This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

Cooperative Edge Caching in User-Centric Clustered Mobile Networks

Shan Zhang, *Member, IEEE,* Peter He, *Member, IEEE,* Katsuya Suto, *Member, IEEE,* Peng Yang, *Student Member, IEEE,* Lian Zhao, *Senior Member, IEEE,* and Xuemin (Sherman) Shen, *Fellow, IEEE*

Abstract—With files proactively stored at base stations (BSs), mobile edge caching enables direct content delivery without remote file fetching, which can reduce the end-to-end delay while relieving backhaul pressure. To effectively utilize the limited cache size in practice, cooperative caching can be leveraged to exploit caching diversity, by allowing users served by multiple base stations under the emerging user-centric network architecture. This paper explores delay-optimal cooperative edge caching in large-scale user-centric mobile networks, where the content placement and cluster size are optimized based on the stochastic information of network topology, traffic distribution, channel quality, and file popularity. Specifically, a greedy content placement algorithm is proposed based on the optimal bandwidth allocation, which can achieve $(1-1/e)$-optimality with linear computational complexity. In addition, the optimal user-centric cluster size is studied, and a condition constraining the maximal cluster size is presented in explicit form, which reflects the tradeoff between caching diversity and spectrum efficiency. Extensive simulations are conducted for analysis validation and performance evaluation. Numerical results demonstrate that the proposed greedy content placement algorithm can reduce the average file transmission delay up to 45% compared with the non-cooperative and hit-ratio-maximal schemes. Furthermore, the optimal clustering is also discussed considering the influences of different system parameters.

Index Terms—Mobile edge caching, cooperative coded caching, user-centric networks, content placement

1 Introduction

To accommodate the ever increasing mobile traffic demand, small cell base stations (SBSs) are expected to be ultra-densely deployed for extensive spatial spectrum reuse in the next generation (5G) networks and beyond [2]. As networks are further densified, deploying ideal backhaul for each SBS becomes impractical due to the high cost, leading to possible backhaul congestions and performance degradation [3], [4]. To relieve the backhaul pressure, mobile edge caching has been proposed to store contents at the edge of networks (e.g., SBSs or end devices) in addition to remote servers, whereby contents can be directly delivered through wireless transmission without backhaul or core network transmissions [5], [6], [7], [8], [9]. By exploiting the similarity of requested contents, mobile edge caching has the potential to reduce backhaul capacity requirement to 35% [10]. In practical systems, the performance of mobile edge caching can be constrained by the limited cache size [11], [12]. A straightforward solution is to design more effective content placement schemes, by exploring the information of content popularity and user preferences [13]. Furthermore, cooperative caching can be leveraged to enlarge the set of cached files, which enables users served by multiple SBSs to exploit caching diversity in space [14], [15]. In fact, mobile networks are now evolving from the conventional cellular topology to a de-cellular user-centric structure, where each user can be served by a dynamically formed cluster of SBSs for quality of experience (QoE) enhancement [16], [17], [18], [19]. In this case, different SBSs can store diverse contents in a cooperative manner, such that users have higher probability to obtain the requested contents locally from the caches of clustered SBSs [20]. However, cooperative caching may sacrifice spectrum efficiency due to the enlarged transmission distance. Specifically, increasing the cluster size improves the content hit ratio, but users are served by farther SBSs with higher path loss, degrading spectrum efficiency. The tradeoff relationship between caching diversity and spectrum efficiency in terms of the cluster size can bring significant challenges for edge caching, which has not been well studied in literature [21].

This paper investigates delay-optimal cooperative edge caching in large-scale user-centric clustered mobile networks considering the constrained cache size and radio resources, based on the stochastic information of network topology, traffic distribution, channel quality, and file popularity. Specifically, we focus on two fundamental problems: 1) content placement, and 2) SBS clustering. Coded caching is adopted, whereby each user can fetch coded segments for decoding from the caches of the candidate SBSs in cluster. For the content placement design, an optimization problem is formulated to determine the ratio of cached segments for different files,

- *Shan Zhang, Peter He, Katsuya Suto and Xuemin (Sherman) Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1.*
  E-mail: {s327zhan, z85he, ksuto, sshen}@uwaterloo.ca
- *Peng Yang is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China.*
  E-mail: yangpeng@hust.edu.cn
- *Lian Zhao is with the Department of Electrical and Computer Engineering, Ryerson University, Ontario, Canada, M5B 2K3.*
  E-mail: l5zhao@ryerson.ca

Part of this work has been accepted by IEEE GLOBECOM 2017 [1].

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

aiming at minimizing the average file transmission delay. In addition, the bandwidth allocation is jointly optimized for load balancing, since different content placement results in different traffic distributions. The formulated problem is challenging for following reasons. Firstly, the accurate form of average file transmission delay cannot be derived due to the multi-dimensional randomness of network topology, traffic distribution, and channel quality. Secondly, the problem is a NP-hard mixed integer programming problem, as content placement and bandwidth allocation determines the average file transmission rate in a coupled manner. Thirdly, the average file transmission rate has a piecewise structure with respect to content placement, introducing additional complexity for delay performance analysis. By applying the theory of stochastic geometry, we obtain the average transmission rate with conservative approximation, whereby two subproblems need to be addressed, i.e., bandwidth allocation and content placement. The bandwidth allocation subproblem is convex and thus is solved by employing Lagrange multiplier method. To address the non-convex content placement subproblem, we prove the piecewise objective function has monotone submodular property, and thus propose a greedy algorithm which can achieve $(1-\frac{1}{e})$-optimality with linear complexity. For the SBS clustering, an explicit condition constraining the maximal cluster size is obtained, which reveals the tradeoff between content diversity and spectrum efficiency with respect to system parameters (including SBS density, backhaul delay and content popularity). Extensive simulations are conducted to validate the analysis as well as evaluate the performance of proposed algorithm, based on both real-world YouTube data trace and the classic Zipf content popularity distribution. Two typical caching schemes, non-cooperative and hit-ratio-maximal, are adopted to demonstrate the tradeoff between content diversity and spectrum efficiency in user-centric clustered caching. Simulation results show that the non-cooperative scheme guarantees highest spectrum efficiency with lowest content ratio, which performs well in case of low network density or large cache size; whereas the hit-ratio-maximal caching scheme provides maximal hit rate but sacrifices the spectrum efficiency, which is more advantageous in dense networks with congested backhaul and small cache size. The proposed greedy caching scheme always outperforms the other two schemes, by adjusting content placement and cluster size to balance content diversity and spectrum efficiency. Furthermore, the influences of important system parameters are also studied, including cache size, SBS density, backhaul delay.

The main contributions of this work are as follows.

1) The content placement, SBS clustering, and bandwidth allocation have been jointly optimized in large-scale user-centric mobile networks considering the radio resource constraints, based on the stochastic information of network topology, traffic distribution, channel quality, and file popularity.
2) The average file transmission delay is derived with conservative approximation, and proved to have monotone submodular property with respect to content placement. Accordingly, a low-complexity greedy content placement algorithm is proposed with $(1-1/e)$-optimality performance guarantee.
3) The bandwidth allocation is optimized to minimize average file transmission delay, which can match radio resource to the traffic load distribution resulting from content placement.
4) The delay-optimal SBS clustering for cooperative caching is obtained, and the influences of important system parameter are analyzed.
5) Both analytical and numerical results reveal the underneath tradeoff relationship between content caching diversity and spectrum efficiency in cooperative caching, which can provide insightful guidelines to dynamic SBS clustering and content placement with the variations of traffic loads and operational scenarios.

The remainder is organized as follows. Section 2 reviews related work on cooperative caching, and Section 3 introduces the system model and problem formulation. The average file transmission delay is analyzed in Section 4, based on which the optimal bandwidth allocation is derived. In Section 5, a greedy content placement algorithm is proposed. Finally, Section 6 shows simulation results, and Section 7 concludes the paper.

2 Literature Review

Although the content placement problem has been extensively studied in wired networks, the design of mobile edge caching is relatively underdevelopped, due to the features of user mobility, link connectivity and channel quality [14], [22]. When each BS provides service independently, the popularity-based cache placement scheme (i.e., each BS stores the most popular contents) has been widely adopted to maximize the content hit ratio [23]. The cooperation among BSs can further enhance caching efficiency, which is also more challenging since the caching decision of one BS can be influenced by neighboring BSs [24]. Existing works on cooperative mobile edge caching can be classified into two categories based on the utilized system information, i.e., complete priori information and stochastic information.

With complete priori information of user connectivity and channel quality, both centralized and distributed cooperative content placement schemes have been designed [25], [26], [27], [28]. Shanmugam *et al.* have investigated the centralized cooperative content placement problem to minimize the average file downloading delay in a network consisting of one MBS and several cache-enabled SBSs, considering both coded and uncoded caching [25]. Li *et al.* have proposed a distributed belief propagation algorithm for content placement, by applying the factor graph to describe network topology [26]. Tran *et al.* have further combined the edge-based BS caching with core-based cloud caching in the Cloud-based Radio Access Networks (C-RANs), and designed a hierarchical content placement scheme to reduce file fetching delay and backhaul pressure [27]. The cooperative caching among BSs and users has been studied in [28], which can be

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

formulated as an integer-linear programming problem and solved by the hierarchical primal-dual decomposition method. In addition to popularity-based caching, other works have also exploited the user mobility information for proactive content fetching and cache update, to provide seamless handover experience [29]. Furthermore, caching schemes jointly consider content popularity and user mobility have been also designed [30]. By utilizing the complete information of network status, these algorithms can significantly improve network performance in terms of transmission delay, content hit rate, backhaul load, and user QoE. However, the limitation is that the cached contents need to be updated frequently with the variations of user requests or channel conditions, which may introduce considerable overhead and backhaul loads in practice. Furthermore, these algorithms are usually designed on a case-by-case basis for small-scale networks, which cannot provide general design guidelines for practical networks.

Cooperative caching has also been studied based on stochastic network information. By modeling the cache-enabled BSs as a homogeneous Poisson Point Process (PPP), the caching probabilities of different files have been optimized to maximize the content delivery success probability, where each user is served by the nearest BS with requested contents in cache [31]. Our previous work has shown the tradeoff between content diversity and spectrum efficiency in cooperative caching, when the user can be steered to the second nearest SBSs (i.e., user-centric cluster size set to 2) [1]. For multi-tier heterogeneous networks, an iterative heuristic content placement algorithm has been proposed, where each network tier makes decisions based on the contents cached at other tiers [32]. Furthermore, the long-term cache instance deployment problem has been studied to determine the optimal cache size of SBSs and MBSs for the given storage resource budget, where the SBS-tier cache the most popular contents and the MBS-tier cache the less popular ones for load balance [33]. Very recent works have explored the cooperative caching in user-centric clustered networks [20], [21], where insightful analytical results have been obtained on the tradeoff between caching diversity and channel diversity. However, [20], [21] only considered the single user case and ignored the constraints of radio resources. In fact, radio resource management has great impact on the performance of edge caching, which should be jointly optimized to further enhance system performance [34], [35], [36].

This paper investigates the joint optimization of content placement, SBS clustering, and bandwidth allocation in large-scale user-centric mobile networks, based on the stochastic network information. The novelty is three-fold. Firstly, by jointly optimizing bandwidth allocation, we conduct network-level cooperative caching design with the constraint of radio resources taken into consideration, which has been ignored in the existing clustered cooperative caching studies focusing on user-aspect performance [20], [21]. Secondly, through the design of content placement and cluster size, we reveal the underneath tradeoff relationship between content diversity and spectrum efficiency in an analytical way, which has not been well investigated in existing literature. Thirdly, both the analytical and numerical results demonstrate that the cluster size should be adjusted based on the traffic load and network status to balance the content diversity and spectrum efficiency, whereas most existing works have adopted the constant cluster size only based on the received signal strength.

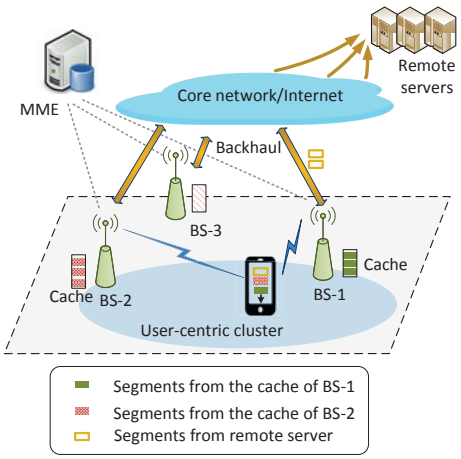


Fig. 1: Mobile edge caching with user-centric clustered services.

3 System Model and Formulation

We consider a homogeneous mobile network with edge caching, where contents can be partially or completely stored at each SBS after being coded into segments, as shown in Fig. 1. The contents are characterized by popularity based on the hit rate, corresponding to diverse mobile services such as videos streaming, HD map, social media, news, software update. When a user raises a content request, it can be served by a cluster of candidate SBSs, depending on the content caching states. If the requested content is cached, the user fetches the coded segments directly from candidate SBSs in ascending order of transmission distance, until the obtained segments are sufficient for file decoding. In addition, if the obtained segments from the caches of all candidate SBSs are insufficient for decoding, the nearest SBS will fetch the remaining ones from remote servers through backhaul, and then deliver to the user through wireless transmission. If the requested content is not stored in cache, the nearest SBS will fetch the whole content from remote servers. Therefore, the transmission delay depends on the content placement. The main notations are listed in Table 1.

3.1 Cooperative Content Caching

The distributions of SBSs and end users are modeled as independent PPPs of densities ρ and λ, respectively, for tractable analysis [37]. Denote by $\mathcal{F} = \{1, 2, \cdots, f, \cdots, F\}$ the file library, where F is the total number of files. Denote by $\mathcal{Q} = \{q_1, q_2, \cdots, q_f, \cdots, q_F\}$ the file popularity distribution, where $q_f > 0$ is the probability that the requested content is file-f ($\sum_{f=1}^{F} q_f = 1$). We exploit rateless fountain coding for cooperative caching [38]. Each

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

TABLE 1: Notation table

Notation	Definition/Description
ρ	Density of SBSs
λ	Density of active users
F	The number of files in the file library
f	File index, $f = 1, \cdots, F$
q_f	File popularity of file-f
L	Length of each coded
s_f	The number of coded segments of file-f
c_f	The number of coded segments of file-f cached
C	Cache size of each SBS
K	SBS cluster size
B_k	The kth nearest SBS, $k = 1, 2, \cdots, K$
$P_{k,f}$	Downloading ratio at B_k if file-f is requested
Ω_k	Average ratio of traffic served by B_k
W	Available system bandwidth
φ_k	Ratio of bandwidth allocated to Group-k users
N_k	Number of Group-k users served by a SBS
ζ	Spectrum efficiency of Group-k users
R_k	Transmission rate of Group-k users
P_{T}	Transmit power of SBSs
d_k	Transmission distance of Group-k users
α	Path loss coefficient
σ^2	Addictive Gaussian noise power density
I_k	Inter-cell interference received by Group-k users
$\bar{D}$	Average transmission delay

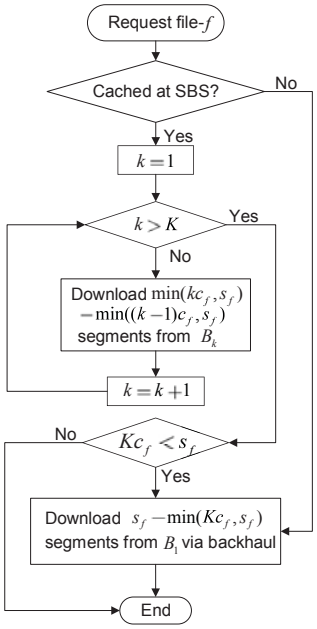


Fig. 2: Service process with user-centric clusters.

file is encoded by fountain coding into independent segments, which are equally deployed to all SBSs, i.e., each SBS holds c_f encoded segments of file-f[1]. Users can decode file-f by collecting s_f encoded segments from SBSs. The coded segments are considered to have the same size of L in bits, and thus $s_f L$ reflects the size of file-f in bits. Due to the limited cache size, each SBS can store C coded segments at most, i.e., $\sum_{f=1}^{F} c_f \leq C$.

Consider a typical user u, served by K SBSs which have the shortest transmission distance. As the SBSs are homogeneous with the same system parameters like transmit power, the distance-based SBS clustering can provide highest transmission rate on average. In practical systems, the transmission distance can be estimated by the user-side channel measurement, and the information is reported to the Mobility Management Entity (MME) which makes decisions on user association and service. Denote by $\mathcal{B}_u = \{B_1, B_2, \cdots, B_k, \cdots, B_K\}$ the set of candidate SBSs, which are sorted according to transmission distance in ascending order without loss of generality. Suppose user-u requests file-f, and it will fetch the cached segments from SBS-B_1 to SBS-B_K successively, until the number of obtained segments reaches s_f. If $Kc_f < s_f$, SBS-B_1 will fetch the remaining $(s_f - Kc_f)$ segments from remote servers via backhaul, which does not break cell load balancing. In addition, such service scheme can minimize the average transmission delay, without the instant information of channel status or cell load. User-u can get $[\min(kc_f, s_f) - \min((k-1)c_f, s_f)]$ segments from SBS-B_k, and SBS-1 needs to fetch $[s_f - \min(Kc_f, s_f)]$ segments from remote servers. The service flow chart is given as Fig. 2. When the cluster size is $K = 2$, Fig. 3 illustrates the detailed service process with different cache ratio, corresponding to Fig. 1. Notice that the average downloading delay is lower at SBS-B_1 compared with that at SBS-B_2, which will be analyzed in details later. In addition, the delay will further increase when fetching the remaining segments from remote servers, due to the backhaul transmission. Thus, increasing c_f helps to reduce average transmission delay of file-f, as shown in Fig. 3. As a results, the content placement should be well designed to minimize delay, under the constraint of cache size $\sum_{f=1}^{F} c_f \leq C$.

TABLE 2: Example of user-centric cluster

	user-1	user-2	user-3	user-4	user-5
SBS-1	B_1	B_2	B_1	$\times$	$\times$
SBS-2	B_2	$\times$	$\times$	B_1	B_2
SBS-3	$\times$	B_1	B_2	B_2	B_1

Denote by $P_{k,f}$ the ratio of segments that user-u can get from the cache of SBS-B_k, given by

$$P_{k,f} = \frac{1}{s_f}\left[\min(kc_f, s_f) - \min((k-1)c_f, s_f)\right]. \quad (1)$$

For notation simplicity, denote by $P_{K+1,f}$ the ratio of remaining contents fetched from remote servers:

$$P_{K+1,f} \triangleq 1 - \sum_{k=1}^{K} P_{k,f} = 1 - \min(Kc_f, s_f)\frac{1}{s_f}. \quad (2)$$

Therefore, the average content hit ratio at the kth candidate SBS is given by

$$\Omega_k = \sum_{f=1}^{F} q_f P_{k,f}, \quad (3)$$

where $k = 1, 2, \cdots, K$. Denote by Ω_{K+1} the average content miss ratio: $\Omega_{K+1} = 1 - \sum_{k=1}^{K} \Omega_k$.

For a typical SBS, it can hold different rankings to the served users, varying from the 1st to the $(K+1)$th. Thus, from the perspective of SBSs, the served users can be classified into $(K+1)$ groups, based on the ranking of the

1. The identical file caching at SBSs is adopted since the SBSs are homogeneous with the same system parameters (transmit power, bandwidth, cache size and backhaul capacity).

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

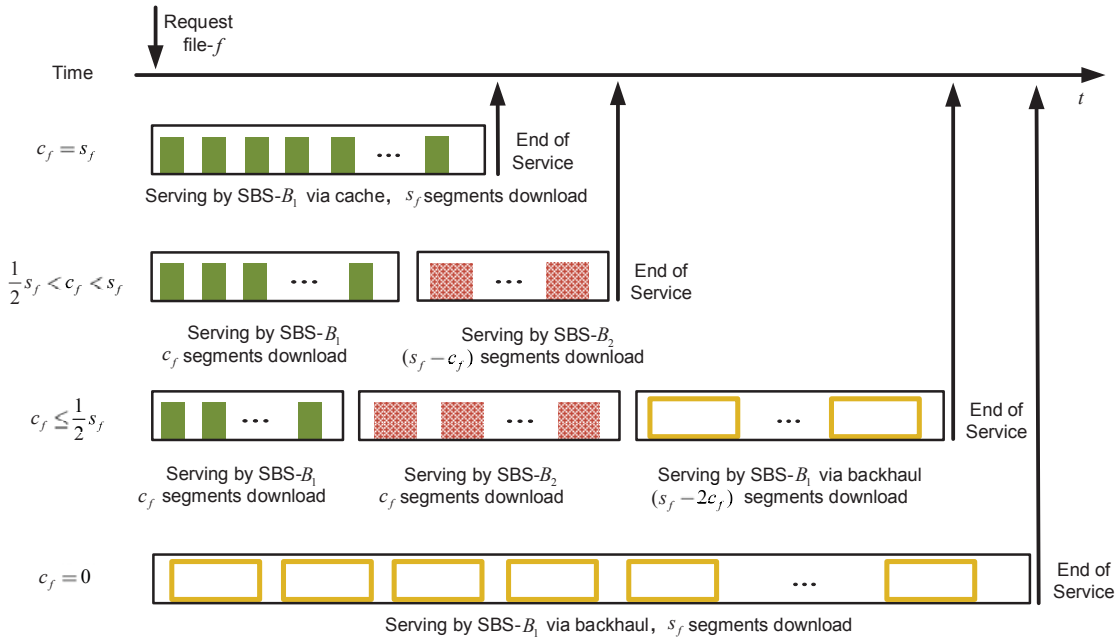


Fig. 3: Segment downloading illustration when cluster size is 2.

serving SBSs. From the perspective of an individual user, it can belong to different groups at different time, since it fetches segments from different SBSs successively. User-u belongs to Group-k when it is fetching file from SBS B_k, i.e., the kth candidate SBS. Table 2 gives an example of network topology when there are 3 SBSs serving 5 users with cluster size of 2, where B_k means the SBS is the kth nearest SBS for the corresponding user ($\times$ means the SBS is not connected). For illustration, SBS-2 is the second nearest SBS for user-1 and user-5, and is the nearest SBS for users-4. As a result, SBS-2 can serve three users, user-4 belongs to the first user group when downloading cached segments, user-1 and user-5 belong to the second user group, while user-4 belongs to the third group when downloading remaining segments via the backhaul.

The main difference between the groups is the wireless transmission rate. Specifically, the transmission distance increases with the group index k, leading to higher path loss and lower transmission rate on average. According to Eq. (3), Ω_k also denotes the probability that a user is being served by the kth candidate SBS, i.e., belonging to Group-k. As the user distribution follows a PPP of density λ, the distribution of Group-k users also follows a PPP of density $\Omega_k\lambda$. Define $\{\Omega_k\}$ as the *group load distribution*, to denote the traffic load distribution among different groups, which is a key factor influencing the average file transmission delay.

3.2 Wireless Transmission

The transmission rate should be guaranteed for successful file delivery. When user-u is served by SBS-B_k, the file transmission rate is given by

$$R_k = w_{u,k}\log_2\left(1+\zeta_k\right), \tag{4}$$

where $w_{u,k}$ is the allocated bandwidth to user-u and ζ_k is the received signal to interference and noise ratio (SINR)[2]:

$$\zeta_k = \frac{P_{\mathrm{T}}{d_k}^{-\alpha}}{\sigma^2 + I_k}, \tag{5}$$

P_{T} is the SBS transmit power density, d_k is the distance from user-u to SBS-B_k, α is the path loss exponent, σ^2 is the additive Gaussian noise power density, and I_k denotes the inter-cell interference when served by SBS-B_k. With the advanced interference mitigation techniques (like spectrum reuse, power control, and interference alignment), the inter-cell interference can be treated as constant noise [39]. Then, I_k can be interpreted as the efficiency of interference mitigation. For example, $I_k = 0$ if the interference is canceled completely, which usually requires perfect channel information. Larger I_k means less effective interference mitigation, which can happen with incomplete channel information, lower received signal strength, extensive spatial spectrum use [40].

Denote by W the total available system bandwidth, and ϕ_k the ratio of bandwidth allocated to Group-k users, where $\phi_k \ge 0$ and $\sum_{k=1}^{K+1}\phi_k \le 1$. When user-$u$ is being served by SBS B_k, it belongs to Group-k and the corresponding transmission rate is given by

$$R_k = \frac{\phi_k W}{N_k}\log_2\left(1+\zeta_k\right), \tag{6}$$

where N_k is the number of Group-k users being served by SBS-B_k. Due to the random network topology and user location, N_k and ζ_k are random variables. Thus, the average file transmission delay is given by

$$\bar{D} = \sum_{k=1}^{K}\frac{\Omega_k\bar{S}L}{\mathop{\mathbb{E}}_{\{N_k,d_k\}}[R_k]} + \left(\frac{\bar{S}L}{\mathop{\mathbb{E}}_{\{N_{K+1},d_{K+1}\}}[R_k]} + \bar{D}_{\mathrm{BH}}\right)\Omega_{K+1}, \tag{7}$$

2. $\zeta_{K+1} = \zeta_1$.

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

where $\mathop{\mathbb{E}}_{\{N_k,d_k\}}[R_k]$ denotes the expected transmission rate of Group-k users in respect to the random variables N_k and d_k, $\bar{D}_{\mathrm{BH}}$ is the average delay of backhaul file fetching, and $\bar{S}$ is the average length of requested files, i.e., $\bar{S} = \sum_{f=1}^{F} q_f s_f$. The probability distribution of N_k depends on the group load distribution $\{\Omega_k\}$, which varies with content placement $\{c_f\}$ according to Eq. (3). Accordingly, the bandwidth allocation $\{\phi_k\}$ should be adjusted to Ω_k for load balancing and effective resource utilization, based on the content placement. Therefore, we jointly optimize the content placement and bandwidth allocation, to minimize the average file transmission delay.

3.3 Problem Formulation

The delay-optimal cooperative caching problem can be formulated as follows:

$$\min_{\{c_f\},\{\phi_k\}} \quad \bar{D}, \tag{8a}$$

$$\text{(P1)} \quad \text{s.t.} \quad \sum_{f=1}^{F} c_f \le C, \tag{8b}$$

$$\sum_{k=1}^{K+1} \phi_k \le 1, \tag{8c}$$

$$c_f \in \{0,1,\cdots,s_f\}, \ \ \forall f \in \mathcal{F}, \tag{8d}$$

$$\phi_k \ge 0, \ \ k = 1,2,\cdots,K+1, \tag{8e}$$

where (8b) reflects the limitation of cache capacity and (8c) is due to the constrained system bandwidth. The optimization of content placement $\{c_f\}$ should consider the tradeoff between content diversity and spectrum efficiency. Increasing content diversity can improve content hit ratio, and the average backhaul transmission delay can be reduced as more users can get requested contents directly from the candidate SBSs. However, the spectrum efficiency degrades due to the increased transmission distance, introducing higher delay of wireless transmission. The optimization of bandwidth allocation (i.e., $\{\phi_k\}$) is to balance the resource and traffic demand among different groups, since different content placement leads to different group load distributions (i.e., $\{N_k\}$).

Problem (P1) has two-fold challenges. Firstly, the average delay cannot be derived in closed form, due to the multi-dimensional randomness of network topology, traffic distribution, as well as file popularity. Accordingly, the relationship between average delay and content placement cannot be obtained directly. Secondly, as $\{c_f\}$ takes discrete values while $\{\phi_k\}$ is continuous, (P1) is a mixed integer programing problem and cannot be solved in polynomial time. To deal with the first challenge, we apply the stochastic geometry and analyze the average delay approximately based on the high SINR requirements of practical networks. As for the second challenge, (P1) can be solved in two steps: 1) *Bandwidth Allocation Subproblem*: deriving the optimal bandwidth allocation for any given content placement; and 2) *Content Placement with Optimal Bandwidth Allocation*: optimizing the caching problem by substituting the obtained optimal bandwidth allocation into (P1). In the following sections, the analysis of average file transmission delay, bandwidth allocation and content placement subproblems will be presented in details.

4 Average Delay Analysis

In this section, we analyze the average file transmission delay via stochastic geometry, based on which the bandwidth allocation subproblem is solved by applying the Lagrange multiplier method.

4.1 File Transmission Delay

Due to the multi-dimensional randomness of cell load, transmission distance and coverage area, the accurate transmission rate cannot be derived, posing challenges to the design of cooperative caching. However, a lower bound of average file transmission rate can be approximately obtained in closed form, based on the theory of stochastic geometry, given as Lemma 1.

Lemma 1. The average file transmission rate for Group-k users has a lower bound,

$$\mathbb{E}[R_k] \ge \frac{\phi_k W \rho}{\lambda \Omega_k}\left[\log_2 \frac{P_{\mathrm{T}}(\pi\rho)^{\frac{\alpha}{2}}}{\sigma^2 + I_k} + \frac{\alpha}{2\ln 2}(\gamma - \sum_{m=1}^{k-1}\frac{1}{m})\right], \tag{9}$$

where $\gamma \approx 0.577$ denotes the Euler-Mascheroni constant. In addition, the equality holds when $\frac{\sigma^2+I_k}{P_{\mathrm{T}}} \to 0$.

Proof. As the received SINR and cell load can be considered as independent variables, the average file transmission rate can be approximated as:

$$\mathbb{E}[R_k] = \frac{\phi_k W}{\ln 2}\frac{\mathbb{E}\left[\ln\left(1+\zeta_k\right)\right]}{\mathbb{E}[N_k]}, \tag{10}$$

where $\mathbb{E}[N_k] = \lambda\Omega_k/\rho$ since both users and SBSs follow PPPs. Furthermore,

$$\mathbb{E}\left[\ln\left(1+\zeta_k\right)\right] \ge -\alpha\mathbb{E}\left[\ln\left(d_k\right)\right] + \ln\left(\frac{P_{\mathrm{T}}}{I_k+\sigma^2}\right), \tag{11}$$

where the equality holds if $\frac{I_k+\sigma^2}{P_{\mathrm{T}}} \to 0$ (i.e., high SINR). The cumulative distribution function (CDF) of the transmission distance d_k is given by

$$\mathbb{P}\left\{d_k \le D\right\} = 1 - \sum_{m=0}^{k-1}\frac{(\pi D^2\rho)^m}{m!}e^{-\pi D^2\rho}, \tag{12}$$

i.e., the probability that there are at least k SBSs within the circular of radius d_k centered at a user. Thus,

$$\begin{aligned}
\mathbb{E}\left[\ln\left(d_k\right)\right] &= \int_0^\infty \ln D \ \mathrm{d}(\mathbb{P}\left\{d_k \le D\right\}) \\
&= \int_0^\infty 2\pi\rho D e^{-\pi\rho D^2}\ln D \ \mathrm{d}D \\
&\quad - \int_0^\infty \ln D \ \mathrm{d}\left(\sum_{m=1}^{k-1}\frac{(\pi D^2\rho)^m}{m!}e^{-\pi D^2\rho}\right) \\
&= \int_0^\infty \frac{1}{2}e^{-z}\ln z dz - \int_0^\infty \frac{1}{2}e^{-z}\ln\pi\rho dz - \sum_{m=1}^{k-1}\frac{(\pi D^2\rho)^m}{m!} \\
&\quad \cdot e^{-\pi D^2\rho}\ln D\Big|_0^\infty + \sum_{m=1}^{k-1}\int_0^\infty \frac{(\pi\rho D^2)^m}{m!}e^{-\pi\rho D^2}\mathrm{d}\left(\ln D\right) \\
&= -\frac{\gamma}{2} - \frac{\ln\pi\rho}{2} + \sum_{m=1}^{k-1}\int_0^\infty \frac{(\pi\rho)^m}{m!}D^{2m-1}e^{-\pi\rho D^2}\mathrm{d}D \\
&= -\frac{\gamma}{2} - \frac{\ln\pi\rho}{2} + \sum_{m=1}^{k-1}\frac{1}{2m},
\end{aligned} \tag{13}$$

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

where γ is the Euler-Mascheroni constant with numerical value of 0.577215664902... [41]. Substituting (13) into (11) and (10), Lemma 1 can be proved. ∎

The lower bound abstracts wireless transmission rate with respect to physical layer parameters and the group load distributions. As the group load distribution results from the cooperative caching schemes, Lemma 1 can be applied to analyze the performance of cooperative caching while taking into account wireless transmission features. The lower-bound approximation is quite accurate in high SINR region, and hence supports conservative analysis. In practical systems, most users are guaranteed with high SINR due to the reliable communication requirement, which can be achieved through advanced interference mitigation techniques. The accuracy of the lower-bound will be validated in the simulation section. Lemma 1 indicates that the file transmission rate of Group-k varies with the allocated bandwidth ϕ_k and traffic load Ω_k. For notation simplicity, the average file transmission rate is rewritten as

$$\mathbb{E}[R_k] \approx \frac{\phi_k}{\Omega_k} W \tau_k, \tag{14}$$

where

$$\tau_k \triangleq \frac{\rho}{\lambda}\left[\log_2 \frac{P_{\mathrm{T}}(\pi\rho)^{\frac{\alpha}{2}}}{\sigma^2 + I_k} + \frac{\alpha}{2\ln 2}(\gamma - \sum_{m=1}^{k-1}\frac{1}{m})\right], \tag{15}$$

representing the average spectrum efficiency of the kth candidate SBS. Specifically, the physical meaning of $W\tau_k$ is the average file transmission rate when all users are served by their kth candidate SBS with all available system bandwidth. τ_k is irrelevant with content placement and bandwidth allocation, whereas depends on the overall traffic load (i.e., λ/ρ) and network resources (such as cell density, and transmitted SINR). Furthermore, τ_k decreases with k, due to high transmission distance and path loss.

4.2 Bandwidth Allocation Subproblem

Substituting the results of Lemma 1 into Eq. (7), the average file transmission delay can be rewritten as

$$\bar{D} = \left[\sum_{k=1}^{K+1} \frac{{\Omega_k}^2}{W\tau_k\phi_k}\right]\bar{S}L + \bar{D}_{\mathrm{BH}}\Omega_{K+1}. \tag{16}$$

For the given content placement $\{c_f\}$, $[\Omega_1, \Omega_2, \cdots, \Omega_{K+1}]$ can be derived based on Eqs. (1), (2), and (3). As a result, the average file transmission delay only varies with $\{\phi_k\}$ in Eq. (16). The bandwidth allocation subproblem can be formulated as follows:

$$\min_{\{\phi_k\}} \quad \left[\sum_{k=1}^{K+1} \frac{{\Omega_k}^2}{W\tau_k\phi_k}\right]\bar{S}L + \bar{D}_{\mathrm{BH}}\Omega_{K+1}, \tag{17a}$$

$$\text{(SP1)} \quad \text{s.t.} \quad \sum_{k=1}^{K+1}\phi_k \le 1, \tag{17b}$$

$$\phi_k \ge 0, \quad k = 1, 2, \cdots, K+1. \tag{17c}$$

(SP1) is a convex optimization with respect to $\{\phi_k\}$, since $\frac{\partial \bar{D}^2}{\partial^2 \phi_k} = \frac{\Omega_k^2}{2W\tau_k{\phi_k}^3} > 0$. The optimal bandwidth allocation is given as Proposition 1, obtained with the Lagrange multiplier method.

Proposition 1. For the given content placement, the optimal bandwidth allocation is given as

$$\hat{\phi}_k = \frac{\frac{\Omega_k}{\sqrt{\tau_k}}}{\sum_{j=1}^{K+1}\frac{\Omega_j}{\sqrt{\tau_j}}}. \tag{18}$$

Proof. The Lagrange function of bandwidth allocation subproblem is given by

$$\begin{aligned} &G(\phi_1, \cdots, \phi_{K+1}) \\ &= \sum_{k=1}^{K+1}\frac{{\Omega_k}^2\bar{S}L}{W\phi_k\tau_k} + \bar{D}_{\mathrm{BH}}\Omega_{K+1} + \xi_0\left(\sum_{k=1}^{K+1}\phi_k - 1\right) - \sum_{k=1}^{K+1}\xi_k\phi_k, \end{aligned} \tag{19}$$

where $\xi_0 \ge 0$ is the Lagrange multiplier for constraint $\sum_{k=1}^{K+1}\phi_k \le 1$, $\xi_k \ge 0$ is the multiplier for constraint $\phi_k \ge 0$ and $k = 1, 2, \cdots, K+1$. Taking the derivative of the Lagrange function, the optimal solution should satisfy

$$\frac{\partial G(\phi_1, \cdots, \phi_{K+1})}{\partial \phi_k} = -\frac{{\Omega_k}^2\bar{S}L}{W\tau_k{\phi_k}^2} + \xi_0 - \sum_{k=1}^{K+1}\xi_k = 0, \tag{20}$$

which is equivalent to

$$\phi_k = \Omega_k\sqrt{\frac{\bar{S}L}{W\tau_k(\xi_0 - \sum_{k=1}^{K+1}\xi_k)}}. \tag{21}$$

Due to the complementary slackness of a constraint and its optimal Lagrange multiplier, $\phi_k > 0$ implies $\mu_k = 0$, $\forall k$. As a result, the optimal bandwidth allocation should also satisfy $\xi_0\left(\sum_{k=1}^{K+1}\phi_k - 1\right) = 0$, i.e.,

$$\xi_0\left(\sum_{k=1}^{K+1}\Omega_k\sqrt{\frac{\bar{S}L}{W\tau_k\xi}} - 1\right) = 0. \tag{22}$$

Therefore, $\sqrt{\xi_0} = \sum_{k=1}^{K+1}\Omega_k\sqrt{\frac{\bar{S}L}{W\tau_k}}$, and Proposition 1 is thus proved. ∎

Substitute (18) into (16), the average file transmission delay can be rewritten as

$$\begin{aligned} \bar{D} &= \left(\sum_{k=1}^{K+1}\frac{\Omega_k}{\sqrt{\tau_k}}\right)^2\frac{\bar{S}L}{W} + \bar{D}_{\mathrm{BH}}\Omega_{K+1} \\ &= \left(\sum_{k=1}^{K}\frac{\Omega_k}{\sqrt{\tau_k}} + \frac{1-\sum_{k=1}^{K}\Omega_k}{\sqrt{\tau_1}}\right)^2\frac{\bar{S}L}{W} + \bar{D}_{\mathrm{BH}}\left(1 - \sum_{k=1}^{K}\Omega_k\right) \\ &= \left(\sum_{k=2}^{K}\left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_1}}\right)\Omega_k + \frac{1}{\sqrt{\tau_1}}\right)^2\frac{\bar{S}L}{W} + \bar{D}_{\mathrm{BH}}\left(1 - \sum_{k=1}^{K}\Omega_k\right). \end{aligned} \tag{23}$$

The two parts can be interpreted as the average wireless transmission delay and backhaul delay, respectively. Notice that placing more diverse files in cache increases the content hit ratio Ω_k ($k = 1, 2, \cdots, K$). Based on Eq. (23), the average backhaul delay decreases, whereas the average wireless transmission delay increases with Ω_k, where $k = 2, 3, \cdots, K$. This result reflects the tradeoff between content diversity gain and spectrum efficiency degradation under cooperative caching, which is influenced by the

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

cache size. Specifically, larger cluster size brings higher content diversity gain, but results in lower spectrum efficiency. Take derivative of Eq. (23) with respect to Ω_k:

$$\begin{aligned}\frac{\partial \bar{D}}{\partial \Omega_k} =& \frac{2\bar{S}L}{W}\left(\sum_{k=2}^{K}\left(\frac{1}{\sqrt{\tau_k}}-\frac{1}{\sqrt{\tau_1}}\right)\Omega_k+\frac{1}{\sqrt{\tau_1}}\right)\\ &\cdot\left(\frac{1}{\sqrt{\tau_k}}-\frac{1}{\sqrt{\tau_1}}\right)-\bar{D}_{\mathrm{BH}},\end{aligned} \tag{24}$$

where $k = 2, 3, \cdots, K$. If $\frac{\partial \bar{D}}{\partial \Omega_k} \leq 0$, the backhaul delay dominants, and we should enhance content diversity to reduce average file transmission delay. Otherwise, the wireless transmission delay dominants, and using the kth candidate SBS may even increase average transmission delay due to the large path loss. To balance the gain of content diversity and degradation of spectrum efficiency, $\frac{\partial \bar{D}}{\partial \Omega_k} \leq 0$ should be always guaranteed, and a sufficient condition is given in Proposition 2.

Proposition 2. If

$$\frac{2\bar{S}L}{W\sqrt{\tau_K}}\left(\frac{1}{\sqrt{\tau_K}}-\frac{1}{\sqrt{\tau_1}}\right)\leq \bar{D}_{\mathrm{BH}}, \tag{25}$$

$\frac{\partial \bar{D}}{\partial \Omega_k} \leq 0$ can be guaranteed $\forall \Omega_k \in [0,1]$, where $k = 2, 3, \cdots, K$.

Proof. As $\tau_{K+1} = \tau_1$ and $\tau_K \leq \tau_k$ (for $k = 1, 2, \cdots, K+1$), we have

$$\begin{aligned}&\sum_{k=2}^{K}\left(\frac{1}{\sqrt{\tau_k}}-\frac{1}{\sqrt{\tau_1}}\right)\Omega_k+\frac{1}{\sqrt{\tau_1}}\\ &=\sum_{k=1}^{K+1}\left(\frac{1}{\sqrt{\tau_k}}-\frac{1}{\sqrt{\tau_1}}\right)\Omega_k+\frac{1}{\sqrt{\tau_1}}\sum_{k=1}^{K+1}\Omega_k\\ &=\sum_{k=1}^{K+1}\frac{1}{\sqrt{\tau_k}}\Omega_k\leq\frac{1}{\sqrt{\tau_K}}.\end{aligned} \tag{26}$$

Furthermore,

$$\frac{1}{\sqrt{\tau_k}}-\frac{1}{\sqrt{\tau_1}}\leq\frac{1}{\sqrt{\tau_K}}-\frac{1}{\sqrt{\tau_1}}, \tag{27}$$

Substitute Eqs. (26) and (27) into Eq. (24):

$$\frac{\partial \bar{D}}{\partial \Omega_k}\leq\frac{2\bar{S}L}{W}\frac{1}{\sqrt{\tau_K}}\left(\frac{1}{\sqrt{\tau_K}}-\frac{1}{\sqrt{\tau_1}}\right)-\bar{D}_{\mathrm{BH}}. \tag{28}$$

Therefore, Eq. (25) provides a sufficient condition to $\frac{\partial \bar{D}}{\partial \Omega_k} \leq 0$, and Proposition 2 is thus proved. ∎

Notice that the left part of Eqn. (25) can be interpreted as the maximal degradation of spectrum efficiency when the cluster size is K[3].. Therefore, Proposition 2 indicates that the degradation of spectrum efficiency should be no larger than the backhaul delay. Otherwise, the delay of remote file fetching is even smaller than the delay of fetching files from SBS B_K, and B_K should not be involved into cluster. Proposition 2 offers a guideline for SBS clustering. As the left part of Eqn. (25) (i.e., the spectrum efficiency degradation) increases with K, Eqn. (25) constraints the maximal cluster size. Furthermore, the maximal cluster size depends on network parameters. For example, larger backhaul delay indicates larger cluster size. In addition, the cluster size can increase in dense networks, as τ_K increases with shorter transmission distance.

3. "Eqn". is short for "inequation", and "Eq." is short for "Equation"

5 Content Placement with Optimal Bandwidth Allocation

Based on the result of optimal bandwidth allocation, the content placement subproblem can be formulated as follows:

$$\min_{\{c_f\}} \ \frac{\bar{S}L}{W}\left(\sum_{k=1}^{K+1}\frac{\Omega_k}{\sqrt{\tau_k}}\right)^2+\bar{D}_{\mathrm{BH}}\Omega_{K+1}, \tag{29a}$$

$$\text{(SP2)} \quad \text{s.t.} \ \sum_{f=1}^{F}c_f\leq C, \tag{29b}$$

$$c_f\in\{0,1,\cdots,s_f\}, \ \ \forall f\in\mathcal{F}. \tag{29c}$$

The complexity of this integer programing problem increases exponentially with the number of files and file size. In addition, the average delay is piecewise with respect to caching placement based on Eqs. (1), (2) and (3), posing additional challenges. Therefore, a sub-optimal algorithm with low complexity should be devised for practical network operations. To this end, we first analyze the delay improvement by adding a segment into cache, and then design a greedy algorithm based on the submodular properties.

5.1 Delay Improvement with Segment Placement

When a segment is added into cache, the content hit ratio increases, influencing average file transmission delay. Suppose the number of segments cached in the SBS is $\mathcal{C} = \{c_1, c_2, \cdots, c_f, \cdots, c_F\}$, whose corresponding content hit ratio and group load distribution is $\{P_{k,f}\}$ and $\{\Omega_k\}$, respectively. Assume file-h is not completely stored (i.e., $c_h < s_h$), and we add a segment of file-h into cache. Accordingly, the number of cached segments becomes $\mathcal{C}' = \{c'_1, c'_2, \cdots, c'_f, \cdots, c'_F\}$, where $c'_h = c_h + 1$, $c'_f = c_f$, $\forall f \in \mathcal{F}$ and $f \neq h$. Denote by $\{P'_{k,f}\}$ and $\{\Omega'_k\}$ the updated content hit ratio and group load distribution, respectively. To evaluate the delay performance improved by caching the segment, we define $V(\mathcal{C}, h)$ as the marginal gain:

$$V(\mathcal{C},h)=\bar{D}\Big|_{\mathcal{C}}-\bar{D}\Big|_{\mathcal{C}'}, \tag{30}$$

where $\bar{D}\Big|_{\mathcal{C}}$ and $\bar{D}\Big|_{\mathcal{C}'}$ denote the average file transmission delays when the cached segment set is $\mathcal{C}$ and $\mathcal{C}'$, respectively. Substitute Eq. (23) into Eq. (30), and the marginal gain is given by

$$\begin{aligned}V(\mathcal{C},h)=&\bar{D}_{\mathrm{BH}}(\Omega_{K+1}-\Omega'_{K+1})\\ &+\frac{\bar{S}L}{W}\left(\sum_{k=1}^{K+1}\frac{\Omega_k+\Omega'_k}{\sqrt{\tau_k}}\right)\left(\sum_{k=1}^{K+1}\frac{\Omega_k-\Omega'_k}{\sqrt{\tau_k}}\right),\end{aligned} \tag{31}$$

indicating that the key factor of the marginal gain is the variation of group load distribution.

Denote by $\delta_k = \Omega'_k - \Omega_k$ and $\Delta = [\delta_1, \delta_2, \cdots, \delta_{K+1}]$, i.e., the variations of group load distribution. The group

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

load distribution depends on the content hit ratio, according to Eq. (3). For the other files except h, the number of cached segment remains the same, and thus the content hit ratio does not change, i.e., $P'_{k,f} = P_{k,f}$ for $f \neq h$. Therefore, the variations of group load distribution is given by $\delta_k = q_h(P'_{k,h} - P_{k,h})$ for $k = 1, 2, \cdots, K+1$. As the content hit ratio is piecewise with respect to the SBS index k according to Eq. (1) and (2), δ_k also has a piecewise structure. The result of δ_k can be classified into the following two cases, depending on the number of cached segments c_h [4].

Case 1. If $c_h \leq \frac{s_h}{K} - 1$, the variation of group load distribution is given by

$$\delta_k = \begin{cases} \frac{q_h}{s_h}, & k = 1, 2, \cdots, K, \\ -\frac{K q_h}{s_h}, & k = K+1. \end{cases} \tag{32}$$

Case 2. If $c_h \geq \frac{s_h}{K}$, the variation of group load distribution is given by

$$\begin{cases} \delta_k = \frac{q_h}{s_h}, & k = 1, 2, \cdots, \check{K}, \\ \delta_k < 0, & k = \check{K}+1, \cdots, \hat{K}, \\ \delta_k = 0, & k = \hat{K}+1, \cdots, K+1, \end{cases} \tag{33}$$

where $\sum_{k=1}^{\hat{K}} \delta_k = 0$, $\check{K} = \lfloor \frac{s_h}{c_h+1} \rfloor$, and $\hat{K} = \lceil \frac{s_h}{c_h} \rceil$.

When the number of cached segments is small, extra segments need to be fetched from remote servers for file decoding. Accordingly, adding a new segment into cache can reduce content miss ratio, which is the situation of Case 1. However, when the number of cached segments exceeds some threshold, the requested file can be decoded by downloading files locally from the candidate SBSs, i.e., no content miss in Case 2. Therefore, adding a new segment into cache no longer improves total content hit ratio, but users can finish file downloading from closer SBSs with higher transmission rate. For example, suppose there are 6 candidate SBSs and consider a file-h requiring 100 segments to decode. If each SBS caches 24 segments, users need to download segments from SBS B_1-B_5 for file decoding. If each SBS caches 25 segments, users can finish file decoding only from SBS B_1-B_4. In this case, $\check{K} = 4$, $\hat{K} = 5$, and $\Delta = [\frac{1}{200}, \frac{1}{200}, \frac{1}{200}, \frac{1}{200}, -\frac{1}{50}, 0, 0]$ if the popularity is $q_h = 0.5$.

The tradeoff between content diversity and spectrum efficiency influences the marginal gain differently in the two cases. In Case 1, adding a new segment enhances content diversity as well as content hit ratio. Accordingly, the backhaul transmission delay can be reduced, but the wireless transmission delay increases due to degraded spectrum efficiency. In Case 2, adding a new segment cannot enhance content diversity, and brings no backhaul delay improvement. However, wireless transmission delay can be reduced with improved spectrum efficiency, since users can download file from BSs in proximity.

5.2 Submodular Property Analysis

In what follows, we analyze the marginal gain of content placement with respect to $\mathcal{C}$ and h. Specifically, the marginal gain can be proved to have monotone submodular properties under the condition of Proposition 2, given as Proposition 3.

Proposition 3. For any two feasible cache placement $\mathcal{C} = \{c_1, \cdots, c_F\}$ and $\mathcal{C}' = \{c'_1, \cdots, c'_F\}$ satisfying $c_f \leq c'_f$ ($\forall f \in \mathcal{F}$), we have $V(\mathcal{C}, h) \geq V(\mathcal{C}', h) > 0$ under the condition of Proposition 2, where $c_h \leq c'_h < s_h$.

Proof. The proof can be conducted in two steps: 1) the marginal gain is positive, i.e., $V(\mathcal{C}, h) > 0$; and 2) the marginal gain decreases with the number of cached contents, i.e., $V(\mathcal{C}', h) < V(\mathcal{C}, h)$. For the proof of step 2), we prove $V(\mathcal{C}', h) < V(\mathcal{C}, h)$ under a special case: suppose $\mathcal{C}$ is the initial cache placement where $c_h \leq s_h - 2$, $\mathcal{C}'$ is the cache placement when a segment of file-h is added, and $\mathcal{C}''$ is the cache placement when two segments of file-h are added. For the general case of $c'_f \geq c_f$, $V(\mathcal{C}', h) < V(\mathcal{C}, h)$ can be proved through recursion. Furthermore, the proof needs to be conducted under both cases, respectively.

(1) $V(\mathcal{C}, h) > 0$ **in Case 1:**

Substituting Eq. (32) into (31), the marginal value can be rewritten as

$$\begin{aligned} &V(\mathcal{C}, h) \\ &= K\bar{D}_{\mathrm{BH}} \frac{q_h}{s_h} - \frac{\bar{S}L}{W} \left(\sum_{k=1}^{K+1} \frac{\Omega_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1} \frac{\Omega'_k}{\sqrt{\tau_k}} \right) \left(\sum_{k=1}^{K+1} \frac{\delta_k}{\sqrt{\tau_k}} \right) \\ &= K\bar{D}_{\mathrm{BH}} \frac{q_h}{s_h} - \frac{\bar{S}L q_h}{W s_h} \left(\sum_{k=1}^{K+1} \frac{\Omega_k + \Omega'_k}{\sqrt{\tau_k}} \right) \sum_{k=1}^{K} \left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_{K+1}}} \right) \end{aligned} \tag{34}$$

As $\tau_{K+1} = \tau_1 > \tau_2 > \cdots > \tau_K$, we have

$$\sum_{k=1}^{K} \left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_1}} \right) < K \left(\frac{1}{\sqrt{\tau_K}} - \frac{1}{\sqrt{\tau_1}} \right), \tag{35}$$

and

$$\frac{2}{\sqrt{\tau_K}} > \sum_{k=1}^{K+1} \frac{\Omega_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1} \frac{\Omega'_k}{\sqrt{\tau_k}}, \tag{36}$$

due to $\sum_{k=1}^{K+1} \Omega_k = \sum_{k=1}^{K+1} \Omega'_k = 1$. According to (25),

$$\begin{aligned} K\bar{D}_{\mathrm{BH}} &\geq 2K \frac{\bar{S}L}{W} \frac{1}{\sqrt{\tau_K}} \left(\frac{1}{\sqrt{\tau_K}} - \frac{1}{\sqrt{\tau_1}} \right) \\ &> \frac{\bar{S}L}{W} \left(\sum_{k=1}^{K+1} \frac{\Omega_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1} \frac{\Omega'_k}{\sqrt{\tau_k}} \right) \sum_{k=1}^{K} \left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_1}} \right). \end{aligned} \tag{37}$$

Hence, $V(\mathcal{C}, h) > 0$.

(2) $V(\mathcal{C}, h) > 0$ **in Case 2:**

Substituting Eq. (33) into (31), the marginal value can be rewritten as

$$V(\mathcal{C}, h) = -\frac{\bar{S}L}{W} \left(\sum_{k=1}^{K+1} \frac{\Omega_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1} \frac{\Omega'_k}{\sqrt{\tau_k}} \right) \left(\sum_{k=1}^{\hat{K}-1} \frac{\delta_k}{\sqrt{\tau_k}} \right). \tag{38}$$

4. Assume s_f/K is an integer without losing generality.

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

As $\delta_1 = \delta_2 = \cdots = \delta_{\check{K}} = q_h/s_h > 0 > \delta_{\check{K}+1} \geq \cdots \geq \delta_{\hat{K}}$ and $\tau_1 > \tau_2 > \cdots > \tau_{\hat{K}}$, we have

$$\begin{aligned}\sum_{k=1}^{\hat{K}} \frac{\delta_k}{\sqrt{\tau_k}} &= \sum_{k=1}^{\check{K}} \frac{\delta_k}{\sqrt{\tau_k}} + \sum_{k=\check{K}+1}^{\hat{K}} \frac{\delta_k}{\sqrt{\tau_k}} \\ &< \sum_{k=1}^{\check{K}} \frac{\delta_k}{\sqrt{\tau_{\check{K}}}} + \sum_{k=\check{K}+1}^{\hat{K}} \frac{\delta_k}{\sqrt{\tau_{\check{K}}}} = 0.\end{aligned} \tag{39}$$

Hence, $V(\mathcal{C}, h) > 0$.

(3) $V(\mathcal{C}', h) < V(\mathcal{C}, h)$ **in Case 1:**

Denote by $\{\Omega_k\}$, $\{\Omega'_k\}$ and $\{\Omega''_k\}$ the group load distributions, corresponding to cache placement is $\mathcal{C}$, $\mathcal{C}'$ and $\mathcal{C}''$, respectively. Denote by $\delta_k = \Omega'_k - \Omega_k$, and $\delta'_k = \Omega''_k - \Omega'_k$, for $k = 1, \cdots, K+1$. According to Eq. (34) and $\tau_{K+1} = \tau_1$, we have

$$\begin{aligned}&V(\mathcal{C}', h) - V(\mathcal{C}, h) \\ =&\frac{\bar{S}L}{W}\frac{q_h}{s_h}\left(\sum_{k=1}^{K+1}\frac{\Omega_k - \Omega''_k}{\sqrt{\tau_k}}\right)\sum_{k=1}^{K}\left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_1}}\right).\end{aligned} \tag{40}$$

As $\Omega_k - \Omega''_k = -2q_h/s_h$ for $k = 1, \cdots, K$ and $\Omega_{K+1} - \Omega''_{K+1} = 2Kq_h/s_h$, we have

$$\sum_{k=1}^{K+1}\frac{\Omega_k - \Omega''_k}{\sqrt{\tau_k}} = -2\frac{q_h}{s_h}\sum_{k=1}^{K+1}\left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_1}}\right). \tag{41}$$

Therefore,

$$\begin{aligned}&V(\mathcal{C}', h) - V(\mathcal{C}, h) \\ =&-\frac{2\bar{S}L}{W}\left(\frac{q_h}{s_h}\right)^2\left[\sum_{k=1}^{K+1}\left(\frac{1}{\sqrt{\tau_k}} - \frac{1}{\sqrt{\tau_1}}\right)\right]^2 < 0.\end{aligned} \tag{42}$$

(4) $V(\mathcal{C}', h) < V(\mathcal{C}, h)$ **in Case 2:**

According to Eq. (38), we have

$$\begin{aligned}&V(\mathcal{C}', h) - V(\mathcal{C}, h) \\ =&\frac{\bar{S}L}{W}\left[\left(\sum_{k=1}^{K+1}\frac{\Omega_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1}\frac{\Omega'_k}{\sqrt{\tau_k}}\right)\left(\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}}\right)\right. \\ &\left. - \left(\sum_{k=1}^{K+1}\frac{\Omega'_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1}\frac{\Omega''_k}{\sqrt{\tau_k}}\right)\left(\sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}}\right)\right],\end{aligned} \tag{43}$$

where $\hat{K} = \lceil \frac{s_h}{c_h} \rceil$ and $\hat{K}' = \lceil \frac{s_h}{c_h+1} \rceil$. Notice that $\Omega''_k - \Omega_k$ corresponds to the variations of group load distributions when two segments of file-h are added, which has a similar structure of δ_k:

$$\begin{cases}\Omega''_k - \Omega_k = \frac{2q_h}{s_h}, & k = 1, 2, \cdots, \check{K}'', \\ \Omega''_k - \Omega_k < 0, & k = \check{K}'' + 1, \cdots, \hat{K}'', \\ \Omega''_k - \Omega_k = 0, & k = \hat{K}'' + 1, \cdots, K+1,\end{cases} \tag{44}$$

where $\sum_{k=1}^{\hat{K}''}(\Omega''_k - \Omega_k) = 0$, $\check{K}'' = \lfloor \frac{s_h}{c_h+2} \rfloor$, and $\hat{K}'' = \lceil \frac{s_h}{c_h} \rceil$. Due to the same reason of Eqn. (39),

$$\sum_{k=1}^{K+1}\frac{\Omega''_k - \Omega_k}{\sqrt{\tau_k}} < 0. \tag{45}$$

Therefore,

$$\sum_{k=1}^{K+1}\frac{\Omega_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1}\frac{\Omega'_k}{\sqrt{\tau_k}} > \sum_{k=1}^{K+1}\frac{\Omega'_k}{\sqrt{\tau_k}} + \sum_{k=1}^{K+1}\frac{\Omega''_k}{\sqrt{\tau_k}} > 0. \tag{46}$$

In what follows, we prove $\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} < \sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}} < 0$. As $c_h \geq \frac{s_h}{K}$ in Case 2, we have $\frac{s_h}{c_h} - \frac{s_h}{c_h+1} = \frac{s_h}{(c_h+1)c_h} < \frac{K^2}{s_h}$. In practical systems, the number of candidate SBSs K (usually smaller than 10) is generally much smaller than s_h (usually hundreds or thousands). Thus, we assume $\frac{s_h}{c_h} - \frac{s_h}{c_h+1} \leq 1$. Therefore, $\hat{K}' \leq \hat{K} \leq \hat{K}' + 1$, and $\check{K} \leq \hat{K} \leq \check{K} + 2$. Similarly, $\check{K}' \leq \check{K} \leq \check{K}' + 1$, $\check{K}' \leq \hat{K}' \leq \check{K}' + 2$, and $\check{K}' \leq \hat{K} \leq \check{K}' + 2$.

If $\check{K} = \check{K}'$, we have $\hat{K}' = \check{K}' + 1$, $\delta_k = \delta'_k = q_h/s_h$ for $k = 1, 2, \cdots, \check{K}$, $\delta_{\check{K}+1} < 0$, $\delta_{\check{K}+2} \leq 0$, $\delta'_{\check{K}+1} < 0$, $\delta'_{\check{K}+2} = 0$, and $\delta_k = \delta'_k = 0$ for $k = \check{K} + 3, \cdots, K$. Thus,

$$\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} = \sum_{k=1}^{\check{K}}\frac{\delta_k}{\sqrt{\tau_k}} + \frac{\delta_{\check{K}+1}}{\sqrt{\tau_{\check{K}+1}}} + \frac{\delta_{\check{K}+2}}{\sqrt{\tau_{\check{K}+2}}}, \tag{47}$$

and

$$\sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}} = \sum_{k=1}^{\check{K}}\frac{\delta'_k}{\sqrt{\tau_k}} + \frac{\delta'_{\check{K}+1}}{\sqrt{\tau_{\check{K}+1}}}. \tag{48}$$

As $\sum_{k=1}^{\check{K}+2}\delta_k = \sum_{k=1}^{\check{K}+1}\delta'_k = 0$, $\delta_{\check{K}+1} + \delta_{\check{K}+1} = \delta'_{\check{K}+1} = -\check{K}q_h/s_h < 0$, we further have

$$\begin{aligned}&\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} - \sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}} \\ =&\frac{\delta_{\check{K}+1}}{\sqrt{\tau_{\check{K}+1}}} + \frac{\delta_{\check{K}+2}}{\sqrt{\tau_{\check{K}+2}}} - \frac{\delta_{\check{K}+1} + \delta_{\check{K}+2}}{\sqrt{\tau_{\check{K}+1}}} \\ =&\delta_{\check{K}+2}\left(\frac{1}{\sqrt{\tau_{\check{K}+2}}} - \frac{1}{\sqrt{\tau_{\check{K}+1}}}\right) \leq 0,\end{aligned} \tag{49}$$

since $\tau_{\check{K}+1} > \tau_{\check{K}+2}$.

If $\check{K} > \check{K}'$, we have $\check{K} = \check{K}' + 1$, $\hat{K}' \leq \check{K}' + 2$, and $\hat{K} = \check{K} + 1$. Thus, $\delta_k = \delta'_k = q_h/s_h$ for $k = 1, 2, \cdots, \check{K}'$, $\delta_{\check{K}'+1} = q_h/s_h$, $\delta_{\check{K}'+2} < 0$, $\delta'_{\check{K}'+1} < 0$, $\delta'_{\check{K}'+2} \leq 0$, and $\delta_k = \delta'_k = 0$ for $k = \check{K}' + 3, \cdots, K$. Accordingly,

$$\begin{aligned}&\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} - \sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}} \\ =&\frac{\delta_{\check{K}'+1}}{\sqrt{\tau_{\check{K}'+1}}} + \frac{\delta_{\check{K}'+2}}{\sqrt{\tau_{\check{K}'+2}}} - \frac{\delta'_{\check{K}'+1}}{\sqrt{\tau_{\check{K}'+1}}} - \frac{\delta'_{\check{K}'+2}}{\sqrt{\tau_{\check{K}'+2}}} \\ <&\frac{\delta_{\check{K}'+1} + \delta_{\check{K}'+2}}{\sqrt{\tau_{\check{K}'+2}}} - \frac{\delta'_{\check{K}'+1} + \delta'_{\check{K}'+2}}{\sqrt{\tau_{\check{K}'+2}}} = 0,\end{aligned} \tag{50}$$

since $\delta_{\check{K}'+1} + \delta_{\check{K}'+2} = \delta'_{\check{K}'+1} + \delta'_{\check{K}'+2} = -\check{K}'q_h/s_h$. With Eqns. (49) and (50), we have $\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} < \sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}}$. Furthermore, we can prove $\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} < 0$ and $\sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}} < 0$ in the same way as Eqn. (39). Combining $\sum_{k=1}^{\hat{K}}\frac{\delta_k}{\sqrt{\tau_k}} < \sum_{k=1}^{\hat{K}'}\frac{\delta'_k}{\sqrt{\tau_k}} < 0$ with Eqn. (46), $V(\mathcal{C}', h) < V(\mathcal{C}, h)$ can be proved in Case 2.

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

Algorithm 1 The Proposed Greedy Algorithm

Input: $\mathcal{F}$: file library; $\{q_f\}$: file popularity distribution; $\{s_f\}$: file length; C: cache size of SBSs; ρ: SBS density; K: cluster size; $\bar{D}_{\rm BH}$: average backhaul delay;

Output: $\{c_f\}$: number of cached segments; $\{\hat{\phi}_f\}$: the optimal bandwidth allocation;

1: Calculate τ_k based on Eq. (15), for $k = 1, 2, \cdots, K+1$;
2: Set $\mathcal{Y}_f = \emptyset,\ c_f = 0, \forall f \in \mathcal{F}$;
3: **while** $\sum_{f=1}^{F} c_f < C$ **do**
4: Calculate the group load distribution $\{\Omega_k\}$ with Eq. (52);
5: $h^* = \arg\max_h V(\{c_f\}, h)$, based on Eq. (31);
6: $\mathcal{Y}_{h^*} \leftarrow \mathcal{Y}_{h^*} \cup \{x_{c_{h^*}+1}\},\ c_{h^*} = c_{h^*} + 1$;
7: **end while**
8: Calculate $\{\hat{\phi}_f\}$ with Proposition 1;
9: Return $\{c_f\}$ and $\{\hat{\phi}_f\}$;

Therefore, $0 < V(\mathcal{C}', h) < V(\mathcal{C}, h)$ holds for both cases, and Proposition 3 can be proved. ∎

Remark. Proposition 3 reveals two facts: 1) adding a segment in cache can always improve the delay performance; and 2) the marginal gain of adding segments in cache decreases when more segments are stored in cache. With this monotone submodular property, greedy algorithms can be designed with near-optimal performance guarantee.

5.3 Greedy Content Placement Algorithm

The content placement problem can be transferred into a submodular optimization problem. Denote by $\mathcal{X}_f$ a set of coded segments which can support the decoding of file-f, and $|\mathcal{X}_f| = s_f$. Denote by $\mathcal{Y}_f \subseteq \mathcal{X}_f$ the set of segments placed in cache, and $|\mathcal{Y}_f| = c_f$ in problem (SP2) [5]. The content placement can be reformulated as (SP3).

$$\max_{\{\mathcal{Y}_f\}} \left[\frac{\bar{S}L}{W\tau_1} + \bar{D}_{\rm BH}\right] - \left[\frac{\bar{S}L}{W}\left(\sum_{k=1}^{K+1}\frac{\Omega_k}{\sqrt{\tau_k}}\right)^2 + \bar{D}_{\rm BH}\Omega_{K+1}\right],$$

$$\text{(SP3)}\ \text{s.t.}\ \sum_{f=1}^{F}|\mathcal{Y}_f| \le C, \tag{51a}$$

$$\mathcal{Y}_f \subseteq \mathcal{X}_f,\ \forall f \in \mathcal{F}, \tag{51b}$$

where the group load distribution is given by

$$\Omega_k = \sum_{f=1}^{F}\frac{q_f}{s_f}\left[\min\left(k|\mathcal{Y}_f|, s_f\right) - \min\left((k-1)|\mathcal{Y}_f|, s_f\right)\right], \tag{52}$$

for $k = 1, 2, \cdots, K$, and $\Omega_{K+1} = 1 - \sum_{k=1}^{K}\Omega_k$. The objective function represents the reduced delay with cooperative caching, where $\frac{\bar{S}L}{W\tau_1} + \bar{D}_{\rm BH}$ denotes the average delay without content caching. A greedy content placement algorithm is proposed to solve problem (SP3) as described in Algorithm 1, where $\mathcal{Y}_f$ denotes the set of segments in cache, c_f is the number of segments cached for file-f. Each time a segment is placed in cache until the caching capacity C is achieved, and the one providing maximal marginal gain is always selected to minimize average delay. The reformulated content placement problem (SP3) can be proved to be monotone submodular based on the result of Proposition 3. Therefore, the proposed greedy algorithm can achieve $(1 - \frac{1}{e})$-optimality [42]. The computation complexity of the proposed greedy algorithm is $\mathcal{O}(FC)$. Specifically, $\mathcal{O}(F)$ denotes the computation complexity to find the file segment which can bring the maximal marginal gain, and $\mathcal{O}(C)$ corresponds to the process that C segments are selected for caching.

5. $|\cdot|$ denotes the cardinality of a set.

TABLE 3: Simulation parameters [11]

Parameter	Value	Parameter	Value
$P_{\rm T}$	1 W	α	4
W	10 MHz	$D_{\rm BH}$	200 ms
ρ	50 /km^2	λ	500 /km^2
σ^2	-105 dBm/MHz	I_1	-75 dBm/MHz
I_2	-70 dBm/MHz	I_3	-68 dBm/MHz
F	1000	ν	1
s_f	1000	L	1000 bit

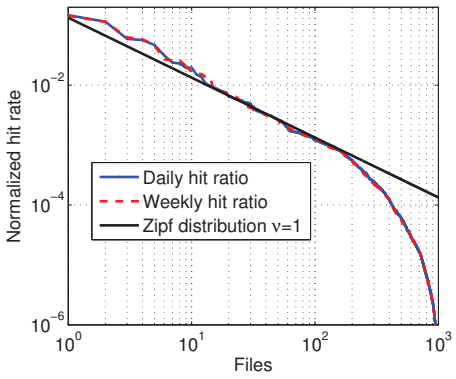


Fig. 4: YouTube video popularity illustration.

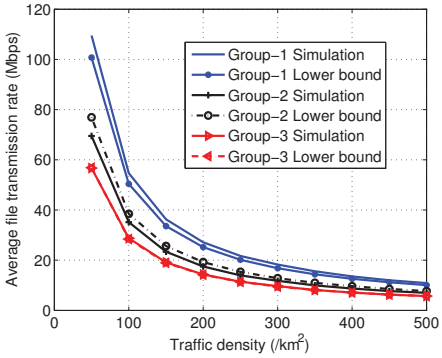


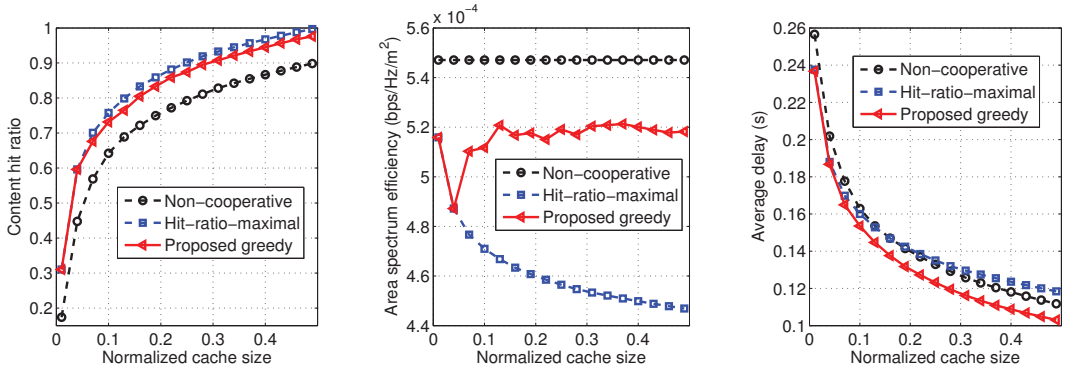
Fig. 5: Evaluation of average file transmission rate.

6 Simulation Results

In this section, we validate the obtained analytical results of average rates based on extensive system-level simulations, evaluate the performance of the proposed greedy caching algorithm, and study the influence of system parameters as well as cluster size. Important simulation parameters are set as Table 3. Both real-world data trace and Zipf

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing



(a) Content hit ratio (b) Spectrum efficiency (c) Average delay

Fig. 6: Performance evaluation of greedy content placement.

popularity distribution are adopted. The real-world data trace is crawled from YouTube where some video owners made their video view statistics open to public, and the view amount information is recorded on a daily basis. We randomly crawled 1000 videos on May 2017. The most popular video has been watched over 9 million times during the May of 2017, while the least popular one has been rarely viewed. The view amounts (in both daily and weekly scales) of the 1000 videos are normalized to represent the hit ratio, illustrated as Fig. 4. Although the popularity of a specific file may change with time, the overall popularity distribution of the whole file library is shown to be stable in both daily and weekly scales. We adopt the normalized daily content popularity distribution to evaluate the proposed greedy content placement algorithm. The Zipf distribution has been widely adopted to model the content popularity distribution [43]:

$$q_f = \frac{1/f^{\nu}}{\sum_{h=1}^{F} 1/h^{\nu}}, \tag{53}$$

where $\nu \geq 0$ represents the skewness of popularity distribution, and a larger ν corresponds to more concentrated file requests. Thus, we vary the skewness parameter ν in simulation to depict diverse applications and contents. As shown in Fig. 4, the normalized content hit rate of the top 10% of real trace YouTube videos can be approximated by the Zipf distribution with $\nu = 1$. The other files can be outdated ones and rarely requested, causing the mismatch. As these files do not need to be considered for edge caching, the Zipf distribution can be applied to model the popularity distribution of real-world data trace.

6.1 Analytical Results Evaluation

The derived lower bound of average file transmission rates for radio access is validated, shown as Fig. 5, where Group-k in the legend represents the results when users are associated with the kth candidate SBS. The results of lower bound is based on Eqn. (9). The simulation results are calculated based on the Monte Carlo method, whereby the SBS topology, user locations, and channel fading are generated randomly according to the corresponding probability distribution functions. Fig. 5 shows that the derived lower bounds of average transmission rate are quite close to the simulation results for different groups of users, and both decrease with the traffic density due to the limited radio resources. Therefore, Lemma 1 is validated, and the derived lower bound can be applied to approximate the average file transmission rate for theoretical analysis.

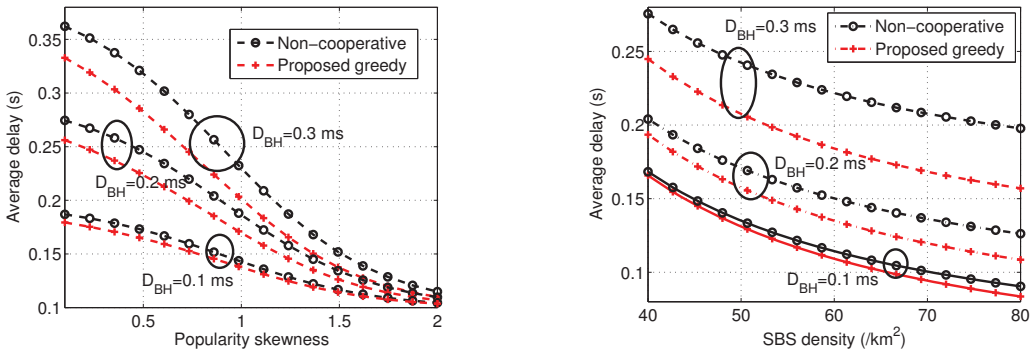
6.2 Tradeoff between Content Diversity and Spectrum Efficiency

Fig. 6 demonstrates the tradeoff between content diversity and spectrum efficiency, where different caching schemes are compared in terms of content hit ratio, spectrum efficiency and average delay. The cluster size is set as 2, and the daily YouTube video hit ratio is used. Two caching schemes are adopted as benchmarks. Under the non-cooperative scheme, users are only served by the nearest SBS to achieve the high transmission rate, where each SBS cache s_f segments of the C/s_f most popular files. On the contrary, the hit-ratio-maximal scheme exploits caching diversity to maximize local content hit ratio, where each SBS cache s_f/K segments of the CK/s_f most popular files. In fact, maximizing the content hit ratio has been considered as a key objective of edge caching, which is favorable in dealing with backhaul congestion and cache size limitations [21], [32].

As shown in Figs. 6 (a) and (b), the hit-ratio-maximal scheme always achieves the highest content hit ratio with the lowest spectrum efficiency. Specifically, the content hit ratio increases with the cache size, whereas the spectrum efficiency decreases. The reason is that more users fetch file segments from the cache of farther SBSs as cache size increases, introducing larger wireless transmission delay. On the contrary, the non-cooperative caching scheme is shown to maintain the highest spectrum efficiency but the content hit rate is the lowest, regardless of the cache size. This result is rational since all users are only served by the home SBSs, which can provide maximal average transmission rate. Compared with the other two schemes, the proposed greedy scheme is shown to balance content hit rate and spectrum efficiency under different cache sizes. Fig. 6 (c) demonstrates that the average file transmission delay decreases with cache size, under all the three schemes. The greedy scheme presents the minimal

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing



(a) Content popularity (b) Small cell density

Fig. 7: Influence of system parameters.

transmission delay, while the hit-ratio-maximal algorithm performs better than the non-cooperative scheme only when the cache size is smaller than some threshold.

Figs. 6 (a), (b) and (c) are consistent with the analysis. Specifically, the proposed greedy scheme balances content diversity and spectrum efficiency to minimize the average file transmission delay. When the cache size is small, enhancing content diversity is more important due to backhaul congestion, and the proposed greedy scheme is equivalent to the hit-ratio-maximal scheme. When the cache size further increases, the backhaul congestion has been significantly relieved with local caching, and increasing content hit ratio through traffic steering is no longer advantageous considering the degraded spectrum efficiency. In this case, the proposed greedy scheme maintains the spectrum efficiency at certain level, by slightly sacrificing the content hit rate.

6.3 Impact of System Parameters

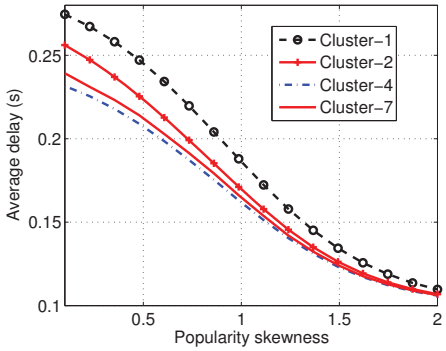
To evaluate the performance of the proposed greedy scheme under different traffic demands and use scenarios, we further study the influences of key system parameters, including content popularity distribution, network density, and backhaul delay, illustrated as Fig. 7. The Zipf popularity distribution is adopted in Fig. 7 (a), where popularity skewness varies to represent different applications. Fig. 7 (b) studies the influence of SBS density based on daily trace, corresponding to different network scenarios like dense urban or sparse rural networks. The results reveal three facts. Firstly, the performance gain of greedy algorithm decreases with the content popularity skewness, shown as Fig. 7 (a). This is because that the content hit ratio gain brought by cooperative caching degrades when the content requests are more concentrated. Secondly, the performance gain of greedy algorithm increases with cell density, shown as Fig. 7 (b). As the network gets more densified, the cost of traffic steering decreases due to reduced transmission distance and path loss, enhancing the benefit of cooperative caching. Thirdly, the greedy algorithm is even more advantageous when backhaul delay is high, as shown in both Figs. 7 (a) and (b). This result actually indicates the effectiveness of cooperative caching on reducing backhaul transmissions. With the three observed facts, we can conclude that the proposed greedy algorithm can be more effective to cache the less concentrated contents in denser networks with higher backhaul delay.

6.4 Optimal Cluster Size

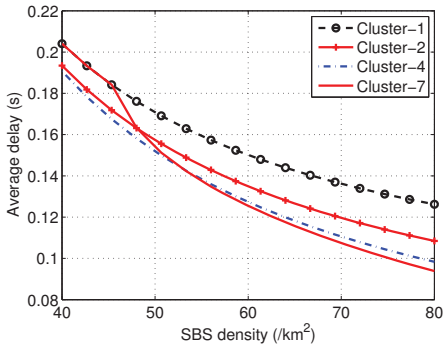
As analyzed in previous context, the cluster size also influences the content diversity and spectrum efficiency. Fig. 8 reveals the system performance with different cluster sizes. As shown in Fig. 8(a), the average delay can be minimized when the cluster size is 4. In addition, when the cluster size exceeds 7, the average delay decreases with the cluster size, indicating that the benefit of cooperative caching will vanish if the cluster size is too large. In this case, it is optimal to store the most popular files as a whole, to avoid users served by SBSs farther away. The simulation results are consistent with the analytical ones. In fact, the cluster size also influences the tradeoff between caching diversity and spectrum efficiency. As cluster size increases, users can fetch files from more SBSs, which increases caching diversity whereas degrades spectrum efficiency. Accordingly, the optimal cluster size should balance the tradeoff to minimize average transmission delay. The optimal cluster size also depends on the SBS density, shown as Fig. 8(b). Specifically, larger cluster sizes are more advantageous in denser networks, due to the reduced transmission distance and path loss on average. For example, the optimal cluster size is shown to be 4 when the SBS density is smaller than 54 /km^2, but becomes 7 when the SBS density further increases. Consider another example when cluster size equals to 7. The average delay is shown to be no smaller than that of the non-cooperative caching when the SBS density is smaller than 45 /km^2. In this case, the condition of Proposition 2 (i.e., Eqn. (25)) cannot be satisfied, as the caching diversity gain is overwhelmed by the spectrum efficiency degradation due to the long transmission distance. However, the average delay decreases significantly when the SBS diversity exceeds 45 /km^2. Then, the condition of Proposition 2 holds with reduced spectrum efficiency degradation, and thus cooperative caching improves delay performance. Notice that the optimal cluster size also increases with the backhaul delay, shown as Fig. 9. The reason is that increasing content hit ratio is more advantageous when

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing



(a) With respect to content popularity



(b) With respect to small cell density

Fig. 8: Performance with different cluster size K.

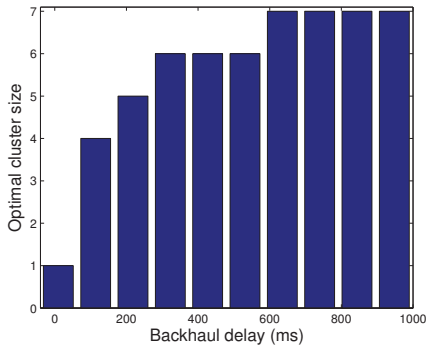


Fig. 9: Optimal cluster size with respect to backhaul delay.

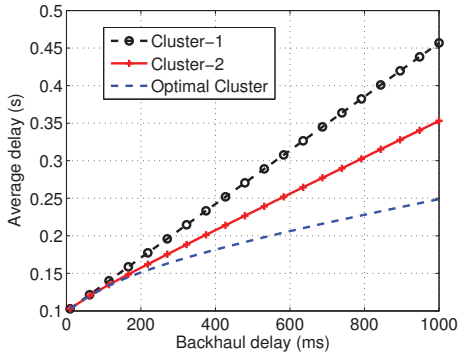


Fig. 10: Delay enhancement with optimal clustering (cluster size denoted as K).

backhaul delay is higher, which can be realized by increasing cluster size.

Fig. 10 further demonstrates the effectiveness of optimizing cluster size. The average transmission delay increases with the backhaul delay regardless of cluster size, whereas the increasing rates are different. When the cluster size is a constant, the average transmission delay is shown to increase almost linearly with backhaul delay. When the cluster size is optimized, the average transmission delay increases sub-linearly. For illustration, compared with the non-cooperative caching (cluster size set to 1), the average delay can be reduced by around 25% and 45% through cluster size optimization, when the backhaul delay is 400 ms and 1 s, respectively. The important insights for application is that the cluster size should be adjusted based on the system parameter and status. For instance, when the backhaul is congested during rush hours, the user-centric cluster size should enlarge to increase content hit ratio and reduce backhaul pressure. Instead, when the traffic load decreases at midnight, users can just fetch files from home SBSs with the cluster size shrink to 1.

7 Conclusions and Future Work

In this paper, the cooperative edge caching has been investigated in large-scale user-centric clustered mobile networks, aiming at minimizing the average file transmission rates with caching size and bandwidth constraints. Based on the optimal bandwidth allocation obtained by the Lagrange multiplier method, a linear-complexity greedy content placement algorithm has been proposed with guaranteed performance. In addition, an explicit condition constraining the maximal cluster size has been obtained, which offers a guideline for user-centric clustering in practical networks. The results of the optimal content placement and SBS clustering have both revealed the tradeoff relationship between content diversity and spectrum efficiency with cooperative edge caching. For our future work, we will investigate the impact of user mobility and unknown content popularity on mobile edge caching. Furthermore, it is also interesting to design advanced cooperative caching scheme with instant information of cell load and channel condition in heterogeneous networks with diversified backhaul capacity.

References

[1] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Traffic steering assisted mobile edge caching: Exploiting spatial content diversity gain," in *IEEE GLOBECOM'17*, Singapore, Dec. 2017.

[2] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, Jun. 2015.

[3] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov. 2014.

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

[4] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of green information-centric networking: Research issues and challenges." *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1455–1472, Feb. 2015.

[5] K. Poularakis and L. Tassiulas, "Code, cache and deliver on the move: A novel caching paradigm in hyper-dense small-cell networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 675–687, Mar. 2017.

[6] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[7] 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.

[8] P. Yang, N. Zhang, Y. Bi, L. Yu, and X. S. Shen, "Catalyzing cloud-fog interoperation in 5G wireless networks: An SDN approach," *IEEE Netw.*, vol. 31, no. 5, pp. 14–20, Sep. 2017.

[9] K. Wang, F. R. Yu, and H. Li, "Information-centric virtualized cellular networks with device-to-device communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9319–9329, Nov. 2016.

[10] "Mobile-edge computing - introductory technical white paper," European Telecommunications Standards Institute, Tech. Rep., Sep. 2014, accessed Mar. 27, 2017. [Online]. Available: https://portal.etsi.org/

[11] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[12] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "On the delay of geographical caching methods in two-tiered heterogeneous networks," in *IEEE SPAWC'16*, Edinburgh, UK, Aug. 2016, pp. 1–6.

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

[14] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 1, pp. 77–89, Jan. 2006.

[15] C.-Y. Chow, H. V. Leong, and A. T. Chan, "GroCoca: Group-based peer-to-peer cooperative caching in mobile environment," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 1, pp. 179–191, Jan. 2007.

[16] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: challenges, methodologies, and directions," *IEEE Commun. Mag.*, vol. 23, no. 2, pp. 78–85, Apr. 2016.

[17] W. Bao and B. Liang, "Stochastic geometric analysis of handoffs in user-centric cooperative wireless networks," in *IEEE INFOCOM'16*, San Francisco, USA, Apr. 2016, pp. 1–9.

[18] W. Song and W. Zhuang, "Performance analysis of probabilistic multipath transmission of video streaming traffic over multi-radio wireless devices," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1554–1564, Apr. 2012.

[19] W. Nie, F. C. Zheng, X. Wang, W. Zhang, and S. Jin, "User-centric cross-tier base station clustering and cooperation in heterogeneous networks: Rate improvement and energy saving," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1192–1206, May 2016.

[20] J. Song, H. Song, and W. Choi, "Optimal caching placement of caching system with helpers," in *IEEE ICC'15*, London, UK, Jun. 2015, pp. 1825–1830.

[21] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[22] J. He and W. Song, "Optimizing video request routing in mobile networks with built-in content caching," *IEEE Trans. Mobile Comput.*, vol. 15, no. 7, pp. 1714–1727, Jul. 2016.

[23] 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.

[24] X. Li, X. Wang, and V. C. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[25] 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.

[26] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[27] T. X. Tran and D. Pompili, "Octopus: A cooperative hierarchical caching strategy for cloud radio access networks," in *IEEE MASS'13*, Brasilia, Brazil, Oct. 2016, pp. 154–162.

[28] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1382–1393, May 2017.

[29] F. Zhang, C. Xu, Y. Zhang, K. K. Ramakrishnan, S. Mukherjee, R. Yates, and T. Nguyen, "Edgebuffer: Caching and prefetching content at the edge in the mobilityfirst future internet architecture," in *IEEE WoWMoM'16*, Boston, MA, USA, June 2015, pp. 1–9.

[30] X. Vasilakos, V. A. Siris, and G. C. Polyzos, "Addressing niche demand based on joint mobility prediction and content popularity caching," *Computer Networks*, vol. 110, pp. 306–323, 2016.

[31] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Oct. 2016.

[32] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," in *IEEE WCNC'17*, San Francisco, USA, Mar. 2017, pp. 1–7.

[33] S. Zhang, N. Zhang, P. Yang, and X. S. Shen, "Cost-effective cache deployment in mobile heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017, DOI: 10.1109/TVT.2017.2724547, to appear.

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

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

[36] M. Ismail, A. Abdrabou, and W. Zhuang, "Cooperative decentralized resource allocation in heterogeneous wireless access medium," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 714–724, Feb. 2013.

[37] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[38] D. J. MacKay, "Fountain codes," *IEEE Proc. Commun.*, vol. 152, no. 6, pp. 1062–1068, Dec. 2005.

[39] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[40] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "A distributed numerical approach to interference alignment and applications to wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3309–3322, Jun. 2011.

[41] E. W. Weisstein, "Euler-Mascheroni constant," 2002, accessed Oct. 10, 2017. [Online]. Available: http://mathworld.wolfram.com/Euler-MascheroniConstant.html

[42] A. Krause and D. Golovin, *Submodular function maximization*. Cambridge University Press, Feb. 2014.

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

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2780834, IEEE Transactions on Mobile Computing

Shan Zhang (S'13-M'16) received her Ph.D. degree in Department of Electronic Engineering from Tsinghua University and B.S. degree in Department of Information from Beijing Institute Technology, Beijing, China, in 2016 and 2011, respectively. She is currently a post doctoral fellow in Department of Electronical and Computer Engineering, University of Waterloo, Ontario, Canada. Her research interests include resource and traffic management for green communication, intelligent vehicular networking, and software defined networking. Dr. Zhang received the Best Paper Award at the Asia-Pacific Conference on Communication in 2013.

Peter He (S'14-M'15) received the M.A.Sci. degree in electrical engineering from McMaster University, Hamilton, ON, Canada, in 2009, and the Ph.D. degree in electrical engineering from Ryerson University, Toronto, ON, Canada, in 2015. Since then, he has been a Post-Doctoral Research Fellow with the Broadband Communications Research Group Laboratory, University of Waterloo, Waterloo, ON, Canada, and Ryerson University. His current research interests include union of large-scale optimization, information theory, communications, and smart power grid. Dr. He was a recipient of the Graduate Research Excellence Award from the Department of Electrical and Computer Engineering, Ryerson University, the Ontario Graduate Scholarship awards, the Natural Sciences and Engineering Research Council Post-Doctoral Fellowship in 2016, and the Best Paper Award of the 2013 International Conference on Wireless Communications and Signal Processing.

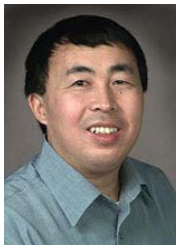
Katsuya Suto (M'16) received the B.Sc. degree in computer engineering from Iwate University, Morioka, Japan, in 2011, and M.Sc. and Ph.D. degrees in information science from the Tohoku University, Japan, in 2013 and 2016, respectively. He is currently a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include resilient networking, software defined networking, and mobile cloud computing. He received the Best Paper Award at the IEEE 79th Vehicular Technology Conference in 2013, the IEICE Academic Encouragement Award in 2014, the IEEE VTS Japan 2015 Young Researchers Encouragement Award, the Best Paper Award at the IEEE/CIC International Conference on Communications in China in 2015, and the Best Paper Award at the IEEE International Conference on Communications in 2016.

Peng Yang received his B.E. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2013. Currently, he is pursuing his Ph.D. degree in the School of Electronic Information and Communications, HUST. From Sep. 2015, he is also a visiting Ph.D. student in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His current research interests include next generation wireless networking, software defined networking and fog computing.

Lian Zhao (S'99-M'03-SM'06) received the Ph.D. degree from the Department of Electrical and Computer Engineering (ELCE), University of Waterloo, Canada, in 2002. She joined the Department of Electrical and Computer Engineering at Ryerson University, Toronto, Canada, in 2003 and a Professor in 2014. Her research interests are in the areas of wireless communications, radio resource management, power control, cognitive radio and cooperative communications, optimization for complicated systems.

She received the Best Land Transportation Paper Award from IEEE Vehicular Technology Society in 2016; Top 15 Editor in 2015 for IEEE Transaction on Vehicular Technology; Best Paper Award from the 2013 International Conference on Wireless Communications and Signal Processing (WCSP) and Best Student Paper Award (with her student) from Chinacom in 2011; the Ryerson Faculty Merit Award in 2005 and 2007; the Canada Foundation for Innovation (CFI) New Opportunity Research Award in 2005, and Early Tenure and promotion to Associate Professor in 2006. She has been an Editor for IEEE TRANSACTION ON VEHICULAR TECHNOLOGY since 2013; workshop co-chair for IEEE/CIC ICCC 2015; local arrangement co-chair for IEEE Infocom 2014; co-chair for IEEE Global Communications Conference (GLOBECOM) 2013 Communication Theory Symposium.

She served as a committee member for NSERC (Natural Science and Engineering Research Council of Canada) Discovery Grants Evaluation Group for Electrical and Computer Engineering since 2015; an Associate Chair at the Department of Electrical and Computer Engineering at Ryerson University 2013-2015. She is a licensed Professional Engineer in the Province of Ontario, a senior member of the IEEE Communication and Vehicular Society.



Xuemin (Sherman) Shen (M'97-SM'02-F'09) received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is also the Associate Chair for Graduate Studies. Dr. Shens research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is an elected member of IEEE ComSoc Board of Governor, and the Chair of Distinguished Lecturers Selection Committee. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom16, Infocom14, IEEE VTC10 Fall, and Globecom07, the Symposia Chair for IEEE ICC10, the Tutorial Chair for IEEE VTC11 Spring and IEEE ICC08, the General CoChair for ACM Mobihoc15, Chinacom07 and QShine06, the Chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/served as the Editor-in-Chief for IEEE Network, Peer-to-Peer Networking and Application, and IET Communications; an Associate Editor-in-Chief for IEEE Internet of Things Journal, a Founding Area Editor for IEEE Transactions on Wireless Communications; an Associate Editor for IEEE Transactions on Vehicular Technology, Computer Networks, and ACM/Wireless Networks, etc.; and the Guest Editor for IEEE JSAC, IEEE Wireless Communications, IEEE Communications Magazine, and ACM Mobile Networks and Applications, etc. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo, the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, 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.

1536-1233 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.