and the average fronthaul transmission delay and buffer delay can be reduced. The other reason is that there is a trade-off between cache capacity and buffer capacity according to equation (6). The special structure of the optimal solution indicated in Theorem 1 shows that the optimal solution of Problem (13) has a finite solution space. By comparing all these possibly optimal solutions, the optimal cache placement of Problem (13) is given in the following theorem.

Theorem 2: The optimal $\boldsymbol{s}_m^*$ of Problem (13) is one of the following $F_{m1}-F_{m2}+1$ potential solutions with the highest objective value (13a): $(\mathbf{1}_{f-1}, s_{mf}^*, \mathbf{0}_{F-f})$, $f=F_{m2},\cdots,F_{m1}$, where $\mathbf{1}_{f-1}$ and $\mathbf{0}_{F-f}$ are defined in (B.6),

$$F_{m1}=\begin{cases}\min_{f\in\mathcal{F},\sum_{l=1}^f L_l>C_m} f\\ \text{if there exits no } f\in\mathcal{F} \text{ satisfying } \sum_{l=1}^f L_l=C_m\\ \min_{f\in\mathcal{F},\sum_{l=1}^f L_l=C_m} f\\ \text{if there exits } f\in\mathcal{F} \text{ satisfying } \sum_{l=1}^f L_l=C_m\end{cases}, \tag{15}$$

$$F_{m2}=\max_{f\in\mathcal{F},C_m\ge\sum_{l=1}^f L_l+\sum_{l=f+1}^F\frac{q_lL_l}{q_f}} f, \tag{16}$$

and

$$s_{mf}^*=\begin{cases}s_{mf}^{\max}\\ \text{if } C_m\ge\sum_{l=1}^f L_l+\sum_{l=f+1}^F\frac{q_lL_l}{q_f}\\ \arg\min_{s_{mf}\in\{s_{mf}(1),s_{mf}^{\max}\}} g_{mf}(s_{mf})\\ \text{if } C_m<\sum_{l=1}^f L_l+\sum_{l=f+1}^F\frac{q_lL_l}{q_f} \text{ and } 0<s_{mf}(1)<1\\ \arg\min_{s_{mf}\in\{0,s_{mf}^{\max}\}} g_{mf}(s_{mf})\\ \text{if } C_m<\sum_{l=1}^f L_l+\sum_{l=f+1}^F\frac{q_lL_l}{q_f} \text{ and } s_{mf}(1)\le 0\\ s_{mf}^{\max}\\ \text{if } C_m<\sum_{l=1}^f L_l+\sum_{l=f+1}^F\frac{q_lL_l}{q_f} \text{ and } s_{mf}(1)>s_{mf}^{\max}\end{cases}, \tag{17}$$

with $s_{mf}^{\max}$, $g_{mf}(s_{mf})$ and $s_{mf}(1)$ defined in (C.2), (C.1a) and (C.5), respectively.

Proof: Please refer to Appendix C. □

Theorem 2 shows the impact of file popularity, file length and storage capacity on the optimal caching strategy. Furthermore, Theorem 2 points out that the optimal solution is one of the $F_{m1}-F_{m2}+1$ potential solutions.

V. Joint Bandwidth and Cache Optimization

In this section, we jointly optimize bandwidth allocation and cache placement to solve Problem (9). The optimal bandwidth allocation can be obtained in closed from by checking the KKT conditions. Based on the result of the optimal bandwidth allocation, the original Problem (9) is equivalent to a problem with only cache placement variables. To solve the equivalent

nonconvex problem, we derive one suboptimal algorithm with low complexity.

A. Optimal Bandwidth Allocation

With given $\boldsymbol{s}$, Problem (9) becomes the following problem.

$$\min_{\boldsymbol{w}} \quad \sum_{m\in\mathcal{M}} \frac{\sum_{f\in\mathcal{F}} b_m q_f L_f (1-s_{mf})}{w_m} \tag{18a}$$

$$\text{s.t.} \quad \sum_{m\in\mathcal{M}} w_m \le W \tag{18b}$$

$$w_m \ge 0, \quad \forall m \in \mathcal{M}. \tag{18c}$$

Since the objective function (18a) is convex and the constraints (18b)-(18c) are all linear, Problem (18) is a convex problem, which can be globally optimal solved via the KKT conditions. Thus, the following theorem is provided.

***Theorem** 3:* The optimal bandwidth allocation to Problem (18) is

$$w_m = \frac{\sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f (1-s_{mf})}W}{\sum_{m\in\mathcal{M}} \sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f (1-s_{mf})}}, \quad \forall m \in \mathcal{M}. \tag{19}$$

Proof: Please refer to Appendix D. □

B. Cache Optimization

Substituting the optimal bandwidth allocation (19) into Problem (9) yields

$$\min_{\boldsymbol{s}} \ \frac{1}{W}\left(\sum_{m\in\mathcal{M}} \sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f (1-s_{mf})}\right)^2 + \sum_{m\in\mathcal{M}} \frac{\sum_{f\in\mathcal{F}} q_f L_f D (1-s_{mf})}{C_m - \sum_{f\in\mathcal{F}} L_f s_{mf}} \triangleq U(\boldsymbol{s}) \tag{20a}$$

$$\text{s.t.} \quad \sum_{f\in\mathcal{F}} s_{mf} L_f \le C_m, \quad \forall m \in \mathcal{M} \tag{20b}$$

$$0 \le s_{mf} \le 1, \quad \forall m \in \mathcal{M}, f \in \mathcal{F}. \tag{20c}$$

Due to the nonconvex objective function (20a), Problem (20) is a nonconvex problem. To solve nonconvex Problem (20), we provide the following theorem to exploit the optimal structure of the optimal solution.

***Theorem** 4:* In the optimal cache placement $\boldsymbol{s}^*$ to Problem (20), for each pico BS $m \in \mathcal{M}$, at most one file f_m has the optimal cache probability with the range $s^*_{mf_m} \in (0,1)$, and for all the other files, i.e., $s^*_{ml} \in \{0,1\}$, $\forall l \in \mathcal{F}\backslash\{f\}$. Besides, the optimal cache placement $\boldsymbol{s}^*$ satisfies $s^*_{m1} \ge s^*_{m2} \ge \cdots \ge s^*_{mF}$, for all $m \in \mathcal{M}$.

Proof: Please refer to Appendix E. □

According to Theorem 4, it is optimal to store the high-popularity files with high priority as indicated from Theorem 1. Theorem 4 reflects the structure of the optimal solution of Problem (20). Since the cache strategies for different pico BSs are coupled in the objective function (20a), it is hard to obtain the globally optimal solution of nonconvex Problem (20). In the following, we propose an iterative cache placement (ICP) algorithm to solve Problem (20) in Algorithm 1.

Algorithm 1 Iterative Cache Placement (ICP)

1: Initialize $\boldsymbol{s}_1^{(0)}, \cdots, \boldsymbol{s}_M^{(0)}$. Set $k = 1$, and maximal iteration number $K_{\max}$.
2: **for** $m = 1, 2, \cdots, M$ **do**
3: Calculate the optimal $\boldsymbol{s}_m^{(k)}$ to Problem (20) with given $(\boldsymbol{s}_1^{(k)}, \cdots, \boldsymbol{s}_{m-1}^{(k)}, \boldsymbol{s}_{m+1}^{(k-1)}, \cdots, \boldsymbol{s}_M^{(k-1)})$.
4: **end for**
5: If $k > K_{\max}$ or the objective function (20a) converges, output $\boldsymbol{s}^* = (\boldsymbol{s}_1^{(k)}, \cdots, \boldsymbol{s}_M^{(k)})$, and terminate. Otherwise, set $k = k + 1$ and go to step 2.

In the ICP algorithm, to calculate the optimal $\boldsymbol{s}_m$ to Problem (20) with given cache strategies of other pico BSs, we provide the following theorem.

***Theorem** 5:* The optimal cache placement $\boldsymbol{s}^*_m = [s^*_{m1}, \cdots, s^*_{mF}]$ of Problem (20) with given $[s_{11}, \cdots, s_{(m-1)F}, s_{(m+1)1}, \cdots, s_{MF}]^T$ is one of the following F_{m1} potential solutions with the highest objective value (20a): $(\mathbf{1}_{f-1}, s^*_{mf}, \mathbf{0}_{F-f})$, $f = 1, \cdots, F_{m1}$, F_{m1} is defined in (15), and

$$s^*_{mf} = \arg \min_{s_{mf}\in\left\{0, s^{\max}_{mf}, s_{mf}(1), \cdots, s_{mf}(k)\right\}} y_{mf}(s_{mf}), \tag{21}$$

where $s^{\max}_{mf}$ is defined in (C.2), $k \in \{0,1,2,3,4,5\}$, $s_{mf}(i) = \frac{b_m \sum_{l=f}^{F} q_l L_l - (x(i))^2}{b_m q_f l_f}$, $i = 1, \cdots, k$, and $x(1), \cdots, x(k)$ are k roots in interval $(\sqrt{b_m \sum_{l=f}^{F} q_l L_l - b_m q_f L_f s^{\max}_{mf}}, \sqrt{b_m \sum_{l=f}^{F} q_l L_l})$ to equation (F.3).

Proof: Please refer to Appendix F. □

Theorem 5 indicates that the optimal $\boldsymbol{s}^*_m$ has a finite solution space with F_{m1} potential solutions, which ensures that the optimal cache placement for each pico BS can be effectively obtained by comparing finite solutions.

C. Complexity Analysis

For the ICP algorithm, the major complexity in each iteration lies in calculating the optimal $\boldsymbol{s}_m^{(k)}$ to Problem (20) with given cache placement of other pico BSs. The optimal $\boldsymbol{s}_m^{(k)}$ is one of the F_{m1} potential solutions according to Theorem 5. The complexity of obtaining each potential solution is $\mathcal{O}(T)$, where T is the complexity of calculating the roots to equation (F.3) via root-finding algorithm for polynomials [37]. According to [37], T equals to the number of iterations via the Newton's method. As a result, the total complexity of the ICP algorithm is $\mathcal{O}(KT\sum_{m\in\mathcal{M}} F_{m1})$, where K is the number of iterations of the ICP algorithm.

D. Convergence Analysis

***Theorem** 6:* Assuming $K_{\max} \to \infty$, the sequence $\boldsymbol{s}$ generated by Algorithm 1 converges.

Proof: Since the average delay (20a) is nonincreasing in each iteration according to Theorem 5, which states that the optimal cache placement of one pico BS is obtained with given

cache placement of other pico BSs, and average delay (20a) is lower-bounded by 0, the IULP algorithm must converge. □

E. Simulated Annealing Algorithm

Due to the nonconvexity of Problem (20), we propose a simulated annealing (SA) [38]–[41] based algorithm as described in Algorithm 2. SA, which is a stochastic search technique, was initially proposed in [38] to solve combinatorial problems. The main advantage of SA is the possibility to find a new optimal point after a local optimum to the objective function has been found, accepting solutions for which the objective function value is even worse than the current solution.

In Algorithm 2, the cooling rate κ takes a value in $[0.50, 0.99]$ [42], and control parameter T_0 can be determined by following the similar way as in [43]. According to [41], a new neighboring cache placement $\boldsymbol{s}^{(l)}$ of $\boldsymbol{s}^{(0)}$ can be obtained in the following two steps. In the first step, randomly choose one variable $s_{mf}^{(0)}$ in vector $\boldsymbol{s}^{(0)}$. In the second step, set $s_{nk}^{(l)} = [s_{nk}^{(0)}$ for $n \neq m$ or $n = m, k \neq f$, and update $s_{mf}^{(l)} = [s_{mf}^{(0)} \pm z s_{\text{step}}]_0^1$ with equal probability, where z is a random variable uniformly distributed in [0,1], $[a]_b^c = \min\{\max\{a, b\}, c\}$, and s_{step} is a constant step. Note that when the updated $\boldsymbol{s}^{(0)}$ does not satisfy constraint (20b), set $\boldsymbol{s}^{(l)} = \boldsymbol{s}^{(0)}$.

Algorithm 2 SA Based Cache Placement (SACP)

```
Initialize cooling rate κ, iteration number l = 0, maximal iteration number S_max, control parameter T_0, and cache placement s^(0).
while l < S_max do
    l = l + 1.
    Generate a new neighboring cache placement s^(l) from s^(0).
    Calculate δU = U(s^(l)) − U(s^(0)).
    if ΔU ≤ 0 then
        s^(0) = s^(l), U(s^(0)) = U(s^(l)).
    else
        if exp(ΔU/T_{l−1}) > random[0, 1] then
            s^(0) = s^(l), U(s^(0)) = U(s^(l)).
        end if
    end if
    Update T_l = κ T_{l−1}.
end while
```

VI. Numerical Results

In this section, numerical results are presented to evaluate the performance of the proposed ICP algorithm. In the simulations, we consider a circular macrocell with three pico BSs, i.e., $M = 3$. The macro cell has radius $r_0 = 1$ km, and the coverage area of each pico BS is a circle area with radius $r_1 = \cdots = r_M = 150$ m. The macro BS is located at the origin, and the pico BSs are located at $(-339, 741)$, $(218, -230)$, $(561, -457)$. The path-loss exponent is set as $\alpha = 3.76$, and the distribution of users is modeled as independent PPP of density $\lambda = 500$ /km^2. We assume equal file length, i.e., $L_1 = \cdots = L_F = L$, and equal storage capacity for each pico BS, i.e., $C_1 = \cdots = C_M = C$. The number of files is $F = 1000$, and the Zipf distribution is adopted to model the content popularity distribution [44]

$$q_f = \frac{1/f^{\nu}}{\sum_{l=1}^{F} 1/l^{\nu}}, \quad \forall f \in \mathcal{F}, \tag{22}$$

where $\nu \geq 0$ stands for the skewness of popularity distribution, and larger ν represents more centralized file request. Unless specified otherwise, system parameters are set as $\nu = 0.8$, $W = 10$ MHz, $D = 5$ s, and $C = 1000$ Mbits. The stopping criteria for the SA algorithm are based on the maximum number of iterations $S_{\max}$. We define $S_{\max}$ as the number of iterations after which a given minimum temperature level, i.e., $T_{S_{\max}} = 0.01 \times T_0$, can be reached. As a result, $S_{\max}$ can be calculated based on the equation $T_{S_{\max}} = (\kappa)^{S_{\max}} T_0$, where the cooling rate κ is a constant of 0.99. The step of cache variable in the SA algorithm is $s_{\text{step}} = 0.01$. The solution obtained by the ICP algorithm is set as the initial point of the SACP algorithm.

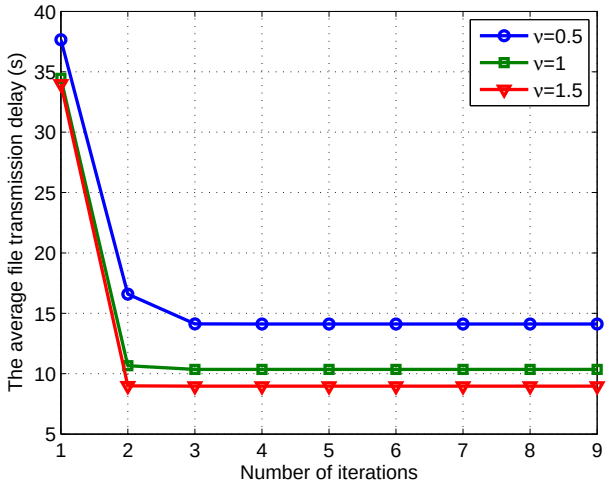

Fig. 4. Convergence behavior of the ICP algorithm under different values of parameter ν.

We compare the proposed ICP algorithm with the optimal cache placement with equal bandwidth allocation (labeled as 'OCEB') algorithm, where the optimal cache placement is obtained from Theorem 2 and $w_1 = \cdots = w_M = \frac{W}{M}$, and the optimal cache placement algorithm [25] with fixed buffer capacity (half of the storage capacity is left for buffering) and optimized bandwidth allocation obtained from the ICP (labeled as 'OCFBOB').

The convergence behavior of the ICP algorithm is illustrated in Fig. 4. From this figure, the average file transmission delay monotonically decreases and converges rapidly. Note that only two or three iterations are sufficient for the algorithm to converge, which shows the effectiveness of the proposed algorithm.

In Fig. 5, we show the average file transmission delay versus different parameters ν in (22). From Fig. 5, it is observed that the ICP outperforms OCEB and OCFBOB in terms of delay. This is because the ICP jointly optimizes bandwidth allocation

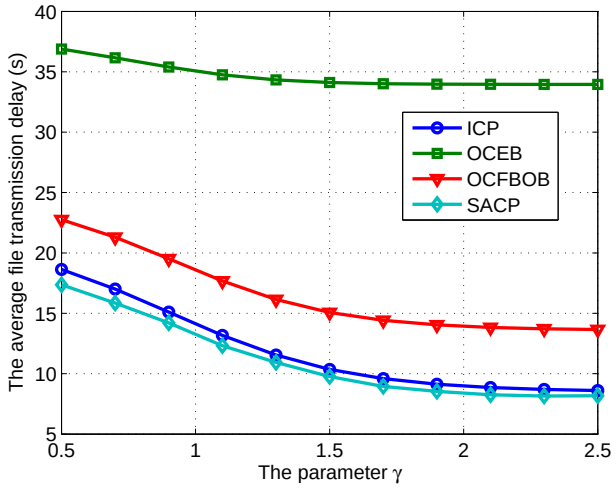

Fig. 5. The average file transmission delay versus the parameter ν.

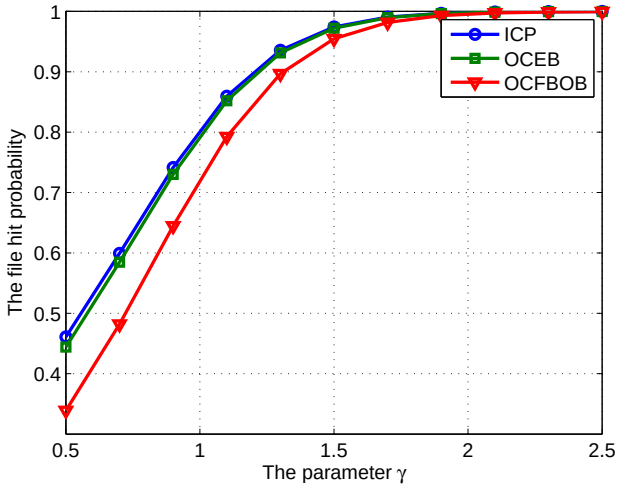

Fig. 6. The file hit probability versus the parameter ν.

and cache placement, while the OCEB only optimizes cache placement and the OCFBEB ignores the tradeoff of cache and buffer. The OCEB yields the largest delay among all algorithms, which shows that the optimization of bandwidth allocation can greatly reduce the delay. The SACP algorithm outperforms the ICP algorithm, which shows the benefits of SA for possibility finding a new optimal point after a local optimum to the objective function has been found.

The file hit ratio versus different parameters ν is depicted in Fig. 6, where the file hit ratio ρ is defined as the average successful probability that a user can fetch files from the pico BSs, i.e., $\rho = \sum_{m\in\mathcal{M}} \frac{1}{M} \sum_{f\in\mathcal{F}} q_f s_{mf}$. According to Fig. 6, the file hit probabilities of the ICP and OCEB are larger than that of the OCFBEB. This is because that the ICP and OCEB consider the tradeoff of cache and buffer to cache more files by reducing the buffer capacity, while the OCFBEB assumes fixed buffer capacity and the cache capacity cannot be further improved. From Fig. 6, we find that the file hit probability of the ICP is slightly larger than that of the OCEB, while the average delay of the ICP is significantly superior over that of the OCEB according to Fig. 5. This is because that the delay performance not only depends on the users' hit performance, but also on the transmission rate and buffer delay. Combining Fig. 5 and Fig. 6, we can conclude that the ICP achieves the best delay performance as well as the highest file hit probability through considering two tradeoffs: cache placement versus bandwidth allocation, and cache capacity versus buffer capacity, i.e., the ICP outperforms the OCFBEB in terms of delay through increasing file hit probability, and the ICP outperforms the OCEB in terms of delay through proper bandwidth allocation.

As shown in Fig. 7, we illustrate the average file transmission delay versus the total system bandwidth W. It is found that the average file transmission delay decreases with the total system bandwidth, since large bandwidth leads to large file transmission rate. It is also observed that the ICP outperforms OCEB and OCFBOB, and the delay is greatly reduced by using the ICP compared to the OCEB when the

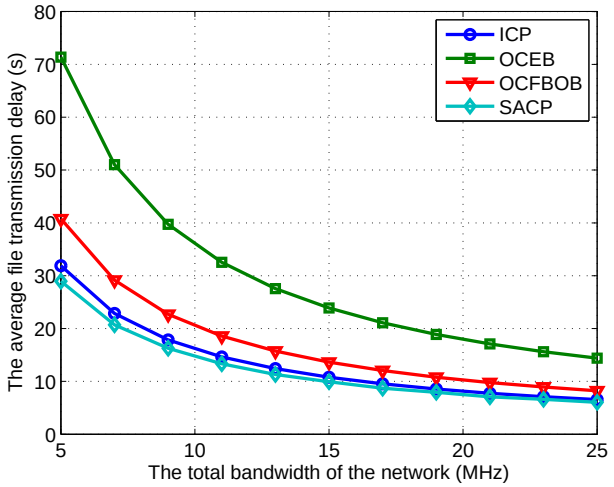

Fig. 7. The average file transmission delay versus the total system bandwidth.

total system bandwidth is small. This is due to the fact that the bandwidth is optimally allocated in the ICP, which results in good performance especially for limited system bandwidth resource.

Fig. 8 demonstrates the average file transmission delay versus the buffer delay D per time. It can be seen that the average file transmission delay increases with the buffer delay for all algorithms. It is also found that the ICP yields the best performance in terms of delay, and the delay is greatly improved by using the SACP compared to the OCFBEB for large buffer delay. The reason is that the ICP can dynamically allocate cache capacity and buffer capacity to reduce the delay based on the value of buffer delay, while the OCFBEB assumes fixed cache capacity and buffer capacity allocation.

In Fig. 9 and Fig. 10, we show the average file transmission delay and file hit probability versus different storage capacities C, respectively. According to Fig. 9, the average file transmission delay monotonically decreases with the increase of the storage capacity. The is due to the following two reasons. One

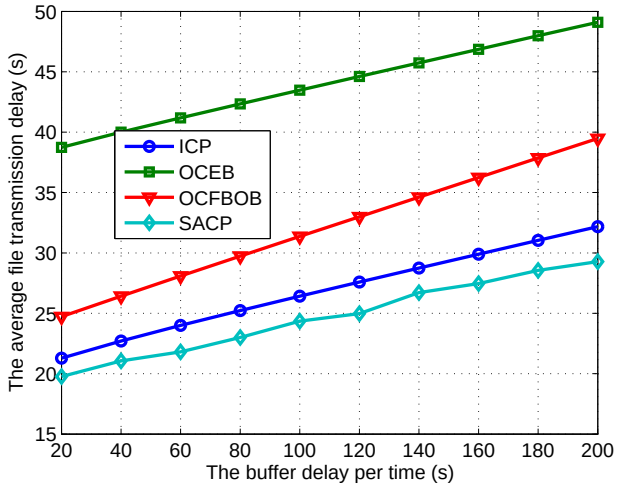

Fig. 8. The average file transmission delay versus the buffer delay per time.

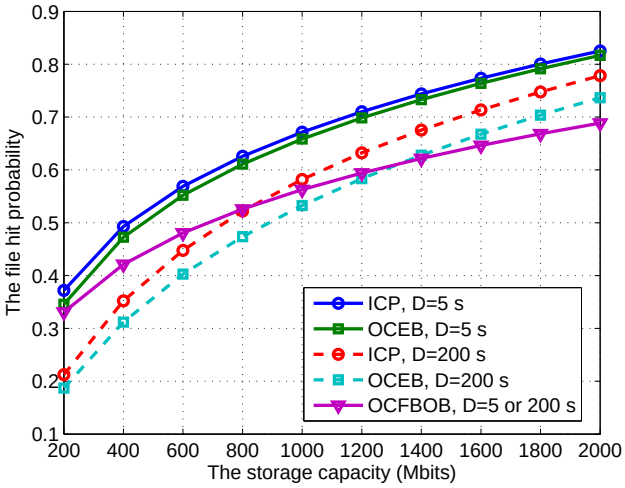

Fig. 10. The file hit probability versus the storage capacity

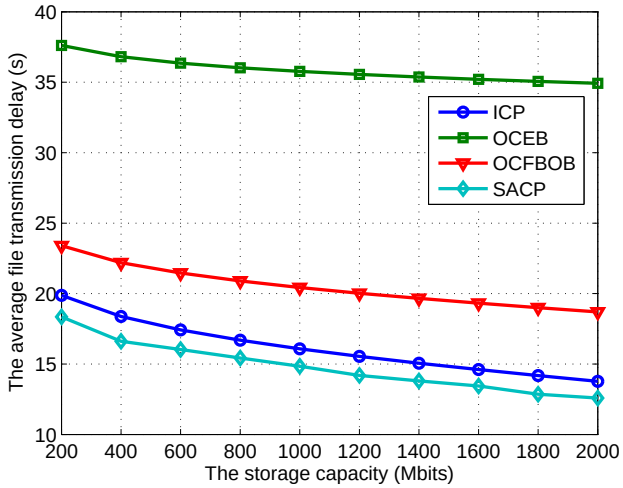

Fig. 9. The average file transmission delay versus the storage capacity.

reason is that with the increase of storage capacity, more files can be cached in the pico BSs and more users can download files directly from the cache of the pico BSs. The other reason is that larger storage capacity can lead to larger buffer capacity, which reduces the buffer time. Fig. 10 illustrates that the file hit probability increases with the storage capacity, since the pico BSs can cache more popular files for larger storage capacity. For small buffer delay D, the file hit probability of the ICP or OCEB is larger than that of the OCFBEB. This is because the average delay mainly lies in the transmission delay for small D and large cache capacity is allocated by the ICP and OCEB. For large buffer delay D and small storage capacity C, the file hit probability of the OCFBEB is superior over that of the ICP and OCEB. This is due to the fact that the buffer delay consumption dominates the transmission delay for large D and limited C, which allows to allocate more capacity to buffer by the ICP and OCEB.

VII. Conclusions

In this paper, we investigated the tradeoff of cache capacity and buffer capacity via joint bandwidth allocation and cache placement to minimize the average file transmission delay. By analyzing the KKT conditions of the nonconvex delay minimization problem, we show that it is optimal to cache the files with high popularity first rather than the files with large size. We proposed an iterative algorithm to obtain a suboptimal solution with low complexity. Through dynamically allocating cache capacity and buffer capacity, the proposed algorithm is superior over the existing caching strategy with fixed buffer capacity. It tends to allocate more cache capacity for low buffer delay per time and high storage capacity, while more capacity should be shifted to the buffer capacity for high buffer delay per time and low storage capacity. The coordination between pico BSs for cache enabled networks as our future work are left for our future work,

Appendix A
Proof of Lemma 1

According to (3) and (4), we have

$$a_m = \mathbb{E}_{U_m,h_{0i},\cdots,h_{Mi},\xi}\frac{W}{R_{mi}(\xi)} = \mathbb{E}_{U_m}U_m\mathbb{E}_{h_{0i},\cdots,h_{Mi},\xi}\frac{1}{\bar{R}_{mi}(\xi)}, \tag{A.1}$$

for all $m \in \mathcal{M}\cup\{0\}$, where

$$\bar{R}_{mi}(\xi) = \log_2\left(1+\frac{P_m|h_{mi}|^2(d_{mi}(\xi))^{-\alpha}}{\sigma^2+\sum_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}}P_n|h_{ni}|^2(d_{ni}(\xi))^{-\alpha}}\right). \tag{A.2}$$

Since users follow independent PPP with density λ, we have

$$\mathbb{P}(U_m=k) = \mathrm{e}^{-\lambda A_m}\frac{(\lambda A_m)^k}{k!}, \quad k=0,1,2,\cdots \tag{A.3}$$

for all $m\in\mathcal{M}\cup\{0\}$, where $A_m=\pi r_m^2$ for all $m\in\mathcal{M}$, and $A_0=\pi r_0^2-\sum_{m\in\mathcal{M}}r_m^2$. Based on (A.3), we can obtain

$$\mathbb{E}_{U_m}U_m = \lambda A_m. \tag{A.4}$$

$$
\begin{aligned}
&\mathbb{P}\left[\frac{1}{\bar{R}_{mi}(\xi)} < r\right] = \mathbb{P}\left[\bar{R}_{mi}(\xi) > \frac{1}{r}\right] \\
&= \mathbb{P}\left[|h_{mi}|^2 > \frac{\left(2^{\frac{1}{r}}-1\right)\left(\sigma^2 + \sum_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} P_n|h_{ni}|^2 (d_{ni}(\xi))^{-\alpha}\right)}{P_m (d_{mi}(\xi))^{-\alpha}}\right] \\
&\overset{\text{(a)}}{=} \int_{\xi\in\mathcal{A}_m} \mathbb{E}_{h_{ni},\forall n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \exp\left(-\frac{\left(2^{\frac{1}{r}}-1\right)\left(\sigma^2 + \sum_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} P_n|h_{ni}|^2 (d_{ni}(\xi))^{-\alpha}\right)}{P_m (d_{mi}(\xi))^{-\alpha}}\right) \lambda \,\mathrm{d}\xi \\
&= \int_{\xi\in\mathcal{A}_m} \exp\left(-\frac{\left(2^{\frac{1}{r}}-1\right)\sigma^2}{P_m (d_{mi}(\xi))^{-\alpha}}\right) \prod_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \mathcal{L}_{|h_{ni}|^2}\left(\frac{\left(2^{\frac{1}{r}}-1\right) P_n (d_{ni}(\xi))^{-\alpha}}{P_m (d_{mi}(\xi))^{-\alpha}}\right) \lambda \,\mathrm{d}\xi \\
&\overset{\text{(b)}}{=} \int_{\xi\in\mathcal{A}_m} \exp\left(-\frac{\left(2^{\frac{1}{r}}-1\right)\sigma^2}{P_m (d_{mi}(\xi))^{-\alpha}}\right) \prod_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \frac{P_m (d_{mi}(\xi))^{-\alpha}}{\left(2^{\frac{1}{r}}-1\right) P_n (d_{ni}(\xi))^{-\alpha} + P_m (d_{mi}(\xi))^{-\alpha}} \lambda \,\mathrm{d}\xi,
\end{aligned} \tag{A.4}
$$

To calculate $\mathbb{E}_{h_{0i},\cdots,h_{Mi},\xi}\frac{1}{\bar{R}_{mi}(\xi)}$, we first calculate the cumulative distribution function (CDF) as equation (A.4) at the front of this page, where both (a) and (b) follow from that $|h_{ni}|^2 \sim \exp(1)$, $\forall n \in \mathcal{M}\cup\{0\}$, $\mathcal{L}_{|h_{ni}|^2}(\cdot)$ is the Laplace transform of $|h_{ni}|^2$. Based on (A.4), the probability density function (PDF) of $\frac{1}{\bar{R}_{mi}(\xi)}$ is

$$
\begin{aligned}
f_{\frac{1}{\bar{R}_{mi}(\xi)}}(r) = &\int_{\xi\in\mathcal{A}_m} \exp\left(-\frac{\left(2^{\frac{1}{r}}-1\right)\sigma^2}{P_m (d_{mi}(\xi))^{-\alpha}}\right) \\
&\prod_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \frac{P_m (d_{mi}(\xi))^{-\alpha}}{\left(2^{\frac{1}{r}}-1\right) P_n (d_{ni}(\xi))^{-\alpha} + P_m (d_{mi}(\xi))^{-\alpha}} \\
&\lambda\left(\frac{(\ln 2) 2^{\frac{1}{r}}\sigma^2}{r^2 P_m (d_{mi}(\xi))^{-\alpha}} + \sum_{n\in\mathcal{M}\cup\{0\}\setminus\{m\}} \right. \\
&\left. \frac{(\ln 2) 2^{\frac{1}{r}} P_n (d_{ni}(\xi))^{-\alpha}}{r^2\left(\left(2^{\frac{1}{r}}-1\right) P_n (d_{ni}(\xi))^{-\alpha} + P_m (d_{mi}(\xi))^{-\alpha}\right)}\right) \mathrm{d}\xi.
\end{aligned} \tag{A.5}
$$

Combining (A.1), (A.4) and (A.5), a_m can be expressed as (10).

From (5), we can obtain

$$
\begin{aligned}
b_m &= w_m \mathbb{E}_{h_m} \frac{1}{R_m} = \mathbb{E}_{h_m} \frac{1}{\log_2\left(1 + \frac{P_0 |h_m|^2 d_m^{-\alpha}}{\sigma^2}\right)} \\
&= \int_0^\infty \frac{1}{\mathrm{e}^x \log_2\left(1 + \frac{P_0 d_m^{-\alpha}}{\sigma^2} x\right)} \,\mathrm{d}x.
\end{aligned} \tag{A.6}
$$

According to (A.6), it is difficult to calculate the precious value of b_m in closed form. In the following, we provide one lower bound of b_m. From (A.6), we have

$$
\begin{aligned}
b_m &= \mathbb{E}_{h_m} \frac{1}{\log_2\left(1 + \frac{P_0 |h_m|^2 d_m^{-\alpha}}{\sigma^2}\right)} \\
&\geq \frac{1}{\mathbb{E}_{h_m} \log_2\left(1 + \frac{P_0 |h_m|^2 d_m^{-\alpha}}{\sigma^2}\right)} = \frac{1}{\bar{R}_m},
\end{aligned} \tag{A.7}
$$

where the inequality follows from the convexity of function $\frac{1}{x}$, and

$$
\begin{aligned}
\bar{R}_m &= \mathbb{E}_{h_m} \log_2\left(1 + \frac{P_0 |h_m|^2 d_m^{-\alpha}}{\sigma^2}\right) \\
&= \int_0^\infty \log_2\left(1 + \frac{P_0 d_m^{-\alpha}}{\sigma^2} x\right) \mathrm{e}^{-x} \,\mathrm{d}x \\
&\overset{\text{(c)}}{=} -\log_2\left(1 + \frac{P_0 d_m^{-\alpha}}{\sigma^2} x\right) \mathrm{e}^{-x}\Big|_0^\infty \\
&\quad + \int_0^\infty \frac{\mathrm{e}^{-x}}{(\ln 2)(x + \sigma^2/(P_0 d_m^{-\alpha}))} \,\mathrm{d}x \\
&\overset{\text{(d)}}{=} -\frac{1}{\ln 2} \mathrm{e}^{\frac{\sigma^2}{P_0 d_m^{-\alpha}}} \mathrm{Ei}\left(-\frac{\sigma^2}{P_0 d_m^{-\alpha}}\right),
\end{aligned} \tag{A.8}
$$

where we obtain (c) by using integration by parts, $\mathrm{Ei}(\cdot)$ is the exponential integral function and (d) follows from [36, equation (3.352.4)]. Based on (A.7) and (A.8), one lower bound of b_m is given by (11).

Appendix B
Proof of Theorem 1

Introducing auxiliary variable t_m, Problem (13) is equivalent to

$$
\begin{aligned}
\min_{\boldsymbol{s}_m, t_m} \quad & \left(\frac{b_m}{w_m} + \frac{D}{C_m - t_m}\right) \sum_{f\in\mathcal{F}} q_f L_f (1 - s_{mf}) && \text{(B.1a)} \\
\text{s.t.} \quad & \sum_{f\in\mathcal{F}} s_{mf} L_f \leq t_m && \text{(B.1b)} \\
& t_m \leq C_m && \text{(B.1c)} \\
& 0 \leq s_{mf} \leq 1, \quad \forall f \in \mathcal{F}, && \text{(B.1d)}
\end{aligned}
$$

since constraint (B.1b) always holds with equality for the optimal solution. With given t_m, the Lagrangian function of

Problem (B.1) is

$$\mathcal{L}_1(\boldsymbol{s}_m,\theta_m,\boldsymbol{\phi}_m,\boldsymbol{\psi}_m)=\left(\frac{b_m}{w_m}+\frac{D}{C_m-t_m}\right)\sum_{f\in\mathcal{F}}q_fL_f(1-s_{mf}) +\theta_m\left(\sum_{f\in\mathcal{F}}s_{mf}L_f-t_m\right) +\sum_{f\in\mathcal{F}}\phi_{mf}(-s_{mf})+\sum_{f\in\mathcal{F}}\psi_{mf}(s_{mf}-1), \tag{B.2}$$

where θ_m, $\boldsymbol{\phi}_m = [\phi_{m1},\cdots,\phi_{mF}]^T$ and $\boldsymbol{\psi}_m = [\psi_{m1},\cdots,\psi_{mF}]^T$ are nonnegative Lagrangian multipliers associated with corresponding constraints of Problem (B.1). According to [45], the optimal solution should satisfy the following KKT conditions of Problem (B.1):

$$\frac{\partial\mathcal{L}_1}{\partial s_{mf}}=-\left(\frac{b_m}{w_m}+\frac{D}{C_m-t_m}\right)q_fL_f+\theta_mL_f-\phi_{mf}+\psi_{mf}=0. \tag{B.3}$$

From (B.3), we have

$$-\left(\frac{b_m}{w_m}+\frac{D}{C_m-t_m}\right)q_f+\theta_m=\frac{\phi_{mf}-\psi_{mf}}{L_f},\quad\forall f\in\mathcal{F}. \tag{B.4}$$

Due to the fact that $\frac{b_m}{w_m}+\frac{D}{C_m-t_m}>0$ and $q_1>q_2>\cdots>q_F>0$, we can obtain

$$\frac{\phi_{m1}-\psi_{m1}}{L_1}<\frac{\phi_{m2}-\psi_{m2}}{L_2}<\cdots<\frac{\phi_{mF}-\psi_{mF}}{L_F}. \tag{B.5}$$

Since $L_f>0$ for all $f\in\mathcal{F}$, we consider the following four cases.

1) If $\frac{\phi_{m1}-\psi_{m1}}{L_1}>0$, we have $\phi_{mf}-\psi_{mf}>0$ for all $f\in\mathcal{F}$ according to (B.5). Since $\psi_{mf}\geq 0$, we further can obtain $\phi_{mf}>0$ for all $f\in\mathcal{F}$. Based on the complementary slackness condition $\phi_{mf}(-s_{mf})=0$, we have $s_{mf}=0$ for all $f\in\mathcal{F}$.
2) If $\frac{\phi_{mF}-\psi_{mF}}{L_F}<0$, we have $\phi_{mf}-\psi_{mf}<0$ for all $f\in\mathcal{F}$ according to (B.5). Since $\phi_{mf}\geq 0$, we further can obtain $\psi_{mf}>\phi_{mf}\geq 0$ for all $f\in\mathcal{F}$. Based on the complementary slackness condition $\psi_{mf}(1-s_{mf})=0$, we have $s_{mf}=1$ for all $f\in\mathcal{F}$.
3) If there exists one $f\in\mathcal{F}$ such that $\phi_{mf}-\psi_{mf}=0$. For $l\in\{1,\cdots,f-1\}$, we have $\frac{\phi_{ml}-\psi_{ml}}{L_l}<\frac{\phi_{mf}-\psi_{mf}}{L_f}=0$ from (B.5). Considering $\phi_{ml}\geq 0$ and the complementary slackness condition, we can obtain $s_{mf}=1$ for all $l\in\{1,\cdots,f-1\}$. For $l\in\{f+1,\cdots,F\}$, we have $\frac{\phi_{ml}-\psi_{ml}}{L_l}>\frac{\phi_{mf}-\psi_{mf}}{L_f}=0$ from (B.5). Considering $\psi_{ml}\geq 0$ and the complementary slackness condition, we can obtain $s_{mf}=0$ for all $l\in\{f+1,\cdots,F\}$.
4) If there exists one $f\in\mathcal{F}$ such that $\phi_{mf}-\psi_{mf}<0$ and $\phi_{m(f+1)}-\psi_{m(f+1)}>0$. For $l\in\{1,\cdots,f\}$, we have $\frac{\phi_{ml}-\psi_{ml}}{L_l}<\frac{\phi_{mf}-\psi_{mf}}{L_f}<0$ from (B.5). Considering $\phi_{ml}\geq 0$ and the complementary slackness condition, we can obtain $s_{mf}=1$ for all $l\in\{1,\cdots,f\}$. For $l\in\{f+1,\cdots,F\}$, we have $\frac{\phi_{ml}-\psi_{ml}}{L_l}\geq\frac{\phi_{m(f+1)}-\psi_{m(f+1)}}{L_{f+1}}>0$ from (B.5). Considering $\psi_{ml}\geq 0$ and the complementary slackness condition, we can obtain $s_{mf}=0$ for all $l\in\{f+1,\cdots,F\}$.

Based on the above analysis, the optimal solution of Problem (B.1) with any given t_m has the structure $(\mathbf{1}_{f-1},s^*_{mf},\mathbf{0}_{F-f})$, where

$$\mathbf{1}_{f-1}=[\underbrace{1,\cdots,1}_{f-1}],\mathbf{0}_{F-f}=[\underbrace{0,\cdots,0}_{F-f}], \tag{B.6}$$

$s^*_{mf}\in[0,1]$, and $f\in\mathcal{F}$. Since Problem (13) is equivalent to Problem (B.1), Theorem 1 is proved.

Appendix C
Proof of Theorem 2

Based on Theorem 1, the optimal solution of Problem (13) has the structure $(\mathbf{1}_{f-1},s^*_{mf},\mathbf{0}_{F-f})$ with $s^*_{mf}\in[0,1]$ and $f\in\mathcal{F}$. As a result, the optimal solution of Problem (13) is one the F solutions, $(s^*_{m1},\mathbf{0}_{F-1})$, $(1,s^*_{m2},\mathbf{0}_{F-2})$, $\cdots$, $(\mathbf{1}_{F-1},s^*_{mF})$, with the best objective value. For the f-th solution $(\mathbf{1}_{f-1},s^*_{mf},\mathbf{0}_{F-f})$, constraint (13b) should be satisfied, i.e., $\sum_{l=1}^{f-1}L_l\leq C_m$ and then s^*_{mf} can be obtained by substituting the optimal values of other $F-1$ variables into Problem (13), i.e., s^*_{mf} is the optimal solution of the following problem:

$$\min_{s_{mf}}\ \left(\frac{b_m}{w_m}+\frac{D}{C_m-\sum_{l=1}^{f-1}L_l-s_{mf}L_f}\right)\left(q_fL_f(1-s_{mf})+\sum_{l=f+1}^{F}q_lL_l\right)\triangleq g_{mf}(s_{mf}) \tag{C.1a}$$

$$\text{s.t. } 0\leq s_{mf}\leq s^{\max}_{mf}, \tag{C.1b}$$

where

$$s^{\max}_{mf}=\min\left\{1,\frac{C_m-\sum_{l=1}^{f-1}L_l}{L_f}\right\}. \tag{C.2}$$

The first-order derivative of the objective function (C.1a) is

$$g'_{mf}(s_{mf})=\frac{DL_f\left(q_fL_f(1-s_{mf})+\sum_{l=f+1}^{F}q_lL_l\right)}{\left(C_m-\sum_{l=1}^{f-1}L_l-s_{mf}L_f\right)^2} -q_fL_f\left(\frac{b_m}{w_m}+\frac{D}{C_m-\sum_{l=1}^{f-1}L_l-s_{mf}L_f}\right). \tag{C.3}$$

Setting the first-order derivative (C.3) with 0 yields

$$-b_mq_f\left(C_m-\sum_{l=1}^{f-1}L_l-s_{mf}L_f\right)^2 +w_mD\left(-q_fC_m+\sum_{l=f+1}^{F}q_lL_l+\sum_{l=1}^{f}q_fL_l\right)=0. \tag{C.4}$$

To solve (C.4), we consider the following two cases.

1) If $-q_fC_m+\sum_{l=f+1}^{F}q_lL_l+\sum_{l=1}^{f}q_fL_l\leq 0$, the left term of equation (C.4) is always nonpositive, i.e., $g'_{mf}(s_{mf})\leq 0$ for all $s_{mf}\geq 0$. The objective function $g_{mf}(s_{mf})$ monotonically decreases with s_{mf}, and the optimal $s^*_{mf}=s^{\max}_{mf}$.

2) If $-q_f C_m+\sum_{l=f+1}^{F} q_l L_l+\sum_{l=1}^{f} q_f L_l > 0$, there exists two different roots to equation (C.4), i.e.,

$$s_{mf}(1)=\frac{(C_m-\sum_{l=1}^{f-1} L_l)\sqrt{b_m q_f}-\sqrt{\alpha}}{L_f\sqrt{b_m q_f}}, \quad \text{(C.5)}$$

where $\alpha=w_m D\left(-q_f C_m+\sum_{l=f+1}^{F} q_l L_l+\sum_{l=1}^{f} q_f L_l\right)$, and

$$s_{mf}(2)=\frac{(C_m-\sum_{l=1}^{f-1} L_l)\sqrt{b_m q_f}+\sqrt{\alpha}}{L_f\sqrt{b_m q_f}}. \quad \text{(C.6)}$$

Since the objective function $g_{mf}(s_{mf})$ decreases with s_{mf} when $s_{mf} < s_{mf}(1)$ and $s_{mf} > s_{mf}(2)$ and increases with s_{mf} when $s_{mf}(1) \le s_{mf} \le s_{mf}(2)$. If $0 < s_{mf}(1) < s_{mf}^{\max}$, we have the optimal $s_{mf}^* = \arg\min_{s_{mf}\in\{s_{mf}(1),s_{mf}^{\max}\}} g_{mf}(s_{mf})$. If $s_{mf}(1) \le 0$, $s_{mf}^* = \arg\min_{s_{mf}\in\{0,s_{mf}^{\max}\}} g_{mf}(s_{mf})$. If $s_{mf}(1) \ge s_{mf}^{\max}$, we have the optimal $s_{mf}^* = s_{mf}^{\max}$.

To further reduce the number of candidate solutions for the optimal cache placement, we investigate the property of the candidate solutions in the following lemma.

Lemma 2: For two solutions $(\mathbf{1}_{f-1}, \mathbf{0}_{F-f+1})$ and $(\mathbf{1}_{f-1}, s_{mf}^*, \mathbf{0}_{F-f})$ with $s_{mf}^* \in (0,1]$, solution $(\mathbf{1}_{f-1}, s_{mf}^*, \mathbf{0}_{F-f})$ yields better objective value (13a) than solution $(\mathbf{1}_{f-1}, \mathbf{0}_{F-f+1})$.

Proof: According to the above proof of Theorem 2, s_{mf}^* is the optimal cache strategy with given $s_{ml}=1$ for $l<f$ and $s_{ml}=0$ for $l>f$, which shows that the objective function (13a) of solution $(\mathbf{1}_{f-1}, s_{mf}^*, \mathbf{0}_{F-f})$ is smaller than that of solution $(\mathbf{1}_f, \mathbf{0}_{F-f})$, i.e., Lemma 2 is proved. □

Denoting $z(f)=\sum_{l=1}^{f} L_l+\sum_{l=f+1}^{F}\frac{q_l L_l}{q_f}$, we have

$$\begin{aligned} z(f)-z(f-1) &= L_f+\sum_{l=f+1}^{F}\frac{q_l L_l}{q_f}-\sum_{l=f}^{F}\frac{q_l L_l}{q_{f-1}} \\ &> L_f+\sum_{l=f+1}^{F}\frac{q_l L_l}{q_f}-\sum_{l=f}^{F}\frac{q_l L_l}{q_f}=0, \quad \text{(C.7)} \end{aligned}$$

where the inequality follows from that $q_f < q_{f-1}$. Based on (16) and (C.7), we have $C_m \ge z(F_{m2}) > z(F_{m2}-1) > \cdots > z(1)$. From Theorem 1, the first F_{m2} potentially optimal solutions can be expressed by $(\mathbf{1}_{f-1}, s_{mf}^{\max}, \mathbf{0}_{F-f})$, $f=1,\cdots,F_{m2}$. Since (16) implies that $\frac{C_m-\sum_{l=1}^{f-1} L_l}{L_f} \ge 1$, we have $s_{mf}^{\max}=1$ for $f \le F_{m2}$ according to (C.2). As a result, the first F_{m2} potentially optimal solutions are $(\mathbf{1}_{f-1}, 1, \mathbf{0}_{F-f})$, $f=1,\cdots,F_{m2}$.

Based on Lemma 2, the number of potentially optimal solutions can be reduced to $F_{m1}-F_{m2}+1$, which can largely simplify the computation of obtaining the optimal solution.

APPENDIX D
PROOF OF THEOREM 3

Denoting by χ the Lagrange multiplier associated to (18b), the Lagrange function of Problem (18) is

$$\begin{aligned} \mathcal{L}_2(\boldsymbol{w},\chi) &= \sum_{m\in\mathcal{M}}\frac{\sum_{f\in\mathcal{F}} b_m q_f L_f(1-s_{mf})}{w_m} \\ &\quad +\chi\left(\sum_{m\in\mathcal{M}\cup\{0\}} w_m - W\right). \quad \text{(D.1)} \end{aligned}$$

The optimal solution should satisfy the following KKT conditions of Problem (18):

$$\frac{\partial\mathcal{L}_2}{\partial w_m}=-\frac{\sum_{f\in\mathcal{F}} b_m q_f L_f(1-s_{mf})}{w_m^2}+\chi=0, \quad \forall m\in\mathcal{M}, \quad \text{(D.2)}$$

which yields

$$w_m=\frac{\sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f(1-s_{mf})}}{\sqrt{\chi}}, \quad \forall m\in\mathcal{M}. \quad \text{(D.3)}$$

Substituting (D.3) into (18b), we can obtain

$$\sqrt{\chi}=\frac{\sum_{m\in\mathcal{M}}\sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f(1-s_{mf})}}{W}. \quad \text{(D.4)}$$

Substituting (D.4) into (D.3) yields (19).

APPENDIX E
PROOF OF THEOREM 4

Introducing auxiliary variable $\boldsymbol{t}=[t_1,\cdots,t_M]^T$, Problem (20) is equivalent to

$$\begin{aligned} \min_{\boldsymbol{s},\boldsymbol{t}} \ & \frac{1}{W}\left(\sum_{m\in\mathcal{M}}\sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f(1-s_{mf})}\right)^2 \\ & +\sum_{m\in\mathcal{M}}\frac{\sum_{f\in\mathcal{F}} q_f L_f D(1-s_{mf})}{C_m-t_m} \quad \text{(E.1a)} \\ \text{s.t.} \ & \sum_{f\in\mathcal{F}} L_f s_{mf}\le t_m, \quad \forall m\in\mathcal{M} \quad \text{(E.1b)} \\ & t_m\le C_m, \quad \forall m\in\mathcal{M} \quad \text{(E.1c)} \\ & 0\le s_{mf}\le 1, \quad \forall m\in\mathcal{M}, f\in\mathcal{F}. \quad \text{(E.1d)} \end{aligned}$$

With given $\boldsymbol{t}$, the Lagrangian function of Problem (E.1) is

$$\begin{aligned} \mathcal{L}_3\ (\boldsymbol{s},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}) &= \frac{1}{W}\left(\sum_{m\in\mathcal{M}}\sqrt{\sum_{f\in\mathcal{F}} b_m q_f L_f(1-s_{mf})}\right)^2 \\ &\quad +\sum_{m\in\mathcal{M}}\frac{\sum_{f\in\mathcal{F}} q_f L_f D(1-s_{mf})}{C_m-t_m} \\ &\quad +\sum_{m\in\mathcal{M}}\theta_m\left(\sum_{f\in\mathcal{F}} L_f s_{mf}-t_m\right)+\sum_{m\in\mathcal{M}}\sum_{f\in\mathcal{F}}\phi_{mf}(-s_{mf}) \\ &\quad +\sum_{m\in\mathcal{M}}\sum_{f\in\mathcal{F}}\psi_{mf}(s_{mf}-1), \quad \text{(E.2)} \end{aligned}$$

where $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_M]^T$, $\boldsymbol{\phi} = [\phi_{11}, \cdots, \phi_{1F}, \cdots, \phi_{MF}]^T$ and $\boldsymbol{\psi} = [\psi_{11}, \cdots, \psi_{1F}, \cdots, \psi_{MF}]^T$ are nonnegative Lagrangian multipliers associated with corresponding constraints of Problem (E.1). The optimal solution should satisfy the following KKT conditions of Problem (E.1):

$$\frac{\partial \mathcal{L}_3}{\partial s_{mf}} = -\left(\frac{\sum_{n\in\mathcal{M}} b_m v_n}{W v_m} + \frac{D}{C_m - t_m}\right) q_f L_f + \theta_m L_f - \phi_{mf} + \psi_{mf} = 0, \quad \forall m \in \mathcal{M}, f \in \mathcal{F}, \tag{E.3}$$

where $v_m = \sqrt{\sum_{f\in\mathcal{F}} q_f L_f b_m (1 - s_{mf})}$, $\forall m \in \mathcal{M}$. From (B.3), we have

$$-\left(\frac{\sum_{n\in\mathcal{M}} b_m v_n}{W v_m} + \frac{D}{C_m - t_m}\right) q_f + \theta_m = \frac{\phi_{mf} - \psi_{mf}}{L_f}. \tag{E.4}$$

Due to the fact that $\frac{\sum_{n\in\mathcal{M}} b_m v_n}{W v_m} + \frac{D}{C_m - t_m} > 0$ and $q_1 > q_2 > \cdots > q_F > 0$, we can obtain

$$\frac{\phi_{m1} - \psi_{m1}}{L_1} < \frac{\phi_{m2} - \psi_{m2}}{L_2} < \cdots < \frac{\phi_{mF} - \psi_{mF}}{L_F}, \quad \forall m \in \mathcal{M}. \tag{E.5}$$

Similar to the analysis in Appendix B, the optimal solution of Problem (E.1) with any given $\boldsymbol{t}$ must have the structure

$$(\mathbf{1}_{f_1 - 1}, s^*_{1f_1}, \mathbf{0}_{F-f_1}, \cdots, \mathbf{1}_{f_m - 1}, s^*_{mf_m}, \mathbf{0}_{F-f_m}, \cdots, \mathbf{1}_{f_M - 1}, s^*_{Mf_M}, \mathbf{0}_{F-f_M}), \tag{E.6}$$

where $s^*_{mf_m} \in [0, 1]$, $f_m \in \mathcal{F}$, $m \in \mathcal{M}$. Since Problem (13) is equivalent to Problem (E.1), Theorem 4 is proved.

Appendix F
Proof of Theorem 5

According to Theorem 4, the optimal $\boldsymbol{s}_m$ for pico BS m to Problem (20) with given $\boldsymbol{s}_{-m}$ is one of the F solutions, $(s^*_{m1}, \mathbf{0}_{F-1})$, $(1, s^*_{m2}, \mathbf{0}_{F-2})$, $\cdots$, $(\mathbf{1}_{F-1}, s^*_{mF})$, with the best objective value. For the f-th solution $(\mathbf{1}_{f-1}, s^*_{mf}, \mathbf{0}_{F-f})$, $\sum_{l=1}^{f-1} L_l \le C_m$ should be first satisfied from (20b) and s^*_{mf} is the optimal solution of the following problem according to (20):

$$\min_{s_{mf}} \ \frac{1}{W}(u_m + \sqrt{b_m w_f - b_m q_f L_f s_{mf}})^2 + \frac{w_f D - q_f L_f D s_{mf}}{C_m - \sum_{l=1}^{f-1} L_l - L_f s_{mf}} \triangleq y_{mf}(s_{mf}) \tag{F.1a}$$

$$\text{s.t. } 0 \le s_{mf} \le s^{\max}_{mf}, \tag{F.1b}$$

where $u_m = \sum_{n\in\mathcal{M}\setminus\{m\}} \sqrt{\sum_{l\in\mathcal{F}} b_n q_l L_l (1 - s_{nl})}$, $w_f = \sum_{l=f}^{F} q_l L_l$, $s^{\max}_{mf}$ is defined in (C.2). The first-order derivative of the objective function (F.1a) is

$$y'_{mf}(s_{mf}) = \frac{-b_m q_f L_f (u_m + \sqrt{b_m w_f - b_m q_f L_f s_{mf}})}{W\sqrt{b_m w_f - b_m q_f L_f s_{mf}}} + \frac{\left(w_f - q_f C_m + q_f \sum_{l=1}^{f-1} L_l\right) L_f D}{\left(C_m - \sum_{l=1}^{f-1} L_l - L_f s_{mf}\right)^2}. \tag{F.2}$$

Setting $y'_{mf}(s_{mf}) = 0$ and $x = \sqrt{b_m w_f - b_m q_f L_f s_{mf}}$ to (F.2) yields

$$z_5 x^5 + z_4 x^4 + z_3 x^3 + z_2 x^2 + z_1 x + z_0 = 0, \tag{F.3}$$

where

$$\begin{aligned}
z_5 &= b_m q_f l_f, \\
z_4 &= b_m u_m q_f l_f, \\
z_3 &= -2 b_m^2 q_f l_f \left(q_f C_m - q_f \sum_{l=1}^{f-1} L_l - w_f\right), \\
z_2 &= 2 b_m^2 u_m q_f l_f \left(q_f C_m - q_f \sum_{l=1}^{f-1} L_l - w_f\right), \\
z_1 &= b_m^3 q_f l_f \left(q_f C_m - q_f \sum_{l=1}^{f-1} L_l - w_f\right)^2 \\
&\quad - b_m^2 q_f^2 L_f D W \Big(w_f - q_f C_m + q_f \sum_{l=1}^{f-1} L_l\Big), \\
z_0 &= b_m^2 u_m \left(q_f C_m - q_f \sum_{l=1}^{f-1} L_l - w_f\right)^2.
\end{aligned}$$

Having obtained x from (F.3), s_{mf} can be presented by $s_{mf} = \frac{b_m w_f - x^2}{b_m q_f L_f} = \frac{b_m \sum_{l=f}^{F} q_l L_l - x^2}{b_m q_f L_f}$. Due to that $s_{mf} \in (0, s^{\max}_{mf})$, x should be in the interval $(\sqrt{b_m \sum_{l=f}^{F} q_l L_l - b_m q_f L_f s^{\max}_{mf}}, \sqrt{b_m \sum_{l=f}^{F} q_l L_l})$. According to Abel-Ruffini theorem [46], there is no algebraic expression for general quintic equations over the rationals in terms of radicals. The roots located in $(\sqrt{b_m \sum_{l=f}^{F} q_l L_l - b_m q_f L_f s^{\max}_{mf}}, \sqrt{b_m \sum_{l=f}^{F} q_l L_l})$ to equation are numerically calculated using root-finding algorithm for polynomials [37]. Since the optimal solution of Problem (F.1) either lies in the boundary or in the extreme point, the optimal s^*_{mf} can be presented in (21).

References

[1] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, Jun. 2012.

[2] V. W. Wong and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge university press, 2017.

[3] L. Ying, Z. Liu, D. Towsley, and C. H. Xia, "Distributed operator placement and data caching in large-scale sensor networks," in *Proc. IEEE Int. Conf. Compt. Commun.*, Phoenix, AZ, USA, Apr. 2008.

[4] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[5] M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li, "Enhancement for content delivery with proximity communications in caching enabled wireless networks: Architecture and challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 70–76, Aug. 2016.

[6] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[7] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[8] A. Argyriou, K. Poularakis, G. Iosifidis, and L. Tassiulas, "Video delivery in dense 5G cellular networks," *IEEE Network*, vol. 31, no. 4, pp. 28–34, Jul. 2017.

[9] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[10] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, Dec. 2015.

[11] S. W. Jeon, S. N. Hong, M. Ji, and G. Caire, "Caching in wireless multihop device-to-device networks," in *Proc. IEEE Int. Conf. Commun.*, London, UK, Jun. 2015, pp. 6732–6737.

[12] A. Afzal, S. A. R. Zaidi, D. McLernon, and M. Ghogho, "On the analysis of cellular networks with caching and coordinated device-to-device communication," in *Proc. IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[13] 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, Sep. 2016.

[14] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6284–6297, Sep. 2016.

[15] ——, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2956–2970, Jul. 2017.

[16] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[17] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[18] R. G. Stephen and R. Zhang, "Green OFDMA resource allocation in cache-enabled CRAN," in *Proc. IEEE Online Conf. Green Commun.*, Piscataway, NJ, USA, Nov. 2016, pp. 70–75.

[19] W. Wen, Y. Cui, F. C. Zheng, and S. Jin, "Random caching based cooperative transmission in heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun.*, Pais, France, May 2017, pp. 1–6.

[20] J. Wen, K. Huang, S. Yang, and V. O. K. Li, "Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5939–5952, Sep. 2017.

[21] Y. Pan, C. Pan, H. Zhu, Q. Z. Ahmed, M. Chen, and J. Wang, "On consideration of content preference and sharing willingness in D2D assisted offloading," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 4, pp. 978–993, Apr. 2017.

[22] K. Li, C. Yang, Z. Chen, and M. Tao, "Optimization and analysis of probabilistic caching in N-tier heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–1, 2017.

[23] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[24] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *Proc. IEEE Global Commun. Conf.*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[25] D. Liu and C. Yang, "Optimizing caching policy at base stations by exploiting user preference and spatial locality," *arXiv preprint arXiv:1710.09983*, 2017.

[26] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao *et al.*, "Cooperative edge caching in user-centric clustered mobile networks," *arXiv preprint arXiv:1710.08582*, 2017.

[27] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, V. W. S. Wong, and J. Wang, "Cross-layer optimization of fast video delivery in cache- and buffer-enabled relaying networks," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2017.

[28] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun.*, London, UK, Jun. 2015, pp. 3358–3363.

[29] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[30] N. Jindal, P. R. Panda, and S. R. Sarangi, "Reusing trace buffers to enhance cache performance," in *Design, Automation Test in Europe Conference Exhibition*, Mar. 2017, pp. 572–577.

[31] W.-H. Kang, S.-W. Lee, and B. Moon, "Flash-based extended cache for higher throughput and faster recovery," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1615–1626, 2012.

[32] S. Huang, Q. Wei, D. Feng, J. Chen, and C. Chen, "Improving flash-based disk cache with lazy adaptive replacement," *ACM Trans. Storage*, vol. 12, no. 2, p. 8, 2016.

[33] S. Jiang and X. Zhang, "Lirs: an efficient low inter-reference recency set replacement policy to improve buffer cache performance," *ACM SIGMETRICS Performance Evaluation Review*, vol. 30, no. 1, pp. 31–42, 2002.

[34] J. D. Little, "A proof for the queuing formula: $l = \lambda w$," *Operations research*, vol. 9, no. 3, pp. 383–387, 1961.

[35] T. Islam, A. Ikhlef, R. Schober, and V. K. Bhargava, "Diversity and delay analysis of buffer-aided BICM-OFDM relaying," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5506–5519, Nov. 2013.

[36] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic Press, 2014.

[37] K. Madsen, "A root-finding algorithm based on newton's method," *BIT Numerical Mathematics*, vol. 13, no. 1, pp. 71–75, 1973.

[38] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[39] L. P. Qian, Y. Wu, H. Zhou, and X. Shen, "Joint uplink base station association and power control for small-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5567–5582, Sep. 2017.

[40] S. Dadallage, C. Yi, and J. Cai, "Joint beamforming, power, and channel allocation in multiuser and multichannel underlay MISO cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3349–3359, May 2016.

[41] J. P. F. Trovo, V. D. N. Santos, P. G. Pereirinha, H. M. Jorge, and C. H. Antunes, "A simulated annealing approach for optimal power source management in a small EV," *IEEE Trans. Sustainable Energy*, vol. 4, no. 4, pp. 867–876, Oct. 2013.

[42] A. J. Monticelli, R. Romero, and E. N. Asada, "Fundamentals of simulated annealing," *Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems*, pp. 123–146, 2007.

[43] D. R. Thompson and G. L. Bilbro, "Sample-sort simulated annealing," *IEEE Trans. Sys., Man, Cybern., B, Cybern.*, vol. 35, no. 3, pp. 625–632, 2005.

[44] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.

[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[46] P. Pesic, *Abel's proof*. MIT Press, 2003.