QoE-Driven Transmission-Aware Cache Placement and Cooperative Beamforming Design in Cloud-RANs

Ruijin Sun, *Student Member, IEEE*, Ying Wang, *Member, IEEE*, Nan Cheng, *Member, IEEE*, Ling Lyu, *Student Member, IEEE*, Shan Zhang, *Member, IEEE*, Haibo Zhou, *Senior Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Pre-caching popular videos at base stations (BSs) is a cost-effective way to significantly alleviate the backhaul pressure. With the video caching, the cache placement and the transmission strategy are intertwined with each other and jointly affect the system performance. Furthermore, the cache placement is updated in a much longer timescale than the transmission strategy. In this paper, the long-term transmission-aware cache placement and the short-term transmission strategy are designed to enhance the quality of experience (QoE) for the video streaming in cloud radio access networks (cloud-RANs). Specifically, consider a cache-enabled cloud-RAN, video contents are cached at BSs, and user requests are cooperatively satisfied by multiple BSs via the cooperative beamforming. To improve the weighted sum of users' QoE, the long-term transmission-aware caching problem in the caching stage and the short-term transmission problem in the delivery stage are respectively formulated, taking into account the backhaul capacity constraint, the transmission power constraint, and the storage size constraint. For the caching problem, the sample average approach is first used to approximate the long-term average QoE value. Then, cache placement strategies are devised in both centralized and distributed manner. For the transmission problem, the full-cooperative beamforming scheme is studied with the optimized cache placement, and an iterative algorithm is proposed. Simulation results show that our proposed transmission-aware cache placement and transmission strategies can achieve higher QoE performance than other cache placement and transmission strategies.**

***Index Terms*—QoE-driven, short-term transmission strategy, transmission-aware cache placement, non-convex optimization.**

I. Introduction

RECENTLY, we have witnessed the unprecedent growth of the multimedia data, in the form of text, audio, image, video streaming and their combinations. Among which, videos have been developed from standard-definition (SD) and high-definition (HD) to data-craving Ultra HD (UHD), such as 4 K, 8 K, etc. It is reported that the video traffic has reached 6,130 petabytes every month in 2016 and will exceed 40 ZB by 2020 [2]. Together with the rapid upgrading of mobile phones, UHD videos can already be supported by the phone manufacture like Apple and Huawei. The video streaming has also become the main source of the mobile data traffic. Specifically, it will account for 82% of the total traffic by the end of 2022 [3].

To cater for the exponential increase of the mobile video traffic, great efforts have been devoted to improving the network capacity [4]–[6]. From the perspective of network architecture evolution, cloud radio access networks (cloud-RANs) are proposed to enhance the system capacity [7]. In cloud-RANs, baseband signals of all base stations (BSs) are jointly processed in a central processor (CP), which is connected to BSs via backhaul links. Nevertheless, to enable this full cooperation, not only the channel state information (CSI) but also the traffic data should be shared among different BSs, leading to heavy pressure on capacity-limited backhaul links [8].

One way to overcome this issue is to deploy more high-capacity backhaul links, however, the infrastructure cost will be huge. As the price of storage devices drops, pre-caching popular videos at BSs is a cost-effective way to significantly relieve the backhaul pressure [9]. By storing the traffic-dominant popular videos at BSs in advance, pre-caching can avoid the redundant transmission and reduce the backhaul capacity requirement to 35% [10], [11]. Moreover, by bringing video contents more closer to users, pre-caching can also considerably reduce the end-to-end delay [12]. Generally, video contents delivery in wireless caching networks consists of two stages, i.e., the caching stage and the delivery stage. In the caching stage, BSs or local storages prefetch the popular video contents during the

Manuscript received May 5, 2019; revised September 21, 2019 and November 1, 2019; accepted November 6, 2019. Date of publication November 11, 2019; date of current version January 15, 2020. This work was supported by the National Nature Science Foundation of China Project 61372112 and China Scholarships Council (CSC). This article was presented in part at the IEEE Conference on Communications, Kansas City, MO, May 2018. The review of this article was coordinated by Prof. J. Joung. *(Corresponding author: Ying Wang.)*

R. Sun and Y. Wang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: sunruijin@bupt.edu.cn; wangying@bupt.edu.cn).

N. Cheng is with the School of Telecommunications Engineering, State Key Laboratory of Integrated Services Network, Xidian University, Xi'an 710071, China (e-mail: dr.nan.cheng@ieee.org).

L. Lyu is with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China (e-mail: linglyu@yeah.net).

S. Zhang is with the School of Computer Science and Technology, Beihang University, Beijing 100191, China (e-mail: zhangshan18@buaa.edu.cn).

H. Zhou is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China (e-mail: haibozhouuw@gmail.com).

X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TVT.2019.2952726

off-peak hours. In the delivery stage, users can directly acquire their requesting video contents from the local storage if these videos are cached. Correspondingly, there are challenging issues in each stage.

In the caching stage, the major challenging issue is how to rationally devise caching policies [13]–[19]. Compared with wired networks, the caching policy in wireless networks is more complicated, because of the broadcast nature of the wireless channel and the user mobility. For a single BS, the most popular contents should be stored at the cache storage to improve the cache hit ratio [13]. However, if multiple BSs cooperatively serve a set of users, the caching policies of these BSs are mutually influenced [14]. In [15], with the given network topology and content popularity distribution, Shanmugam *et al.* jointly optimized file caching policies of multiple cache-enabled helpers to reduce the average file downloading time. In [16], the probability of successful file discovery is maximized to design the cache placement in device-to-device networks, where users with different file preferences are formed as different interest groups. In [15] and [16], an assumption is made that the file can be successfully delivered within the communication region, without considering the variability of the wireless channel and the interference. To be more practical, with the consideration of the random channel fading and the interference, the probability of file success delivery and the area spectral efficiency are maximized by optimizing the caching policy in heterogeneous networks, where users associate with the best BS caching the requested contents [17].

In the delivery stage, the major challenging issue is how to reasonably allocate wireless resources with the given caching policy [1], [20]–[22]. Different from traditional wireless networks, the transmission strategy in wireless caching networks should not only consider the channel conditions, but also the cache status of BSs and the user requests. For example, to reduce the access delay or alleviate the backhaul payload, the user may prefer to associate with the BS caching the requested content instead of that with the best channel condition. In [20], a cache-induced coordinated multipoint (CoMP) scheme is proposed, in which the interference network can be transformed into a CoMP scenario if the requesting file of a user is concurrently stored and transmitted by multiple BSs. With this scheme, the average power consumption can be greatly reduced. In [21], with the given cache state, beamforming vectors are optimized to enhance the network spectrum efficiency in cloud-RANs. As different users may submit the same request with a high probability, the multicast transmission is also adopted in [22]. All these papers reveal that, with a proper pre-defined caching policy, the transmission strategy can be further optimized, which greatly benefits the system performance.

Generally, the update frequency of the video content popularity is much slower than that of the wireless channel fading. For example, a hot news may last for several days, while the wireless channel fading changes at ms level. Thus, compared with the transmission strategy, the caching policy is a long-term problem and can significantly affect transmission strategies of multiple transmission slots with different channel fading. Therefore, how to design an efficient caching policy considering flexible transmission strategies is a valuable problem. In [23], the transmission-aware caching placement is optimized to reduce the transmission power, taking into account the backhaul capacity constraint. It is demonstrated that with the proposed cache placement, the transmission power is decreased by 10%, in comparison with the optimized transmission strategy with random caching policy. Notice that the aim of previous works is to optimize the quality of service (QoS) metrics, such as the transmission power minimization, the throughput enhancement and the end-to-end delay reduction. However, the data-craving video streaming service is the main source of the future mobile traffic and the main type of cached files. Compared with QoS, quality of user experience (QoE), considering both the objective network QoS and the subjective user experience, is gradually become a major assess metric for the video streaming. The QoE enhancement is also a key means for video content providers, such as YouTube, Tencent, Alibaba Group etc, to develop the differentiated competitiveness and improve the user viscosity [24]–[26].

In this paper, we design both the long-term transmission-aware cache placement and the short-term transmission strategy for video streaming services in cloud-RANs, and adopt QoE as the performance metric. Specifically, a downlink cloud-RAN scenario is considered, where each BS is equipped with a limited storage. In our preliminary work [1], the short-term transmission strategy is designed with the given cache status. As an extension, in this paper, both the long-term transmission-aware caching problem and the short-term transmission problem are studied to maximize the weighted sum of users' QoE. Iterative algorithms are respectively proposed to solve these two problems. Simulations are conducted to show the convergence of proposed algorithms. The impacts of both cache placement and transmission strategies on the QoE performance are also evaluated. The main contributions of this paper are summarized as follows.

- Since the cache placement and the transmission strategy are updated in different timescales and the cache placement has a profound impact on the transmission strategy, the long-term transmission-aware cache placement in the caching stage and the short-term transmission strategy in the delivery stage are devised to maximize the weighted sum of users' QoE for mobile video streaming services, with the transmission power constraint, the backhaul capacity constraint and the local storage capacity constraint.
- In the caching stage, cache placement is designed to maximize the long-term weighted sum-QoE based on the file popularity distribution (FPD) and the channel distribution information (CDI). To deal with this stochastic problem, the sample average approach (SAA) is adopted to estimate the long-term QoE performance. Then, a centralized algorithm is presented to tackle this non-convex problem. Furthermore, to enhance the scalability, a distributed algorithm based on the alternating direction method of multipliers (ADMM) is also proposed.
- In the delivery stage, the full-cooperative beamforming scheme is optimized to enhance the weighted sum-QoE with the optimized cache placement. To handle this non-convex problem, an iterative algorithm with the successive convex approximation (SCA) method is proposed.

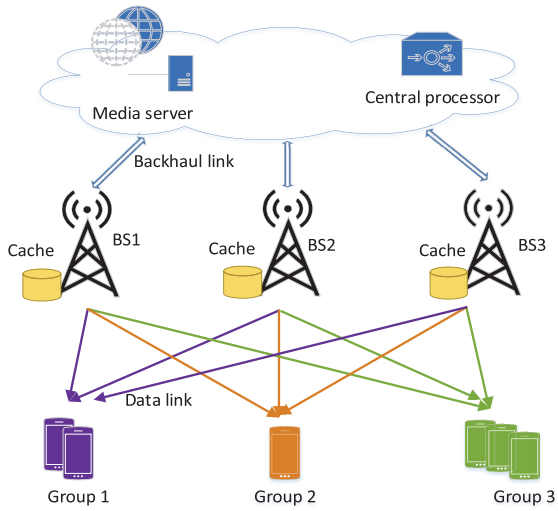


Fig. 1. An example of multicast transmission in a cache-enabled cloud-RAN. All users are formed as three multicast groups, which are cooperatively served by all BSs.

- Simulation results show that all our proposed algorithms can converge fast. In addition, the proposed long-term transmission-aware cache placement and short-term transmission strategy outperform other cache placement and transmission strategies in terms of the QoE performance.

The remainder of the paper is organized as follows. The system model is given in Section II. The long-term transmission-aware caching problem and the short-term transmission problem are formulated in Section III. The solutions to these two problems are given in Sections IV and V, respectively. Section VI presents the simulation results and Section VII concludes this paper.

Notation: Bold lower case letter and bold upper case letter are used to denote column vector and matrix, respectively. The superscripts $\mathbf{h}^H$ is the (Hermitian) conjugate transpose of $\mathbf{h}$. $\mathbb{E}\{x\}$ is the expectation of x and $\mathcal{R}e\{x\}$ is the real part of complex number x.

II. System Model

A. Network Model

The downlink cloud-RAN scenario is considered as Fig. 1, where M cache-enabled BSs, each with N_T antennas, cooperatively serve K single-antenna users. The BS set and the user set are denoted as $\mathcal{M} = \{1, \ldots, M\}$ and $\mathcal{K} = \{1, \ldots, K\}$, respectively. All BSs are connected to the CP via capacity-limited backhaul links. Denote by F the number of video files with equal size in the library and $\mathcal{F} = \{1, \ldots, F\}$ the set of video files. Files of different sizes can be split into equal-size pieces [22] and the file size is normalized to 1 without loss of generality. The main notations are listed in Table I.

The content delivery in the cache-enabled system consists of two stages, namely, the caching stage and the delivery stage. In the first stage, during the off-peak hours, BSs store popular videos in advance from the media server to update their caches based on the designed caching policy. Then, in the second stage, users can directly acquire requested files from BSs without the backhaul capacity consumption if these files are cached.

Note that the video popularity distribution changes much slower than the wireless channel fading [20]. Hence, the caching policy is usually updated in a long-term timescale, which includes many channel-varying short-term transmission time slots. Furthermore, the cache policy and the transmission strategy are respectively updated in the upper network layer and the PHY layer. Therefore, the main structure of caching policy and transmission strategy is shown as Fig. 2. Considering a long-term caching updated time interval with the fixed FPD and CDI, the cache placement remains unchanged for the following L transmission slots. In the caching stage, the caching policy is designed in the upper layer based on the long-term FPD and CDI. Then, the updated cache placement information is transferred from the upper layer to the PHY layer. In the delivery stage, at each time slot, users submit their requests independently. With the given cache placement optimized in the caching stage, the transmission strategy is designed in the PHY layer according to the instantaneous user request (IUR) and the CSI.

TABLE I
Main Notations

Notation	Description
M	The number of BSs
$\mathcal{M}$	The set of BSs
K	The number of users
$\mathcal{K}$	The set of users
N_T	The number of antennas equipped with each BS
F	The number of video files in the library
$\mathcal{F}$	The set of video files
$\mathbf{u}$	The user request vector
N	The number of formulated multicast groups
$\mathcal{N}$	The set of multicast groups
$\mathcal{G}_n$	The set of users within multicast group n
$\mathbf{w}_{m,n}$	The beamforming vector for BS m serving multicast group n
$\mathbf{h}_k$	The aggregated channel vector from all BSs to user k
γ_k^n	The signal-to-interference-plus-noise ratio of user k
$R^n_{\{\mathbf{u},\mathbf{H}\}}$	The transmission rate of multicast group n
p_f	The probability that users request file f
β	The positive parameter in Zipf's distribution
$\mathbf{C}$	The caching matrix
$c_{f,m}$	The indicator of BS m caching file f
S_m	The finite cache size of BS m
$Q^n_{k,\{\mathbf{u},\mathbf{H}\}}$	The QoE value of user k in multicast group n
$\hat{R}_k^n$	The desired rate of user k in multicast group n
a_n, b_n	Parameters related to video types in the QoE model
$\bar{Q}_k^n$	The long-term average QoE
L	The number of real transmission slots
T	The number of training transmission slots
η_k	The priority of user k
C_{Bm}	The backhaul capacity of BS m
P_m	The transmission power of BS m
γ_n, t_n, v_k	The introduced variables to tackle problem $\mathcal{P}_T$
$\tilde{c}_{f,m}$	The relaxed indicator of cache placement
$\mathcal{V}$	The variable set in problem $\mathcal{P}_C1$

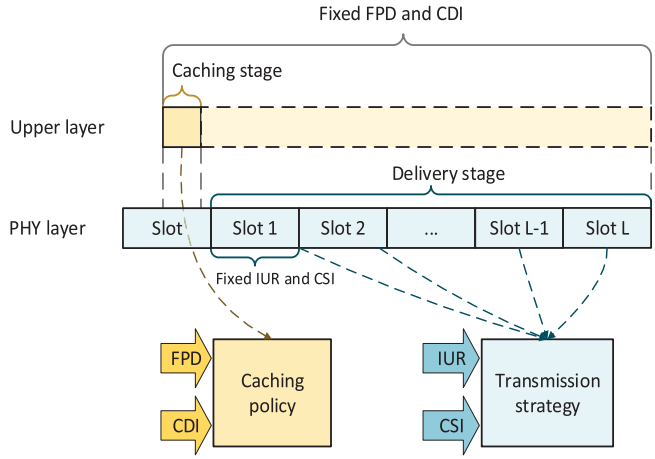


Fig. 2. The main structure of our proposed caching policy and transmission strategy. The caching policy is updated in the upper layer based on the statistical information and the transmission strategy is updated in the PHY layer based on the instantaneous information.

B. User Request Model and Cache Model

Sort all F files in the media server in a descending order based on their request probabilities. Then, denote by $\mathbf{p} = [p_1, p_2, \ldots, p_F]$ the file popularity vector, where p_f is the probability that users request file f. The file popularity is modeled as Zipf's distribution [27],

$$p_f = \frac{1/f^{\beta}}{\sum_{i=1}^{F} 1/i^{\beta}}, \forall f, \tag{1}$$

where β is a positive parameter. A higher β leads to more concentrated requests of the most popular files.

As for the cache model, it is denoted by the binary matrix $\mathbf{C} = [c_{f,m}]^{F\times M}$ the caching status, where $c_{f,m}$ denotes whether file f is cached by BS m or not. $c_{f,m} = 1$ if file f is cached by BS m and 0 otherwise. Besides, $\sum_{f=1}^{F} c_{f,m} \leq S_m$ is satisfied due to the limited cache size, where S_m is finite cache size of BS m.

C. Transmission Model

When a transmission interval starts, all users submit their video requests[1] based on the FPD $\mathbf{p}$. Each user is allowed to request at most one video file. Denote by $\mathbf{u} = [u_1, \ldots, u_k, \ldots, u_K]$ the IUR vector in a transmission interval, where $u_k = f \in \mathcal{F}$ denotes user k requesting file f. According to the user request $\mathbf{u}$, users are scheduled as different groups and each group has the same video request. Hence, the multicast beamforming is cooperatively conducted by all BSs to improve the transmission efficiency. Define N and $\mathcal{N} = \{1, \ldots, N\}$ as the number of formed multicast groups and the set of multicast groups, respectively. The set of users within group n, $\forall n \in \mathcal{N}$ is denoted as $\mathcal{G}_n$. In addition, the block fading channel is considered, i.e., the wireless channel fading is approximately unchanged for each transmission slot but different and independent from one slot to another.

Let $d_n \in \mathbb{C}$ with $\mathbb{E}[|d_n|^2] = 1$ denote the data symbol[2] of the video file requested by multicast group n. Hence, the signal delivered by BS m is expressed as $\mathbf{x}_m = \sum_{n=1}^{N} \mathbf{w}_{m,n} d_n$, where $\mathbf{w}_{m,n} \in \mathbb{C}^{N_T\times 1}$ is the beamforming vector from BS m to multicast group n. The received signal at user k, $k \in \mathcal{G}_n$, is expressed as

$$y_k^n = \sum_{m\in\mathcal{M}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,n} d_n + \sum_{i\in\mathcal{N}, i\neq n} \sum_{m\in\mathcal{M}} \mathbf{h}_{m,k}^H \mathbf{w}_{m,i} d_i + n_k, \quad \forall k \in \mathcal{G}_n, \tag{2}$$

where $\mathbf{h}_{m,k} \in \mathbb{C}^{N_T\times 1}$ is the channel vector from BS m to user k and $n_k \sim \mathcal{C}N(0, \sigma_k^2)$ is the additive white Gaussian noise (AWGN) received by user k with the noise power σ_k^2. For simplicity, (2) can also be written as

$$y_k^n = \mathbf{h}_k^H \mathbf{w}_n d_n + \sum_{i\in\mathcal{N}, i\neq n} \mathbf{h}_k^H \mathbf{w}_i d_i + n_k, \forall k \in \mathcal{G}_n, \tag{3}$$

where $\mathbf{h}_k = [\mathbf{h}_{1,k}^H, \mathbf{h}_{2,k}^H \ldots, \mathbf{h}_{M,k}^H]^H \in \mathbb{C}^{MN_T\times 1}$ is the aggregated channel vector from all BSs to user k and $\mathbf{w}_n = [\mathbf{w}_{1,n}^H, \mathbf{w}_{2,n}^H \ldots, \mathbf{w}_{M,n}^H]^H \in \mathbb{C}^{MN_T\times 1}$ is the aggregated beamforming vector from all BSs to multicast group n.

Then, according to (3), the signal-to-interference-plus-noise ratio (SINR) received by user k is given as

$$\gamma_k^n = \frac{\left|\mathbf{h}_k^H \mathbf{w}_n\right|^2}{\sum_{i\in\mathcal{N}, i\neq n} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sigma_k^2}, \forall k \in \mathcal{G}_n. \tag{4}$$

Notice that the transmission rate of multicast group is limited by the lowest transmission rate of users in the group. Denote by B the wireless bandwidth, the transmission rate of group n with the given IUR $\mathbf{u}$ and CSI $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$ is expressed as

$$R_{\{\mathbf{u},\mathbf{H}\}}^n = B \log\left(1 + \min_{k\in\mathcal{G}_n} \gamma_k^n\right). \tag{5}$$

It is worth pointing out that multicast groups $\mathcal{G}_n$ in $R_{\{\mathbf{u},\mathbf{H}\}}^n$ are scheduled based on the user request $\mathbf{u}$ and $\mathcal{G}_n$ can be regarded as the function of $\mathbf{u}$.

D. QoE Model

Traditionally, QoS metrics, such as throughput, outage probability and jitter, are used to measure the objective performance of wireless networks. However, for the user-centric video streaming services, subjective user perception is a more important metric for video providers to retain customers. Therefore, QoE, considering both the objective network QoS and the subjective user perception, is adopted in this paper as the performance metric for video streaming transmission.

[1]Note that the successful transmission of a video content may consume multiple transmission slots. If a video content is not completely transmitted in a transmission slot, it can be regarded that users submit the same video request in the next transmission slot until the transmission completes.

[2]It should be noticed that d_n denoting the data symbol of the video is used for the analysis of the baseband signal processing. In real FDD LTE system with bandwidth B = 20 MHz, each second can transfer 100.8 M data symbols like d_n [28].

As QoE has a logarithmic nature based on the Weber- Fechner Law (WFL) [29], the popular logarithmic law is widely adopted to depict the video quality experienced by users [24], [25]. It is also mentioned in [26] that the logarithmic model is suitable for correlating QoE to the perceivable QoS resource like bandwidth or bit rate. As the mapping between the QoE and rate is considered in this paper, the logarithmic QoE model is used. Let Q_k^n denote the QoE of user k in multicast group n and $\hat{R}_k^n$ represent the desired rate of user k in group n, we have

$$Q_{k,\{\mathbf{u},\mathbf{H}\}}^n = a_n \ln\left(b_n \frac{R_{\{\mathbf{u},\mathbf{H}\}}^n}{\hat{R}_k^n}\right), \tag{6}$$

where $a_n > 0$ and $b_n > 0$ are introduced parameters. The value of QoE is usually measured by the mean opinion score (MOS) scaling from 1 to 5, where 5 represents the excellent user experience and 1 represents the poor user experience. The value of a_n and b_n are related to the type of videos and the value of $\hat{R}_k^n$ depends on users' screen size, preference, etc. In [24], the empirical values of these three parameters are given via real experiments. Observing from (6) that, when $R_{\{\mathbf{u},\mathbf{H}\}}^n < \hat{R}_k^n$, the QoE value will grow rapidly with the increase of the actual transmission rate $R_{\{\mathbf{u},\mathbf{H}\}}^n$. However, when $R_{\{\mathbf{u},\mathbf{H}\}}^n > \hat{R}_k^n$, the growth trend becomes much slower.

III. Two-Stage Problem Formulation

Generally, the cache placement is updated in a long timescale, while the transmission strategy is updated in a relatively shorter time span. Furthermore, the cache placement has significant impact on the transmission strategy. Thus, the long-term transmission-aware caching problem in the caching stage and the short-term transmission problem in the delivery stage are respectively formulated in this section.

A. Long-Term Caching Problem Formulation

The optimization of caching problem is non-trivial. If all BSs cache the same popular video files, these BSs can cooperatively transmit the cached files to requesting users, and thus the transmit diversity gain can be achieved. However, this will lead to the low cache diversity and the low cache hit ratio, degrading the overall system performance. On the contrary, if all BSs cache different popular contents, the cache hit ratio can be greatly improved, but the transmission diversity will be degraded. Actually, the cache placement and the transmission strategy are intertwined with each other. Therefore, not only the file popularity, but also the transmission strategy, should be taken into account when devising the long-term cache placement.

The long-term average QoE is adopted to measure the long-term network performance. For user k, the long-term average QoE is given by

$$\bar{Q}_k^n = \mathbb{E}\left\{a_n \ln\left(b_n \frac{R_{\{\mathbf{u},\mathbf{H}\}}^n}{\hat{R}_k^n}\right)\right\}, \tag{7}$$

where the expectation is taken over the user request and the channel information. Recall that users are scheduled to multicast groups according to the IUR $\mathbf{u}$ in each transmission slot and $\mathbf{u}$ is independently generated based on the FPD $\mathbf{p}$. Thus, the user request is implicitly reflected in the multicast group $\mathcal{G}_n$ in $R_{\{\mathbf{u},\mathbf{H}\}}^n$. Besides, the channel information is reflected in $\mathbf{h}_k$ in $R_{\{\mathbf{u},\mathbf{H}\}}^n$. Therefore, both the user request and the channel information are random system parameters and whose distributions, FPD and CDI, are assumed to be fixed during the cache updated time period. One common approach to deal with (7) is to calculate its analytical expression based on the knowledge of FPD and CDI. However, due to the fact that each BS has multiple antennas and the interferences in our paper are random parameters, the accurate analytical expressions of (7) is hard to derive.

To address this problem, we resort to the SAA method in this paper to approximate the expected value [30], [31]. Note that the expected value is naturally defined as the time average of many samples. With the SAA method, the long-term expected QoE value in this paper is approximatively estimated by averaging multiple short-term QoE values. Specifically, suppose that T transmission time slots are considered. IUR $\mathbf{u}(t)$ and CSI $\mathbf{H}(t)$ of each time slot t are independently generated and then the short-term QoE value can be obtained. By the SAA method,[3] we can approximate the expected QoE value of user k by averaging the short-term QoE values of these T time slots. That is,

$$\bar{Q}_k^n \approx \frac{1}{T}\sum_{t=1}^{T} a_n \ln\left(b_n \frac{R^n(t)}{\hat{R}_k^n}\right), \tag{8}$$

where $R^n(t)$ is a simplified form of $R_{\{\mathbf{u}(t),\mathbf{H}(t)\}}^n(t)$ denoting the transmission rate at time slot t.

Therefore, the caching problem with the aim to maximize the long-term weighted sum-QoE is formulated as

$$\mathcal{P}_C: \max_{\mathbf{C},\{\{\mathbf{w}_{m,n}(t)\},\forall t\}} \quad \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \eta_k a_n \ln\left(b_n \frac{R^n(t)}{\hat{R}_k^n}\right) \tag{9a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} \|\mathbf{w}_{m,n}(t)\|_2^2 \le P_m, \forall m, \forall t, \tag{9b}$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}(1 - c_{f(n),m}) B \log_2\left(1 + \min_{k\in\mathcal{G}_n} \gamma_k^n(t)\right) \le C_{Bm}, \forall m, \tag{9c}$$

$$\sum_{f=1}^{F} c_{f,m} \le S_m, c_{f,m} \in \{0,1\}, \tag{9d}$$

where $\mathbf{w}_{m,n}(t)$ and $\gamma_k^n(t)$ are the beamforming vector and SINR at time slot t, respectively; η_k is the weighted QoE factor; P_m is the maximum allowable transmission power of BS m; C_{Bm} is the backhaul capacity. (9a) is the long-term average weighted sum-QoE value. Users desiring better QoE can pay

[3]Due to the adoption of the SAA method, our proposed cache placement algorithms in Section V also work in other user request models in addition to the Zipf distribution.

more money to network operators and get higher priority with larger weighting factors. If all users have the same priority, the weighted sum-QoE maximization can be reduced as the proportional fairness schedule, due to the logarithmic QoE model. (9b) is the BS transmission power constraint for each time slot. (9c) is the long-term average backhaul capacity constraint, where the subscript $f(n)$ in $c_{f(n),m}$ denotes that users within group n request file f, and $c_{f(n),m}$ is a binary cache placement indicator representing whether file $f(n)$ is cached at BS m. According to (9c), if the required file of group n, $f(n)$, is cached by BS m, i.e., $c_{f(n),m} = 1$, the consumed backhaul capacity for transmitting file $f(n)$ will be zero. That means the cached file $f(n)$ can be directly severed by BS. However, if $c_{f(n),m} = 0$, file $f(n)$ needs to be first acquired from the media sever to BS m with the backhaul capacity consumption, and then delivered by BS m. (9d) is the storage capacity constraint. Due to the redundant transmission, files with higher popularity will consume larger backhaul capacity if not cached at BSs. Observing from problem $\mathcal{P}_c$, (9c) would force BSs to cache more popular files. Therefore, problem $\mathcal{P}_C$ is both transmission-aware and popularity-aware.

In practice, the future IUR and CSI information of these T time slots can be predicted based on the historical data or the stochastic information of FPD and CDI. This method to handle the unknown future IUR and CSI can be treated as a "train and test" scheme [23], [31]. In the caching stage, the average QoE value, defined for T training user request and channel samples, is maximized to train a long-term cache placement matrix, $\mathbf{C}$. Then, in the delivery stage, the trained cache placement matrix is tested in L real user request and channel samples. Since the trained T samples and the real L samples both follow the same probability distribution function, when T and L go to infinity, the performance difference between the training process and the test process will go to zero. Thus, it is reasonable to use the historical data or generated data to train the cache placement matrix if T and L are enough large.

Remark: By adopting the SAA method, the long-term expected QoE value is approximated by averaging T QoE values derived from training samples. For each sample, the user requests and the channel vectors are either predicted by the historical data or respectively generated by FPD $\mathbf{p}$ and CDI. With the given sample of user requests and channel vectors, the beamforming vectors need to be designed to obtain the short-term QoE value. Thus, in the long-term caching problem $\mathcal{P}_C$, both the caching matrix $\mathbf{C}$ and the beamforming vectors $\{\mathbf{w}_{m,n}(t), \forall t\}$ for each sample are unknown variables to be optimized. The beamforming vectors for each sample $\{\mathbf{w}_{m,n}(t), \forall t\}$ are designed to help obtain the caching matrix $\mathbf{C}$ and only the caching matrix are the output in the caching stage.

It can be observed that problem $\mathcal{P}_C$ is a non-convex mixed integer nonlinear programming (MINLP) problem, owing to the complicated non-convex objective function, and the coupling of the binary cache placement indicator $\{c_{f(n),m}\}$ and the continuous beamforming vectors $\{\mathbf{w}_{m,n}(t)\}$. Thus, problem $\mathcal{P}_C$ is very hard to directly handle and iterative algorithms are put forward to efficiently solve it in both centralized and distributed way.

B. Short-Term Transmission Problem Formulation

With the optimized cache placement in the caching stage, the transmission problem in the delivery stage is then formulated. At each time slot, users independently request their preferred video files, and then BSs satisfy these requirements by cooperative beamforming based on both the channel conditions and the cache status. Assume that the IUR and CSI are acquired by the CP at the start of each transmission slot. With the given cache status, the short-term QoE maximization problem is formulated as

$$\mathcal{P}_T: \max_{\{\mathbf{w}_{m,n}\}} \quad \sum_{k=1}^{K} \eta_k a_n \ln\left(b_n \frac{B\log_2(1+\min_{k\in\mathcal{G}_n}\gamma_k^n)}{\hat{R}_k^n}\right) \tag{10a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|_2^2 \le P_m, \forall m, \tag{10b}$$

$$\sum_{n=1}^{N} (1-c_{f(n),m}) B\log_2\left(1+\min_{k\in\mathcal{G}_n}\gamma_k^n\right) \le C_{Bm}, \forall m, \tag{10c}$$

where (10a) and (10b) are the weighted sum-QoE value and the power constraint, respectively. (10c) is the backhaul capacity constraint, where $c_{f(n),m}$ is a given cache placement indicator, optimized in the caching stage.

Notice that problem $\mathcal{P}_T$ is also non-convex, owing to the objective function and the backhaul capacity constraint. According to [32], finding global maximizers to nonconvex problems is a daunting task. Generally, even verifying a feasible point is a local minimizer is NP-hard. Thus, problem $\mathcal{P}_T$ is NP-hard. In this paper, we propose an iterative algorithm to efficiently solve problem $\mathcal{P}_T$. Comparing the long-term caching problem $\mathcal{P}_C$ with the short-term transmission problem $\mathcal{P}_T$, it can be observed that the solution to problem $\mathcal{P}_C$ is on the basis of the solution to problem $\mathcal{P}_T$. So problem $\mathcal{P}_C$ is also NP-hard. In the following, we first present the solution to problem $\mathcal{P}_T$ in Section IV and then the solution to problem $\mathcal{P}_C$ in Section V.

IV. Solution to Short-Term Transmission Problem

This section solves the short-term transmission problem $\mathcal{P}_T$ and designs the transmission strategy with the given cache status. For each time slot, the short-term transmission strategy is designed to enhance the weighted sum-QoE value based on the CSI and IUR. To deal with the non-convex weighted sum-QoE objective function and the non-convex backhaul capacity constraint, the SCA method is adopted to iteratively solve problem $\mathcal{P}_T$.

A. Problem Reformulation

To make problem $\mathcal{P}_T$ tractable, we first introduce positive variables $\{\gamma_n\}$ to replace $\min_{k\in\mathcal{G}_n}\gamma_k^n$ in both the objective function and the backhaul capacity constraint. Consequently,

problem $\mathcal{P}_T$ is equivalent to problem $\mathcal{P}_T1$.

$$\mathcal{P}_T1: \max_{\{\mathbf{w}_{m,n}\},\{\gamma_n\}} \quad \sum_{k=1}^{K}\eta_k a_n \ln\left(b_n \frac{B\log_2(1+\gamma_n)}{\hat{R}_k^n}\right) \quad (11a)$$

$$\text{s.t.} \quad \frac{\left|\mathbf{h}_k^H\mathbf{w}_n\right|^2}{\sum_{i\in\mathcal{N},i\neq n}\left|\mathbf{h}_k^H\mathbf{w}_i\right|^2+\sigma_k^2} \geq \gamma_n, \forall k\in\mathcal{G}_n, \forall n, \quad (11b)$$

$$\sum_{n=1}^{N}(1-c_{f(n),m})B\log_2(1+\gamma_n) \leq C_{Bm}, \forall m, \quad (11c)$$

$$\text{and } (10b), \quad (11d)$$

where $\gamma_n = \min_{k\in\mathcal{G}_n}\gamma_k^n$ and constraint (11b) is the SINR constraint. Obviously, problem $\mathcal{P}_T1$ is still non-convex, due to (11a), (11b) and (11c).

To deal with the non-convex problem $\mathcal{P}_T1$, positive slack variables $\{t_n\}$ and $\{v_k\}$ are further introduced to replace the actual transmission rate in (11a) and the interference in (11b), respectively. Thus, problem $\mathcal{P}_T1$ is equivalently rewritten as

$$\mathcal{P}_T2: \max_{\substack{\{\mathbf{w}_{m,n}\},\{\gamma_n\},\\ \{t_n\},\{v_k\}}} \quad \sum_{k=1}^{K}\eta_k a_n \ln t_n + \sum_{k=1}^{K}\eta_k a_n \ln\frac{b_n}{\hat{R}_k^n} \quad (12a)$$

$$\text{s.t.} \quad B\log_2(1+\gamma_n) \geq t_n, \forall n, \quad (12b)$$

$$\frac{\left|\mathbf{h}_k^H\mathbf{w}_n\right|^2}{v_k} \geq \gamma_n, \forall k\in\mathcal{G}_n, \forall n, \quad (12c)$$

$$\sum_{i=1,i\neq n}^{N}\left|\mathbf{h}_k^H\mathbf{w}_i\right|^2+\sigma_k^2 \leq v_k, \forall k\in\mathcal{G}_n, \forall n, \quad (12d)$$

$$\sum_{n=1}^{N}(1-c_{f(n),m})B\log_2(1+\gamma_n) \leq C_{Bm}, \forall m, \quad (12e)$$

$$\text{and } (10b), \quad (12f)$$

which is convex except for constraints (12c) and (12e). The equivalence of these two problems is proved in [1].

B. SCA for Non-Convex Constraints

To efficiently tackle the non-convex constraints (12c) and (12e) in problem $\mathcal{P}_T2$, the SCA method is adopted [33]. The basic idea is to successively make the problem convex, by replacing the non-convex part with its first-order Tayor expansion. Observing from (12c) that, $\frac{|\mathbf{h}_k^H\mathbf{w}_n|^2}{v_k}$ is a quasi-convex function [34]. Thus, let $\mathbf{w}_n^{(i)}$ and $v_k^{(i)}$ respectively denote values of $\mathbf{w}_n$ and v_k obtained from iteration i, at iteration $i+1$, we have

$$\frac{\left|\mathbf{h}_k^H\mathbf{w}_n\right|^2}{v_k} \geq \frac{2\mathcal{R}e\left\{\left(\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_n^{(i)}\right)^H\mathbf{w}_n\right\}}{v_k^{(i)}} - \frac{\left|\mathbf{h}_k^H\mathbf{w}_n^{(i)}\right|^2}{\left(v_k^{(i)}\right)^2}v_k,$$

$$\forall k\in\mathcal{G}_n, \forall n, \quad (13)$$

Algorithm 1: SCA-Based Full-Cooperative Multicast Beamforming.

1: Initialize variable $\mathbf{w}_{m,n}^{(0)}$ and calculate $v_k^{(0)}$, and $\gamma_n^{(0)}$;
2: **while** the backhaul capacity constraint is not satisfied **do**
3: $\quad\mathbf{w}_{m,n}^{(0)} = 0.5\mathbf{w}_{m,n}^{(0)}$;
4: $\quad$Calculate $v_k^{(0)}$ and update $\gamma_n^{(0)}$;
5: **end while**
6: Set $i := 1$;
7: **while** $\frac{u^{(i)}-u^{(i-1)}}{u^{(i-1)}} \geq 10^{-3}$ **do**
8: $\quad$Solve problem $\mathcal{P}_T3$ by CVX to obtain $\mathbf{w}_{m,n}^{(i)}$, $v_k^{(i)}$ and $\gamma_n^{(i)}$;
9: $\quad$Set $i := i+1$;
10: **end while**
11: **return** $\mathbf{w}_{m,n}^{(i)}$.

where the right side is the first-order Tayor expansion [35] and also a lower bound of the left side.

Similarly, the left side of (12e) is concave, and its first-order Tayor expansion is a lower bound. Hence, at iteration $i+1$, we have

$$\sum_{n=1}^{N}(1-c_{f(n),m})B\log_2(1+\gamma_n) \leq \sum_{n=1}^{N}(1-c_{f(n),m})B \times\left(\log_2\left(1+\gamma_n^{(i)}\right)+\frac{1}{\left(1+\gamma_n^{(i)}\right)\ln 2}\left(\gamma_n-\gamma_n^{(i)}\right)\right), \forall m, \quad (14)$$

where $\gamma_n^{(i)}$ is the value of γ_n obtained from iteration i.

As a result, at iteration $i+1$, the approximated convex problem is given as

$$\mathcal{P}_T3: \max_{\substack{\{\mathbf{w}_{m,n}\},\{\gamma_n\},\\ \{t_n\},\{v_k\}}} \quad \sum_{k=1}^{K}\eta_k a_n \ln t_n + \sum_{k=1}^{K}\eta_k a_n \ln\frac{b_n}{\hat{R}_k^n} \quad (15a)$$

$$\text{s.t.} \quad (12b),(12d),(10b), \quad (15b)$$

$$\frac{2\mathcal{R}e\left\{\left(\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_n^{(i)}\right)^H\mathbf{w}_n\right\}}{v_k^{(i)}} - \frac{\left|\mathbf{h}_k^H\mathbf{w}_n^{(i)}\right|^2}{\left(v_k^{(i)}\right)^2}v_k \geq \gamma_n, \quad \forall k\in\mathcal{G}_n, \forall n, \quad (15c)$$

$$\sum_{n=1}^{N}(1-c_{f(n),m})B\Bigg(\log_2(1+\gamma_n^{(i)}) + \frac{1}{(1+\gamma_n^{(i)})\ln 2}(\gamma_n-\gamma_n^{(i)})\Bigg) \leq C_{Bm}, \forall m, \quad (15d)$$

which can be optimally solved by the CVX [36].

The proposed algorithm to solve the short-term transmission problem $\mathcal{P}_T$ is outlined in Algorithm 1, in which $u^{(i)}$ is the weighted sum-QoE of problem $\mathcal{P}_T3$ achieved in iteration i. In short, Algorithm 1 has two procedures: 1), with the given $\gamma_n^{(i)}$, $\mathbf{w}_{m,n}^{(i)}$ and $v_k^{(i)}$, the off-the-shelf CVX is adopted to handle

problem $\mathcal{P}_T3$; 2), update these variables based on the previous iteration. These two procedures are iteratively updated until the stopping criteria is satisfied.

C. Convergence and Complexity Analysis

Notice that the linear approximation of (15c) and (15d) is conservative, and thus the obtained solution of Algorithm 1 can converge to the KKT point of problem $\mathcal{P}_T2$ [33]. Recall that problem $\mathcal{P}_T2$ is equivalent to problem $\mathcal{P}_T$, Therefore, Algorithm 1 can return a local optimal solution to the short-term transmission problem $\mathcal{P}_T$. Simulation results in Section VI also demonstrate that Algorithm 1 can converge fast.

Note that the computational complexity of Algorithm 1 is mainly in the solving of problem $\mathcal{P}_T3$, i.e., line 8 in Algorithm 1. Let N_v be the number of variables, the computational complexity of using interior-point method is $O(N_v^{3.5})$ [37]. Thus, the computational complexity of using the interior-point method within CVX to solve problem $\mathcal{P}_T3$ is $O(((N_TM+2)N+K)^{3.5})$, where $(N_TM+2)N+K$ is the number of variables in problem $\mathcal{P}_T3$. If the algorithm needs N_{ite} number of iterations to satisfy the stopping criteria, the total computational complexity of Algorithm 1 is $O(((N_TM+2)N+K)^{3.5}N_{ite})$.

V. Solution to Long-Term Caching Problem

In this section, the long-term weighted average sum-QoE maximization problem $\mathcal{P}_C$ is solved and the cache placement is designed. Both the centralized and ADMM-based distributed cache placement are proposed.

A. Centralized Cache Placement

Notice that problem $\mathcal{P}_C$ is a MINLP problem, to deal with the binary variable $c_{f,m}$, we first relax it to a continuous one within 0 to 1, i.e., $\tilde{c}_{f,m}\in[0,1]$. Then, in the similar manner with the short-term transmission problem $\mathcal{P}_T$, problem $\mathcal{P}_C$ can be reformulated as

$$\mathcal{P}_C1:\max_{\mathcal{V}}\ \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\eta_k a_n\ln t_n(t)+\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\eta_k a_n\ln\frac{b_n}{\hat{R}_k^n} \tag{16a}$$

$$\text{s.t.}\quad B\log_2(1+\gamma_n(t))\geq t_n(t),\forall n,\forall t, \tag{16b}$$

$$\frac{\left|\mathbf{h}_k^H(t)\mathbf{w}_n(t)\right|^2}{v_k(t)}\geq\gamma_n(t),\forall k\in\mathcal{G}_n,\forall n,\forall t, \tag{16c}$$

$$\sum_{i=1,i\neq n}^{N}\left|\mathbf{h}_k^H(t)\mathbf{w}_i(t)\right|^2+\sigma_k^2 \leq v_k(t),\forall k\in\mathcal{G}_n,\forall n,\forall t, \tag{16d}$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}(1-\tilde{c}_{f(n),m})B\log_2(1+\gamma_n(t)) \leq\frac{1}{T}\sum_{t=1}^{T}\mathrm{C}_{Bm},\forall m, \tag{16e}$$

$$\sum_{n=1}^{N}\|\mathbf{w}_{m,n}(t)\|_2^2\leq P_m,\forall m,\forall t, \tag{16f}$$

$$\sum_{f=1}^{F}\tilde{c}_{f,m}\leq S_m,\tilde{c}_{f,m}\in[0,1], \tag{16g}$$

where $\{t_n(t)\}$, $\{\gamma_n(t)\}$ and $\{v_k(t)\}$ are introduced variables and $\mathcal{V}=\{\{\tilde{c}_{f(n),m}\},\{\mathbf{w}_{m,n}(t)\},\{\gamma_n(t)\},\{t_n(t)\},\{v_k(t)\},\forall t\}$ is the variable set.

Note that problem $\mathcal{P}_C1$ is still non-convex due to the SINR constraint (16c) and the backhaul capacity constraint (16e). To handle the non-convex (16c), similar to (13), the SCA method can be adopted to iteratively solve problem $\mathcal{P}_C1$. However, for the non-convex backhaul capacity constraint (16e), the SCA method in (14) can not be directly used, due to the coupling of $\tilde{c}_{f(n),m}$ and $\gamma_n(t)$. To tackle this obstacle, at iteration $i+1$, we replace the variable $\gamma_n(t)$ in (16e) with the fixed $\gamma_n^{(i)}(t)$, which is the value of $\gamma_n(t)$ from iteration i. As a result, the backhaul capacity constraint (16e) becomes a convex linear constraint with respect to $\tilde{c}_{f(n),m}$ at each iteration.

Suppose that, at iteration $i+1$, $\mathbf{w}_n^{(i)}(t)$, $v_k^{(i)}(t)$ and $\gamma_n^{(i)}(t)$ are given. Then, the approximated convex optimization problem at iteration $i+1$ is given as

$$\mathcal{P}_C2:\max_{\mathcal{V}}\ \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\eta_k a_n\ln t_n(t)+\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\eta_k a_n\ln\frac{b_n}{\hat{R}_k^n} \tag{17a}$$

$$\text{s.t.}\quad \frac{2\mathcal{R}e\left\{\left(\mathbf{h}_k(t)\mathbf{h}_k^H(t)\mathbf{w}_n^{(i)}(t)\right)^H\mathbf{w}_n(t)\right\}}{v_k^{(i)}(t)} -\frac{\left|\mathbf{h}_k^H(t)\mathbf{w}_n^{(i)}(t)\right|^2}{\left(v_k^{(i)}(t)\right)^2}v_k(t)\geq\gamma_n(t),\forall k\in\mathcal{G}_n,\forall n,\forall t, \tag{17b}$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}(1-\tilde{c}_{f(n),m})B\log_2(1+\gamma_n^{(i)}(t)) \leq\frac{1}{T}\sum_{t=1}^{T}\mathrm{C}_{Bm},\forall m, \tag{17c}$$

$$(16b),(16d),(16f),(16g), \tag{17d}$$

which is convex and can be solved by the general purpose interior-point method within CVX in a centralized manner. Notice that the size of problem $\mathcal{P}_C2$ is about T times of the short-term transmission problem $\mathcal{P}_T3$. Similar to Algorithm 1 in the previous section, the computational complexity of the centralized caching algorithm is $O(N_v^{3.5}N_{ite})$, where $N_v=FM+((MN_T+2)N+K)T$ is the number of variables in problem $\mathcal{P}_C2$. The convergence of the centralized algorithm is demonstrated in Section VI.

Since we have relaxed the achieved cache placement indicator to be a continuous real number within 0 and 1, the obtained

objective value of problem $\mathcal{P}_C1$ serves as an upper bound of problem $\mathcal{P}_C$. Interestingly, the fractional cache placement indicator can be meaningfully interpreted as the ratio of each media file that is cached at the BS. Besides, it can also be explained as the probability that each media file is preferred to be cached at the BS. Not only the file popularity, but also the transmission strategy, are considered in this cache probability. To recover to binary cache placement indicator in this paper, we treat the fractional indicator as the cache probability. Each BS caches media files in sequence based on the obtained cache probability until the cache capacity is full. Since the caching policy is just trained based on the historical data or generated data, there is no need to make the recovered binary cache placement indicator strictly satisfy the feasible region of caching problem $\mathcal{P}_C$. However, in the delivery stage, with the real data and the given binary cache placement matrix, the optimized transmission variables should strictly be within the feasible region of the short-term transmission problem $\mathcal{P}_T$.

B. ADMM-Based Distributed Cache Placement

Although problem $\mathcal{P}_C2$ can be directly solved in a centralized way, it may lack the scalability. To tackle this scalability problem, we further propose an ADMM-based distributed algorithm to effectively solve problem $\mathcal{P}_C2$. The basic idea is to decompose the original problem into several independent subproblems by creating multiple local variables as copies of coupled variables [38].

Toward this end, the first step is to reformulate problem $\mathcal{P}_C2$ as the ADMM consensus form. Observing from problem $\mathcal{P}_C2$ that, only variable $\tilde{c}_{f,m}$ is coupled in different time slots in the backhaul capacity constraint (17c) and the cache capacity constraint (16g). In order to decouple this variable and parallelize the problem $\mathcal{P}_C2$, the local variable of $\tilde{c}_{f,m}$ for each time slot needs to be introduced. Define

$$\tilde{c}_{f,m}(t) = \tilde{c}_{f,m}, \forall f, \forall m, \forall t, \tag{18}$$

where $\tilde{c}_{f,m}(t)$ is the local copy of $\tilde{c}_{f,m}$ for each time slot t. Then, for each time slot t, the local variable set, $\mathcal{V}(t) = \{\{\tilde{c}_{f,m}(t)\}, \{\mathbf{w}_{m,n}(t)\}, \{\gamma_n(t)\}, \{t_n(t)\}, \{v_k(t)\}\}$, should satisfy the following feasible region:

$$\mathcal{R}(t) = \left\{ \begin{array}{l} (16b), (16d), (16f), (17b), \\ \frac{1}{T}\sum\limits_{t=1}^{T}\sum\limits_{n=1}^{N}(1-\tilde{c}_{f(n),m}(t))R_n^{(i)}(t) \le \frac{1}{T}\sum\limits_{t=1}^{T}\mathrm{C}_{Bm}, \\ \sum\limits_{f=1}^{F}\tilde{c}_{f,m}(t) \le S_m, \tilde{c}_{f,m}(t) \in [0,1], \end{array} \right\}, \tag{19}$$

where $R_n^{(i)}(t) = B\log_2(1+\gamma_n^{(i)}(t))$. The corresponding objective function of each time slot t is represented as

$$Q(t) = \left\{ \begin{array}{l} -\sum_{k=1}^{K}\eta_k a_n \ln t_n(t) - \sum_{k=1}^{K}\eta_k a_n \ln \frac{b_n}{\hat{R}_k^n}, \\ \mathcal{V}(t) \subset \mathcal{R}(t), \\ \infty, \mathcal{V}(t) \not\subset \mathcal{R}(t). \end{array} \right. \tag{20}$$

Thus, problem $\mathcal{P}_C2$ can be equivalently rewritten as

$$\mathcal{P}_C3: \min_{\mathcal{V}(t),\{\tilde{c}_{f,m}\}} \quad \sum_{t=1}^{T} Q(t) \tag{21a}$$

$$\text{s.t.} \quad \tilde{c}_{f,m} = \tilde{c}_{f,m}(t), \forall f, \forall m, \forall t. \tag{21b}$$

It can be observed that the objective function in problem $\mathcal{P}_C3$ is separable across different time slots, and the constraints (21b) are linear equations. Therefore, problem $\mathcal{P}_C3$ is a standard ADMM consensus form. Then, the local variable set $\mathcal{V}(t)$ and the global variable $\{\tilde{c}_{f,m}\}$ can be alternatively optimized by the ADMM method.

Denote by ρ the penalty parameter and $\lambda_{f,m}(t)$ the multiplier of the constraint (21b). Similar to [38], the scaled form of augmented Lagrangian function of problem $\mathcal{P}_C3$ is represented as

$$\mathcal{L}_\rho(\{\mathcal{V}(t)\}, \{\tilde{c}_{f,m}\}; \{\lambda_{f,m}(t)\}) = \sum_{t=1}^{T} Q(t) + \frac{\rho}{2}\sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{m=1}^{M}|\tilde{c}_{f,m} - \tilde{c}_{f,m}(t) + \lambda_{f,m}(t)|^2. \tag{22}$$

With the ADMM method, the Lagrangian function (22) is minimized by alternatively optimizing the local variable set $\{\mathcal{V}(t)\}$, the global variable $\{\tilde{c}_{f,m}\}$ and the dual variable $\{\lambda_{f,m}(t)\}$. In what follows, the optimization of all these variable sets are respectively given, each of which can be solved in parallel.

With the given dual variable $\{\lambda_{f,m}(t)\}$ and the global variable $\{\tilde{c}_{f,m}\}$, the local variables update can be casted as

$$\min_{\substack{\mathcal{V}(t)\subset\mathcal{R}(t),\\ \forall t}} \sum_{t=1}^{T} Q(t) + \frac{\rho}{2}\sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{m=1}^{M} \times |\tilde{c}_{f,m} - \tilde{c}_{f,m}(t) + \lambda_{f,m}(t)|^2. \tag{23}$$

This problem can be decomposed into T independent subproblems, one for each time slot t:

$$\mathcal{P}_C4: \min_{\mathcal{V}(t)\subset\mathcal{R}(t)} Q(t) + \frac{\rho}{2}\sum_{f=1}^{F}\sum_{m=1}^{M}|\tilde{c}_{f,m} - \tilde{c}_{f,m}(t) + \lambda_{f,m}(t)|^2, \tag{24}$$

which is a convex and can be solved by CVX. The problem size is much smaller than the original problem $\mathcal{P}_C2$.

Then, with the given local variable set and the dual variable, the global variable $\{\tilde{c}_{f,m}\}$ update is given as

$$\min_{\substack{0\le\tilde{c}_{f,m}\le 1,\\ \forall f,\forall m}} \sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{m=1}^{M}|\tilde{c}_{f,m} - \tilde{c}_{f,m}(t) + \lambda_{f,m}(t)|^2. \tag{25}$$

This problem can also be decomposed into FM independent subproblems:

$$\min_{0\le\tilde{c}_{f,m}\le 1} \sum_{t=1}^{T}|\tilde{c}_{f,m} - \tilde{c}_{f,m}(t) + \lambda_{f,m}(t)|^2, \tag{26}$$

Algorithm 2: Admm-Based Distributed Long-Term Caching Algorithm.

1: Initialize variable $\mathbf{w}_{m,n}^{(0)}$ and calculate $v_k^{(0)}$, $\gamma_n^{(0)}$;
2: **while** the backhaul capacity constraint is not satisfied **do**
3: $\mathbf{w}_{m,n}^{(0)} = 0.5\mathbf{w}_{m,n}^{(0)}$;
4: Calculate $v_k^{(0)}$ and update $\gamma_n^{(0)}$;
5: **end while**
6: Set $i := 1$;
7: **while** $\frac{u^{(i)}-u^{(i-1)}}{u^{(i-1)}} \geq 10^{-3}$ **do**
8: Initialize $c_{f,m}^{(i,0)}$, $\lambda_{f,m}^{(i,0)}(t)$ and penalty parameter ρ;
9: Set $j := 1$;
10: **while** the convergence condition is not met **do**
11: Update the local variable set $\mathcal{V}^{(i,j)}(t)$ by solving problem $\mathcal{P}_C4$;
12: Update global variable $\tilde{c}_{f,m}^{(i,j)}$ via (27);
13: Update dual variable $\lambda_{f,m}^{(i,j)}(t)$ via (28);
14: Set $j := j + 1$;
15: **end while**
16: Update $\mathbf{w}_{m,n}^{(i)}$, calculate $v_k^{(i)}$ and $\gamma_n^{(i)}$;
17: Set $i := i + 1$;
18: **end while**
19: **return** $\tilde{c}_{f,m}$.

whose solution is

$$\tilde{c}_{f,m} = \frac{\sum_{t=1}^{T}(\tilde{c}_{f,m}(t) - \lambda_{f,m}(t))}{T}. \tag{27}$$

At last, with the given local variable set and the global variable, the dual variable can be updated as follows:

$$\lambda_{f,m}(t) = \lambda_{f,m}(t) + \tilde{c}_{f,m}(t) - \tilde{c}_{f,m}, \forall f, \forall m, \forall t. \tag{28}$$

To sum up, the proposed distributed algorithm consists of two loops: 1), in the outer loop, similar to the centralized algorithm, SCA method is adopted to iteratively make the long-term caching problem $\mathcal{P}_C$ convex; 2), in the inner loop, the convex problem $\mathcal{P}_C2$ is solved by ADMM in parallel. For the ADMM, the global variables update, the local variable sets update and the multipliers update are repeated until the stopping criteria is met. The distributed algorithm is outlined in Algorithm 2, in which $u^{(i)}$ is the weighted sum-QoE achieved in iteration i of the outer loop. Note that the achieved cache placement indicator of Algorithm 2 is also a continuous value, and the binary cache placement recovery policy is the same with Subsection A, which is omitted here.

Since problem $\mathcal{P}_C3$ is a convex problem and has two optimization variable blocks, i.e., the local variable set $\mathcal{V}(t)$ and the global variable $\{\tilde{c}_{f,m}\}$, the ADMM can achieve the optimal solution of problem $\mathcal{P}_C2$ or $\mathcal{P}_C3$ [38]. That means, the inner loop (line 10–14) in Algorithm 2 can obtain the optimal solution. Furthermore, as shown in Section VI, the outer loop of our proposed Algorithm 2 can also converge fast.

Although the total computational complexity of distributed Algorithm 2 is roughly the same as that of the centralized algorithm in Subsection A, Algorithm 2 can disperse the computational task to T different agents to implement. Thus, the computational complexity for each agent can be significantly reduced, which results in short execution time. Specifically, the computational complexity for each agent is dominated by the solving of problem $\mathcal{P}_C4$, and its computational complexity is $O((FM + (MN_T + 2)N + K)^{3.5})$.

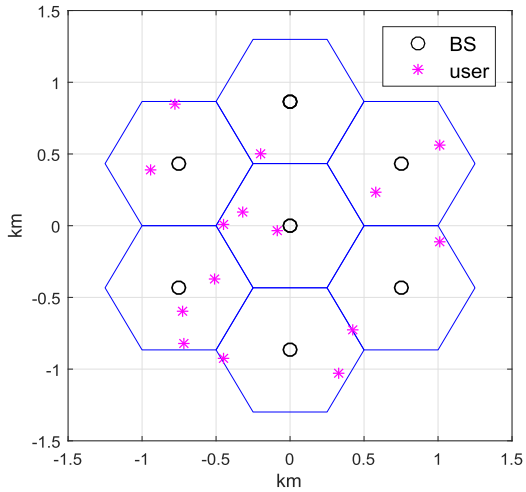


Fig. 3. Simulation scenario with 7 BSs and 15 randomly generated users.

VI. Numerical Results

This section shows the numerical results of our proposed transmission-aware cache placement and transmission strategies. The simulation scenario is shown in Fig. 3, where the numbers of BSs and users are respectively set as $M = 7$ and $K = 15$. The radius of each cell is set as 500 m. Each BS is in the center of its corresponding cell, and users are uniformly distributed within the system. The media server has $F = 50$ video contents and the bandwidth is $B = 10$ MHz. The cache capacity, the backhaul capacity and the number of antennas of each BS are respectively set as $S_m = 10$, $C_{Bm} = 50$ Mbps and $N_T = 2$. The power spectral density of noise is -174 dBm/Hz and the transmission power is 40 dBm. For the wireless channel realization, both the large-scale and the small-scale fading are considered, which are modeled as $PL = 148.1 + 37.6\text{lg}(d(\text{km}))$ in dB and Rayleigh fading, respectively. For simplicity, all users are assumed to have the same priority and thus the same the weighting factor, i.e., $\eta_k = 1$. Furthermore, we consider a Zipf's distribution with $\beta = 1$ in our simulation.

In [24], the QoE function related parameters, a_n, b_n and $\hat{R}_k^n$ are given based on the experimental study. In that paper, parameters a_n and b_n depend on the media type, and the desired rate of user k within multicast group n, $\hat{R}_k^n$, depends on the media type, the media outlet (e.g. screen size) and the user preference. Normally, $\hat{R}_k^n$ can be modeled as a uniform distribution within $[r_0, r_n]$, where r_0 and r_n are respectively the lower bound and the upper bound of $\hat{R}_k^n$. Table II in [24] lists the experimental values of a_n, b_n and r_0, r_n for several different types of videos. In our simulation, we randomly select the values of a_n, b_n and r_0, r_n for group n from that table. Then, the required transmission rate

TABLE II
SIMULATION PARAMETERS

Description	Value
The number of BSs, M	7
The number of users, K	15
The radius of each cell	500 m
The number of video contents, F	50
The bandwidth, B	10 MHz
The cache capacity, S_m	10
The backhaul capacity, C_{Bm}	50 Mbps
The number of antennas at each BS, N_T	2
The power spectral density	-174 dBm/Hz
The transmission power of BSs, P_m	40 dBm
The weighting factor, η_k	1
The parameter in Zipf distribution, β	1
The video type related parameters, a_n, b_n, r_0, r_n	randomly chosen from TABLE II in [24]
The desired rate of user k within group n, $\hat{R}_k^n$	uniform distribution within $[r_0, r_n]$
The number of training time slots, T	100

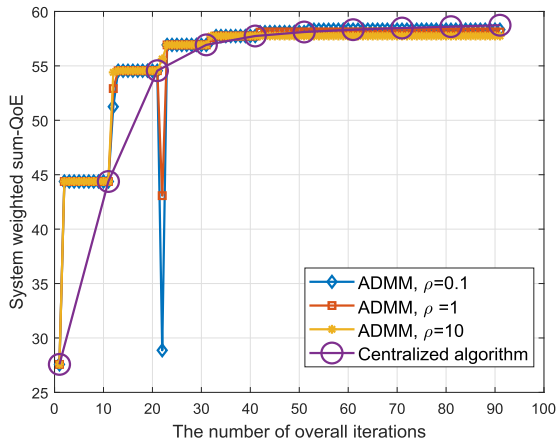


Fig. 4. Convergence behaviour of our proposed ADMM-based distributed cache placement algorithm with different ρ.

$\hat{R}_k^n$ for the user k within the group n is randomly chosen from $[r_0, r_n]$ based on the uniform distribution.

As mentioned in Subsection III-A, the long-term system QoE performance is estimated by averaging T short-term QoE values. We assume that $T = 100$. To overcome the unknown future data (i.e., CSI and IUR of each short-term time slot) problem, the data for training the cache placement matrix in our simulation is independently generated based on the CDI and the FPD. Then, in the delivery stage, all results are the average value of 100 simulation trials to make the curve smooth. To be reliable, the user location, the user request and the channel realization are independently generated in each trial. The simulation parameters are summarized in Table II.

A. The Convergence Behaviour of Proposed Algorithms

In Fig. 4, the impact of the value of the penalty parameter, ρ, on the convergence behaviour of ADMM-based distributed cache placement algorithm is shown. Observing from Algorithm 2, the

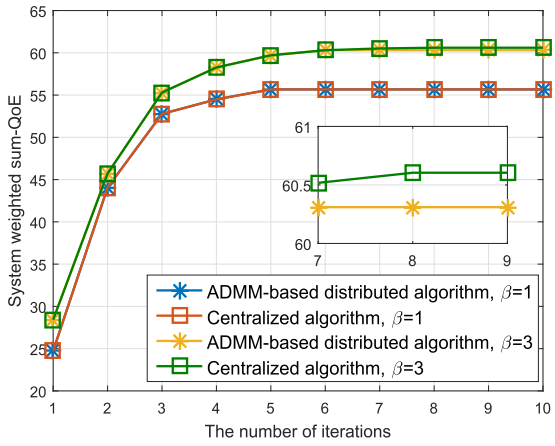


Fig. 5. Convergence behaviour of our proposed caching algorithms with $P_m = 40$ dBm.

SCA method is used in the outer loop to make problem $\mathcal{P}_C1$ convex in each iteration and the ADMM method is used in the inner loop to distributively solve convex problem $\mathcal{P}_C2$. To show the impact of different ρ in the ADMM method on the convergence behaviour, both the outer iterations and the inner iterations are plotted in Fig. 4. The curve of the centralized algorithm shows the convergence behaviour of the outer loop when the centralized algorithm is used. The curve of ADMM algorithm within each outer loop shows the convergence behaviour of the inner loop when the ADMM algorithm is adopted. It can been seen from Fig. 4 that, results with different ρ finally converges to almost the same QoE value with only a small gap. However, $\rho = 10$ can provide higher convergence rate than $\rho = 1$ and $\rho = 0.1$. Thus, $\rho = 10$ is adopted in the following simulation.

In Fig. 5, convergence behaviours of our proposed centralized and distributed caching algorithms with $\beta = 1$ and $\beta = 3$ are shown. It shows that both centralized and ADMM-based distributed algorithms can converge within several iterations. Moreover, the performance gap between these two algorithms is extremely narrow, which verifies the effectiveness of our proposed ADMM-based distributed algorithm.

Then, the convergence behaviours of the proposed iterative transmission algorithm are shown in Fig. 6. It can be seen that our proposed SCA-based full-cooperative multicast beamforming algorithm can converge within several iterations for five independent channel realizations.

B. The Effect of Cache Placement on the System Performance

The impact of cache placement on the weighted sum-QoE utility is investigated in this subsection. For comparison, four commonly seen cache placement strategies are considered as benchmarks, which are specified in the following.

- *The most popular cache placement:* The most popular files are stored by all BSs according to the user request probability. Hence, it is a file popularity aware cache placement strategy. If each BS has the same shortage capacity, the cached files in each BS will be the same, which can improve the opportunity for the cooperative transmission. However,

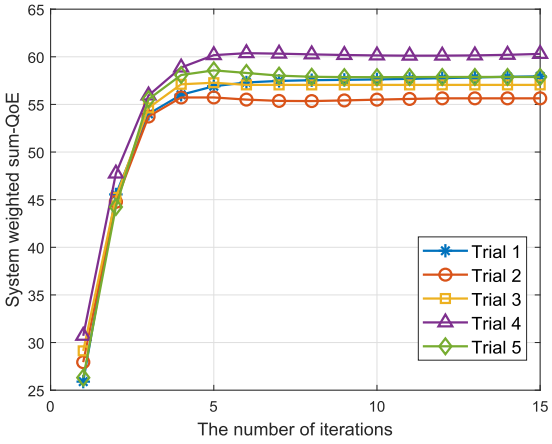


Fig. 6. Convergence behaviour of our proposed transmission algorithms with $P_m = 40$ dBm.

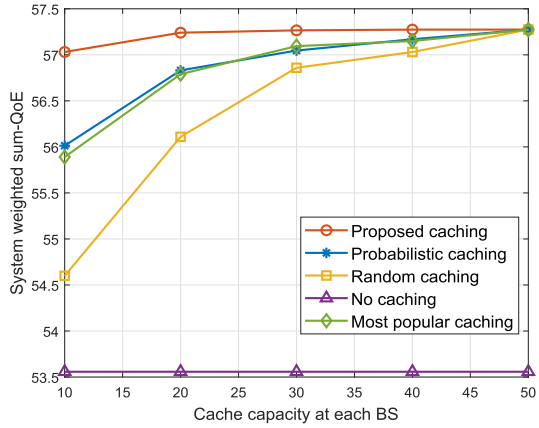


Fig. 8. The impact of cache placement on the system weighted QoE performance with different cache capacities and $P_m = 40$ dBm.

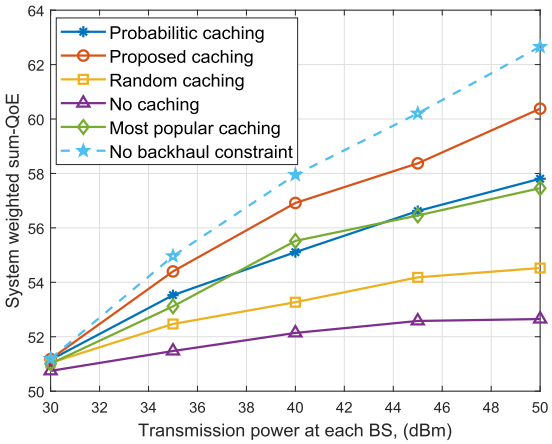


Fig. 7. The effect of cache placement on the system weighted sum-QoE with different transmission powers and $S_m = 10$.

the lack of content diversity leads to low cache hit ratio, which finally brings huge burden on the backhaul link, especially when the file popularity is equal.

- *The probabilistic cache placement:* Each BS caches a video content randomly with probability depending on Zipf distribution. The more popular the content is, the more likely it will be cached in each BS.
- *The random cache placement:* Each BS randomly selects cached media files with equal probability and a file can only be cached once. The file popularity is not considered in this cache placement.
- *The cache placement with no backhaul constraint:* The backhaul capacity constraint (9c) is excluded in the caching problem $\mathcal{P}_C$. It also means that the backhaul capacity of each BS goes to infinity, or the storage of each BS is large enough to cache all files in the media server. Thus, this cache placement is an upper bound of the QoE performance.

In Fig. 7, the effect of cache placement on the system weighted sum-QoE is shown with the variation of the transmission power. The cache capacity for each BS is set as $S_m = 10$. In Fig. 7, with the improvement of the transmission power, the QoE performances of all cache placement strategies are increasing continuously. In addition, the cache placement with no backhaul constraint and the no caching policy respectively serve as an upper bound and a lower bound in term of the QoE performance. And our proposed cache placement greatly outperforms the most popular cache placement, the probabilistic cache placement and the random cache placement. This is because that, our proposed cache placement is aware of both the file popularity and the transmission strategy. We can conclude from Fig. 7 that, the QoE performance in the backhaul capacity limited system can be extremely enhanced with a reasonable caching policy.

Fig. 8 shows the impact of cache placement on the system QoE performance for different cache capacities with $P_m = 40$ dBm. As expected, the QoE performance of all cache placement strategies increases with the expanding of the cache capacity except for the no caching policy, and our proposed cache placement can achieve a large performance gain as compared with others. More importantly, as the cache capacity approaches to the total number of files in the media server, the performance gap gradually becomes narrow and the performance of all cache placement strategies is extremely close to the upper bound performance.

The impact of the number of users on the QoE performance is shown in Fig. 9. As expected, with the increase of the number of users, the system QoE performance is improved. Then, the impact of the number of video contents on the QoE performance is shown in Fig. 10. When the number of video contents increases, the system QoE performance is gradually decreasing. This is because that, the user request is becoming dispersive with the increased number of video contents, which makes users hard to form multicast transmission and thus leads to lower QoE performance.

C. The Effect of Transmission Strategies on the System Performance

We investigate the influence of transmission strategies on the system performance. The cache placement in this subsection is decided by our proposed caching policy. For comparison, two transmission strategies are considered as baselines.

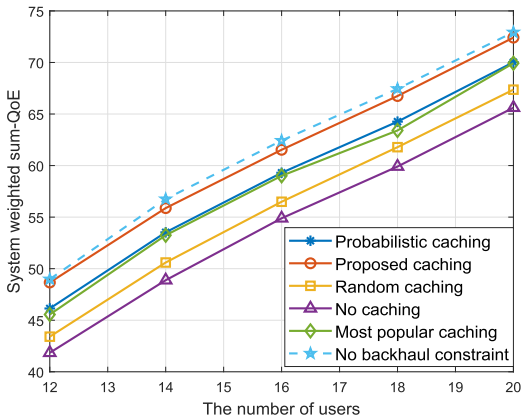


Fig. 9. The impact of cache placement on the system weighted sum-QoE with different number of users.

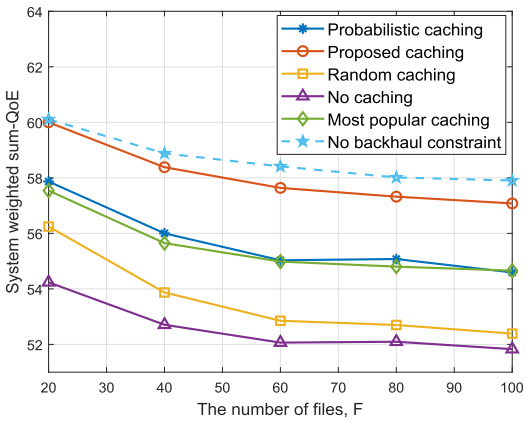


Fig. 10. The impact of cache placement on the system weighted sum-QoE with different number of video contents.

- *The unicast strategy:* Different users are independently served by different beamformers regardless of their requested media files.
- *The rate maximization strategy:* The aim of this strategy is to maximize the system weighted sum-rate performance. In specific, the rate maximization transmission problem is a variant of problem $\mathcal{P}_T$, where the weighted sum-QoE objective function is replaced by the weighted sum-rate function. For simplicity, the weighting factor for each user is also assumed to be 1.

To verify the effectiveness of our proposed multicast transmission, Fig. 11 shows the system QoE performance of the multicast strategy and the unicast strategy for both $k = 15$ and $k = 20$ cases. It can be seen that, the proposed multicast strategy is superior than the unicast strategy, especially when the number of users becomes large. This is because that, the multicast strategy can exploit the file popularity among different users and the broadcast nature of the wireless channel. Furthermore, a group of multiple users is served by the same beamformer, which also fully utilizes the spatial dimension to boost the system performance. With the increase of the number of users, different users have more opportunity to request the same content and form a multicast group. Thus, the performance gain becomes large.

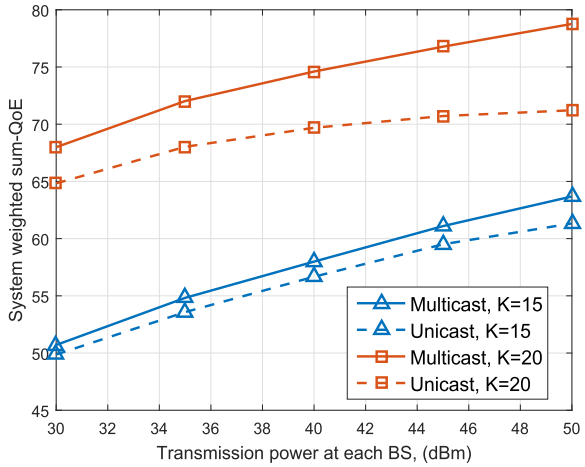


Fig. 11. The impact of transmission strategies (multicast and unicast) on the system weighted QoE with $S_m = 10$.

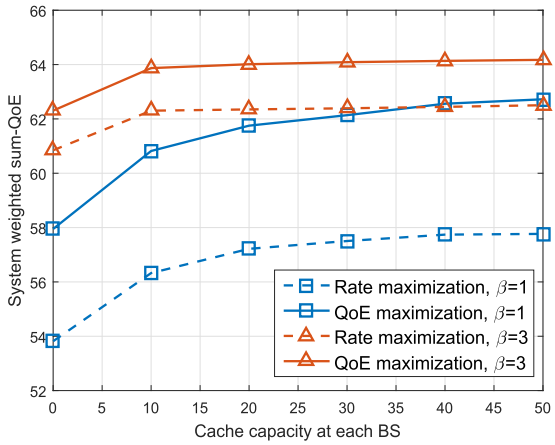


Fig. 12. The impact of transmission strategies (QoE maximization and Rate maximization) on the system weighted QoE with $P_m = 40$ dBm.

Eventually, the performance of our proposed QoE maximization strategy and the rate maximization strategy is shown in Fig. 12 with $P_m = 40$ dBm. For the user request model, Zipf distributions with $\beta = 1$ and $\beta = 3$ are considered. It is easy to find that our proposed QoE maximization strategy achieves much better QoE performance than that of the rate maximization strategy with both $\beta = 1$ and $\beta = 3$. This is mainly because the logarithmic law of the QoE function. Specifically, for the rate maximization strategy, more resources are allocated to users with better channel conditions, but the increase of these users' QoE is slower. Nevertheless, the QoE performance of worse users degrades severely. In addition, the QoE performance gap between these two strategies is increasing when the storage size becomes large. The reason is that, with the increase of the storage size, the above mentioned phenomenon is more serious. Furthermore, it can also be seen from Fig. 12 that, a larger skewness parameter β of Zipf distribution results in a better system sum-QoE performance. This is because with larger β, the user file requests become more concentrated, which brings a higher chance for the multicast transmission.

VII. Conclusion

In this paper, we have designed QoE-driven long-term transmission-aware cache placement and short-term transmission strategy for video services in a cloud-RAN network, where multiple users are cooperatively served by multiple BSs via the multicast transmission. The weighted sum-QoE has been maximized, taking into account the backhaul capacity constraint, the transmission power constraint and the storage capacity of each BS. In the caching stage, the cache placement has been optimized based on the long-term system information, and both the centralized and distributed cache placement strategies have been proposed. In the delivery stage, with the optimized cache placement in the previous stage, the multicast beamforming has been designed. Simulation results have demonstrated that both the cache placement and transmission strategies have significant effect on the QoE, and the proposed transmission-aware cache placement and transmission strategy can outperform others in terms of the QoE. In the future, we will consider the user mobility in wireless caching networks when designing the cache placement strategy.

References

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

[2] S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. Iyengar, "Multimedia big data analytics: A survey," *ACM Comput. Surv.*, vol. 51, no. 1, Apr. 2018, Art. no. 10.

[3] I. Cisco Visual Networking, "Global mobile data traffic forecast update, 2017–2022 White Paper," Feb. 2019. [Online]. Available: http://goo.gl/ylTuVx

[4] N. Cheng *et al.*, "Big data driven vehicular networks," *IEEE Netw.*, vol. 32, no. 6, pp. 160–167, Dec. 2018.

[5] L. Lyu, C. Chen, Z. Shanying, and X. Guan, "5G enabled co-design of energy-efficient transmission and estimation for industrial IoT systems," *IEEE Trans. Ind. Inform.*, vol. 14, no. 6, pp. 2690–2704, Jun. 2018.

[6] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[7] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[8] D. Wang, Y. Wang, R. Sun, and X. Zhang, "Robust C-RAN precoder design for wireless fronthaul with imperfect channel state information," in *Proc. IEEE Wireless Commun. Netw. Conf.*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[9] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[10] M. Patel *et al.*, "Mobile-edge computing introductory technical white paper," White Paper, Mobile-edge ComputingIndustry Initiative, Sep. 2014.

[11] S. Zhang, N. Zhang, P. Yang, and X. Shen, "Cost-effective cache deployment in mobile heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 264–11 276, Dec. 2017.

[12] 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.

[13] J. Gong, S. Zhou, Z. Zhou, and Z. Niu, "Policy optimization for content push via energy harvesting small cells in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 717–729, Feb. 2017.

[14] X. Li, X. Wang, and V. C. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *Proc. IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[15] 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.

[16] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching under heterogeneous file preferences," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 444–457, Jan. 2017.

[17] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2699–2714, Jun. 2017.

[18] 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.

[19] W. Wu, N. Zhang, N. Cheng, Y. Tang, K. Aldubaikhy, and X. Shen, "Beef up mmWave dense cellular networks with D2D-assisted cooperative edge caching," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3890–3904, Apr. 2019, doi 10.1109/TVT.2019.2896906.

[20] A. Liu and V. K. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[21] S.-H. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

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

[23] L. Xiang, D. W. K. Ng, R. Schober, and V. W. Wong, "Cache-enabled physical layer security for video streaming in Backhaul-limited cellular networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 736–751, Feb. 2018.

[24] W. Zhang, Y. Wen, Z. Chen, and A. Khisti, "QoE-driven cache management for http adaptive bit rate streaming over wireless networks," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1431–1445, Oct. 2013.

[25] W. Huang, L. Ding, D. Meng, J.-N. Hwang, Y. Xu, and W. Zhang, "QoE-based resource allocation for heterogeneous multi-radio communication in software-defined vehicle networks," *IEEE Access*, vol. 6, pp. 3387–3399, 2018.

[26] Y. Wang *et al.*, "A data-driven architecture for personalized QoE management in 5G wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 102–110, Feb. 2017.

[27] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *Proc. IEEE Int. Conf. Comput. Commun.*, New York, NY, USA, Mar. 1999, pp. 126–134.

[28] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. New York, NY, USA: Academic, 2013.

[29] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment," in *Proc. IEEE Int. Conf. Commun.*, May 2010, pp. 1–5.

[30] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *Math. Program.*, vol. 157, no. 2, pp. 515–545, Jun. 2016.

[31] R. Sun, H. Baligh, and Z.-Q. Luo, "Long-term transmit point association for coordinated multipoint transmission by stochastic optimization," in *Proc. IEEE 14th Workshop Signal Process. Adv. Wireless Commun.*, Darmstadt, Germany, Jun. 2013, pp. 330–334.

[32] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Math. Program.*, vol. 39, no. 2, pp. 117–129, Jun. 1987.

[33] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, Oct. 2010.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[35] A. Hjørungnes, *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[36] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined convex programming, version 1.22," Aug. 2012. [Online]. Available: http://cvxr.com/cvx

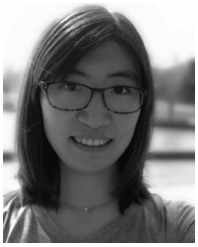
[37] Y. Ye, *Interior Point Algorithms: Theory and Analysis (Interscience Series in Discrete Mathematics and Optimization)*. New York, NY, USA: Wiley, 1997.

[38] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

Ruijin Sun (S'16) received the B.S. degree from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2013, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. She is currently working as a joint Postdoctoral Fellow with Peng Cheng Laboratory and Tsinghua University, Beijing, China. She was a Visiting Student with the University of Waterloo, Canada (September 2017–September 2018). Her research interests are in the area of MIMO and wireless caching networks.

Ying Wang (M'03) received the Ph.D. degree in circuits and systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2003. She is currently a Professor with BUPT and the Director of the Radio Resource Management Laboratory, Wireless Technology Innovation Institute, BUPT. Her research interests are in the area of the cooperative systems and radio resource management in wireless communications. She has authored more than 100 papers in international journals and conferences proceedings. She is an active in standardization activities of 3GPP and ITU. She took part in performance evaluation work of the Chinese Evaluation Group, as a Representative of BUPT. She was the recipient of first prizes of the Scientific and Technological Progress Award by the China Institute of Communications in 2006 and 2009, respectively, and a 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.

Nan Cheng (S'12–M'16) received the B.E. and the M.S. degrees from Tongji University, Shanghai, China, in 2009 and 2012, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2016. He is currently a Professor with the School of Telecommunication Engineering, and with State Key Lab of ISN, Xidian University, Shaanxi, China. He worked as a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, from 2017 to 2018. His current research focuses on space-air-ground integrated system, big data in vehicular networks, and self-driving system. His research interests also include performance analysis, MAC, opportunistic communication, and application of AI for vehicular networks. He is also interested in space-air-ground integrated networks.



Ling Lyu (S'16) received the B.S. degree in telecommunication engineering from Jinlin University, Changchun, China, in 2013; and the Ph.D. degree in control theory and control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. She joined Dalian Maritime University, China, in 2019, where she is currently an Associate Professor with the Department of Telecommunication. She was a Visiting Student with the University of Waterloo, Canada (September 2017–September 2018). Her current research interests include wireless sensor and actuator network and application in industrial automation, the joint design of communication and control in industrial cyber-physical systems, estimation and control over lossy wireless networks, machine type communication enabled reliable transmission in the fifth-generation network, resource allocation, energy efficiency.

Shan Zhang (S'13–M'16) received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. She is currently an Assistant Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. She was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, from 2016 to 2017. Her research interests include mobile edge computing, wireless network virtualization and intelligent management. She was the recipient of the Best Paper Award at the Asia-Pacific Conference on Communication in 2013.

Haibo Zhou (M'14–SM'18) received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. From 2014 to 2017, he was a Postdoctoral Fellow with the Broadband Communications Research Group, Department of Electrical and Computer Engineering, University of Waterloo. He is currently an Associate Professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include resource management and protocol design in cognitive networks.

Xuemin (Sherman) Shen (M'97–SM'02–F'09) received the B.Sc. degree from Dalian Maritime University, Dalian, China, in 1982 and the M.Sc. and Ph.D. degrees from Rutgers University, New Brunswick, NJ, USA, in 1987 and 1990, respectively, all in electrical engineering. He is a University Professor and an Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research focuses on resource management, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He was as the Technical Program Committee Chair/Co- Chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom'07, the Symposia Chair for IEEE ICC'10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC'08, the General Co-Chair for ACM Mobihoc'15, Chinacom'07 and the Chair for IEEE Communications Society Technical Committee on Wireless Communications. He also serves/served as an Editor-in-Chief for the IEEE INTERNET OF THINGS JOURNAL, IEEE NETWORK, *Peer-to-Peer Networking and Application*, and *IET Communications*; a Founding Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS; an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *Computer Networks*, and *ACM/Wireless Networks*, etc., and the Guest Editor for the IEEE JSAC, IEEE WIRELESS COMMUNICATIONS, and IEEE COMMUNICATIONS MAGAZINE, etc. He was the recipient of the Excellent Graduate Supervision Award in 2006, and the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.