IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 10, 2019, accepted May 16, 2019, date of publication May 22, 2019, date of current version June 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918334

User-Centric Delay-Aware Joint Caching and User Association Optimization in Cache-Enabled Wireless Networks

WENPENG JING[1], XIANGMING WEN[1], ZHAOMING LU[1], AND HAIJUN ZHANG[2], (Senior Member, IEEE)

[1]Beijing Laboratory of Advanced Information Networks, Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Wenpeng Jing (jingwenpeng@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China, under Grant 61801036.

ABSTRACT The mobile edge caching is a promising way to reduce the user-perceived delay and improve the transmission data rates for the wireless networks. However, the cache capacities of base stations (BSs) tend to be limited and users' interests for the contents are diverse, which makes the content placement decision critical for the network performance optimization. Besides, due to the flexible user-BS association, it is more complicated to optimize the content delivery and placement which are coupled with each other. This paper investigates the content placement and content delivery strategies in the cache-enabled wireless networks. In particular, the *effective capacity*, which can characterize the end-to-end user-perceived delay and data rates simultaneously, is introduced as the user's utility metric. As the content caching and content delivery operate in different time-scales, they are investigated separately. For the content caching, a content placement problem is formulated, where both users' different active levels and diverse content preferences are considered. Due to the NP-hard nature, the problem is decomposed into two sub-problems, and an iterative association-aware content placement algorithm is proposed. For the content delivery, the user-BS association problem is formulated, and a cache-aware user-BS association algorithm is designed. The performance of the proposed algorithms is evaluated based on the simulations. The numerical results show that the proposed algorithms have a better capability to cope with users' diverse active levels, and achieve a better performance in terms of effective capacity and fairness level, compared with the existing algorithms.

INDEX TERMS Mobile edge caching, content placement, user-BS association, resources allocation.

I. INTRODUCTION

The mobile data traffic has experienced an explosive growth in recent years, driven by the rapid advance of the wireless networks, smart devices, and mobile Internet services. Specifically, it is reported that the mobile video has accounted for more than half of the global mobile traffic [1]. Usually, different video contents have different popularities. For example, it is reported that 80% of the overall video requests in the YouTube are contributed by only 10% of the most popular contents [2]. Hence, a large portion of the mobile video traffic is generated by the duplicate transmission of the same popular video contents. To cope with the repetitive transmission, a new paradigm, i.e., mobile edge caching, is proposed and regarded as one of the most promising solutions to reduce the traffic burden of the mobile networks [3]. The main idea of mobile edge caching is to equip the base stations (BSs) with storage devices to cache the popular contents [4], [5]. In particular, the BSs can proactively fetch the popular contents during the off-peak traffic hour, and serve users' requests locally during the peak traffic hour. This could relieve the backhaul traffic congestion and improve users' data rates significantly. Besides, if the requested contents have been cached, users' perceived delay can be reduced significantly without fetching from the remote content server [6].

However, owing to the huge number of BSs, to equip each of them with large cache capacity would bring a great financial burden for the operator. Instead, the relative small capacity storage device for each BS would be a more economic solution. As a result, only a small number of video

The associate editor coordinating the review of this manuscript and approving it for publication was Guanding Yu.

contents would be cached for every BS. On the other hand, the volume of the video contents is growing exponentially nowadays, with the rise of the video websites and applications. Considering the contradiction between the number of video contents and BSs' limited cache capacity, it becomes a critical problem to choose the proper contents for each BS to cache.

Usually, traditional content distribution network (CDN) operator owns a limited number of cache servers, but each cache server has large-sized storage capacity and can response millions of requests simultaneously [7], [8]. Hence, the popularity based caching strategies have the good hit ratio performance and are common to be used. However, this is totally different from the mobile edge caching scenarios, where each BS only covers a small area and serves a few users. Due to users' diverse interests and preferences for the video contents, the demand agglomeration of a small number of users would deviate from the global popularity of the contents. Hence, the popularity based content caching strategies would not be suitable for the cache-enabled wireless networks. Instead, each BS should make the content placement decision based on the fine-granularity statistics information, i.e., the individual user's interest and preference for the contents.

What's more, with the densification trend of wireless network, the users have high probabilities to be covered by multiple BSs [9], [10]. Hence, the user-BS association could be adjusted according to their specific contents requests. Users' experience associated with certain BS would be impacted by not only the channel quality and load condition of the BS [11]–[14], but also the local availability of the requested contents, i.e., whether the contents have been cached or not. In principle, the user-BS association could be utilized to squeeze the performance gain of mobile edge caching. However, the user-BS association flexibility for the content delivery would bring new challenges for the content placement, as the contents demands of each BS would be more difficult to be estimated without fixed user-BS association. Hence, the user-BS association and content placement would be coupled with each other and should be considered and optimized jointly.

A. RELATED WORKS

Ever since the mobile edge caching was proposed, the joint optimization of the content placement and user-BS association has attracted significant attentions [15].

Many existing works focus on the maximization of the hit ratio, or the backhaul traffic reduction, and make the joint optimization of content caching and delivery in cache-enabled wireless networks. In [3], the wireless edge caching system named FemtoCaching is proposed, and the content placement is optimized to maximize the hit ratio considering the adaptive user-BS association. In [16], the fraction of content requests served by the small-cell BSs (SBSs) is maximized by the joint optimization of user-BS association and content caching. In [17], a novel content placement strategy is proposed based on a hybrid collaborative filtering model, and the backhaul traffic offloading is maximized by the user-BS association adaption. In [18], a green content caching and user-BS association mechanism is proposed for energy-harvesting enabled small-cell networks, to maximize the number of users' content requests served by the SBSs.

Besides, considering mobile edge caching's impact on the content transmission delay, some recent works investigate the delay-based content placement and delivery optimization. In [19], the content caching and user-BS association are investigated to minimize the total time to satisfy the average demands of users, where a cached-enabled heterogeneous networks with wireless backhaul is considered. In [20], the amount of time required to satisfy all users' requests is minimized by joint optimization of caching, routing, and channel assignment. In [21], the average download delay including both the wireless link delay and the backhaul link delay is optimized based on the joint optimization of the content caching and user-BS association.

Moreover, instead of single utility, some papers take multiple utilities into account simultaneously, and optimize the content placement and content delivery based on composite utilities. In [22], the content placement and user-BS association algorithm is proposed based on a weighted sum utility of users' data rates and backhaul traffic reduction. In [23], a two-sided matching game based user-BS association algorithm and the fluctuation popularity estimation based content placement algorithm are proposed, respectively, in order to optimize the weighted sum of users data rates and the cache hit ratios.

The joint content placement and delivery optimisation have been widely investigated based on a variety of utilities, however, some factors are overlooked in most of the literature.

On one hand, the unilateral objective used by most of the existing works, e.g., the backhaul traffic reduction or the delay reduction, can characterize the caching's impact on the content delivery from only one perspective. Besides, the composite utility function based optimization, which is in form of the weighted sum of two utilities, always has the manual tuned weight parameter, which cannot capture the interplay of the two kinds of utilities exactly. In contrast, the concept of effective capacity [24], which is an efficient criterion to characterize the delay and data rates of the content transmission, would be a more suitable metric for the cache-enabled wireless networks. Specifically, in [25], the effective capacity has been introduced as the utility for the cloud RAN scenarios with mobile edge caching. However, [25] derives the content caching solution based on the novel users' demand and users' mobility prediction method, and does not consider the interaction between the content caching and the user-BS association. This cannot reap the performance gain brought by the flexible user-BS association sufficiently.

On the other hand, the users are always with different active levels in the realistic wireless network scenarios. Specifically, it is reported that the network traffic generated by 20% of the users could account for at most 80% of the overall

traffic. Different active levels mean that the users are not statistically equivalent to each other any more in the content placement decision process. Instead, each user should be assigned different weights when the caching interests of users may conflict. In [26], the differences of users' active levels are considered in the optimization of the mobile edge caching scheme. However, the proposed caching algorithm is based on a fixed user-BS association, where each user would be associated with the nearest BS that has cached the requested content. This would limit the algorithm to reaping the benefits of flexible user-BS association, which can not achieve the optimal solution.

Hence, how to take into account users' different active levels and flexible user-BS association, and design the efficient content placement and user-BS association algorithm remains to be an open problem.

B. THE CONTRIBUTION AND ORGANIZATION

In this paper, we aim to maximize the utilities of users' in the cache-enabled wireless networks, by joint optimization of content placement and user-BS association. Specifically, a more realistic user behavior model is considered, where users have discrepancies on the active levels and content preferences. Besides, the effective capacity is introduced as user's performance metric, which can provide a user-centric model to capture the impact of the mobile edge caching on the data rates and user-perceived delay comprehensively.

Due to the mismatch in decision time-scales, the content placement and content delivery problem are investigated separately, but their interaction is integrated into each other. For the **content delivery phase**, a user-BS association optimization problem is formulated, where the user-BS association would be optimized based on users' requests, cache status and each BS's traffic load. A variable relaxation technique is adopted to convert the original problem into a convex optimization problem. Based on Lagrangian dual method, a cache-aware user-BS association algorithm is derived. For the **content placement phase**, a content placement optimization problem is formulated, where the users' active levels and the user-BS association flexibility are considered. Due to the NP-hard nature of the problem, an iterative scheme is adopted and an association-aware content placement algorithm is proposed. Finally, the performance of the proposed algorithms is validated by the simulation. Specifically, the numerical results confirm that the proposed algorithms have a better performance in terms of both users' average effective capacity and the fairness level, compared with the existing algorithms.

The remainder of the paper is organized as follows: in Section II, the effective capacity model is presented, and the content delivery and content placement optimization problems are also formulated, respectively. Section III proposes the cache-aware user-BS association algorithm and association-aware content placement algorithm, respectively. In Section IV, the simulation results are given, and the performance of the proposed algorithms is evaluated. Finally, the conclusion is given in Section V.

II. SYSTEM MODEL

A. NETWORK SCENARIO

In this paper, the downlink transmission of a cache-enabled wireless small-cell network is considered, which consists of M BSs and K users. The set of BSs is denoted as $\mathcal{M}=\{1,\cdots,m,\cdots,M\}$, and the set of users is denoted as $\mathcal{K}=\{1,\cdots,k\cdots,K\}$. Each of the BSs and users is assumed to be equipped with only one antenna. The system bandwidth is BW, and is reused by all the BSs in the network. Suppose that the overall contents library consists of F contents, which is denoted by the set $\mathcal{F}=\{1,\cdots,f,\cdots,F\}$. For ease of illustration, all the contents are assumed to have the same size. Besides, each BS is equipped with the mobile edge cache. The cache capacity of BS m is denoted as X_m^{MAX}, which means that at most X_m^{MAX} contents can be stored.

B. USER BEHAVIOR OF REQUESTING CONTENTS

Different from the existing works, in this paper a more realistic user request pattern is considered.

- Each user has its own content interest and preference, which may not be in line with the global popularity of the contents. Denote $\boldsymbol{p}^k=\left[p^{k,1},\cdots,p^{k,f},\cdots,p^{k,F}\right]$ as user k's preference for all the contents, where $p^{k,f}$ is the probability that a request of user k is for the content f. Specifically, the $p^{k,f}$ could be predicted based on the machine learning techniques [27], [28], and is assumed to be known a priori during the content placement phase.
- The active levels of users' are different from each other. This is in more accordance with the practical situation, where a portion of the users are addicted to the mobile video, and would send content requests more frequently, while the other users would be less active, and would send fewer content requests. Specifically, define $\boldsymbol{\lambda}=\left[\lambda^1,\cdots,\lambda^k,\cdots,\lambda^K\right]$ as users' active level set, where λ^k denotes the probability of user k sending content request at one time slot. Note that the λ^k can be predicted based on the statistics of each user's request logs, and is assumed to be a priori during the content placement phase.

C. TRANSMISSION MODEL

If the user k is associated with BS m and is allocated with the overall system bandwidth BW, the maximum data rates that can be achieved by user k is

$$R_m^k = BW\log_2(1+\frac{p_m h_m^k}{\sigma^2+I_m^k}), \tag{1}$$

where p_m is the transmit power of BS m, h_m^k is the channel power gain from BS m to user k, $I_m^k=\sum_{m'\in\mathcal{M}_m^k} p_{m'}h_{m'}^k$ is the power of inter-cell interference, $\mathcal{M}_m^k$ is the BSs set that would interfere with user k, and σ^2 is the power of the additive white Gaussian noise (AWGN).

The introduction of the mobile edge caching would divide the content transmission into two phases: the content

placement phase and the content delivery phase. During the content placement phase, the BSs would prefetch and cache the contents that may be requested in the future. Then, during the content delivery phase, the contents requested by the users would be delivered from the cache directly if they have been cached. This would reduce the congestion probability of the transmission link between the BS and the remote server. Besides, without multiple-hop transmission, the end-to-end delay could be reduced. The improvement of data rates and delay would be of great benefit to users' experience.

Note that although whether the content has been cached would have a direct impact on the end-to-end delay and data rates of the content transmission, however, this impact cannot be characterized by the traditional metrics such as Shannon capacity or the transmission delay. In contrast, effective capacity [24], which is a channel capacity model and is able to characterize the maximum data rates of the transmission link under a statistical delay constraint, is more suitable to be adopted as the performance metric for the cache-enabled wireless networks.

Denote $EC_m^k\left(\theta_m^k\right)$ as the effective capacity of user k associated with BS m, and it can be calculated by

$$EC_m^k\left(\theta_m^k\right) = -\frac{1}{\theta_m^k T}\ln \mathbb{E}\left[e^{-\theta_m^k T R_m^k}\right] = -\frac{1}{\theta_m^k T}\ln \mathbb{E}_{I_m^k, I_m^k}\left[e^{-\theta_m^k TBW\log_2(1+\frac{p_m h_m^k}{\sigma^2+I_m^k})}\right], \quad (2)$$

where $\mathbb{E}_x(y)$ means the expectation of y over x, T is time duration of the transmission, and θ_m^k is the quality of service (QoS) exponent that characterizes the steady-state delay violation probability of a transmission link. According to the effective capacity theory, the QoS exponent and the steady-state delay violation probability for a tandem transmission link with constant packet size L, constant sending data rates v from the video content server, and N_h hops, would satisfy the following equation

$$-\theta_m^k = \lim_{D^{k,\max}\to\infty}\frac{\log \Pr\left(D_m^k > D^{k,\max}\right)}{D^{k,\max} - N_h L/v} \quad (3)$$

where $D^{k,\max}$ is the maximum tolerable delay of user k, D_m^k is the perceived delay of user k associated with BS m, and $\Pr\left(D_m^k > D^{k,\max}\right)$ is the steady-state delay violation probability [29]. For a large $D^{k,\max}$, equation (3) can be further derived as

$$\Pr\left(D_m^k > D^{k,\max}\right) \approx e^{-\theta_m^k\left(D^{k,\max} - N_h L/v\right)}. \quad (4)$$

Based on equation (4), it can be seen that a smaller QoS exponent θ_m^k defines a looser delay constraint while a larger θ_m^k imposes a more stringent delay constraint.

Combined equation (2) and (4), the effective capacity model provides a concise metric characterizing the maximum data rates that a transmission link can support, under the statistical QoS delay constraint. As for the cache-enabled wireless networks, if the content requested by one user has been cached, it would be delivered from BS's cache directly and the transmission is one-hop. On the contrary, if the content has not been cached, it should be fetched from the remote content server firstly, and the overall transmission would be multiple hops [6]. According to equation (4), whether or not the content has been cached would make the difficulty of delay guarantee totally different. Denote $\theta_m^{k(c)}$ as the QoS exponent corresponding to the one-hop transmission from BSs' cache, and $\theta_m^{k(b)}$ corresponding to the multiple hops transmission from the remote content server, respectively. Then, we have the following proposition.

Proposition 1: To guarantee the equivalent end-to-end delay experience, $\theta_m^{k(c)}$ and $\theta_m^{k(b)}$ would have the following relation

$$\theta_m^{k(b)} = \frac{D^{k,\max} - \frac{L}{v}}{D^{k,\max} - \frac{N_h L}{v}}\theta_m^{k(c)}. \quad (5)$$

Proof: To guarantee the equivalent end-to-end delay experience, the delay violation probability corresponding to the situations that the content has been cached and not cached should be equal, i.e.,

$$e^{-\theta_m^{k(b)}\left(D^{k,\max} - N_h L/v\right)} = e^{-\theta_m^{k(c)}\left(D^{k,\max} - N_h' L/v\right)}, \quad (6)$$

where N_h is the hop number of the content delivery link when the content has not been cached, while N_h' is the counterpart when the content has been cached. Note that when the content has been cached at the BS's cache, the transmission link has only one hop, which leads to $N_h' = 1$. Then, the equation (5) can be easily derived, which completes the proof of Proposition 1. ■

From Proposition 1, it is easy to derive that $\theta_m^{k(b)} > \theta_m^{k(c)}$ would always hold. This means that with the same statistical delay guarantee requirement, the one hop transmission link would have larger effective capacity and can support larger maximum data rates than those of the multi hop transmission link. In general, based on equation (2)-(5), effective capacity provides a concise end-to-end delay-aware data rates performance metric, and can characterize mobile edge caching's impact on the content transmission comprehensively. Hence, it is a more suitable performance metric for the cache-enabled wireless networks.

III. PROBLEM FORMULATION

The overall goal of this paper is to design efficient strategies for the content placement phase and content delivery phase, respectively. As the content placement and content delivery operate with different time scales, the corresponding problems are formulated and solved separately and respectively.

A. CONTENT PLACEMENT PROBLEM

The content placement phase usually happens at the off-peak hour before any contents request and delivery happen. The target of the content placement scheme is to obtain the optimal set of contents for each BS to cache, by taking into account users' diverse preferences and active levels. In particular, it should be noted that during the content placement phase, the user-BS association relation is not known.

Due to the flexible user-BS association relation, as well as diverse users' active levels, it would be complicated to derive the optimal content placement decision. Hence, the user-BS association should be integrated into the content placement problem.

For notational convenience, denote $\overline{EC}_m^k$ and $\underline{EC}_m^k$ as

$$\overline{EC}_m^k = EC_m^k\left(\theta_m^{k(c)}\right) = -\frac{1}{\theta_m^{k(c)}T}\ln\mathbb{E}\left[e^{-\theta_m^{k(c)}TR_m^k}\right] \quad (7)$$

and

$$\underline{EC}_m^k = EC_m^k\left(\theta_m^{k(b)}\right) = -\frac{1}{\theta_m^{k(b)}T}\ln\mathbb{E}\left[e^{-\theta_m^{k(b)}TR_m^k}\right], \quad (8)$$

respectively. Besides, define $x_m^f \in \{0,1\}$ as the variable indicating whether content f would be cached by BS m or not. Then, the content placement optimization problem is formulated as

$$\text{P1:} \max_{\left\{x_m^f, \tilde{a}_m^k\right\}} \sum_{k\in\mathcal{K}}\lambda^k\sum_{m\in\mathcal{M}}\tilde{a}_m^k\bar{U}_m^k \quad (9a)$$

$$\text{s.t.}: \quad (9b)$$

$$C1.1: \sum_{f\in\mathcal{F}} x_m^f \le X_m^{MAX},\ \forall m\in\mathcal{M}, \quad (9c)$$

$$C1.2: x_m^f\in\{0,1\},\ \forall m\in\mathcal{M},\ f\in\mathcal{F} \quad (9d)$$

$$C1.3: \tilde{a}_m^k\in\{0,1\},\ \forall m\in\mathcal{M},\ k\in\mathcal{K}, \quad (9e)$$

$$C1.4: \sum_{m\in\mathcal{M}}\tilde{a}_m^k \le 1,\ \forall k\in\mathcal{K}, \quad (9f)$$

where

$$\begin{aligned}\bar{U}_m^k &= \sum_{f\in\mathcal{F}} p^{k,f}\log\frac{\left[x_m^f\overline{EC}_m^k + \left(1-x_m^f\right)\underline{EC}_m^k\right]}{\sum_{k\in\mathcal{K}}\lambda^k\tilde{a}_m^k}\\ &= \sum_{f\in\mathcal{F}} p^{k,f}\log\left[x_m^f\overline{EC}_m^k + \left(1-x_m^f\right)\underline{EC}_m^k\right]\\ &\quad - \sum_{f\in\mathcal{F}} p^{k,f}\log\left(\sum_{k\in\mathcal{K}}\lambda^k\tilde{a}_m^k\right)\end{aligned} \quad (10)$$

denotes the expected utility of user k associated with BS m, and a logarithmic utility function is adopted, which is able to provide proportional fairness among users [30]. Note that the $\sum_{k\in\mathcal{K}}\lambda^k a_m^k$ in equation (10) denotes the expected number of users associated with BS m, and the λ^k is introduced because each user is not active all the times, but would send request with a certain probability for one time slot. Hence, $\frac{\left[x_m^f\overline{EC}_m^k+\left(1-x_m^f\right)\underline{EC}_m^k\right]}{\sum_{k\in\mathcal{K}}\lambda^k\tilde{a}_m^k}$ denotes the expected effective capacity of user k associated with BS m when requesting content f. Besides, the $\tilde{a}_m^k$ is not the realistic user-BS association decision for the content delivery phase, but the auxiliary variable indicating the possible user-BS association relation, which reflects the impact of the content delivery on the content caching.

B. CONTENT DELIVERY PROBLEM

The content delivery is assumed to operate in a time-slot manner. For one specific time slot, a portion of users would send their content requests, and the possibility of one user sending request is characterized by the active level $\{\lambda^k\}$. Besides, it is assumed that each user would send at most one content request during one time slot. Given BSs' cache status and users' content requests, users' effective capacity performance would be impacted by the traffic load of the associated BS. Hence, the user-BS association would be optimized to provide high effective capacity for users.

Define $a_m^k \in \{0,1\}$ as the user-BS association indicator variable denoting whether user k would be associated with BS m or not. For one specific time slot of the content delivery phase, aiming at improving each user's effective capacity performance, a user-BS association optimization problem is formulated as

$$\text{P2:} \max_{\{a_m^k\}} \sum_{k\in\mathcal{K}_A}\sum_{m\in\mathcal{M}} a_m^k U_m^k$$

$$\text{s.t.}:$$

$$C2.1: \sum_{m\in\mathcal{M}} a_m^k \le 1, \quad \forall k\in\mathcal{K}_A,$$

$$C2.2: a_m^k\in\{0,1\},\ \forall m\in\mathcal{M},\ k\in\mathcal{K}_A, \quad (11)$$

where

$$\begin{aligned}U_m^k &= \sum_{f\in\mathcal{F}} s^{k,f}\log\frac{\left[x_m^f\overline{EC}_m^k + \left(1-x_m^f\right)\underline{EC}_m^k\right]}{\sum_{k\in\mathcal{K}_A} a_m^k}\\ &= \sum_{f\in\mathcal{F}} s^{k,f}\log\left[x_m^f\overline{EC}_m^k + \left(1-x_m^f\right)\underline{EC}_m^k\right]\\ &\quad - \log\sum_{k\in\mathcal{K}_A} a_m^k\end{aligned} \quad (12)$$

is the utility of user k if it is associated with BS m, $\mathcal{K}_A$ denotes the set of active users for the considered time slot, and $s^{k,f}\in\{0,1\}$ is the parameter indicating whether content f is requested by user k. Besides, note that the $s^{k,f}$ and $\mathcal{K}_A$ would be known before the user-BS association optimization is implemented at each time slot.

IV. SOLUTIONS TO THE OPTIMIZATION PROBLEMS

In this section, we would solve the content placement problem, i.e., P1, and content delivery problem, i.e, P2, respectively. In particular, we focus on the content delivery phase at first, and design a cache-aware user-BS association algorithm for problem P2. Then, we investigate the content placement problem, and propose an association-aware content placement algorithm.

A. CACHE-AWARE USER-BS ASSOCIATION STRATEGY

For each time slot of the content delivery, the cache status, i.e., $\left\{x_m^f\right\}$, and the content requests of the users, i.e., $\left\{s^{k,f}\right\}$

are fixed. Denote

$$w_m^k = \sum_{f\in\mathcal{F}} s^{k,f} \log\left[x_m^f \overline{EC}_m^k + \left(1-x_m^f\right)\underline{EC}_m^k\right], \quad (13)$$

then w_m^k would be a known parameter before the user-BS association optimization.

Owing to the integer nature of the variables, the user-BS association problem is an integer programming problem, which is complicated to be solved. In order to make it more tractable, the variable relaxation technique is firstly used, where the $a_m^k \in \{0,1\}$ is relaxed to be be continuous in the range of $[0,1]$. Besides, an auxiliary variable $K_m = \sum_{k\in\mathcal{K}_A} a_m^k$ is introduced, which denotes the number of users associated with BS m.

Then, the problem P2 can be transformed into

$$\begin{aligned} \text{P2.1}: \max_{\{a_m^k, K_m\}} & \sum_{k\in\mathcal{K}_A}\sum_{m\in\mathcal{M}} a_m^k w_m^k - \sum_{m\in\mathcal{M}} K_m \log K_m \\ \text{s.t.}: & \\ & C2.1:, \sum_{m\in\mathcal{M}} a_m^k \le 1, \quad \forall k\in\mathcal{K}_A, \\ & C2.3: \sum_{k\in\mathcal{K}_A} a_m^k = K_m, \quad \forall m\in\mathcal{M}, \\ & C2.4: a_m^k \in [0,1], \quad \forall m\in\mathcal{M},\ k\in\mathcal{K}_A. \end{aligned} \quad (14)$$

Note that the problem P2.1 is a convex optimization problem, which can be solved by the Lagrangian dual method in the dual domain. Specifically, the Lagrangian function of P2.1 with constraint C2.3 can be denoted as

$$\begin{aligned} L\left(\left\{a_m^k\right\},\{K_m\},\{\beta_m\}\right) &= \sum_{k\in\mathcal{K}_A}\sum_{m\in\mathcal{M}} a_m^k w_m^k \\ &- \sum_{m\in\mathcal{M}} K_m \log K_m + \sum_{m\in\mathcal{M}} \beta_m \left(K_m - \sum_{k\in\mathcal{K}_A} a_m^k\right) \end{aligned} \quad (15)$$

where $\{\beta_m\}$ is the Lagrangian multipliers for constraint C2.3. Then, the Lagrangian dual function is

$$g\left(\{\beta_m\}\right) = \max_{\{a_m^k, K_m\}} L\left(\left\{a_m^k\right\},\{K_m\},\{\beta_m\}\right), \quad (16)$$

and the dual problem is

$$\min_{\{\beta_m\}} g\left(\{\beta_m\}\right). \quad (17)$$

On the condition that $\{\beta_m\}$ is given, the optimal user-BS association solution can be deduced as

$$a_m^{k*} = \begin{cases} 1, & \text{if } m = \arg\max\limits_{m'\in\mathcal{M}} \left(w_{m'}^k - \beta_{m'}\right), \\ 0, & \text{else.} \end{cases} \quad (18)$$

Besides, the optimal number of users associated with BS m, i.e., K_m^* can be deduced by

$$\frac{\partial L\left(\left\{a_m^k\right\},\{K_m\},\{\beta_m\}\right)}{\partial K_m} = 0, \quad (19)$$

which leads to

$$K_m^* = e^{\beta_m - 1}. \quad (20)$$

Furthermore, the optimal Lagrangian multiplier β_m^* can be obtained by subgradient method [11]. During the tth step of the iterations, β_m could be updated according to

$$\beta_m(t+1) = \beta_m(t) - \Gamma(t)\left(K_m(t) - \sum_{k\in\mathcal{K}_A} a_m^k(t)\right), \quad (21)$$

where $\Gamma(t)$ is the step size.

Note that the user-BS association solution, i.e., equation (18) is similar with the counterpart in [22], but there are significant differences. The weighted sum of data rates and backhaul traffic reduction is adopted in [22] as the metric to decide user-BS association, and the metric is network-based where the weight parameter needs to be adjusted manually to balance the network throughput and backhaul traffic reduction. On the contrary, the effective capacity based utility in this paper is user-centric, and is able to characterize not only the end-to-end delay and data rates simultaneously, but also their mutual influence in a concise manner. As a result, it is advantageous for deciding the user-BS association in a cache-enabled wireless network.

Based on the derivation, we design a cache-aware user-BS association algorithm, which is shown in the following **Algorithm 1**.

Algorithm 1 Cache-Aware User-BS Association Algorithm

Initialize $\{\boldsymbol{\beta}\}$;
repeat
  Obtain the $\left\{a_m^{k*}\right\}$ based on equation (18);
  Obtain the $\left\{K_m^*\right\}$ based on equation (20);
  Update $\{\boldsymbol{\beta}\}$ based on subgradient method;
until Convergence;

B. ASSOCIATION-AWARE CONTENT PLACEMENT STRATEGY

The content placement scheme would be designed by solving the problem P1. As there are two kinds of integer variables $\left\{x_m^f\right\}$ and $\left\{\tilde{a}_m^k\right\}$ that are coupled with each other, it is an integer programming problem, which is prohibitive to find the optimal solution. Hence, we adopt an iterative scheme that make the optimization of $\left\{x_m^f\right\}$ and $\left\{\tilde{a}_m^k\right\}$ iteratively and separately.

Firstly, when the $\left\{x_m^f\right\}$ is fixed, the overall problem of P1 would be reduced into the following form:

$$\begin{aligned} \text{P1.1}: \max_{\{\tilde{a}_m^k\}} & \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} \lambda^k \tilde{a}_m^k \bar{w}_m^k - \sum_{m\in\mathcal{M}} \lambda^k \tilde{a}_m^k \log \sum_{k\in\mathcal{K}} \lambda^k \tilde{a}_m^k \\ \text{s.t.}: & \ C1.3, C1.4 \end{aligned} \quad (22)$$

where $\bar{w}_m^k = \sum_{f\in\mathcal{F}} p^{k,f} \log\left[x_m^f \overline{EC}_m^k + \left(1-x_m^f\right)\underline{EC}_m^k\right]$ would be invariable.

Denote $b_m^k = \lambda^k \tilde{a}_m^k$, the problem P1.1 can be further transformed into the following form:

$$\begin{aligned} \text{P1.2:} \quad & \max_{\{b_m^k\}} \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} b_m^k \bar{w}_m^k - \sum_{m\in\mathcal{M}} b_m^k \log \sum_{k\in\mathcal{K}} b_m^k \\ & \text{s.t.}: \\ & C1.5: \sum_{m\in\mathcal{M}} b_m^k \le 1, \quad \forall k \in \mathcal{K}, \\ & C1.6: b_m^k \in \left\{0, \lambda^k\right\}, \ \forall m \in \mathcal{M}, \ k \in \mathcal{K}, \end{aligned} \tag{23}$$

The problem P1.2 is a user-BS association optimization problem, which is similar with problem P2. The solution of $\{b_m^k\}$ can be obtained based on Algorithm 1 with minor modification. Specifically, as the feasible region of the variable b_m^k is $\left\{0, \lambda^k\right\}$, the optimal solution of $\{b_m^k\}$ can be denoted as

$$b_m^{k*} = \begin{cases} \lambda^k, & \text{for } m = \arg\max_{m'} \left(\bar{w}_{m'}^k - \bar{\beta}_{m'}\right), \\ 0, & \text{else.} \end{cases} \tag{24}$$

where the $\bar{\beta}_m$ is the Lagrangian multiplier of problem P1.2.

When the variable $\left\{b_m^k\right\}$ is decided, the problem P1 is simplified into a problem with the variable $\{x_m^f\}$, i.e.,

$$\begin{aligned} \text{P1.3:} \quad & \max_{\left\{x_m^f\right\}} \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}}\sum_{f\in\mathcal{F}} b_m^k p^{k,f} U_m^{k,f} - VB \\ & \text{s.t.:} \quad C1.1, C1.2 \end{aligned} \tag{25}$$

where $U_m^{k,f} = \log\left[x_m^f \overline{EC}_m^k + \left(1 - x_m^f\right)\underline{EC}_m^k\right]$ and $VB = \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} b_m^k \log \sum_{k\in\mathcal{K}} b_m^k$. Specifically, the VB is invariable when the $\{b_m^k\}$ is given.

It can be found that different BSs' content placement has no relation with each other. Hence, the problem P1.3 can be decomposed into M subproblems, and each can be denoted as

$$\begin{aligned} \text{P1.4:} \quad & \max_{\left\{x_m^f\right\}} \sum_{k\in\mathcal{K}}\sum_{f\in\mathcal{F}} b_m^k p^{k,f} \log\left[x_m^f \overline{EC}_m^k + \left(1 - x_m^f\right)\underline{EC}_m^k\right] \\ & \text{s.t.}: \quad C1.1, C1.2, \end{aligned} \tag{26}$$

Due to the 0-1 integer nature of x_m^f, the objective function of problem P1.4 is equivalent to the following form

$$\max_{\left\{x_m^f\right\}} \sum_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}} b_m^k p^{k,f} \left[x_m^f \left(\log \overline{EC}_m^k - \log \underline{EC}_m^k\right) + \log \underline{EC}_m^k\right] \tag{27}$$

Then the optimal content placement solution is derived as

$$x_m^f = \begin{cases} 1, & \text{if } f \in \left\{\Delta U_{m(1)}^f, \cdots, \Delta U_{m(X_m^{MAX})}^f\right\}, \\ 0, & \text{else,} \end{cases} \tag{28}$$

where $\Delta U_m^f = \sum_{k\in\mathcal{K}} b_m^k p^{k,f} \left(\log \overline{EC}_m^k - \log \underline{EC}_m^k\right)$ and $\Delta U_{m(i)}^f$ denotes the ith largest item in $\left\{\Delta U_m^f\right\}$. Note that the equation (28) is intuitive because the optimal content placement scheme would be to cache the contents that can provide the largest utility improvement.

Based on the derivation above, an association-aware content placement algorithm is designed, which is shown in the following **Algorithm 2**. Note that the algorithm would be implemented iteratively, and would not stop until no performance improvement can be achieved.

Algorithm 2 Association-Aware Content Placement Algorithm

Initialize $\left\{x_m^f\right\}$ based on the global popularity, i.e, each BS caches the most popular contents;
repeat
  Obtain the $\left\{\tilde{b}_m^{k*}\right\}$ based Algorithm 1;
  Obtain $\left\{x_m^{f*}\right\}$ based on equation (28);
until No performance improvement could be obtained.

C. IMPLEMENTATION AND COMPLEXITY ANALYSIS

In this subsection, the implementation and complexity issues of the proposed algorithms would be analyzed.

During the content placement phase, the content provider would implement the association-aware content placement algorithm, i.e., Algorithm 2, to update the contents of each BS's cache. The Algorithm 2 needs the information of each user's preference and active level. These information can be estimated based on the log of users' content requests in the past, which are collected by the content provider during the daily operation. After obtaining the content placement solution $\{x_m^f\}$, the corresponding contents that need to be cached would be pushed to BSs' cache from the content server. As for the content delivery phase, the user-BS association would be optimized based on Algorithm 1 after the users send the content requests. The Algorithm 1 needs to be implemented in a real-time manner, based on the information of BSs' cache status, the users content requests, as well as users channel condition that could be estimated by the BSs.

Besides, the complexity of Algorithm 1 can be denoted by $\mathcal{O}\left(M^2K/\left(1/\epsilon^2\right)\right)$, where ϵ is the convergence accuracy when the subgradient method is adopted to obtain β_m^*. Besides, denote $I_{\max}$ as the maximum iteration number of the Algorithm 2, then the complexity of Algorithm 2 can be characterized by $\mathcal{O}\left[I_{\max}\left(M^2K/\left(1/\epsilon^2\right) + MF^2\right)\right]$. Hence, both the Algorithm 1 and the Algorithm 2 are with polynomial time-complexity, which facilitates their implementation in the practical wireless network scenarios.

V. SIMULATION RESULTS

In this section, the performance of the proposed algorithms is evaluated based on Monte Carlo-based simulations. All experiments are conducted with Matlab R2018b on a X64-based laptop. The laptop is equipped with a 4-core Intel(R) Core(TM) i7-8550U CPU of speed 1.80 GHz and a memory of 16 GB.

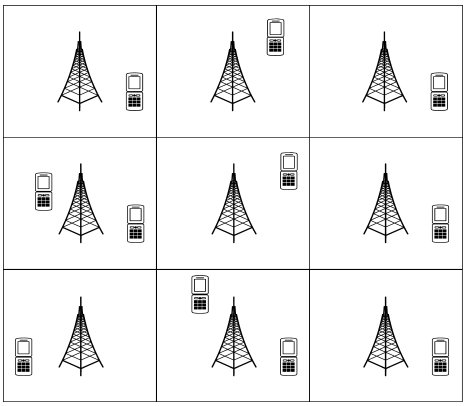

FIGURE 1. Network scenario used in simulations.

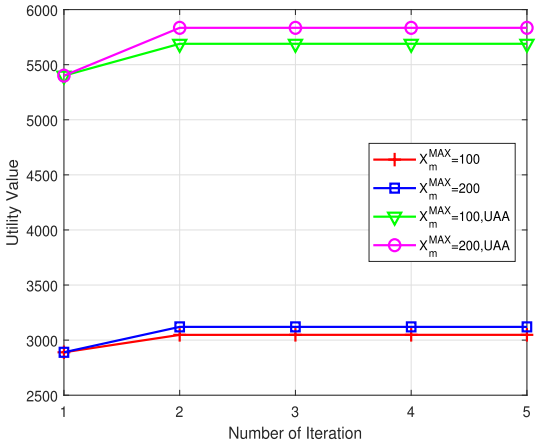

FIGURE 2. The convergence of Algorithm 2.

Specifically, a small-cell network scenario with grid topology is considered in the simulation, where 9 small-cell BSs (SBSs) are deployed in a 60m×60m area as illustrated in Fig. 1. Besides, 400 users are uniformly distributed in the considered area. The overall system bandwidth is 20MHz and is reused by all the SBSs. The transmit power of each SBS is fixed as 23dBm. The wireless channel is assumed to experience Rayleigh fading with mean 1. The $L(d) = 37 + 30\ log(d)$ is used as the pathloss model, where d is the distance between the user and the SBS. The power density of AWGN is -174 dBm. The $\theta_m^{k(b)}$ and $\theta_m^{k(c)}$ are set to be 10^{-1} and 10^{-4}, respectively. The overall number of the contents in the library is set as 1000. The probability of a request sent by user k that is for content f is characterized by $\frac{(\phi(k))^{-\gamma}}{\sum_{f' \in \mathcal{F}} f'^{-\gamma}}$ where the skewness parameter γ is set as 1. Besides, note that users' diverse preferences for contents can be modeled by $\phi(k)$, which is a function that would return a random permutation of $[1, 2, \cdots, F]$ for each user [31]. Furthermore, without otherwise specified, the users' active level, i.e., λ^k, is assumed to follow a uniformly distribution on the interval [0,1].

Three baseline algorithms are also simulated for comparison with the proposed algorithms in this paper (which would be referred to proposed algorithms, or PA for short in the following).

- Baseline Algorithm 1: The baseline algorithm 1 (or BA 1, for short) adopts the local-popularity based content placement scheme and SINR-MAX based user-BS association scheme. Specifically, based on the SINR-MAX user-BS association scheme, each user would be associated with the SBS with the highest SINR [32]. Based on the local popularity-based content placement scheme, the contents with the highest value of $\sum_{k \in \mathcal{K}_m} \lambda^k p^{k,f}$ would be cached by SBS m until the cache is full, where $\mathcal{K}_m$ denotes the users set which would be associated with SBS m. Besides, the wireless frequency bandwidth would be allocated equally among the active users that send the content requests at each time slot of the content delivery phase.
- Baseline Algorithm 2: The algorithm proposed in [11] is adopted as the baseline algorithm 2 (or BA 2, for short). Specifically, the BA 2 is also an iterative content placement and user-BS association algorithm, which aims at maximizing the weighted sum of users' data rates and the backhaul traffic reduction (The value of the weight is set as 10) [22]. However, based on BA 2, the user would be associated with one certain SBS during the content delivery process without adaptive adjustment.
- Baseline Algorithm 3: The baseline algorithm 3 (or BA 3, for short) is an algorithm which is similar with BA 2, but with some differences in the user-BS association procedure of the content delivery phase. Specifically, during the content placement phase, the BA 3 would implement the same procedures with BA 2 to obtain the content placement decision. However, at each time slot of the content delivery phase, the BA 3 would do the user-BS association optimization again according to users' specific content requests, which could adjust the user-BS association adaptively to maximize the weighted sum of users' data rates and the backhaul traffic reduction.

A. THE CONVERGENCE OF ALGORITHM 2

As the Algorithm 2 is a heuristic iterative algorithm, its convergence performance is evaluated firstly. The value of the function $\sum_{k \in \mathcal{K}} \lambda^k \sum_{m \in \mathcal{M}} \tilde{a}_m^k \bar{U}_m^k$, i.e., the objective function of problem P1, is shown in Fig. 2 with different network setups. Specifically, the cache capacity is set as 100 and 200, respectively. Besides, an extreme situation where all the users are always active (or UAA, for short), i.e., the λ^k equals to 1 for all the users, is also simulated. It can be seen that the network utility improves rapidly and then remains stable within 2 or 3 iterations, no matter what the network setup is. This validates a fast convergence speed of Algorithm 2. Therefore, it can be concluded that the Algorithm 2 has a fast and stable convergence property no matter what the network scenario is.

B. THE IMPACT OF THE CACHE CAPACITY

The impact of the cache capacity on the performance is evaluated and the corresponding results are shown in Fig.3-Fig.5. Specifically, all the results in the figures are averaged over 1000 experiments. Each experiment includes both the content placement phase and content delivery phase, where the content delivery phase consists of 20 time slots.

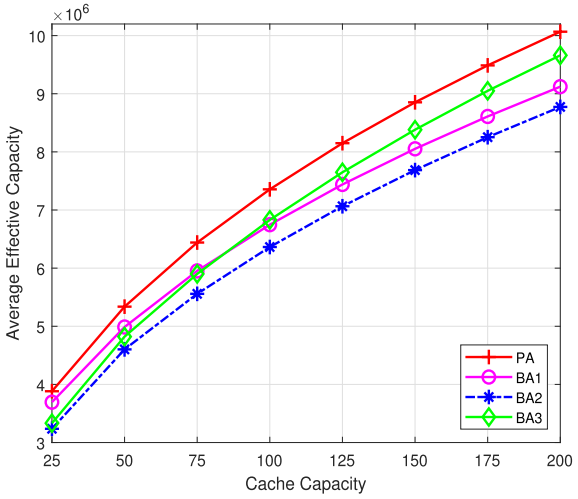

FIGURE 3. The average effective capacity corresponding to different cache capacities.

Fig. 3 shows users' average effective capacity corresponding to different cache capacities. Specifically, the cache capacity is varied from 25 to 200, which corresponds to the situation where the 2.5%-20% of the overall contents could be cached. From Fig. 3, it can be seen that with the increase of the cache capacity, the average effective capacity of all the four algorithms is improved. This is intuitive, because larger cache capacity means more contents can be cached by each BS and delivered locally without fetching from the remote server. As a result, the effective capacity performance of the overall users can be improved. Besides, the Fig. 3 shows that the proposed algorithms always have performance advantages over the baseline algorithms, which means that the proposed algorithms could support larger transmission data rates, under the same statistical delay QoS requirements. Besides, note that the BA 1 adopts a SINR-MAX user-BS association, and BA 2 adopts the static user-BS association decision, which means that both the BA 1 and the BA 2 use the fixed user-BS association during the content delivery phase. However, the BA 1 has a better performance compared with the BA 2, no matter what the cache capacity is. Note that the content placement of BA 1 is based on local-popularity, which takes users' active levels into account. Hence, the performance advantage of BA 1 over BA 2 is enabled by its user-active-level-aware property. Furthermore, the BA 1 has better performance than BA 3 when the cache capacity is not larger than 75, but loses the performance advantage when the cache capacity becomes larger. This is because when the cache capacity is small, the content placement scheme would

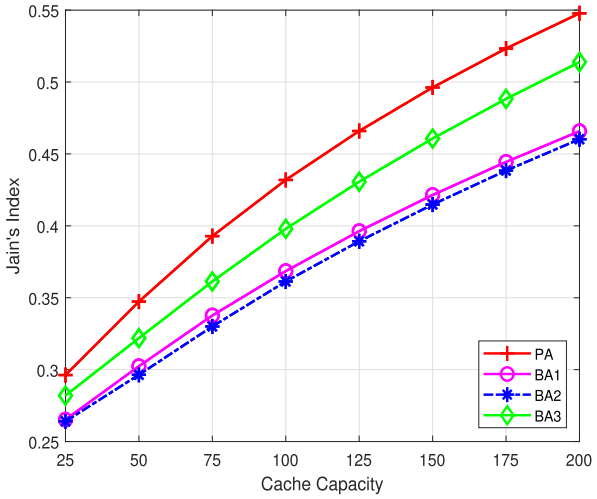

FIGURE 4. The fairness level of different algorithms.

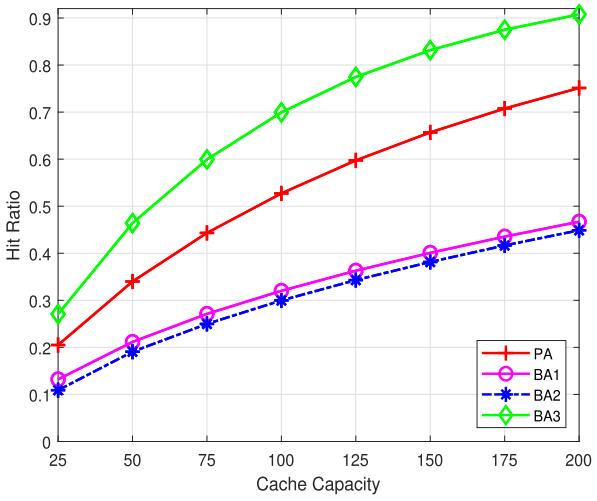

FIGURE 5. The hit ratio of different algorithms corresponding to different active ratios.

play an important role for the performance. As the user-active-level-aware property of BA 1, it has a performance advantage over BA 2. However, with the increase of the cache capacity, the impact of the content placement on the performance would be reduced, while the flexible user-BS association scheme of BA 3 would play a more important role on the performance gap. In contrast with all the baseline algorithms, the proposed algorithms employ both the user-active-level-aware caching scheme and flexible user-BS association scheme, which can reap their benefits simultaneously.

The fairness level of users' effective capacity performance is evaluated and the results are shown in Fig. 4. Specifically, the Jain's index is adopted as the metric, which is calculated by $\eta=\frac{\left[\sum_{k\in\mathcal{K}_A} EC^k(t)\right]^2}{|\mathcal{K}_A|\sum_{k\in\mathcal{K}_A}[EC^k(t)]^2}$, where the $EC^k(t)$ denotes the effective capacity of the user k at time slot t. It can be seen from Fig. 4 that the Jain's index of all the algorithms

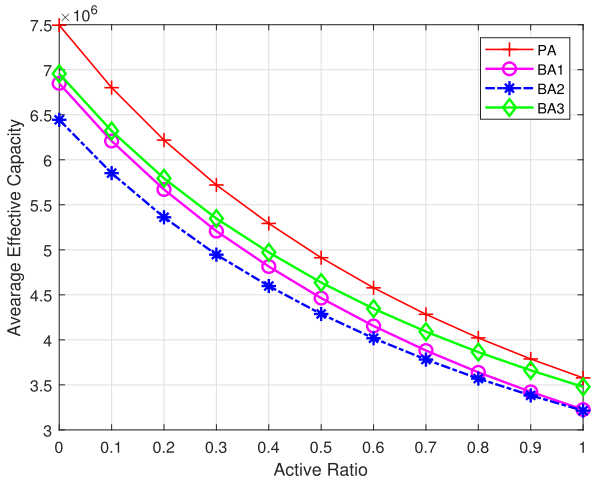

FIGURE 6. The average effective capacity of different algorithms corresponding to different active ratios.

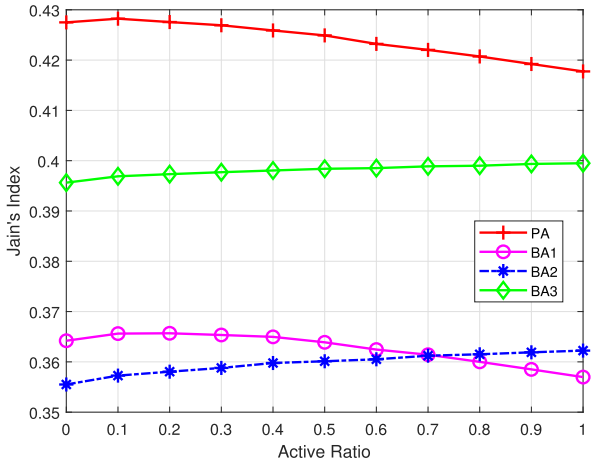

FIGURE 7. The fairness level of different algorithms corresponding to different active ratios.

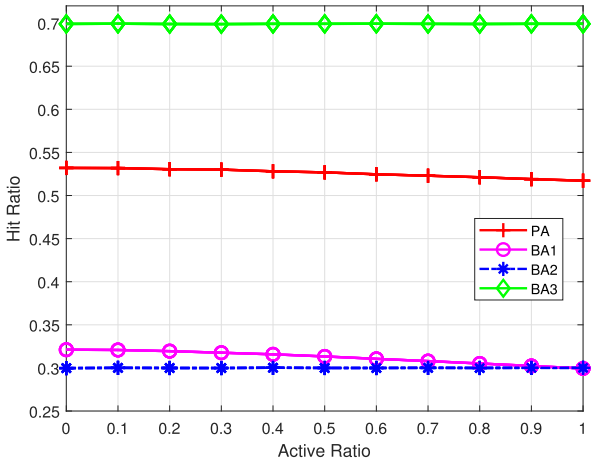

FIGURE 8. The hit ratio of different algorithms corresponding to different active ratios.

increases with the cache capacity becoming larger, which means that the fairness level among users is improved. This improvement is enabled by the fact that larger cache capacity means more users can be served locally and can benefit from the mobile edge caching. As a result, their corresponding effective capacities would be increased and the differences among users' performance are reduced, which means that the fairness level among users is improved. Besides, the Fig. 4 shows that the proposed algorithms have the largest Jain's index, which means the highest fairness level. This validates the proposed algorithms' better capability of guaranteeing users' fairness. Combined with the Fig. 3, we can conclude that the proposed algorithms can not only achieve higher effective capacity performance, but also provide better fairness guarantee among users.

Furthermore, the hit ratio of all the algorithms is shown in Fig. 5. It can be seen that larger cache capacity leads to higher hit ratio for all the algorithms, which is in line with the intuition. Note that the hit ratio of the proposed algorithms is higher than that of the BA 1 and the BA 2 significantly, but is smaller than that of the BA 3. However, combined with Fig. 2 and Fig. 3, it can be deduced that the higher hit ratio of BA 3 does not lead to users' effective capacity utilities improvement. This is because based on BA 3, more users would be associated with the BSs that have been cached their requesting contents. But these BSs may be over-loaded, which are not the best user-BS association solutions. Hence, the higher hit ratio does not equal to the better user performance. On the contrary, we can deduce that the proposed algorithms do not pursue the cache hit maximization unilaterally, but would make the user-centric utility optimization, by considering the users' content requests, cache status, as well as each BS's traffic load jointly.

C. THE IMPACT OF THE USERS' ACTIVE LEVEL

The impact of users' active levels on algorithms' performance is evaluated in this subsection. Specifically, in order to differentiate the overall active levels of the users, we assume that a portion of users are always active, i.e., the λ_k equals to 1 for these users, and they would send requests at each time slot of the content delivery phase. The percentage of the users who are always active (which is referred to active ratio in the following) are varied, and the corresponding results are shown in Fig. 6 - Fig.8. Besides, all the results in the figures are obtained based on the implementation similar with the experiments in Subsection V.B.

The average effective capacity performance is shown in Fig. 6 as the active ratio is varied. It can be seen that with the increase of the active ratio, the average effective capacity performance of all the algorithms decreases. This is because there would be more active users who would always send content requests when the active ratio becomes larger. This leads to the situation that the wireless spectrum would be shared by more users at each time slot, hence, the percentage of wireless spectrum occupied by each user would be reduced. As a result, the corresponding effective capacity of each user's would decline. Besides, note that

the effective capacity of the proposed algorithms' is always larger than that of the other algorithms, which means that the proposed algorithms can perform better no matter what the overall active level is. This advantage is brought by the fact that users' active levels are integrated into the content placement problem, and then the Algorithm 2 proposed in this paper is an active-level-aware content placement algorithm. But it should be noted that the performance gap of the proposed algorithms over the BA 3 becomes narrow with the increase of the active ratio. The reason is that larger active ratio not only means more users would be active, but also means that the similarity of users' active levels increases. As a result, the advantages of the proposed algorithms brought by the active-level-aware content placement are reduced and the corresponding performance gap becomes narrow.

Besides, the fairness levels are shown in Fig. 7 with different active ratios. Obviously, the proposed algorithms have a higher Jain's index value, which means the proposed algorithms achieve a higher fairness level. Combined with the Fig. 6, we can conclude that the proposed algorithms maintain a better performance compared with existing algorithms in terms of not only the average effective capacity, but also the fairness level among users.

Furthermore, the hit ratio performance is shown in Fig. 8. It can be seen that no matter what the active ratio is, the BA 3 achieves the highest hit ratio, which is similar with the phenomenon of Fig. 5. Considering the worse effective capacity of BA 3 shown in Fig. 6, it can be deduced that BA 3 achieves higher hit ratio, i.e., more backhaul traffic reduction, at the cost of users' effective capacity performance degradation. In contrast, although the hit ratio of the proposed algorithms is smaller than that of the BA 3, the effective capacity of the proposed algorithm is higher than that of the BA 3. This means that proposed algorithms make a better balance between backhaul traffic reduction and the effective capacity improvement, which is a more satisfying solution from the perspective of users' utility optimization.

VI. CONCLUSION

In this paper, the maximization of users' utilities is investigated by joint optimization of content placement and content delivery in a cache-enabled wireless network. In particular, the effective capacity, which can characterize mobile edge caching's impact on the data rates and delay of the content transmission simultaneously, is introduced as user's performance metric. The content placement problem and content delivery problem are formulated, respectively, where the users' diverse preferences and active levels are taken into account. Then, the association-aware content placement algorithm and cache-aware user-BS association algorithm are proposed, respectively. Simulation results demonstrate that the proposed algorithms can achieve larger average effective capacity and provide higher-level fairness guarantee for users, compared with the existing algorithms.

REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and trends, 2017–2022," Cisco, White Paper, 2017.

[2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, New York, NY, USA, 2007, pp. 1–14. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298309

[3] 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.

[4] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[5] H. Zhang, Y. Qiu, X. Chu, K. Long, and V. C. M. Leung, "Fog radio access networks: Mobility management, interference mitigation, and resource optimization," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 120–127, Dec. 2017.

[6] 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.

[7] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Commun. ACM*, vol. 49, no. 1, pp. 101–106, 2006. [Online]. Available: http://doi.acm.org/10.1145/1107458.1107462

[8] M. A. Salahuddin, J. Sahoo, R. Glitho, H. Elbiaze, and W. Ajib, "A survey on content placement algorithms for cloud-based content delivery networks," *IEEE Access*, vol. 6, pp. 91–114, Feb. 2018.

[9] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[10] G. Yu, Z. Zhang, F. Qu, and G. Y. Li, "Ultra-dense heterogeneous networks with full-duplex small cell base stations," *IEEE Netw.*, vol. 31, no. 6, pp. 108–114, Nov./Dec. 2017.

[11] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[12] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100–1113, Jun. 2014.

[13] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

[14] G. Yu, Y. Jiang, L. Xu, and G. Y. Li, "Multi-objective energy-efficient resource allocation for multi-RAT heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2118–2127, Oct. 2015.

[15] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, 3rd Quart., 2018.

[16] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[17] D. Huang, X. Tao, C. Jiang, Y. Li, and J. Lu, "Latency-efficient video streaming in metropolis: A caching framework," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[18] F. Guo, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Joint optimization of caching and association in energy-harvesting-powered small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6469–6480, Jul. 2018.

[19] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.

[20] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[21] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *IEEE Access*, vol. 4, pp. 8625–8633, 2016.

[22] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 3521–3525.

[23] R. Haw, S. M. A. Kazmi, K. Thar, M. G. R. Alam, and C. S. Hong, "Cache aware user association for wireless heterogeneous networks," *IEEE Access*, vol. 7, pp. 3472–3485, 2019.
[24] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
[25] M. Chen, W. Saad, C. Yin, and M. Debbah, "Echo state networks for proactive caching in cloud-based radio access networks with mobile users," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3520–3535, Jun. 2017.
[26] D. Liu and C. Yang, "Caching at base stations with heterogeneous user demands and spatial locality," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1554–1569, Feb. 2019.
[27] S. Müller, O. Atan, M. van der Schaar, and A. Klein, "Context-aware proactive content caching with service differentiation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1024–1036, Feb. 2017.
[28] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *CoRR*, vol. abs/1707.07435, Jul. 2017. [Online]. Available: http://arxiv.org/abs/1707.07435
[29] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," in *Proc. 1st Int. Conf. Broadband Netw.*, Oct. 2004, pp. 527–536.
[30] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, 1998.
[31] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commu.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.
[32] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

WENPENG JING received the B.S. degree in communication engineering from Shandong University, in 2012, and the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, in 2017, where he is currently holding a postdoctoral position with the Beijing University of Posts and Telecommunications. His research interests include mobile edge caching, radio resource allocation, energy efficiency optimization, and interference management in heterogeneous networks.

XIANGMING WEN received the M.S. and Ph.D. degrees in information and communication engineering from the Beijing University of Posts and Telecommunications. He is currently the Director of the Beijing Key Laboratory of Network System Architecture and Convergence, where he has managed several projects related to open wireless networking. He is also the Vice President of the Beijing University of Posts and Telecommunications. His current research interests focus on radio resource and mobility management, software-defined wireless networks, and broadband multimedia transmission technology.

ZHAOMING LU received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2012. He joined the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, in 2012. His research interests include open wireless networks, QoE management in wireless networks, software-defined wireless networks, and cross-layer design for mobile video applications.

HAIJUN ZHANG (M'13–SM'17) is currently a Full Professor with the University of Science and Technology Beijing, China. He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia (UBC), Canada. He received the IEEE CSIM Technical Committee Best Journal Paper Award, in 2018, and the IEEE ComSoc Young Author Best Paper Award, in 2017. He serves as the Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS NETWORKING, and the IEEE COMMUNICATIONS LETTERS.