Wireless Personal Communications manuscript No.
(will be inserted by the editor)

Edge Caching and Computation Offloading for Fog-enabled Radio Access Network

Xiaohuan Rao[1] · **Liqiang Zhao**[1] · **Kai Liang**[1] · **Kezhi Wang**[2]

Received: date / Accepted: date

Abstract Cooperation between the fog and cloud in fog-enabled radio access network (F-RAN) could enormously improve the computation offloading services both in terms of latency and energy consumption, while edge caching allowing for low-latency transmission of the cached files, is essential for cooperation. It is necessary to select the best method that minimizes the energy consumption with latency tolerance guaranteed, to offload the task. In this paper, we study the problem of content placement for edge caching and the cooperative computation offloading between the fog and the cloud by optimizing the cooperative offloading decisions and caching strategies. The problem is formulated to minimize the total energy consumption of all the UEs by optimizing the transmit power of the user equipments (UEs), caching strategies and the *cooperation factor* (the ratio of the fog execution of a file to the whole file), which is a non-convex programming problem. An approximate solution is obtained by using successive convex approximation (SCA) strategy and implicit enumeration method. Numerical results show that our proposed algorithm can achieve better performance compared with other schemes, e.g., cloud computing and cache most popular (CMP) strategies.

Xiaohuan Rao
E-mail: rxh_0601@yeah.net

Liqiang Zhao
E-mail: lqzhao@mail.xidian.edu.cn

Kai Liang
E-mail: kliang@xidian.edu.cn

Kezhi Wang
E-mail: kezhi.wang@northumbria.ac.uk

[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xian, Shaanxi, China, 710071

[2] Department of Computer and Information Sciences Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom

Keywords Computation offloading · edge caching · cloud computing · fog computing

1 Introduction

Recently, smart user equipments (UEs) have made a great impact on people's everyday life and usually need to run complex applications, which demands prodigious computation capacity and leads to intensive energy consumption. However, UEs are usually resource-constrained, suffering from limited battery lifetime and computation capacity, which can't satisfy the demand of high-computational applications. In this situation, great attention has been attracted to cloud computing which possesses enormous computation and storage capacity, and it has been applied to a substantial number of applications [1–5]. Nevertheless, because of its remote location from the user terminals[6], it is costly in both time latency and energy consumption. To overcome the shortcomings above, an evolved architecture which is known as fog-enabled radio access network (F-RAN) has been proposed [7–9]. In an F-RAN, fog node is equipped with storage and computation capacity, which makes it possible to achieve co-processing between cloud and fog, and jointly optimize the cooperative offloading decisions and caching strategies.

A promising approach to reduce the time latency and save energy is to relieve unnecessary traffic pressure by jointly optimizing the cooperative offloading decisions and caching strategies in the F-RAN [7]. Caching in the fog node can alleviate backhaul load as well as eliminate the backhaul bottleneck by pre-storing parts of files in the fog node at off-peak periods, which greatly improves the quality of servise (QoS) of users [10]. During peak traffic periods, the cached files can be processed in the fog node and then conveyed to the user terminals. Due to the limited storage and computation capacity in the fog node, we consider that the request files are cooperatively processed in the fog and cloud before conveying to the user terminals. If the request files are stored in both cloud and fog node, one part will be processed in the fog, and the remain will be computed in the cloud. If the request files only exist in the cloud, the files will be processed totally in the cloud.

Earlier efforts have been done on the edge caching for F-RAN. The authors in [11] analyzed unencoded and encoded micro-caches to minimize the total average latency of all users. [12] proposes the most popular content (MPC) strategy and maximizing file density caching strategy, as well as cooperative transport in cluster-centric small networks. In [13], a random distributed model of base station with caching capability is established, and the expressions of outage probability and average transmission rate are obtained. [14] has proposed another random framework in D2D communication with cache capability, and studied the performance indicators of local and global distribution of quantitative service content requests. [15] studies the problem of optimizing the system cost in the case of considering the fronthaul cost when meeting the goal of signal to interference plus noise ratio (SINR) constraint. In [16], the au-

thors considered the computing resources in the fog nodes and proposed a joint resource allocation and coordinated offloading method for F-RANs. [10] studied the joint design of cloud and edge processing for the downlink of an F-RAN and compared the performance of different caching strategies. [17, 18] studied the trade-off between caching, fronthaul capacity and delivery latency. Nevertheless, the works aforementioned adopted fixed caching strategies and the performance needs to be further improved. [19] proposed an energy-efficient caching scheme for fog computing and compared the performance with no caching scheme and random caching scheme. The authors in [20] proposed a game theoretic approach for the computation offloading decision making problem for mobile-edge cloud computing (MEC). [21] investigated the problem of the joint optimization of radio and computational resources aiming at minimizing MUs' energy consumption under latency and power budget constraints. The caching and computation resources at MEC server were studied in [22].

However, the cooperative computation offloading between the fog and cloud, combined with the optimized caching strategies in the F-RAN, has not been well investigated. In this paper, we optimize the cooperative offloading decisions and caching strategies by taking full advantage of the caching and computation capacity of the fog. Compared with the existing literatures, our main contributions are summarized as follows:

- We consider the computation and storage resources in the fog so that the request files can be cached and computed in the fog, which can achieve co-processing between cloud and fog. In addition, according to different conditions we can obtain corresponding caching strategies, which can achieve to minimize the energy consumption and take full use of the total cache capacity of the fog node.
- The problem is formulated to minimize the total energy consumption of all UEs by jointly optimizing the transmit power of the UEs, caching strategies and the cooperation factor, which is a non-convex programming problem. An approximate solution is obtained by using successive convex approximation (SCA) strategy. Through the convex optimization algorithm, we find the optimal cooperation factor and caching strategy that minimize the energy consumption.
- We provide simulations to show the performance of the proposed algorithm. We compare the proposed algorithm with the cloud computing and cache most popular (CMP) strategies, and it indicates that the cooperation between the fog and cloud can achieve better performance.

2 System Model

As shown in Fig .1, the system model introduced in this paper is a three-tier network architecture, consisting of cloud layer, fog layer, and user terminal layer. Denote the i^{th} file in the cloud as f_i, where i is the popularity rank of files. The fog node equipped with storage and computation capacity, can

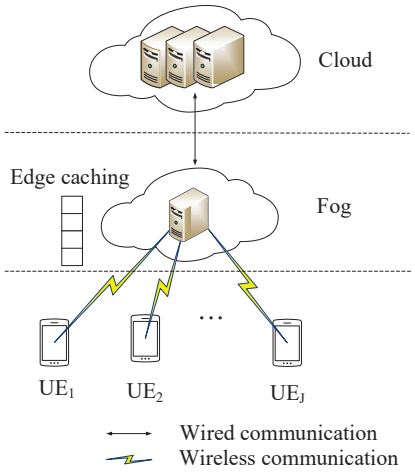

Fig. 1 System model.

process the UEs' request files and deliver them to the UEs through the wireless link. The cloud server with strong computing capacity, can process the request files after receiving the requesting information from the fog. We make an assumption that the cloud server is connected to the large data center, so that all the request files can be stored in the cloud. However due to the limited cache capacity, the fog node can only store a part of the files. The UEs may send request information to their serving fog node when requesting files, and then the fog will send the request information to the cloud. If the requested file is cached in the fog, it will be collaboratively processed by cloud and fog. Otherwise, it will be processed by the cloud only. For simplicity, we assume that there is only one cloud server, one fog node, and J UEs in the network architecture. Denote B_i as the size of the request information, I_i as the calculation quantity for processing the request file f_i and G_i as the size of a request file f_i. Assume that there are F different files requested by all UEs and each UE requests F_{UE} of the F files, also all files are of different sizes.

We define binary requesting variables $q_{i,j}$ as

$$q_{i,j} = \begin{cases} 1, & \text{if the file } f_i \text{ is requested by the } j^{th} \text{ UE} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We emphasize coordinated computation offloading between the cloud and the fog in this paper. In particular, each request file needs to be downloaded from a cloud server or a fog server. There are three data transmission stages in this paper including the wireless transmission stage, the wired transmission stage and the data processing stage.

In the wireless transmission stage for the uplink, the UE transmits the request information to its serving fog node. We assume that the perfect channel state information (CSI) is known at the fog node. The transmission rate of the

j^{th} UE is give by [7]

$$R_j^{UL} = W\log_2\left(1 + \frac{p_j|h_j v_j|^2}{W\sigma^2}\right) \tag{2}$$

where h_j represents the wireless channel between j^{th} UE and the fog for uplink, suffering a joint path loss and multipath fading; σ^2 is the power of additive white Gaussian noise (AWGN); p_j denotes the transmission power of j^{th} UE; v_j indicates CoMP-ZF signal detecting vector [23, 24]; W is the bandwidth of the wireless channel, following the Rayleigh distribution with zero mean parameter and unit variance matrix.

Similar to the transmission rate for the uplink, the wireless transmission rate for the downlink is given by

$$R_j^{DL} = W\log_2\left(1 + \frac{P_{FC}|g_j v_j|^2}{W\sigma^2}\right) \tag{3}$$

where g_j is the wireless channel between j^{th} UE and the fog for downlink, suffering a joint path loss and multipath fading; P_{FC} indicates the the transmission power of the fog node; other notations remain the same as (2).

2.1 Content Caching Rules

It is reported that the popularity of the contents follow a Zipf distribution which is a kind of power law distribution [25]. There are different kinds of caching strategies, e.g., cache most popular (CMP), cache distinct (CD) and fractional cache distinct (FCD) [10]. However, the content popularity often updates, so adopting fixed caching strategies can't achieve to minimize the energy consumption and neither take full use of the total cache capacity of the fog node. In this paper, we propose an algorithm to obtain the best caching strategy that can minimize the energy consumption.

We define binary caching variables d_i as

$$d_{i,j} = \begin{cases} 1, & if\ the\ file\ f_i\ is\ cached\ in\ fog\ node \\ 0, & otherwise \end{cases} \tag{4}$$

while satisfying the cache capacity of the fog node as

$$\sum_{j=1}^{J}\sum_{i=1}^{F} d_i G_i \leq S \tag{5}$$

where S represents the cache capacity of the fog node.

2.2 The Latency and Energy Analysis of Different Transmission Stage

2.2.1 Wireless Transmission Stage

When having a demand for the file, the UE sends the request information to the fog node through wireless transmission, whose transmission latency is given by

$$t_{1,i,j} = \frac{q_{i,j} B_i}{R_j^{UL}} \tag{6}$$

The corresponding energy consumption is given by

$$e_{1,i,j} = t_{1,i,j} p_j \tag{7}$$

For the wireless transmission stage in the downlink, the processed request file is conveyed to the UE from the fog with the latency and energy consumption:

$$t_{2,i,j} = \frac{q_{i,j} G_i}{R_j^{DL}} \tag{8}$$

$$e_{2,i,j} = t_{2,i,j} P_{FC} \tag{9}$$

2.2.2 Data Processing Stage

When receiving the user's request information, if the request file is only processed in the fog, the corresponding latency and energy consumption are given by

$$t_{3,i,j} = \frac{q_{i,j} I_i}{S_{f,j}} \tag{10}$$

$$e_{3,i,j} = t_{3,i,j} P_F \tag{11}$$

where $S_{f,j}$ represents the computing resource allocated for j^{th} UE in the fog; P_F stands for the power consumption of processing the request files in the fog.

After receiving the request information from the fog, if the request file is computed in the cloud, the corresponding latency and energy consumption are respectively given by

$$t_{4,i,j} = \frac{q_{i,j} I_i}{S_{c,j}} \tag{12}$$

$$e_{4,i,j} = t_{4,i,j} P_C \tag{13}$$

where $S_{c,j}$ indicates the computing resource allocated for j^{th} UE in the cloud; P_C stands for the power consumption of processing the request files in the cloud.

2.2.3 Wired Transmission Stage

For the wired transmission stage in the uplink, the fog server conveys the request information to the cloud server through wired transmission, whose latency and energy consumption are respectively given by

$$t_{5,i,j} = \frac{q_{i,j}B_i}{c} \tag{14}$$

$$e_{5,i,j} = t_{5,i,j}P_{FC} \tag{15}$$

where c is the wired link bandwidth allocated to the user.

In the downlink, the processed request file is sent back to the fog with the latency and energy consumption:

$$t_{6,i,j} = \frac{q_{i,j}G_i}{c} \tag{16}$$

$$e_{6,i,j} = t_{6,i,j}P_{CF} \tag{17}$$

where P_{CF} is the power consumption of transmitting the processed request file from the cloud to the fog.

3 Problem Formulation and Solution

3.1 Problem Formulation

In this work, we aim at minimizing the energy consumption of the entire communication process by jointly optimizing the cooperative offloading decisions and the allocation of computation and storage resources for the F-RAN. In the following we will first discuss the latency and energy consumption caused by fog computing, cloud computing and cooperation between the fog and cloud, respectively.

3.1.1 Fog Computing

If all the request files are dealt with in the fog and then delivered to the UEs. The corresponding latency and energy consumption are respectively given by

$$t_{f,i,j} = t_{1,i,j} + t_{2,i,j} + t_{3,i,j} \tag{18}$$

$$e_{f,i,j} = e_{1,i,j} + e_{2,i,j} + e_{3,i,j} \tag{19}$$

3.1.2 Cloud Computing

If all the request files are processed in the cloud and then delivered to the UEs. The corresponding latency and energy consumption are respectively given by

$$t_{c,i,j} = t_{1,i,j} + t_{2,i,j} + t_{4,i,j} + t_{5,i,j} + t_{6,i,j} \tag{20}$$

$$e_{c,i,j} = e_{1,i,j} + e_{2,i,j} + e_{4,i,j} + e_{5,i,j} + e_{6,i,j} \tag{21}$$

3.1.3 Cooperation between the Fog and Cloud

In this paper, we consider that the request files of the UEs are processed collaboratively in both cloud and fog, which has the energy consumption of

$$e_{i,j} = \lambda_i d_i e_{f,i,j} + (1 - \lambda_j d_{i,j}) e_{c,i,j} \tag{22}$$

where λ_i is the cooperation factor which is defined as the ratio of the fog execution of a file to the whole file. Obviously, λ_i is limited to

$$0 \leq \lambda_i \leq 1. \tag{23}$$

Substituting (19) and (21) into (22), we can get the following expression

$$e_{i,j} = e_{1,i,j} + e_{2,i,j} + \lambda_i d_i e_{3,i,j} + (1 - \lambda_i d_{i,j}) (e_{4,i,j} + e_{5,i,j} + e_{6,i,j}) \tag{24}$$

The j^{th} UE with the transmit power of p_j has a tolerance for the latency as follows:

$$\sum_{i=1}^{F} \{\lambda_i t_{f,i,j} + (1 - \lambda_i d_i) t_{c,i,j}\} \leq T \tag{25}$$

where T represents the maximum latency tolerance of the j^{th} UE in the entire communication process.

ubstituting (18) and (20) into (25), we can get the following expression

$$\sum_{i=1}^{F} (t_{1,i,j} + t_{2,i,j} + \lambda_i d_i t_{3,i,j} + (1 - \lambda_i d_i) (t_{4,i,j} + t_{5,i,j} + t_{6,i,j})) \leq T \tag{26}$$

In this section, we assume that the wireless link rate assigned to the j^{th} UE is always less than the rate of the wired link. The constraint on the link rate is given by

$$R_j^{UL} \leq c \tag{27}$$

$$\sum_{j=1}^{J} R_j^{UL} \leq C_{FC} \tag{28}$$

where C_{FC} is the total bandwidth of the wired link between the fog and the cloud server, and $c \leq C_{FC}/J$.

Because of the limited cache capacity of the fog, the files cached in the fog node should be no more than the cache capacity of the fog server, based on which the constraint is given by (5).

With respect to the transmission power of j^{th} UE , the following constraint is satisfied:

$$0 \leq p_j \leq P_m \tag{29}$$

where P_m is the maximum transmission power of each UE.

In this section, we focus on the energy consumption of the entire communication system. Therefore, the optimization problem can be regarded as minimizing the energy consumption of the whole system under the constraint of satisfying the latency of each user and so on. According to the analysis of the latency and energy consumption in the previous section, and the description of the constraints (23), (26), (27), (28) and (29). In this section, the problem can be formulated by the following:

$$\begin{aligned}
&\min_{p_j,\lambda_i,d_i} \sum_{j=1}^{J}\sum_{i=1}^{F}\{e_{1,i,j}+e_{2,i,j}+\lambda_i d_i e_{3,i,j}+(1-\lambda_i d_{i,j})\left(e_{4,i,j}+e_{5,i,j}+e_{6,i,j}\right)\}\\
&s.t.\quad C1:\sum_{i=1}^{F}\{t_{1,i,j}+t_{2,i,j}+\lambda_i d_i t_{3,i,j}+(1-\lambda_i d_i)\left(t_{4,i,j}+t_{5,i,j}+t_{6,i,j}\right)\}\le T\\
&\quad C2:\ R_j^{UL}\le c\\
&\quad C3:\sum_{j=1}^{J}R_j^{UL}\le C_{FC}\\
&\quad C4:\sum_{i=1}^{F}d_i G_i\le S\\
&\quad C5: 0\le p_j\le P_m\\
&\quad C6: 0\le\lambda_i\le 1\\
&\quad C7: d_i=0\ or\ 1\\
&\quad j=1,...,J, i=1,...,F
\end{aligned}\tag{30}$$

where $e_{1,i,j}$ is a concave function, $C3$ is a convex constraint and d_i is 0-1 caching variables. Therefore, problem (30) is a non-convex problem. It is difficult to solve problem (30) for the optimal solution. In the next subsection, we will propose optimization algorithms by using SCA strategy and implicit enumeration method to solve problem (30) and get an approximate solution.

3.2 Problem Solution and Optimization Algorithm

In the objective function described in problem (30), $e_{1,i,j}$ is a concave function, d_i is 0-1 caching variables and the rest of the objective function are either a constant or a convex function. We notice that

$$e_{1,i,j}\le\tilde{e}_{1,i,j}=\frac{P_m B_i}{R_j^{UL}}\tag{31}$$

$\tilde{e}_{1,i,j}$ is convex and approximate to $e_{1,i,j}$, therefore we can substitute $\tilde{e}_{1,i,j}$ for $e_{1,i,j}$. Let $\tilde{E}$ represent the objective function after substitution and T_j represent the time consumption of j^{th} UE, that is $\tilde{E}=\sum_{j=1}^{J}\sum_{i=1}^{F}\{\tilde{e}_{1,i,j}+e_{2,i,j}+\lambda_i d_i e_{3,i,j}+(1-\lambda_i d_i)\left(e_{4,i,j}+e_{5,i,j}+e_{6,i,j}\right)\}$, and $T_j=\sum_{i=1}^{F}\{t_{1,i,j}+t_{2,i,j}+\lambda_i d_i t_{3,i,j}+(1-\lambda_i d_i)\left(t_{4,i,j}+t_{5,i,j}+t_{6,i,j}\right)\}$.

Since the constraint $C3$ is non-convex [26], the problem (30) is not easy to solve. Constraint $C2$ limits the upper bound of the transmission rate of each UE, so we can use the access control method in the network in [27] to ensure that the number of users in the network will not exceed $\frac{C_{FC}}{c}$. In effect, this ensures that we can ignore the constraints $C3$.

The problem (30) is still a non-convex programming through the above treatment due to the 0-1 caching variables d_i. We use SCA strategy and implicit enumeration method to solve this problem, and the corresponding algorithms are summarized in Algorithm 1 and Algorithm 2, respectively. Algorithm 1 jointly optimizes the cooperative offloading decisions and caching strategies by using SCA strategy and Lagrangian methods. Algorithm 2 invoked by Algorithm 1, solves for binary caching variables d_i.

Algorithm 1 Joint optimization of edge caching and computation offloading algorithm

Input: B_i,I_i,G_i,P_F,P_C,P_{FC},P_{CF},h_i,g_j,S,$S_{c,j}$,$S_{f,j}$,$q_{i,j}$
Output: p_j, λ_i, d_i
Initialize the cooperative factor $\lambda_i^0 \in [0,1]$, transmit power $p_j^0 \in [0, P_m]$; Set $k = 1$;
loop
  Update d_i^k by invoking Algorithm 2 with fixed λ_i^{k-1}, p_j^{k-1}, $\forall i, j$;
  Update λ_i^k and p_j^k by solving problem (32) using Lagrangian methods with fixed d_i^k, $\forall i, j$;
  if certain stopping criterion is satisfied **then**
    $\lambda_i^* = \lambda_i^k$, $p_j^* = p_j^k$, $d_i^* = d_i^k$,break;
  end if
  $k = k + 1$;
end loop.

Algorithm 2 The algorithm of solving for binary caching variables d_i

Input: B_i,I_i,G_i,P_F,P_C,P_{FC},P_{CF},h_j,g_j,S,$S_{c,j}$,$S_{f,j}$,$q_{i,j}$,λ_i,p_j
Output: d_i,$\tilde{E}$
Utilize heuristics to find an initial feasible solution d_i^0, and compute the corresponding energy consumption $\tilde{E}^0$; Set $k = 0$;
loop
  Add a filter condition $\tilde{E}^k < \tilde{E}^{k-1}$ to the original question (32);
  if $\tilde{E}^k < \tilde{E}^{k-1}$ **then**
    check whether other constraints meet the requirements;
    if all other constraints are satisfied **then**
      Update d_i^0 with d_i^k, continue;
    end if
  end if
  $k = k + 1$;
  if all possible combinations of d_i are checked **then**
    d_i^*=d_i^0;
  end if
end loop

We obtain the 0-1 caching variables d_i through Algorithm 2 and then the problem (30) becomes convex with the solved d_i as follows

$$\begin{aligned} &\min_{p_j,\lambda_i} \ \tilde{E} \\ &s.t. \quad C1, C2, C5, C6 \end{aligned} \tag{32}$$

Upon fixing the caching variables d_i, we can solve problem (32) by using Lagrangian methods.

The Lagrangian function of (32) is given by

$$\mathcal{L} = \tilde{E} + \sum_{j=1}^{J} \beta_j \left(T_j - T\right) + \sum_{j=1}^{J} \gamma_j \left(R_j^{UL} - c\right) + \sum_{i=1}^{F} \mu_i \left(\lambda_i - 1\right) \tag{33}$$

where β_j, γ_j and μ_i are the scalar Lagrange multipliers associated with C1, C2 and C6, respectively.

Applying the Karush-Kuhn-Tucker (KKT) optimality conditions to (32) and the optimal can be found as

$$p_j^* = \min\left(\frac{\left(2^{\frac{c}{W}} - 1\right) W\sigma^2}{\left|h_j v_j^{CoMP-ZF}\right|^2}, P_m\right) \tag{34}$$

If $\beta_j \neq 0$, then

$$\begin{aligned} &\sum_{i=1}^{F} \lambda_i^* d_i \left(t_{3,i,j} - t_{4,i,j} - t_{5,i,j} - t_{6,i,j}\right) \\ &= T - \sum_{i=1}^{F} \left(t_{1,i,j} + t_{2,i,j} + t_{4,i,j} + t_{5,i,j} + t_{6,i,j}\right) \end{aligned} \tag{35}$$

otherwise $\lambda_i^* = 1$.

In Algorithm 1, the first invoke Algorithm 2 to obtain the 0-1 caching variables d_i. Then utilize Lagrangian methods with solved d_i to solve the optimization problem. By constantly updating the power or choosing to continually update the cooperation factor, we can obtain the optimal solution when the convergence conditions are satisfied.

We can easily analyze through Algorithm 1 and Algorithm 2 that the proposed algorithm has complexity of $O(JF_{UE})$, where J is the number of the UEs served by one fog node and F_{UE} is the number of files requested by each UE. Since the number of UEs served by the fog node is limited and the number of files that need to be processed of each UE is not large, the complexity of the proposed algorithm is relatively low. On the other hand, by following a similar approach as that presented in [28], we will present the average iteration numbers of the proposed algorithm in the next section, which can further reflect the relatively fast convergence speed of the proposed algorithm.

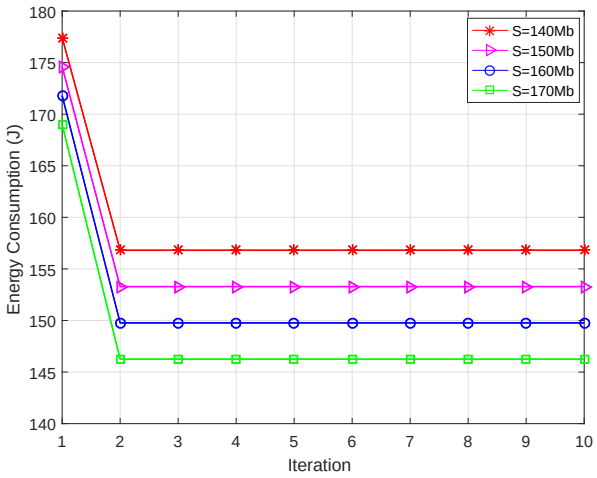

Fig. 2 Average iteration numbers

4 Numerical Results

In this section, we present system simulation results to evaluate the performance of the proposed algorithm. In the simulation, the basic parameters unless otherwise specified, are as follows. For convenience, we assume that there is one fog node serving 4 UEs, altogether requesting $F = 10$ different files and all files are of different sizes with randomly given value between 15 Mb and 50 Mb. Each UE requests $F_{UE} = 4$ files. The maximum transmission power of each UE is $P_m = 0.3W$ and the power of the channel noise is $\sigma^2 = -90dBm$. The average distance between each UE and its serving fog node is $D = 50$ meters and the path loss exponent is $\alpha = 2$. The bandwidth of wireless channel is $W = 10MHz$. We assume the computing resources allocated for each UE from the fog and cloud are the same respectively, $S_f = 25$ million instructions/s (MIPS) and $S_c = 250$ MIPS. The power consumption of processing the files in the fog and the cloud are $P_F = 1W$ and $P_C = 4W$ respectively.The transmitting power of the fog and cloud are $P_{FC} = 20W$ and $P_{CF} = 40W$ respectively. The maximum cache capacity of the fog node is 200 Mbit. The total bandwidth of the wired link between the fog and the cloud is $C_{FC} = 100Mbps$, and the latency tolerance of each UE is $T = 20s$.

The average iteration numbers of the proposed algorithm are depicted in Fig .2. It is clear that the optimal solutions can be obtained within 2 iterations under different cache capacity of fog, which shows fast convergence.

Fig .3 shows the energy consumption of different algorithms regarding to cache capacity of fog. The proposed algorithm is compared with cloud computing and CMP caching strategy. It is obvious that the energy consumption of the cloud computing remains constant with the increasing cache capacity, because all the request files are processed in the cloud and the energy con-

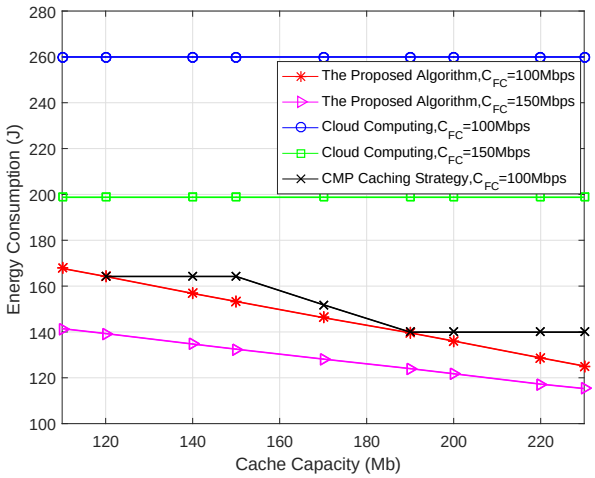

Fig. 3 Energy consumption versus cache capacity in the fog

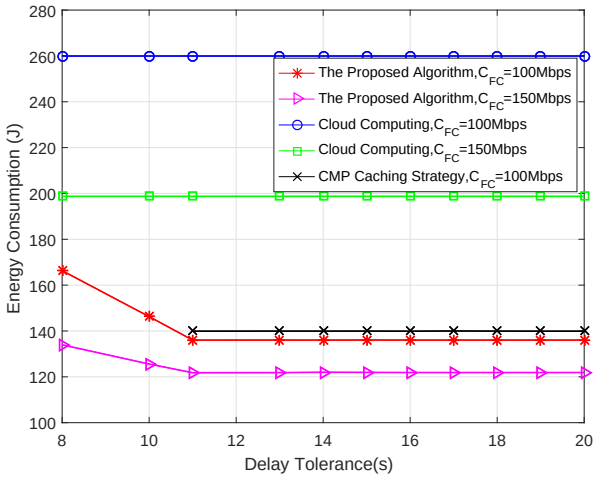

Fig. 4 Energy consumption versus delay tolerance

sumption is not related to the cache capacity. The energy consumption of the proposed algorithm and CMP caching strategy shows a downward trend with the increasing cache capacity, and the proposed algorithm consumes the lowest energy among the three algorithms. This is because as the cache capacity increases, more request files will be cached in the fog, which contributes to the cooperation between the cloud and fog. For the CMP caching strategy, the energy consumption in a certain period remains constant, because the request file can't be cached in the fog when the size of this request file is greater than the growth step of the cache capacity, and it can't reduce the energy

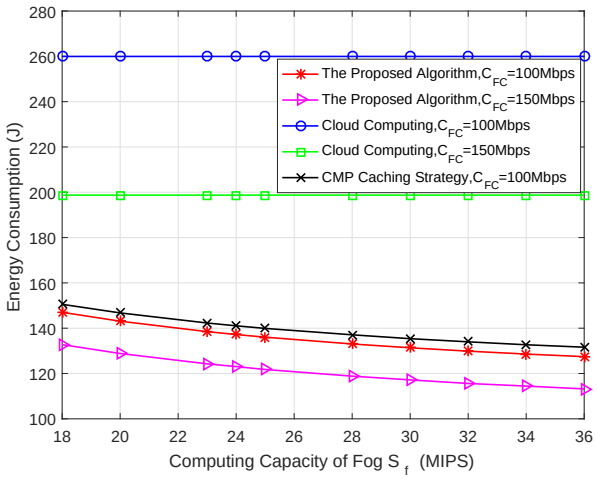

Fig. 5 Energy consumption versus computing capacity of fog

consumption. We can also see that the larger total bandwidth C_{FC} consumes lower energy than the smaller C_{FC} because larger C_{FC} needs less time to convey the request files and on the other hand the transmit power of the cloud and fog remains constant.

Fig .4 depicts the energy consumption of different algorithms regarding to delay tolerance. We can see that the energy consumption of the cloud computing remains unchanged as the delay tolerance increases, because with other conditions unchanged the energy consumption of the cloud computing is constant, which is independent of delay tolerance. The energy consumption of the proposed algorithms first decreases and then stays the same, since when the delay tolerance is small, the energy consumption is limited to the delay, and when the delay tolerance becomes larger the energy consumption will be limited to the cache capacity of fog node. On the other hand, when the delay tolerance becomes larger, more request files will be processed in the fog to reduce the energy consumption. The CMP caching strategy needs more time latency, so that it is infeasible when the delay tolerance is small.

Fig .5 shows the energy consumption versus the various computing capacity of fog node S_f. It can be seen from the figure that the energy consumption of the cloud computing remains constant as S_f increases, because all the request files are processed in the cloud and it is not related to the computing capacity of fog node. The energy consumption of the proposed algorithm and CMP caching strategy decreases with the increasing S_f, since the larger S_f is, the less time it will take to execute the request files and the corresponding energy consumption will be lower. We can also see that the proposed algorithm consumes the lowest energy among the three algorithms.

5 Conclusion

In this paper, we study the problem of caching strategies for edge caching and the cooperative computation offloading between the fog and cloud by optimizing the cooperative offloading decisions and caching strategies. It is considered to minimize the energy consumption of the system with the constraints satisfied, which is a non-convex programming problem. We use the SCA strategy and implicit enumeration method to obtain an approximate solution. The analysis of numerical results show that the proposed algorithm can reduce the energy consumption with the constraints satisfied compared with other schemes.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (61771358), National Natural Science Foundation of Shaanxi Province (2018JM6052), and the 111 Project (B08038).

References

1. R. Pitchai, S. Jayashri, and J. Raja, "Searchable Encrypted Data File Sharing Method Using Public Cloud Service for Secure Storage in Cloud Computing," *Wireless Personal Communications*, vol. 90, no. 2, pp. 947-960, 2016.
2. David S. Linthicum, "Cloud Computing Changes Data Integration Forever: What's Needed Right Now," *IEEE Cloud Computing*, vol. 4, no. 3, pp. 50-53, 2017.
3. Hwan-Seok Yang, and Seung-Jae Yoo, "A Study on Smartwork Security Technology Based on Cloud Computing Environment," *Wireless Personal Communications*, vol. 94, no. 3, pp. 445-454, 2017.
4. X. Wang, K. Wang, S. Wu, et al. "Dynamic Resource Scheduling in Mobile Edge Cloud with Cloud Radio Access Network," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 11, pp. 2429-2445, 2018.
5. H. Mei, K. Wang, and K. Yang, "Multi-Layer Cloud-RAN With Cooperative Resource Allocations for Low-Latency Computing and Communication Services," *IEEE Access*, vol. 5, pp. 19023 - 19032, 2017.
6. Nur Idawati Md Enzai, and Maolin Tang, "A Taxonomy of Computation Offloading in Mobile Cloud Computing," *2014 2nd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, pp. 19-28, 2014.
7. Mugen Peng, Shi Yan, Kecheng Zhang, and Chonggang Wang, "Fog-computing-based Radio Access Networks: Issues and Challenges," *IEEE Network*, vol. 30, no. 4, pp. 46-53, 2016.
8. S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless Communications in the Era of Big Data," arXiv:1508.06369, Aug. 2016.

9. China Mobile, "Next Generation Fronhtaul Interface," White Paper, Oct. 2015.
10. SH. Park, O. Simeone, and SS. Shitz, "Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 11, pp. 7621-7632, 2016.
11. Karthikeyan Shanmugam, Negin Golrezaei, et al. "FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers," *IEEE Trans. on Information Theory*, vol. 59, no. 12, pp. 8402-8413, 2013.
12. Zheng Chen, Jemin Lee, Tony Q. S. Quek, et al. "Cooperative Caching and Transmission Design in Cluster-Centric Small Cell Networks," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 5, pp. 3401-3415, 2017.
13. Ejder Bastug, Mehdi Bennis, and Merouane Debbah, "Cache-enabled Small Cell Networks: Modeling and Tradeoffs," *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, pp. 649-653, 2014.
14. Andres Altieri, Pablo Piantanida, Leonardo Rey Vega, et al. "On Fundamental Trade-offs of Device-to-Device Communications in Large Wireless Networks," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 9, pp. 4958-4971, 2015.
15. Yuanming Shi, Jun Zhang, and Khaled B. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 5, pp. 2809 - 2823, 2014.
16. Kai Liang, Liqiang Zhao, Xiaohui Zhao, et al. "Joint Resource Allocation and Coordinated Computation Offloading for Fog Radio Access Networks," *China Communications*, vol. 13, pp. 131-139, 2016.
17. A Sengupta, R Tandon, and O Simeone, "Cloud RAN and Edge Caching: Fundamental Performance Trade-Offs," *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1-5, 2016.
18. R. Tandon and O. Simeone, "Cloud-Aided Wireless Networks with Edge Caching: Fundamental Latency Trade-Offs in Fog Radio Access Networks," *IEEE International Symposium on Information Theory*, pp. 2029 - 2033, 2016.
19. Siming Wang, Xumin Huang, Yi Liu, and Rong Yu, "CachinMobile: An Energy-Efficient Users Caching Scheme for Fog Computing," *2016 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1-6, 2016.
20. Xu Chen, Lei Jiao, Wenzhong Li, and Xiaoming Fu, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795-2808, 2016.
21. S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89-103, 2015.
22. Anselme Ndikumana, Saeed Ullah, Tuan LeAnh, Nguyen H. Tran, and Choong Seon Hong, "Collaborative Cache Allocation and Computation Offloading in Mobile Edge Computing," *Network Operations and Management*

Symposium (APNOMS), pp.366-369, Sept. 2017.
23. J. Park, G. Lee, Y. Sung, and M. Yukawa, "Coordinated Beamforming With Relaxed Zero Forcing: The Sequential Orthogonal Projection Combining Method and Rate Control," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3100-3112, 2013.
24. W. Yu, T. Kwon, and C. Shin, "Multicell Coordination via Joint Scheduling, Beamforming, and Power Spectrum Adaptation," *IEEE Trans. Commun.*, vol. 12, no. 7, pp. 1-14, 2013.
25. A. Tatar et al., "A Survey on Predicting the Popularity of Web Content," *Springer J. Internet Services and Applications*, vol. 5, no. 1, pp. 1-20, 2014. "Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 11, pp. 7621-7632, 2016.
26. S. Yamada, T. Tanino, M. Inuiguchi, and K. Tatsumi, "An Inner Approximation Method for a Reverse Convex Programming Problem," *IEEE SMC 1999*, pp. 521-526, 1999.
27. Yuanchao Shu, Yu Jason Gu, and Jiming Chen, "Dynamic Authentication with Sensory Information for the Access Control Systems," *IEEE Trans. on Parallel and Distributed Systems*, vol. 25, no. 2, pp.427-436, 2014.
28. Kai Liang, Liqiang Zhao, Kun Yang, et al. "Online Power and Time Allocation in MIMO Uplink Transmissions Powered by RF Wireless Energy Transfer," *IEEE Trans. on Veh. Technol.*, vol. 66, no. 8, pp. 6819-6830, 2017.