Content-Centric Heterogeneous Fog Networks Relying on Energy Efficiency Optimization

Kunlun Wang, *Member, IEEE*, Jun Li, *Senior Member, IEEE*, Yang Yang, *Fellow, IEEE*, Wen Chen, *Senior Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Next-generation wireless communications are expected to provide reliable high-rate connectivity. The recent concept of fog-enabled architecture comes to rescue by invoking a substantial amount of data storage close to end-user devices for meeting these challenges. Thus, caching popular contents at heterogeneous devices, e.g., fog nodes (FNs) or fog access points (FAPs), constitutes a promising technique of reducing both the traffic and the energy consumption of the backhaul links. Therefore, in this paper, we propose an energy-efficient caching and node association algorithm for cache-aided fog networks. First, we solve the problem of energy-efficient content caching and delivery in the FNs/FAPs in conjunction with a fixed node association strategy, where the FNs communicate either with other FNs by node-to-node (N2N) communications or with the FAPs in their proximity. In both caching scenarios, we investigate the relationship between the caching probability of the file and the energy-efficient content delivery by formulating the associated energy efficiency (EE) optimization problem under caching memory constraints. Then, we derive a joint modulation mode allocation strategy and caching policy for each content caching node and conceive a joint node association and caching algorithm for maximizing the EE. Finally, we quantify both the overall EE and throughput for demonstrating that the proposed caching and transmission strategy achieves significant performance improvements.**

***Index Terms*—Fog networks, content caching, energy efficiency, node-to-node communications, access points**

The work of K. Wang was supported by the National Natural Science Foundation of China (NSFC) under grant 61801463. This work of J. Li was supported by National Key R&D Program under Grant 2018YFB1004800. The work of Y. Yang was supported in part by the National Key Research and Development Program of China under grant 2019YFB1803304, and the National Development and Reform Commission of China (NDRC) under grant "5G Network Enabled Intelligent Medicine and Emergency Rescue System for Giant Cities". The work of W. Chen was supported by the NSFC under grant 61671294. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/N004558/1, EP/P034284/1, EP/P034284/1, EP/P003990/1 (COALESCE), of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom.

K. Wang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China. E-mail: wangkl2@shanghaitech.edu.cn. Y. Yang is with Shanghai Institute of Fog Computing Technology (SHIFT), ShanghaiTech University, Shanghai 201210, China, and the Research Center for Network Communication, Peng Cheng Laboratory, Shenzhen 518000, China. E-mail: yangyang@shanghaitech.edu.cn. J. Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, CHINA, and the School of Computer Science and Robotics, National Research Tomsk Polytechnic University, Tomsk, 634050, RUSSIA. Email: jun.li@njust.edu.cn. W. Chen is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: wenchen@sjtu.edu.cn. L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U. K. E-mail: lh@ecs.soton.ac.uk. (Corresponding author: Jun Li.) Part of this work was presented at IEEE VTC2020-Fall, Victoria, BC Canada, Oct. 2020.

I. Introduction

WIRELESS traffic has experienced a tremendous growth in recent years due to the proliferation of hand-held devices connected to the Internet. This traffic increase is expected to continue steadily during the forthcoming years, hence in excess of 4.8 zettabytes of mobile traffic is forecast worldwide for the year 2022 [1]. With the exponential growth of data traffic, energy efficiency (EE) has emerged as a key figure of merit and has become the design metric of green wireless communication systems [2]. On the other hand, given the fact that often a large fraction of content requests aim for downloading cached files, caching at the wireless edge has become a popular content delivery technique [3, 4]. Specifically, caching the most popular content is capable of improving the quality of experience (QoE) for the users, whilst simultaneously improving both the throughput and the EE, as demonstrated in [5, 6]

Regarding the exponential increase in demand for popular content, next-generation wireless systems must provide fast, reliable and sustainable wireless connection. The most promising solutions of meeting these challenges are constituted by device caching and small-cell base station (SBS) caching, which have attracted intensive research attention [7–9]. In case of device-based caching, the user equipment (UE) communicate directly with the aid of device-to-device (D2D) communication. D2D communication relies on direct information exchange between devices without their data being routed to the evolved NodeB (eNB), hence offloading the traffic from the base stations (BS). These D2D user equipments (DUEs) are capable of using the same resources as the cellular users under the control of the eNB, i.e., underlaying the coverage of the cellular BSs. As a benefit, D2D communication is capable of improving the area spectral efficiency of the system [7], whilst maintaining efficient load balancing. Therefore, D2D communication constitutes a promising technique of achieving the ambitious goals of next-generation wireless networks. On the other hand, driven by the development of heterogeneous cell sizes of micro-, pico- and femto-cells, the large number of traditional SBSs, may be exploited as fog access points (FAPs) in the process of the ongoing small-cell densification, which allows a more intense spatial reuse of the resources. Consequently, the beneficial concept of fog computing networks has been proposed [10], which enhances the cloud radio access network (C-RAN) architecture by allowing the remote radio heads (RRHs) to be equipped with caching and signal processing functionalities. The resultant fog caching is a promising

technique of alleviating the heavy network traffic. Clearly, analyzing the network's tele-traffic and proactively caching popular content locally both at the FNs and FAPs is capable of significantly reducing the backhaul traffic, especially when the conventional network is inundated with requests for the cached content. In recent years fog networking has also gained popularity in a diverse range of vehicular scnearios [11–14].

In this context, our proposed framework is focused on the general family of collaborative fog networks relying on heterogeneous nodes, e.g., FNs and FAPs, sharing storage and spectral resources. These nodes cache a subset of popular contents depending on their limited storage and computing capabilities. Fog computing systems are envisioned to act as a promising enabler of real-time applications due to the proximity of fog nodes to the IoT/end-user devices, which can fill the gap between the cloud and the things to enable a service continuum [10, 15].

A. Related Work

The idea of caching popular content at the edge of the network is gaining momentum as one of the most promising enablers of next-generation networks [8, 16, 17]. The content can be cached at the UEs, which is termed as device-based caching [18]. Generally speaking, device-based caching mitigates the cellular traffic by offloading the most popular content from the BSs, thereby increasing the access rate of the service requests and reducing the energy consumption of the BSs [8]. More particularly, device caching enhances the experience of the UEs by reducing the delays, because the cached content can be reached near-instantaneously through D2D communication from local caches. Although the explicit offloading throughput benefits have indeed been shown in previous research efforts relying on device caching [18–21], the EE has not been explicitly considered. To characterize the bandwidth efficiency of a cache-enabled D2D network, the authors of [18] have shown that it scales linearly with the network size, provided that their content requests are not uniformly distributed. The problem of UE throughput maximization was shown analyzed in [19], provided that the size of the UE cache is larger than the library and that only a low outage probability is allowed. Consequently, the throughput is scaling proportionally with the UE cache size. When considering the D2D link scheduling and power allocation, the authors of [7] solved the problem of system throughput maximization. By contrast, the problem of offloading maximization was solved by the authors of [21], relying on an interference-aware reactive caching mechanism.

In addition to device-based caching, the content can also be stored at SBSs. By relying on SBS caching, we can reduce the number of transmissions from the core network and alleviate the backhaul constraint of the small-cells. In [22], an SBS-caching scheme has been proposed, Shanmugam *et al.* analyzed the optimum way of assigning files to the SBSs for minimizing the expected downloading time for files. However, the energy consumption imposed by downloading files was not considered. Hajri and Assaad in [23] proposed an optimization framework for deriving the optimal active SBS density vector in order to maximize the EE. Liu and Yang in [5] have found the condition of when EE can indeed benefit from caching as well as EE-memory relationship, and the maximal attainable EE of caching. In [24], Gregori *et al.* determined the prefetching and local caching gains either by caching at the SBSs, or directly by the user devices, which determines the optimal transmission and caching policies that minimize a specific cost function, such as the energy or throughput attained. In [25], Zhao *et al.* jointly optimized the resource allocation and remote radio heads (RRH) association, Gabry *et al.* in [26] considered the minimization of two fundamental metrics: the expected backhaul rate and the energy consumption, the content caching in user devices is not considered in both of the above literature. In [15, 27, 28], fog computing-based content delivery and task scheduling wireless networks have been proposed in order to improve the overall system performance in terms of the service delay and EE. However, joint energy-efficient content caching and delivery have not been considered in these contributions. In contrast to most of the existing literature, our proposed energy-efficient caching model relies on general fog networks, including caching at the FNs and FAPs, which are capable of simultaneously arranging for energy-efficient node association, content caching and delivery with the aid of sharable storage and spectral resources.

B. Main Contributions

Although the above discussions have demonstrated the benefits of edge caching, the joint optimization of content caching and transmission maximizing the EE in heterogeneous fog frameworks has not been considered to the best of our knowledge, even though it is a key figure of merit in wireless networks [27, 29–31], which is particularly important for battery-limited FNs or FAPs. The proposed heterogeneous fog architecture includes all heterogeneous nodes, e.g., wearable devices, mobile phones and vehicular terminals. The total energy consumption includes both the transmission energy and circuit energy. The circuit energy consumption is the energy consumed by the circuit blocks along the signal path. Additionally, to characterize the node association effects in heterogeneous fog networks, we employ a joint node association and content caching policy for modelling the EE. Based on fixed node association, we aim for studying how heterogeneous fog nodes tackle the challenge of numerous delivery requests fetching a few popular content files. Specifically, we first aim for exploiting the optimal cache placement strategy in order to minimize the traffic burden and then maximize the EE of fog networks. Secondly, we would like to characterize the relationship of content placement and energy-efficient delivery strategy. Finally, we aim for jointly optimizing the node-association and content placement-delivery policy.

Against the above backdrop, our contributions can be summarized as follows:

- In terms of the system model, we propose a novel energy-efficient caching framework for heterogenous fog networks defined by a group of heterogeneous nodes with sharable storage and spectral resources to cache the popular files. Furthermore, we propose a problem

formulation for maximizing the EE of cache-aided fog networks, which offers a new approach to content caching and delivery in fog networks.

- In terms of the mathematical framework and theoretical analysis, we propose an adaptive content delivery strategy for optimizing the EE, which can select the most energy-efficient modulation mode for information transmissions. To the best of our knowledge, considering the modulation alphabet-size of probabilistic caching in energy-efficient fog networks has not been considered in the open literature.
- Additionally, we optimize the node association policy based on the fixed caching policy. Based on the optimal solutions advocated, we propose an algorithm for jointly optimizing the content caching and node association, which maximizes the EE of heterogeneous fog networks.
- We evaluate the performance of the proposed caching and delivery strategy through extensive simulations. Our simulation results demonstrate that the proposed strategy achieves significantly better EE than the traditional strategies subject to realistic caching constraints and relying on diverse system parameters.

C. Paper Organization

The rest of the paper is organized as follows. Section II introduces the system model, while Section III presents our problem formulation and analysis. Section IV is focused on caching at the FNs, including the EE analysis and EE optimization. In Section V, we optimize caching for the FAPs. Our EE analysis is provided in Section V-A and the resultant EE optimization problem is solved in Section V-B. Section VI presents our joint node association and content caching algorithm. In Section VII, we provide our simulation results, and our conclusions are offered in Section VIII. Table I lists the frequently used notations.

II. System Model

A. Network Model

As shown in Fig. 1, we consider a heterogeneous fog network consisting of FNs and FAPs, where the active FNs either rely on N2N communications or are served by the FAPs for content exchange, let us denote the location set of FAPs by $\mathcal{M}$. Each FN and FAP has a local cache memory of size C_{u} and a cache capacity of Q_j units for storing popular files, respectively. Let us assume that the FNs are distributed according to a homogeneous Poisson Point Process (PPP) Φ_{u} [32], where the intensity of Φ_{u} is λ_{u}. Specifically, a FN may be classified as (1) a content request node (RN), (2) an inactive neighbor helper node (HN), and (3) an idle node (IN) with spare spectrum. For every FN of Φ_{u}, the probability of active requests for a file is $\rho \in [0, 1]$, and the HNs will serve as potential transmitters. Let us denote the location sets of the HNs and the RNs by $\mathcal{N}$ and $\mathcal{N}_r$, respectively. [1]

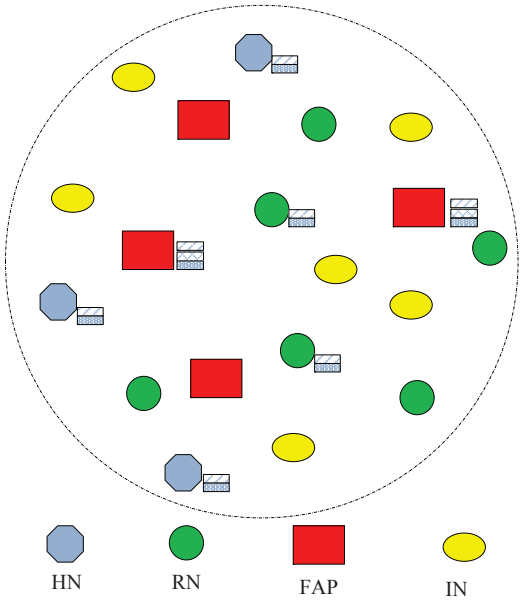


Fig. 1. System model.

We consider a finite set of content files $\mathcal{F} = \{1, \cdots, F\}$ consisting of F files each with N_f bits that all FNs in the network may request, which are indexed in descending order according to their popularity, i.e., the fth file is the fth most popular file. Every FN retains a cache of C_{u} files, and the library size is larger than C_{u}. We assume that the files requested by the RNs are from the content library, and the probability of their requests obeys the Zipf distribution [33–35]. According to this model, after ordering the files according to their decreasing popularity, the probability that the fth file is requested is

$$p_f = \frac{1}{f^{\beta} \sum_{n \in \mathcal{F}} n^{-\beta}}, \tag{1}$$

where β is the popularity exponent of the Zipf distribution. This parameter characterizes the skew of the popularity distribution. Considering the geographic caching strategy of [32] at both the HNs and FAPs, the fth file has a certain probability of $c_{f,k}$ and $z_{f,j}$ to be independently cached by each HN k, $k \in \mathcal{N}$, and FAP j, $j \in \mathcal{M}$. Following the probabilistic caching approach of [36], each user device applies the geographic caching strategy independently caching file with a certain probability. Additionally, the caching probabilities have to satisfy $z_{f,j} \in [0, 1]$ and $c_{f,k} \in [0, 1]$. Due to the limited size of the caching storage, we have $\sum_{f \in \mathcal{F}} c_{f,k} \leq C_{\mathrm{u}}$, $\forall k \in \mathcal{N}$ and $\sum_{f \in \mathcal{F}} z_{f,j} \leq Q_j$.

[1] In the cloud computing systems such as C-RAN, the edge nodes, e.g., HNs, are connected to the cloud processor by fronthaul links. While the RNs request a file, it causes large latencies due to fronthaul transmission. Recognizing that next generation wireless networks are expected to cater to a broad range of quality of service requirements for mobile broadband communication, a hybrid architecture of fog computing systems is proposed [10]. Fog computing organizes and manages the local data storage, computing, communication and networking. In this case, the content of the popular files is prefetched in a cache-enabled fog network with the help of FNs and FAPs, which can enhance the overall system efficiency.

TABLE I
FREQUENTLY USED NOTATION

Definition	Notation	Definition	Notation
Location set of FAPs	$\mathcal{M}$	Cache memory size of FN	C_u
Cache memory size of FAP	Q_j	Location set of HNs	$\mathcal{N}$
Location set of RNs	$\mathcal{N}_r$	Set of content files	$\mathcal{F}$
Total number of files	F	File size	N_f
Popularity exponent of the Zipf distribution	β	Set of resource blocks	$\mathcal{N}_m$
Bandwidth of a resource block	W	Noise power	σ^2
Circuit power consumption	P_0	Battery lifetime of HN k	T_k
The probability that the FAP caches the fth file	$z_{f,j}$	Transmit power of FAP j destined to the RN i	$P_{i,j}$
modulation scheme	$b_{i,j}$	binary association flag between RN i and HN k	$y_{i,k}$
The probability that the HN k caches the fth file	$c_{f,k}$	The request probability of the fth file from the HN k	$p_{f,k}$

As illustrated above, the RN can fetch a file from the content library in three realistic ways. Firstly, the requested file may be cached in its own storage, hence it is retrieved locally with a negligible delay; The second case is via N2N transmission, when the requested file is cached in one of the nearby HNs within a certain distance R_{d}, but not cached in its own storage, provided that the N2N link can be established. Generally speaking, if more than one HNs have the requested file, the nearest one will be selected to transmit the file. In the third scenario the file is performed at the FAP, when the RNs' requested file is not locally available.

B. EE of HN Caching

Let us denote the set of resource blocks (RBs) by $\mathcal{N}_m = \{1, 2, \cdots, N_{\mathrm{m}}\}$, where N_{m} is the total number of RBs from the INs. Note that a RB is formed by a single time-slot as well as 12 subcarriers, and we assume that the bandwidth of a RB is W. Upon involving RB resource reuse, the rth RB allocated to IN l is reused by HN k. In order to transmit the cached file from HN k on the rth RB, the transmit power is set to $P_{k,r}$. Let σ^2 denote the noise power, and the resulting average data rate of the HN k on the rth RB is denoted as $R_{k,r}$.

When aiming for energy-efficient transmission, the battery lifetime is beneficially extended. In quantifying the battery lifetime both the circuit power and caching power dissipation play important roles. Based on [2], we can estimate the circuit power consumption of all the circuit blocks. In this work, we assume that all the nodes have the same constant circuit power consumption P_0, and the caching power consumption is proportional to the cache capacity expressed as $P_{\mathrm{ca}}C_{\mathrm{u}}N_f$, where P_{ca} is the power coefficient in Watt/bit and the value of P_{ca} strongly depends on the caching hardware technology. Similar to the traditional EE metric, the transmission power of the HN is also considered. Thus, both the circuit power and the transmission power are taken into account in the total power consumption. Furthermore, in order to capture the non-linear effect of battery charge on EE, Peukert's law can be used for modelling the battery lifetime T [37], which can be expressed as

$$T = \frac{B_{\mathrm{c}}}{D_{\mathrm{c}}^{\alpha}}, \tag{2}$$

where B_{c} and D_{c} are the battery capacity and the discharge current, respectively, while the exponent α is a constant around 1.3. If the average operating voltage is V_0, we can express the expected battery lifetime T_k of HN k as

$$T_k = \frac{B_{\mathrm{c},k}V_0^{\alpha}}{(P_{k,r} + P_0 + P_{\mathrm{ca}}C_{\mathrm{u}}N_f)^{\alpha}}. \tag{3}$$

Our goal is to maximize the expected amount of transmitted data during the battery lifetime. Then, the EE of content exchange from RN i associated with HN k can be expressed as

$$U_{i,k} = R_{k,r}T_k = \frac{R_{k,r}B_{\mathrm{c},k}V_0^{\alpha}}{(P_{k,r} + P_0 + P_{\mathrm{ca}}C_{\mathrm{u}}N_f)^{\alpha}}, \forall i \in \mathcal{N}_r. \tag{4}$$

Consequently, there is a fundamental tradeoff between the total power consumption and the average data transmission rate. Note that EE is the ratio of the throughput to the power consumption, hence (4) is also an EE metric. The reason why we use (4) instead of the traditional definition of EE is that capturing the non-linear effects of power consumption is beneficial for the lifetime of battery-driven devices. In comparison to the traditional instantaneous EE, this metric also characterizes the network's average amount of data transmitted during the nodes' lifetime, which is better in term of characterising the EE of caching systems.

C. EE of FAP Caching

Similar to HN caching, a probabilistic caching policy is considered in the FAP. We consider N_j RNs requesting files selected by the FAP j to be served simultaneously, and each FAP has L antennas, with $L \geq N_j$. Similarly, the FAP independently selects files from the set $\mathcal{F}$ for caching according to a specific probability distribution. To unify our analysis, we denote the probability that the FAP caches the fth file by $0 \leq z_{f,j} \leq 1$. When $\boldsymbol{z}_j = [z_{f,j}]_{f\in\mathcal{F}}$ is given, the FAP determines which particular files should be cached using the method of [32].

Remark 1. *In fog computing networks, several FAPs have the same non-orthogonal spectral resources. However, in this contribution we assume that each FAP is assigned orthogonal resources, so that the FAPs have non-overlapping coverage areas. Hence, energy efficient solutions can be obtained according to our specific EE optimization problem for each FAP. Although using cooperative transmission among the FAPs having overlapping coverage areas and activating multicast transmissions to different FAPs is capable of achieving beneficial extra caching gains, coordinating the design of the FAPs' caching policies is beyond the scope of this treatise.*

An uncoded caching strategy is utilized for the files' transmission over the links, where the FAP will send the requested files to the RNs. Let us denote the precoding vector and the transmit power of FAP j destined to the RN i by $\mathbf{w}_{i,j} \in \mathbb{C}^{L\times 1}$ and $P_{i,j}$, respectively. Then the signal-to-interference-plus-noise ratio of RN i associated with FAP j can be formulated as

$$\gamma_{i,j} = \frac{|\mathbf{h}_{i,j}^H \mathbf{w}_{i,j}|^2}{\sum_{l\neq i} |\mathbf{h}_{l,j}^H \mathbf{w}_{i,j}|^2 + \sigma^2}. \tag{5}$$

As a benefit of its low computational complexity, zero-forcing (ZF) beamforming is considered. Since the direction of the beamforming vectors is defined by the ZF, only the transmit power of each beam has to be optimized. Given the ZF beamforming vector, the precoding vector of RN i is given as $\mathbf{w}_{i,j} = \sqrt{P_{i,j}}\mathbf{a}_{i,j}$, where $\mathbf{a}_{i,j}$ is the ZF beamforming vector of RN i associated with FAP j, which is the ith column of the matrix $\mathbf{H}_j^H(\mathbf{H}_j\mathbf{H}_j^H)^{-1}$, with $\mathbf{H}_j = [\mathbf{h}_{1,j}, \cdots, \mathbf{h}_{N_j,j}]^T$, where $\mathbf{h}_{i,j}$ denotes the channel vector between FAP j and the RN i.

Given the ZF beamforming, we can obtain the downlink signal-to-noise ratio (SNR) at RN i as

$$\gamma_{i,j} = \frac{P_{i,j}\delta_{i,j}}{\sigma^2}, \quad \forall \delta_{i,j} \in \boldsymbol{\delta}, \tag{6}$$

where $\boldsymbol{\delta} = [\delta_{1,j}, \delta_{2,j}, \cdots, \delta_{N_j,j}]$ is the eigenvalue vector of $\mathbf{H}_j\mathbf{H}_j^H$. Based on the Chernoff upper bound [38], we can approximate the symbol error rate (SER) ϵ as:

$$\epsilon = 2(1-2^{-b_{i,j}/2})e^{-\frac{3}{2^{b_{i,j}}-1}\frac{\gamma_{i,j}}{2}}. \tag{7}$$

Upon substituting (6) into (7), the average transmit power of the FAP j for the RN i can be approximated as

$$P_{i,j} = \frac{\sigma^2}{\delta_{i,j}} \frac{2(2^{b_{i,j}}-1)}{3} \ln \frac{2(1-2^{-b_{i,j}/2})}{\epsilon}, \tag{8}$$

where $b_{i,j}$ characterizes the modulation scheme. Similarly, the circuit power and the caching power consumption of FAP j are given by P_0 and $P_{\text{ca}}Q_jN_f$, respectively. Similarly, the expected transmission time T_j of FAP j can be expressed as

$$T_j = \frac{B_{\text{c},j}V_{0,j}^{\alpha}}{(P_{i,j}+P_0+P_{\text{ca}}Q_jN_f)^{\alpha}}, \tag{9}$$

where $B_{\text{c},j}$ and $V_{0,j}$ are the battery capacity and the average operating voltage of FAP j, respectively. Then, the EE of content delivery from FAP j to RN i can be expressed as

$$U_{i,j} = \bar{R}_{i,j}T_j, \tag{10}$$

where $\bar{R}_{i,j}$ is the average data rate from FAP j to RN i.

III. Problem Formulation and Analysis

A. Problem Formulation

In this section, we discuss the node association, the content placement constraints and formulate the related optimization problem. Then, the EE is formulated to maximize the system level benefits.

Let $x_{i,j} \in \{0,1\}$ be a binary variable, which can be defined as

$$x_{i,j} = \begin{cases} 1, & \text{if RN } i \text{ is associated to FAP } j, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Similarly, $y_{i,k} \in \{0,1\}$ denotes the binary association flag between RN i and HN k. We assume that every RN can only be served by a single transmitter, which is formulated as

$$\sum_{j\in\mathcal{M}} x_{i,j} + \sum_{k\in\mathcal{N}} y_{i,k} \leq 1, \forall i \in \mathcal{N}_{\text{r}}. \tag{12}$$

Similarly, for N2N-aided communications, only a single RN can be linked to one HN, which can be formulated as

$$\sum_{i\in\mathcal{N}_r} y_{i,k} \leq 1, \forall k \in \mathcal{N}. \tag{13}$$

Let $U_{i,k}$ and $U_{i,j}$ denote the EE between RN i and HN k or FAP j. Then the EE of the associated fog networks can be formulated as

$$U_{\text{ee}} = \sum_{i\in\mathcal{N}_r}\sum_{k\in\mathcal{N}} y_{i,k}c_{f,k}U_{i,k} + \sum_{i\in\mathcal{N}_{\text{r}}}\sum_{j\in\mathcal{M}} x_{i,j}z_{f,j}U_{i,j}. \tag{14}$$

Hence, the EE optimization problem of fog networks is expressed as:

$$\begin{aligned} \max_{x_{i,j},y_{i,k},z_{f,j},c_{f,k},b_{k,r},b_{i,j}} \quad & U_{\text{ee}}, \\ \text{s.t.} \quad \textstyle\sum_{j\in\mathcal{M}} x_{i,j} + \sum_{k\in\mathcal{N}} y_{i,k} &\leq 1, \forall i \in \mathcal{N}_{\text{r}}, \\ \textstyle\sum_{i\in\mathcal{N}_{\text{r}}} y_{i,k} &\leq 1, \forall k \in \mathcal{N}, \\ b_{k,r}, b_{i,j} &\in [b_{\min}, b_{\max}], \\ x_{i,j}, y_{i,k} &\in \{0,1\}, \forall i \in \mathcal{N}_{\text{r}}, \\ z_{f,j}, c_{f,k} &\in [0,1], \forall f \in \mathcal{F}, \\ \textstyle\sum_{f\in\mathcal{F}} z_{f,j} &\leq Q_j, \forall i \in \mathcal{N}_{\text{r}}, \forall j \in \mathcal{M}, \\ \textstyle\sum_{f\in\mathcal{F}} c_{f,k} &\leq C_{\text{u}}, \forall i \in \mathcal{N}_{\text{r}}, \forall k \in \mathcal{N}, \end{aligned} \tag{15}$$

where $b_{k,r}$ and $b_{i,j}$ represent the modulation scheme selected for the HN k at the rth RB and the modulation scheme of the RN i associated with the FAP j, respectively.

B. Problem Analysis

With the problem formulation described above, we can analyze Problem (15), where $x_{i,j}$ and $y_{i,k}$ are binary variables, hence the feasible set of Problem (15) is non-convex. The objective function is not convex due to the product-based relationship between binary variables. The global optimum of a mixed discrete and non-convex optimization problem is challenging to find [39]. Thus, we have to simplify Problem (15).

By exploiting the fact that the RN-FAP and RN-HN association constraints and the caching constraints are separable, we propose an iterative algorithm for optimizing the content placement at each node assuming a fixed node association policy, as described below.

1) Caching Relying on a Fixed Association Policy: Given the fixed associations $\{x'_{i,j}\}$ and $\{y'_{i,k}\}$, $\forall i \in \mathcal{N}_{\text{r}}$, $\forall k \in \mathcal{N}$, the caching problem is coupled among the FAPs and HNs. In particular, the content placement problem formulated for HNs can be written as

$$\begin{aligned} \max_{z_{f,k},b_{k,r}} \quad & U_{\text{ee}}^{\text{D}} = \textstyle\sum_{i\in\mathcal{N}_{\text{r}}}\sum_{k\in\mathcal{N}} y'_{i,k}c_{f,k}U_{i,k}, \\ \text{s.t.} \quad b_{k,r} &\in [b_{\min}, b_{\max}], \forall k \in \mathcal{N}, \forall r \in \mathcal{N}_{\text{m}}, \\ c_{f,k} &\in [0,1], \forall f \in \mathcal{F}, \forall k \in \mathcal{N}, \\ \textstyle\sum_{f\in\mathcal{F}} c_{f,k} &\leq C_{\text{u}}, \forall k \in \mathcal{N}, \end{aligned} \tag{16}$$

where $U_{\rm ee}^{D}$ and $c_{f,k}$ represent the EE of the HNs-based caching network and the caching result for the fth file at the HN k, respectively.

Similarly, the content placement problem of FAPs can be written as

$$\begin{array}{lll} \max\limits_{z_{f,j},b_{i,j}} & U_{\rm ee}^{\rm B} = \sum_{i\in\mathcal{N}_{\rm r}}\sum_{j\in\mathcal{M}} x'_{i,j} z_{f,j} U_{i,j}, \\ \text{s.t.} & b_{i,j} \in [b_{\min}, b_{\max}], j\in\mathcal{M}, \\ & z_{f,j} \in [0,1], \forall f\in\mathcal{F}, j\in\mathcal{M}, \\ & \sum\limits_{f\in\mathcal{F}} z_{f,j} \leq Q_j, \forall j\in\mathcal{M}, \end{array} \tag{17}$$

where $U_{\rm ee}^{B}$, $b_{i,j}$, $z_{f,j}$ and $\mathcal{M}$ represent the EE of the FAP-based caching network, the modulation scheme selection by the FAP j, the caching result of the fth file at the FAP j and the set of FAPs, respectively.

2) Node Association with Fixed Caching Policy: Given the fixed content placement at each HN and FAP as $\{c'_{f,k}\}$ and $\{z'_{f,j}\}$, the association problem can be stated as

$$\begin{array}{lll} \max\limits_{x_{i,j},y_{i,k}} & \sum\limits_{i\in\mathcal{N}_{\rm r}}\sum\limits_{k\in\mathcal{N}} y_{i,k}c'_{f,k}U_{i,k} + \sum\limits_{i\in\mathcal{N}_{\rm r}}\sum\limits_{j\in\mathcal{M}} x_{i,j}z'_{f,j}U_{i,j}, \\ \text{s.t.} & \sum_{j\in\mathcal{M}} x_{i,j} + \sum_{k\in\mathcal{N}} y_{i,k} \leq 1, \forall i\in\mathcal{N}_{\rm r}, \\ & \sum_{i\in\mathcal{N}_{\rm r}} y_{i,k} \leq 1, \forall k\in\mathcal{N}, \\ & \sum_{i\in\mathcal{N}_{\rm r}} x_{i,j} = N_j, \forall j\in\mathcal{M}, \\ & x_{i,j}, y_{i,k} \in \{0,1\}, \forall i\in\mathcal{N}_{\rm r}. \end{array} \tag{18}$$

Observe that Problem (18) is a mixed integer programming problem, hence obtaining the globally optimal solution is challenging. As a result, we adopt a similar strategy to that used in [40] for solving a partially dualized version of the original problem with respect to the constraints.

IV. Maximal EE of HN Caching

In this section, we conceive an energy-efficient content caching and delivery policy for the HNs relying on fixed user association.

A. N2N Transmission Probability Analysis

It should be noted that if a random RN requests a file, it can fetch it from its local cache. However, if the requested content is not cached in the local storage, the RN would get the content from neighbour HN via N2N transmission. Under this condition, the probability of fetching the requested file from the HN is given by the probability that the RN relies on the N2N transmission mode.

Under probabilistic caching, the probability of finding a file cached by the HN strongly depends on the popularity of the file and on the distance between them. Next, upon denoting the fth requested file by f_f, the caching probability of the fth file from the HN k by $c_{f,k}$ and the request probability of the fth file from the HN k by $p_{f,k}$, respectively, the probability of finding the requested file cached in the HN k within distance R_d can be calculated as [41]

$$\zeta_{k,f} = 1 - e^{-\pi(1-\rho)\lambda_{\rm u} c_{f,k} R_{\rm d}^2}, \tag{19}$$

where ρ represents the active probability of the neighbor HNs. Note that the content library has F files, and each file can be cached by HNs based on its request probability. Upon averaging over all the files, we arrive at the content exchange probability from the HN k as $\zeta_k = \sum_{f\in\mathcal{F}} p_{f,k}(1-c_{f,k})\zeta_{k,f}$, which is given by

$$\zeta_k = \sum_{f\in\mathcal{F}} p_{f,k}(1-c_{f,k})\left(1 - e^{-\pi(1-\rho)\lambda_{\rm u} c_{f,k} R_{\rm d}^2}\right). \tag{20}$$

B. Power and Throughput Analysis

In order to derive the EE, we should first calculate the power consumption of content delivery. The probability that the fth file requested by the RN will be found within the HN k, but not cached in its own local cache is quantified by $\zeta_{k,f}$, which is given by (19). Based on the content caching probability of the HN k, the transmit power of the fth file can be quantified as $P_{k,r}\zeta_{k,f}$, where $P_{k,r}$ denotes the transmit power of the HN k on the rth RB. Averaging over all files in the content library with the aid of the N2N request probability, we can derive the average transmit power of the HN k as

$$\begin{aligned} P_k &= \sum_{f\in\mathcal{F}} P_{k,r} p_{f,k}(1-c_{f,k})\zeta_{k,f}, \\ &= P_{k,r}\zeta_k. \end{aligned} \tag{21}$$

We assume that an adaptive modulation scheme is used by each HN, which has been actively used in wireless standards [29]. It is widely recognized that [42] the SER ϵ of M-ary quadrature amplitude modulation (MQAM) having an alphabet size $2^{b_{k,r}}$, is given by

$$\epsilon = 2(1-2^{-b_{k,r}/2})Q\left(\sqrt{\frac{3}{2^{b_{k,r}}-1}\gamma_{k,r}}\right), \tag{22}$$

where $Q(\cdot)$ is the complementary cumulative distribution function (CCDF) of the standard Gaussian random variable and $b_{k,r}$ is the modulation alphabet size for HN k on the rth RB. In order to derive the average caching probability distribution for the content, we assume that the channel experiences an inlarge second-order path loss law [38]. Let us denote the transmit power of the IN l on the corresponding rth RB by $P_{l,r}$. Then we can compute the received signal-to-interference-plus-noise ratio (SINR) $\gamma_{k,r}$ for the HN k on the rth RB as

$$\gamma_{k,r} = \frac{P_{k,r}\phi(d_{k,k})}{P_{l,r}\phi(d_{l,k}) + \sigma^2}, \tag{23}$$

where $\phi(d_{k,k}) = K_{\rm dB} - 10\gamma\log_{10}\frac{d_{k,k}}{d_0}$, and the parameters $K_{\rm dB}$, γ and d_0 represent the shadowing effects, path loss exponent and reference distance, respectively [38]. Based on the Chernoff upper bound, the transmission symbol error rate can be approximated as (in the high SINR regime)

$$\epsilon \leq 2(1-2^{-b_{k,r}/2})e^{-\frac{3}{2^{b_{k,r}}-1}\frac{\gamma_{k,r}}{2}}. \tag{24}$$

Therefore upon substituting (23) into (24), the closed-form expression of the transmission power of the HN k is given by

$$P_{k,r} \approx \frac{I_{k,r}\left(P_{l,r}\phi(d_{l,k})+\sigma^2\right)}{\phi(d_{k,k})}\frac{2(2^{b_{k,r}}-1)}{3}\ln\frac{2(1-2^{-b_{k,r}/2})}{\epsilon}, \tag{25}$$

where $I_{k,r} \in \{0,1\}$ is the indicator of the rth RB allocated to the HN k. Thus, the average transmit power of the HN k can be approximated as

$$P_{k,r} = \frac{I_{k,r}\zeta_k \left(P_{l,r}\phi(d_{l,k}) + \sigma^2\right)}{\phi(d_{k,k})} \frac{2(2^{b_{k,r}} - 1)}{3} \ln \frac{2(1 - 2^{-b_{k,r}/2})}{\epsilon}. \tag{26}$$

We assume that each packet of the content exchange contains L bits. Let W denote the transmission bandwidth. Then the transmission time per cached packet from HN k can be obtained as

$$t_{\mathrm{L}} = \frac{L}{W b_{k,r} \zeta_k}. \tag{27}$$

In this continuation, the cached packet contains an overhead of L_{p} bits. Since not all the transmitted data in the packet are information bits, we can get the successful transmission probability for a cached packet as $S_{\mathrm{p}} = (\zeta_k(1-\epsilon))^{\frac{L}{b_{k,r}}}$. Taking into account the overheads, the effective throughput $R_{k,r}$ can be defined as the payload information of cached files that can be correctly received per second, which is given as [43–45]:

$$R_{k,r} = I_{k,r}\frac{L_{\mathrm{p}} S_{\mathrm{p}}}{t_{\mathrm{L}}} = I_{k,r}\frac{L_{\mathrm{p}}}{L} W b_{k,r} \zeta_k^{(1+\frac{L}{b_{k,r}})} (1-\epsilon)^{\frac{L}{b_{k,r}}}. \tag{28}$$

C. Optimizing Content Caching and Delivery

Substituting (26) and (28) into (4) and denoting ζ_k as $\varphi(\mathbf{c})$, we arrive at the EE of the caching networks relying on the HNs as

$$U_{\mathrm{ee}}^{\mathrm{D}} = \sum_{k\in\mathcal{N}} \sum_{r\in\mathcal{R}} \frac{I_{k,r}(\varphi(\mathbf{c}))^{1+\frac{L}{b_{k,r}}}\theta_{k,r}}{\left[I_{k,r}\varphi(\mathbf{c})\left(g_{k,r} + P_0 + P_{\mathrm{ca}} C_{\mathrm{u}} N_f\right)\right]^{\alpha}}, \tag{29}$$

where $\theta_{k,r} = \frac{L_{\mathrm{p}}}{L} b_{k,r} W B_{\mathrm{c},k} V_0^{\alpha} (1-\epsilon)^{\frac{L}{b_{k,r}}}$ and $g_{k,r} = \frac{\left(P_{l,r}\phi(d_{l,k}) + \sigma^2\right)}{\phi(d_{k,k})} \frac{2(2^{b_{k,r}}-1)}{3} \ln \frac{2(1-2^{-b_{k,r}/2})}{\epsilon}$. We assume that the resource allocation of the HNs has already been accomplished by the central controller, and that the rth RB allocated to the HN k results in $I_{k,n} = 0$ for $n \neq r$, which means that each HN can only reuse one RB and this RB cannot be reused by other HNs.

To achieve energy efficient communications, the modulation alphabet size and the caching probability distribution have to be determined for each N2N pair. Given the cache capacity C_{u} of each HN, the number of cached files should satisfy $\sum_{f\in\mathcal{F}} z_{f,k} \leq C_{\mathrm{u}}$. We assume using adaptive modulation and coding (AMC).

Lemma 1. *The optimization Problem* (16) *is equivalent to optimizing the problem*

$$\max_{c_{f,k}, b_{k,r}} \frac{I_{k,r}(\varphi(\mathbf{c}))^{1+\frac{L}{b_{k,r}}}\theta_{k,r}}{\left[I_{k,r}\varphi(\mathbf{c})\left(g_{k,r} + P_0 + P_{ca} C_u N_f\right)\right]^{\alpha}}. \tag{30}$$

Proof: The proof is given in Appendix A. ∎

In order to solve the optimization Problem (30), we first study the characteristics of the N2N transmission probability $\varphi(\mathbf{c})$ via the following lemma.

Lemma 2. $[\varphi(\mathbf{c})]^{\alpha - \frac{L}{b_{k,r}}}$ *is a convex function of* $c_{f,k}$, $\forall f \in \mathcal{F}$.

Proof: The proof is given in Appendix B. ∎

By analyzing Problem (30), we have

Theorem 1. *For the HN k, the optimal caching probability $c_{f,k}^*$, $\forall f \in \mathcal{F}$, and modulation $b_{k,r}^*$ of Problem* (30) *exists, which achieve the optimal EE for the RN-HN transmissions.*

Proof: The proof is given in Appendix C. ∎

Algorithm 1 EE optimization for HNs-based caching

1: For $b_{k,r} = b_{\min} : b_{\max}$

Step 1): Initialization: Let $\mathbf{c}_1$ be a feasible point and $\delta_1 = \frac{I_{k,r}(\varphi(\mathbf{c}_1))^{1+\frac{L}{b_{k,r}}}\theta_{k,r}}{\left[I_{k,r}\varphi(\mathbf{c}_1)(g_{k,r}+P_0+P_{ca}C_{\mathrm{u}}N_f)\right]^{\alpha}}$. Let $v = 1$.

Step 2): Invoking convex programming to solve the following problem:

$$\eta(\delta_v) = \max_{c_{f,k}} \Big\{ I_{k,r}(\varphi(\mathbf{c}))^{1+\frac{L}{b_{k,r}}}\theta_{k,r} - \delta_v \left[I_{k,r}\varphi(\mathbf{c})(g_{k,r} + P_0 + P_{\mathrm{ca}}C_{\mathrm{u}}N_f)\right]^{\alpha} \Big\}. \tag{31}$$

With the aid of convex optimization, we find the solution point $\mathbf{z}_{v+1}$.

Step 3): If the solution $\eta(\delta_v) = 0$, stop and $\mathbf{q}_v$ is optimal. Otherwise, set $\delta_{v+1} = \frac{I_{k,r}(\varphi(\mathbf{c}_{v+1}))^{1+\frac{L}{b_{k,r}}}\theta_{k,r}}{\left[I_{k,r}\varphi(\mathbf{c}_{v+1})(g_{k,r}+P_0+P_{\mathrm{ca}}C_{\mathrm{u}}N_f)\right]^{\alpha}}$, and $v = v+1$, and go to step 2.

2: Energy-efficient caching probability: $c_{f,k} = c_{f,k}^*$, and the modulation mode is: $b_{k,r} = b_{k,r}^*$.

Based on Algorithm 1, the globally optimal solution $b_{k,r}^*$ and $\mathbf{z}^*$ can be found by a one-dimensional search and $\mathbf{c}^*$ is given by a closed-form expression. Hence, the solution has a low complexity.

V. Maximal EE of FAP Caching

Similarly to the previous section, this section considers the energy-efficient content caching and delivery policy in FAPs in conjunction with fixed node association.

A. EE Analysis

According to the content exchange probability relying on FAP caching, the average data rate from FAP j to RN i can be expressed as

$$\bar{R}_{i,j} = \sum_{f\in\mathcal{F}} p_{f,i} z_{f,j} e^{-\pi(1-\rho)\lambda_{\mathrm{u}} z_{f,j} R_{\mathrm{d}}^2} W b_{i,j},\ \forall i \in \mathcal{N}_r,\ \forall j \in \mathcal{M}, \tag{32}$$

where $p_{f,i}$ and R_{d} are the request probability of the fth file from the RN i and the N2N search distance, respectively. Based on (8), (9), (10) and (32), the average EE, which is defined as the expected amount of transmitted data during the FAP lifetime, is given by

$$U_{ee}^{\mathrm{B}} = \sum_{j\in\mathcal{M}} \sum_{i\in\mathcal{N}_r} \sum_{f\in\mathcal{F}} \frac{x'_{i,j} p_{f,i} z_{f,j} e^{-\pi(1-\rho)\lambda_{\mathrm{u}} z_{f,j} R_{\mathrm{d}}^2} W b_{i,j} B_{\mathrm{c},j} V_{0,j}^{\alpha}}{\left(P_{i,j} + P_0 + P_{\mathrm{ca}} Q_j N_f\right)^{\alpha}}. \tag{33}$$

Our aim is to jointly design the transmission policy $\mathbf{b} = \{b_{1,1}, \cdots, b_{N_j,M}\}$ at the FAP, and the local caching policy

$\mathbf{z}_j = \{z_{1,j}, \cdots, z_{F,j}\}$ at the FAP, for maximizing the EE of fog networks. By exploiting that the content caching constraints and modulation constraints are separable, Problem (17) can be decoupled into the sub-problems (Q1), (Q2) and thus can be efficiently solved via optimizing these two problems independently, where (Q1) can be expressed as

$$\begin{aligned} \max_{\mathbf{z}} \quad & S(\mathbf{z}) = \sum_{j\in\mathcal{M}} \sum_{i\in\mathcal{N}_r} \sum_{f\in\mathcal{F}} x'_{i,j} p_{f,i} z_{f,j} e^{-\pi(1-\rho)\lambda_{\mathrm{u}} z_{f,j} R_{\mathrm{d}}^2} b_{i,j}, \\ \text{s.t.} \quad & 0 \leq z_{f,j} \leq 1, f = 1, \cdots, F, \\ & \sum_{f\in\mathcal{F}} z_{f,j} \leq Q_j, \forall j \in \mathcal{M}, \end{aligned} \tag{Q1}$$

and (Q2) is shown at the bottom of this page.

We then have the following remark.

Remark 2. *Based on the problem formulation, the energy-efficient caching problem can be partitioned into a transmission strategy maximizing the EE and caching strategy maximizing the delivery success probability for the RNs. From a heterogeneous fog network perspective, the maximum probability of satisfying RN demand only depends on the cache size, and it is independent of the FAP density, as well as of the content transmit power.*

B. Optimizing Content Caching and Delivery

Lemma 3. *The optimal caching policy* $\mathbf{z}$ *of Problem (Q1) can be obtained as follows:*

$$z_{f,j}^* = \min\left\{ \left[\frac{\mathcal{W}(\frac{\eta_f e}{\sum_{i\in\mathcal{N}_r} x'_{i,j} p_{f,i} b_{i,j}^*}) - 1}{-\pi(1-\rho)\lambda_u R_d^2} \right]^+, 1 \right\}, \tag{34}$$

for $\forall f \in \mathcal{F}$*, where* $[\varpi]^+ = \max\{0, \varpi\}$ *and* $\mathcal{W}$ *is from the definition of the Lambert function [46].*

Proof: The proof is given in Appendix D. ∎

Remark 3. *Since the modulation mode* $b_{i,j}$ *used for each transmission stream is an integer and the number of modulation modes is limited for realistic next-generation standards, we can find the globally optimal solution by enumerating* $b_{i,j}$ *until the value of* $E(\mathbf{b})$ *achieves the maximum. Then, we can derive the optimal* $z_{f,j}^*$, $\forall f \in \mathcal{F}$*, with the optimal* $b_{i,j}^*$, $\forall i \in \mathcal{N}_r$*. Then we achieve the maximum of* $S(\mathbf{z})$ *in (Q1) under the constraints considered.*

Interestingly, it turns out that the parameter $\mathcal{W}(\frac{\eta_f e}{\sum_{i\in\mathcal{N}_r} x'_{i,j} p_{f,i} b_{i,j}^*})$, which depends both on the Lagrange multiplier and on the modulation mode, characterizes the caching policy: if we have

$$\mathcal{W}(\frac{\eta_f e}{\sum_{i\in\mathcal{N}_r} x'_{i,j} p_{f,i} b_{i,j}^*}) \geq 1, \tag{35}$$

then the fth file, $\forall f \in \mathcal{F}$, is not cached at the FAP; while, if we have

$$\mathcal{W}(\frac{\eta_f e}{\sum_{i\in\mathcal{N}_r} x'_{i,j} p_{f,i} b_{i,j}^*}) = 1 - \pi(1-\rho)\lambda_{\mathrm{u}} R_{\mathrm{d}}^2, \tag{36}$$

the fth file, $\forall f \in \mathcal{F}$, is completely cached at the FAP; and, finally, if

$$1 - \pi(1-\rho)\lambda_{\mathrm{u}} R_{\mathrm{d}}^2 < \mathcal{W}(\frac{\eta_f e}{\sum_{i\in\mathcal{N}_r} x'_{i,j} p_{f,i} b_{i,j}^*}) < 1, \tag{37}$$

the FAP j caches the fth file, $\forall f \in \mathcal{F}$, with a probability of $z_{f,j}^*$.

Note that the above energy-efficient content caching policy relies on known node association. In practical situations in which the contents are cached during off-peak hours and accessed during peak hours, the optimization of content caching would require predicting the specific node association.

VI. Optimizing Node Association and Caching Policy

In this section, we study the node association policy with the aid of fixed content caching for the heterogeneous fog networks considered. Then we investigate the joint energy-efficient node association and content placement with the aid of the energy-efficient solutions obtained.

A. Optimizing Node Association

Since problem (18) is a mixed integer programming problem, it is challenging to obtain its globally optimal solution. In this case, we adopt the partially dualized method of [47] to solve Problem (18). Let us introduce dual variables $\boldsymbol{\upsilon} = [\upsilon_1, \cdots, \upsilon_N]$ for the second constraint of Problem (18), and $\boldsymbol{\psi} = [\psi_1, \cdots, \psi_M]$ for the third constraint of Problem (18). The Lagrangian function is given by

$$\begin{aligned} L(\{x_{i,j}\}, \{y_{i,k}\}, \boldsymbol{\psi}, \boldsymbol{\upsilon}) = & \sum_{i\in\mathcal{N}_{\mathrm{r}}} \sum_{k\in\mathcal{N}} y_{i,k} c'_{f,k} U_{i,k} \\ & + \sum_{i\in\mathcal{N}_{\mathrm{r}}} \sum_{j\in\mathcal{M}} x_{i,j} z'_{f,j} U_{i,j} - \sum_{j\in\mathcal{M}} \psi_j \left(\sum_{i\in\mathcal{N}_{\mathrm{r}}} x_{i,j} - N_j \right) \\ & - \sum_{k\in\mathcal{N}} \upsilon_k \left(\sum_{i\in\mathcal{N}_{\mathrm{r}}} y_{i,k} - 1 \right). \end{aligned} \tag{38}$$

In this context, the partially dualized problem of the original Problem (18) is given by

$$\begin{aligned} f(\boldsymbol{\psi}, \boldsymbol{\upsilon}) = & \max_{x_{i,j}, y_{i,k}} \quad L(\{x_{i,j}\}, \{y_{i,k}\}, \boldsymbol{\psi}, \boldsymbol{\upsilon}), \\ \text{s.t.} \ & \textstyle\sum_{j\in\mathcal{M}} x_{i,j} + \sum_{k\in\mathcal{N}} y_{i,k} \leq 1, \forall i \in \mathcal{N}_{\mathrm{r}}. \end{aligned} \tag{39}$$

Then, we can obtain explicit analytical solutions, that are

$$x_{i,j}^* = \begin{cases} 1, & \text{if } U_{i,j^*} - \psi_{j^*} > U_{i,k^*} - \upsilon_{k^*}, \\ 0, & \text{otherwise}, \end{cases} \tag{40}$$

$$\begin{aligned} \max_{\mathbf{b}} \quad & E(\mathbf{b}) = \sum_{j\in\mathcal{M}} \sum_{i\in\mathcal{N}_r} \frac{x'_{i,j} b_{i,j} B_{\mathrm{c},j} V_{0,j}^{\alpha}}{\left(\frac{\sigma^2}{\delta_{i,j}} \frac{2(2^{b_{i,j}}-1)}{3} \ln \frac{2(1-2^{-b_{i,j}/2})}{\epsilon} + P_0 + P_{\mathrm{ca}} Q_j N_f \right)^{\alpha}}, \\ \text{s.t.} \ & b_{i,j} \in [b_{\min}, b_{\max}], \end{aligned} \tag{Q2}$$

and

$$y_{i,k}^{*}=\begin{cases}1, & \text{if } U_{i,k^{*}}-\upsilon_{k^{*}}\geq U_{i,j^{*}}-\psi_{j^{*}},\\ 0, & \text{otherwise,}\end{cases} \tag{41}$$

where $j^{*}=\arg\max_{j\in\mathcal{M}}(U_{i,j}-\psi_j)$ and $k^{*}=\arg\max_{k\in\mathcal{N}}(U_{i,k}-\upsilon_k)$. The optimal dual variables ψ_j^{*} and υ_k^{*} can be found by the subgradient method of [40].

B. Optimizing Node Association and Content Caching

Before proceeding to our optimization results, we would like to mention that the optimal solutions of Problem (39) may not be the optimal solution of Problem (15), because Problem (15) has the nonconvex nature. However, given that the utility-maximization based association approach often produces excellent solutions for the overall problem [40], we assume that the suboptimal solutions (40) and (41) of Problem (39) are acceptable in practice.

Based on the above analysis, we can solve the joint content placement and node association problem. The optimization process is summarized in Algorithm 2, which describes the joint node association and caching design.

Remark 4. *Both the node association and the content placement steps of Algorithm 2 aim for increasing the EE of heterogeneous fog networks. As a result, the overall algorithm is guaranteed to converge. By contrast, due to the nonconvex nature of the original Problem* (15)*, the converged solutions are not capable of achieving the global optimum.*

Algorithm 2 Energy Efficient Node Association and Content Caching

Initialization: Set the initial node association $\{x_{i,j}\}$ and $\{y_{i,k}\}$ without considering content caching;
repeat
Step 1: Fix the node association policy $\{x_{i,j}\}$ and $\{y_{i,k}\}$. Then find the energy efficient content caching policy $\{c_{f,k}^{*}\}$ and $\{z_{f,j}^{*}\}$ according to the optimal solutions from Algorithm 1 and (34).
Step 2: Fix the content caching policy $\{c_{f,k}^{*}\}$ and $\{z_{f,j}^{*}\}$ at each HN and FAP from Step 1. Then update the node association policy $\{x_{i,j}\}$ and $\{y_{i,k}\}$ by solving (39).
until convergence.

C. Complexity Analysis

This subsection briefly analyzes the computational complexity of the proposed algorithms. After obtaining the content caching policy $\{c_{f,k}^{*}\}$ and $\{z_{f,j}^{*}\}$ from Algorithm 1 and (34), the corresponding contents that need to be cached would be pushed to the FAP and HN from the content server. As for the content delivery phase, the node association policy would be optimized based on (39) after the RNs send the content requests. For simplicity, we assume that the numbers of HNs and FAPs are the same M, the number of RNs is K.

Under our fixed node association policy, the energy efficient content caching Algorithm 1 has a complexity order of $O(MF^2)$. Furthermore, if we denote $n_{\max}$ the maximum number of iterations in Algorithm 2, then Algorithm 2 has a complexity order of $O(n_{\max}(MF^2+M^2K))$. Hence, Algorithm 2 exhibits a polynomial time-complexity.

VII. Performance Evaluation

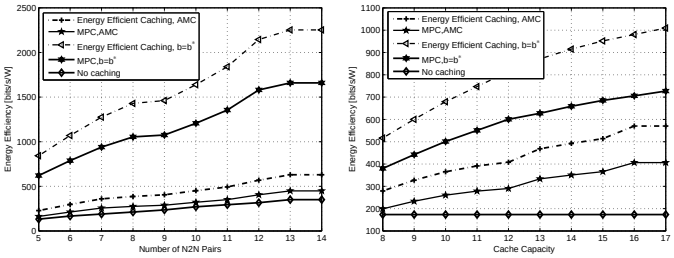
This section characterizes the performance of content caching and delivery both at the HNs and at the FAPs.

A. System Parameters

We assume that the HN cache capacity and FAP cache capacity are $C_{\mathrm{u}}=8$ files and $Q_j=16$ files, respectively, $\forall j\in\mathcal{M}$, and the content library has a size of $F=25$ files. We assume furthermore that there are 5 FAPs in the heterogeneous fog networks considered. Based on the request probabilities of $\mathbf{p}=[p_1,\cdots,p_F]$, $\rho=50\%$ of the FNs will request a random file from $\mathcal{F}$ as RNs, and the request probability obeys the Zipf distribution with a parameter of $\beta=1.3$ in (1). The potential HNs come from the remaining 50% of FNs, which would serve the RNs with data requests. The N2N search distance is $R_{\mathrm{d}}=70$m. The other parameters are listed in Table I.

TABLE II
Parameters

Minimum modulation size $b_{\min}$	4
Maximum modulation size $b_{\max}$	8
Circuit power P_0	60mW
Power spectral density of noise	-90dBm/Hz
Packet size L	320bits
α	1.3



(a) (b)

Fig. 2. Energy efficiency, for $Q_j=16$ and $\beta=1.3$.

B. Performance of HN caching

Fig. 2(a) and Fig. 2(b) show the EE for different number of N2N pairs and cache sizes, respectively, where we compare our energy-efficient caching strategy to the conventional technique of caching the most popular content (MPC). Explicitly, in the MPC caching, each node simply stores the most popular contents according to its individual request probability. We can see that as expected, our energy efficient caching strategy can always offer better EE than that of the MPC strategy for different number of N2N pairs and different cache sizes, respectively. Furthermore, we can observe that the optimal modulation mode selection offers better EE than that of the

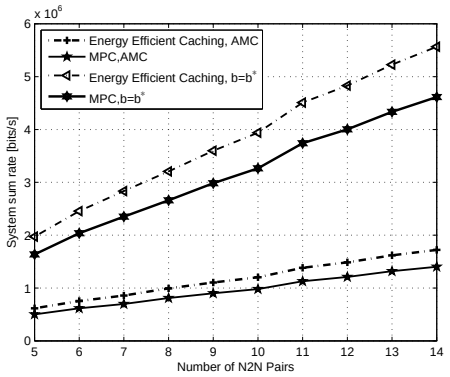


Fig. 3. System throughput with number of N2N pairs, for $Q_j = 16$ and $\beta = 1.3$.

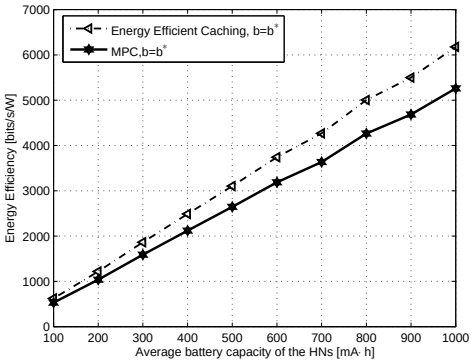
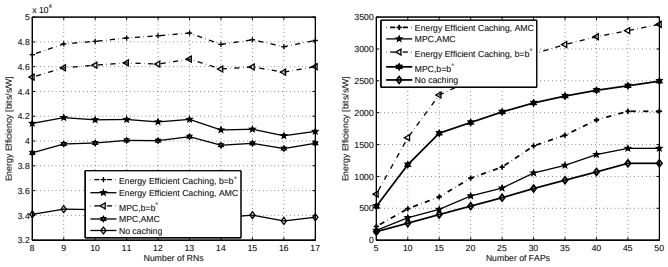


Fig. 4. EE versus average battery capacity of the HNs, for $N_m = 12$ and $\beta = 1.3$.

conventional adaptive modulation of coding (AMC) strategy, which confirms the analytical results. Another interesting remark is that upon increasing the number of N2N pairs, the EE is increased. It is obvious that if the number of N2N pairs is higher, then the EE of the proposed strategy will be higher, implying that it is more beneficial for increasing the system throughput. These results further indicate that the EE can be improved for N2N transmissions by placing different files associated with different probabilities in the HNs' cache. It can also be observed from the figure that the EE is nonincreasing, when the number of N2N pairs increases from 13 to 14, which



(a) (b)

Fig. 5. Energy efficiency, for $C_u = 8$ and $Q_j = 16$.

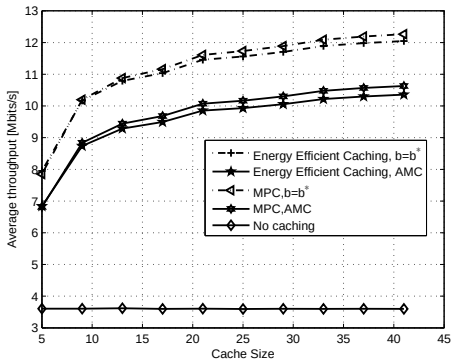


Fig. 6. Average throughput versus the cache size, for $M = 5$ and $N_m = 12$.

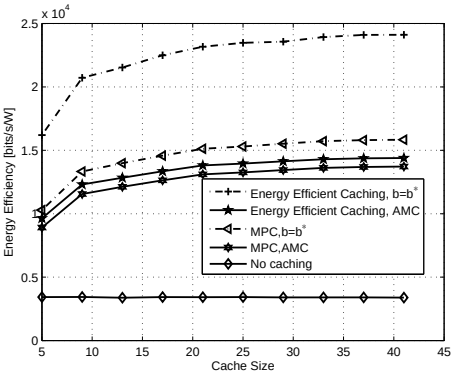


Fig. 7. EE versus the cache size, for $M = 5$ and $N_m = 12$.

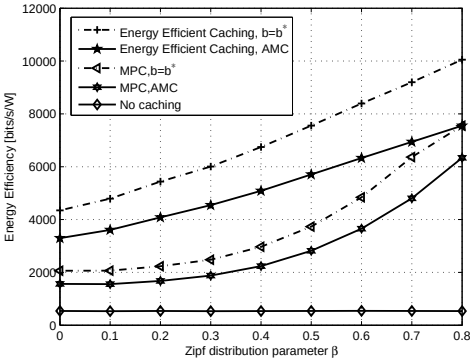


Fig. 8. EE versus different values of the Zipf distribution parameter β, for $C_u = 8$ and $Q_j = 16$.

is due to the increased circuit power consumption. Based on these simulation results, it is intuitive that there is a tradeoff between the EE and the number of active N2N pairs. In Fig. 2(b), we observe that the EE is increased with increasing the cache capacity. It is obvious that if the cache size is larger, then the EE of the proposed strategy will be higher, implying that it is more beneficial for caching more files.

For the comparison of different caching strategies, we also plot the system throughput when applying the energy-efficient caching strategy $\mathbf{q}^*$. Fig. 3 shows the system throughput versus

the number of N2N pairs N_d. We observe that the system sum rate increases upon increasing N_d, due to having more N2N transmissions. We can see that the throughput of the energy-efficient caching strategy can be higher than that of the MPC strategy associated with the same modulation strategy. These results indicate that the system throughput can be guaranteed for N2N transmissions with energy-efficient caching strategy. As expected, the optimal modulation selection strategy offers indeed a higher EE than that of AMC. In summary, the energy-efficient caching strategy can make the N2N transmissions energy-efficient without degrading its throughput.

In Fig. 4, we illustrate the EE of the proposed caching strategy versus the battery capacity. It can be observed that the energy-efficient caching strategy has the best performance regardless of the average battery capacity, which confirms the analytical results. Furthermore, the EE increases upon increasing the average battery capacity of the HNs. This is due to the fact that in case of a high battery capacity, each HN tends to cache the most popular files with an increased probability in order to increase the EE. In case of cache-aided N2N communications, our results indicate that it is necessary to take into account the battery capacity, when aiming for energy-efficient content placement.

C. Performance of the FAP caching

As shown analytically in Section V, the caching gain of the EE policy over the MPC policy decreases both with the node intensity λ_u and the node activity probability ρ, while it is not affected by the library size. In this section, we demonstrate the impact of λ_u, ρ, C_u, Q_j and β on the EE by means of simulations.

In Fig. 5(a) and Fig. 5(b), we illustrate the EE versus the number of RNs and FAPs for different caching strategies. Compared to traditional systems operating without cache, the no-caching strategy serves as a benchmark. We can see that as expected, our energy-efficient caching strategy can always offer better EE than that of the MPC strategy for different number of RNs and FAPs, respectively. We can observe from this figure that the energy-efficient caching strategy relying on the optimal modulation selection scheme always achieves significant performance improvements over the traditional caching strategy. On the other hand, the optimal modulation selection strategy performs much better than AMC using the same caching strategy. Note that even if the energy-efficient caching strategy is used, AMC is outperformed by the modulation strategy associated with MPC. This is because the optimal caching strategy increases both the system's throughput as well as its power consumption, it reduces the average power consumption per bit. Additionally, we observe from Fig. 5(b) that upon increasing the number of FAPs, the EE is increased. It is obvious that if the number of FAPs is higher, then the EE of the proposed strategy will be higher, implying that it is more beneficial for increasing the system throughput. These results further indicate that the EE can be improved by placing different files associated with different probabilities in the FAPs' cache.

In Fig. 6, we illustrate the relationship between the average throughput and the cache size associated with different caching strategies. As expected, we can observe that the energy-efficient caching strategy using the optimal modulation selection scheme is capable of guaranteeing the throughput. At the same time, the average throughput is increased upon increasing the cache size of the FAPs, which implies that the delivery success probability is increased. Furthermore, the MPC strategy performs better than the energy-efficient caching strategy using the same modulation mode selection strategy. This is due to the fact that the MPC strategy achieves the maximal delivery success probability for the model of (1). In Fig. 7, we illustrate the relationship between the EE and the cache size associated with different caching strategies. Since the optimal modulation selection scheme realizes the most energy-efficient transmission, the energy-efficient caching s-trategy using the optimal modulation selection scheme always performs best in EE, which confirms the analytical results. Furthermore, the MPC strategy with optimal modulation mode selection strategy performs better than the energy-efficient caching strategy with AMC strategy. This is due to the fact that the optimal modulation mode selection achieves the maximal EE in content delivery. Another interesting remark is that upon increasing the cache size, the EE converges to a stable value. This is due to the fact that the size of the content library is limited..

Next, in Fig. 8, we plot the EE versus the file popularity parameter β. We can observe from this figure that the EE is dramatically increased when β is increased. Therefore, as β increases, the figure suggests to allocate more power for each transmission stream in order to increase the average network throughput. These results further indicate that the EE can be improved for caching more popular files. We can also observe that for different modulation selection strategies, there is a crossing point at $\beta = 0.8$ between the performances of the energy-efficient caching strategy and the MPC strategy. When β is small, the energy-efficient caching strategy performs better than the MPC strategy. By contrast, when β is large, the performance of the MPC strategy converges to that of the energy-efficient caching strategy. We believe that this is due to the fact that the popularity parameter has a significant influence on the EE at a large β. Hence, the optimal solution is the MPC strategy.

VIII. Conclusions

We investigated the energy-efficient caching and node association strategy of heterogeneous fog networks. Two different scenarios have been considered, where the popular files are cached at the HNs and the FAPs, respectively. We formulated the optimization problem as an EE problem operating under the relevant caching and association constraints, and derived the joint modulation mode allocation and caching schemes for maximizing the EE of the heterogeneous fog networks considered. Finally, we conceive the joint node association and caching design. It has been shown that the proposed strategy achieves a better EE than the traditional caching strategy, when the modulation modes and caching are jointly optimized. Moreover, the results show that the proposed strategy can also guarantee a high network throughput. In particular for HNs-based caching, the simulation results indicate that both the

number of active HNs and the average battery capacity have positive impact on the system's performance, when using our energy-efficient caching strategy. For future work, we will consider the joint content caching and task offloading in the proposed framework.

APPENDIX A
PROOF OF LEMMA 1

Conditioned on deterministic RB allocation r_k^* for each HN k, we only have to solve the following problem instead of solving the problem in (16),

$$\max_{c_{f,k},b_{k,r}} \sum_{k\in\mathcal{N}} \frac{I_{k,r_k^*}(\varphi(\mathbf{c}))^{1+\frac{L}{b_{k,r_k^*}}}\theta_{k,r_k^*}}{\left[I_{k,r_k^*}\varphi(\mathbf{c})\left(g_{k,r_k^*}+P_0+P_{\text{ca}}C_{\text{u}}N_f\right)\right]^{\alpha}}. \quad (42)$$

Specifically, for every HN, only one RB can be used, and this RB cannot be used by other HNs. Therefore, if $k \neq k_{(1)}$, then $r(k) \neq r(k_{(1)})$. As a result, for any $k \in \mathcal{N}$ and $r \in \mathcal{N}_{\text{m}}$, the optimization problem of (42) is equivalent to independently optimizing the following N_{d} problems,

$$\max_{c_{f,k},b_{k,r}} \frac{I_{k,r}(\varphi(\mathbf{c}))^{1+\frac{L}{b_{k,r}}}\theta_{k,r}}{\left[I_{k,r}\varphi(\mathbf{c})\left(g_{k,r}+P_0+P_{\text{ca}}C_{\text{u}}N_f\right)\right]^{\alpha}}. \quad (43)$$

APPENDIX B
PROOF OF LEMMA 2

By taking the second-order derivative of $\varphi(\mathbf{c})$ with respect to $c_{f,k}$, $\forall f \in \mathcal{F}$, we have

$$\begin{aligned}\frac{\partial^2\varphi(\mathbf{c})}{\partial c_{f,k}^2} = \sum_{f\in\mathcal{F}} \Big(&-p_{i,f}e^{-\pi(1-\rho)\lambda_{\text{u}}c_{f,k}R_{\text{d}}^2}\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 \\ &-p_{i,f}\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 e^{-\pi(1-\rho)\lambda_{\text{u}}c_{f,k}R_{\text{d}}^2} \\ &-p_{i,f}(1-z_{f,k})\pi^2(1-\rho)^2\lambda_{\text{u}}^4R_{\text{d}}^4 e^{-\pi(1-\rho)\lambda_{\text{u}}c_{f,k}R_{\text{d}}^2}\Big). \quad (44)\end{aligned}$$

Then the second-order derivative is strictly negative. We can obtain the result that $\varphi(\mathbf{c})$ is positive and also a concave function of $c_{f,k}$, $\forall f \in \mathcal{F}$.

Since α is a constant around 1.3, $\frac{L}{b_{k,r}} - \alpha > 1$, we have proved that $\varphi(\mathbf{c})$ is positive and also a concave function of $c_{f,k}$, $\forall f \in \mathcal{F}$, then $(\varphi(\mathbf{c}))^{-1}$ is positive and a convex function. As a result, $\left[(\varphi(\mathbf{c}))^{-1}\right]^{\frac{L}{b_{k,r}}-\alpha}$ is a convex function of $c_{f,k}$.

APPENDIX C
PROOF OF THEOREM 1

Since the modulation alphabet size $b_{k,r}$ is an integer, we can find the globally optimal solution by first finding the optimal probability distribution $c_{f,k}$, $\forall f \in \mathcal{F}$, for any given $b_{k,r}$ and then enumerating $b_{k,r}$ until the value of U_{ee} achieves its maximum under the constraints considered.

Based on Lemma 2, $[\varphi(\mathbf{c})]^{\alpha-\frac{L}{b_{k,r}}}$ is a convex function with respect to $c_{f,k}$, $\forall f \in \mathcal{F}$. Then, the numerator of (43) is a concave function, and the denominator of (43) is a convex function. In all, by exploiting an iterative algorithm known as Dinkelbach's method [48], the optimization problem (43) can be solved by Algorithm 1.

APPENDIX D
PROOF OF LEMMA 3

The second-order derivative of the objective function is strictly negative, thus $S(\mathbf{z})$ is a concave function of $z_{f,j}$, $\forall f \in \mathcal{F}$. By constructing the Lagrangian function of (Q1), we have

$$\begin{aligned}L(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\eta}) = S(\mathbf{z}) &+ \sum_{f\in\mathcal{F}}\mu_f z_{f,j} + \sum_{f\in\mathcal{F}}\lambda_f(1-z_{f,j}) \\ &+ \sum_{f\in\mathcal{F}}\eta_f(Q_j - \sum_{f\in\mathcal{F}} z_{f,j}), \quad (45)\end{aligned}$$

where μ_f, λ_f and η_f are the non-negative Lagrange multipliers associated with the constraints. Then, we have

$$\begin{aligned}\frac{\partial L(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial z_{f,j}} = \sum_{i\in\mathcal{N}_r}\Big(&x'_{i,j}p_{f,i}e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}R_{\text{d}}^2}b_{i,j} - \\ &x'_{i,j}p_{f,i}z_{f,j}\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}R_{\text{d}}^2}b_{i,j}\Big) + \mu_f - \lambda_f - \eta_f. \quad (46)\end{aligned}$$

The Karush-Kuhn-Tucker (KKT) conditions can be written as

$$\frac{\partial L(\mathbf{z}^*,\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial z_{f,j}^*} = 0, \quad \forall f \in \mathcal{F}, \quad (47)$$

$$\mu_f z_{f,j}^* = 0, \quad \lambda_f(1-z_{f,j}^*) = 0, \quad \forall f \in \mathcal{F}, \quad (48)$$

$$\eta_f(Q_j - \sum_{f\in\mathcal{F}} z_{f,j}^*) = 0, \quad \forall f \in \mathcal{F}, \quad (49)$$

$$\sum_{f\in\mathcal{F}} z_{f,j}^* = Q_j, \quad 0 \le z_{f,j}^* \le 1, \quad \forall f \in \mathcal{F}. \quad (50)$$

Based on (46) and (47), we arrive at

$$\begin{aligned}\eta_f = V_i\Big(&e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_d^2} - \\ &z_{f,j}^*\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2}\Big) + \mu_f - \lambda_f, \forall f \in \mathcal{F}, \quad (51)\end{aligned}$$

where $V_i = \sum_{i\in\mathcal{N}_r} x'_{x,j}p_{f,i}b_{i,j}$. As a result, the optimal solution can be summarized as follows:

- Under the condition that

$$\begin{aligned}\eta_f \ge V_i\Big(&e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2} - \\ &z_{f,j}^*\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2}\Big), \quad (52)\end{aligned}$$

  we have $z_{f,j}^* = 0$, $\lambda_f = 0$ and $\mu_f \ge 0$ based on (48).
- Under the condition that

$$\begin{aligned}\eta_f \le V_i\Big(&e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2} - \\ &z_{f,j}^*\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2}\Big), \quad (53)\end{aligned}$$

  we have $z_{f,j}^* = 1$, $\mu_f = 0$ and $\lambda_f \ge 0$ based on (48).
- If $0 < z_{f,j}^* < 1$, we have $\mu_f = \lambda_f = 0$ based on (48). Thus, we have

$$\begin{aligned}\eta_f = V_i\Big(&e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2} - \\ &z_{f,j}^*\pi(1-\rho)\lambda_{\text{u}}R_{\text{d}}^2 e^{-\pi(1-\rho)\lambda_{\text{u}}z_{f,j}^*R_{\text{d}}^2}\Big). \quad (54)\end{aligned}$$

Let $\psi = -\pi(1-\rho)\lambda_{\mathrm{u}} z^*_{f,j} R^2_{\mathrm{d}}$, we have $V_i(e^{\psi}+\psi e^{\psi}) = \eta_f$. The solution of $\psi e^{\psi} = a$ is $\psi = \mathcal{W}(a)$ by the definition of the Lambert $\mathcal{W}$ function [46] and then the optimal solution can be derived as $z^*_{f,j} = \frac{\mathcal{W}(\frac{\eta_f e}{\sum_{i\in\mathcal{N}_r} x^f_{i,j} p_{f,i} b^*_{i,j}})-1}{-\pi(1-\rho)\lambda_{\mathrm{u}} R^2_{\mathrm{d}}}$, $\forall f \in \mathcal{F}$.

REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and trends, 2017-2022 white paper," *white paper*, Feb. 2019.

[2] K. Wang and W. Chen, "Energy-efficient communications in MIMO systems based on adaptive packets and congestion control with delay constraints," *IEEE Trans. Wireless Commun.*, vol. 14, no.4, pp. 2169–2179, Apr. 2015.

[3] 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.

[4] K. Wang, J. Li, Y. Yang, W. Chen, and L. Hanzo, "Energy-efficient multi-tier caching and node association in heterogeneous fog networks," in *Proc. of the IEEE VTC2020-Fall*, (Victoria, BC Canada), pp. 1–5, Oct. 2020.

[5] 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.

[6] 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.

[7] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no.6, pp. 2438–2452, Jun. 2016.

[8] 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.

[9] T. Liu, J. Li, F. Shu, M. Tao, W. Chen, and Z. Han, "Design of contract-based trading mechanism for a small-cell caching system," *IEEE Trans. Wireless Commun.*, vol. 16, no.10, pp. 6602–6617, Oct. 2017.

[10] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no.6, pp. 854–864, Dec. 2016.

[11] Y. Sun, M. Peng, S. Mao, and S. Yan, "Hierarchical radio resource allocation for network slicing in fog radio access networks," *IEEE Trans. Veh. Technol*, vol. 68, no.4, pp. 3866–3881, Apr. 2019.

[12] K. Xiong, S. Leng, J. Hu, X. Chen, and K. Yang, "Smart network slicing for vehicular fog-RANs," *IEEE Trans. Veh. Technol*, vol. 68, no.44, pp. 3075–3085, Apr. 2019.

[13] Y. Zhang, C. Wang, and H. Wei, "Parking reservation auction for parked vehicle assistance in vehicular fog computing," *IEEE Trans. Veh. Technol*, vol. 68, no.4, pp. 3126–3139, Apr. 2019.

[14] J. Du, L. Zhao, X. Chu, F. R. Yu, J. Feng, and C.-L. I, "Enabling low-latency applications in LTE-A based mixed fog/cloud computing systems," *IEEE Trans. Veh. Technol*, vol. 68, no.2, pp. 1757–1771, Feb. 2019.

[15] Y. Yang, K. Wang, G. Zhang, X. Chen, X. Luo, and M.-T. Zhou, "MEETS: Maximal energy efficient task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, vol. 5, pp. 4076–4087, Oct. 2018.

[16] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no.2, pp. 74–80, Feb. 2014.

[17] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no.10, pp. 6650–6678, Oct. 2017.

[18] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no.7, pp. 4286–4298, July 2014.

[19] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no.12, pp. 6833–6859, Dec. 2015.

[20] A. Altieri, P. Piantanida, L. R. Vega, and C. G. Galarza, "On fundamental trade-offs of device-to-device communications in large wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no.9, pp. 4958–4971, Sept. 2015.

[21] J. Jiang, S. Zhang, and B. Li, "Maximized cellular traffic offloading via device-to-device content sharing," *IEEE J. Sel. Areas Commun.*, vol. 34, no.1, pp. 82–91, Jan. 2016.

[22] 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.

[23] S. E. Hajri and M. Assaad, "Energy efficiency in cache enabled small cell networks with adaptive user clustering," *IEEE Trans. Wireless Commun.*, vol. 17, no.2, pp. 955–968, Feb. 2018.

[24] M. Gregori, J. Gomez-Vilardebo, J. Matamoros, and D. Gunduz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no.5, pp. 1222–1234, May 2016.

[25] 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.

[26] F. Gabry, V. Bioglio, and I. Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no.12, pp. 3288–3298, May 2016.

[27] S. Zhao, Y. Yang, Z. Shao, X. Yang, H. Qian, and C.-X. Wang, "FEMOS: Fog-enabled multitier operations scheduling in dynamic wireless networks," *IEEE Internet Things J.*, vol. 5, no.2, pp. 1169–1183, Apr. 2018.

[28] Y. Yang, S. Zhao, W. Zhang, Y. Chen, X. Luo, and J. Wang, "DEBTS: Delay energy balanced task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, pp. 1–1, 2018.

[29] K. Wang, M. Tao, W. Chen, and Q. Guan, "Delay-aware energy-efficient communications over Nakagami-m fading channel with MMPP traffic," *IEEE Trans. Commun.*, vol. 63, no.8, pp. 3008–3020, Aug. 2015.

[30] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, pp. 72–80, Aug. 2017.

[31] D. W. K. Ng, Y. Wu, and R. Schober, "Power efficient resource allocation for full-duplex radio distributed antenna networks," *IEEE Trans. Wireless Commun.*, vol. 15, no.4, pp. 2896–2911, Apr. 2016.

[32] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. of the IEEE ICC*, (London, U.K.), pp. 3358–3363, Jun. 2015.

[33] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no.1, pp. 176–189, Jan. 2016.

[34] 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.

[35] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networks," *IEEE Trans. Commun.*, vol. 64, no.10, pp. 4365–4380, Oct. 2016.

[36] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no.3, pp. 584–587, Mar. 2017.

[37] R. Rao, S. Vrudhula, and D. N. Rakhmatov, "Battery modeling for energy aware system design," *Computer*, vol. 36, no.12, pp. 77–87, Dec. 2003.

[38] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE Jour. Select. Aeras in Commun*, vol. 22, no.6, pp. 1089–1098, Aug. 2004.

[39] X. Xiao, X. Tao, and J. Lu, "QoS-aware energy-efficient radio resource scheduling in multi-user OFDMA system," *IEEE Commun. Lett.*, vol. 17, no.1, pp. 75–78, Jan. 2013.

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

[41] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc. of the IEEE ICC*, (Kuala Lumpur, Malaysia), pp. 1–6, May 2016.

[42] A. J. Goldsmith, *Wireless Communications*. New York, NY: Cambridge University Press, 2005.

[43] M. B. Purslet and J. M. Shea, "Adaptive nonuniform phase-shift-key modulation for multimedia traffic in wireless networks," *IEEE J. Select. Aeras Commun*, vol. 18, no.8, pp. 1394–1407, Aug. 2000.

[44] R. J. Lavery, "Throughput optimization for wireless data transmission," Master's thesis, Polytechnic University, June 2001.

[45] S. Catreux, P. F. Driessen, and L. J. Greenstein, "Data throughputs using multiple-input multiple-output (MIMO) techniques in a noise-limited cellular environment," *IEEE Trans. Wireless Commun*, vol. 1, no.2, pp. 226–235, Apr. 2002.

[46] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.
[47] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[48] A. Roubi, "Method of centers for generalized fractional programming," *Jounal of Optimization Theory and Applications*, vol. 107, pp. 123–143, Oct. 2000.

Kunlun Wang (M'19) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016. From 2016 to 2017, he was with Huawei Technologies Company, Ltd., where he was involved in energy efficiency algorithm design. From 2017 to 2019, he was with the Key Lab of Wireless Sensor Network and Communication, SIMIT, Chinese Academy of Sciences, Shanghai, China. Since March 2019, he has been a Research Assistant Professor with the School of Information Science and Technology, ShanghaiTech University. He is also with the Shanghai Institute of Fog Computing Technology (SHIFT). His current research interests include energy efficient communications, fog computing networks, resource allocation, and optimization algorithm.

Jun Li (M'09-SM'16) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From January 2009 to June 2009, he worked in the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a Research Scientist. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2012 to June 2015, he is a Research Fellow at the School of Electrical Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a visiting professor at Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning, and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and industrial Internet of things. He has co-authored more than 200 papers in IEEE journals and conferences, and holds 1 US patents and more than 10 Chinese patents in these areas. He was serving as an editor of IEEE Communication Letters and TPC member for several flagship IEEE conferences. He received Exemplary Reviewer of IEEE Transactions on Communications in 2018, and best paper award from IEEE International Conference on 5G for Future Wireless Networks in 2017.

Yang Yang (S'99-M'02-SM'10-F'18) received the B.S. and M.S. degrees in Radio Engineering from Southeast University, Nanjing, China, in 1996 and 1999, respectively, and the PhD degree in Information Engineering from the Chinese University of Hong Kong in 2002. He is currently a full professor at ShanghaiTech University, China, serving as the Executive Dean of School of Creativity and Art and the Director of Shanghai Institute of Fog Computing Technology (SHIFT). He is also an Adjunct Professor with the Research Center for Network Communication, Peng Cheng Laboratory, China. Before joining ShanghaiTech University, he has held faculty positions at the Chinese University of Hong Kong, Brunel University, U.K., University College London (UCL), U.K., and SIMIT, CAS, China. Yangs current research interests include fog computing networks, service-oriented collaborative intelligence, wireless sensor networks, IoT applications, and advanced testbeds and experiments. He has published more than 200 papers and filed more than 80 technical patents in these research areas. He has been the Chair of the Steering Committee of Asia-Pacific Conference on Communications (APCC) since January 2019. In addition, he is a General Co-Chair of the IEEE DSP 2018 conference and a TPC Vice-Chair of the IEEE ICC 2019 conference.

Wen Chen (M'03-SM'11) is a fellow of Chinese Institue of Electronics and a distinguished lecturer of IEEE Communications Society. He received BS and MS from Wuhan University, China in 1990 and 1993 respectively, and PhD from University of Electro Communications, Tokyo, Japan in 1999. He was a JSPS fellow in the University of Electro Communications from 1999 through 2001. In 2001, he joined University of Alberta, Canada, starting as a PIMS post-doctoral fellow and then as a research associate. Since 2006, he has been a full professor in the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is also the director of Institute for Signal Processing and Systems. Dr. Chen is the editors of IEEE Transactions on Wireless Communications, IEEE Transctions on Communications, IEEE Access, and IEEE Open Journal of Vehicular Technology, He is the chair of IEEE VTS Shanghai Chapter. He has published 98 papers in IEEE journals and more than 120 papers in IEEE conferences. His interests cover multiple access, coded cooperation, green heterogeneous networks, and vehicular communications.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.orgwikiLajos_Hanzo) (FIEEE'04, Fellow of the Royal Academy of Engineering F(REng), of the IET and of EURASIP), received his Master degree and Doctorate in 1976 and 1983, respectively from the Technical University (TU) of Budapest. He was also awarded the Doctor of Sciences (DSc) degree by the University of Southampton (2004) and Honorary Doctorates by the TU of Budapest (2009) and by the University of Edinburgh (2015). He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He has published 1900+ contributions at IEEE Xplore, 19 Wiley-IEEE Press books and has helped the fast-track career of 123 PhD students. Over 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry.