Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8392511, 10 pages
https://doi.org/10.1155/2022/8392511

Research Article

Energy-Efficient Hierarchical Collaborative Scheme for Content Delivery in Mobile Edge Computing

Chao Fang,[1,2] **Xiaojie Huang**,[3] **Jingjing Huang**,[4] **Zhaoming Hu**,[1] **Yanhua Sun**,[1] **Jun Cai**,[5] **Zhuwei Wang**,[1] **Huamin Chen**,[1] **Jianchuan Zhang**,[2] **and Fangmin Xu**[3]

[1]*Faculty of Information Technology, Beijing University of Technology, Beijing, China*
[2]*Purple Mountain Laboratory: Networking, Communications and Security, Nanjing, China*
[3]*Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China*
[4]*Beijing Saixi Technology Development Company with Limited Liability, Beijing, China*
[5]*China Mobile Group Hunan Company Limited, Changsha, China*

Correspondence should be addressed to Jingjing Huang; hjj187264726@126.com

Received 12 October 2021; Accepted 1 March 2022; Published 1 April 2022

Academic Editor: Antonio Guerrieri

With the rapid growth of Internet traffic and smart mobile terminals, ultradense networks are adopted as the key technology of the fifth generation to enhance resource utilization and content distribution while causing serious energy efficiency problem. Mobile edge computing has recently drawn great attention for its advantages in reducing transmission delay and network energy consumption by implementing caching and computing abilities at the edge of mobile networks. To improve network energy efficiency and content transmission, in this paper, we propose a novel energy-efficient hierarchical collaborative scheme by considering the in-network caching, request aggregation, and joint allocation of caching, computing, and communication resources in a layered heterogeneous network including mobile users, small base stations, macro base stations, and the cloud. We formulate the energy consumption problem as a queuing theory-based centralized model, where the same content requests can be aggregated in the queue of each base station. Then, the optimal solution is analyzed based on the distribution characteristic of content popularity at the base stations. Simulation results show that the performance of our proposed model is much better than the existing cloud-edge cooperation solutions without considering the deployment of caching resource and request aggregation policies.

1. Introduction

With the rapid growth of Internet traffic represented by multimedia (e.g., Youtube and Netflix) and smart mobile terminals (e.g., smartphones and tablets) [1], as the key technology of the fifth generation (5G) mobile communication system, ultradense networks can efficiently improve the resource utilization and efficiency of content distribution [2]. However, the consequent energy consumption problem has been increasingly prominent caused by the density deployment of base stations (BSs) and strict quality of service (QoS) of mobile traffic. Therefore, it is urgent and challenging to enhance network energy efficiency while ensuring the efficiency of content transmission [3, 4].

Although improving content delivery by shortening the distance between content requesters and providers, cloud computing brings extra deployment and operation costs and huge energy consumption [5, 6]. As a lightweight extension of cloud computing, mobile edge computing (MEC) can further reduce transmission delay and network energy consumption by implementing caching and computing abilities at the edge of mobile networks, e.g., base stations (BSs), access routers, and switches [7]. Considering the obvious advantages and limited service capacity of MEC compared with cloud computing, hierarchical cloud-edge cooperation schemes have recently drawn great attention of academic and industry to improve energy efficiency of mobile networks and content delivery [8].

Cloud-edge collaboration solutions have initially been explored in a two-layer network of edge devices and the cloud, ignoring the computing and caching capabilities of the BSs [9–11]. Zhang et al. [12] investigate energy-saving wireless backhaul bandwidth allocation and power allocation in heterogeneous small cellular networks and propose an optimal energy efficiency model, which can be solved iteratively. Then, distributed frameworks and centralized processing capabilities are developed to achieve green networks, in particular, by reducing energy consumption through BS sleep policies. Qi and Wang [9] discuss interference-aware user association problem under cell sleeping for heterogeneous cloud cellular networks. Han et al. [13] propose an energy sharing-based energy and user joint allocation method between macro base stations (MBSs) and small base stations (SBSs) in a heterogeneous network. Wang et al. [14] consider improving network performance in mobile networks by joint management and allocation of caching, computing, and communication (3C) resources. Li et al. [15] propose an energy-efficient resource allocation scheme by orchestrating the delay-sensitive tasks in a resource-constrained cloud-edge-end collaboration system. However, most current works mainly focus on the cooperative optimization of two kinds of resources above [7, 16]. Kai et al. [17] design a cooperative computation offloading policy under the constrained serving and transmitting capabilities, where the tasks are partially conducted at the mobile devices, edge servers, and the cloud. Chen et al. [18] develop a reinforcement learning-based resource allocation framework, which can leverage energy consumption and network latency while satisfying power and computation constraints. Xu et al. [19] propose an IC-IoT network architecture based on software-defined networking paradigm and use deep Q-network (DQN) model to optimize the allocation of computing and cache resources to propagate IC-IoT processes. Zhang et al. [20] formulate the joint offloading and resource allocation problem as a Markov decision process (MDP) to maximize the number of unloading tasks while reducing energy consumption.

Although cloud-edge cooperation can improve energy efficiency and content distribution, most works are studied in a three-tier topology of mobile users (MUs), BSs, and the cloud, without considering the influence of in-network caching, request aggregation [21–24] and network heterogeneity, and mainly focused on the joint allocation optimization of two kinds of resources. In this paper, we propose an energy-efficient hierarchical collaborative solution to improve content delivery, where in-network caching, request aggregation, and the cooperative allocation of 3C resources are analyzed in a layered heterogeneous network including MUs, SBSs, MBSs, and the cloud. The main contributions of the paper are as follows.

(i) We formulate the energy efficiency problem as a queuing theory-based centralized model in a cloud-edge cooperation network, where in-network caching is considered and the same content requests can be aggregated in the queue of each base station. Energy consumption is minimized by jointly optimizing the allocation of 3C resources while ensuring QoS of content delivery

(ii) We analyze the minimal energy consumption problem based on the distribution characteristic of content popularity and present the key factors that affect network performance

(iii) We evaluate the proposed energy efficiency model in heterogeneous network environments. Simulation results show that the performance of our proposed solution is much better than the existing cloud-edge cooperation schemes not considering the deployment of caching resource and request aggregation policies

The rest of this paper is organized as follows. In Section 2, the energy models of different network components are presented. In Section 3, the minimal energy consumption problem is formulated and analyzed. Simulation results are presented and discussed in Section 4. Finally, we conclude this study in Section 5.

2. System Model

In this section, we formulate the cloud-edge cooperation problem to minimize energy consumption and improve content delivery by jointly allocating 3C resources in a mobile network. Figure 1 describes a cloud-edge cooperation scenario, where both MBSs and SBSs have caching and computing capabilities. Content requests from a MU can be satisfied by its connected BS and the cloud in sequence. As shown in Figure 1, the total energy consumption consists of MUs, BSs, the cloud, and network links, where energy models of different BSs and their accessed MUs are discussed separately in the system.

2.1. Energy Model of Mobile Users. We assume that there are N_m MBSs in the network, and the ith MBS is directly accessed by $N_{i,s}$ SBSs and $N_{i,m}$ MUs. $N_{ij,m}$ is the number of MUs connected to the jth SBS of the ith MBS. The number of different network contents available in our system is F. Thus, the energy consumption of MUs, E_{MU}, can be written as

$$E_{MU} = \sum_{i=1}^{N_m}\sum_{j=1}^{N_{i,s}}\sum_{m=1}^{N_{ij,m}}\sum_{k=1}^{F} E_{ij,m}^{k} + \sum_{i=1}^{N_m}\sum_{j=1}^{N_{i,m}}\sum_{k=1}^{F} E_{i,m}^{k}. \tag{1}$$

2.1.1. Energy of Mobile Users in SBSs. The energy about a content request k consumed by the mth user of the ijth SBS, $E_{ij,m}^{k}$, is the product of its power and the delay between sending this request and receiving the corresponding data, which can be written as

$$E_{ij,m}^{k} = P_{ij,m}^{k} T_{ij,m}^{k}, \tag{2}$$

$$T_{ij,m}^{k} = T_{ij}^{k} X_{ij}^{k} + \left(1 - X_{ij}^{k}\right)\left[T_{i}^{k} X_{i}^{k} + T_{c}^{k}\left(1 - X_{i}^{k}\right)\right], \tag{3}$$

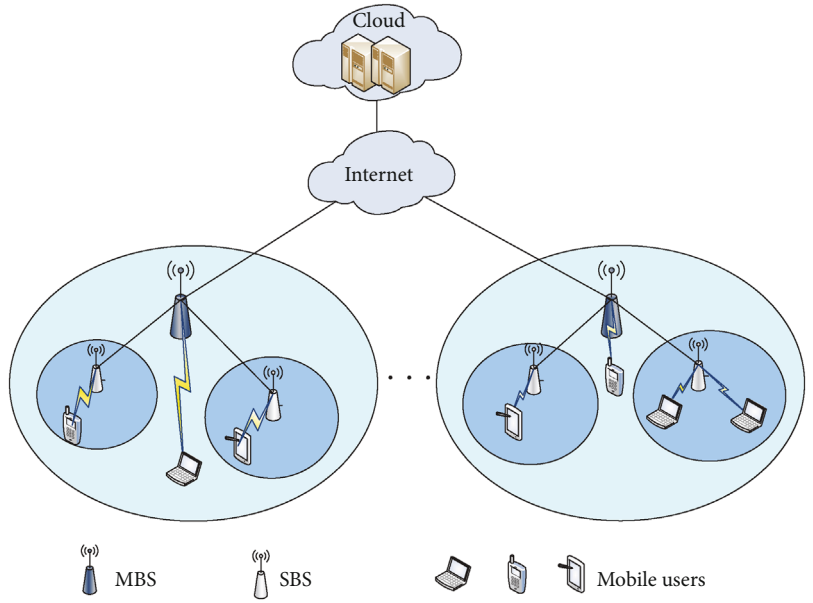

Figure 1: A cloud-edge cooperation scenario.

where $P_{ij,m}^k$ is the power about the content request k consumed by the mth user of the ijth SBS, X_{ij}^k and X_i^k are boolean variables and indicate whether the request about content k is satisfied at the ijth SBS and ith MBS, respectively. X_{ij}^k is equal to be 1 if the ijth SBS buffers content k, and 0 otherwise. X_i^k is equal to be 1 if the ith MBS caches content k, and 0 otherwise. T_{ij}^k, T_i^k, and T_c^k are the delay for the content request k to be met at the ijth SBS, ith MBS, and cloud, respectively.

To solve (3), a new $M/M/k_s$ queuing model with different service rates and arrival rates is designed, where in-network caching and request aggregation are considered [3, 25]. In our proposed queuing model, the same requests arriving at a network node can be aggregated to a request to be forwarded to fetch the corresponding content. If this node stores the data in its cache, the content will be distributed to the end-users along the routing path in the opposite direction. Therefore, both in-network caching and request aggregation can reduce the response time caused by content requests and further improve energy efficiency of our system. We assume that λ_{ij}, μ_{ij}, and k_{ij} are request arrival rate and service rate of the ijth SBS after the introduction of in-network caching and request aggregation policies [3], the number of servers at the ijth SBS, respectively. Therefore, the utilization of the ijth SBS can be written as

$$\rho_{ij} = \frac{\lambda_{ij}}{k_{ij}\mu_{ij}}. \tag{4}$$

The probability that a request from a MU of the ijth SBS will have to queue at this SBS when its servers are occupied can be written as

$$P_{Q_{ij}} = \pi_{0_{ij}} \frac{k_{ij}^{k_{ij}}}{k_{ij}!} \frac{\rho_{ij}^{k_{ij}}}{1-\rho_{ij}}, \tag{5}$$

where $\pi_{0_{ij}}$ is the steady-state probability that zero request tasks exist in the ijth SBS.

Therefore, on basis of (4) and (5), the response time that a content request k is satisfied at the ijth SBS can be written as

$$E\left[T_{ij}^k\right] = \frac{1}{\mu_{ij}} + \frac{\rho_{ij}}{\lambda_{ij}\left(1-\rho_{ij}\right)} P_{Q_{ij}}. \tag{6}$$

When the content k is stored at the ijth SBS, the total delay that a MU obtains this content consists of uplink and downlink transmission delay between the ijth SBS and this MU, execution time, and response delay in the ijth SBS, which can be derived as

$$T_{ij}^k = \frac{S_c^k}{v_{ij}} + \frac{\alpha_c s^k}{l_{ij,\text{up}}} + E\left[T_{ij}^k\right] + \frac{(1-\alpha_c)s^k}{l_{ij,\text{down}}}, \tag{7}$$

where the $l_{ij,\text{up}}$ and $l_{ij,\text{down}}$ indicate uplink and downlink transmission rate between the ijth SBS and the accessed MU, v_{ij} denotes the CPU clock speed in the ijth SBS, S_c^k is

the number of CPU clocks to deal with the task about content request k, s^k is the size of content k, and α_c is the proportion of the requested data in the total traffic generated by the task about content request k.

Similarly, we assume that λ_i, μ_i, and k_i are request arrival rate and service rate of the ith MBS after the introduction of in-network caching and request aggregation policies, the number of servers at the ith MBS, respectively. λ_c, μ_c, and k_c are request arrival rate and service rate of the cloud considering in-network caching and request aggregation, the number of servers in the cloud. Thus, the total delay of a content request k from the SBS satisfied in the ith MBS and the cloud can be written as

$$T_i^k = \frac{S_c^k}{v_i} + T_{ij}^k + \frac{\alpha_c s^k}{l_{i,\text{up}}} + E\left[T_i^k\right] + \frac{(1-\alpha_c)s^k}{l_{i,\text{down}}}, \tag{8}$$

$$T_c^k = \frac{S_c^k}{v_c} + T_i^k + \frac{\alpha_c s^k}{l_{C,\text{up}}} + E\left[T_c^k\right] + \frac{(1-\alpha_c)s^k}{l_{C,\text{down}}}, \tag{9}$$

where the $l_{i,\text{up}}$ and $l_{i,\text{down}}$ are uplink and downlink transmission rate between the ijth SBS and ith MBS, $l_{C,\text{up}}$ and $l_{C,\text{down}}$ are uplink and downlink transmission rate between MBSs and the cloud, v_i and v_c are the CPU clock speed in the ith MBS and cloud, and $E[T_i^k]$ and $E[T_c^k]$ are the response time that a content request k is satisfied at the ith MBS and cloud, respectively.

2.1.2. Energy of Mobile Users in MBSs. Based on the above analysis, the energy about the content request k consumed by the mth user of the ith MBS, $E_{i,m}^k$, and its corresponding delay, $T_{i,m}^k$, can be written as

$$E_{i,m}^k = P_{i,m}^k T_{i,m}^k, \tag{10}$$

$$T_{i,m}^k = T_i^k X_i^k + \left(1 - X_i^k\right)\left\{T_{ij}^k\left[1 - \prod_{j=1}^{N_{i,s}}\left(1 - X_{ij}^k\right)\right] + T_c^k \prod_{j=1}^{N_{i,s}}\left(1 - X_{ij}^k\right)\right\}, \tag{11}$$

where $P_{i,m}^k$ is the power consumed by the mth MU of the ith MBS when obtaining content k.

2.2. Energy Model of Base Stations. In the mobile network, the energy consumption of BSs can be written as

$$E_{BS} = \sum_{i=1}^{N_m}\sum_{j=1}^{N_{i,s}} P_{ij} T_s + \sum_{i=1}^{N_m} P_i T_s, \tag{12}$$

where T_s represents running time of the system, and P_{ij} and P_i are the power consumption of the ijth SBS and ith MBS, respectively.

2.2.1. Energy of SBSs. The total power consumption of the ij th SBS can be written as

$$P_{ij} = P_{ij}^t + P_{ij}^c, \tag{13}$$

where P_{ij}^t and P_{ij}^c represent the traditional power and cache power at the ijth SBS.

The traditional power consumption $P_{i,j}^t$ of the ijth SBS can be written as

$$P_{ij}^t = P_{0ij} + \frac{\Delta_{P_{ij}}}{\gamma_{ij}}\left(2^{\mu_{ij}/B_{ij}} - 1\right)\sum_{k=1}^{F}\left(1 - X_{ij}^k\right) q_{ij}^k s^k, \tag{14}$$

where γ_{ij}, B_{ij}, and q_{ij}^k are the signal-to-noise ratio and the maximal transmission rate, the amount of network requests about data k at the ijth SBS, and P_{0ij} and $\Delta_{P_{ij}}$ are the static power consumption of the ijth SBS in the active mode and its corresponding slop parameter [26].

The cache power at the ijth SBS consists of two parts: cache retrieval power and content caching power, which can be written as

$$P_{ij}^c = \sum_{k=1}^{F}\left(q_{ij}^k P_{ij,r}^k + X_{ij}^k s^k w_{ca}\right), \tag{15}$$

where $P_{ij,r}^k$ presents the retrieval power consumption about content k in the buffer of the ijth SBS, and w_{ca} is the power efficiency parameter depending on storage hardware technologies [27].

2.2.2. Energy of MBSs. Similarly, the total power consumption of the ith MBS can be written as

$$P_i = P_i^t + P_i^c, \tag{16}$$

where P_i^t and P_i^c are the traditional power and cache power consumption of the ith MBS, respectively.

Therefore, the expressions of P_i^t and P_i^c can be written as

$$P_i^t = P_{0i} + \frac{\Delta_{P_i}}{\gamma_i}\left(2^{\mu_i/B_i} - 1\right)\sum_{k=1}^{F}\left(1 - X_i^k\right) q_i^k s^k, \tag{17}$$

$$P_i^c = \sum_{k=1}^{F}\left(q_i^k P_{i,r}^k + X_i^k s^k w_{ca}\right), \tag{18}$$

where γ_i, B_i, and q_i^k are the signal-to-noise ratio, the maximal transmission rate, and the amount of network requests about content k at the ith MBS, P_{0i} and Δ_{P_i} are the static power consumption of the ith MBS in the active mode and its corresponding slop parameter, and $P_{i,r}^k$ is the retrieval power consumption about content k in the buffer of the ith MBS.

2.3. Energy Model of Cloud. The model of cloud is consisted with static power P_s and the processing power of requests which can not satisfied by MBS.

$$P_c = P_s + \left\{ \sum_{i=1}^{N_m} \sum_{k=1}^{F} \sum_{j=1}^{N_{i,s}} q_{ij}^k \left(1 - X_i^k\right) \prod_{j}{}^{'=1} N_{i,s} \left(1 - X_{ij'}^k\right) + \sum_{i=1}^{N_m} \sum_{k=1}^{F} q_i^k \left(1 - X_i^k\right) \prod_{j=1}^{N_{i,s}} \left(1 - X_{ij}^k\right) \right\} P_{c,r}^k, \tag{19}$$

where $P_{c,r}^k$ is the retrieval power consumption about content request k in the cloud.

2.4. Energy Model of Network Wired Links. As shown in Figure 1, the total energy consumption of wired link can be written as

$$E_L = \sum_{i=1}^{N_m} \sum_{j=1}^{N_{i,s}} P_{L_{i,j}} T_s + \sum_{i=1}^{N_m} P_{L_i} T_s, \tag{20}$$

where $P_{L_{i,j}}$ is the power consumption about traffic transmitting between the ijth SBS to the ith MBS, P_{L_i} is the power consumption about traffic distributing between the ith MBS to the cloud [27, 28].

3. Problem Formulation and Analysis

In this section, we formulate the minimal energy consumption problem as a cooperative cloud-edge resource allocation model for content delivery services. Then, this model is analyzed theoretically to present how to obtain the optimal solution.

3.1. Problem Formulation. Based on the energy models presented in Section 2, the energy-efficient hierarchical collaborative problem can be formulated as Eq. (21), where C_i and C_{ij} are the maximal caching capacities of MBS i and its accessed SBS j, respectively. Therefore, the optimization objective of Eq. (21) is to minimize energy consumption of the system by designing optimal caching and routing strategies under the limited caching, computation, and communication capacities.

In the constraints, C_1-C_2 are the maximal cache size constraints of SBSs and MBSs, C_3-C_5 present the utilization constraints of SBSs, MBSs, and the cloud, the boolean variables related to caching decision can only be 0 or 1 in C_6, and C_7 limits the proportion of the requested data in the total traffic generated by a request task between 0 and 1.

$$\begin{aligned} \text{Min} \sum_{i=1}^{N_m} \sum_{j=1}^{N_{i,s}} \sum_{m=1}^{N_{ij,m}} \sum_{k=1}^{F} P_{ij,m}^k T_{ij,m}^k + \sum_{i=1}^{N_m} \sum_{m=1}^{N_{ij,m}} \sum_{k=1}^{F} P_{i,m}^k T_{i,m}^k + \left[\sum_{i=1}^{N_m} \sum_{j=1}^{N_{i,s}} \left(P_{ij} + P_{L_{i,j}}\right) + \sum_{i=1}^{N_m} \left(P_i + P_{L_i}\right) + P_c \right] T_s \text{s.t.} \quad C_1 \\ : \sum_{k=1}^{F} X_{ij}^k s^k \le C_{ij}, \forall i \in N_m, j \in N_{i,s} C_2 : \sum_{k=1}^{F} X_i^k s^k \le C_i, \forall i \in N_m C_3 : \rho_{ij} \le 1, \forall i \in N_m, j \in N_{i,s} C_4 : \rho_i \le 1, \forall i \in N_m C_5 : \rho_c \le 1 C_6 \\ : X_i^k, X_{ij}^k \in \{0, 1\}, \forall i \in N_m, j \in N_{i,s}, k \in F C_7 : \alpha_c \in [0, 1]. \end{aligned} \tag{21}$$

3.2. Model Analysis. In this part, we analyze the optimal solution of the proposed hierarchical energy consumption problem (21) based on the distribution characteristic of content popularity, which will provide a benchmark for online solutions to obtain near-optimal results in mobile heterogeneous networks.

We assume that network content popularity follows the Zipf's distribution and the BSs cache network contents according to the rank of content popularity [29]. Zipf's law is a statistical distribution in certain data sets and states that the relative request probability for the kth most popular content is calculated by $k^{-\alpha}/\sum_{f=1}^{F} f^{-\alpha}$, where F is the number of different network contents, and α is the skewness factor. A large value of α indicates that more requests are sent for the popular data. In our model, the vertical collaborative caching between a MBS and its followed SBSs is utilized to optimize the cache hit rate to improve energy efficiency, which means that the most popular contents are stored in the SBSs and the less ones are cooperatively cached in their upper MBS. Therefore, the utilization of ijth SBS and ith MBS with optimal caching can be written as

$$\rho_{ij}' = \frac{\lambda_{ij}}{k_{ij}\mu_{ij}} \frac{\sum_{k=C_{ij}+1}^{F} k^{-\alpha_{ij}}}{\sum_{k=1}^{F} k^{-\alpha_{ij}}}, \tag{22}$$

$$\rho_i' = \frac{\lambda_i}{k_i\mu_i} \frac{\sum_{k=C_{ij}+C_i+1}^{F} k^{-\alpha_i}}{\sum_{k=1}^{F} k^{-\alpha_i}}, \tag{23}$$

where α_{ij} and α_i are the skewness factors of content popularity at the ijth SBS and ith MBS. The number of popular contents arriving at the BSs increases with the growth of the skewness factor.

Based on the rewritten utilization expressions in (22) and (23), the minimal delay can be obtained under the optimal caching. Similarly, we can rewrite the formulas related to network power in Section 2 on basis of content popularity. Therefore, the minimal energy consumption can be achieved to improve content delivery in the system.

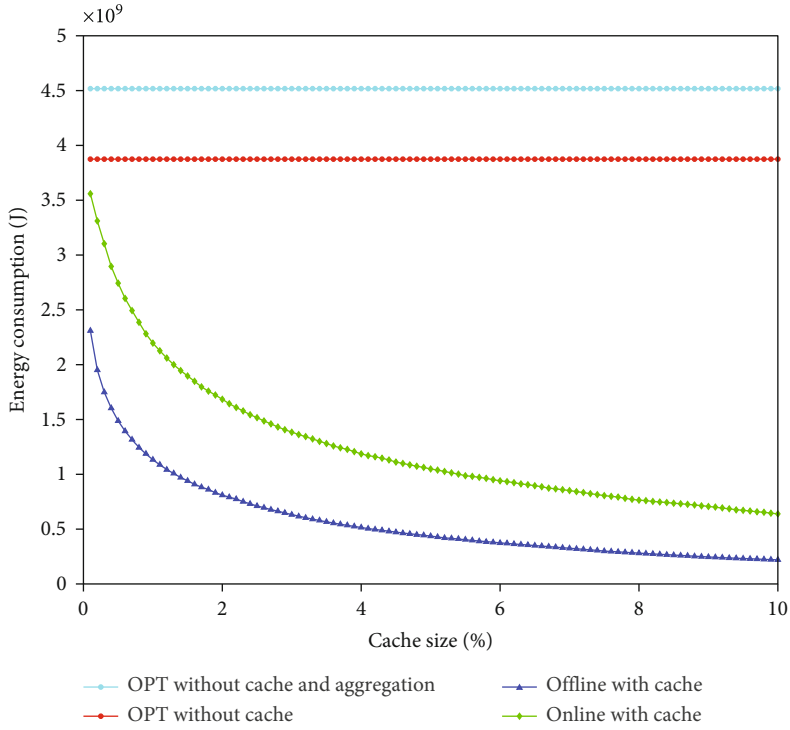

Figure 2: Network energy consumption versus cache size when $\alpha = 1.0$, $F = 1000$, and $\lambda = 10$.

4. Simulation and Results

In this section, we evaluate the performance of our proposed model in heterogeneous scenarios, e.g., cache size, content popularity, the number of different contents, and the arrival rate of network requests. In the simulation, cache size of a BS is abstracted as a proportion and is equal to the relative size to the amount of different contents, which is from 0.1% to 10% [30, 31]. The content popularity follows the Zipf distribution, where the range of the skewness factors α_i and α_{ij} varies from 0.6 to 1.5 [29, 32]. In addition, the request arrival rate of a MBS is twice that of its accessed SBSs. The default value of arrival rate about a SBS is set to be 10 in our simulation. The simulation is carried out to demonstrate the advantages of the proposed solutions "Offline with Cache" and "Online with Cache" compared with "OPT without Cache and Aggregation" and "OPT without Cache" schemes. "Offline with Cache" and "Online with Cache" are the corresponding offline and online models of Eq. (21), which operate under the limited 3C resources capacities by using optimal and least recently used (LRU) caching policies while considering request aggregation. "OPT without Cache and Aggregation" indicates that the existing cloud-edge cooperation scheme does not consider the deployment of caching resource and request aggregation policies, while request aggregation is adopted in "OPT without Cache" [3, 33]. The advantages of in-network caching and aggregation have been discussed in Section II-A. Through performance comparison, we can find their influence on energy consumption in the simulation, respectively.

Figure 2 shows the network energy consumption of the four solutions under different cache sizes. As shown in Figure 2, we can see that the performance of our proposed model performs much better than "OPT without Cache" and "OPT without Cache and Aggregation," where energy consumption of the latter always remains the same. The reason is that there are no caches deployed in the access networks, and all their requests are routed to the cloud to fetch data. Due to that the same content request can be aggregated in the BSs, compared with "OPT without Cache and Aggregation," the network delay in "OPT without Cache" is significantly improved. As the cache size increases, more popular contents are buffered at the network edge, which reduces the performance gap between "Offline with Cache" and "Online with Cache." With the growth of cache size, energy efficiency of BSs converges to a stable state in the proposed model, because additional cached unpopular contents have little effect on energy efficiency of the system.

Figure 3 shows the network energy consumption of the four solutions under different content popularity. As shown in Figure 3, we can see that content popularity has no effect on the performance of "OPT without Cache and Aggregation," because each request must be routed to the cloud to obtain the corresponding content. As content popularity grows, energy efficiency is improved in "Offline with Cache" and "Online with Cache." The reason is that a larger Zipf skewness parameter means more popular contents are requested by the MUs, which makes the majority of requests directly satisfied by the cached data in the SBSs and MBSs. Moreover, the performance of "OPT without Cache" is improved as content popularity increases. The reason is that more content requests are aggregated in the queue.

Figure 4 shows the network energy consumption of the four solutions when the number of different contents varies.

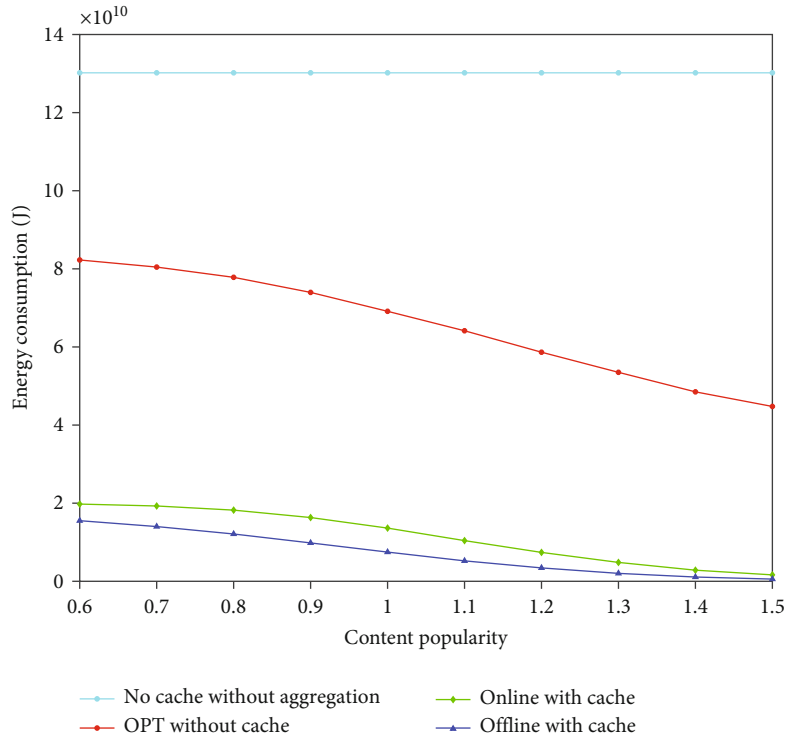

Figure 3: Network energy consumption versus content popularity when $C_{ij} = 0.4\%$, $C_i = 0.8\%$, $F = 1000$, and $\lambda = 10$.

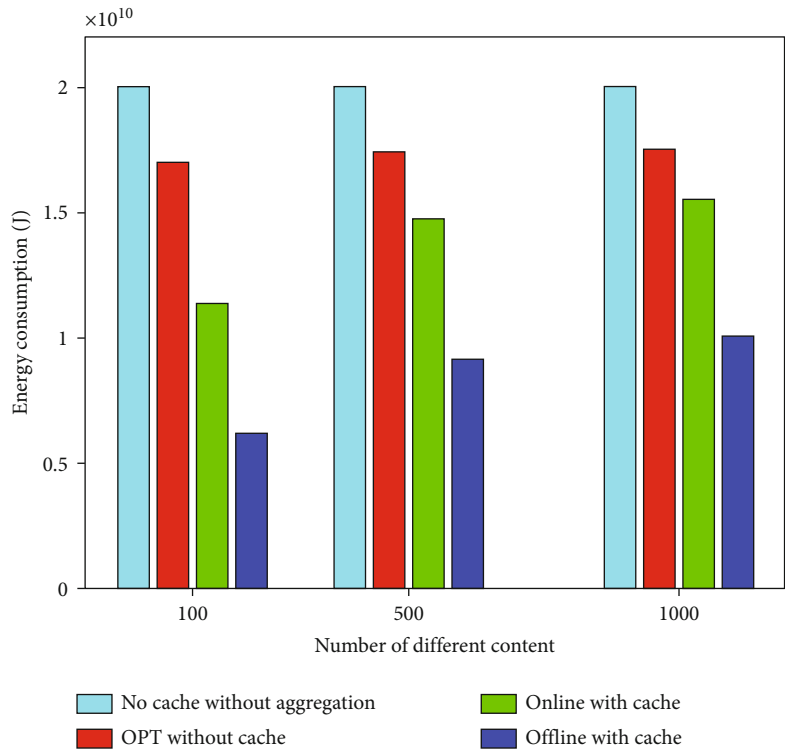

Figure 4: Network energy consumption versus the number of different contents when $C_{ij} = 0.4\%$, $C_i = 0.8\%$, $\alpha = 1.0$, and $\lambda = 10$.

As shown in Figure 4, we can see that the energy consumption of our proposed solution increases with the growth of the number of different contents. The reason is that a larger amount of different contents means more unpopular contents are requested by MUs, which makes more requests unsatisfied in the BSs with limited cache capacity and obtain

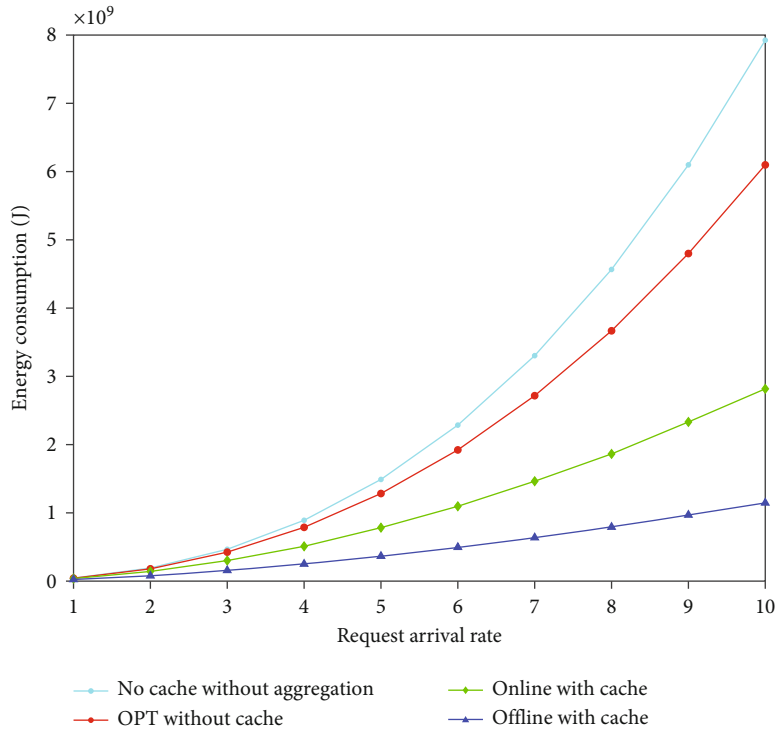

FIGURE 5: Network energy consumption versus request arrival rate when $C_{ij} = 0.4\%$, $C_i = 0.8\%$, $\alpha = 1.0$, and $F = 1000$.

their contents from the cloud. Besides, a larger content diversity can bridge the performance gap between "Offline with Cache" and "Online with Cache," because the increase of more unpopular requests reduces the impact of caching resources on network performance to a certain extent. Compared with in-network caching, the growth of content diversity has limited effect on the request aggregation, which makes the energy consumption of "OPT without Cache" increase slowly. Due to that, each request fetches the corresponding content from the cloud in "OPT without Cache and Aggregation," its performance is not affected by the number of different contents.

Figure 5 shows the network energy consumption of the four solutions under different request arrival rates. As shown in Figure 5, we can see that energy efficiency of the four schemes declines when the request arrival rate increases, because a larger queuing delay results in the growth of energy consumption. Due to the fact that popular contents are always cached at the network edge in "Offline with Cache," the performance gap between "Offline with Cache" and other solutions enlarges with the growth of the request arrival rate. The performance between "OPT without Cache" and "OPT without Cache and Aggregation" has the similar trend. The reason is that the effect of request aggregation is improved with the increase of arrival rate.

5. Conclusions

In this paper, we propose a novel energy-efficient hierarchical collaborative model for content delivery services by considering the in-network caching, request aggregation, and cooperation allocation of caching, computing, and communication resources in a layered heterogeneous network of MUs, SBSs, MBSs, and the cloud. Firstly, we formulate the energy efficiency problem as a centralized model, which can aggregate the same content requests in the queue of each BS, and achieve minimal energy consumption by jointly optimizing different resources on basis of cloud-edge cooperation and request queue. Then, the optimal energy efficiency problem is analyzed based on the distribution characteristic of content popularity. Simulation results show that the performance of our proposed model is much better than the existing cloud-edge cooperation scheme without considering the deployment of caching resource and request aggregation policies.

In future work, we will design a more efficient online scheme to approximate the optimal solution, e.g., predictive and proactive caching about network contents. Besides, we intend to develop a weighted model between network power and delay to achieve their prominent tradeoff. Moreover, the mobile behaviors of network users will be considered to improve our proposed model. Finally, the full-dimensional collaboration problem will be investigated to verify the system performance in the more complex network environment.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partially supported by the Scientific Research Plan of Beijing Municipal Commission of Education under Grant KM201910005026, Beijing Nova Program of Science and Technology Z191100001119094, and Beijing Natural Science Foundation L202016.

References

[1] *Cisco Visual Networking Index: Forecast and Methodology: 2016-2021*, Technical report, Cisco, 2017.

[2] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: a survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2134–2168, 2018.

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

[4] X. He, K. Wang, H. Huang, T. Miyazaki, Y. Wang, and S. Guo, "Green resource allocation based on deep reinforcement learning in content-centric iot," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 3, pp. 781–796, 2020.

[5] X. He, K. Wang, and X. Wenyao, "Qoe-driven content-centric caching with deep reinforcement learning in edge-enabled iot," *IEEE Computational Intelligence Magazine*, vol. 14, no. 4, pp. 12–20, 2019.

[6] L. Haodong, X. He, D. Miao, X. Ruan, Y. Sun, and K. Wang, "Edge qoe: computation offloading with deep reinforcement learning for internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9255–9265, 2020.

[7] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: research issues and challenges in content placement and delivery strategies," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1710–1732, 2018.

[8] H. Guo, J. Liu, and J. Zhang, "Computation offloading for multi-access mobile edge computing in ultra-dense networks," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 14–19, 2018.

[9] Y. Qi and H. Wang, "Interference-aware user association under cell sleeping for heterogeneous cloud cellular networks," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 242–245, 2017.

[10] J. Ren, Y. Guanding, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 5031–5044, 2019.

[11] W. Dapeng, R. Bao, Z. Li, H. Wang, H. Zhang, and R. Wang, "Edge-cloud collaboration enabled video service enhancement: a hybrid human-artificial intelligence scheme," *IEEE Transactions on Multimedia*, vol. 23, pp. 2208–2221, 2021.

[12] H. Zhang, H. Liu, J. Cheng, and V. C. M. Leung, "Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1705–1716, 2018.

[13] D. Han, S. Li, Y. Peng, and Z. Chen, "Energy sharing-based energy and user joint allocation method in heterogeneous network," *IEEE Access*, vol. 8, pp. 37077–37086, 2020.

[14] C. Wang, F. Ying He, Y. Richard, Q. Chen, and L. Tang, "Integration of networking, caching, and computing in wireless systems: a survey, some research issues, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 7–38, 2017.

[15] Z. Li, N. Zhu, D. Wu, H. Wang, and R. Wang, "Energy-efficient mobile edge computing under delay constraints," *IEEE Transactions on Green Communications and Networking*, 2021.

[16] F. Guo, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Joint optimization of caching and association in energy-harvesting-powered small-cell networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6469–6480, 2018.

[17] C. Kai, H. Zhou, Y. Yi, and W. Huang, "Collaborative cloud-edge-end task offloading in mobile-edge computing networks with limited communication capability," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 2, pp. 624–634, 2021.

[18] Q. Chen, Z. Kuang, and L. Zhao, "Multi-user computation offloading and resource allocation for cloud-edge heterogeneous network," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[19] X. Fangmin, F. Yang, S. Bao, and C. Zhao, "Dqn inspired joint computing and caching resource allocation approach for software defined information-centric internet of things network," *IEEE Access*, vol. 7, pp. 61987–61996, 2019.

[20] F. Zhicai Zhang, Y. Richard, F. Fang, Q. Yan, and Z. Wang, "Joint offloading and resource allocation in mobile edge computing systems: an actor-critic approach," in *2018 IEEE global communications conference (GLOBECOM)*, pp. 1–6, Abu Dhabi, United Arab Emirates, 2018.

[21] M. F. Bari, S. R. Chowdhury, R. Ahmed, R. Boutaba, and B. Mathieu, "A survey of naming and routing in information-centric networks," *IEEE Communications Magazine*, vol. 50, no. 12, pp. 44–53, 2012.

[22] C. Fang, C. Liu, X. Hang, Z. Wang, H. Chen, Y. Sun, H. Xiaoyan, D. Zeng, and M. Dong, Eds., "Q-learning based delay-aware content delivery in cloud-edge cooperation networks," in *2021 7th International Conference on Computer and Communications (ICCC)*, pp. 1458–1462, Chengdu, China, 2021.

[23] C. Fang, H. Yao, Z. Wang, W. Wu, X. Jin, and F. R. Yu, "A survey of mobile information-centric networking: research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2353–2371, 2018.

[24] M. Mangili, F. Martignon, S. Paris, and A. Capone, "Bandwidth and cache leasing in wireless information-centric networks: a game-theoretic study," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 679–695, 2016.

[25] A. Alnoman and A. S. Anpalagan, "Computing-aware base station sleeping mechanism in h-cran-cloud-edge networks," *IEEE Transactions on Cloud Computing*, vol. 9, no. 3, pp. 958–967, 2019.

[26] C. Fang, H. Yao, C. Zhao, and Y. Liu, "Modeling energy-delay tradeoffs in single base station with cache," *International Journal of Distributed Sensor Networks*, vol. 2015, 5 pages, 2015.

[27] C. Fang, C. Liu, X. Sheng, M. Zhai, M. Zhang, X. Wen, and Z. Wang, Eds.F. Xu, "An edge cache-based power-efficient content delivery scheme in mobile wireless networks," in *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 294–299, Ho Chi Minh City, Vietnam, 2019.

[28] Y. Jiao and I. Joe, "Energy-efficient resource allocation for heterogeneous cognitive radio network based on two-tier crossover genetic algorithm," *Journal of Communications and Networks*, vol. 18, no. 1, pp. 112–122, 2016.

[29] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *1999 IEEE International Conference on Computer Communications (INFOCOM)*, vol. 1, pp. 126–134, New York, NY, USA, 1999.

[30] J. Li, B. Liu, and W. Hao, "Energy-efficient in-network caching for content-centric networking," *IEEE Communications Letters*, vol. 17, no. 4, pp. 797–800, 2013.

[31] H. Xie, G. Shi, and P. Wang, "TECC: towards collaborative in-network caching guided by traffic engineering," in *2012 IEEE International Conference on Computer Communications (INFOCOM)*, Orlando, FL, USA, 2012.

[32] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In network caching effect on optimal energy consumption in content-centric networking," in *2012 IEEE International Conference on Communications (ICC)*, pp. 2889–2894, Ottawa, ON, Canada, 2012.

[33] C. Fang, C. Liu, Z. Wang et al., "Cache-assisted content delivery in wireless networks: a new game theoretic model," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2653–2664, 2021.