Joint wireless and computational resource allocation for ultra-dense mobile-edge computing networks

Junyi Liu[1,2], Hongbing Huang[1], Yijun Zhong[1], Jiale He[1], Tiancong Huang[2*], Qian Xiao[2] and Weiheng Jiang[2]

[1] State Grid Zhejiang Electric Power Co.,Ltd Information & Telecommunication Branch Zhejiang, 310007, China
[e-mail: shshzaa@163.com, james_hb@163.com, zhongyijun1984@163.com, nyhejiale@163.com]
[2] College of Communication Engeering, Chongqing University, Chongqing, 400044, China
[e-mail: 303753773@qq.com, qxiao@cqu.edu.cn,whjiang@cqu.edu.cn]
*Corresponding author: Tiancong Huang

Received July 24, 2018; revised May 10, 2019; accepted May 21, 2020; published July 31, 2020

Abstract

In this paper, we study the joint radio and computational resource allocation in the ultra-dense mobile-edge computing networks. In which, the scenario which including both computation offloading and communication service is discussed. That is, some mobile users ask for computation offloading, while the others ask for communication with the minimum communication rate requirements. We formulate the problem as a joint channel assignment, power control and computational resource allocation to minimize the offloading cost of computing offloading, with the precondition that the transmission rate of communication nodes are satisfied. Since the formulated problem is a mixed-integer nonlinear programming (MINLP), which is NP-hard. By leveraging the particular mathematical structure of the problem, i.e., the computational resource allocation variable is independent with other variables in the objective function and constraints, and then the original problem is decomposed into a computational resource allocation subproblem and a joint channel assignment and power allocation subproblem. Since the former is a convex programming, the KKT (Karush–Kuhn–Tucker) conditions can be used to find the closed optimal solution. For the latter, which is still NP-hard, is further decomposed into two subproblems, i.e., the power allocation and the channel assignment, to optimize alternatively. Finally, two heuristic algorithms are proposed, i.e., the Co-channel Equal Power allocation algorithm (CEP) and the Enhanced CEP (ECEP) algorithm to obtain the suboptimal solutions. Numerical results are presented at last to verify the performance of the proposed algorithms.

Keywords: Mobile edge computing, computation offloading, resources allocation, power control

This work has been partly supported by the by the Chongqing Frontier and Applied Basic Research Project under Grant cstc2015jcyjA40021.

http://doi.org/10.3837/tiis.2020.07.022 **ISSN : 1976-7277**

1. Introduction

As a key technology of the fifth generation (5G), mobile edge computing (MEC) has received intensive attentions recently [1, 2]. Its core idea is to provide IT and cloud computing capabilities within the wireless access network [2]. Based on MEC, mobile devices can offload computation-intensive applications such as video transcoding [3], AR/VR image rendering [1], to the MEC server to reduce the energy consumption of the mobile devices. In addition, MEC can reduce the dependence on remote cloud for the large-scale deployed IoT applications [4]. Despite with many advantages, the computation and communication are inter-coupled in the MEC system, and thus the perceived performance of the users (such as the latency and energy consumption, etc.) depends on the joint allocation of computational and communication resources. Therefore, with the available resource and predefined user requirements, how to optimize the resource allocation to maximize the performance is critical and open problem for the MEC networks.

Currently, many researchers have discussed the resource allocation problem for the MEC networks under different task models and scenarios. From the perspective of offloading task model, these studies can be classified into two categories, i.e., the binary offloading task and the partial offloading task [1]. For binary offloading task, the mobile device can perform it locally or wholly offload to a MEC server. In such case, the focus of resource allocation is offloading decision [5-10], including whether to offload and how to offload. For partial offloading task, we can segment the task to at least two parts, and each part can be locally performed, or offloaded to the MEC server. Then in such case, the focus is the computation distribution between mobile devices and the MEC, and resources allocation among multiple users or multiple tasks [11-14, 16].

From the perspective of the system scenarios, different system models have been discussed yet, i.e., from single-user [5, 6, 10, 12, 13], multi-users [6, 9, 11, 14, 16-20], to D2D [15] and heterogeneous MEC networks [16] et al. In a single-user single-task MEC system, since the computation demand is much smaller than the computation capacity of the MEC server. Therefore, the execution delay of the task on the MEC server is generally negligible. In such circumstance, the studies focus on the radio resource allocation to improve user perceived performance defined as the weighted sum of delay and energy consumption, such as the optimization of uplink transmission rate [5], uplink transmission time [6], and transmission power [12]. For multi-users scenario, there is a resource competition among these users and how does these users access the radio and computation resource will affect their perceived performance. For the radio resource, orthogonal access schemes such as TDMA [11], OFDMA [8, 11, 16, 17], and non-orthogonal access mechanisms such as CDMA [18, 19], and NOMA [20], are the candidates. For the computation resource, depending on the number of available virtual machines (VMs) at the MEC servers, multi-users or multi-tasks can access to computational resources in serial [21, 22] or parallel [8-20, 23]. The research has also extended to the D2D network with inter-user collaboration [15] and heterogeneous MEC networks [16]. In addition, the system performance criteria used in the existing MEC networks includes offloading delay [12, 23], energy consumption [11, 12, 14, 17] and the delay-energy tradeoff [9, 10, 18, 19].

Different from the existing research, we study the joint resource allocation for the ultra-dense MEC networks with two different service demands at the mobile users. On the one hand, as the key network technology of 5G, ultra-dense networks are deployed through small cells to realize spatial multiplexing of radio resource [24], thereby achieving a 10~1000-fold increase in user rates. Now, the ultra-dense network is widely studied in the field such as industrial and medical [24]. In an ultra-dense MEC network, dense deployment of cells will bring serious co-channel interference, which makes the radio resource allocation coupled over these users and we should perform a joint channel allocation and power control [24-26]. In addition, the competition for computing resource among multiple users from multiple cells will further complicate the resource allocation problem. However, less work has done for the resource allocation of the ultra-dense MEC networks [26]. On the other hand, in 5G and future mobile networks, the supported type of services will become more abundant, including both computing service and communication, such as AR/VR, online games. However, most existing researches on MEC networks only consider a single service type, i.e., computing service [4-15, 17- 23]. Thus, the hybrid service scenario should be considered [27], and the resulted QoS requirements for multiple services will complicate the resource allocation problem [16]. Therefore, this paper studies the resource allocation problem for ultra-dense MEC networks. In which, orthogonal frequency division multiplexing access (OFDMA) is adopted by the ultra-dense cells, and the users in the system asking for computation or communication service. While the users asking for communication service have the minimum communication rate requirements, and the computing offloading users aim at minimizing the weighted sum of delay and power consumption. We model and analyze the problem of joint channel assignment, power control, and computational resource allocation in this scenario. Since the problem is NP-hard, two heuristic algorithms are proposed to obtain the suboptimal solutions. Specifically, our contributions are as follows.

1) We study the joint resource allocation problem for an OFDMA-based ultra-dense MEC network with hybrid traffic demands. In which, the communication users have minimum transmission rate constraints, and the computing users aim at minimizing the tradeoff of delay and energy. As far as we known that, only the literature [16] has studied the scenario with hybrid service demands. However, the focus of [16] is the precoding at the mobile users. In this paper, we focus on the joint channel assignment, power control and the computational resource allocation.
2) Since the formulated joint resource allocation problem is a mixed-integer non-linear programming (MINLP), which proved to be NP-hard and is difficult to solve. Thus, based on the structural features of the original problem, it is decomposed to two subproblems, i.e., computation resource allocation, and a joint subchannel assignment and power control. The former is a convex problem, thus we can obtain a closed-form solution by using the KKT conditions. For the latter, since the subproblem is still a MINLP, two heuristic algorithms to achieve sub-optimal solution are proposed, i.e., CEP and ECEP. Meanwhile, the computation complexity of the proposed algorithms are analyzed.
3) The performance of the proposed algorithms is analyzed at last. We analyze how the weight sum of offloading cost is varying with the user density, task complexity, or delay-energy consumption weight. The results explicitly verify that the ECEP outperforms CEP on the offloading cost performance.

The remainder of this paper is organized as follows. Section II introduces the system and formulates the resource allocation problem. Section III discusses and solves the formulated resource allocation problem by proposing two suboptimal algorithms, i.e., CEP and ECEP. Simulation results are presented in Section IV and we conclude the paper at last.

2. System Model

2.1 Scenario model

Network model: We consider a network composed of M cells and N mobile users (MUs), as shown in **Fig. 1**. The MUs cloud be phones or IoT based sensors. The base stations (BS) over these cells share the same spectrum and access to a common MEC server through the wired backhaul link. Both the MUs and BSs are equipped with single antenna. Herein, the orthogonal frequency division multiplexing access (OFDMA) is adopted by these BSs, i.e., the MUs in the same cell are allocated with orthogonal subchannels and the MUs associated with different BSs may work on the same subchannel and interfere against each other.

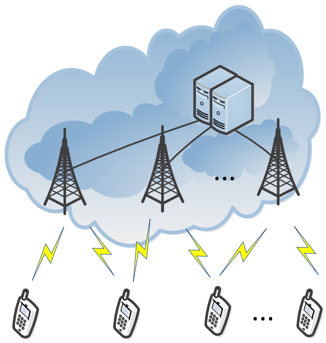


Fig. 1. Ultra-dense mobile-edge computing networks

According to the traffic demands of these MUs, the MUs can be divided into two subsets: the computation offloading MUs $\mathcal{N}_{off}$ and the communication MUs $\mathcal{N}_c$. All the MUs in $\mathcal{N}_{off}$ have computation-intensive tasks to be processed. Due to the limited computing capability or battery capacity on these MUs, their tasks must be offloaded to the MEC server to implement [8, 16]. While MUs in $\mathcal{N}_c$ just ask for network access but with the minimum transmission rate requirements. Define $\mathcal{N}=\mathcal{N}_{off}\cup\mathcal{N}_c$, then all MUs in the network compete for wireless communication resources, but only MUs in $\mathcal{N}_{off}$ compete for computational resources. In this paper, we discuss the problem of joint subchannel assignment, power control and computation resource allocation so that all MUs in $\mathcal{N}_{off}$ can minimize their offloading costs and the transmission rate requirements of the MUs in $\mathcal{N}_c$ are satisfied.

Task model: For MU $n\in\mathcal{N}_{off}$, its offloading task is characterized by $A\left(L_{in}^n, X_n, L_{out}^n\right)$ [29], where L_{in}^n and L_{out}^n denote the number of data bits for input and output of its task, respectively, and X_n denotes the computation complexity of this offloading task.

For the computing offloading MUs, we use the weighted sum of the offloading delay and energy consumption to characterize its offloading performance. However, as [19], the delay caused by task offloading transmission delay from BS to MEC server, and the results feedback from MEC server to the MUs are ignored. In addition, the energy consumption for results feedback from BS to MUs, i.e., the energy consumption for signal receiving at the MUs, is also not considered herein. Following these assumptions, we take MU n allocated with the subchannel k in the mth cell as an example to illustrate the communication, computation and offloading model.

Communication model: As aforementioned, OFDMA is adopted by the BS. Specifically, each BS divides the bandwidth B_0 [Hz] equally into K subchannels with bandwidth B [30], i.e. $B = B_0/K$, and the subchannels set is $\mathcal{K} = \{1,2,\ldots,K\}$. In order to ensure the orthogonally of the uplink transmission of MUs in the same cell, the subchannels are exclusively assigned to MUs. Define subchannel assignment variable as

$$x_n(m,k), \forall n \in \mathcal{N}, m \in \mathcal{M}, k \in \mathcal{K} \tag{1}$$

Where $x_n(m,k) = 1$ indicates that MU n is assigned the nth subchannel in the mth cell, otherwise $x_n(m,k) = 0$.

We define subchannel assignment decision as

$$\mathcal{G} = \left\{ x_n(m,k) \middle| x_n(m,k) = 1, \forall n \in \mathcal{N}, m \in \mathcal{M}, k \in \mathcal{K} \right\} \tag{2}$$

In order to ensure that all MUs' are serviced, the following constraints of the subchannel assignment must be satisfied:

$$\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} x_n(m,k) = 1, \forall n \in \mathcal{N}. \tag{3}$$

Let $p_n \in (0, p_{\max}], n \in \mathcal{N}$ denote the transmit power of MU. $p_{\max}$ is the maximal allowable transmission power for all MUs. Thus the signal to interference plus noise ratio (SINR) of the received signal at its associated BS is

$$\gamma_{n,m,k} = \frac{p_n h_{n,m,k}}{I_m^k + \sigma^2}. \tag{4}$$

Where $h_{m,n,k}$ denotes the uplink channel gain between BS m and MU n, and we assume it keeps unchanged during a resource allocation period, i.e., the channel is a quasi-static channel. And $I_m^k = \sum_{r=1, r \neq m}^{M} \sum_{n=1}^{N} x_n(r,k) p_n h_{n,r,k}^m$ denotes inter-cell interference from co-channel users in other cells. The power of additive white Gaussian noise (AWGN) is denoted by σ^2. Therefore, the transmission rate of MU n is

$$R_n = B \log_2(1 + \gamma_{m,n,k}) \tag{5}$$

In addition, all MUs in $\mathcal{N}_c$ have the same transmission rate constraint $R_{\min}$, i.e.,

$$R_n \geq R_{\min}, \forall n \in \mathcal{N}_c \tag{6}$$

As mentioned above, the number of offloading task input bits for MU n is L_{in}^n, then the offloading transmission delay for MU n is

$$T_n^{off} = \sum_{m \in \mathcal{M}} \frac{x_{n,m} L_{in}^n}{R_n}, \forall n \in \mathcal{N}_{off}. \tag{7}$$

Where $x_{n,m} = \sum_{k \in \mathcal{K}} x_n(m,k), \forall n \in \mathcal{N}_{off}, m \in \mathcal{M}$. Given the transmit power $p_n, n \in \mathcal{N}_{off}$, the energy consumption of the task offloading for MU n is

$$E_n = T_n^{off} p_n. \tag{8}$$

Computation model: We assume that the MEC server allocates the orthogonal computing resource over multiple accessed MUs by virtualization [1]. Define F_s as the maximum CPU frequency of the MEC server, and f_n denotes the CPU frequency allocated to MU $n \in \mathcal{N}_{off}$, which satisfies $f_n > 0, \forall n \in \mathcal{N}_{off}$, $f_n = 0, \forall n \notin \mathcal{N}_{off}$. Then, the computational resource allocation decision is denoted by $\mathcal{F} = \{ f_n \mid n \in \mathcal{N}_{off} \}$, thus we have

$$\sum_{n \in \mathcal{N}_{off}} f_n \le F_s \tag{9}$$

Given the computational resource f_n, the execution delay of the MU n's task is

$$T_n^{exe} = \frac{L_{in}^n X_n}{f_n}, \forall n \in \mathcal{N}_{off} \tag{10}$$

Offloading cost: Given the subchannel assignment $\mathcal{G}$, transmit power $\mathcal{P} = \{ p_n, n \in \mathcal{N} \}$ and the computational resource f_n, the total offloading delay of MU n is

$$T_n = T_n^{off} + T_n^{exe} = \sum_{m \in \mathcal{M}} \frac{x_{n,m} L_{in}^n}{R_n} + \frac{L_{in}^n X_n}{f_n}, \forall n \in \mathcal{N}_{off} \tag{11}$$

Therefore, the offloading cost for offloading MUs which is characterized by the weighted sum of offloading delay and energy consumption, is

$$J_n = \gamma_t T_n + \gamma_e E_n, \forall n \in \mathcal{N}_{off} . \tag{12}$$

Where $\gamma_t, \gamma_e \in [0,1]$ and $\gamma_t + \gamma_e = 1, \forall n \in \mathcal{N}_{off}$, γ_t and γ_e together characterize the MUs' preference over offloading delay and energy consumption, respectively.

2.2 Joint resource allocation problem

As mentioned earlier, for our considered system, we discuss the joint subchannel assignment, power control and computational resource allocation with the objective of minimizing the weight sum of offloading cost for all offloading MUs,

$$J(\mathcal{G}, \mathcal{P}, \mathcal{F}) = \sum_{n \in \mathcal{N}_{off}} a_n J_n \tag{13}$$

In which, J_n is defined in (12) and $a_n > 0, n \in \mathcal{N}_{off}$ denotes system's preference for MUs. To sum up, we have the following constrained optimization problem,

$$\begin{aligned}
&\min_{\mathcal{G},\mathcal{P},\mathcal{F}} J(\mathcal{G}, \mathcal{P}, \mathcal{F}) \\
&s.t.\ C1: x_n(m,k) \in \{0,1\}, \forall n \in \mathcal{N}, m \in \mathcal{M}, k \in \mathcal{K}. \\
&\quad C2: \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} x_n(m,k) = 1, \forall n \in \mathcal{N}. \\
&\quad C3: \sum_{n \in \mathcal{N}} x_n(m,k) \le 1, \forall m \in \mathcal{M}, k \in \mathcal{K}. \\
&\quad C4: 0 < p_n \le p_{\max}, n \in \mathcal{N}. \\
&\quad C5: R_n \ge R_{\min}, \forall n \in \mathcal{N}_c. \\
&\quad C6: f_n > 0, \forall n \in \mathcal{N}_{off}. \\
&\quad C7: \sum_{n \in \mathcal{N}_{off}} f_n \le F.
\end{aligned} \qquad OR$$

In which, C1 and C2 together indicates that each MU should be and at most be assigned with one subchannel. C3 means that each subchannel is exclusively assigned to only one MU. C4 defines the transmission power constraint for the MUs. C5 defines the minimum transmission rate constraint for the communication MUs. C6 promises that each offloading MU will be allocated the computational resource. C7 restricts the available computational resources. Obviously, problem OR is a mixed integer non-linear programming (MINLP) [28], thus it is NP-hard. Therefore, the suboptimal polynomial complexity algorithms are proposed below.

3. Joint sub-channel assignment and resource allocation algorithm

Since problem OR involves optimization of both integer variables (subchannel assignment) and continuous variables (power control and computational resource allocation). In addition, the objective function and constraint C5 are both non-convex. Therefore, it is NP-hard and difficult to solve it directly. In spite of this, we note that, for problem OR, the computational resource allocation is independent with the power control and channel allocation. With this in mind, the original problem OR is decomposed into two subproblems, i.e., the computation resource allocation, and the joint subchannel assignment and power control. The first one is a convex optimization problem, which can be solved by interior point method [28]. However, the second subproblem is still NP-hard. In order to solve this problem and propose low complexity algorithm, it is further decomposed into two subproblems, i.e., channel assignment problem and the power control problem. Finally, an iteration based optimization is proposed to completely solve the joint optimization problem.

3.1 Problem decomposition

Firstly, for problem OR, we have the following proposition.

Proposition 1: Problem OR can be decomposed into two independent subproblems CP and JP as presented below. In which, the former only involves computational resource allocation and the latter is the joint channel assignment and power control.

Proof: For the objective function of OR, we have

$$\begin{aligned} J(\mathcal{G},\mathcal{P},\mathcal{F}) &= \sum_{n\in\mathcal{N}_{off}} a_n\left(\gamma_t T_n + \gamma_e E_n\right) \\ &= \sum_{n\in\mathcal{N}_{off}} a_n\left(\gamma_t\left(\sum_{m\in\mathcal{M}}\frac{x_{n,m}L_{in}^n}{R_n}+\frac{L_{in}^n X_n}{f_n}\right)+\gamma_e p_n\sum_{m\in\mathcal{M}}\frac{x_{n,m}L_{in}^n}{R_n}\right) \\ &= \sum_{n\in\mathcal{N}_{off}}\frac{a_n\gamma_t L_{in}^n X_n}{f_n}+\sum_{n\in\mathcal{N}_{off}} a_n\left(\gamma_t\left(\sum_{m\in\mathcal{M}}\frac{x_{n,m}L_{in}^n}{R_n}\right)+\gamma_e p_n\sum_{m\in\mathcal{M}}\frac{x_{n,m}L_{in}^n}{R_n}\right) \end{aligned}$$

Wherein, the first part involves computation resource allocation, and the second part is about joint subchannel assignment and power control. In addition, for problem OR, C6 and C7 are the constraints for the computational resource allocation, and C1 to C5 are the constraints for the joint subchannel assignment and power control. Therefore, the problem can be decomposed into the following two subproblems, i.e., the computation resource allocation problem CP, and the joint channel allocation and power control problem JP,

$$\begin{aligned} &\min_{\mathcal{F}} J_f\left(f_1,\ldots,f_{|\mathcal{N}_{off}|}\right)=\sum_{n\in\mathcal{N}_{off}}\frac{a_n\gamma_t L_{in}^n X_n}{f_n} \\ &s.t.\quad C6-C7. \end{aligned} \qquad CP$$

$$\begin{aligned} &\min_{\mathcal{G},\mathcal{P}} J_{x,p}=\sum_{n\in\mathcal{N}_{off}} a_n\left(\gamma_t\left(\sum_{m\in\mathcal{M}}\frac{x_{n,m}L_{in}^n}{R_n}\right)+\gamma_e p_n\sum_{m\in\mathcal{M}}\frac{x_{n,m}L_{in}^n}{R_n}\right) \\ &s.t.\quad C1-C5. \end{aligned} \qquad JP$$

Therefore, we have the conclusion.

Based on Proposition 1, the computational resource allocation subproblem and the joint channel assignment and power control subproblem will be separately discussed below.

3.2. Computational resource allocation

Lemma 1: The objective function of CP is a convex function.

Proof: In order to get this conclusion, we have to prove that the Hession matrix of the objective function is a positive definite matrix [28]. Since the Hession matrix of the objective function is

$$\mathbf{H}=\begin{bmatrix} \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial^2 f_1} & \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_1 \partial f_2} \cdots & \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_1 \partial f_{|\mathcal{N}_{off}|}} \\ \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_2 \partial f_1} & \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial^2 f_2} \cdots & \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_2 \partial f_{|\mathcal{N}_{off}|}} \\ \vdots & \vdots \quad \ddots & \vdots \\ \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_{|\mathcal{N}_{off}|} \partial f_1} & \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_{|\mathcal{N}_{off}|} \partial f_2} \cdots & \frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial^2 f_{|\mathcal{N}_{off}|}} \end{bmatrix}$$

Wherein,

$$\frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial^2 f_n}=\frac{\partial^2 \frac{a_n\gamma_t L_{in}^n X_n}{f_n}}{\partial^2 f_n}=2a_n\gamma_t L_{in}^n X_n f_n^{-3}, \forall n\in\mathcal{N}_{off}, \tag{14}$$

$$\frac{\partial^2 J_f\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)}{\partial f_n \partial f_m}=0, \forall n,m\in\mathcal{N}_{off}, n\neq m.$$

Since $a_n, \gamma_t, L_{in}^n, X_n$ and f_n are all positive, thus the Hession matrix is a positive definite matrix. That is, the objective function of CP is a convex function [28]. □

Following the Lemma 1, we can conclude that CP is a convex optimization problem, thus its optimal solution can be characterized by the following proposition.

Proposition 2: The optimal solution of CP is $f_n^*=\frac{\sqrt{a_n\gamma_t L_{in}^n X_n}}{\sum_{n\in\mathcal{N}_{off}}\sqrt{a_n\gamma_t L_{in}^n X_n}}F, \forall n\in\mathcal{N}_{off}$.

Proof: Since CP is a convex optimization problem and the Slater conditions hold [28]. Therefore, the KKT conditions are both the sufficient and necessary conditions for the optimal solution of the problem CP [25]. Defining the Lagrange function of the problem CP as

$$L(\mathbf{f},\lambda,\mu)=J_f(\mathbf{f})+\sum_{n\in\mathcal{N}_{off}}\lambda_n g_n(\mathbf{f})+\mu h(\mathbf{f}).$$

Wherein, $\mathbf{f}=\left(f_1,...,f_{|\mathcal{N}_{off}|}\right)$, $g_n(\mathbf{f})=-f_n, \forall n\in\mathcal{N}_{off}$, $h(\mathbf{f})=\sum_{n\in\mathcal{N}_{off}} f_n-F$. In which, $g_n(\mathbf{f})<0$ and $h(\mathbf{f})<0$ are derived from the inequality constraints C6 and C7, respectively. We let λ_n and μ denote the Lagrange multipliers for C6 and C7 respectively, and $\lambda_n,\mu\geq 0$. Therefore, the optimal solution must satisfy the following KKT conditions,

$$\left.\frac{\partial L(\mathbf{f},\lambda_n,\mu)}{\partial f_n}\right|_{f_n=f_n^*}=0, \qquad \text{(a)}$$

$$\lambda_n g_n\left(\mathbf{f}^*\right)=0, \qquad \text{(b)}$$

$$\mu h\left(\mathbf{f}^*\right)=0, \qquad \text{(c)}$$

$$g_n\left(\mathbf{f}^*\right)<0, \qquad \text{(d)}$$

$$h\left(\mathbf{f}^*\right)\le 0. \qquad \text{(e)}$$

Where (a) is the necessary condition for the Lagrange function to take extreme values, (b) and (c) are the complementary relaxation conditions, and (d) and (e) are the inequality constraints. From (b) and (d), we know that $\lambda_n=0$. Furthermore, based on the Lagrange function of the problem CP, we have

$$\begin{aligned}\frac{\partial L(f,\lambda_n,\mu)}{\partial f_n}&=-\frac{a_n\gamma_t L_{in}^n X_n}{f_n^2}-\lambda_n+\mu\\&=-\frac{a_n\gamma_t L_{in}^n X_n}{f_n^2}+\mu.\end{aligned}$$

Then based on equation (a), the optimal solution for MU n is

$$f_n^*=\sqrt{\frac{a_n\gamma_t L_{in}^n X_n}{\mu}}.$$

In which, μ is determined by the complementary relaxation condition (c), and the optimal solution must satisfy the following condition

$$\sum_{n\in\mathcal{N}_{off}} f_n^*=F.$$

Therefore, we have

$$\mu=\left(\sum_{n\in\mathcal{N}_{off}}\sqrt{a_n\gamma_t L_{in}^n X_n}\Big/F\right)^2.$$

To sum up, we conclude that the solution of CP is $f_n^*=\dfrac{\sqrt{a_n\gamma_t L_{in}^n X_n}}{\sum_{n\in\mathcal{N}_{off}}\sqrt{a_n\gamma_t L_{in}^n X_n}}F,\forall n\in\mathcal{N}_{off}$.

3.3. CEP algorithm

In this section, we discuss and solve the joint subchannel assignment and power control subproblem. As mentioned earlier, this problem is still NP-hard. This comes from the fact that, 1) the subchannel assignment is a combinational problem, and 2) both the objective function and the constraint C5 are non-convex. Therefore, based on the alternation optimization, a low complexity heuristic algorithm is proposed herein. Specifically, the original problem JP is further decomposed to two subproblems, i.e., the subchannel assignment, and the power allocation. At first, given the subchannel assignment, the transmission power of all MUs are optimized. Then, based on the result of power allocation, the subchannel assignments are updated. Finally, these two steps are alternatively performed until the termination conditions are met. The details are summarized as follows.

Step 1 (Initialization): Initialize a feasible subchannel assignment;

Step 2 (Power Allocation): Given the subchannel assignment, we optimize the transmit power for all MUs. Specifically, we consider that the computing offloading MUs who accessed to the same subchannel have equal transmit power, and this is the same for the communication MUs. Then the equal transmit power for both offloading MUs and communication MUs are optimized by searching to minimize the offloading costs for offloading MUs under the transmission power constraints for the communication MUs.

Step 3 (subchannel assignment): Given power allocation, subchannel assignment is updated. Specifically, each MU cyclically accesses to other subchannels and retains the subchannel assignment has the minimum system offloading cost.

Step 4: Repeat the above steps until the terminal conditions are satisfied.

3.3.1 Initialization

Since all MUs in $\mathcal{N}_c$ have the transmission rate constraints, in order to obtain an initial feasible solution, it is necessary to initialize a feasible subchannel allocation and then perform power adjustment. Since the offloading cost for MU *n* in $\mathcal{N}_{off}$ is denoted by

$$J_n = a_n \frac{L_{in}^n}{R_n}\left(\gamma_t + \gamma_e p_n\right) \tag{15}$$

Given the power allocation, $L_{in}^n\left(\gamma_t + \gamma_e p_n\right)$ is constant, thus the offloading cost for MU *n* is inversely proportional to R_n. In addition, since its uplink transmission rate is

$$R_n = B\log_2\left(1 + \frac{h_{n,m,k}\, p_n}{I_n + N_0}\right) \tag{16}$$

From (15) and (16) and the given power allocation, we know that the largest $h_{n,m,k} / (I_n + N_0)$ will bring the least offloading cost. Therefore, following criteria is adopted in the subchannel assignment,

$$x_n\left(m,k\right) = 1\Big|_{(m,k)=\arg\max EIR_{n,m,k}} \forall n. \tag{17}$$

Wherein,

$$\mathrm{EIR}_{n,m,k} = \frac{h_{n,m,k}}{\sum_{r\in\mathcal{M}, r\neq m} h_{n,m,k}^r}, \forall m \in \mathcal{M}, k \in \mathcal{K} \tag{18}$$

And $\mathrm{EIR}_{n,m,k}$ denotes the effective interference ratio of MU *n* on subchannel *k* over cell *m*.

As mentioned earlier, we first assign the subchannels to the communication MUs so that their transmission rates are satisfied, and then consider the computation offloading MUs. We summarize the details of the subchannel assignment initialization in the **Algorithm 1** presented below.

Algorithm 1. Subchannel assignment initialization

1: Initialization: $p_n = p_{\max}, r_n = 0, \forall n \in \mathcal{N}_c$.

Phase 1: Subchannel assignment initialization for communication MUs

2: Initialize $\text{channelAlloc} = 1$, $\text{EC} = 0$, $\text{CoUsr} = 0$, $\text{UoCh} = 0$, compute $\text{EIR}(n,m,k)$, $\forall n \in \mathcal{N}_c, m \in \mathcal{M}, k \in \mathcal{K}$ by (18).

3: For $\forall n \in \mathcal{N}_c$, implements the following steps:

3.1：Compute the allocated subchannel (m,k) of MU *n* by (17).

3.2：If $\text{EC}(m,k) < \text{EIR}(n,m,k)$, implements the following steps.

3.2.1：If $\text{UoCh}(m,k) \neq 0$, update $\text{CoUsr}(1, \text{UoCh}(m,k)) = 0$;

3.2.2：Update the state matrix $\text{CoUsr}(:,n) = (m;k)$.

4: If $\min(\text{CoUsr}) > 0$, update $\text{channelAlloc} = 0$.

5: If $\text{channelAlloc} = 1$, jump to step 3. Otherwise, go to step 6.

6: For $\forall n \in \mathcal{N}_c$, compute R_n by (5).

7: For $\forall n \in \mathcal{N}_c$, if $R_n < R_{\min}$, , the sub-channel assignment successes; otherwise, fails.

Phase 2: Subchannel assignment initialization for computation offloading MUs

8: Similar with phase 1.

9: Output $(\mathcal{G})$.

Where $\mathbf{p} = \left(p_1 \cdots p_{|\mathcal{N}_c|}\right)$ and $\mathbf{r} = \left(r_1 \cdots r_{|\mathcal{N}_c|}\right)$ record the power allocation for communication MUs and computation offloading MUs, respectively. channelAlloc=1 indicates that the MU isn't assigned any subchannel, otherwise MU has been assigned a subchannel. The EC keeps the EIR value of the currently accessed MU for each subchannel, CoUsr is a $2 \times |\mathcal{N}_c|$ matrix who keeps the MUs' assigned subchannels, while the first row records the index of the base station, and the second row records the subchannel. UoCh records the accessed MUs of each subchannel.

3.3.2 Power allocation

Based on the subchannel assignment, then the power allocation on each subchannel can be independently discussed. Thus, for subchannel $k \in \mathcal{K}$, the power allocation problem is

$$\begin{aligned} &\min_{\mathcal{P}} J(\mathcal{P}) \\ &s.t.\ \ 0 < p_n \leq p_{\max}, n \in \mathcal{N}^k. \\ &\qquad R_n \geq R_{\min}, \forall n \in \mathcal{N}_c^k. \end{aligned} \tag{19}$$

In which, $\mathcal{N}^k = \mathcal{N}_{off}^k \cup \mathcal{N}_c^k$, $\mathcal{N}_{off}^k$ and $\mathcal{N}_c^k$ denote the computation offloading MUs and communication MUs on subchannel *k*, respectively. Obviously, (19) is a typical interference channel (IC) power control problem and it is non-convex. Therefore, we propose a suboptimal power control mechanism for this problem, i.e., the MUs with the same service type transmit with equal power. Following that, (19) is simplified as follows,

$$\begin{aligned} & J^k = \min_{\mathcal{P}^k} J^k\left(\mathcal{P}^k\right) \\ & s.t.\ \ 0 < p_{off} \le p_{\max}. \\ & \qquad 0 < p_c \le p_{\max}. \\ & \qquad p_n = p_{off}, \forall n \in \mathcal{N}_{off}^k. \\ & \qquad p_n = p_c, \forall n \in \mathcal{N}_c^k. \\ & \qquad R_n \ge R_{\min}, \forall n \in \mathcal{N}_c^k. \end{aligned} \tag{20}$$

Where p_{off} and p_c are the transmit power of computation offloading MUs and communication MUs in subchannel *k*, respectively. Now, the power allocation problem (19) which has $\left|\mathcal{N}^k\right|$ variables is simplified to (20) with only two decision variables, i.e., p_{off} and p_c. Then we use power iteration to minimize the weighted sum of offloading cost for all offloading MUs in subchannel *k*. Specifically, given the step λ and at each iteration, the offloading MUs adjust the transmit power by $p_{off} = p_{off} + \lambda$, then we calculate the weighted sum of offloading cost and the result is compared with the previous iteration. Finally, we keep the power allocation with less weighted sum of offloading cost. In the following, we first present two useful conclusions and then give the details of the algorithm.

Lemma 2: Given the transmission power p_{off} for the computation offloading MUs, the minimum weighted sum of offloading cost for offloading MUs can be achieved at the point that the co-channel communication MUs transmit with the minimum power but their transmission rates are satisfied.

Proof: From (4) and (5), and given the transmission power of the computation offloading MUs, we know that, less transmission power of the co-channel communication MUs, then the smaller co-channel interference they will bring to the offloading MUs, and thus the offloading MUs can obtain larger transmission rate. From (12), we know that in turn, it will decrease the weighted sum of offloading cost for offloading MUs. □

Proposition 3: Given the transmit power p_{off} for the computation offloading MUs, the lower bound of the transmission power for co-channel communication MUs to satisfy their transmission rate constraints is

$$p_c = \max\left\{p_c\left(i\right), \forall i \in \mathcal{N}_c^k\right\}. \tag{21}$$

Where $p_c\left(i\right) = \alpha p_{off} + \beta$, $\alpha = \gamma \sum_{s \in \mathcal{N}_{off}} h_{s,d,k}^j \Big/ \left(h_{i,j,k} - \gamma \sum_{v \in \mathcal{N}_c} p_{off} h_{v,r,k}^j\right)$, $\beta = \gamma N_0$, and $\gamma = 2^{R_{\min}/B} - 1, \forall i \in \mathcal{N}_c^k$.

Proof: Define the transmission power of communication MU *i* as $p_c\left(i\right), \forall i \in \mathcal{N}_c^k$. To ensure the feasibility of the problem, the rate constraint of communication MUs must be satisfied, that is,

$$B\log_2\left(\frac{p_c\left(i\right)h_{i,j,k}}{\sum_{s\in\mathcal{N}_{off}^k} p_c\left(i\right)h_{s,d,k}^j + \sum\limits_{v\in\mathcal{N}_c^k} p_v h_{v,r,k}^j + N_0} + 1\right) \ge R_{\min}, \forall i \in \mathcal{N}_c^k. \tag{22}$$

Let the transmission power of computation offloading MUs be p_{off}, then the minimum transmission power will be obtained if the equation (22) takes equal sign. Following that, after some simplifications, we have $p_c\left(i\right)$ as follows,

$$p_c\left(i\right) = \alpha p_{off} + \beta, \forall i \in \mathcal{N}_c^k. \tag{23}$$

The SINR of communication MUs is

$$S_c(i)=\frac{p_c(i)h_{i,j,k}}{\sum_{s\in\mathcal{N}_{off}^k} p_c(i)h_{s,d,k}^j+\sum_{v\in\mathcal{N}_c^k} p_{off}h_{v,\mathrm{r},k}^j+N_0},\forall i\in\mathcal{N}_c^k$$

And its first derivative with respect to $p_c(i)$ is

$$\frac{\partial S_c(i)}{\partial p_c(i)}=\frac{h_{i,j,k}\left(\sum_{v\in\mathcal{N}_c^k} p_{off}h_{v,\mathrm{r},k}^j+N_0\right)}{\left(\sum_{s\in\mathcal{N}_{off}^k} p_c(i)h_{s,d,k}^j+\sum_{v\in\mathcal{N}_c^k} p_{off}h_{v,\mathrm{r},k}^j+N_0\right)^2}$$

Since $p_c(i),p_{off}>0$, the first derivative of $S_c(i)$ is positive, that is, $S_c(i)$ is monotonously increasing for $p_c(i)>0$. In addition, the communication rate increases as the SINR increase, and the communication rate increases as the power increase. Therefore, when p_c takes the maximum power of co-channel communication MUs, all communication MUs exactly meet the minimum rate requirement. To sum up, we have the conclusion. □

With **Lemma 2** and **Proposition 3**, and given the transmission power p_{off} of computation offloading MUs, the minimum weighted sum of offloading cost is achieved when the communication MUs set their transmit power by (21). Then the offloading MUs adjust the transmission power by $p_{off}=p_{off}+\lambda$, and the power which has the smallest weighted sum of offloading cost is reserved. The terminal condition is 1) $p_{off}>\mathrm{p}_{\max}$; or 2) $p_c>p_{\max}$. Condition 1) indicates that the power iteration completes, and condition 2) indicates that the rate constraint can not be satisfied. The algorithm is ended when all subchannels have been traversed. In summary, we have the corresponding **Algorithm 2**.

Algorithm 2: equal power allocation for co-channel MUs with the same service type

1: Set step size λ, $J_{\mathrm{now}}=0$, and initialize $\mathcal{G}$, $\mathcal{P}$, and J^{sub}.

2: For $\forall k\in\mathcal{K}$, implements the following steps:

2.1: Initialize $p_{off}=\lambda$, and compute the MUs $\mathcal{N}^k$ assigned the k subchannel.

2.2: If $p_{off}\le\mathrm{p}_{\max}$, compute by (21); otherwise, jump to step 3.

2.3: If $p_c\le p_{\max}$, implement the following steps; otherwise, jump to step 3.

2.3.1: compute the user utility sum J_{now} of $\mathcal{N}_k$ by (13).

2.3.2: If $J_{\mathrm{now}}<J_k$, update $J_k^{sub}=J_{\mathrm{now}}$, $\mathcal{P}_n=p_{off},\forall n\in\mathcal{N}_{off}^k$, $\mathcal{P}_n=p_c,\forall n\in\mathcal{N}_c^k$.

2.4: Update $p_{off}=p_{off}+\lambda$.

3: Compute the system utility $J=\sum_{k=1}^K J_k^{sub}$.

4: Output $(\mathcal{G},\mathcal{P},J)$.

Where the matrix J^{sub} stores the weighted sum of offloading cost for each subchannel. When this algorithm is performed for the first time, the output of Algorithm 1 is used as the input of this algorithm. The user power is initialized to $p_n=p_{\max},n\in\mathcal{N}_c^k$, $p_n=\lambda,n\in\mathcal{N}_{off}^k$.

3.3.3 Subchannel assignment

Given the power allocation, the subchannel assignment problem becomes

$$J(\mathcal{G})=a_n\sum_{n\in\mathcal{N}_{off}}\frac{L_{in}^n}{R_n}(\gamma_t+\gamma_e p_n)$$
$$s.t.\ \ C1-C3、C5. \tag{24}$$

Since this is a combinatorial optimization problem, the size of the solution space is 2^n and $n = M\times K\times N$. Obviously, it is too complexity to directly search the optimal solution. Therefore, a low-complexity heuristic algorithm is proposed to obtain a suboptimal solution, that is, only one subchannel allocation is adjusted per cycle. In the system, subchannels can be divided into occupied subchannels and idle subchannels. If the occupied subchannels are involved, the MU needs to exchange channels with the MU who is occupying this channel; otherwise, the MU switch to the subchannel directly, and the channel can be switched. The adjustment is presented below, i.e., Exchange and switch channels operation.

Exchange and switch operation

exchange$\left(n,(m,k)\right)$

For $i\in\mathcal{M}, j\in\mathcal{K}$

  $\mathcal{G}\leftarrow\mathcal{G}\backslash\left\{x_n(i,j)\right\}$

End

For $r\in\mathcal{N}$

  $\mathcal{G}\leftarrow\mathcal{G}\backslash\left\{x_r(m,k)\right\}$

End

Set $\mathcal{G}\leftarrow\mathcal{G}\cup\left\{x_n(m,k),x_r(i,j)\right\}$

Output: $\mathcal{G}$

switch$\left(n,(m,k)\right)$

For $i\in\mathcal{M}, j\in\mathcal{K}$

  $\mathcal{G}\leftarrow\mathcal{G}\backslash\left\{x_n(i,j)\right\}$

End

Set $\mathcal{G}\leftarrow\mathcal{G}\cup\left\{x_n(m,k)\right\}$

Output: $\mathcal{G}$

Based on the exchange and switch operation, the subchannel assignment process is summarized as follows: The MU cyclically accesses the other subchannels, according to (13), the weighted sum of offloading cost is calculated and then keep the subchannel assignment has the minimum weighted sum of offloading cost. The procedure will continue until all MUs have been traversed. The **Algorithm 3** presented below summarizes the details.

Algorithm 3: Subchannel assignment

1: Initialize $J_{\text{now}} = 0$, $\mathcal{G}$, $\mathcal{P}$, and J.

2: For $\forall n \in \mathcal{N}$, implements the following steps:

2.1: For $\forall m \in \mathcal{M}, \forall k \in \mathcal{K}$, if (m,k) is not assigned, $\text{switch}(n,(m,k))$; otherwise, $\text{exchange}(n,(m,k))$, and $\mathcal{G}_0$ records the temporary subchannel assignment.

2.2: Compute the rate of communication MUs by (5), if C5 is satisfied:

2.2.1: Compute the system utility J_{now} by (13).

2.2.2: If $J > J_{\text{now}}$, update $\mathcal{G} = \mathcal{G}_0$, $J = J_{\text{now}}$.

3: Output $(\mathcal{G}, \mathcal{P}, J)$.

Herein, the input of Algorithm 3 is initialized by the output of Algorithm 2. In addition, the initial value of J is based on (13).

Algorithm 4 summarizes the joint subchannel assignment and power allocation. Since the co-channel MUs who asking for the same service have equal transmission power, thus the Algorithm 4 is named the co-channel Equal Power Allocation (CEP), and the details are illustrated as below.

Algorithm 4: CEP

1: Initialize $t = 0$;

2: Perform Algorithm 1;

3: Perform Algorithm 2;

4: Perform Algorithm 3;

5: Update $t = t + 1$;

6: If $t \le T$, jump to step3; otherwise, jump to step 7.

7: Output $(\mathcal{G}, \mathcal{P}, \mathcal{F}, J(\mathcal{G}, \mathcal{P}, \mathcal{F}))$.

The complexity of algorithm CEP is characterized by the following proposition.

Proposition 4: The computational complexity of CEP is upper bounded by

$$\mathcal{O}\left(KMN + |\mathcal{N}_c|\, p_{\max}/\lambda\right).$$

Proof: The computational complexity of CEP is mainly depended on the power allocation and subchannel assignment. As algorithm 2, the computing intensive operations are power iteration and sort operation of communication MUs, which has the complexity of $\mathcal{O}(p_{\max}/\lambda)$ and $\mathcal{O}(|\mathcal{N}_c|)$, respectively. Therefore, the complexity of the power allocation is upper bounded by $\mathcal{O}(|\mathcal{N}_c|\, p_{\max}/\lambda)$. For Algorithm 3, the computational complexity mainly comes from the subchannel adjustment with the complexity of $\mathcal{O}(NMK)$. To sum up, the computational complexity of the algorithm CEP is $\mathcal{O}(KMN + |\mathcal{N}_c|\, p_{\max}/\lambda)$.

3.4 Enhanced CEP algorithm

One may note that, CEP algorithm ignores the differences among the MUs and equal transmission power is used by the MUs who shared the same channel over different cells. To improve the performance of the CEP, we propose the Enhanced CEP (ECEP). The core idea of the ECEP is that, based on the CEP, we further performs power iteration for each MU and keeps the power allocation with the minimum weighted sum offloading cost as the final solution. Specifically, given a step λ, the power iteration for each MU is ended if one of the following conditions is satisfied, 1) the system weighted sum offloading cost of the current iteration is larger than the previous iteration; 2) the constraint C4 is violated; 3) if all MUs have been traversed. The **Algorithm 5** presented below summarizes the details.

Algorithm 5: ECEP

1: Initialize the step size λ, J, $J_{\text{now}} = 0$, $P^{tmp} = \mathcal{P}$ and $\mathcal{G}$;

2: For $\forall n \in \mathcal{N}$, implements the following steps:

2.1: For $\forall d \in \mathcal{D}$, implements the following steps:

2.1.1: If $0 < P_n^{tmp} \le p_{\max}$, Go to the next step; otherwise, jump to step 3;

2.1.2: Compute the rate of communication MUs by (5), if C5 is satisfied, go to the next step; otherwise, jump to step 3.

2.1.3: Compute the system utility J_{now} by (13).

2.1.4: If $J_{\text{now}} < J$, update $P_n = P_n^{tmp}$, $J = J_{\text{now}}$, $P_n^{tmp} = P_n^{tmp} + d\lambda$;

3: Output $(\mathcal{G}, \mathcal{P}, J)$.

In which, P^{tmp} is used to temporarily record the power allocation result. $\mathcal{P}$ and $\mathcal{G}$ record the results of the power allocation and subchannel assignment, respectively. The initial point of this algorithm is the output of Algorithm 2, and which promises the feasibility of the algorithm at the beginning. In addition, $\mathcal{D}$ indicates the iteration direction. To sum up, we have the ECEP algorithm presented below, i.e., the **Algorithm 6**.

Algorithm 6: ECEP

1: Initialize $t = 0$;

2: Perform Algorithm 1;

3: Perform Algorithm 5;

4: Perform Algorithm 3;

5: Update $t = t + 1$;

6: If $t \le T$, jump to step 4; otherwise, jump to step 8;

7: Output $\left(\mathcal{G}, \mathcal{P}, \mathcal{F}, J\left(\mathcal{G}, \mathcal{P}, \mathcal{F}\right)\right)$.

Similarly, we have the complexity conclusion for algorithm ECEP as follows.

Proposition 5: The computational complexity of the ECEP is upper bounded by

$$\mathcal{O}\left(NKM + Np_{\max}/\lambda\right).$$

Proof: Due to space limitation, the details of the proof are omitted herein. However, this conclusion can be easily proved by the same way as that used for the Proposition 4.

4. Numerical results

In this section, we evaluate the performance of the proposed algorithms, i.e., the CEP and ECEP by numerical simulations and the results are presented below. The scenario parameters used in the simulation are shown in **Table 1** [8, 11], that is, we have 7 cells, the coverage of each cell is 50m and the MUs are uniformly distributed over multiple cells. The path-loss is characterized by $L[\text{dB}] = 140.7 + 36.7\log_{10} d[\text{km}]$ and the small-scale fading is Rayleigh fading.

Without any further statements, the results shown below are averaged over 1000 independent simulations.

Table 1. Simulation parameters [8,11]

Parameters	Value
Number of cells	7
Coverage of the cell	50m
Number of subchannels	10
Subchannel bandwidth	5KHz
Noise Power	-110dBm/Hz
Maximal transmit power at the MUs	200mW
Average input data size of offloading MUs' tasks	2000Kbits
Average complexity of offloading MUs' tasks	1000Megacycles/bit
Available CPU frequency at MEC server	20GHz
Rate requirements for communication MUs	100Kbps

At first, we study the weighted sum offloading cost and the average per MU offloading cost of CEP and ECEP against different number of computation offloading MUs, and the results are shown in **Fig. 2** and **Fig. 3**, respectively. In which, the number of offloading MUs is varying from 7 to 56. From **Fig. 2**, we note that with the increase of the number of computation offloading MUs, as expected, the total offloading cost of both CEP and ECEP increases. We can explain this phenomenon as follows. On the one hand, the system total offloading cost is defined as the weighted sum of delay and energy consumption for all computation offloading MUs. Thus, it will increase as the number of MUs increase and obviously, this is consistent with the intuition. On the other hand, the more MUs in the system, the competition between MUs becomes more serious, and which results the increase of average per MU offloading cost, as shown in **Fig. 3**. In addition, the result in **Fig. 2** confirms our analysis before that the offloading performance of ECEP is better than CEP.

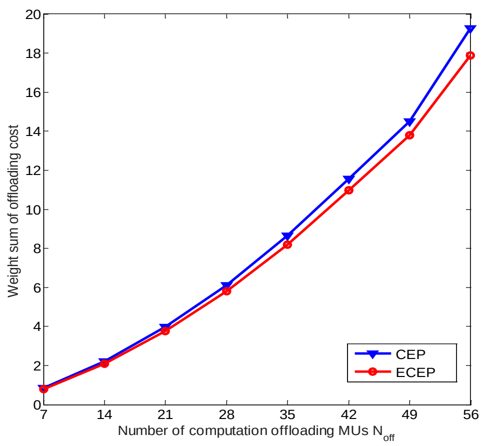


Fig. 2. Sum offloading cost versus the numbers of computing offloading MUs.

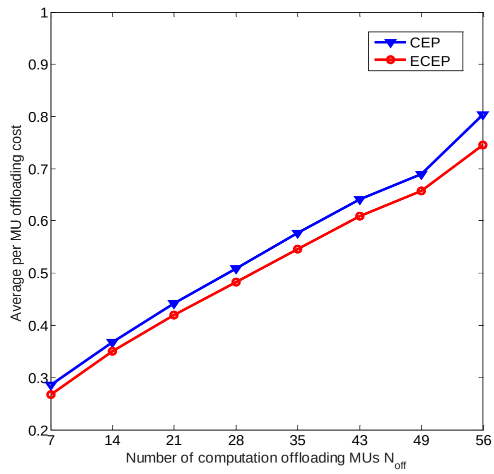


Fig. 3. Average per MU offloading cost versus the numbers of computing offloading MUs.

From **Fig. 3**, we note that the average per MU offloading cost increases linearly with the increase of the number of computation offloading MUs. This due to the fact that, as more computing MUs present in the system, the competition among MUs becomes more serious for limited computing resource at the MEC server and the larger average co-channel interference for each MU, which in turn increases the transmission delay and energy consumption, thus the average per MU offloading cost increases. We also find that the gap of the offloading cost between these two algorithms increases as the number of offloading MUs becomes larger. This performance gain comes from the fact that ECEP has more degree of freedoms in power adjustment.

Then, we analyze how does the system weighted sum of offloading cost is varying with the task complexity of the offloading MUs, and the result is presented in **Fig. 4**. Herein, the parameters used in the simulation are the same at that used in **Fig. 2**, except that right now, we fix the number of communication MUs as 6 and the number of computing offloading MUs is 18. The task complexity takes value from the set [0.5 1 2 4 8 16]G cycles per bit. From **Fig. 4**, one can observe that, the weighted sum of offloading cost of the system slowly increases with the task complexity. That is, these two algorithms are both not sensitive with the task complexity. In fact, the offloading cost now is mainly determined by the computing resource allocation, thus the tendency of these curves can be concluded from Proposition 2.

In **Fig. 5**, the average system delay and energy consumption under different delay weights are evaluated. The parameters used in the simulation is the same as that used in **Fig. 4**. While the delay weight is varying from 0.1 to 0.9. One can observe that, as expected, the average system delay decreases and the average system energy consumption increases with the increase of delay weight. Obviously, the increase of weight for delay means that the contribution of delay in the system offloading cost becomes larger, thus the delay will be reduced when the offloading cost is minimized. In addition, one can observe that the reduction of average user delay is at the cost of more energy consumption.

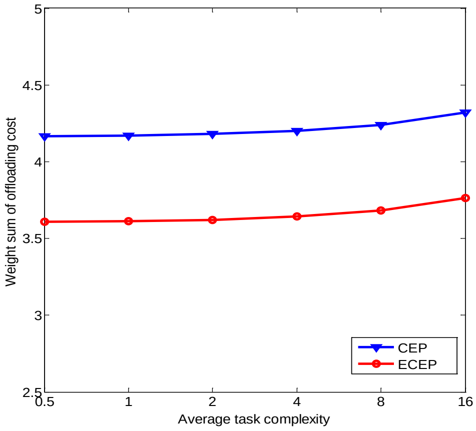


Fig. 4. Offloading cost versus the complexities of the tasks

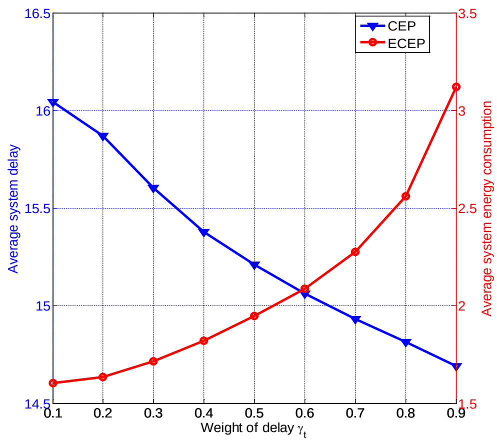


Fig. 5. The delay and energy consumption with the delay weights

At last, we study how does the user offloading cost performance is affected by the user weights, and the result is shown in **Fig. 5**. Herein, the weight of a particular MU, i.e., the MU 1, varies over [1/125 1/25 1 5 25 125], the weights of the other MUs are all equal to 1. The other parameters used the simulation are the same as that used in **Fig. 4**. One can note that, as the weight of MU 1 increases, its offloading cost of both algorithms decrease. This phenomenon is consistent with intuition and analysis as that, increasing the weight of MU 1, the contribution of MU 1 in the system offloading cost becomes larger, then in order to minimize the system offloading cost, we have to significant reduce its offloading cost. In addition, we also note that the curve of ECEP is steeper than CEP. This comes from the fact that for CEP, the power adjustment of each MU should consider the co-channel MUs' offloading costs due to the equal power allocation rule. This is not the case for ECEP where the power of each MU is independently adjusted. Therefore, when the user's weight is small, the user utility of ECEP is bigger than CEP and vice versa.

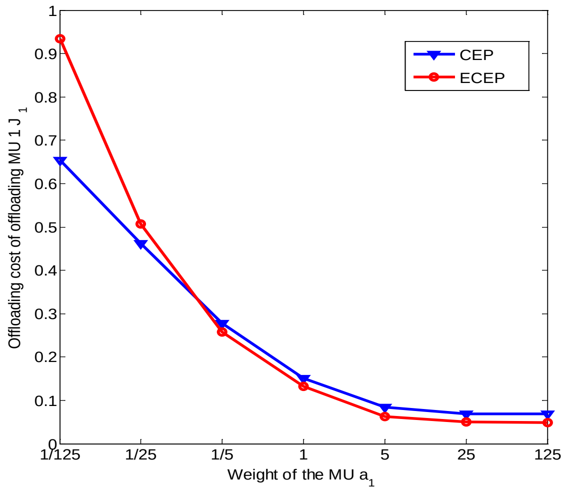


Fig. 6. User offloading cost with the user weights.

5. Conclusion

In this paper, the joint resource allocation for ultra-dense MEC network has been discussed. In particular, we study a scenario that some MUs ask for computation offloading service while the others demand communication service. We formulate the problem as a joint channel assignment, power control and computational resource allocation to minimize the offloading cost of computing offloading MUs, with the precondition that the transmission rate of communication MUs are satisfied. Since the considered problem is a mixed-integer non-linear program (MINLP) which is NP-hard, then two heuristic algorithms are proposed to obtain the suboptimal solutions, i.e., CEP and ECEP. Our simulation results confirm that the ECEP outperforms CEP in offloading cost performance. Future interests are about the scenario with multiple MEC servers and with partial offloading tasks on the users.

References

[1] W. Jiang, Y. Gong, Y. Cao, X. Wu, and Q. Xiao, "Energy-delay-cost Tradeoff for Task Offloading in Imbalanced Edge Cloud Based Computing," *arXiv:1805.02006*, 2018. Article (CrossRef Link).
[2] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322-2358, 2017. Article (CrossRef Link).
[3] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 4, pp.54-61, Apr. 2017. Article (CrossRef Link).

[4] C. Long, Y. Cao, T. Jiang and Q. Zhang, "Edge Computing Framework for Cooperative Video Processing in Multimedia IoT Systems," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1126-1139, May 2018. Article (CrossRef Link).
[5] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-Optimal Mobile Cloud Computing under Stochastic Wireless Channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4569-4581, Sep 2013. Article (CrossRef Link).
[6] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating While Computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 45-55, 2014. Article (CrossRef Link).
[7] J. Cheng, Y. Shi, B. Bai, et al, "Computation offloading in cloud-RAN based mobile cloud computing system," in *Proc. of IEEE International Conference on Communications. IEEE*, 1-6, 2016. Article (CrossRef Link).
[8] Y. Yu, J. Zhang, and K. B. Letaief, "Joint subcarrier and CPU time allocation for mobile edge computing," in *Proc. of IEEE Globecom, Washington, DC*, Dec. 2016. Article (CrossRef Link).
[9] X. Wang, J. Wang, X. Wang, and X. Chen, "Energy and Delay Tradeoff for Application Offloading in Mobile Cloud Computing," *IEEE Systems Journal*, vol. 11, no. 2, pp. 858-867, Jun. 2017. Article (CrossRef Link).
[10] Dinh T Q, Tang J, La Q D, et al, "Offloading in Mobile Edge Computing: Task Allocation and Computational Frequency Scaling," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3571-3584, 2017. Article (CrossRef Link).
[11] C. You, K. Huang, H. Chae, and et al, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397-1411, 2017. Article (CrossRef Link).
[12] Y. Wang, M. Sheng, X. Wang, and J. Li, "Mobile-Edge Computing: Partial Computation Offloading Using Dynamic Voltage Scaling," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268-4282, 2016. Article (CrossRef Link).
[13] X. Cao, F. Wang, J. Xu, R. Zhang, S. Cui, "Joint Computation and Communication Cooperation for Mobile Edge Computing," *CoRR abs/1704.06777*, 2017. Article (CrossRef Link).
[14] F. Wang, J. Xu, Z. Ding, "Optimized Multiuser Computation Offloading with Multi-antenna NOMA," in *Proc. of 2017 IEEE Globecom Workshops (GC Wkshps)*, 2017. Article (CrossRef Link).
[15] N. Fernando, S. W. Loke, W. Rahayu, "Computing with Nearby Mobile Devices: a Work Sharing Algorithm for Mobile Edge-Clouds," *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 329-343, 2019. Article (CrossRef Link).
[16] S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Transactions on Signal & Information Processing Over Networks*, vol. 1, no. 2, pp. 89-103, 2015. Article (CrossRef Link).
[17] A. Al-Shuwaili, O. Simeone, A. Bagheri, et al, "Joint Uplink/Downlink Optimization for Backhaul-Limited Mobile Cloud Computing with User Scheduling," *IEEE Transactions on Signal & Information Processing Over Networks*, vol. 3, no. 4, pp. 787 -802, Dec. 2017. Article (CrossRef Link).
[18] X. Chen, "Decentralized Computation Offloading Game for Mobile Cloud Computing," *IEEE Transactions on Parallel & Distributed Systems*, vol. 26, no. 4, pp. 974-983, 2014. Article (CrossRef Link).
[19] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient mutli-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*,vol. 24, no. 5, pp. 2795–2808, Oct. 2016. Article (CrossRef Link).
[20] F. Wang, J. Xu, Z. Ding, "Optimized Multiuser Computation Offloading with Multi-antenna NOMA," in *Proc. of 2017 IEEE Globecom Workshops (GC Wkshps)*, 2017. Article (CrossRef Link).
[21] J. Guo, Z. Song, Ying Cui, Z. Liu and Y. Ji, "Energy-efficient resource allocation for multi-user mobile edge computing," in *Proc. of IEEE Global Communications Conference (GLOBECOM), Singapore*, Dec. 2017. Article (CrossRef Link).

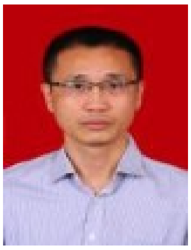
[22] Y. Mao, J. Zhang, and K. B. Letaief, "Joint task offloading scheduling and transmit power allocation for mobile-edge computing systems," in *Proc. of IEEE Wireless Commun. Networking Conf. (WCNC), San Francisco, CA*, Mar. 2017. Article (CrossRef Link).
[23] O. Munoz, A. Pascual Iserte, J. Vidal, M. Molina, and Ieee, "Energy-Latency Trade-off for Multiuser Wireless Computation Offloading," in *Proc. of 2014 Ieee Wireless Communications and Networking Conference Workshops (IEEE Wireless Communications and Networking Conference Workshops)*, pp. 29-33, 2014. Article (CrossRef Link).
[24] P.-H. Kuo, A. Mourad, "User-Centric Multi-RATs Coordination for 5G Heterogeneous Ultra-Dense Networks," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 6-8, Jan. 2018. Article (CrossRef Link).
[25] T. Zhou, N. Jiang, Z. Liu, C. Li, "Joint Cell Activation and Selection for Green Communications in Ultra-Dense Heterogeneous Networks," *IEEE Access*. vol. 6, pp. 1894-1904, 2018. Article (CrossRef Link).
[26] M. Chen, Y. Hao, "Task Offloading for Mobile Edge Computing in Software Defined Ultra-Dense Network," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 587-597, Mar. 2018. Article (CrossRef Link).
[27] Z. Chang, Z. Zhou, S. Zhou, T. Chen, T. Ristaniemi, "Towards Service-Oriented 5G: Virtualizing the Networks for Everything-as-a-Service," *IEEE Access*. vol. 6, pp. 1480-1489, 2018. Article (CrossRef Link).
[28] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University press, 2004. Article (CrossRef Link).
[29] O. Munoz, A. Pascual-Iserte, and J. Vidal, "Optimization of Radio and Computational Resources for Energy Efficiency in Latency-Constrained Application Offloading," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4738-4755, Oct 2015. Article (CrossRef Link).
[30] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-advanced for mobile broadband*, Academic press, 2013.

Junyi Liu received his MS degree in Zhejiang University in March 2014. He is an engineer in State Grid Zhejiang Electric Power Co.,Ltd. His research fields include network communication analysis and image processing. shshzaa@163.com

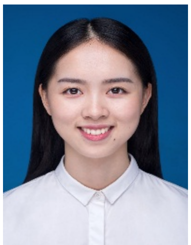
Hongbing Huang received his MS degree in electrical engineering from Wuhan University in 2003. He is currently working in State Grid Zhejiang Information & Telecommunication Company. His research interests include electric power system communication, image processing technology. james_hb@163.com

Yijun Zhong received his MS degree in Zhengjiang University in June 2008. He is currently working in state grid Zhejiang electric power co., LTD. His research fields include information and telecommunication. zhongyijun1984@163.com

Jiale He received his master's degree in Tianjin University of Technology in 2016. He obtained his Bachelor of Engineering from the North China Electric Power University in china. His current major research direction is optical fiber communication. nyhejiale@163.com



Tiancong Huang received the Ph.D. degree from Chongqing University (CQU), Chongqing, China, in 2010. He is currently with the School of Microelectronics and Communication Engineering, CQU. He has authored or coauthored more than 25 articles in journals and conference proceedings. His research interests include electric power communication and cooperative communication.



Qian Xiao received the B.S. degree in communication engineering, in 2015, from Chongqing University, Chongqing, China, where she is currently working toward the M.S. degree in communication engineering. Her research interests include wireless resource allocation and management, physical layer security, and mobile edge computing.

Weiheng Jiang (Member, IEEE) received the Ph.D. degree in communication engineering from Chongqing University, Chongqing, China, in July 2015. He is currently an Assistant Professor with the School of Microelectronics and Communication Engineering, Chongqing University. His research interests include a broad range of areas from signal processing to wireless communications and networking, include 5G and beyond, big data algorithms, machine learning, and deep learning. He is also a reviewer of the IEEE Transactions on Signal Processing, IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, IEEE Systems Journal, IEEE Access and some other journals.