IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 18, 2018, accepted December 1, 2018, date of publication December 12, 2018,
date of current version January 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886361

Energy Efficient Downlink Resource Allocation for D2D-Assisted Cellular Networks With Mobile Edge Caching

YUANFEI LIU, YING WANG, RUIJIN SUN, SACHULA MENG, AND RUNCONG SU
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
Corresponding author: Ying Wang (wangying@bupt.edu.cn)

This work was supported by the National Key Project under Grant 2017ZX03001009.

ABSTRACT In this paper, a downlink device-to-device (D2D)-assisted cellular networks with mobile edge caching, where most popular video files are independently cached in D2D users and cellular base station (BS), are studied. In the considered system model, each user may obtain the requested video from the cache of BS or/and D2D users surrounding them. According to the different collaborative schemes of BS caching and D2D caching, it can be divided into two different resource allocation schemes. In the hybrid caching transmission scheme, users could adopt the BS caching mode or alternatively the D2D caching mode. In the joint caching transmission scheme, each user may obtain the requested files from the BS server and the adjacent D2D users, simultaneously. By taking the required data rate and the interference constraint into account, we formulate two joint resource allocation problems integrating link selection, channel allocation, and power control to maximize the system energy efficiency (EE). Leveraging on the Dinkelbach method, the EE optimization problems are transformed into mixed-integer nonlinear programming problems and can be decomposed into three subproblems: link selection, channel allocation, and power control. To solve these complicated problems, we propose two optimization algorithms that consist of a modified branch and bound method as well as Lagrange dual decomposition approach. The simulation results demonstrate the superiority of these two proposed algorithms in improving system throughput and EE compared with other algorithms.

INDEX TERMS MEC, D2D caching, link selection, subchannel allocation, power allocation.

I. INTRODUCTION

With the proliferation of smart phones and emergence of various applications, mobile data has been explosively increasing in recent years. As one of the most popular services among mobile users, it is predicted that video service will account for more than 78% of the total mobile data traffic in 2021 [1]. Such tremendous growth in the video traffic bring a stringent challenge in next generation networks. In addition, video services set a higher request for low latency, which make the resource allocation extremely complex [2]. In order to cope with these ongoing increasing demands for high rate and low delay, it is of great significance to achieve the efficient transmission of a large amount of video and reduce latency under the limited resource conditions.

Mobile edge caching (MEC) has recently been deemed as one of key technologies for wireless content delivery networks to reduce peak-time traffic, latency and the requirement for expensive high capacity backhaul links [3]–[5]. Its main idea is to provide cloud-caching capabilities at the network edge, either at the base stations (BSs) or/and user devices to bring the cached content much closer to mobile users. Mobile edge caching can be divided into BS caching and device-to-device (D2D) caching according to geographical location. With the widespread caching, users can get cached contents from adjacent edge servers instead of remote cloud servers [6]. Thus, high-volume and latency-sensitive video contents can benefit the most from MEC [7]. Similarly, according to the mode of the content caching, MEC can be divided into coded caching [8], [9] and uncoded caching. Coded caching is mainly applied to multicast and can reduce delivery rate. However, it has to be designed carefully in the content placement. Uncoded caching is a conventional caching approach. In this paper, we are concentrating on resource allocation during the caching delivery phase, using unicast transmissions, so only uncoded caching is considered.

The implemention of BS caching shortens the wireless transmission distance and leads to a lower delay, which improves the service quality of users dramatically compared with the traditional content delivery networks [10]. Device-to-device (D2D) caching can not only unload the cache task on the edge devices, thereby saving a large amount of air interface bandwidth and energy consumption, but also make use of resource-rich user devices to meet the needs of real-time interaction and low latency transmission [11].

The combination of BS caching and D2D caching will bring many clear advantages, such as reducing the traffic in backhaul and improving the energy efficiencies (EE) and network scalability [12], [13]. Unfortunately, designing such integrated mechanism is challenging due to following aspects. First, using caching technology can improve the rate of video transmission, but it also brings the increase of power consumption due to the co-channel interference. It is necessary to consider the balance between the increase of date rate and power consumption. And with the different conditions of BS caching and D2D caching, it gets even more complex. Second, D2D caching can cause inevitable interference to BS caching as a result of spectrum reuse. The existence of mutual interference results in the decline of system performance, which makes the resource management problem much more complicated. Third, the jointly optimizing of link selection, channel allocation and power control is a multi-variable fractional programming, which is extremely difficult to be solved.

To address all aforementioned issues, in this paper, we consider a downlink D2D-caching-assisted cellular networks with BS caching and propose two different resource allocation schemes, i.e., hybrid caching transmission scheme and joint caching transmission scheme. In the considered system model, the MEC server is deployed closer to the cellular base station which provides a centralized large-capacity video caching and request users can also get video contents from each other. In the hybrid caching transmission scheme, users could adopt the BS caching mode or alternatively the D2D caching mode. In the joint caching transmission scheme, each user may obtain the requested files from the BS server and adjacent D2D users simultaneously. We formulate two resource allocation problems to maximize the system EE. However, the optimization problems are NP-hard and non-convex. We can obtain the mixed integer programming according to Dinkelbach method. Then, we intend to select the proper link with modified branch and bound method. After that, the optimization problems can be converted to convex problems by using variable relaxation and variable substitution. Through Lagrange algorithm, they can be decomposed into two-step maximization problems. Finally, we obtain the optimal solutions of channel allocation and power control. Simulation results indicate that the proposed resource allocation algorithms are capable of offering a better performance than other algorithms in system throughput and EE.

A. RELATED WORKS

In the future wireless communication system, mobile cellular network architecture is evolving from BS-centric system to content-centric network, and the center of gravity moves from the core network to the edge [14]. With the evolution of base stations and low cost storage units, it is possible to deploy caching on macro base stations and small base stations [15]. Recently, extensive researches have been devoted to the issues related to BS caching. Some papers study the allocation of network resources through game theory [16]–[19]. Pantisano *et al.* [18] propose a novel cache-aware user association and backhaul allocation algorithm which is a one-to-many matching game between small base stations and users. Hoiles *et al.* [19] consider the problem of distributed caching with limited capacity in a content distribution network and formulate the problem as a noncooperative repeated game using the estimated request probabilities. Besides, some papers also use convex optimization to study the resource allocation problem in caching network [20]–[23]. Liang and Yu [22] present a joint bandwidth provisioning and caching strategies, whose goal is to maximize the performance gains under the limitations of backhaul, spectrum, and cache. Tran and Le [23] analyze the joint resource allocation and content caching problem, which tries to minimize the maximum content request rejection rate of users in virtualized wireless networks. Some papers also use reinforcement learning algorithm to study the resource allocation problem in caching network. In [24] and [25], a novel algorithm based on the machine learning is proposed to predict the users content request distribution and mobility patterns using users contexts. However, these articles only consider BS caching and does not take advantage of the caching resource on D2D users. Meanwhile, EE is a better option among multiple optimization goals because it takes account of both the high speed rate of BS caching and the low power consumption of D2D caching.

Wireless caching using D2D communications has attracted substantial research attention to offload traffic from the infrastructure network [26]–[28]. Some papers only consider D2D caching, or use a temporary BS transmission as a supplement to D2D caching. Golrezaei *et al.* [29] analyze the D2D caching in distributed caching network and focus on the problem to maximize frequency reuse with the constraints of collaboration distance and interference. Wang *et al.* [30] focus on the virtual resource allocation to maximize the utility functions which includes the revenue of received data rate, the cost of consumed radio bandwidth and so on. In [29] and [30], one subchannel can be allocated to only one potential link and the mutual interference between users is not taken into account. Wu *et al.* [31] focus on resource allocation to improve the system efficiency in terms of jointly considered spectrum efficiency and energy efficiency. However, when the required content is not available, user will be served by the BS through allocating dedicated spectrum, which greatly reduces the spectrum utilization.

Unlike the previously mentioned works, some papers [32], [33] adopt the integrated architecture of BS caching and D2D caching, where users may obtain the requested videos from distributed cache videos through D2D transmission, and/or from centralized cache videos through communication links to BS server. Zhang *et al.* [32] propose a joint D2D link scheduling and power allocation algorithm to maximize the system throughput. However, it only considers the mutual interference among different D2D links. In our work, we also give consideration to the interference between the BS server and D2D devices. Furthermore, we consider the joint link selection, channel allocation and power management in our work. In [33], Changyan Yi *et al.* formulate a welfare maximization problem integrating benefits from content sharing, total power costs of all UEs and the BS, and penalties for potential service dissatisfactions and propose a joint resource management method. But only one transmission link is permitted in [33]. Our paper is concerned with the joint scheduling of BS caching content and D2D cache content. A hybrid caching transmission scheme and a joint caching transmission scheme are proposed. When users connect to the BS server for requested video, they can decide whether to connect the D2D caching device, which means that users can connect to both MEC server and D2D caching devices at the same time.

B. MAIN CONTRIBUTIONS

In our paper, two resource allocation algorithms with the integrated mechanism of BS caching and D2D caching in D2D-assisted cellular caching networks are proposed. The major contributions of this paper are listed as follows:

1) A collaboration architecture of centralized caching content and distributed caching content is proposed in D2D-assisted cellular caching networks. According to the different cooperation schemes, it can be divided into two schemes: hybrid caching transmission scheme and joint caching transmission scheme.
2) By considering transmission rate constraint, interference level constraint, two EE maximization problems are formulated. Hybrid caching transmission algorithm and joint caching transmission algorithm with heterogeneous video caching are proposed. Dinkelbach method is utilized to transform the fractional optimization problems into mixed-integer nonlinear programming problems. The resource allocation problems are solved by using the modified branch and bound method, variable substitution and Lagrange approach.
3) Numerical results indicate that the proposed resource allocation algorithms significantly improve the performances of system throughput and EE compared with other algorithms. Besides, the selections of the Dinkelbach method, and the modified branch and bound method in our optimization solution are justified. The relationships among transmission power, video transmission ratio factor and transmission rate is also investigated.

The remainder of this paper is organized as follows. In Section II, the system model and problem formulation are described. The specific hybrid caching transmission algorithm is proposed in Section III. The joint caching transmission algorithm is proposed in Section IV. Simulation results are demonstrated in Section V. Section VI concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we will first introduce the system model for hybrid caching transmission scheme and joint caching transmission scheme, and then present the channel model and cache transmission model. Problem formulation is proposed in the last. For convenience, Table 1 lists some important notations used in this paper.

TABLE 1. Important notations in this paper.

Symbol	Description
$p_{i,0}^m$, $p_{i,j}^m$	Transmission power of base station, the helper v_j to u_i on subchannel m
$h_{i,0}^m$, $h_{i,j}^m$	Channel gain of base station, the helper v_j to u_i on subchannel m
$\gamma_{i,0}^m$, $\gamma_{i,j}^m$	SINR of base station, the helper v_j to u_i on subchannel m
$R_{i,0}^m$, $R_{i,j}^m$	Transmission rate of base station, the helper v_j to u_i on subchannel m
$H_{i,0}^m$, $H_{i,j}^m$	SINR of the subchannel m used by the base station, the helper v_j with unit power
$d_{v,j} = 1$	Video caching factor of the helper v_j to u_i
$l_{d,i}$, $l_{b,i}$	D2D caching link selection factor, BS caching link selection factor
$c_{i,0}^m$, $c_{i,j}^m$	Channel m allocation factor of base station, the helper v_j to u_i
C_i^H, C_D^H, C_B^H	Transmission video rate of u_i, at D2D caching link, at BS caching link in hybrid caching transmission scheme
C_i^J, C_D^J, C_B^J	Transmission video rate of u_i, at D2D caching link, at BS caching link in joint caching transmission scheme
C_{total}^H, C_{total}^J	The total throughput of hybrid caching transmission, joint caching transmission
P_{total}^H, P_{total}^J	The total power consumption of hybrid caching transmission, joint caching transmission
EE_{total}^H, EE_{total}^J	The EE of hybrid caching transmission, joint caching transmission

A. SYSTEM MODEL

We consider the downlink video transmission system where D2D caching is implemented as an underlay of cellular network with mobile edge caching. There is a MEC server located in the cellular base station. We assume that the MEC server stores all the videos requested by users to provide the low-latency video service. Meanwhile, user devices cache part of videos, which can be transmitted to the users who request it. Users in the system are divided into two groups, namely, the requesters which requests the cached video, and the helpers providing cached video. MEC server can obtain cache information and request information of all users, so as to control the cache video transmission and resource allocation of the whole system. In this paper, we mainly consider the phases of caching transmission and resource allocation. Suppose users can get the required content through BS

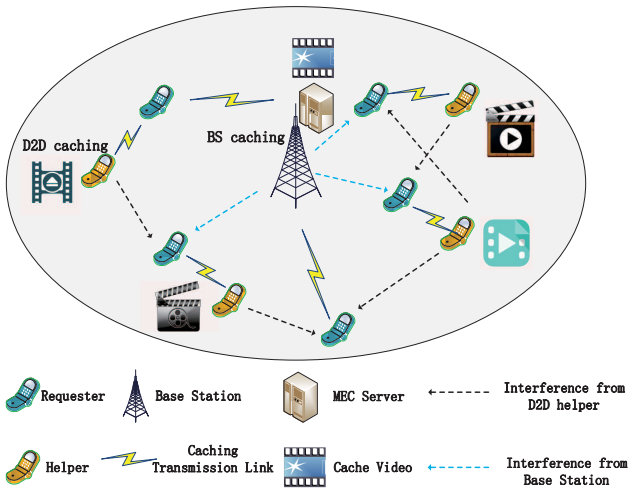


FIGURE 1. Network model.

caching or/and D2D caching, otherwise it will be converted to the traditional video requesting scenario. In the D2D caching transmission, users may be served by multiple helpers, and the helper which can provide the highest transmission rate will be selected as the transmitter of D2D caching.

According to the different video transmission modes, it can be divided into two schemes: hybrid caching transmission scheme and joint caching transmission scheme. In hybrid caching transmission scheme, users can obtain the required videos through the cellular base station or helpers. Choosing the cellular base station or a helper depends on which transmission link the user prefers. In joint caching transmission scheme, the requested videos can be transmitted to users through the transmission links of the cellular base station and helpers.

We define u_i as the requester which can be served by the base station or/and helpers. v_j is defined as the corresponding helper providing cached video. $\mathcal{I}$ is the set of requesters where I is the number of requesters. And $\mathcal{J}$ is the set of helpers where J is the number of helpers. The total number of subchannels is M, which could be shared by all users for channel allocation. Assuming that all the subchannels are block fading, and channel gains are independent identically distributed, that is, there is no correlation at subchannel gains in different time slot [34].

B. CHANNEL MODEL

Let $p_{i,0}^m$ denote the transmission power of the base station on the subchannel m when the requester u_i is connected to the base station. $p_{i,j}^m$ is the transmission power of the helper v_j on the subchannel m when the requester u_i is connected to the helper v_j. $h_{i,j}^m$ is defined as the channel gain on the subchannel m between the requester u_i and the helper v_j, and $h_{i,0}^m$ is defined as the channel gain on the subchannel m between the requester u_i and the base station. When the user u_i selects the base station link to transmit the cache video, the signal to interference plus noise ratio (SINR) $\gamma_{i,0}^m$ of u_i on the subchannel m is as follows:

$$\gamma_{i,0}^m = \frac{p_{i,0}^m h_{i,0}^m}{I_{i,0}^m + N_0 B}, \tag{1}$$

where N_0 is the white Gauss noise, and B is the bandwidth of each subchannel. $I_{i,0}^m = \sum_{j=1}^{J} I_{i,0,j}^m$ indicates the co-channel interference of the requester u_i when helpers occupies the subchannel m to transmit videos. $I_{i,0,j}^m$ is the interference caused by the helper v_j to the requester u_i on the subchannel m. For the sake of simplicity, the SINR of the subchannel m used by the base station with unit power [35] is defined as follows:

$$H_{i,0}^m = \frac{h_{i,0}^m}{I_{i,0}^m + N_0 B}. \tag{2}$$

When the user selects the D2D link to transmit the cached video, the SINR $\gamma_{i,j}^m$ of u_i on the subchannel m is as follows:

$$\gamma_{i,j}^m = \frac{p_{i,j}^m h_{i,j}^m}{I_{i,j}^m + N_0 B}, \tag{3}$$

where $I_{i,j}^m = I_{i,j,0}^m + \sum_{k=1,k\neq j}^{J} I_{i,j,k}^m$ is the interference of the requester u_i which is composed of the co-channel interference from the base station and the interference from other helpers occupying the same subchannel. Then, the SINR of the subchannel m used by the helper v_j with unit power is

$$H_{i,j}^m = \frac{h_{i,j}^m}{I_{i,j}^m + N_0 B}. \tag{4}$$

The channel allocation factor $c_{i,0}^m$ can either be 1 or 0 indicating whether the requestor u_i and the base station are connected through the subchannel m or not. $c_{i,j}^m$ is used to indicate whether the requestor u_i and the helper v_j are connected through the subchannel m.

When the requester u_i and the base station are connected, the transmission rate of the video on the subchannel m is

$$R_{i,0}^m = c_{i,0}^m \log_2\left(1 + \gamma_{i,0}^m\right) = c_{i,0}^m \log_2\left(1 + p_{i,0}^m H_{i,0}^m\right). \tag{5}$$

For simplicity, we use $f(x)$ to denote the logarithmic function $\log_2(1+x)$ with respect to x. Then, the transmission rate of u_i is converted to

$$R_{i,0}^m = c_{i,0}^m f\left(\gamma_{i,0}^m\right) = c_{i,0}^m f\left(p_{i,0}^m H_{i,0}^m\right). \tag{6}$$

When the requester u_i and the helper v_j are connected, the transmission rate of the video on the subchannel m is

$$R_{i,j}^m = c_{i,j}^m f\left(\gamma_{i,j}^m\right) = c_{i,j}^m f\left(p_{i,j}^m H_{i,j}^m\right). \tag{7}$$

C. CACHE TRANSMISSION MODEL

It is assumed that there are V cached videos in the system, and the size of each cached video is s_v, $v = 1, \ldots, V$.

The MEC server at the base station has all the videos, whose popularity probability follows Zipf distribution [36], i.e.,

$$f_\eta = \frac{\frac{1}{\eta^\beta}}{\sum_{v=1}^{V} \frac{1}{v^\beta}}, \quad 1 \le \eta \le V, \tag{8}$$

where η is the file index, and β is the file request coefficient and controls the popularity distribution of files [32]. Each user randomly and independently caches one out of V files in its memory, according to the Zipf-distribution. If these caches contain the video that the requester needs, these users can provide cached video services as helpers. The video list of helpers is $D_{V\times J}$. If the helper v_j cache the video v, then $d_{v,j} = 1$.

The caching request factor is a_i^v, indicating the requestor u_i request for caching video v. For the sake of simplicity, it is assumed that the requester u_i can request only one of the cached videos in each time. Meanwhile, the user device can connect a base station and a helper device at most, for caching content transmission.

In the hybrid caching transmission scheme, $l_{d,i}$ is defined to indicate whether the requester u_i is connected to other D2D users and $l_{b,i}$ is defined to indicate whether the requester u_i is connected to the base station. And there is a constraint that $l_{d,i} + l_{b,i} = 1$, i.e., $l_{b,i} = 1 - l_{d,i}$.

In a certain time slot, the transmission video rate of the requestor u_i is as follows:

$$\begin{aligned} C_i^H &= C_D^H + C_B^H \\ &= \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i} d_{v,j} R_{i,j}^m + \sum_{m=1}^{M} l_{b,i} R_{i,0}^m \\ &= \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i} d_{v,j} R_{i,j}^m + \sum_{m=1}^{M} \left(1 - l_{d,i}\right) R_{i,0}^m, \end{aligned} \tag{9}$$

where C_D^H and C_B^H are the cached video rates when the requestor u_i is connected to other users and the base station in the hybrid caching transmission scheme.

In the joint caching transmission scheme, there is no constraint between $l_{d,i}$ and $l_{b,i}$. The requester can obtain the cached video from helpers and the base station simultaneously, that is, $l_{b,i} = l_{d,i} = 1$. Define β_d and β_b as the video ratio factors that the requester needs to obtain from the helper and the base station. The cached video capacity received by the requester u_i from the helper v_j is $s_v\beta_d$ and the cached video capacity received by the requester u_i from the base station is $s_v\beta_b$, where $\beta_d + \beta_b = 1$.

In a certain time slot, the transmission video rate of the requestor u_i is

$$\begin{aligned} C_i^J &= C_D^J + C_B^J \\ &= \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i} d_{v,j} R_{i,j}^m + \sum_{m=1}^{M} l_{b,i} R_{i,0}^m \\ &= \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j} R_{i,j}^m + \sum_{m=1}^{M} R_{i,0}^m, \end{aligned} \tag{10}$$

where C_D^J and C_B^J are the cached video rates when the requestor u_i is connected to helpers and the base station in the joint caching transmission scheme.

D. PROBLEM FORMULATION

1) HYBRID CACHING TRANSMISSION ALGORITHM

As mentioned before, the MEC server at the base station caches all the videos. At the beginning, the MEC server collects link information and caching information to allocate subchannel and power resource for D2D transmission and base station transmission.

This paper contains two kinds of video transmission algorithms. Under the constraints of limited power and channel resource, it is necessary to effectively accomplish the video transmission and maximize the resource utilization, that is, the energy efficiency.

In the hybrid caching transmission scheme, the throughput of the entire network system is defined as

$$C_{total}^H = \sum_{i\in\mathcal{I}} C_i^H. \tag{11}$$

The total power consumption is expressed as follows:

$$P_{total}^H = \sum_{m=1}^{M}\sum_{i\in\mathcal{I}} \left(l_{d,i} c_{i,j}^m p_{i,j}^m + \left(1 - l_{d,i}\right) c_{i,0}^m p_{i,0}^m\right). \tag{12}$$

The EE of the whole system can be expressed as

$$EE_{total}^H = \frac{C_{total}^H}{P_{total}^H}. \tag{13}$$

For the requester u_i, the transmission throughput should be greater than the transmission speed of the video v.

$$\begin{aligned} &\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i} d_{v,j} c_{i,j}^m f\left(p_{i,j}^m H_{i,j}^m\right) \\ &\qquad + \sum_{m=1}^{M} \left(1 - l_{d,i}\right) c_{i,0}^m f\left(p_{i,0}^m H_{i,0}^m\right) \ge r_i^{s_v}, \end{aligned} \tag{14}$$

where $r_i^{s_v} = \frac{a_i^v s_v}{T_{ave}}$ is the transmission speed of the video v and T_{ave} is the average transmission time.

The interference level constraint of the requester u_i can be written as

$$\sum_{m=1}^{M} \left(l_{d,i} c_{i,j}^m I_{i,j}^m + \left(1 - l_{d,i}\right) c_{i,0}^m I_{i,0}^m\right) \le I_{i,th}, \tag{15}$$

where $I_{i,th}$ is the interference threshold of the requester u_i.

At the same time, the channel allocation factor and power should meet the following constraints.

$$c_{i,j}^m \in \{0, 1\}, c_{i,0}^m \in \{0, 1\}, \tag{16}$$

$$0 \le p_{i,j}^m \le P_{\max}^D, 0 \le p_{i,0}^m \le P_{\max}^B, \tag{17}$$

where $P_{\max}^D$ represents the maximum transmission power of helpers and $P_{\max}^B$ represents the maximum transmission power of the base station.

$$
\begin{aligned}
P1: \max_{\mathcal{L},\mathcal{C},\mathcal{P}} EE^H_{total} &= \frac{C^H_{total}}{P^H_{total}} = \frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c^m_{i,j}f\left(p^m_{i,j}H^m_{i,j}\right)+\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(1-l_{d,i}\right)c^m_{i,0}f\left(p^m_{i,0}H^m_{i,0}\right)}{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(l_{d,i}c^m_{i,j}p^m_{i,j}+\left(1-l_{d,i}\right)c^m_{i,0}p^m_{i,0}\right)} \\
s.t.\ C1: &\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c^m_{i,j}f\left(p^m_{i,j}H^m_{i,j}\right)+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c^m_{i,0}f\left(p^m_{i,0}H^m_{i,0}\right)\geq r_i^{s_v}, \\
C2: &\sum_{m=1}^{M}\left(l_{d,i}c^m_{i,j}I^m_{i,j}+\left(1-l_{d,i}\right)c^m_{i,0}I^m_{i,0}\right)\leq I_{i,th}, \\
C3: &\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}c^m_{i,j}+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c^m_{i,0}=1, \\
C4: &\ l_{d,i}\in\{0,1\}, \\
C5: &\ c^m_{i,j}\in\{0,1\}, c^m_{i,0}\in\{0,1\}, \\
C6: &\ 0\leq p^m_{i,j}\leq P^D_{\max}, 0\leq p^m_{i,0}\leq P^B_{\max},
\end{aligned}
\tag{20}
$$

And the user access factor $l_{d,i}$ has to satisfy the conditions that

$$l_{d,i}\in\{0,1\}, \tag{18}$$

$$\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}c^m_{i,j}+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c^m_{i,0}=1, \tag{19}$$

where (19) denotes that there is only one transmission link which is assigned one subchannel for any requster u_i.

The optimization problem of the hybrid caching transmission algorithm is to maximize the system EE via joint link selection, subchannel allocation and power control, which can be summarized in (20), as shown at the top of this page, where $\mathcal{L}$ is the link selection set, $\mathcal{C}$ is the channel allocation set and $\mathcal{P}$ is the power allocation set.

2) JOINT CACHING TRANSMISSION ALGORITHM

In the joint caching transmission scheme, the cached video requested by users is transmitted through helpers and the base station simultaneously. Besides the constraints of subchannel and power resource, it is also necessary to consider the date rate constraints. The throughput of the entire network system is defined as

$$C^J_{total}=\sum_{i\in\mathcal{I}} C^J_i. \tag{21}$$

The total power consumption is expressed as follows:

$$
\begin{aligned}
P^J_{total} &= \sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(l_{d,i}c^m_{i,j}p^m_{i,j}+l_{b,i}c^m_{i,0}p^m_{i,0}\right) \\
&= \sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(c^m_{i,j}p^m_{i,j}+c^m_{i,0}p^m_{i,0}\right).
\end{aligned}
\tag{22}
$$

The system EE is given by

$$EE^J_{total}=\frac{C^J_{total}}{P^J_{total}}. \tag{23}$$

For the requester u_i, the throughput rate transmitted through the base station and the throughput rate transmitted through the helper v_j should meet the following constraints that

$$\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j}c^m_{i,j}f\left(p^m_{i,j}H^m_{i,j}\right)\geq\beta_d r_i^{s_v}, \tag{24}$$

$$\sum_{m=1}^{M} c^m_{i,0}f\left(p^m_{i,0}H^m_{i,0}\right)\geq\left(1-\beta_d\right)r_i^{s_v}. \tag{25}$$

The optimization problem of the joint caching transmission algorithm can be formulated as

$$
\begin{aligned}
&\max_{\mathcal{C},\mathcal{P}} EE^J_{total}=\frac{C^J_{total}}{P^J_{total}} \\
&=\frac{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} d_{v,j}c^m_{i,j}f\left(p^m_{i,j}H^m_{i,j}\right)+\sum_{m=1}^{M}\sum_{i\in\mathcal{I}} c^m_{i,0}f\left(p^m_{i,0}H^m_{i,0}\right)}{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(c^m_{i,j}p^m_{i,j}+c^m_{i,0}p^m_{i,0}\right)} \\
s.t.\ C7: &\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j}c^m_{i,j}f\left(p^m_{i,j}H^m_{i,j}\right)\geq\beta_d r_i^{s_v}, \\
C8: &\sum_{m=1}^{M} c^m_{i,0}f\left(p^m_{i,0}H^m_{i,0}\right)\geq\left(1-\beta_d\right)r_i^{s_v}, \\
C9: &\sum_{m=1}^{M}\left(c^m_{i,j}I^m_{i,j}+c^m_{i,0}I^m_{i,0}\right)\leq I_{i,th}, \\
C10: &\ c^m_{i,j}\in\{0,1\}, c^m_{i,0}\in\{0,1\}, \\
C11: &\ 0\leq p^m_{i,j}\leq P^D_{\max}, 0\leq p^m_{i,0}\leq P^B_{\max}.
\end{aligned}
\tag{26}
$$

III. HYBRID CACHING TRANSMISSION ALGORITHM

This research investigates the optimization problem of the hybrid caching transmission scheme by studying the system EE maximization problem as P1 shows.

A. REFORMULATION OF PROBLEM P1

This is a mixed integer fractional programming problem [37] since both binary variables and real variables are involved, which is a NP-hard problem. To obtain the joint optimal link scheduling, channel allocation and power control solution, the complexity of the exhaustive search is quite high. Alternatively, we seek to obtain a suboptimal solution of P1 with reasonable complexity. This nonconvex mixed-integer nonlinear programming problem can be decomposed into three subproblems of link selection, channel allocation and power control respectively.

To solve this problem, a nonnegative parameter λ is introduced, and P1 is converted to the following formula.

$$P2: F(\lambda) = \max_{\mathcal{L},\mathcal{C},\mathcal{P}} C_{total}^{H} - \lambda P_{total}^{H}. \tag{27}$$

According to the following ***Theorem 1*** [38], the transformed formula (27) is the equivalent non-fractional form for the fractional programming problem P1.

Theorem 1: Suppose $\mathcal{G}$ is the feasible solution set due to C1-C6, there exists maximum energy efficiency λ^* such that

$$\lambda^* = \frac{C_{total}^{H}(\mathcal{L}^*, \mathcal{C}^*, \mathcal{P}^*)}{P_{total}^{H}(\mathcal{L}^*, \mathcal{C}^*, \mathcal{P}^*)} = \max_{(\mathcal{L},\mathcal{C},\mathcal{P})\in\mathcal{G}} \frac{C_{total}^{H}(\mathcal{L}, \mathcal{C}, \mathcal{P})}{P_{total}^{H}(\mathcal{L}, \mathcal{C}, \mathcal{P})} \tag{28}$$

if and only if

$$\begin{aligned} &\max_{(\mathcal{L},\mathcal{C},\mathcal{P})\in\mathcal{G}} \left\{ C_{total}^{H}(\mathcal{L}, \mathcal{C}, \mathcal{P}) - \lambda^* P_{total}^{H}(\mathcal{L}, \mathcal{C}, \mathcal{P}) \right\} \\ &= C_{total}^{H}\left(\mathcal{L}^*, \mathcal{C}^*, \mathcal{P}^*\right) - \lambda^* P_{total}^{H}\left(\mathcal{L}^*, \mathcal{C}^*, \mathcal{P}^*\right) = 0, \end{aligned} \tag{29}$$

where $(\mathcal{L}^*, \mathcal{C}^*, \mathcal{P}^*)$ reflects the optimal link selection, channel allocation and power control solution. The proof of the theorem can be obtained similar to the proof given in [39].

According to Dinkelbach method in Algorithm 1, we can obtain the value of $F(\lambda)$ with an initial λ by solving the problem P2 in a finite number of iterations.

B. OPTIMIZATION OF OUTER LAYER

The problem P2 is a mixed integer programming problem which is difficult to solve due to the integer constraints (4) and (5). The general method to solve integer programming problems is time-sharing method. By relaxing integer variables into continuous variables, effective linear/nonlinear optimization methods can be adopted. Because all feasible solutions of the original problem fall into the solution space of the relaxation problem, the optimal solution of the relaxation problem is always the upper bound of the original problem P2 [35].

It can be seen that the optimization problem P2 consists of two layers. The outer layer is the selection process of the video transmission mode, which involves the 0-1 integer optimization problem. The inner layer can be further divided into two independent sub-problems. The first sub-problem is channel allocation and the other is power control.

Algorithm 1 Dinkelbach Method

1: *Step 1* **Initialization:**
2: Set maximum iteration index of Dinkelbach algorithm I_{Dink} and Dinkelbach algorithm precision ε_{Dink};
3: $l_{d,i}$, $c_{i,j}^{m}$ and $c_{i,0}^{m}$: The initial link selection index and the initial channel allocation indexs.
4: $p_{i,j}^{m}$ and $p_{i,0}^{m}$: The initial power of the helper v_j and the base station on the subchannel m.
5: $k = 1$ and $\lambda(0) = 1$.
6: *Step 2* **Iteration:**
7: **while** $k < I_{Dink}$ and $\lambda(k) - \lambda(k-1) > \varepsilon_{Dink}$
8: Solve problem (29) for optimal link selection, channel allocation and power allocation.
9: $F(k+1) = C_{total}^{H} - \lambda(k) P_{total}^{H}$
10: $\lambda(k+1) = \frac{C_{total}^{H}}{P_{total}^{H}}$.
11: $k = k + 1$.

In the process of solving the outer layer, the branch and bound method is used to solve the 0-1 integer optimization problem. First, the binary variables $l_{d,i}$ is relaxed to real continuous variables [0, 1], and the optimization problem is converted to P3. The relaxed variables can be considered as the time domain sharing factors. If $l_{d,i}$ represents the proportion of time slot τ occupied by the helper link, $l_{d,i} = 0.1$ represents that the requester adopts the D2D caching mode within 0.1τ, and adopts the BS caching mode from 0.1τ to τ. And the converted problem P3 can be expressed as follows:

$$\begin{aligned} P3: F(\mathcal{L}, \mathcal{C}, \mathcal{P}, \lambda) = &\max_{\mathcal{L}} C_{total}^{H} - \lambda P_{total}^{H} \\ &s.t.\ C1, C2, C3, C5, C6, \\ &C4': l_{d,i} \in [0, 1], \end{aligned} \tag{30}$$

Consider $l_{D,i}^{*}$ and $F^*(\mathcal{L}, \mathcal{C}, \mathcal{P}, \lambda)$ as the optimal solution to P3 and the optimal value of the objective function of P3, respectively. If each element of $l_{D,i}^{*}$ is an integer, then the problem P3 obtains the optimal solution. Otherwise, the branching strategy is applied to P3, and each optimization problem is branched into two sub-problems. Similar to [40], a non-integer $l_{D,i'}$ is chosen as the branching variable that

$$i' = \arg\max \left\{ d_{v,j} c_{i,j}^{m} f\left(p_{i,j}^{m} H_{i,j}^{m}\right), c_{i,0}^{m} f\left(p_{i,0}^{m} H_{i,0}^{m}\right) \right\}. \tag{31}$$

The maximum partial derivative of the objective function of P3 with respect to $l_{d,i}$ is used as the branching variable, which leads to faster convergence and less operation. Along with $l_{D,i'}$, P3 is decomposed into two sub-problems,

P3-1 and P3-2.

$$
\begin{aligned}
&P3-1: F(\mathcal{L},\mathcal{C},\mathcal{P},\lambda)\\
&\quad=\max_{\mathcal{L}}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^{m}f\left(p_{i,j}^{m}H_{i,j}^{m}\right)\\
&\qquad+\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(1-l_{d,i}\right)c_{i,0}^{m}f\left(p_{i,0}^{m}H_{i,0}^{m}\right)\\
&\qquad-\lambda\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(l_{d,i}c_{i,j}^{m}p_{i,j}^{m}+\left(1-l_{d,i}\right)c_{i,0}^{m}p_{i,0}^{m}\right)\\
&\qquad s.t.\ C1, C2, C3, C5, C6,\\
&\qquad l_{d,i}\in[0,1],\forall i\in I\backslash\left\{i'\right\},\\
&\qquad l_{d,i'}=0,
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
&P3-2: F(\mathcal{L},\mathcal{C},\mathcal{P},\lambda)\\
&\quad=\max_{\mathcal{L}}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^{m}f\left(p_{i,j}^{m}H_{i,j}^{m}\right)\\
&\qquad+\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(1-l_{d,i}\right)c_{i,0}^{m}f\left(p_{i,0}^{m}H_{i,0}^{m}\right)\\
&\qquad-\lambda\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(l_{d,i}c_{i,j}^{m}p_{i,j}^{m}+\left(1-l_{d,i}\right)c_{i,0}^{m}p_{i,0}^{m}\right)\\
&\qquad s.t.\ C1, C2, C3, C5, C6,\\
&\qquad l_{d,i}\in[0,1],\forall i\in I\backslash\left\{i'\right\},\\
&\qquad l_{d,i'}=1.
\end{aligned}
\tag{33}
$$

The branch and bound process will be repeated until the relaxed sub-problem satisfies all integer constraints and obtains the maximum value of the objective function. Then, the optimization problem P3 can be converted to

$$
\begin{aligned}
&P4: F(\mathcal{C},\mathcal{P},\lambda)\\
&\quad=\max_{\mathcal{C},\mathcal{P}}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^{m}f\left(p_{i,j}^{m}H_{i,j}^{m}\right)\\
&\qquad+\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(1-l_{d,i}\right)c_{i,0}^{m}f\left(p_{i,0}^{m}H_{i,0}^{m}\right)\\
&\qquad-\lambda\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(l_{d,i}c_{i,j}^{m}p_{i,j}^{m}+\left(1-l_{d,i}\right)c_{i,0}^{m}p_{i,0}^{m}\right)\\
&\qquad s.t.\ C1, C2, C3, C5', C6.
\end{aligned}
\tag{34}
$$

C. OPTIMIZATION OF INNER LAYER

In the process of solving the inner layer, the variables $c_{i,0}^{m}$ and $c_{i,j}^{m}$ are relaxed, that is, the constraint condition C5 is converted to the constraint condition $C5'$. We introduce new variables $s_{i,0}^{m}$ and $s_{i,j}^{m}$, where $s_{i,0}^{m}=c_{i,0}^{m}p_{i,0}^{m}$ and $s_{i,j}^{m}=c_{i,j}^{m}p_{i,j}^{m}$. The optimization problem P4 can be converted to P5 which is given by

$$
\begin{aligned}
&P5: F(\mathcal{C},\mathcal{S},\lambda)\\
&\quad=\max_{\mathcal{C},\mathcal{S}}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^{m}f\left(\frac{s_{i,j}^{m}}{c_{i,j}^{m}}H_{i,j}^{m}\right)\\
&\qquad+\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(1-l_{d,i}\right)c_{i,0}^{m}f\left(\frac{s_{i,0}^{m}}{c_{i,0}^{m}}H_{i,0}^{m}\right)\\
&\qquad-\lambda\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(l_{d,i}s_{i,j}^{m}+\left(1-l_{d,i}\right)s_{i,0}^{m}\right)\\
&\qquad s.t.\ C1': \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^{m}f\left(\frac{s_{i,j}^{m}}{c_{i,j}^{m}}H_{i,j}^{m}\right)\\
&\qquad+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c_{i,0}^{m}f\left(\frac{s_{i,0}^{m}}{c_{i,0}^{m}}H_{i,0}^{m}\right)\geq r_{i}^{s_v},\\
&\qquad C2: \sum_{m=1}^{M}\left(l_{d,i}c_{i,j}^{m}I_{i,j}^{m}+\left(1-l_{d,i}\right)c_{i,0}^{m}I_{i,0}^{m}\right)\leq I_{i,th},\\
&\qquad C3: \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}c_{i,j}^{m}+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c_{i,0}^{m}=1,\\
&\qquad C5': c_{i,j}^{m}\in[0,1], c_{i,0}^{m}\in[0,1],\\
&\qquad C6': 0\leq s_{i,j}^{m}\leq P_{\max}^{D}, 0\leq s_{i,0}^{m}\leq P_{\max}^{B},
\end{aligned}
\tag{35}
$$

where $\mathcal{S}$ is the set of new variable s.

Proposition: The problems P4 and P5 are equivalent problems and the problem P5 is a convex problem.

Proof: With the exception of some special points, like $c_{i,0}^{m}=0$ and $c_{i,j}^{m}=0$, the equivalent problem P5 can be converted back the problem P4. So this correspondence is not a one-to-one mapping relationship. However, when $c_{i,0}^{m}=0$ and $c_{i,j}^{m}=0$, the requestor does not receive cached video through the subchannel m, so the MEC server does not need to allocate power on this channel. Because of the nature of the optimization problem, these points where $c_{i,0}^{m}=0$ and $c_{i,j}^{m}=0$ do not affect the optimization problem. When the mapping between $\left\{c_{i,0}^{m},p_{i,0}^{m},c_{i,j}^{m},p_{i,j}^{m}\right\}$ and $\left\{c_{i,0}^{m},s_{i,0}^{m},c_{i,j}^{m},s_{i,j}^{m}\right\}$ is defined as the following expression, the problems P4 and P5 become equivalent problems.

$$
p_{i,0}^{m}=\begin{cases}\dfrac{s_{i,0}^{m}}{c_{i,0}^{m}} & \text{if } c_{i,0}^{m}>0,\\ 0 & \text{if otherwise,}\end{cases}\quad
p_{i,j}^{m}=\begin{cases}\dfrac{s_{i,j}^{m}}{c_{i,j}^{m}} & \text{if } c_{i,j}^{m}>0,\\ 0 & \text{if otherwise.}\end{cases}
\tag{36}
$$

According to convex optimization theory [41], a convex optimization problem is to maximize a convex objective function or to minimize a concave objective function on a convex set, so we only need to prove that the constraint condition and objective function of P5 satisfy these conditions. Because the constraints $C2$, $C3$, $C5'$, and $C6'$ are linear, their feasible sets are convex sets. It is only necessary to prove that

the objective function and the constraint condition $C1'$ are convex functions.

First, we should prove the continuity of the function $g(t,x) = x\log\left(\frac{t}{x}\right), t \geq 0, x \geq 0$ at the point where $x = 0$. Define $y = \frac{t}{x}$, and we can get $g(t,0) = \lim_{x\to 0} x\log\left(\frac{t}{x}\right) = \lim_{y\to\infty} \frac{t}{y}\log y = t\lim_{y\to\infty}\frac{\log y}{y} = 0$. The function $g(t,x) = x\log\left(\frac{t}{x}\right), t \geq 0, x \geq 0$ is the perspective function of the logarithmic function [42]. Since the function $c\log\frac{s}{c}$ is the perspective function of the function $\log s$, and the function $\log s$ is a convex function of the variable s, so the function $c\log\frac{s}{c}$ is also a convex function with respect to c and s. It can be concluded that the objective function of P5 is the weighted sum of a series of convex functions and linear functions, and it is a convex function. Meanwhile, the constraint condition $C1'$ is also a convex function with respect to c and s. Finally, the feasible solution set of the problem P5 has been proved to be a convex set and the objective function is a convex function, so the problem P5 is a convex problem. □

With the given values of $c_{i,0}^m$ and $c_{i,j}^m$, $s_{i,0}^m$ and $s_{i,j}^m$ can be obtained. Based on $s_{i,0}^m$ and $s_{i,j}^m$, optimal subchannel allocation indexes are obtained. Through multiple iterations, the values of $s_{i,0}^m$, $s_{i,j}^m$, $c_{i,0}^m$ and $c_{i,j}^m$ will converge, and the optimization problem P5 gets its optimal solution.

The convex problem P5 satisfies Slaters condition. Therefore, the solution to the original problem can be obtained by solving the dual problem [38]. The optimization problem P5 can be solved by the Lagrange algorithm. With the Lagrange multipliers of μ_1, ν_1, ς_1, φ_1^m and ϕ_1^m, P5 can be transformed into

$$P6: \min_{\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m} \max_{\mathcal{C},\mathcal{S}} L\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m,\mathcal{C},\mathcal{S}\right). \quad (37)$$

The problem P6 can be solved by the iteration of a master problem and multiple subproblems. The subproblem is as follows:

$$D\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m\right) = \max_{\mathcal{C},\mathcal{S}} L\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m,\mathcal{C},\mathcal{S}\right). \quad (38)$$

And the expression of $L\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m,\mathcal{C},\mathcal{S}\right)$ is given by

$$\begin{aligned}
&L\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m,\mathcal{C},S\right)\\
&=\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^m f\left(\frac{s_{i,j}^m}{c_{i,j}^m}H_{i,j}^m\right)\\
&\quad+\sum_{m=1}^{M}\left(\left(1-l_{d,i}\right)c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right)\right)\\
&\quad-\lambda\sum_{m=1}^{M}\left(l_{d,i}s_{i,j}^m+\left(1-l_{d,i}\right)s_{i,0}^m\right)\\
&\quad+\mu_1\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^m f\left(\frac{s_{i,j}^m}{c_{i,j}^m}H_{i,j}^m\right)\\
&\quad+\mu_1\sum_{m=1}^{M}\left(\left(1-l_{d,i}\right)c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right)\right)\\
&\quad+\nu_1\left(I_{i,th}-\sum_{m=1}^{M}\left(l_{d,i}c_{i,j}^m I_{i,j}^m+\left(1-l_{d,i}\right)c_{i,0}^m I_{i,0}^m\right)\right)\\
&\quad+\varsigma_1\left(\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}c_{i,j}^m+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c_{i,0}^m-1\right)\\
&\quad+\sum_{m=1}^{M}\varphi_1^m\left(P_{\max}^D-s_{i,j}\right)+\sum_{m=1}^{M}\phi_1^m\left(P_{\max}^B-s_{i,0}\right)-\mu_1 r_i^{s_v}.
\end{aligned} \quad (39)$$

It can be decomposed into a two-step maximization problem. The optimal solutions of $s_{i,0}^m$ and $s_{i,j}^m$ can be obtained through the Karush-Kuhn-Tucker (KKT) condition, and the expressions are as follows:

$$s_{i,j}^{m*} = \frac{(1+\mu_1)\left(l_{d,i}d_{v,j}c_{i,j}^m\right)}{\left(\lambda l_{d,i}+\varphi_1^m\right)\ln 2} - \frac{c_{i,j}^m}{H_{i,j}^m}, \quad (40)$$

$$s_{i,0}^{m*} = \frac{(1+\mu_1)\left(\left(1-l_{d,i}\right)c_{i,0}^m\right)}{\left(\lambda\left(1-l_{d,i}\right)+\phi_1^m\right)\ln 2} - \frac{c_{i,0}^m}{H_{i,0}^m}. \quad (41)$$

Substituting these optimal values of $s_{i,0}^m$ and $s_{i,j}^m$ into $D\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m\right)$, we can obtain

$$k^* = \arg\max_k\left(L\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m,c_{i,j}^k,s_{i,j}^*,s_{i,0}^*\right)\right), \quad (42)$$

$$d^* = \arg\max_d\left(L\left(\mu_1,\nu_1,\varsigma_1,\varphi_1^m,\phi_1^m,c_{i,0}^d,s_{i,j}^*,s_{i,0}^*\right)\right). \quad (43)$$

And we can also get these optimal channel allocation values as follows:

$$c_{i,j}^k = \begin{cases}1, & k=k^*,\\ 0, & otherwise,\end{cases} \quad (44)$$

$$c_{i,0}^d = \begin{cases}1, & d=d^*,\\ 0, & otherwise.\end{cases} \quad (45)$$

The master problem is to solve the Lagrange multipliers. A gradient descent method is adopted to update the Lagrange multipliers, and the expressions of μ_1 and ν_1 are shown at the bottom of the next page.

$$\begin{aligned}
\varsigma_1(i+1) = \varsigma_1(i) - \delta_{\varsigma_1}&\left(\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}c_{i,j}^m\right.\\
&\left.+\sum_{m=1}^{M}\left(1-l_{d,i}\right)c_{i,0}^m-1\right).
\end{aligned} \quad (48)$$

$$\varphi_1^m(i+1) = \varphi_1^m(i) - \delta_{\varphi_1^m}\left(P_{\max}^D - s_{i,j}\right). \quad (49)$$

$$\phi_1^m(i+1) = \phi_1^m(i) - \delta_{\phi_1^m}\left(P_{\max}^B - s_{i,0}\right). \quad (50)$$

where i is the iteration index, δ_{μ_1}, δ_{ν_1}, δ_{ς_1}, $\delta_{\varphi_1^m}$ and $\delta_{\phi_1^m}$ are positive step sizes.

And the procedure of the proposed hybrid caching transmission algorithm is illustrated in Algorithm 2.

Algorithm 2 Hybrid Caching Transmission Algorithm

Step 1 **Initialization:**
Set maximum iteration index of Dinkelbach algorithm I_{Dink}, Dinkelbach algorithm precision ε_{Dink} and maximum iteration index of Lagrange duality method I_{La};
$l_{D,i}$, $c_{i,j}^m$ and $c_{i,0}^m$: The initial link selection index and the initial channel allocation indexes.
$p_{i,j}^m$ and $p_{i,0}^m$: The initial power of the helper v_j and the base station on the subchannel m.
$k = 1, \lambda(0) = 1, t = 1$.
Step 2 **Iteration:**
While $k < I_{Dink}$ and $\lambda(k) - \lambda(k-1) > \varepsilon_{Dink}$
  Initialize the problem list with the root problem P3 and set its lower bound as $LB(F(L)) = 0$;
  While the problem list is not empty **Do** Select the problem from the problem list that has the largest upper bound by applying bounding strategy. Obtain its optimal mode selection solution $l_{d,i'}$ and upper bound $UB\left(F\left(l_{d,i'}\right)\right)$;
  If $l_{d,i'}$ is infeasible or $UB\left(F\left(l_{d,i'}\right)\right) < LB(F(L))$, discard the problem. **Else if** all elements in $l_{D,i'}$ are integers and $UB\left(F\left(l_{d,i'}\right)\right) > LB(F(L))$, set $l_{d,i}^* = l_{d,i'}$ and $LB(F(L)) = UB\left(F\left(l_{d,i'}\right)\right)$ then DISCARD the problem. **Otherwise**, branch the problem into two new subproblems along the determinate split index and add these new sub-problems to the problem list;
  End while
  Repeat
    Compute $p_{i,j}^{m*}$ and $p_{i,0}^{m*}$ according to (36), (40) and (41).
    Set $p_{i,j}^m = p_{i,j}^{m*}$ and $p_{i,0}^m = p_{i,0}^{m*}$, then obtain the optimal channel allocation indexes $c_{i,j}^m$ and $c_{i,0}^m$ according to (42), (43), (44) and (45).
    Update μ_1, ν_1, ς_1, φ_1^m and ϕ_1^m and $t = t + 1$.
  until Lagrangian multipliers convergence
  $F(k+1) = C_{total}^H - \lambda(k) P_{total}^H$
  $\lambda(k+1) = \frac{C_{total}^H}{P_{total}^H}$.
  $k = k + 1$.
End while
Step 3 **Finalization:**
The final $l_{D,i}^*$, $c_{i,j}^{m*}$, $c_{i,0}^{m*}$, $p_{i,j}^{m*}$ and $p_{i,0}^{m*}$ is obtained.

The base station with a MEC server stores the popular videos to improve transmission efficiency. The MEC server could obtain the caching information and request information of users. Based on the above information, the server completes the specific resource allocation to achieve cached videos transmission.

By using the modified branch and bound method, the result of link selection can be determined. Then, the Lagrange dual algorithm is used to obtain the optimal solutions of subchannel and power allocation. After a certain number of iterations, the hybrid caching transmission algorithm obtains the final results of link selection, channel allocation and power control.

The proposed hybrid caching transmission algorithm based on the Dinkelbach Method, the modified branch and bound method and the Lagrange dual decomposition approach solve three subproblems, i.e., problem P1, P3 and P5. In the worst case, the computational complexity of the modified branch and bound method is $O(2MIJ)$ [40]. The computational complexity of the Algorithm 2 is $O(2I_{Dink}I_{La}M^2I^2J^2)$, where I_{Dink} and I_{La} are the numbers of iterations required for convergence of Dinkelbach method and dual decomposition technique, respectively.

IV. JOINT CACHING TRANSMISSION ALGORITHM

Consider the joint transmission of BS caching and D2D caching. The optimization problem of the joint caching transmission algorithm is formulated as equation (25).

Similar to the hybrid video transmission algorithm, the optimization problem can be converted to P7 by using Dinkelbach algorithm, variable relaxation and variable substitution.

$$
\begin{aligned}
P7: F(\mathcal{C},\mathcal{S},\lambda) &= \max_{\mathcal{C},\mathcal{S}} \sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} d_{v,j}c_{i,j}^m f\left(\frac{s_{i,j}^m}{c_{i,j}^m}H_{i,j}^m\right) \\
&+ \sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right)\right) - \lambda\sum_{m=1}^{M}\sum_{i\in\mathcal{I}}\left(s_{i,j}^m + s_{i,0}^m\right) \\
C7': &\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j}c_{i,j}^m f\left(\frac{s_{i,j}^m}{c_{i,j}^m}H_{i,j}^m\right) \geq \beta_d r_i^{s_v}, \\
C8': &\sum_{m=1}^{M} c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right) \geq (1-\beta_d) r_i^{s_v}, \\
C9: &\sum_{m=1}^{M}\left(c_{i,j}^m I_{i,j}^m + c_{i,0}^m I_{i,0}^m\right) \leq I_{i,th}, \\
C10': &\ c_{i,j}^m \in [0,1], c_{i,0}^m \in [0,1], \\
C11': &\ 0 \leq s_{i,j}^m \leq P_{\max}^D, 0 \leq s_{i,0}^m \leq P_{\max}^B.
\end{aligned} \tag{51}
$$

$$
\mu_1(i+1) = \mu_1(i) - \delta_{\mu_1}\left(\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} l_{d,i}d_{v,j}c_{i,j}^m f\left(p_{i,j}^m H_{i,j}^m\right) + \sum_{m=1}^{M}\left(\left(1-l_{d,i}\right)c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right)\right) - r_i^{s_v}\right). \tag{46}
$$

$$
\nu_1(i+1) = \nu_1(i) - \delta_{\nu_1}\left(I_{i,th} - \sum_{m=1}^{M}\left(l_{d,i}c_{i,j}^m I_{i,j}^m + \left(1-l_{d,i}\right)c_{i,0}^m I_{i,0}^m\right)\right). \tag{47}
$$

The constraints $C9$, $C10'$ and $C11'$ are linear, so their feasible domain is a convex set. And the constraints $C7'$ and $C8'$ are convex functions. The objective function of P7 is a weighted sum of a series of convex functions and linear functions, so it is a convex function. It has been proved that the feasible domain of the problem P7 is a convex set and the objective function is a convex function, so the problem P7 is a convex problem.

By using the Lagrange algorithm, P7 can be transformed into

$$P8: \min_{\mu_2,\nu_2,\varsigma_2,\varphi_2^m,\phi_2^m} \max_{\mathcal{C},\mathcal{S}} L\left(\mu_2, \nu_2, \varsigma_2, \varphi_2^m, \phi_2^m, \mathcal{C}, \mathcal{S}\right). \quad (52)$$

Similarly, the problem P8 can be solved by the iteration of a master problem and multiple subproblems. The subproblem is as follows:

$$\begin{aligned} &D\left(\mu_2, \nu_2, \varsigma_2, \varphi_2^m, \phi_2^m, \mathcal{C}, \mathcal{S}\right) \\ &= \max_{\mathcal{C},\mathcal{S}} L\left(\mu_2, \nu_2, \varsigma_2, \varphi_2^m, \phi_2^m, \mathcal{C}, \mathcal{S}\right) \\ &= \max_{\mathcal{C},\mathcal{S}} \sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j}c_{i,j}^m f\left(\frac{s_{i,j}^m}{c_{i,j}^m}H_{i,j}^m\right) \\ &\quad + \sum_{m=1}^{M}\left(c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right)\right) - \lambda\sum_{m=1}^{M}\left(s_{i,j}^m + s_{i,0}^m\right) \\ &\quad + \mu_2\left(\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j}c_{i,j}^m f\left(\frac{s_{i,j}^m}{c_{i,j}^m}H_{i,j}^m\right) - \beta_d r_i^{s_v}\right) \\ &\quad + \nu_2\left(\sum_{m=1}^{M}\left(c_{i,0}^m f\left(\frac{s_{i,0}^m}{c_{i,0}^m}H_{i,0}^m\right)\right) - (1-\beta_d)\, r_i^{s_v}\right) \\ &\quad + \varsigma_2\left(I_{i,th} - \sum_{m=1}^{M}\left(c_{i,j}^m I_{i,j}^m + c_{i,0}^m I_{i,0}^m\right)\right) \\ &\quad + \sum_{m=1}^{M}\varphi_2^m\left(P_{\max}^D - s_{i,j}^m\right) + \sum_{m=1}^{M}\phi_2^m\left(P_{\max}^B - s_{i,0}^m\right). \end{aligned} \quad (53)$$

Through the KKT condition, we can obtain

$$s_{i,j}^{m*} = \frac{(1+\mu_2)\left(d_{v,j}c_{i,j}^m\right)}{\left(\lambda+\varphi_2^m\right)\ln 2} - \frac{c_{i,j}^m}{H_{i,j}^m}, \quad (54)$$

$$s_{i,0}^{m*} = \frac{(1+\nu_2)\left(c_{i,0}^m\right)}{\left(\lambda+\phi_2^m\right)\ln 2} - \frac{c_{i,0}^m}{H_{i,0}^m}. \quad (55)$$

By substituting (54) and (55) into (53), we can get the solution of channel allocation. The optimal channel allocation index can be determined as follows:

$$k^* = \arg\max_k \left(L\left(\mu_2, \nu_2, \varsigma_2, \varphi_2^m, \phi_2^m, c_{i,j}^k, s_{i,j}^{m*}, s_{i,0}^{m*}\right)\right), \quad (56)$$

$$d^* = \arg\max_d \left(L\left(\mu_2, \nu_2, \varsigma_2, \varphi_2^m, \phi_2^m, c_{i,0}^d, s_{i,j}^{m*}, s_{i,0}^{m*}\right)\right). \quad (57)$$

And the optimal channel allocation values are given by

$$c_{i,j}^k = \begin{cases} 1, & k = k^*, \\ 0, & otherwise, \end{cases} \quad (58)$$

$$c_{i,0}^d = \begin{cases} 1, & d = d^*, \\ 0, & otherwise. \end{cases} \quad (59)$$

The expressions of Lagrange multipliers are given as follows:

$$\mu_2(i+1) = \mu_2(i) - \delta_{\mu_2} \times \left(\sum_{m=1}^{M}\sum_{j\in\mathcal{J}} d_{v,j}c_{i,j}^m f\left(p_{i,j}^m H_{i,j}^m\right) - \beta_d r_i^{s_v}\right), \quad (60)$$

$$\nu_2(i+1) = \nu_2(i) - \delta_{\nu_2}\left(\sum_{m=1}^{M} c_{i,0}^m f\left(p_{i,0}^m H_{i,0}^m\right) - (1-\beta_d)\, r_i^{s_v}\right), \quad (61)$$

$$\varsigma_2(i+1) = \varsigma_2(i) - \delta_{\varsigma_2}\left(I_{i,th} - \sum_{m=1}^{M}\left(c_{i,j}^m I_{i,j}^m + c_{i,0}^m I_{i,0}^m\right)\right), \quad (62)$$

$$\varphi_2^m(i+1) = \varphi_2^m(i) - \delta_{\varphi_2^m}\left(P_{\max}^D - s_{i,j}^m\right), \quad (63)$$

$$\phi_2^m(i+1) = \phi_2^m(i) - \delta_{\phi_2^m}\left(P_{\max}^B - s_{i,0}^m\right), \quad (64)$$

where i is the iteration index, δ_{μ_2}, δ_{ν_2}, δ_{ς_2}, $\delta_{\varphi_2^m}$ and $\delta_{\phi_2^m}$ are positive step sizes.

Similar to the hybrid caching transmission algorithm, the procedure of the proposed joint caching transmission algorithm is illustrated in Algorithm 3. The proposed joint caching transmission algorithm based on the Dinkelbach Method, and the Lagrange dual decomposition approach solve subproblems of P2 and P7. The computational complexity of the Algorithm 3 is $O(I_{Dink}I_{La}MIJ)$.

V. PERFORMANCE EVALUATION

A. PARAMETERS

In order to illustrate the performances of the proposed hybrid caching transmission algorithm (HCTA) and joint caching transmission algorithm (JCTA) in solving resource allocation problem based on heterogeneous video caching in D2D-assisted wireless caching networks, numerical simulations are conducted. There is a cellular base station deployed at the center of simulation region, where users are randomly distributed around the base station. Other simulation parameters are shown in Table 2.

B. SIMULATION RESULTS

The proposed algorithms are compared with the base station cache transmission algorithm (BCTA), D2D cache transmission algorithm (DCTA), the joint D2D link scheduling and power allocation algorithm (JDLS) in [32] and the basis transformation method (TBTM) in [33]. In BCTA and DCTA, the resource allocation problems with the objectives of maximizing the system EE are based on BS caching and D2D caching respectively. Note that the system model and the

Algorithm 3 The Joint Caching Transmission Algorithm

Step 1 **Initialization:**
Set maximum iteration index of Dinkelbach algorithm I_{Dink}, Dinkelbach algorithm precision ε_{Dink} and maximum iteration index of Lagrange duality method I_{La};
$c_{i,j}^m$ and $c_{i,0}^m$: The initial channel allocation indexes.
$p_{i,j}^m$ and $p_{i,0}^m$: The initial power of the helper v_j and the base station on the subchannel m.
$k = 1, \lambda(0) = 1, t = 1$.
Step 2 **Iteration:**
while $k < I_{Dink}$ and $\lambda(k) - \lambda(k-1) > \varepsilon_{Dink}$
  $W = 1$;
  while $W == 1$ and $t < I_{La}$ **do**;
    Compute $p_{i,j}^{m*}$ and $p_{i,0}^{m*}$ according to (36), (54) and (55).
    Set $p_{i,j}^m = p_{i,j}^{m*}$ and $p_{i,0}^m = p_{i,0}^{m*}$, then obtain the optimal channel allocation indexes $c_{i,j}^m$ and $c_{i,0}^m$ according to (56), (57), (58) and (58).
    Update μ_2, ν_2, ς_2, φ_2^m and ϕ_2^m.
    if Lagrangian multipliers convergence **then**
      $W = 0$;
    else $t = t + 1$
    end if
  end while
  $F(k+1) = C_{total}^J - \lambda(k) P_{total}^J$
  $\lambda(k+1) = \frac{C_{total}^J}{P_{total}^J}$.
  $k = k + 1$.
end while
Step 3 **Finalization:**
The final $c_{i,j}^{m*}$, $c_{i,0}^{m*}$, $p_{i,j}^{m*}$ and $p_{i,0}^{m*}$ is obtained.

TABLE 2. Simulation parameters

Parameters	Value
Number of base station	1
Number of helpers	$[100:300]$
Number of requesters	$[100:300]$
Number of subchannels	$M = 50$
Subchannel bandwidth	$B = 0.05(MHz)$
Noise power density	$N_0 = 10^{-12}(W/Hz)$
Maximum transmit power of BS	$P_{\max}^B = 30dBm$
Maximum transmit power of helper	$P_{\max}^D = 10dBm$
Video size	randomly from 100 to 200 Mb
Total amount of videos	200
Video ratio factor	$\beta_d = \{0.005, 0.01, 0.3, 0.5\}$
Zipf distribution exponent	0.8

considered conditions in [32] and [33] are different from those in our paper, so the compared algorithms have been modified to adapt to our system settings.

In this subsection, we present numerical results for these algorithms in system throughput and EE. The proposed HCTA algorithm is compared with the random-link-allocation hybrid caching transmission algorithm (no-BnB HCTA) and the hybrid transmission algorithm without Dinkelbach method (no-Dinkelbach HCTA) to justify the selections of the Dinkelbach method, and the modified branch and bound method. The relationship between video transmission ratio factor and transmission rate also is investigated. Note that the ratio factor of cached video β_d is 0.3 when we compare the performance of these six algorithms.

In Fig. 2, it is obvious that the system throughput of the JCTA algorithm is larger than that of other algorithms. The requesters can obtain cached video simultaneously from helpers and the base station, which can greatly improve the system throughput. The throughputs of the HCTA and the TBTM are higher than the maximum values of the BCTA algorithm and the DCTA algorithm. Due to the hybrid link transmission, the co-channel interference is smaller than those of BCTA algorithm and DCTA algorithm, thus the HCTA and the TBTM can achieve greater performance gain. Influenced by pricing factors, the TBTM algorithm solves a welfare maximization problem. Under the simulation settings, more users in TBTM choose to transmit through D2D caching, so the throughput is smaller than HCTA algorithm. Because the JDLS algorithm does not consider the interference between BS server and D2D devices, which leads to the serious co-channel interference, its throughput is smaller. And the growth rate of system throughput of DCTA algorithm remains basically unchanged, because requesters in this algorithm obtain the cached video from helpers surround it.

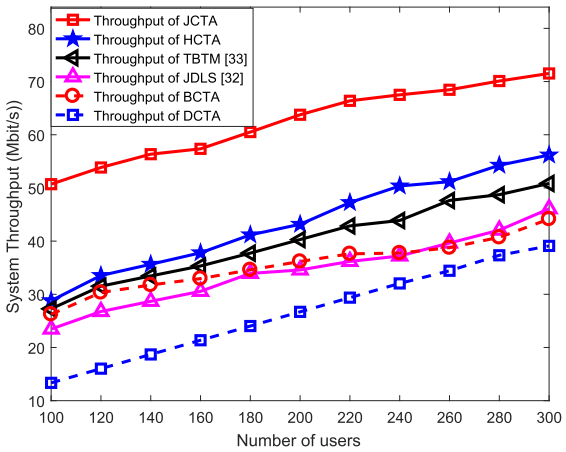


FIGURE 2. System throughput versus the number of users.

Fig. 3 shows the system EE performance for all six algorithms. We observe that the system EE decreases as the number of users increases. This coincides with the intuition that a larger number of users means a stricter condition on the co-channel interference. In DCTA algorithm, requesters get caching videos from adjacent helpers, which are less affected by the increase of users number, so the EE downtrend is relatively small. Similarly, when the number of users is larger, the HCTA algorithm has a better performance relative to the algorithm BCTA. Similar to our HCTA algorithm, the maximization goal of TBTM algorithm is to maximize the system utility function and more users choose to transmit through D2D caching which leads to the fact that EE is relatively

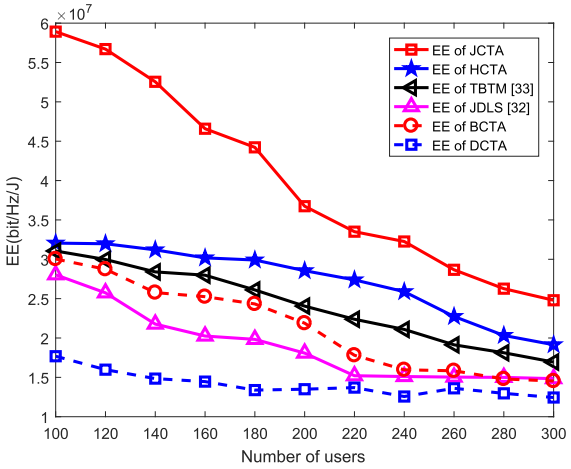


FIGURE 3. System EE versus the number of users.

smaller than that of HCTA algorithm. In the JDLS algorithm, the performance of EE is greatly affected because the interference between BS caching and D2D caching is not considered. When the number of users increases, energy efficiency can be improved because the using of D2D caching.

In Fig. 4 and Fig. 5, it is obvious that the system throughput and EE of the HCTA algorithm are larger than those without Dinkelbach method and branch and bound method. The modified branch and bound method is used to obtain the optimal link selection. Based on randomly link selection, the no-BnB HCTA algorithm can not make full use of wireless resource between the BS caching and D2D caching so that its throughput is smaller compared with the HCTA algorithm. Meanwhile, it is not the optimal link selection, which may increase the total consumed power. In the no-Dinkelbach HCTA algorithm, because power is not the final convergence result, it consumes more power than that of the no-BnB HCTA algorithm. With the number of users increases, the interference becomes more serious than that of the no-BnB HCTA algorithm.

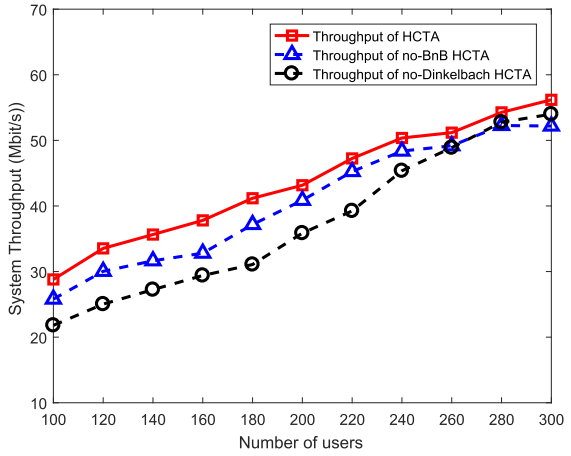


FIGURE 4. Comparison of the system throughput with different scheduling algorithms.

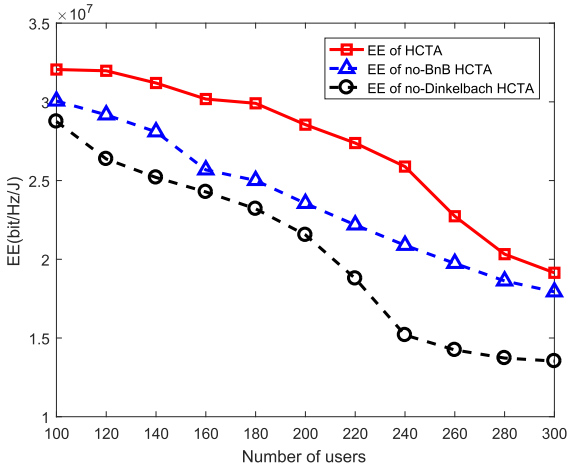


FIGURE 5. Comparison of the system EE with different scheduling algorithms.

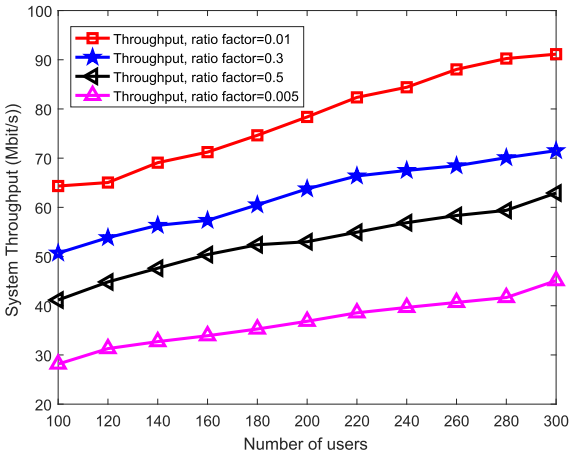


FIGURE 6. Comparison of the system throughput of JCTA algorithm with different β_d.

Fig. 5 and Fig. 6 illustrate the system throughput and the system EE of JCTA algorithm with different β_d. To make the comparison fair, the overall transmission power and other parameters are set as the same. Specifically, the maximum transmit power of base station is $P^B_{\max} = 30dBm$ and the maximum transmit power of D2D helper is $P^D_{\max} = 10dBm$. It can be seen from Fig. 5 and Fig. 6 that the system throughput and system EE with smaller β_d is much higher than that with larger β_d. When β_d is smaller, a larger proportion of cached video can be transmitted through the base station and then the system throughput and EE performance with smaller β_d is better than the larger one. As β_d continues to decrease, the performances of throughput and EE will drop sharply. When β_d is 0, the joint transmission scheme is transformed into the BCTA algorithm. Through theoretical analysis, the optimal video ratio β_d is equaled to the ratio of D2D transmission power and BS transmission power. However, this result is also affected by co-channel interference, so the system performance is not the biggest when the theoretical results are satisfied. The maximum value will occur when the video ratio is close to the transmission power ratio.

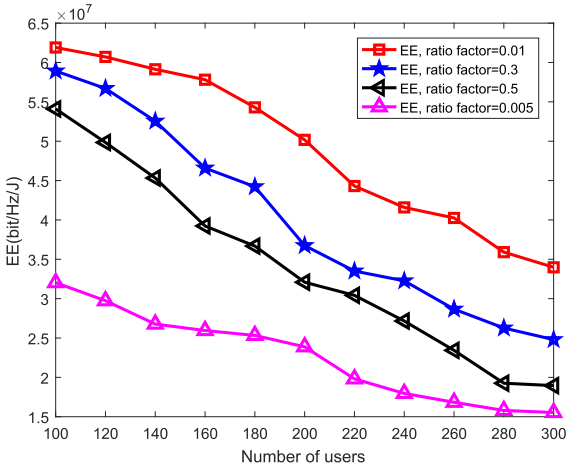


FIGURE 7. Comparison of the system EE with different β_d.

VI. CONCLUSION

Due to the emergence of a large scale of video services and low delay transmission requirement of users, the MEC-enabled network incorporating D2D communication is a good alternative technique to enhance traditional wireless networks. According to the different transmission schemes, we have proposed the hybrid caching transmission algorithm and the joint caching transmission algorithm. By considering transmission rate constraint, interference level constraint, two EE maximization problems have been formulated in this paper. Dinkelbach Method has been utilized to transform the fractional optimization problems into mixed-integer nonlinear programming problems, which can be decomposed into three subproblems: link selection, channel allocation, and power control. The resource allocation problems have been transformed into convex problems by using the modified branch and bound method and variable substitution. Then, two optimization solutions have been developed leveraging on the Lagrange dual decomposition approach. Simulation results have demonstrated the superiority of these proposed algorithms in improving system throughput and EE. Besides, the selections of the Dinkelbach method, and the modified branch and bound method in our optimization solution have been justified. The relationships among transmission power, video transmission ratio factor and transmission rate also have been investigated. These conclusions provide more insight in the network implementation and effective resource allocation in actual network. In future work, we can consider users with high mobile speed. The combination of caching offloading and computing offloading is another potential research direction.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," Cisco Syst. Inc., San Jose, CA, USA, White Paper, Feb. 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf

[2] R. Sun, Y. Wang, N. Cheng, H. Zhou, and S. Shen, "QoE driven BS clustering and multicast beamforming in cache-enabled C-RANs," in *Proc. IEEE Int. Conf. Commun.*, May 2018, pp. 1–6.

[3] J. Liao, K.-K. Wong, Y. Zhang, Z. Zheng, and K. Yang, "Coding, multicast, and cooperation for cache-enabled heterogeneous small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.

[4] 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.

[5] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 8, pp. 1791–1805, Aug. 2018.

[6] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1794–1808, Feb. 2018.

[7] J. Poderys, M. Artuso, C. M. O. Lensbøl, Claus, H. L. Christiansen, and J. Soler, "Caching at the mobile edge: A practical implementation," *IEEE Access*, vol. 6, pp. 8630–8637, 2018.

[8] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2014.

[9] L. Zhang, Z. Wang, M. Xiao, G. Wu, Y.-C. Liang, and S. Li, "Decentralized caching schemes and performance limits in two-layer networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12177–12192, Dec. 2018.

[10] Y. Long, D. Wu, Y. Cai, and J. Qu, "Joint cache policy and transmit power for cache-enabled D2D networks," *IET Commun.*, vol. 11, no. 16, pp. 2498–2506, Nov. 2017.

[11] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility increases the data offloading ratio in D2D caching networks," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–6.

[12] Y. Wang, X. Tao, X. Zhang, and Y. Gu, "Cooperative caching placement in cache-enabled D2D underlaid cellular network," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1151–1154, May 2017.

[13] C. Li, L. Toni, J. Zou, H. Xiong, and P. Frossard, "QoE-driven mobile edge caching placement for adaptive video streaming," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 965–984, Apr. 2018.

[14] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[15] G. Yu, Y. Xu, R. Yin, and F. Qu, "Interference coordination strategy based on Nash bargaining for small-cell networks," *IET Commun.*, vol. 9, no. 13, pp. 1583–1590, Sep. 2015.

[16] J. Li, J. Sun, Y. Qian, F. Shu, M. Xiao, and W. Xiang, "A commercial video-caching system for small-cell cellular networks using game theory," *IEEE Access*, vol. 4, pp. 7519–7531, 2016.

[17] J. Li, W. Chen, M. Xiao, F. Shu, and X. Liu, "Efficient video pricing and caching in heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8744–8751, Oct. 2016.

[18] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Match to cache: Joint user association and backhaul allocation in cache-aware small cell networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 3082–3087.

[19] W. Hoiles, O. N. Gharehshiran, V. Krishnamurthy, N.-D. Đào, and H. Zhang, "Adaptive caching in the YouTube content distribution network: A revealed preference game-theoretic learning approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 1, no. 1, pp. 71–85, Mar. 2015.

[20] M. Gerami, M. Xiao, J. Li, C. Fischione, and Z. Lin, "Repair for distributed storage systems with packet erasure channels and dedicated nodes for repair," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1367–1383, Apr. 2016.

[21] M. Gerami, M. Xiao, and M. Skoglund, "Partial repair for wireless caching networks with broadcast channels," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 145–148, Apr. 2015.

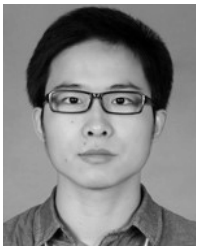
[22] C. Liang and F. R. Yu, "Enhancing mobile edge caching with bandwidth provisioning in software-defined mobile networks," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–6.

[23] T. D. Tran and L. B. Le, "Joint resource allocation and content caching in virtualized multi-cell wireless networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.

[24] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[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] J. Li *et al.*, "On social-aware content caching for D2D-enabled cellular networks with matching theory," *IEEE Internet Things J.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8025784
[27] L. Wang and H. Wu, "Fast pairing of device-to-device link underlay for spectrum sharing with cellular users," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1803–1806, Oct. 2014.
[28] 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.
[29] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.
[30] K. Wang, H. Li, F. R. Yu, and W. Wei, "Virtual resource allocation in software-defined information-centric cellular networks with device-to-device communications and imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10011–10021, Dec. 2016.
[31] H. Wu, L. Wang, T. Svensson, and Z. Han, "Resource allocation for wireless caching in socially-enabled D2D communications," in *Proc. IEEE Int. Conf. Commun.*, May 2016, pp. 1–6.
[32] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE J. Sel. Areas Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.
[33] C. Yi, S. Huang, and J. Cai, "An incentive mechanism integrating joint power, channel and link management for social-aware D2D content sharing and proactive caching," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 789–802, Apr. 2018.
[34] Y. Liu, Y. Wang, R. Sun, and Z. Miao, "Hierarchical power allocation algorithm for D2D-based cellular networks with heterogeneous statistical quality-of-service constraints," *IET Commun.*, vol. 12, no. 5, pp. 518–526, 2018.
[35] S. Wang, M. Ge, and W. Zhao, "Energy-efficient resource allocation for OFDM-based cognitive radio networks," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3181–3191, Aug. 2013.
[36] M. Taghizadeh, K. Micinski, S. Biswas, C. Ofria, and E. Torng, "Distributed cooperative caching in social wireless networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 3, pp. 1037–1053, Jun. 2014.
[37] Y. Xu, R. Q. Hu, L. Wei, and G. Wu, "QoE-aware mobile association and resource allocation over wireless heterogeneous networks," in *Proc. Global Commun. Conf.*, Dec. 2015, pp. 4695–4701.
[38] R. A. Loodaricheh, S. Mallick, and V. K. Bhargava, "Energy-efficient resource allocation for OFDMA cellular networks with user cooperation and QoS provisioning," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6132–6146, Nov. 2014.
[39] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.
[40] Y. Mo, M. Peng, H. Xiang, Y. Sun, and X. Ji, "Resource allocation in cloud radio access networks with device-to-device communications," *IEEE Access*, vol. 5, pp. 1250–1262, 2017.
[41] S. Boyd, L. Vandenberghe, and L. Faybusovich, "Convex optimization," *IEEE Trans. Autom. Control*, vol. 51, no. 11, p. 1859, Nov. 2006.
[42] Y. Cheng and M. Pesavento, "Joint optimization of source power allocation and distributed relay beamforming in multiuser peer-to-peer relay networks," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2962–2973, Jun. 2012.



YUANFEI LIU received the B.S. degree in communications engineering from the University of Science and Technology Beijing, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree with the Wireless Technology Innovation Institute, Beijing University of Posts and Telecommunications, China. His research interests are in the areas of radio resource management, device-to-device communications, and mobile edge caching in future wireless networks.

YING WANG received the Ph.D. degree in circuits and systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2003. In 2004, she was invited to work as a Visiting Researcher with the Communications Research Laboratory (renamed NiCT from 2004), Yokosuka, Japan. She was a Research Associate with The University of Hong Kong, Hong Kong, in 2005. She is currently a Professor with BUPT, where he is also the Director of the Radio Resource Management Laboratory, Wireless Technology Innovation Institute. She is active in the standardization activities of 3GPP and ITU. She took part in the performance evaluation work of the Chinese Evaluation Group as a Representative of BUPT. She has authored over 100 papers in international journals and conferences proceedings. Her research interests are in the areas of cooperative and cognitive systems, radio resource management, and mobility management in 5G systems. She was a recipient of the first prize of the Scientific and Technological Progress Award by the China Institute of Communications, in 2006 and 2009, and the second prize of the National Scientific and Technological Progress Award, in 2008. She was also selected in the New Star Program of Beijing Science and Technology Committee and the New Century Excellent Talents in University, Ministry of Education, in 2007 and 2009, respectively.

RUIJIN SUN received the B.S. degree in communications engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2013. She is currently pursuing the Ph.D. degree with the Wireless Technology Innovation Institute, Beijing University of Posts and Telecommunications, China. Her research interests are in the areas of MIMO, cooperative communications, and energy-harvesting communications in future wireless networks.

SACHULA MENG received the M.S. degree in electronics engineering from Inner Mongolia University, China, in 2014. She is currently pursuing the Ph.D. degree in telecommunication and information system with the Wireless Technology Innovation Institute, Beijing University of Posts and Telecommunications. Her research interests are in the areas of mobile cloud computing and radio resource management in wireless networks.

RUNCONG SU received the B.S. degree in communications engineering from Beijing Jiaotong University, Beijing, China, in 2016. She is currently pursuing the master's degree in telecommunication and information systems with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. Her research interests are in the areas of MIMO, SWIPT, and physical layer security in future wireless networks.