Collaborative Multi-Tier Caching in Heterogeneous Networks: Modeling, Analysis, and Design

Xiuhua Li, *Student Member, IEEE*, Xiaofei Wang, *Member, IEEE*, Keqiu Li, *Senior Member, IEEE*, Zhu Han, *Fellow, IEEE*, and Victor C. M. Leung, *Fellow, IEEE*

***Abstract*— To deal with the explosive growth in multimedia service requests in mobile networks, caching contents at the cells (base stations) is regarded as an effective emerging technique to reduce the duplicated transmissions of content downloads, while heterogeneous networks (HetNets) are regarded as an effective technique to increase the network capacity. Yet, the combination of content caching and HetNets for future networks (i.e., 5G) is still not well explored. In this paper, we propose an efficient collaborative multi-tier caching framework in HetNets. In particular, based on patterns of user requests, link capacities, heterogenous cache sizes, and the derived system topology, we focus on exploring the maximum capacity of the network infrastructure so as to offload the network traffic and support users' content requests locally. Due to the NP-hardness of the complex multi-tier caching problem, we approximately decompose it into some subproblems that focus on the caching cooperation at different tiers by utilizing the derived system topology. Our proposed framework is low-complexity and distributed, and can be used for practical engineering implementation. Trace-based simulation results demonstrate the effectiveness of the proposed framework.**

***Index Terms*— Content caching, traffic load, heterogeneous networks, time complexity.**

I. Introduction

OVER recent years, the demands for multimedia services in mobile networks have been explosively growing, and the induced enormous amount of network traffic load has become a serious concern of mobile network operators (MNOs), especially on radio access networks and backhaul networks [1], [2]. The growth motivates changes to the operations of mobile networks to introduce revolutionary schemes involving new mobile network architectures and advanced data transmission technologies for future networks (i.e., 5G) [3]–[6].

One emerging technique to address the above challenge is to cache popular contents at base stations (BSs) to bring contents closer to users. Caching at BSs can effectively offload the network traffic caused by massive duplicated content downloads from service providers (SPs) over the Internet via backbone networks and cellular links, and improve user quality of service (QoS) such as access delay of contents [2]–[6].

Another effective way of evolving to a more capable mobile networks is to introduce the architecture of heterogeneous networks (HetNets) to bring networks with good link quality closer to users. In a HetNet, there are densely deployed nodes such as macro BSs, micro BSs, pico BSs, femto BSs and relays, forming a multi-tier network architecture. Due to the reduction in the distances between users and their associated BSs, HetNets can effectively improve the area spectral efficiency as well as network capacity. However, the exponential growth in network traffic also requires high-capacity backhaul for the connection of different types of BSs and SPs [7].

Considering the great potentials of the above two techniques, it is beneficial to combine them together, i.e., content caching in HetNets, so as to effectively reduce the explosively growing network traffic load. Due to the limitations of network resources and the practical scale of popular contents, it is impractical to cache all the popular contents inside the network [1]–[7]. Therefore, it is crucial to design proper caching schemes to effectively utilize the network infrastructures.

Though the idea of multi-tier (or hierarchical) caching has been widely used to achieve its benefits such as in web caching systems [8]–[10] and IPTV systems [11], there exists few works in the literature that address the challenges with employing multi-tier caching in mobile networks. In this paper, we are motivated to explore the collaborative multi-tier content caching in HetNets based on the derived topology. Our multi-tier caching topology in HetNets is similar to the widely used hierarchical structures in [8]–[11]. However, the collaborative multi-tier caching problem that we explore in this paper have serval different features. One important feature that distinguishes our work from similar problems is that users are possible to send content requests at any tier in our network topology due to the association between users and BSs at different tiers, while user requests in [8]–[11] are only received at the bottom tier. This feature on user

Manuscript received November 28, 2016; revised May 21, 2017; accepted July 26, 2017. Date of publication August 4, 2017; date of current version October 9, 2017. This work was supported in part by the China Scholarship Council Four Year Doctoral Fellowship, in part by the Canadian NSERC under Grant RGPIN-2014-06119 and Grant RGPAS-462031-2014, and in part by the U.S. National NSF Grants of CNS-1717454, CNS-1731424, CNS-1702850, CNS-1646607, ECCS-1547201, CMMI-1434789, CNS-1443917, and ECCS-1405121. The associate editor coordinating the review of this paper and approving it for publication was T. Taleb. *(Corresponding author: Xiaofei Wang.)*

X. Li and V. C. M. Leung are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: lixiuhua@ece.ubc.ca; vleung@ece.ubc.ca).

X. Wang and K. Li are with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: xiaofeiwang@tju.edu.cn; likeqiu@gmail.com).

Z. Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea (e-mail: zhan2@uh.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2017.2734646

request patterns makes our problem more challenging to design caching strategies. Besides, compared with the web caching systems [8]–[10] where link capacity among web caching servers has little impact and is rarely regarded as the bottleneck, our considered system in HetNets needs to consider the bandwidth consumed for massive content delivery due to the limited-capacity fronthaul/backhaul. Moreover, compared with the user association via wired links in IPTV systems [11], the users in HetNets are associated with BSs at each tier via wireless links that is more difficult to handle due to wireless channel characteristics.

In this paper, we propose an efficient collaborative multi-tier caching framework in HetNets so as to offload the network traffic and support users' content requests locally considering some practical constraints such as user request patterns, link capacities, heterogenous cache sizes and the derived system topology. Our contributions are indicated as follow:

- This is the first study for integrating issues of mobile user link quality and cell association together with tiered collaborative cell caching in HetNets, for practically reducing duplicated in-network traffic as our major objective, towards the future mobile networks;
- We decompose the sophisticated caching problem into subproblems of inter-tier and intra-tier collaboration, and propose the corresponding distributed algorithms with low complexity from the perspective of engineering implementation as well as a content request routing scheme;
- Together with theoretical analysis, numerical simulation and realistic trace-based evaluation, our proposed framework is proved to be with excellent caching performance along with effective tiered collaboration for offloading duplicated traffic significantly while satisfying most demands of mobile users.

The remainder of this paper is organized as follow. Sec. II studies the related work. Sec. III discusses the infrastructure modeling of multi-tier caching. Sec. IV introduces the collaborative caching framework and the corresponding strategies. Sec. V evaluates the performance of the proposed caching framework. Finally, Sec. VI concludes the paper.

II. Related Work

There have been several studies focusing on content caching at BSs in mobile networks. For instance, the surveys in [2] and [12]–[19] explored the potentials of caching in mobile networks. The proposed cell caching in [1], [5], and [20], FemtoCaching in [14] and [21] and AMVS-NDN in [22] focused on cooperative caching popular contents at small BSs to offload the network traffic. The studies in [6] and [23]–[28] proposed the strategies of collaborative caching at BSs to improve users' QoS especially on access delay. The studies in [29]–[33] focused on the analysis of energy efficiency and the design of the corresponding energy-efficient caching schemes in wireless networks. References [34]–[36] proposed wireless content delivery schemes from BSs to users. However, these works only focus on the case of single-tier caching.

In practice, multi-tier caching has been widely used to explore the potentials of the system infrastructures especially in web caching systems [8]–[10] and IPTV systems [11]. However, there exist only few works utilizing the idea of multi-tier caching in mobile networks with hierarchical structures, e.g., HetNets. For instance, the survey in [37] discussed the caching paradigm in two-tier small cell networks based on user social structures. Reference [7] focused on the theoretical performance analysis on content caching in a three-tier HeNets where content sizes were assumed to be identical. However, [7], [37] did not involve the design of caching strategies with practical considerations on some constraints (e.g., limited-capacity fronthaul/backhaul, diversity of content sizes) and specific characteristics of the network topology. In contrast, our work focuses on the modeling, analysis and design of multi-tier caching in HetNets with the practical considerations, which can be applied in practical deployments.

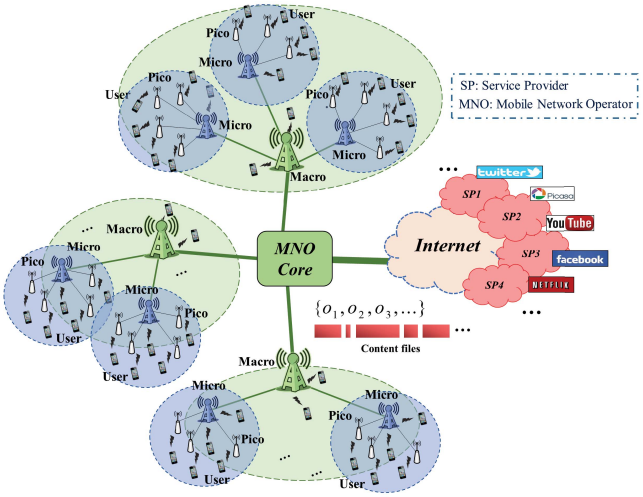


Fig. 1. Illustration of multi-tier caching architecture in HetNets.

III. Infrastructure Modeling

In this section, we introduce and model the infrastructure of multi-tier caching in HetNets.

A. Multi-Tier Caching Architecture

An illustration of the multi-tier caching architecture in a HetNet is shown in Fig. 1. Outside the MNO network, there are some SPs (e.g., YouTube, Facebook, and so on) offering the content files over the Internet while inside the MNO network, there are a great number of macro cells covering the whole service area. In each macro cell, there are several micro cells covering most of the macro cell's area. Besides, the area of each micro cell is mostly covered by several pico cells. Moreover, through limited-capacity cables or optical fibers, each macro BS is connected with its several geographically distributed micro BSs while each micro BS is connected with its several geographically distributed pico BSs. If two BSs at the same tier are connected through limited-capacity cables or optical fibers and thus have direct link, they are defined as neighboring BSs. Due to short geographical distances, a pico BS may be connected with several neighboring pico BSs in the same micro cell, while a micro BS may also be connected with several neighboring micro BSs in the same

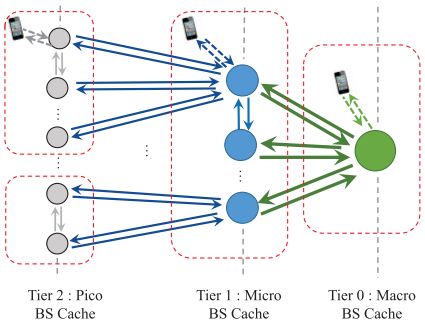


Fig. 2. Topology of multi-tier caching in HetNets.

macro cell. Mobile users are heterogenous due to various channel conditions, and consist of macro users, micro users and pico users which are associated with macro BSs, micro BSs and pico BSs, respectively. Mobile users' content requests are received and served by their associated BSs.

Specifically, all the macro BSs, micro BSs and pico BSs in the HetNet are able to cache some contents with limited cache storage capacity, in order to satisfy users' content requests and reduce the duplicated content transmission and thus network traffic load. The caching functionality of the macro BSs, micro BSs and pico BSs are hierarchical, which forms a three-tier caching topology as shown in Fig. 2. Here, the caching at macro BSs, micro BSs and pico BSs are defined as Tier 0, Tier 1 and Tier 2, respectively.

To satisfy a dynamic content request from a user at any tier in the HetNet, the associated local BS either returns the content if it is locally available, or routes the request to other BSs. The details of the content request routing scheme will be provided in the following section. However, to achieve a distributed caching mechanism, micro BSs in a macro cell can decide how to effectively select popular contents to cache at their connected pico BSs, and can properly update the pico BSs' contents. Besides, a micro BS maintains a list of all the connected pico BSs' cached contents at the cost of a small traffic overhead that can be neglected. In a similar way, the caching at micro BSs is managed by the connected macro BS. In other words, the caching at Tier 2 is controlled by micro BSs while the caching at Tier 1 is controlled by macro BSs. Besides, the caching at Tier 0 is controlled by the MNO core.

Moreover, we assume that the content popularity changes slowly. For instance, short-lifetime popular news with short videos are updated every a few hours, while long-lifetime new movies and new music videos are, respectively, posted weekly and monthly. To reduce the traffic load and avoid possible network traffic congestion especially in busy hours, popular contents especially for long-lifetime contents can be cached in peak-off hours (e.g., late night) [6]. Since the content popularity can be regarded as fixed in a relatively long time, the cost of updating the contents in all the BSs can be neglected. In terms of the content popularity, it can be obtained in advance or predicted by the system learning and analysis from the user behavior and preference [21], and thus we assume that the content popularity is available in the network.

B. System Modeling

In the system model, we consider the case of a single macro cell, equipped with a macro BS, M micro BSs and N_m pico BSs in the m-th micro cell to serve a great number of users. The caching storage sizes of the macro BS, micro BSs and pico BSs are denoted by S_0, $(S_m)_{M\times 1}$ and $\{(S_{1n})_{N_1\times 1}, (S_{2n})_{N_2\times 1}, \ldots, (S_{Mn})_{N_M\times 1}\}$, respectively. Denote $\delta_{kn}^m \in \{0,1\}$ for whether pico BS$_{mk}$ and pico BS$_{mn}$ are neighboring, and $\delta_{tm} \in \{0,1\}$ for whether micro BS$_t$ and micro BS$_m$ are neighboring. As shown in Fig. 2, we assume that the links between any pair of neighboring BSs at the same tier are two-directional connected, i.e., $\delta_{kn}^m = \delta_{nk}^m$ and $\delta_{tm} = \delta_{mt}$, and all the values of (δ_{kn}^m) and (δ_{tm}) are known once the network topology is given. Besides, we set $\delta_{nn}^m = 1$ and $\delta_{mm} = 1$. To simplify the notations, we assume that the maximum available fronthaul link capacity and backhaul link capacity between any two BSs are identical. Denote C_{kn}^m as the maximum available link capacity between pico BS$_{mk}$ and pico BS$_{mn}$, C_n^m as the maximum available link capacity between micro BS$_m$ and pico BS$_{mn}$, C_{tm} as the maximum available link capacity between micro BS$_t$ and micro BS$_m$, and C_m as the maximum available link capacity between micro BS$_m$ and the macro BS.

There is a catalog of F popular contents in the system, denoted by $(o_f)_{F\times 1}$. From the practical perspective, the sizes of all the contents are assumed to be various, denoted by $(s_f)_{F\times 1}$. Here, a content is assumed to be either entirely cached or not cached. The caching strategy at Tier 2 is denoted by a series of binary variables, i.e., $\{(x_{1n}^f)_{N_1\times F}, (x_{2n}^f)_{N_2\times F}, \ldots, (x_{Mn}^f)_{N_M\times F}\}$, where $x_{mn}^f = 1$ means caching content o_f at pico BS$_{mn}$ while $x_{mn}^f = 0$ means no caching. The caching strategy at Tier 1 is also denoted by a series of binary variables, i.e., $(x_m^f)_{M\times F}$, where $x_m^f = 1$ means caching content o_f at micro BS$_m$ while $x_m^f = 0$ means no caching. Denote λ_{mn}, λ_m and λ_0 as the average arrival rates of content requests at pico BS$_{mn}$, micro BS$_m$ and the macro BS, respectively. Note that the content requests arrived at a micro BS are either generated by micro users or routed from the connected pico BSs or neighboring micro BSs, and the content requests arrived at the macro BS are in a similar way. Thus, denote β_m and β_0 as the average arrival rates of content requests from micro users to micro BS$_m$ and from macro users to the macro BS, respectively. Besides, we define the average arrival rate of the requests for content o_f at pico BS$_{mn}$ and micro BS$_m$ as λ_{mn}^f and λ_m^f, respectively. As well, we denote β_m^f and β_0^f as the average arrival rates of the request for content o_f from micro users to micro BS$_m$ and from macro users to the macro BS, respectively. The popularity and normalized popularity of content o_f in the n-th pico cell of the m-th micro cell are represented by p_{mn}^f and r_{mn}^f, respectively. As well, the popularity of content o_f requested by micro users in the m-th micro cell and macro users in the macro cell are q_m^f and q_0^f, respectively. Accordingly, we have

$$r_{mn}^f = \frac{p_{mn}^f}{\sum_{i=1}^F p_{mn}^i} \text{ and } \lambda_{mn}^f = r_{mn}^f \lambda_{mn}, \quad \forall m,\ \forall n,\ \forall f, \quad (1)$$

$$\beta_m^f = \frac{q_m^f \beta_m}{\sum_{i=1}^F q_m^i} \text{ and } \beta_0^f = \frac{q_0^f \beta_0}{\sum_{i=1}^F q_0^i}, \quad \forall m, \ \forall f. \tag{2}$$

Moreover, following [38], we assume that the overall popularity of the contents in the network, denoted as $(P_f)_{F\times 1}$, satisfies the Mandelbrot-Zipf (MZipf) distribution. Specifically, we have

$$P_f = \frac{(rank_f + q)^{-\beta}}{\sum_{i=1}^F (rank_i + q)^{-\beta}}, \quad \forall f, \tag{3}$$

where $rank_f$ is the rank of the content o_f in the descending order of content popularity, $q \geq 0$ is the plateau factor, and $\beta > 0$ is the skewness factor. In particular, larger value of q leads to a more flattened head of the distribution, while larger value of β leads to a smaller number of most popular contents accounting for a majority of content requests. Clearly, we have $P_f = \sum_{m=1}^M \sum_{n=1}^{N_m} p_{mn}^f + \sum_{m=1}^M q_m^f + q_0^f$ for each content.[1] Besides, the content sizes follow a Pareto distribution [39].

C. User Association and Content Request Modeling

By using the HetNet system model as in [40], we assume that the considered three tiers of BSs are distinguished by their spatial density, transmit power and supported data rate. Specifically, the locations of BSs at Tier k ($k = 0, 1, 2$) follow a homogeneous Poisson point process (PPP) Φ_k with intensity φ_k. Each tier has the same path loss exponent denoted by α. Each of the BSs at Tier k uses the same transmit power Γ_k, and has the same signal-to-interference-plus-noise ratio (SINR) threshold Υ_k. The random channel fluctuations are modeled as Rayleigh fading with unit average power. Here, an open access strategy is employed, i.e., a user is allowed to connect to any tier. Besides, a user can be associated with a BS at Tier k only when its SINR w.r.t. the BS is greater than Υ_k. Moreover, users are located according to a homogeneous PPP $\Phi^{(u)}$ with intensity $\varphi^{(u)}$ that is independent of $\{\Phi_k\}$. Thus, according to the analysis in [40], we can get the average proportion of users associated with the BSs at Tier k in the open access strategy as

$$\rho_k = \frac{\varphi_k \Gamma_k^{2/\alpha} \Upsilon_k^{-2/\alpha}}{\sum_{i=0}^2 \varphi_i \Gamma_i^{2/\alpha} \Upsilon_i^{-2/\alpha}}, \quad \forall k \in \{0, 1, 2\}. \tag{4}$$

Specifically, in the considered system with a single macro cell, the ratio among the intensity of BSs in the three tiers can be approximately calculated as

$$\varphi_0 : \varphi_1 : \varphi_2 = 1 : M : \Big(\sum_{m=1}^M N_m\Big). \tag{5}$$

Moreover, since the content requests at each tier are directly proportional to the associated user number, we can get the ratio as

$$\beta_0 : \Big(\sum_{m=1}^M \beta_m\Big) : \Big(\sum_{m=1}^M \sum_{n=1}^{N_m} \lambda_{mn}\Big) = \rho_0 : \rho_1 : \rho_2. \tag{6}$$

[1]From the statistical perspective of the overall network, the overall content popularity $\{P_f\}$ denotes the ratio of the requests of a specific content in the network to the requests of all the contents in the network, while the popularity $\{p_{mn}^f\}$, $\{q_m^f\}$ and $\{q_0^f\}$ denotes the ratio of the requests of a specific content received at the corresponding BS to the requests of all the contents in the network. Besides, the content popularity can be obtained in advance or predicted by the system learning and analysis from the user behavior and preference [21].

IV. Multi-Tier Caching and Request Routing

In this section, we first investigate and decompose the problem of content caching as well as the corresponding request routing, and then analyze the caching cooperation at each tier based on the topology as well as propose the corresponding collaborative caching schemes.

A. Problem Definition and Decomposition

In this paper, we will explore the problem of content caching as well as the corresponding request routing in multi-tier HetNets. Based on user request patterns, link capacities, heterogenous cache sizes and the derived system topology, we aim to explore the maximum capacity of the network infrastructure on offloading the network traffic and satisfying the users' content requests inside the MNO network.

However, as shown in [10], [11], and [41], the content caching problem in the multi-tier caching infrastructure is NP-hard even without considering the link capacity constraints. Besides, popular contents are usually large-scale in the real-world systems, and the corresponding caching problems are hard to solve and even impossible to get the optimal solutions with centralized control. Thus, it is very important to design efficient and distributed caching schemes from the perspective of engineering implementation. To address the challenges, based on the derived system topology, we approximately decompose the complex multi-tier caching problem into some subproblems that focus on the cooperation at different tiers, and propose distributed low-complexity solutions for the subproblems. We firstly explore the direct cooperation among neighboring pico BSs at Tier 2, which is based on the practical consideration that neighboring pico BSs are connected with high-capacity links due to the short geographical distances. We secondly discuss the indirect cooperation among pico BSs at Tier 2 via the connected micro BS based on the content request routing. At last, we investigate the cooperation among micro BSs at Tier 1 as well as the content caching at the macro BS at Tier 0. We will provide efficient content caching and request routing schemes in the following discussions.

B. Direct Cooperation at Tier 2

As shown in Fig. 3, neighboring pico BSs at Tier 2 are able to disseminate contents via their direct links. To reduce the complexity and generated signaling overhead of content management at Tier 2, we assume that any content between two neighboring pico BSs is entirely disseminated or not. As similar to [11], we regard $\lambda_{mn}^f s_f$ as the bandwidth capacity requirement for the requests for content o_f at pico BS_{mn} in one time unit. The notation $\{z_{kn}^{fm}\} \in \{0, 1\}$ is used to denote the routing decision whether content o_f is directly disseminated from pico BS_{mk} to pico BS_{mn} or not. Here, we set $z_{nn}^{fm} = 0$. Besides, if pico BS_{mk} and pico BS_{mn} are not neighboring, i.e., $\delta_{kn}^m = 0$ given (k, n, m), we set $C_{kn}^m = 0$.

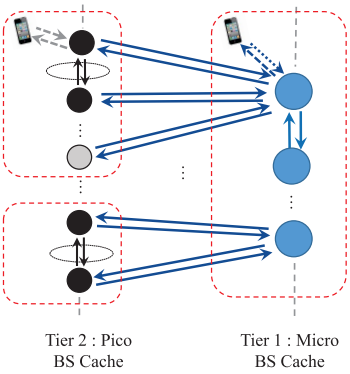


Fig. 3. Topology of direct cooperation at Tier 2. Here, both the BSs and links for cooperation are in black color.

To maximize the amount of supported traffic load at Tier 2 by using the capacity of direct links between neighboring pico BSs, the corresponding problem can be formulated as

$$\max_{\{x^f_{mn}\},\{z^{fm}_{kn}\}} \sum_{f=1}^{F}\sum_{m=1}^{M}\sum_{n=1}^{N_m} x^f_{mn}\lambda^f_{mn}s_f + \sum_{f=1}^{F}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{N_m} z^{fm}_{kn}\lambda^f_{mn}s_f \tag{7a}$$

$$s.t. \quad \sum_{f=1}^{F} x^f_{mn}s_f \le S_{mn}, \quad \forall m, \ \forall n, \tag{7b}$$

$$\sum_{f=1}^{F} z^{fm}_{kn}\lambda^f_{mn}s_f \le C^m_{kn}, \quad \forall m, \ \forall n, \ \forall k, \tag{7c}$$

$$z^{fm}_{kn} \le (\delta^m_{kn}x^f_{mk}) \oplus x^f_{mn} - x^f_{mn}, \quad \forall m, \ \forall n, \ \forall k, \ \forall f, \tag{7d}$$

$$\sum_{k=1}^{N_m} z^{fm}_{kn} \le \bigcup_{k=1}^{N_m} (\delta^m_{kn}x^f_{mk}) - x^f_{mn}, \quad \forall m, \ \forall n, \ \forall f, \tag{7e}$$

$$x^f_{mn} \in \{0,1\}, z^{fm}_{kn} \in \{0,1\}, z^{fm}_{nn} = 0, \quad \forall m, \ \forall n, \ \forall k, \ \forall f, \tag{7f}$$

where the logic calculation $\bigcup_{k=1}^{N_m} (\delta^m_{kn}x^f_{mk}) = (\delta^m_{1n}x^f_{m1}) \oplus (\delta^m_{2n}x^f_{m2}) \oplus \ldots \oplus (\delta^m_{N_m n}x^f_{mN_m}) \in \{0,1\}$. The first part and second part in (7a) denote the supported traffic load by local pico BSs and neighboring pico BSs, respectively. (7b) and (7c) denote the constraints of pico BSs' caching storage and link capacity, respectively. (7d) and (7e) denote the cooperation between neighboring pico BSs and guarantee that any content request will not be routed to neighboring pico BSs if the content is locally available. Clearly, the problem in (7) can be decomposed into M subproblems and solved separately in each micro cell.

In practice, the link capacity C^m_{kn} can be always assumed to be greater than $\sum_{f=1}^{F} z^{fm}_{kn}\lambda^f_{mn}s_f$ for $\forall m, \forall n, \forall k$ based on the practical observation in [11], i.e., the constraint in (7c) always holds. Based on the assumption, as proved in Appendix VI, the problem in (7) can be equivalent to

$$\max_{\{x^f_{mn}\}} \sum_{f=1}^{F}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\bigcup_{k=1}^{N_m} (\delta^m_{kn}x^f_{mk})\lambda^f_{mn}s_f \tag{8a}$$

$$s.t. \quad \sum_{f=1}^{F} x^f_{mn}s_f \le S_{mn}, \ x^f_{mn} \in \{0,1\}, \quad \forall m, \ \forall n, \ \forall f. \tag{8b}$$

According to our previous work [6], the logical calculation $\bigcup$ can be equivalently transformed through *Theorem 1* as follows.

Theorem 1: For a vector $\mathbf{a} \in \{0,1\}^{n\times 1}$, denote $z = \bigcup_{k=1}^{n} a_k \in \{0,1\}$. Then operation $\bigcup$ on $\mathbf{a}$ is equivalent to

$$z = \max_{k\in\{1,2,\ldots,n\}} \{a_k\}, \tag{9}$$

and is also equivalent to two linear inequality conditions as

$$a_k \le z \text{ for } \forall k \in \{1,2,\ldots,n\}, \text{ and } z \le \sum_{k=1}^{n} a_k. \tag{10}$$

Proof: See Appendix B. ∎

Based on *Theorem 1*, we can derive the property of the optimal solutions to the problem in (8) as shown in the following theorem.

Theorem 2: For any optimal solution to the problem in (8), if there exists any (f, m, n, j) such that $\sum_{k=1}^{N_m} \delta^m_{kn}x^f_{mk} > 1$ and $\delta^m_{jn}x^f_{mj} = 1$, we can always generate another optimal solution by setting $x^f_{mj} = 1$ and $x^f_{mk} = 0$ for $\forall k \ne j$ such that $\delta^m_{kn} = 1$.

Proof: See Appendix C. ∎

Theorem 2 provides an important insight that among the neighboring pico BSs with direct links, all the duplicated contents can be removed and each content can be cached at most one copy, i.e., $\sum_{k=1}^{N_m} \delta^m_{kn}x^f_{mk} \le 1$ for $\forall m, \forall n, \forall f$. Thus, all the neighboring pico BSs at Tier 2 can be regarded as a single pico BS with aggregated content requests, storage capacities and link capacities in the following analysis. Moreover, it is not necessary to consider the strategy of content placement with such an aggregation since the content request can be easily routed through direct links among neighboring pico BSs.

C. Indirect Cooperation at Tier 2

For each pico user, if the requested content is available at the local pico BS or its neighboring pico BSs at Tier 2, the request can be immediately satisfied. Otherwise, the pico user's content request needs to be routed to the local micro BS at Tier 1. Here, based on the analysis in Section IV-B, all the neighboring pico BSs are directly collaborative and can be aggregated as a single pico BS. To simply the description in this section, a local pico BS also means a single aggregated pico BS. As well, to maximize the amount of supported traffic load caused by pico users, we explore the indirect cooperation at Tier 2 as shown in Fig. 4. In the indirect cooperation at Tier 2, a pico user's requested content that is not available at the local pico BS can be fetched by the local micro BS at Tier 1 from another pico BS in the local micro cell. Here, we also assume that one content is either entirely fetched from

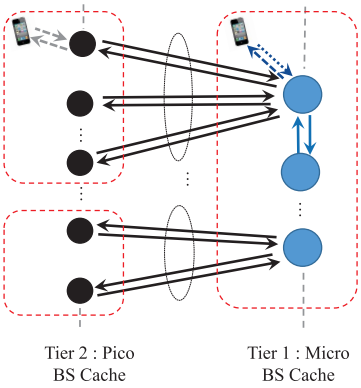


Fig. 4. Topology of indirect cooperation at Tier 2. Here, both the BSs and links for cooperation are in black color.

one pico BS or not, in order to reduce the complexity and generated signaling overhead of content management.

Moreover, the pico BSs in the same micro cell have no direct link due to the corresponding aggregation, i.e., $\delta_{ij}^{m} = \delta_{ji}^{m} = 0$ for $\forall i \neq j$ and $\delta_{ii}^{m} = 1$ for $\forall i$. As well, the notation $\{y_{kn}^{fm}\} \in \{0, 1\}$ is used to denote the routing decision whether content o_f is fetched from pico BS_{mk} to pico BS_{mn} via micro BS_m or not, and we set $y_{nn}^{fm} = 0$. For the pico BS_{mn}, the overall traffic load from pico users supported locally is calculated as $\sum_{f=1}^{F} x_{mn}^{f} \lambda_{mn}^{f} s_f$, while the amount of supported traffic load by other pico BSs in the local micro cell is given by $\sum_{f=1}^{F} \sum_{k=1}^{N_m} y_{kn}^{fm} \lambda_{mn}^{f} s_f$. The indirect cooperation at Tier 2 in each micro cell is independent and can be operated in parallel. Thus, by using the link capacity between pico BSs and the local micro BS, we formulate the problem of indirect cooperation at Tier 2 in the m-th micro cell as

$$\max_{\{x_{mn}^{f}\},\{y_{kn}^{fm}\}} \sum_{f=1}^{F} \sum_{n=1}^{N_m} x_{mn}^{f} \lambda_{mn}^{f} s_f + \sum_{f=1}^{F} \sum_{n=1}^{N_m} \sum_{k=1}^{N_m} y_{kn}^{fm} \lambda_{mn}^{f} s_f \tag{11a}$$

$$s.t. \sum_{f=1}^{F} x_{mn}^{f} s_f \leq S_{mn}, \quad \forall n, \tag{11b}$$

$$\sum_{f=1}^{F} \sum_{n=1}^{N_m} y_{kn}^{fm} \lambda_{mn}^{f} s_f \leq C_k^{m}, \quad \forall k, \tag{11c}$$

$$\sum_{f=1}^{F} \sum_{k=1}^{N_m} y_{kn}^{fm} \lambda_{mn}^{f} s_f \leq C_n^{m}, \quad \forall n, \tag{11d}$$

$$y_{kn}^{fm} \leq x_{mk}^{f} \oplus x_{mn}^{f} - x_{mn}^{f}, \quad \forall n, \ \forall k, \ \forall f, \tag{11e}$$

$$\sum_{k=1}^{N_m} y_{kn}^{fm} \leq \bigcup_{k=1}^{N_m} x_{mk}^{f} - x_{mn}^{f}, \quad \forall n, \ \forall f, \tag{11f}$$

$$x_{mn}^{f} \in \{0, 1\}, y_{kn}^{fm} \in \{0, 1\}, y_{nn}^{fm} = 0, \quad \forall n, \ \forall k, \ \forall f. \tag{11g}$$

(11c) and (11d) denote the constraints of the fronthaul capacity and backhaul capacity between pico BSs and micro BS_m, respectively. (11e) and (11f) denote the cooperation among pico BSs in the m-th micro cell and guarantee that any content request will not be routed to other pico BSs if the content is locally available. Moreover, to equivalently transform the constraints in (11e) and (11f) into linear constraints, except using *Theorem 1*, we provide another equivalent form as

$$y_{kn}^{fm} \leq x_{mk}^{f} \quad \text{and} \quad \sum_{k=1}^{N_m} y_{kn}^{fm} \leq 1 - x_{mn}^{f}, \quad \forall m, \ \forall n, \ \forall k, \ \forall f. \tag{12}$$

Thus, the problem in (11) is a binary integer linear programming (BILP) problem and thus NP-complete. Considering the real-world scale of variables in the BILP problem in (11), it is not practical for engineering implementation to get the optimal solution by using exact methods, e.g., branch and bound methods [42], due to their exponential complexity. Thus, we develop a low-complexity distributed heuristic method to achieve suboptimal solutions. The basic idea of the proposed method is to divide the problem in (11) into two subproblems as follow:

1) Subproblem 1: Optimization of $\{x_{mn}^{f}\}$, which can be formulated as

$$\max_{\{x_{mn}^{f}\}} \sum_{n=1}^{N_m} \sum_{f=1}^{F} x_{mn}^{f} \lambda_{mn}^{f} s_f, \quad s.t. \ (11b), (11g). \tag{13}$$

2) Subproblem 2: Optimization of $\{y_{kn}^{fm}\}$ under achieved $\{x_{mn}^{f}\}$, which can be formulated as

$$\max_{\{y_{kn}^{mf}\}} \sum_{f=1}^{F} \sum_{n=1}^{N_m} \sum_{k=1}^{N_m} y_{kn}^{fm} \lambda_{mn}^{f} s_f,$$
$$s.t. \ (11c), (11d), (11g) \ \text{ and } (12). \tag{14}$$

Subproblem 1 can be decomposed into N_m single knapsack problems and can be separately solved by using the greedy algorithm in [42]. Subproblem 2 can also be solved with a greedy method. The details of the proposed greedy heuristic method for solving the problem in (11) is shown in Algorithm 1 with the complexity $O(T \log(T))$ where $T = F \cdot N_m \cdot N_m$. In the proposed method, solving Subproblem 1 means to cache contents at each pico BS in the m-th micro cell in a greedy manner, while solving Subproblem 2 means to achieve the indirect cooperation among pico BSs in the same local micro cell. Thus, the whole problem of indirect cooperation at Tier 2 can be solved in parallel with Algorithm 1.

D. Cooperation at Tier 1

Given our proposed caching mechanisms at Tier 2, the direct cooperation and indirect cooperation among pico BSs in the same local micro cell have been considered. Then at Tier 1, the average arrival rate of the requests for content o_f at micro BS_m is given as

$$\lambda_m^{f} = \Big[\sum_{n=1}^{N_m} \lambda_{mn}^{f} (1 - x_{mn}^{f} - \sum_{k=1}^{N_m} y_{kn}^{fm})\Big] + \beta_m^{f}, \quad \forall m, \ \forall f, \tag{15}$$

which consists of the content requests from partial pico users and all micro users in the m-th micro cell.

In the following, we will explore the caching cooperation at Tier 1 as shown in Fig. 5. To improve the QoS of users especially on delay of content delivery, we assume that each

Algorithm 1 Greedy Heuristic Method for Solving (11)

Input: m, F, N_m, $\{B_{kn}^{fm}\}$, $\{\lambda_{mn}^{f}\}$, $\{s_f\}$, $\{S_{mn}\}$, $\{C_n^m\}$.
Initialize $x_{mn}^{f}=0$, $y_{kn}^{fm}=0$ for $\forall k, \forall n, \forall f$.
–Procedure 1. [Optimize $\{x_{mn}^{f}\}$]
Optimize $\{x_{mn}^{f}\}$ by solving Subproblem 1 with the greedy method in [42].
–Procedure 2. [Optimize $\{y_{kn}^{fm}\}$ under achieved $\{x_{mn}^{f}\}$]
Set $Cl_n := C_n^m$, $Cd_n := C_n^m$ for $\forall n$, $\mathbf{A}=\boldsymbol{\pi}=\boldsymbol{\omega}:=\mathbf{0}_{(F\cdot Nm)\times 1}$.
Reshape $\{\lambda_{mn}^{f} s_f\}$ into $\mathbf{A}$ where $A_j=\lambda_{mn}^{f} s_f$ satisfying $f=\mathrm{mod}(j-1,F)+1$ and $n=\frac{j-f}{F}+1$.
Sort $\mathbf{A}$ into $\boldsymbol{\pi}$ in a descending order, and label the original indices as $\boldsymbol{\omega}$, where $A_j=\pi_i$ satisfying $j=\omega_i$.
for $j=1$ to $F\cdot N_m$ **do**
  Calculate $f=\mathrm{mod}(j-1,F)+1$ and $n=\frac{j-f}{F}+1$.
  if $(x_{mn}^{f}==0)$ & $(\sum_{i=1}^{N_m} x_{mi}^{f}\geq 1)$ & $(A_j\leq Cd_n)$ **then**
    Find the set $\mathcal{S}\subseteq\{1,2,\ldots,N_m\}$ with all the elements w.r.t. i satisfying $x_{mi}^{f}=1$.
    Find $i^*:=\arg\min_{i\in\mathcal{S}}\{\frac{\lambda_{mi}^{f} s_f}{Cl_i}\}$.
    if $\lambda_{mi^*}^{f} s_f\leq Cl_{i^*}$ **then**
      Set $y_{i^*n}^{fm}=1$, $Cl_{i^*}\leftarrow Cl_{i^*}-A_j$, and $Cd_n\leftarrow Cd_n-A_j$.
    end if
  end if
end for
Output: $\{x_{mn}^{f}\}$ and $\{y_{kn}^{fm}\}$.

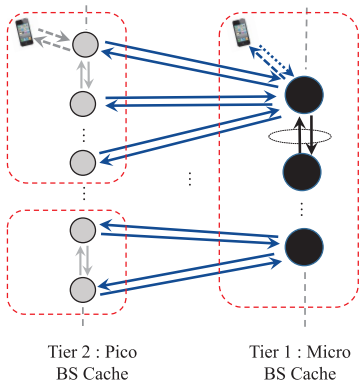


Fig. 5. Topology of cooperation at Tier 1. Here, both the BSs and links for cooperation are in black color.

content request in the caching cooperation at Tier 1 can only be routed from the local micro BS to its neighboring micro BSs. Moreover, to reduce the complexity and generated signaling overhead of content management at Tier 1, we also assume that any content between two neighboring micro BSs is entirely disseminated or not. Besides, the notation $\{y_{mt}^{f}\}\in\{0,1\}$ is used to denote the routing decision whether content o_f is fetched from micro BS$_m$ to micro BS$_t$ entirely or not, and we set $y_{mm}^{f}=0$. Once the topology at Tier 1 is given, the indices $\{\delta_{mt}\}$, denoting whether micro BS$_m$ and micro BS$_t$ are neighboring, will be determined. Denote the set $\mathcal{M}_m:=\{t|\,\delta_{mt}=1,\forall t\neq m\}$ as the neighboring micro BS set of micro BS$_m$. As a result, we set $C_{mt}=C_{tm}=0$ and $y_{mt}^{f}=y_{tm}^{f}=0$ for $\forall t\notin\mathcal{M}_m$.

The main target here is to achieve proper content caching schemes within the neighboring micro BSs at Tier 1 under the constraints of the caching storage and link capacities in order to satisfy heterogeneous content requests. Thus, according to content requests at each micro BS, we then employ a cost-utility function to denote the benefits of caching content o_f at micro BS$_m$ or its neighboring micro BSs as

$$u_m^f=\frac{u_{loc,m}^{f}+u_{nei,m}^{f}}{s_f},\quad \forall m,\ \forall f, \tag{16}$$

where $u_{loc,m}^{f}$ and $u_{nei,m}^{f}$ denote the local utility at micro BS$_m$ and the remote utility at its neighboring micro BSs, respectively. Specifically, the local utility $u_{loc,m}^{f}$ is achieved from whether to cache content o_f at BS$_m$ and deliver content o_f to the neighboring micro BSs, while the remote utility $u_{nei,m}^{f}$ is obtained from whether to deliver content o_f from neighboring micro BSs to micro BS$_m$.

Moreover, to design proper $u_{loc,m}^{f}$ and $u_{nei,m}^{f}$, some important factors need to be considered as: 1) $u_{loc,m}^{f}$ is a monotone increasing function w.r.t. the local content request's average arrival rate λ_m^f and is a monotone decreasing function w.r.t. the neighboring content request's average arrival rate λ_t^f; 2) $u_{nei,m}^{f}$ is a monotone increasing function w.r.t. the local content request's average arrival rate λ_m^f; 3) Both $u_{loc,m}^{f}$ and $u_{nei,m}^{f}$ are related to the content availability of content o_f within micro BS$_m$ and its neighboring micro BSs, thereby indicating the caching cooperation at Tier 1; 4) Both $u_{loc,m}^{f}$ and $u_{nei,m}^{f}$ are limited by the link capacity between micro BS$_m$ and its neighboring micro BSs. Besides, we assume the information among neighboring micro BSs can be periodically exchanged with low costs. Here, the cost of caching content o_f at micro BS$_m$ is the storage cost s_f. Thus, based on the above consideration, we derive the utility function as: 1) if $\mathcal{M}_m=\emptyset$, then $u_{loc,m}^{f}=x_m^f\lambda_m^f s_f$ and $u_{nei,m}^{f}=0$; 2) otherwise, $u_{loc,m}^{f}=x_m^f\lambda_m^f s_f-\frac{1}{|\mathcal{M}_m|}\sum_{t\in\mathcal{M}_m} y_{mt}^{f}\lambda_t^f s_f(1-\exp\{-\frac{\eta\lambda_t^f s_f}{C_{mt}}\})$ and $u_{nei,m}^{f}=\sum_{t\in\mathcal{M}_m} y_{tm}^{f}\lambda_m^f s_f(1-\exp\{-\frac{\eta\lambda_m^f s_f}{C_{tm}}\})$. Here, η is an introduced parameter denoting the sensitivity to the content availability in the neighboring micro BSs. Moreover, to satisfy the above considered factors, we may choose $0<\eta\leq\min_{m,\,\mathcal{M}_m\neq\emptyset}\Big\{\max\{2,\min_{t\in\mathcal{M}_m,\,f}\{\frac{2C_{tm}}{s_f\lambda_m^f},\frac{2C_{mt}}{s_f\lambda_t^f}\}\}\Big\}$.

In terms of the link capacity, the used utility function can indicate the effectiveness of caching cooperation. The increase of link capacity C_{mt} results in a decrease with $1-\exp\{-\frac{\eta\lambda_t^f s_f}{C_{mt}}\}$, which implies that the neighboring micro BS$_t$ with higher link capacity are with higher probability to consider the content delivery from micro BS$_m$. In a similar way, the increase of link capacity C_{tm} leads to a decreasing concern on the local content requests at micro BS$_m$ if the content is cached at the neighboring micro BS$_t$. In order to show the priority diversity of micro BSs at Tier 1 in the system, we assume that different micro BSs have various weighted factors w.r.t. the utility, denoted by $(\omega_m)_{M\times 1}$. In the strategy of caching cooperation at Tier 1, we aim at maximizing the weighted sum utility of all the micro BSs by using the link capacity among neighboring micro BSs. The corresponding

problem is formulated as

$$\max_{\{x_m^f\},\{y_{mt}^f\}} \sum_{m=1}^{M} \left(\omega_m \sum_{f=1}^{F} u_m^f\right) \tag{17a}$$

$$s.t. \sum_{f=1}^{F} x_m^f s_f \le S_m, \quad \forall m, \tag{17b}$$

$$\sum_{f=1}^{F} y_{mt}^f \lambda_t^f s_f \le C_{mt}, \quad \forall m,\ \forall t \in \mathcal{M}_m \ne \emptyset, \tag{17c}$$

$$y_{mt}^f \le x_m^f, \quad \forall m,\ \forall t \in \mathcal{M}_m \ne \emptyset,\ \forall f, \tag{17d}$$

$$\sum_{t \in \mathcal{M}_m} y_{tm}^f \le 1 - x_m^f, \quad \forall m,\ \mathcal{M}_m \ne \emptyset,\ \forall f, \tag{17e}$$

$$x_m^f \in \{0,1\}, y_{mt}^f \in \{0,1\}, y_{mm}^f = 0, \quad \forall m,\ \forall t \in \mathcal{M}_m \ne \emptyset,\ \forall f, \tag{17f}$$

$$y_{mt}^f = 0, \quad \forall m,\ \forall t \notin \mathcal{M}_m,\ \forall f. \tag{17g}$$

Here, (17b) and (17c) denote the constraints of micro BSs' caching storage and the link capacity between a pair of neighboring micro BSs, respectively. (17d) and (17e) denote the cooperation between neighboring micro BSs and guarantee that any content request will not be routed to neighboring micro BSs if the content is locally available. The problem in (17) is also a BILP problem and thus NP-complete.

We define two sets $\mathcal{T}_0 := \{m | \mathcal{M}_m = \emptyset, \forall m\}$ and $\mathcal{T}_1 := \{1, 2, \ldots, M\} \setminus \mathcal{T}_0$. Thus, the cost-utility function in (16) can be rewritten as in (18), as shown at the bottom of this page.

Observed from (18), u_m^f for $\forall m \in \mathcal{T}_0, \forall f$ is only dependent on x_m^f since $y_{tm}^f = y_{mt}^f = 0$ holds for $\forall m \in \mathcal{T}_0, \forall t, \forall f$, while u_m^f for $\forall m \in \mathcal{T}_1, \forall f$ is dependent on $(x_m^f, y_{tm}^f, y_{mt}^f), \forall t \in \mathcal{M}_m$. Besides, the optimization objective function in (17a) can be separated into two parts as $\sum_{m \in \mathcal{T}_0} \left(\omega_m \sum_{f=1}^{F} u_m^f\right) + \sum_{m \in \mathcal{T}_1} \left(\omega_m \sum_{f=1}^{F} u_m^f\right)$. Thus, the problem in (17) can be decomposed into two subproblems based on $\mathcal{T}_0$ and $\mathcal{T}_1$. More specifically, the first subproblem is to maximize $\sum_{m \in \mathcal{T}_0} \left(\omega_m \sum_{f=1}^{F} u_m^f\right)$ w.r.t. $\{x_m^f\}, \forall m \in \mathcal{T}_0, \forall f$ under the constraints in (17b) and (17f), and can be further decomposed into $|\mathcal{T}_0|$ single knapsack problems, which are in the form as

$$\max_{\{x_m^f\},\, m \in \mathcal{T}_0} \sum_{f=1}^{F} x_m^f \lambda_m^f \tag{19a}$$

$$s.t. \sum_{f=1}^{F} x_m^f s_f \le S_m, \quad x_m^f \in \{0,1\},\ \forall f. \tag{19b}$$

Meanwhile, the second subproblem is to maximize $\sum_{m \in \mathcal{T}_1} \left(\omega_m \sum_{f=1}^{F} u_m^f\right)$ w.r.t. $(x_m^f, y_{tm}^f, y_{mt}^f), \forall m \in \mathcal{T}_1, \forall t, \forall f$ under the constraints in (17b)-(17g), which is in a similar form as in (17).

Both of the above two subproblems are usually large-scale BILP problems in the real-world system. Thus, based on the main idea of Algorithm 1, we develop an efficient greedy heuristic method as shown in Algorithm 2 to solve the two subproblems as well as the whole problem. In the procedure of Algorithm 2, Procedure 1 and Procedure 2 aim at caching contents at each micro BS in a greedy manner and can be operated in parallel, while Procedure 3 aims at achieving the caching cooperation among neighboring micro BSs at Tier 1. Besides, the whole procedure of Algorithm 2 takes $O(FM^2 \log(FM^2))$ time.

Algorithm 2 Greedy Heuristic Method for Solving (17)

Input: M, F, η, $\{\omega_m\}$, $\{\lambda_m^f\}$, $\{s_f\}$, $\{S_m\}$, $\{C_{mt}\}$, $\{\mathcal{M}_m\}$, $\mathcal{T}_0$, $\mathcal{T}_1$.
Initialize $x_m^f = 0$, $y_{mt}^f = 0$ for $\forall m, \forall t, \forall f$.
–Procedure 1. [Optimize $\{x_m^f\}$ for $m \in \mathcal{T}_0$]
Optimize $\{x_m^f\}$ for $m \in \mathcal{T}_0$ with the greedy method in [42].
–Procedure 2. [Optimize $\{x_m^f\}$ for $m \in \mathcal{T}_1$]
Optimize $\{x_m^f\}$ for $m \in \mathcal{T}_1$ with the similar method in Procedure 1.
–Procedure 3. [Optimize $\{y_{mt}^f\}$ under $\{x_m^f\}$ for $m \in \mathcal{T}_1$]
Set $\overline{C}_{mt} := C_{mt}$, $A_{mt}^f := \frac{1}{|\mathcal{M}_m|} \omega_m \lambda_t^f (1 - \exp\{-\frac{\eta \lambda_t^f s_f}{C_{mt}}\})$ and $B_{tm}^f := \omega_m \lambda_m^f (1 - \exp\{-\frac{\eta \lambda_m^f s_f}{C_{tm}}\})$ for $\forall m \in \mathcal{T}_1, \forall t \in \mathcal{M}_m, \forall f$. Besides, set $\mathcal{V} := \{(f,m,t) | x_m^f = 1, x_t^f = 0, \forall m \in \mathcal{T}_1, \forall t \in \mathcal{M}_m, \forall f\}$ and $\mathcal{V}_m^f := \{(f,m,t) | x_m^f = 0, x_t^f = 1, \forall t \in \mathcal{M}_m\}$ for $\forall m \in \mathcal{T}_1, \forall f$.
while $\mathcal{V} \ne \emptyset$ **do**
  Find $(f^*, m^*, t^*) := \arg \max_{(f,m,t) \in \mathcal{V}} \left\{\frac{B_{tm}^f - A_{mt}^f}{\lambda_t^f s_f}\right\}$.
  if $(\lambda_{t^*}^{f^*} s_{f^*} \le \overline{C}_{m^* t^*})$ **then**
    Set $y_{m^* t^*}^{f^*} = 1$, $\overline{C}_{m^* t^*} \leftarrow \overline{C}_{m^* t^*} - \lambda_{t^*}^{f^*} s_{f^*}$.
    Set $\mathcal{V} \leftarrow \mathcal{V} \setminus \mathcal{V}_{t^*}^{f^*}$.
  else
    Set $\mathcal{V} \leftarrow \mathcal{V} \setminus \{(f^*, m^*, t^*)\}$.
  end if
end while
Output: $\{x_m^f\}$ and $\{y_{mt}^f\}$.

E. Content Caching at Tier 0

With our proposed caching strategies at Tier 1, the cooperation among neighboring micro BSs in the macro cell has been considered. Thus, the average arrival rate of the requests for content o_f at the macro BS is given as

$$\lambda_0^f = \left[\sum_{m=1}^{M} \lambda_m^f (1 - x_m^f - \sum_{t \in \mathcal{M}_m} y_{tm}^f)\right] + \beta_0^f, \quad \forall f, \tag{20}$$

which consists of the content requests from micro BSs and all macro users in the macro cell.

$$u_m^f = \begin{cases} x_m^f \lambda_m^f, & \forall m \in \mathcal{T}_0,\ \forall f, \\ x_m^f \lambda_m^f - \dfrac{1}{|\mathcal{M}_m|} \displaystyle\sum_{t \in \mathcal{M}_m} y_{mt}^f \lambda_t^f (1 - e^{-\frac{\eta \lambda_t^f s_f}{C_{mt}}}) + \sum_{t \in \mathcal{M}_m} y_{tm}^f \lambda_m^f (1 - e^{-\frac{\eta \lambda_m^f s_f}{C_{tm}}}), & \forall m \in \mathcal{T}_1,\ \forall f. \end{cases} \tag{18}$$

As well, each content is assumed to be either entirely cached at the macro BS or not, denoted by $\{x_f\}$. The strategy of content caching at Tier 0 aims at maximizing the overall traffic load supported locally under the constraint of the macro BS's storage capacity, and the problem can be formulated as

$$\max_{\{x^f\}} \sum_{f=1}^{F} x_f \lambda_0^f s_f, \quad s.t. \sum_{f=1}^{F} x_f s_f \leq S_0, x_f \in \{0,1\}, \quad \forall f. \tag{21}$$

The problem in (21) is a single knapsack problem and can be solved with the low-complexity greedy method in [42] to achieve the content caching strategy at Tier 0.

F. Content Request Routing

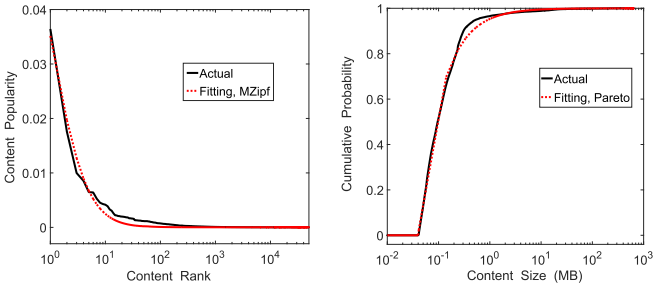
Using the above decomposed content caching cooperation framework in the multi-tier HetNet system, we can get the strategies of joint content caching and request routing with the practical consideration on large-scale content distribution. For any user request for content o_f, the details of the proposed content request routing strategy are given in Algorithm 3.

Algorithm 3 Content Request Routing Strategy

1: **Input**: Achieved content caching results at each tier with $\{x_{mn}^f\}$, $\{z_{kn}^{fm}\}$, $\{y_{kn}^{fm}\}$, $\{x_m^f\}$, $\{y_{mt}^f\}$ and $\{x^f\}$.
2: After receiving a request for content o_f, check the user type.
3: **if** pico user **then**
4: Satisfy the request at the local pico BS$_{mn}$ if $x_{mn}^f = 1$.
5: If not yet satisfied, check the content availability of the neighboring pico BSs of pico BS$_{mn}$, and route the request to pico BS$_{mk}$ if $z_{kn}^{fm} = 1$.
6: If not yet satisfied, route the request to micro BS$_m$, and satisfy the request if $x_m^f = 1$.
7: If not yet satisfied, check the following three options:
8: a) Route the request to pico BS$_{mk}$ if $y_{kn}^{fm} = 1$.
9: b) Otherwise, route the request to micro BS$_t$ if $y_{tm}^f = 1$.
10: c) At last, route the request to the macro BS.
11: **else**
12: **if** micro user **then**
13: Satisfy the request at the local micro BS$_m$ if $x_m^f = 1$.
14: If not yet satisfied, check the following two options:
15: a) Route the request to micro BS$_t$ if $y_{tm}^f = 1$.
16: b) Otherwise, route the request to the macro BS.
17: **else**
18: Route the request to the macro BS, and satisfy the request locally if $x_f = 1$ or download the content.
19: **end if**
20: **end if**

V. Trace-Based Simulation Results

In this section, we evaluate the performance of our collaborative multi-tier caching scheme based on the trace of a real-world proxy caching system, IRCache [43]. The IRCache traces are original for providing operational web caching services by Squid Caching Software, and the logs of HTTP content caching and forwarding with user requests can be utilized for our simulation purpose. We have collected logs for 7 days in June 2013 to represent the user requests of popular Internet contents along with their content sizes. And after filtering out some illegal and incorrect data, we select about 50,000 popular contents, about 5,000 content requesters, and related over 500,000 content requests. Here, the content requesters in IRCache are regarded as mobile users in the HetNet, and their user association and content requests are randomly set as modeled in Section III-C. We use MATLAB to implement a simulator that constructs the three-tier caching topology, in which the macro BS is connected to a number of micro BSs and each micro BS is connected to a number of pico BSs. Based on [44], we set the path loss exponent of α as 4, the BSs' transmit power of $(\Gamma_0, \Gamma_1, \Gamma_2)$ as (46, 40, 30) dBm, all the SINR threshold of $(\Upsilon_0, \Upsilon_1, \Upsilon_2)$ as 0 dB. Besides, we set the ratio of the cache sizes of (S_{mn}, S_m, S_0) as $1:2:4$, and the sensitivity parameter of η as 1.5. Moreover, the scalability of our proposed framework is not restricted by the aforementioned parameters. All codes are written in MATLAB (8.6.0) running on a personal computer (Intel Core i7-2600 CPU @3.40GHz, 8.00 GB RAM).



(a) Content popularity (b) Content size distribution

Fig. 6. Comparison of actual and fitting distribution.

In the following subsections, we will first analyze the actual and fitting distribution w.r.t. the contents in the trace, and then evaluate the performance of the proposed scheme in comparison with some proper baseline schemes when the number of the considered contents is either small-scale or large-scale, i.e., small-scale case and large-scale case, respectively.

A. Actual and Fitting Distribution

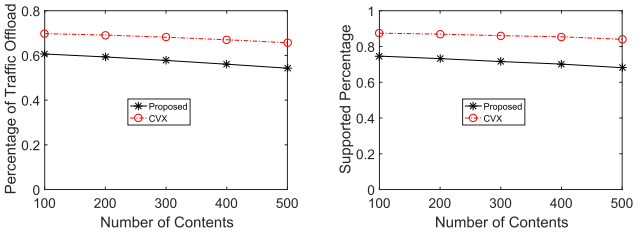
We first analyze the content popularity and content size distribution in the trace as shown in Fig. 6. From Fig. 6(a), it can be observed that the actual content popularity in the trace can be well fitted by the MZipf distribution with $(q, \beta) = (1.56, 1.76)$, which agrees with the used model on the content popularity in (3). Meanwhile, from Fig. 6(b) the actual content size distribution in the trace can be well fitted by a Pareto distribution, which agrees with the conclusion in [39]. Note that the following results are based on the above practical trace.

B. Small-Scale Case

In the small-scale case, we use the trace of the most popular $(100, 200, 300, 400, 500)$ contents accounting for $(0.2\%, 0.4\%, 0.6\%, 0.8\%, 1\%)$ of the entire content set as well as the corresponding set of users that request the contents. For the simulation purpose, we set that the macro BS is connected

TABLE I
CPU COMPUTATION TIME (UNIT: SECOND)

Content Number	100	200	300	400	500
CVX	11.2	611.8	1,858.1	10,636.4	31,335.5
Proposed	2.3	2.6	3.0	3.7	5.2



(a) Traffic offload (b) Supported request percentage

Fig. 7. The percentage of traffic offload and the percentage of supported requests versus different numbers of contents, where the cache size of each pico BS at Tier 2 is 5% of the sum size of the considered contents for each given content number.

to two micro BSs and each micro BS is connected to two pico BSs, and set the topology as: 1) at Tier 2, pico BS_{11} is connected to pico BS_{12}; 2) at Tier 1, micro BS_1 is connected to micro BS_2. Besides, we set the average available link capacity of $(C_{kn}^m, C_n^m, C_{tm})$ as $(20, 20, 40)$ Mbps and the weighted factors of $\boldsymbol{\omega}$ as $(2, 1)$, and set that the cache size of each pico BS at Tier 2 is 5% of the sum size of the considered contents for each given content number. To evaluate the performance gap between the optimal solution and the achieved heuristic solution with our proposed scheme, we use MATLAB CVX Toolbox as a baseline scheme to achieve the optimal solutions to all the considered BILP problems at all the tiers.

Table I and Fig. 7 evaluate the performance of the proposed scheme and the CVX scheme in terms of the CPU computation time and the percentages of both traffic offload[2] and supported requests versus different numbers of the considered contents, respectively. From Table I, as the number of contents increases, the optimal CVX scheme takes exponentially increasing CPU computation time and thus is not practical to be used for solving the problem in the large-scale case, while the proposed scheme has a much lower time complexity and thus can also be used in the large-scale case. From Fig. 7, the proposed scheme can achieve from 83% to 87% and from 81% to 85% of the optimal performance of the CVX scheme in terms of the percentage of traffic load and the percentage of supported requests, respectively. Thus, from Table I and Fig. 7, the proposed scheme has low complexity and can achieve good performance on offloading the traffic and supporting content requests from users in the network, which can be used for practical engineering implementation.

C. Large-Scale Case

In the large-scale case, we use the trace of the entire content set and the entire user set, and set that the macro BS is connected to five micro BSs and each micro BS is connected to five pico BSs. For the simulation purpose, we set the topology as: 1) at Tier 2, pico BS_{m1} is connected to pico BS_{m2} for $\forall m \in \{1, 3, 5\}$; 2) at Tier 1, micro BS_1 is connected to micro BS_3, micro BS_4 and micro BS_5, while micro BS_2 is connected to micro BS_4 and micro BS_5. Besides, we set the average available link capacity of $(C_{kn}^m, C_n^m, C_{tm})$ as $(0.5, 0.5, 1)$ Gbps based on [44], and the weighted factors of $\boldsymbol{\omega}$ as $(2, 2, 1, 1, 1)$. In particular, we compare the proposed collaborative multi-tier caching scheme with two baseline schemes below.

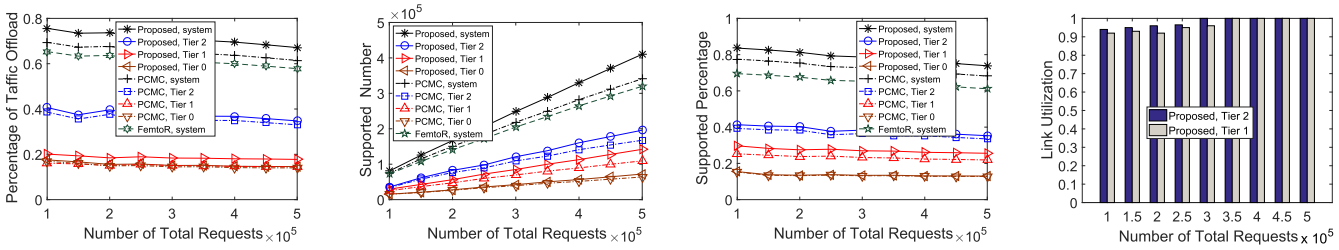
1) Revised FemtoCaching (FemtoR): This scheme is derived by carefully modifying the proposed FemtoCaching scheme in [21] to address the case in this paper where content sizes are not the same but various. In FemtoCaching, the femtocell-like BSs form a distributed single-tier caching network to assist the macro BS by handling content requests and caches based on a greedy algorithm. In FemtoR in this paper, all the BSs are regarded as femtocell-like BSs, and the cache sizes of the macro BS and all the micro BSs as well as well as the link connection among BSs are also set as the same as those in the proposed scheme. Consequently, FemtoR is a collaborative but single-tier caching scheme.

2) Partial-Collaborative Multi-Tier Caching (PCMC): This scheme is also a multi-tier caching scheme with partial cooperation. In PCMC, the cooperation among the BSs at the same tier is not considered and each BS at the same tier greedily caches contents by using the method in [42], while the cooperation among the BSs at different tiers is considered. Thus, compared with the proposed scheme, one caching result of the PCMC scheme is that all the values of $(z_{kn}^{fm}, y_{kn}^{fm}, y_{tm}^{f})$ are zeros.

a) Effects of different numbers of total requests: Fig. 8 evaluates the performance of the proposed collaborative multi-tier caching scheme in terms of the percentage of traffic offload, the number of supported requests, the percentage of supported requests and the link utilization[3] versus different numbers of total requests. Here, the cache size of each pico BS at Tier 2 is set as 5% of the sum size of the total contents. From Fig. 8(a), the proposed scheme can offload the traffic of the entire system by up to 76%, and outperforms the FemtoR scheme with around 16% improvement for the entire system as well as the PCMC scheme with 5% to 26% improvements for the entire system and each tier. Compared with the FemtoR scheme and the PCMC scheme, the performance gap achieved by the proposed scheme is due to the multi-tier caching infrastructure and due to the cooperation at the same tier (i.e., Tier 2 and Tier 1), respectively. From Fig. 8(b) and Fig. 8(c), all the numbers of the supported requests for the entire system in the three schemes and each tier in the proposed scheme and the PCMC scheme go up as the number of total requests increases. Meanwhile, the proposed scheme can support up to 84% of the total requests, and outperforms the FemtoR scheme with around 21% improvement for the entire system as well as the PCMC scheme with 3% to 16% improvements for the entire system and each tier. From Fig. 8(a) and Fig. 8(c), the contribution of all the tiers in the proposed scheme can

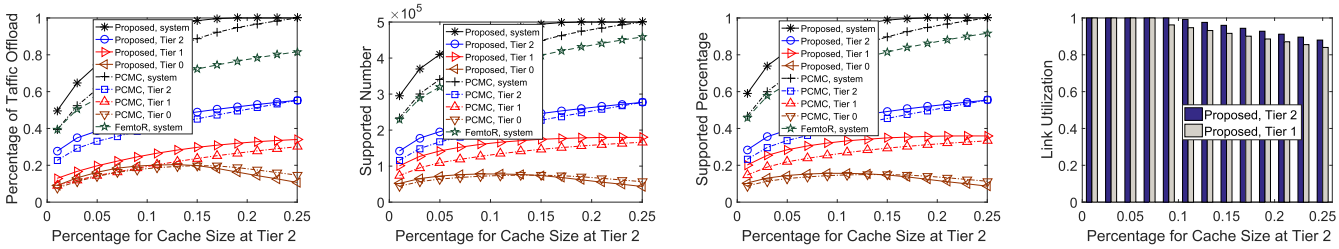
[2]The traffic offload denotes the reduction on the traffic caused by downloading contents from SPs over the Internet to the macro BS via the MNO core.

[3]The link utilization at Tier 1 and Tier 2, respectively, denotes the utilization of intra-links among neighboring micro BSs at Tier 1 and inter-links between pico BSs at Tier 2 and the local micro BSs at Tier 1.



(a) Traffic offload (b) Supported request number (c) Supported request percentage (d) Link utilization

Fig. 8. The percentage of traffic offload, the number of supported requests, the percentage of supported requests and the link utilization versus different numbers of total requests, where the cache size of each pico BS at Tier 2 is 5% of the sum size of the total contents.



(a) Traffic offload (b) Supported request number (c) Supported request percentage (d) Link utilization

Fig. 9. The percentage of traffic offload, the number of supported requests, the percentage of supported requests and the link utilization versus different cache sizes (percentage to the total content size) of each pico BS at Tier 2, with the entire set of 500, 000 requests.

maintain stable performance on traffic offload and the percentage of supported requests with different numbers of total requests, which clearly indicates the stability of our proposed scheme. Besides, from Fig. 8(a), Fig. 8(b) and Fig. 8(c), pico BSs at Tier 2 offload much more traffic and satisfy much more requests than the BSs at both Tier 1 and Tier 0 since most content requests are already served through local pico BSs and the cooperation at Tier 2. Fig. 8(d) shows that in the proposed scheme, the cooperation at both Tier 2 and Tier 1 is sufficiently good initially with over 90% utilization, and that as the number of total requests increases, greater numbers of requests received at different tiers lead to full link utilization for the cooperation at Tier 2 and Tier 1.

b) Effects of different cache sizes: Fig. 9 evaluates the performance of the proposed scheme in terms of the percentage of traffic offload, the number of supported requests, the percentage of supported requests and the link utilization versus different cache sizes (percentage to the total content size) of each pico BS at Tier 2. Here, we use the entire set of 500, 000 requests. From Fig. 9(a), the proposed scheme can offload the traffic of the entire system by at least 50% and even 100% when the cache size of each pico BS is sufficiently large, and outperforms the FemtoR scheme with up to 34% improvement for the entire system as well as the PCMC scheme with up to 20% improvement for the entire system and each tier. From Fig. 9(b) and Fig. 9(c), as the increase of the number of total requests, all the numbers and percentage of the supported requests for the entire system in the three schemes and Tier 1 in the proposed scheme and the PCMC scheme go up and finally stay constant, and those for Tier 2 always increase while those for Tier 0 first increase and then gradually decrease in the proposed scheme and PCMC scheme. Meanwhile, the proposed scheme can support up to 84% of the total requests and outperforms the FemtoR scheme with up to 25% improvement for the entire system as well as the PCMC scheme with 3% to 16% improvements for the entire system and each tier. Besides, from Fig. 9(a), Fig. 9(b) and Fig. 9(c), pico BSs at Tier 2 also offload much more traffic and satisfy much more requests than the BSs at both Tier 1 and Tier 0. From Fig. 9(d), for the cooperation at Tier 2 and Tier 1 in the proposed scheme, small cache sizes of each pico BS lead to full link utilization, while the link utilization gradually decreases with the increase of the cache size when the cache sizes are relatively large.

VI. Conclusion

In this paper, we have proposed a collaborative multi-tier caching framework in HetNets. Specifically, based on user request patterns, link capacities, heterogenous cache sizes and the derived system topology, we have focused on exploring the maximum capacity of the network infrastructure on offloading the network traffic and supporting users' content requests inside the MNO network, approximately decomposed the complex multi-tier caching problem into some subproblems that focus on the caching cooperation at different tiers, and proposed the corresponding low-complexity distributed solutions from the perspective of engineering implementation. Trace-based simulation results have demonstrated the effectiveness of the proposed framework.

Appendix

A. Equivalence Proof of the Problems in (7) and (8)

Based on the assumption that the constraint in (7c) always holds, proving the equivalence of the problems in (7) and (8) can be equivalent to proving that the local/global optimal solution to the problem in (7), denoted by $\left(\{(x_{mn}^{f})^{*}\},\{(z_{kn}^{fm})^{*}\}\right)$,

satisfies

$$(x_{mn}^{f})^{*}+\sum_{k=1}^{N_m}(z_{kn}^{fm})^{*}=\bigcup_{k=1}^{N_m}[\delta_{kn}^{m}(x_{mk}^{f})^{*}],\quad \forall m,\ \forall n,\ \forall f. \quad (22)$$

Note that $\sum_{k=1}^{N_m}(z_{kn}^{fm})^{*}\in\{0,1\}$ and $\bigcup_{k=1}^{N_m}[\delta_{kn}^{m}(x_{mk}^{f})^{*}]\in\{0,1\}$ for $\forall m,\forall n,\forall f$. To further prove that (22) holds, we discuss the local/global optimal solution from two aspects as follow.

- For each fixed (m,n,f) such that $(x_{mn}^{f})^{*}=1$, based on (7d), we can get $(z_{kn}^{fm})^{*}\le[\delta_{kn}^{m}(x_{mk}^{f})^{*}]\oplus(x_{mn}^{f})^{*}-(x_{mn}^{f})^{*}=[\delta_{kn}^{m}(x_{mk}^{f})^{*}]\oplus 1-1=0$, thus $(z_{kn}^{fm})^{*}=0$ for $\forall k$. As a result, we have $\sum_{k=1}^{N_m}(z_{kn}^{fm})^{*}=0$. Besdies, we have $\bigcup_{k=1}^{N_m}[\delta_{kn}^{m}(x_{mk}^{f})^{*}]=1$. Thus, (22) holds.
- For each fixed (m,n,f) such that $(x_{mn}^{f})^{*}=0$, based on (7d), we can get $(z_{kn}^{fm})^{*}\le[\delta_{kn}^{m}(x_{mk}^{f})^{*}]\oplus(x_{mn}^{f})^{*}-(x_{mn}^{f})^{*}=[\delta_{kn}^{m}(x_{mk}^{f})^{*}]\oplus 0-0=\delta_{kn}^{m}(x_{mk}^{f})^{*}$, i.e., $(z_{kn}^{fm})^{*}\le\delta_{kn}^{m}(x_{mk}^{f})^{*}$ for $\forall k$. Besides, based on (7e), we have $\sum_{k=1}^{N_m}(z_{kn}^{fm})^{*}\le\bigcup_{k=1}^{N_m}[\delta_{kn}^{m}(x_{mk}^{f})^{*}]\le 1$. Thus, we further discuss the local/global optimal solution from two aspecs as follow.
  - If $(x_{mk}^{f})^{*}=0$ for $\forall k$, we can get $(z_{kn}^{fm})^{*}=0$ for $\forall k$ and $\bigcup_{k=1}^{N_m}[\delta_{kn}^{m}(x_{mk}^{f})^{*}]=0$, which indicates that (22) holds.
  - Otherwise, i.e., there exists j such that $(x_{mj}^{f})^{*}=1$, then we assume $(z_{kn}^{fm})^{*}=0$ for $\forall k$ and can find another solution by setting $(z_{kj}^{fm})^{*}=1$ such that the finally achieved objevtive value in (7a) is greater than the local/global optimal objective value, which is contradictory with the assumption that $\big(\{(x_{mn}^{f})^{*}\},\{(z_{kn}^{fm})^{*}\}\big)$ is a local/global optimal solution. Thus, we have $\sum_{k=1}^{N_m}(z_{kn}^{fm})^{*}=1$. Besdies, we have $\bigcup_{k=1}^{N_m}[\delta_{kn}^{m}(x_{mk}^{f})^{*}]=1$. Thus, (22) also holds.

Based on the above analysis, the whole proof is complete.

B. Proof of Theorem 1

Clearly, $z=\bigcup_{k=1}^{n}a_k\in\{0,1\}$. If there exists $i\in\{1,2,\ldots,n\}$, such that $a_i=1$, then we can get $z=1$ and $\max_{k\in\{1,2,\ldots,n\}}\{a_k\}=1$, thus (9) holds; otherwise, i.e., $a_k=0$ for $\forall k\in\{1,2,\ldots,n\}$, then we can get $z=0$ and $\max_{k\in\{1,2,\ldots,n\}}\{a_k\}=0$, thus (9) still holds.

If there exists $i\in\{1,2,\ldots,n\}$, such that $a_i=1$, then we can get $y=1\ge a_k$ for $\forall k\in\{1,2,\ldots,n\}$, and $\sum_{k=1}^{n}a_k\ge a_i=1$, thus (10) holds; otherwise, i.e., $a_k=0$ for $\forall k\in\{1,2,\ldots,n\}$, then we can get $y=0=a_k$ for $\forall k\in\{1,2,\ldots,n\}$, and $\sum_{k=1}^{n}a_k=0$, thus (10) still holds.

C. Proof of Theorem 2

With the above condition in an optimal solution, we have $\delta_{in}^{m}x_{mi}^{f}=\delta_{jn}^{m}x_{mj}^{f}=1$ and $i\neq j$. Thus, we have $\bigcup_{k=1}^{N_m}\delta_{kn}^{m}x_{kn}^{f}=1$ and $\delta_{in}^{m}=\delta_{jn}^{m}=x_{mi}^{f}=x_{mj}^{f}=1$. Setting $x_{mj}^{f}=1$ and $x_{mk}^{f}=0$ for $\forall k\neq j$ such that $\delta_{kn}^{m}=1$, we still have $\bigcup_{k=1}^{N_m}\delta_{kn}^{m}x_{kn}^{f}=1$, which generates another feasible solution no worse than the optimal solution and makes the constraint in (8b) more relaxed. Thus, the new solution is also optimal and much better to save the caching storage of neighboring pico BSs.

REFERENCES

[1] X. Wang, X. Li, V. C. M. Leung, and P. Nasiopoulos, "A framework of cooperative cell caching for the future mobile networks," in *Proc. HICSS*, Jan. 2015, pp. 5404–5413.

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

[3] X. Li, X. Wang, C. Zhu, W. Cai, and V. C. M. Leung, "Caching-as-a-service: Virtual caching framework in the cloud-based mobile networks," in *Proc. IEEE INFOCOM Workshops*, Apr. 2015, pp. 372–377.

[4] X. Li, X. Wang, K. Li, and V. C. M. Leung, "CaaS: Caching as a service for 5G networks," *IEEE Access*, vol. 5, pp. 5982–5993, May 2017.

[5] X. Li, X. Wang, and V. C. M. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *Proc. IEEE ICC*, May 2016, pp. 1–6.

[6] X. Li, X. Wang, S. Xiao, and V. C. M. Leung, "Delay performance analysis of cooperative cell caching in future mobile networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 5652–5657.

[7] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[8] P. Rodriguez, C. Spanner, and E. W. Biersack, "Analysis of Web caching architectures: Hierarchical and distributed caching," *IEEE/ACM Trans. Netw.*, vol. 9, no. 4, pp. 404–418, Aug. 2001.

[9] H. Che, Y. Tung, and Z. Wang, "Hierarchical Web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.

[10] K. Poularakis and L. Tassiulas, "On the complexity of optimal content placement in hierarchical caching networks," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2092–2103, May 2016.

[11] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li, "Collaborative hierarchical caching with dynamic request routing for massive content distribution," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2444–2452.

[12] K. Samdanis, T. Taleb, and S. Schmid, "Traffic offload enhancements for eUTRAN," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 3, pp. 884–896, 3rd Quart., 2012.

[13] J. Wu, I. Bisio, C. Gniady, E. Hossain, M. Valla, and H. Li, "Context-aware networking and communications: Part 1," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 14–15, Jun. 2014.

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

[15] W. Han, A. Liu, and V. K. N. Lau, "PHY-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[16] Z. Hu, Z. Zheng, T. Wang, L. Song, and X. Li, "Game theoretic approaches for wireless proactive caching," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 37–43, Aug. 2016.

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

[18] 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.
[19] E. Zeydan *et al.*, "Big data caching for networking: Moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.
[20] K. Kanai *et al.*, "Context-aware proactive content caching for mobile video utilizing transportation systems and evaluation through field experiments," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2102–2114, Aug. 2016.
[21] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.
[22] B. Han, X. Wang, N. Choi, T. K. Kwon, and Y. Choi, "AMVS-NDN: Adaptive mobile video streaming and sharing in wireless named data networking," in *Proc. IEEE INFOCOM Workshop*, Apr. 2013, pp. 375–380.
[23] J.-P. Hong and W. Choi, "User prefix caching for average playback delay reduction in wireless video streaming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 377–388, Jan. 2016.
[24] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.
[25] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *Proc. MobiHoc*, Jun. 2015, pp. 127–136.
[26] Z. Chang, Y. Gu, Z. Han, X. Chen, and T. Ristaniemi, "Context-aware data caching for 5G heterogeneous small cells networks," in *Proc. IEEE ICC*, Jun. 2016, pp. 1–6.
[27] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.
[28] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah. (2016). "On the delay of geographical caching methods in two-tiered heterogeneous networks." [Online]. Available: https://arxiv.org/abs/1605.01110
[29] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 627–640, Apr. 2015.
[30] C. Ge, Z. Sun, N. Wang, K. Xu, and J. Wu, "Energy management in cross-domain content delivery networks: A theoretical perspective," *IEEE Trans. Netw. Service Manage.*, vol. 11, no. 3, pp. 264–277, Sep. 2014.
[31] C. Yang, Z. Chen, Y. Yao, B. Xia, and H. Liu, "Energy efficiency in wireless cooperative caching networks," in *Proc. IEEE ICC*, Jun. 2014, pp. 4975–4980.
[32] A. C. Güngör and D. Gündüz, "Proactive wireless caching at mobile user devices for energy efficiency," in *Proc. ISWCS*, Aug. 2015, pp. 186–190.
[33] 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.
[34] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016. [Online]. Available: http://arxiv.org/abs/1512.06938
[35] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.
[36] X. Li, X. Wang, K. Li, H. Chi, and V. C. M. Leung, "Resource allocation for content delivery in cache-enabled OFDMA small cell networks," in *Proc. IEEE VTC-Fall*, Sep. 2017, pp. 1–6.
[37] 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.
[38] M. Hefeeda and O. Saleh, "Traffic modeling and proportional partial caching for peer-to-peer systems," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1447–1460, Dec. 2008.
[39] A. B. Downey, "The structural cause of file size distributions," in *Proc. IEEE MASCOTS*, Aug. 2001, pp. 361–370.
[40] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.
[41] I. Baev, R. Rajaraman, and C. Swamy, "Approximation algorithms for data placement problems," *SIAM J. Comput.*, vol. 38, no. 4, pp. 1411–1429, 2008.
[42] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*. Hoboken, NY, USA: Wiley, 1990.
[43] National Laboratory for Applied Network Research. *Weekly Squid HTTP Access Logs*. Accessed on Jun. 2013. [Online]. Available: http://www.ircache.net/
[44] 3GPP, "Further advancements for E-UTRA physical layer aspects," 3GPP, France, Tech. Rep. v.9.0.0, Mar. 2010. [Online]. Available: http://www.3gpp.org/

Xiuhua Li (S'12) received the B.S. and M.S. degrees from the Honors School and the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. His current research interests are resource allocation, optimization, distributed antenna systems, cooperative base station caching and traffic offloading in mobile content-centric networks.

Xiaofei Wang (S'06–M'13) received the B.S. degree from the Department of Computer Science and Technology, Huazhong University of Science and Technology, in 2005, and the M.S. and Ph.D. degrees from the School of Computer Science and Engineering, Seoul National University, in 2008 and 2013, respectively. He was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia. He is currently a Professor with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University. His current research interests are social-aware multimedia service in cloud computing, cooperative base station caching, and traffic offloading in mobile content-centric networks.

Keqiu Li (S'04–M'05–SM'12) received the bachelor's and master's degrees from the Department of Applied Mathematics, Dalian University of Technology, in 1994 and 1997, respectively, and the Ph.D. degree from the Graduate School of Information Science, Japan Advanced Institute of Science and Technology, in 2005. He held a post-doctoral position with the University of Tokyo, Japan, for two years.

He is currently a Professor with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, China. He has authored over 100 technical papers in journals, such as the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the *ACM Transactions on Internet Technology*, and the *ACM Transactions on Multimedia Computing, Communications, and Applications*. His current research interests include data center networks, cloud computing, and wireless networks. He is an Associate Editor of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS and the IEEE TRANSACTIONS ON COMPUTERS.

Zhu Han (S'01–M'04–SM'09–F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland at College Park, College Park, MD, USA, in 1999 and 2003, respectively.

From 2000 to 2002, he was a Research and Development Engineer of JDSU, Germantown, MD. From 2003 to 2006, he was a Research Associate with the University of Maryland at College Park. From 2006 to 2008, he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently a Professor with the Electrical and Computer Engineering Department and the Computer Science Department with the University of Houston, Houston, TX, USA. His current research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, the IEEE Leonard G. Abraham Prize in the field of communications systems (best paper award at the IEEE JSAC) in 2016, and several best paper awards in the IEEE conferences. He is currently the IEEE Communications Society Distinguished Lecturer.

Victor C. M. Leung (S'75–M'89–SM'97–F'03) received the B.A.Sc. degree (Hons.) in electrical engineering from The University of British Columbia (UBC) in 1977 and the Ph.D. degree in electrical engineering in 1982. He received the APEBC Gold Medal as the head of the graduating class with the Faculty of Applied Science. He attended the Graduate School, UBC, on a Canadian Natural Sciences and Engineering Research Council Postgraduate Scholarship.

He was a Senior Member of Technical Staff and a Satellite System Specialist with MPR Teltech Ltd., Canada, from 1981 to 1987. In 1988, he was a Lecturer with the Department of Electronics, The Chinese University of Hong Kong. He returned to UBC as a faculty member in 1989, and currently holds the position of Professor and TELUS Mobility Research Chair in advanced telecommunications engineering with the Department of Electrical and Computer Engineering. He has co-authored over 1000 journal/conference papers, 37 book chapters, and co-edited 12 book titles. His research interests are in the broad areas of wireless networks and mobile systems. Several of his papers had been selected for best paper awards.

Dr. Leung is a registered professional engineer with the Province of British Columbia, Canada. He is a fellow of the Royal Society of Canada, the Engineering Institute of Canada, and the Canadian Academy of Engineering. He received the IEEE Vancouver Section Centennial Award and the 2011 UBC Killam Research Prize. He was a recipient of the 2017 Canadian Award for Telecommunications Research. He is a co-author of papers that received the 2017 IEEE ComSoc Fred W. Ellersick Prize and the 2017 IEEE SYSTEMS JOURNAL Best Paper Award. He was a Distinguished Lecturer of the IEEE Communications Society. He is serving on the editorial boards of the IEEE WIRELESS COMMUNICATIONS LETTERS, the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE ACCESS, *Computer Communications*, and several other journals, and has previously served on the editorial boards of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Wireless Communications Series and Series on Green Communications and Networking, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON COMPUTERS, and the *Journal of Communications and Networks*. He has guest-edited many journal special issues, and provided leadership to the organizing committees and technical program committees of numerous conferences and workshops.