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 node 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.

1. 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) communicates 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 scenarios [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 heterogeneous 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) $\Phi_u$ [32], where the intensity of $\Phi_u$ is $\lambda_u$. 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. We consider a finite set of content files $\mathcal{F} = \{1, \ldots, F\}$ consisting of $F$ files each with $N_f$ bits that all FNs in the network

1In 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

<table>
<thead>
<tr>
<th>Definition</th>
<th>Notation</th>
<th>Definition</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location set of FAPs</td>
<td>$\mathcal{M}$</td>
<td>Cache memory size of FN</td>
<td>$C_u$</td>
</tr>
<tr>
<td>Cache memory size of FAP</td>
<td>$Q_j$</td>
<td>Location set of HNs</td>
<td>$\mathcal{N}$</td>
</tr>
<tr>
<td>Location set of RNs</td>
<td>$\mathcal{N}_r$</td>
<td>Set of content files</td>
<td>$\mathcal{F}$</td>
</tr>
<tr>
<td>Total number of files</td>
<td>$F$</td>
<td>File size</td>
<td>$N_f$</td>
</tr>
<tr>
<td>Popularity exponent of the Zipf distribution</td>
<td>$\beta$</td>
<td>Set of resource blocks</td>
<td>$\mathcal{N}_m$</td>
</tr>
<tr>
<td>Bandwidth of a resource block</td>
<td>$W$</td>
<td>Noise power</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Circuit power consumption</td>
<td>$P_0$</td>
<td>Battery lifetime of FN $k$</td>
<td>$T_k$</td>
</tr>
<tr>
<td>The probability that the FAP caches the $f$th file</td>
<td>$z_{f,i}$</td>
<td>Transmit power of FAP $j$ destined to the RN $i$</td>
<td>$P_{t,j}$</td>
</tr>
<tr>
<td>modulation scheme</td>
<td>$h_{t,j}$</td>
<td>binary association flag between RN $i$ and HN $k$</td>
<td>$y_{t,k}$</td>
</tr>
<tr>
<td>The probability that the HN $k$ caches the $f$th file</td>
<td>$c_{f,k}$</td>
<td>The request probability of the $f$th file from the HN $k$</td>
<td>$P_{r,k}$</td>
</tr>
</tbody>
</table>
may request, which are indexed in descending order according to their popularity, i.e., the \( f \)th file is the \( f \)th 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 \( f \)th file is requested is

\[
p_f = \frac{1}{f^\beta \sum_{n \in F} n^{-\beta}} ,
\]

where \( \beta \) 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 \( f \)th file has a certain probability of \( c_{f,k} \) and \( z_{f,j} \) to be independently cached by each HN \( k \), \( k \in \mathcal{N}_c \), 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 \( \sum_{f \in F} c_{f,k} \leq C_u, \forall k \in \mathcal{N}_c \) and \( \sum_{f \in F} z_{f,j} \leq Q_j \).

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 \( \mathcal{N}_m = \{1, 2, \ldots, N_m\} \), where \( N_m \) is the total number of RBs from the RNs. 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 \( \sigma^2 \) 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 \( P_{ca} C_u N_f \), 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

\[
T = \frac{B_c}{D_c^\alpha} ,
\]

where \( B_c \) and \( D_c \) are the battery capacity and the discharge current, respectively, while the exponent \( \alpha \) 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

\[
T_k = \frac{B_c V_0^{\alpha}}{(P_{k,r} + P_0 + P_{ca} C_u N_f)^\alpha} ,
\]

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

\[
U_{i,k} = \frac{R_{k,r} B_c V_0^\alpha}{(P_{k,r} + P_0 + P_{ca} C_u N_f)^\alpha}, \quad \forall i \in \mathcal{N}_r .
\]

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 \( N_f \) RNs requesting files selected by the FAP \( j \) to be served simultaneously, and each FAP has \( L \) antennas, with \( L \geq N_f \). Similarly, the FAP independently selects files from the set \( F \) for caching according to a specific probability distribution. To unify our analysis, we denote the probability that the FAP caches the \( f \)th file by \( 0 \leq z_{f,j} \leq 1 \). When \( z_{f,j} = [z_{f,j}]_{f \in F} \) 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 preceding vector and the transmit power of FAP \( j \) destined to the RN \( i \) by \( w_{i,j} \in \mathbb{C}^{L \times 1} \) and \( P_{i,j} \), respectively. Then the signal-to-interference-plus-noise ratio of
RN $i$ associated with FAP $j$ can be formulated as
\[
\gamma_{i,j} = \frac{|h_{i,j}^Hw_{i,j}|^2}{\sum_{j\neq i}|h_{i,j}^Hw_{i,j}|^2 + \sigma^2}.
\]
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, the precoding vector of RN $i$ is given by $w_{i,j} = \sqrt{P_i}a_{i,j}$, where $a_{i,j}$ is the ZF beamforming vector of RN $i$ associated with FAP $j$, which is the $i$th column of the matrix $H_j^H(H_j^H)\^{-1}$, with $H_j = [h_{i,j}, \ldots, h_{N_j,j}]^T$, where $h_{i,j}$ 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
\[
\gamma_{i,j} = \frac{P_j\delta_{i,j}}{\sigma^2} \quad \forall \delta_{i,j} \in \delta,
\]
where $\delta = [\delta_{i,j}, \delta_{j,2}, \ldots, \delta_{N_j,j}]$ is the eigenvalue vector of $H_j^H$. Based on the Chernoff upper bound [38], we can approximate the symbol error rate (SER) $\epsilon$ as:
\[
\epsilon = 2(1 - 2^{-b_{i,j}/2})e^{-\frac{2\gamma_{i,j}}{b_{i,j}}}. \tag{7}
\]
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 $P_{cQ_j}N_j$, respectively. Similarly, the expected transmission time $T_j$ of FAP $j$ can be expressed as
\[
T_j = \frac{B_{i,j}V_{0,j}^2}{(P_{i,j} + P_0 + P_{cQ_j}N_j)^2}, \tag{9}
\]
where $B_{i,j}$ 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
\[
U_{i,j} = \frac{R_{i,j}T_j}{\sum_{j\in M} x_{i,j} + \sum_{k\in N} y_{i,k}} \leq 1 \quad \forall i \in N_i, \tag{10}
\]
where $R_{i,j}$ 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 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
\[
\sum_{j\in M} x_{i,j} + \sum_{k\in N} y_{i,k} \leq 1 \quad \forall i \in N_i. \tag{12}
\]
Similarly, for N2N-aided communications, only a single RN can be linked to one RN, which can be formulated as
\[
\sum_{i\in N_i} y_{i,k} \leq 1 \quad \forall k \in N. \tag{13}
\]
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
\[
U_{ee} = \sum_{i\in N_i} \sum_{k\in N} y_{i,k}c_{f,k}U_{i,k} + \sum_{i\in N_i} \sum_{j\in M} x_{i,j}z_{f,j}U_{i,j}. \tag{14}
\]
Hence, the EE optimization problem of fog networks is expressed as:
\[
\max \quad U_{ee}, \quad x_{i,j}, y_{i,k}, z_{f,j}, c_{f,k}, b_{k,r}, b_{i,j} \quad \text{s.t.} \quad \sum_{j\in M} x_{i,j} + \sum_{k\in N} y_{i,k} \leq 1, \forall i \in N_i, \tag{15}
\]
where $b_{k,r}$ and $b_{i,j}$ represent the modulation scheme selected for the HN $k$ at the $r$th 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 $\{y_{i,k}\}$, $\forall i \in N_i$, $\forall k \in N$, the caching problem is coupled among the FAPs and HNs. In particular, the content placement problem formulated for HNs
can be written as
\[
\max_{z_{f,k},b_{k,r}} U_{ee}^{D} = \sum_{i \in N} \sum_{k \in N} y_{i,k}^t c_{f,k} U_{i,k},
\]
\[\text{s.t.} \quad b_{k,r} \in [b_{\min}, b_{\max}], \forall k \in N, \forall r \in N_m,
\]
\[c_{f,k} \in [0, 1], \forall f \in F, \forall k \in N,
\]
\[\sum_{f \in F} c_{f,k} \leq C_w, \forall k \in N, \tag{16}\]
where \(U_{ee}^{D}\) and \(c_{f,k}\) represent the EE of the HNs-based caching network and the caching result for the \(f\)th file at the HN \(k\), respectively.

Similarly, the content placement problem of FAPs can be written as
\[
\max_{y_{i,k}} U_{ee}^{B} = \sum_{i \in N} \sum_{j \in M} y_{i,j}^t z_{f,j} U_{i,j},
\]
\[\text{s.t.} \quad b_{i,j} \in [b_{\min}, b_{\max}], j \in M,
\]
\[z_{f,j} \in [0, 1], \forall f \in F, j \in M,
\]
\[\sum_{f \in F} z_{f,j} \leq Q_j, \forall j \in M, \tag{17}\]
where \(U_{ee}^{B}\), \(b_{i,j}\), \(z_{f,j}\) and \(M\) represent the EE of the FAP-based caching network, the modulation scheme selection by the FAP \(j\), the caching result of the \(f\)th 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
\[
\max_{x_{i,j},y_{i,k}} \sum_{i \in N} \sum_{k \in N} y_{i,k}^t c_{f,k} U_{i,k} + \sum_{i \in N} \sum_{j \in M} x_{i,j}^t z_{f,j} U_{i,j},
\]
\[\text{s.t.} \quad \sum_{j \in M} x_{i,j} + \sum_{k \in N} y_{i,k} \leq 1, \forall i \in N_i,
\]
\[\sum_{i \in N_i} y_{i,k} \leq 1, \forall k \in N,
\]
\[\sum_{i \in N_i} x_{i,j} = N_j, \forall j \in M,
\]
\[x_{i,j}, y_{i,k} \in \{0, 1\}, \forall i \in N_i. \tag{18}\]
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 \(f\)th requested file by \(f_j\), the caching probability of the \(f\)th file from the HN \(k\) by \(c_{f,k}\) and the request probability of the \(f\)th 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]
\[
\zeta_{k,f} = 1 - e^{-\pi(1-\rho)\lambda c_{f,k} R_d^2}, \tag{19}\]
where \(\rho\) 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 \(\zeta_k = \sum_{f \in F} p_{f,k}(1 - c_{f,k})\zeta_{k,f}\), which is given by
\[
\zeta_k = \sum_{f \in F} p_{f,k}(1 - c_{f,k}) \left(1 - e^{-\pi(1-\rho)\lambda c_{f,k} R_d^2}\right). \tag{20}\]

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 \(f\)th 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 \(f\)th file can be quantified as \(P_{k,f,\zeta_{k,f}}\), where \(P_{k,f}\) denotes the transmit power of the HN \(k\) on the \(r\)th 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
\[
P_k = \sum_{f \in F} P_{k,f,\zeta_{k,f}}(1 - c_{f,k})\zeta_{k,f}, \tag{21}\]
\[= P_{k,\zeta_k}.\]

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 \(\epsilon\) of M-ary quadrature amplitude modulation (MQAM) having an alphabet size \(2^{b_{k,r}}\), is given by
\[
\epsilon = 2(1 - 2^{-b_{k,r}/2})Q\left(\frac{3}{2^{b_{k,r}} - 1} \gamma_{k,r}\right), \tag{22}\]
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 \(r\)th 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 \( r \)th 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 \( r \)th RB as
\[
\gamma_{k,r} = \frac{P_{l,r} \phi(d_{k,k})}{P_{l,r} \phi(d_{k,k}) + \sigma^2},
\]
where \( \phi(d_{k,k}) = K_{dB} - 10 \gamma \log_{10} \frac{d_{k,k}}{d_0} \), and the parameters \( K_{dB}, \gamma \) 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)
\[
\epsilon \leq 2 \left( 1 - 2^{-b_k/r/2} \right) e^{-\frac{3}{2} \gamma_{k,r}}. \tag{24}
\]
Therefore, upon substituting (23) into (24), the closed-form expression of the transmission power of the HN \( k \) is given by
\[
P_{k,r} \approx \frac{I_{k,r} \phi(d_{k,k})}{\phi(d_{k,k})} \frac{2(2^{b_k/r} - 1)}{3} \ln \left( \frac{1}{2 - 2^{b_k/r}} \right),
\]
where \( I_{k,r} \in \{0,1\} \) is the indicator of the \( r \)th RB allocated to the HN \( k \). Thus, the average transmit power of the HN \( k \) can be approximated as
\[
P_k = \frac{I_k \phi(d_{k,k})}{\phi(d_{k,k})} \frac{2(2^{b_k/r} - 1)}{3} \ln \left( \frac{1}{2 - 2^{b_k/r}} \right).
\]
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
\[
t_{L} = \frac{L}{W b_{k,r} c_k}.
\]
In this continuation, the cached packet contains an overhead of \( L_p \) bits. Since not all the transmitted data is information bits, we can get the successful transmission probability for a cached packet as \( S_p = (\zeta_k(1 - \epsilon))^{L_p} \). 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]:
\[
R_{k,r} = \frac{I_{k,r} L_p S_p}{t_L} = \frac{I_{k,r} L_p}{W b_{k,r} c_k} \zeta_k^{(1 - \frac{1}{r} \epsilon)} (1 - \epsilon) \frac{L}{b_k/r}. \tag{28}
\]

C. Optimizing Content Caching and Delivery

Substituting (26) and (28) into (4) and denoting \( \zeta_k \) as \( \varphi(c) \), we arrive at the EE of the caching networks relying on the HNs as
\[
U_{ee}^D = \sum_{k \in N} \sum_{r \in R} \frac{I_{k,r} \varphi(c)}{[I_{k,r} \varphi(c) (g_k + P + P_{ca} C_a N_f)]^\alpha},
\]
where \( \theta_{k,r} = \frac{I_{k,r} (P_{l,r} \phi(d_{k,k}) + \sigma^2)}{\phi(d_{k,k})} \frac{2(2^{b_k/r} - 1)}{3} \ln \left( \frac{1}{2 - 2^{b_k/r}} \right) \) and \( g_k = g_k = \frac{P_{l,r} \phi(d_{k,k}) + \sigma^2}{\phi(d_{k,k})} \frac{2(2^{b_k/r} - 1)}{3} \ln \left( \frac{1}{2 - 2^{b_k/r}} \right) \). We assume that the resource allocation of the HNs has already been accomplished by the central controller, and that the \( r \)th RB allocated to the HN \( k \) results in \( I_{k,r} = 0 \) for \( r \neq r \), 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 \( C_a \) of each HN, the number of cached files should satisfy \( \sum_{f \in F} z_{f,k} \leq C_a \). We assume using adaptive modulation and coding (AMC).

**Lemma 1:** The optimization Problem (16) is equivalent to optimizing the problem
\[
\max_{c_{f,k}, b_{k,r}} \frac{I_{k,r} \varphi(c)}{[I_{k,r} \varphi(c) (g_k + P + P_{ca} C_a N_f)]^\alpha}. \tag{30}
\]

**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 \( \varphi(c) \) via the following lemma.

**Lemma 2:** \( \varphi(c) \) is a convex function of \( c_{f,k}, \forall f \in F \).

**Proof:** The proof is given in Appendix B.

By analyzing Problem (30), we have

**Theorem 1:** For the HN \( k \), the optimal caching probability \( c_{f,k}^* \) achieves the optimal EE for the RN-HN transmissions.

**Proof:** The proof is given in Appendix C.

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

**Algorithm 1:** EE Optimization for HNs-Based Caching.

1: For \( b_{k,r} = b_{\min} : b_{\max} \\
Step 1: Initialization: Let \( c_1 \) be a feasible point and
\( \delta_1 = \frac{I_{k,r} \varphi(c_1)}{[I_{k,r} \varphi(c_1) (g_k + P + P_{ca} C_a N_f)]^\alpha} \). Let \( v = 1 \).
Step 2: Invoking convex programming to solve the following problem:
\[
\eta(\delta_v) = \max_{c_{f,k}} \{ I_{k,r} \varphi(c)^{1+\frac{1}{\alpha}} \theta_{k,r} \}
- \delta_v [I_{k,r} \varphi(c) (g_k + P + P_{ca} C_a N_f)]^\alpha. \tag{31}
\]
With the aid of convex optimization, we find the solution point \( z_{v+1} \).
Step 3: If the solution \( \eta(\delta_v) = 0 \), stop and \( q_v \) is optimal. Otherwise, set
\( \delta_{v+1} = \frac{I_{k,r} \varphi(c_{v+1})}{[I_{k,r} \varphi(c_{v+1}) (g_k + P + P_{ca} C_a N_f)]^\alpha} \), 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: \( b_{k,r} = b_{k,r}^* \).
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

$$R_{i,j} = \sum_{f \in F} p_{f,i} z_{f,j} e^{-\pi (1-\rho) \mu_{z_f,j} R_{i,j}^2} W_{b_{i,j}} \quad \forall i \in \mathcal{N}_r, \forall j \in \mathcal{M},$$

(32)

where $p_{f,i}$ and $R_d$ are the request probability of the $j$th 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

$$\bar{U}_{ee}^B = \sum_{j \in \mathcal{M}} \sum_{i \in \mathcal{N}_r} \sum_{f \in F} \frac{x_{i,j} p_{f,i} z_{f,j} e^{-\pi (1-\rho) \mu_{z_f,j} R_{i,j}^2} W_{b_{i,j}} B_{c,j} V_{\theta_{ij}}^x}{(P_{i,j} + P_0 + P_{ca} Q_{j} N_f)^\alpha}. \quad (33)$$

Our aim is to jointly design the transmission policy $b = \{b_{1,j}, \ldots, b_{N_r,j}\}$ at the FAP, and the local caching policy $z_j = \{z_{1,j}, \ldots, z_{F,j}\}$ at the FAP, for maximizing the EE of fog networks. By exploiting that the content caching constraints and modulation mode 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

$$\max_z S(z) = \sum_{j \in \mathcal{M}} \sum_{i \in \mathcal{N}_r} \sum_{f \in F} x_{i,j} p_{f,i} z_{f,j} e^{-\pi (1-\rho) \mu_{z_f,j} R_{i,j}^2 b_{i,j}},$$

s.t. $0 \leq z_{f,j} \leq 1, f = 1, \ldots, F,$

$$\sum_{f \in F} z_{f,j} \leq Q_j, \forall j \in \mathcal{M},$$

(34)

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 $z$ of Problem (Q1) can be obtained as follows:

$$z_{f,j}^* = \min \left\{ \left[ \frac{\mathcal{W}(\frac{\eta_f e}{\sum_{i \in \mathcal{N}_r} x_{i,j} p_{f,i} b_{i,j}}) - 1}{-\pi (1-\rho) \lambda_R R_D^2} \right] + 1 \right\},$$

(34)

for $\forall f \in F$, where $[z]^+ = \max\{0, z\}$ 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 $E_b$ achieves the maximum. Then, we can derive the optimal $z_{f,j}^*$, $\forall f \in F$, with the optimal $b_{i,j}$, $\forall i \in \mathcal{N}_r$. Then we achieve the maximum of $S(z)$ 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

$$\mathcal{W}(\frac{\eta_f e}{\sum_{i \in \mathcal{N}_r} x_{i,j} p_{f,i} b_{i,j}}) \geq 1,$$

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

$$\mathcal{W}(\frac{\eta_f e}{\sum_{i \in \mathcal{N}_r} x_{i,j} p_{f,i} b_{i,j}}) = 1 - \pi (1-\rho) \lambda_R R_D^2,$$

the $f$th file, $\forall f \in F$, is completely cached at the FAP; and, finally, if

$$1 - \pi (1-\rho) \lambda_R R_D^2 < \mathcal{W}(\frac{\eta_f e}{\sum_{i \in \mathcal{N}_r} x_{i,j} p_{f,i} b_{i,j}}) < 1,$$

the FAP $j$ caches the $f$th file, $\forall f \in F$, 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]. Let us introduce dual variables $\psi = [\psi_1, \ldots, \psi_M]$ for the third constraint of Problem (18), and $\psi = [\psi_1, \ldots, \psi_M]$ for the first constraint of Problem (18). The Lagrangian function is given by

$$L (\{x_{i,j}\}, \{y_{i,k}\}, \psi, \nu) = \sum_{i,j} x_{i,j} \nu_{j} - \sum_{j \in M} \psi_{j} \left( \sum_{i \in N} x_{i,j} - N_{j} \right)$$

(38)

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

$$f(\psi, \nu) = \max_{x_{i,j}, y_{i,k}} L (\{x_{i,j}\}, \{y_{i,k}\}, \psi, \nu) ,$$

s.t. \( \sum_{j \in M} x_{i,j} + \sum_{k \in N} y_{i,k} \leq 1, \forall i \in N \). (39)

Then, we can obtain explicit analytical solutions, that are

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

(40)

and

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

(41)

where

$$j^{*} = \arg \max_{j \in M} (U_{i,j} - \psi_{j})$$

and

$$k^{*} = \arg \max_{k \in N} (U_{i,j} - \nu_{k}) .$$

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.

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. 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 $n_{\text{max}}$ the maximum number of iterations in Algorithm 2, then Algorithm 2 has a complexity order of $O(n_{\text{max}}(MF^2 + M^2K))$. 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 $C_H = 8$ files and $Q_J = 16$ files, respectively, \( \forall j \in M \), and the content library has a size of $F = 25$ files. We assume further that there are 5 FAPs in the heterogeneous fog networks considered. Based on the request probabilities of $p = [p_1, \ldots, p_F]$, $\rho = 50\%$ of the RNs will request a random file from $F$ 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 RNs, which would serve
the RNs with data requests. The N2N search distance is $R_d = 70$ m. The other parameters are listed in Table I.

### B. Performance of HN Caching

Figs. 2(a) and 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 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 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 $\lambda_u$ and the node activity probability $\rho$, while it is not affected by the library size. In this section, we demonstrate the impact of $\lambda_u$, $\rho$, $C_u$, $Q_j$ and $\beta$ on the EE by means of simulations.
In Figs. 5(a) and 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 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 using the same modulation strategy. This is due to the fact that the optimal modulation mode selection scheme 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 strategy using the optimal modulation selection scheme always performs best in EE, which confirms the analytical results. Furthermore, the MPC strategy with optimal modulation selection strategy 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 $\beta$. We can observe from this figure that the EE is dramatically increased when $\beta$ is increased. Therefore, as $\beta$ 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 $\beta$ is small, the energy-efficient caching strategy performs better than the MPC strategy. By contrast, when $\beta$ 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 $\beta$. Hence, the optimal solution is the MPC strategy.
VIII. CONCLUSION

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),

$$\max_{\ell, k, b, r} \sum_{k \in K} \frac{I_{k, r^*_k}(\varphi(c))}{I_{k, r^*_k}(\varphi(c))} \left[ \frac{1}{\bar{r}^*_k} \theta_k r_k \right]^\alpha. \quad (42)$$

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 $k \in K$ and $r \in \tilde{N}_h$, the optimization problem of (42) is equivalent to independently optimizing the following $N_h$ problems,

$$\max_{\ell, k, b, r} \frac{I_{k, r}(\varphi(c))}{I_{k, r}(\varphi(c))} \left[ \frac{1}{\bar{r}} \theta_k r_k \right]^\alpha. \quad (43)$$

APPENDIX B
PROOF OF LEMMA 2

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

$$\frac{\partial^2 \varphi(c)}{\partial c_{f,k}^2} = \sum_{f \in F} \left( -p_{i,k} e^{-\pi(1-\rho)\alpha c_{f,k} R^2_i} (1-\rho) \lambda_u R^2_\theta - p_{i,k} (1-\rho) \lambda_u R^2_\theta e^{-\pi(1-\rho)\alpha c_{f,k} R^2_i} - p_{i,k} (1-\rho) \lambda_u R^2_\theta e^{-\pi(1-\rho)\alpha c_{f,k} R^2_i} \right). \quad (44)$$

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

Since $\alpha$ is a constant around 1.3, $\frac{1}{b_{k,r}} - \alpha > 1$, we have proved that $\varphi(c)$ is positive and also a concave function of $c_{f,k}$.

\[ \forall f \in \mathcal{F}, \text{then } (\varphi(c))^{-1} \text{ is positive and a convex function. As a result, } \left[(\varphi(c))^{-1}\right]^{-1} \text{ is a convex function of } c_{f,k}. \]

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, $[\varphi(c)]^{-1} \bar{r}^*_k$ 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 $S(z)$ is a concave function of $z_{f,j}, \forall f \in \mathcal{F}$. By constructing the Lagrangian function of (Q1), we have

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

where $\mu_f, \lambda_f$ and $\eta_f$ are the non-negative Lagrange multipliers associated with the constraints. Then, we have

$$\frac{\partial L(z, \mu, \lambda, \eta)}{\partial z_{f,j}} = \sum_{i \in N} \left( x'_{i,j} p_{f,j} e^{-\pi(1-\rho)\alpha z_{f,j} R^2_i b_{i,j}} - x'_{i,j} p_{f,j} e^{-\pi(1-\rho)\alpha z_{f,j} R^2_i b_{i,j}} \right) \right) \mu_f - \lambda_f - \eta_f. \quad (46)$$

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

$$\frac{\partial L(z, \mu, \lambda, \eta)}{\partial z_{f,j}} = 0, \quad \forall f \in \mathcal{F}, \quad \mu_f z_{f,j} = 0, \quad \lambda_f (1 - z_{f,j}^*) = 0, \quad \forall f \in \mathcal{F}, \quad \eta_f (Q_j - z_{f,j}^*) = 0, \quad \forall f \in \mathcal{F}, \quad \sum_{f \in \mathcal{F}} z_{f,j}^* = Q_j, \quad 0 \leq z_{f,j}^* \leq 1, \quad \forall f \in \mathcal{F}. \quad (50)$$

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

$$\eta_f = V_i \left( e^{-\pi(1-\rho)\alpha z_{f,j} R^2_i} \right) \right) \mu_f - \lambda_f, \forall f \in \mathcal{F}, \quad (51)$$

where $V_i = \sum_{j \in N} x'_{i,j} p_{f,j} b_{i,j}$. As a result, the optimal solution can be summarized as follows:
• Under the condition that
\[
\eta_f \geq V_i \left( e^{-\lambda_i} \mu_i x_{f,j} R_i^2 \right)\]
we have \(z_{f,j}^* = 0, \lambda_f = 0\) and \(\mu_f \geq 0\) based on (48).

• Under the condition that
\[
\eta_f \leq V_i \left( e^{-\lambda_i} \mu_i x_{f,j} R_i^2 \right)
\]
we have \(z_{f,j}^* = 1, \lambda_f = 0\) and \(\mu_f \geq 0\) based on (48).

If \(0 < z_{f,j}^* < 1\), we have \(\mu_f = \lambda_f = 0\) based on (48). Thus, we have
\[
\eta_f = \frac{V_i \left( e^{-\lambda_i} \mu_i x_{f,j} R_i^2 \right)}{\left( \sum_{k=1}^{n} \psi_i x_{k,j} R_k^2 \right)^{-1}}
\]

Let \(\psi = \frac{\psi_i}{V_i} = \frac{1}{\eta_f}\). The solution of \(\psi_i = a\) is \(\psi = \psi_i = \psi_i(a)\) by the definition of the Lambert \(\psi_i\) function [46] and then the optimal solution for \(\forall f \in F\).


Kunlun Wang (Member, IEEE) 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 (Senior Member, IEEE) received the 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 2014, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2014 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 U.S. 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 (Fellow, IEEE) received the B.S. and M.S. degrees in radio engineering from Nantong University, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. 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 (Senior Member, IEEE) received the B.S. and M.S. degrees from Wuhan University, China in 1990 and 1993, respectively, and the Ph.D. degree from the 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 Postdoctoral Fellow and then as a Research Associate. Since 2006, he has been a Full Professor at the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is also the Director of Institute for Signal Processing and Systems. He is the editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS 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. He is a Fellow of Chinese Institute of Electronics and a Distinguished Lecturer of IEEE Communications Society.

Lajos Hanzo (Fellow, IEEE) received the Master degree and Doctorate from the Technical University (TU) of Budapest, in 1976 and 1983, respectively. He was also awarded the Doctor of Sciences (D.Sc.) degree by the University of Southampton, in 2004, Honorary Doctorates by the TU of Budapest, in 2009, and by the University of Edinburgh, in 2015. He is Fellow of the Royal Academy of Engineering (FR(Eng)) of the IET and of EURASIP. 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.