154 -C115b4

Joint Offloading and Resource Allocation Optimization for Time-Sensitive Mobile-Edge Computing Network

Jun-Jie Yu, Wen-Tao Li, Mingxiong Zhao

***Abstract*—Mobile Edge Computing MEC is a "hardware + software" system that provides IT service environments and cloud computing capabilities at the edge of the mobile network, enabling mobile devices to offload computationally intensive or latency-critical tasks to attached MEC servers. To effectively reduce the execution delay and energy consumption of mobile devices. The MEC and the traditional cloud server are deployed in different forms. The MEC server is a small server deployed on the side of the wireless access layer, so the requirements for computing resources and wireless networks are relatively high. In this paper, we consider a multi-user MEC server system based on Orthogonal Frequency Division Multiple Access (OFDMA), and we research User resource and MEC computing resource allocation, to minimize the user's task execution time and reduce the total system energy consumption. Through the three-layer optimization method, the original NP-hard problem is transformed into the problem of finding the optimal offloading decision problem, subcarrier allocation and computing resource allocation. Simulation results show that the proposed algorithm achieves excellent performance .**

I. Introduction

Accompanying with the development of technology, a considerable number of novel applications, suck as Virtual Reality(VR), Online Interactive Games, Smart City Services, Autonomous Vehicle and Industrial Internet of Things(IIoT), put forward stringent requirements of the communication delay, performance, reliability, data capacity and other aspects of metrics. For example, VR applications and devices generally demand enormous computation workload while the expected delay should be lower than 1ms[1]. Mobile Edge Computing(MEC), as a key technology of 5G network defined by ETSI, makes it possible for these applications by providing IT infrastructures in close proximity to terminal user equipment(UE) within the Radio Access Network (RAN) [2]. Comparing with offloading to traditional Mobile Cloud Computing(MCC), applications on UE could save the dominated propagation delay and energy consumption by forwarding task requests to MEC servers for computation, thus avoiding core network(CN) congestion. MEC Offloading is the process by which tasks on UE can choose to be executed locally or sent to MEC server for computation, which is particularly helpful for energy constrained devices saving battery-life. In literature, according to whether the task could be partitioned, the Offloading process is divided into binary offloading and partial offloading.

In [3], a Lyapunov optimization-based LODCO algorithm was developed to jointly decide offloading option and CPU frequencies for a MEC system integrated with energy harvesting (EH) technologies, which, however, only considers one mobile devices offloading tasks to multi-access points.

In this paper, we propose a joint offloading and subcarrier allocation strategy. This strategy jointly optimizes the offload strategy, subcarrier and computing resource allocation scheme to reduce the total energy consumption of the entire MEC system. However, due to the NP-hard features and non-convexity of the joint problem, it is difficult to solve,

II. System model and problem formulation

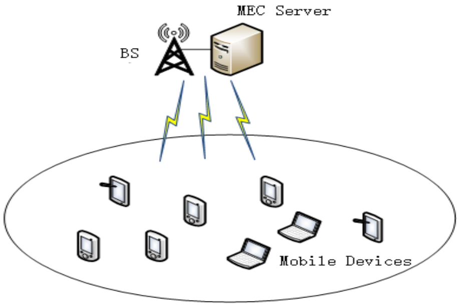

Fig. 1. The MEC system model of multiuser offloading with OFDMA.

We consider an OFDMA-based MEC system with K users and one base-station (BS) integrated with an MEC server to execute the offloaded data of users, and all nodes are equipped with a single antenna. Denote $\mathcal{K} \triangleq \{1, 2, \cdots, K\}$ as the set of users, and let $\mathcal{N} \triangleq \{1, 2, \cdots, N\}$ be the index for multiple orthogonal subcarriers, each of which has bandwidth B and can be assigned to only one user. In this system, we assume that user k has a task described by a tuple of four parameters $\{R_k, c_k, \lambda_k, t_k\}$, where R_k indicates the amount of input data to be processed, c_k represents the number of CPU cycles for computing 1-bit of input data, $\lambda_k \in [0, 1]$ is the proportion of R_k offloading to MEC, while the rest $(1 - \lambda_k)R_k$ bits are processed by its local CPU, and t_k is the maximum tolerable latency. In this paper, it is assumed that the maximum tolerable latency for user k, t_k is no longer than the channel coherence time, such that the wireless channels remain constant during a time slot with length T, i.e., $t_k \leq T, \forall k$, but can vary from time to time. The local CPU frequency of user k is characterized by f_k, and $f_{k,m}$ is the computational speed of the edge cloud assigned to user k, where both of them are measured by the number of CPU cycles per second.

Herein, a practical constraint that the total computing resources allocated to all the associated users must not excess the servers computing capacity is given by $\sum_{k=1}^{K} f_{k,m} \leq F$.

In the following, the time latency and the energy consumption of user k for our considered system are given in details.

A. Time Latency

1) Local Computing at Users: Consider the local computing for executing the residual $(1-\lambda_k)R_k$ input bits at user k, the time consumption for local computing at user k is

$$t_{k,l} = \frac{c_k\left(1-\lambda_k\right)R_k}{f_k}. \tag{1}$$

2) Computation Offloading: According to the OFDMA mechanism, the inter-interference is ignored in virtue of the exclusive subcarrier allocation. Therefore, the aggregated transmission rate to offload $\lambda_k R_k$ input bits from user k to MEC server is expressed as

$$r_k = B\sum\nolimits_{n=1}^{N} x_{k,n}\log_2\left(1+\frac{p_k g_{k,n}}{\sigma_n^2}\right), \tag{2}$$

where $g_{k,n}$ and σ_n^2 are the channel gain between user k and BS, and the variance of the additive white Gaussian noise at BS on subcarrier n, respectively, where we set $\sigma_n^2 = \sigma^2, \forall n$ for simplicity, p_k denotes the transmission power,and it can be allocated by the users. We define the maximum transmission power as p_k^{max}. Apparently, any power optimization solution have a good impact on system performance. For the sake of simplicity, p_k remains at a random level in this paper.[4] Meanwhile, denote $x_{k,n}$ as the channel allocation indicator, specifically $x_{k,n}=1$ means that subcarrier n is assigned to user k, otherwise $x_{k,n}=0$.

The offloading time $t_{k,\text{off}}$ of user k mainly consists of two parts[1]: the uplink transmission time $t_{k,u}$ from user k to MEC server and the corresponding execution time at MEC server $t_{k,m}$. Therefore, the offloading time $t_{k,\text{off}}$ is given by

$$t_{k,\text{off}} = t_{k,u} + t_{k,m} = \frac{\lambda_k R_k}{r_k} + \frac{\lambda_k R_k c_k}{f_{k,m}}. \tag{3}$$

Due to the parallel computing at users and MEC server, the total time latency for user k depends on the the larger one between $t_{k,l}$ and $t_{k,\text{off}}$, and can be expressed as

$$t_k = \max\{t_{k,l}, t_{k,\text{off}}\}. \tag{4}$$

B. Energy Consumption

According to the strategy of computation offloading at user k, the total energy consumption comprises two parts[1]: the energy for local computing and for offloading, given in details as follow.

1) Local Computing Mode: Given the processor's computing speed f_k, the power consumption of the processor is modeled as $\kappa_k f_k^3$ (joule per second), where κ_k represents the computation energy efficiency coefficient related to the processor's chip of user k [7]–[9]. Taking consideration of (1), the energy consumption at this mode is given by

$$E_{k,l} = \kappa_k f_k^3 t_{k,l} = \kappa_k c_k\left(1-\lambda_k\right)R_k f_k^2. \tag{5}$$

2) Computation offloading mode: In this mode, the energy consumption includes the cost of uplink transmitting and remote computing for offloaded $\lambda_k R_k$ input bits, which can be obtained as

$$\begin{aligned} E_{k,\text{off}} &= E_{k,u} + E_{k,m} \\ &= p_k\frac{\lambda_k R_k}{r_k} + \kappa_m \lambda_k c_k R_k f_{k,m}^2, \end{aligned} \tag{6}$$

where κ_m is the computation energy efficiency coefficient related to the processor's chip of MEC server.

Therefore, the total energy consumption for user k related with its computation offloading strategy in our system is

$$E_k = E_{k,l} + E_{k,u} + E_{k,m}. \tag{7}$$

In this paper, we minimize the overall energy consumption of the considered system, which is related to resource allocation on the channel, offloading communication and computation. Mathematically, the energy consumption minimization problem can be written as

$$\begin{aligned} \mathbf{P1}: \ \min_{\boldsymbol{X},\boldsymbol{\lambda},\boldsymbol{f}} \ & \sum\nolimits_{k=1}^{K} E_k & \text{(8a)}\\ \text{s.t.} \ & 0 \leq f_{k,m}, \forall k, & \text{(8b)}\\ & \sum\nolimits_{k=1}^{K} f_{k,m} \leq F, & \text{(8c)}\\ & t_k \leq T, \forall k, & \text{(8d)}\\ & \sum\nolimits_{k=1}^{K} x_{k,n} \leq 1, \forall n, & \text{(8e)}\\ & x_{k,n} \in \{0,1\}, \forall k, n, & \text{(8f)}\\ & 0 \leq \lambda_k \leq 1, \forall k, & \text{(8g)} \end{aligned}$$

where $\boldsymbol{X} \triangleq \{x_{k,n}\}$, $\boldsymbol{\lambda} \triangleq \{\lambda_k\}$ and $\boldsymbol{f} \triangleq \{f_{k,m}\}$. The constraints in the formulation above can be explained as follows: constraint (8b) and (8c) show that MEC server must allocate a positive computing resource to user associated with it, and the sum of which cannot exceed the total computational capability of MEC server; constraint (8d) states that the task of user k must be completely executed within a time slot; constraint (8e) and (8f) enforce that each subcarrier can only be used by one user to avoid the multi-user interference; constraint (8g) shows range of offloading ratio.

III. Resource Allocation Joint Offloading Strategy Algorithm

In view of the problem P1,we can know P1 is nonconvex, finding the optimal solution is usually prohibitively due to the complexity. However, the duality gap becomes zero in multicarrier systems as the number of subcarriers goes to large and the time sharing conditionis satisfied. Thus the optimal solution to a nonconvex resourceallocation problem

[1] In practice, the MEC-integrated BS will provide sufficient transmit power, while the amount of output data from MEC server to user k is usually much less than that of the input data, the time consumed and the transmission energy for delivering the computed results are negligible [5], [6].

in multicarrier system can be obtained in the dual domain. Nevertheless, as we will dicuss later,the traditional Lagrangian decomposition cannot be directly employed to decompose the problem into parallel subproblems with eachsubproblem corresponding to one subcarrier. This is becausethe offloading ratio $\boldsymbol{\lambda}$ appears in the rate expression.

A. *Iterative optimization approach*

The process is repeated until both $\boldsymbol{X}$,$\boldsymbol{\lambda}$ and $\boldsymbol{f}$ converge, which is known as the block coordinate descent(BCD) method.

We define $\boldsymbol{\mathcal{T}}$ as all sets of possible $\boldsymbol{X}$ that satisfy (9e) and (9f), $\boldsymbol{\mathcal{R}}$ as all sets of possible $\boldsymbol{\lambda}$ that satisfy $0 \leq \lambda_k \leq 1$, and $\boldsymbol{\mathcal{F}}$ as all sets of possible that satisfy (9b)

1) To solve task offload ratio: To solve $\boldsymbol{\lambda}$ with given $\boldsymbol{X}$ and $\boldsymbol{f}$, because we can get subproblem P2 of P1 for user k with fixed $\boldsymbol{x}_k$ and $\boldsymbol{f}_{k,m}$

$$\begin{aligned} \mathbf{P2}: \quad & \min_{\boldsymbol{\lambda} \in \boldsymbol{\mathcal{R}}} \; E_k && \text{(9a)} \\ & \text{s.t. } t_{k,l} \leq T_k, && \text{(9b)} \\ & \quad t_{k,off} \leq T_k, && \text{(9c)} \end{aligned}$$

Then we can get the value range of λ_k is expressed as

$$1 - \frac{T_k f_k}{c_k R_k} \leq \lambda_k \leq \frac{T_k r_k f_{k,m}}{R_k f_{k,m} + r_k R_k c_k} \tag{10}$$

and for P2, we have

$$\frac{\partial E_k}{\partial \lambda_k} = \frac{r_k R_k c_k \left(f_{k,m}^2 k_m - f_k^2 k_0 \right) + p_k R_k}{r_k} \tag{11}$$

So, we can get the λ_k^*

$$\lambda_k^* = \begin{cases} \left[1 - \dfrac{T_k f_k}{c_k R_k} \right]^+, & \dfrac{\partial E_k}{\partial \lambda_k} > 0 \\ \min \left\{ \dfrac{T_k r_k f_{k,m}}{R_k f_{k,m} + r_k R_k c_k}, 1 \right\}, & \dfrac{\partial E_k}{\partial \lambda_k} < 0 \end{cases} \tag{12}$$

where $[x]^+ = \max\{0, x\}$ and if $\frac{\partial E_k}{\partial \lambda_k}$ = 0, then the value of E_k is independent of λ_k.

2) To solve resource allocation strategy: To solve $\boldsymbol{X}$ and $\boldsymbol{f}$ with given $\boldsymbol{\lambda}$.

First,According to (13), we can get two special cases, respectively.

- First special case: $\lambda_k = 0$ When $\lambda_k = 0$, this means that user k does not perform computation offloading. So, $X_k = 0, f_{k,m} = 0$.
- Second special case: $\lambda_k = 1$ When $\lambda_k = 1$, this means that user k offload all data to the MEC server. So, we can get new problem P3

$$\begin{aligned} \mathbf{P3}: \quad & \min_{\boldsymbol{X} \in \boldsymbol{\mathcal{T}}, \boldsymbol{f} \in \boldsymbol{\mathcal{F}}} \; \sum_{k=1}^{K} (E_{k,u} + E_{k,s}) && \text{(13a)} \\ & \text{s.t. } t_{k,off} \leq T_k, \forall k && \text{(13b)} \\ & \quad \sum_{k}^{K} f_{k,m} \leq F && \text{(13c)} \end{aligned}$$

Then we can obtain a suboptimal solution by iteratively optimizing $\boldsymbol{X}$ with fixed $\boldsymbol{f}$, and optimizing $\boldsymbol{f}$ with fixed $\boldsymbol{X}$.

To solve $\boldsymbol{X}$ with fixed $\boldsymbol{f}$, then the problem P3 is transformed into Prolem P4

$$\begin{aligned} \mathbf{P4}: \quad & \min_{\boldsymbol{X} \in \boldsymbol{\mathcal{T}}} \; E_{k,u} && \text{(14a)} \\ & \text{s.t. } t_{k,off} \leq T_k && \text{(14b)} \\ & && \text{(14c)} \end{aligned}$$

and $\min_{\boldsymbol{X} \in \boldsymbol{\mathcal{T}}} E_{k,u}$ equivalent to $\max_{\boldsymbol{X} \in \boldsymbol{\mathcal{T}}} \frac{1}{E_{k,u}}$, So we can get $\hat{P}4$

$$\begin{aligned} \hat{\mathbf{P4}}: \quad & \max_{\boldsymbol{X} \in \boldsymbol{\mathcal{T}}} \; \frac{1}{E_{k,u}} && \text{(15a)} \\ & \text{s.t. } t_{k,off} \leq T_k && \text{(15b)} \\ & && \text{(15c)} \end{aligned}$$

The Lagrangian function for $\hat{P}4$ is given by

$$\begin{aligned} & \mathcal{L}(\boldsymbol{X_k}, \alpha_k) = \\ & \frac{r_k}{p_k R_k} - \qquad = \end{aligned} \tag{16}$$

According to $t_{k,l} = t_{k,u} + t_{k,m}$ and (6), the (11d) in P1 can be transformed into $t_{k,l} \leq T_k$, then we can get P2

$$\begin{aligned} \mathbf{P2}: \quad & \min_{\boldsymbol{X} \in \mathcal{T}} \; \sum_{k=1}^{K} E_k && \text{(17a)} \\ & \text{s.t. } t_{k,l} \leq T_k, \forall k && \text{(17b)} \\ & \quad \sum_{n=1}^{N} p_{k,n} x_{k,n} \leq p_k^{max}, \forall k && \text{(17c)} \end{aligned}$$

Because $t_{k,l} = t_{k,u} + t_{k,m}$, so there is

$$\begin{aligned} & E_{k,u} = \\ & (t_{k.l} - t_{k,m}) \sum_{n=1}^{N} x_{k,n} p_{k,n} \\ & = \left[\frac{c_k (1 - \lambda_k) R_k}{f_k} - \frac{\lambda_k R_k c_k}{f_{k,m}} \right] \sum_{n=1}^{N} x_{k,n} p_{k,n} \end{aligned} \tag{18}$$

The Lagrangian function for P2 is given by

$$\begin{aligned} & \mathcal{L}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \\ & \sum_{k=1}^{K} \Bigg\{ k_0 c_k (1 - \lambda_k) R_k f_k^2 + k_m c_k \lambda_k R_k f_{k,m}^2 \\ & + \left[\frac{c_k (1 - \lambda_k) R_k}{f_k} - \frac{\lambda_k R_k c_k}{f_{k,m}} \right] \sum_{n=1}^{N} x_{k,n} p_{k,n} \\ & + \beta_k \frac{c_k (1 - \lambda_k) R_k}{f_k} + \zeta_k \sum_{n=1}^{N} x_{k,n} p_{k,n} \Bigg\} \\ & - \left[\sum_{k=1}^{K} (\beta_k T_k + \zeta_k P_k^{\max}) \right] \\ & = \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg\{ w(\lambda_k, f_{k,m}) + \zeta_k x_{k,n} p_{k,n} + \Bigg[\frac{c_k (1 - \lambda_k) R_k}{f_k} \\ & - \frac{\lambda_k R_k c_k}{f_{k,m}} \Bigg] x_{k,n} p_{k,n} \Bigg\} - \left[\sum_{k=1}^{K} (\beta_k T_k + \zeta_k P_k^{\max}) \right] \end{aligned} \tag{19}$$

and $\boldsymbol{\beta}, \boldsymbol{\zeta}$ are the non-negative Lagrage multipliers. The dual function is then defined as

$$g(\boldsymbol{\beta}, \boldsymbol{\zeta}) = \min_{X \in \mathcal{T}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\zeta}) \tag{20}$$

Thus,the dual problem of (15) is given by

$$\begin{align} &\max g(\boldsymbol{\beta}, \boldsymbol{\zeta}) \tag{21a} \\ &\text{s.t.} \boldsymbol{\beta} \succeq 0, \boldsymbol{\zeta} \succeq 0 \tag{21b} \end{align}$$

For the minmization problem of (14a), Then for given dual variables $\boldsymbol{\beta}, \boldsymbol{\zeta}$, to solve $\boldsymbol{X}$ with fixed $\boldsymbol{\lambda}$ and $\boldsymbol{f}$, suppose that subcarrier n is assigned to user k, we have

$$\mathcal{L} = \sum_{n=1}^{N} \mathcal{L}_n - \left[\sum_{k=1}^{K} (\beta_k T_k + \zeta_k P_k^{\max}) \right] \tag{22}$$

where

$$\begin{split} \mathcal{L}_n = \sum_{k=1}^{K} w(\lambda_k, f_{k,m}) + \zeta_k p_{k,n} + \\ \left[\frac{c_k \left(1 - \lambda_k\right) R_k}{f_k} - \frac{\lambda_k R_k c_k}{f_{k,m}} \right] p_{k,n} \end{split} \tag{23}$$

the function $w(\lambda_k, f_{k,m})$ is expressed as

$$\begin{split} w(\lambda_k, f_{k,m}) = \frac{k_0 c_k (1 - \lambda_k) R_k f_k^2}{N} \\ + \frac{k_m c_k \lambda_k R_k f_{k,m}^2}{N} + \frac{\beta_k c_k (1 - \lambda_k) R_k}{N f_k} \end{split} \tag{24}$$

Thus, the subproblem is given by

$$\min_{X_n \in \mathcal{T}} \mathcal{L}_n \left(\boldsymbol{\lambda}, X_n, \boldsymbol{f} \right) \tag{25}$$

where $\boldsymbol{X}_n = \{x_{k,n}\}_{k=1}^{K}$, and this problem can be solved independently. By maximizing each $\mathcal{L}_n$, the optimal $\boldsymbol{X}$ can be obtained as

$$x_{k,n}^* = \begin{cases} 1, \text{if} \quad k = k^* = \text{argmin}_k \, \mathcal{L}_n \\ 0, \text{ otherwise} \end{cases} \tag{26}$$

3) To solve computing resource allocation: To solve $\boldsymbol{f}$ with given $\boldsymbol{X}$ and $\boldsymbol{\lambda}$, because $\boldsymbol{X}$and $\boldsymbol{\lambda}$ are fixed, this is equivalent to the local energy consumption of computing and uploading data. and due to $t_k = t_{k,l} = t_{k,u} + t_{k,m}$, So we can get subproblem P3 of P1

$$\begin{align} \mathbf{P3}: \; &\min_{\boldsymbol{f} \in \mathcal{F}} \sum_{k=1}^{K} E_{k,s} \tag{27a} \\ \text{s.t.} \; &\sum_{k=1}^{K} f_{k,m} \leq F \tag{27b} \\ &t_{k,u} + t_{k,m} \leq T_k, \forall k \tag{27c} \end{align}$$

The Lagrangian function for P3 is given by

$$\begin{split} &\mathcal{G}(\boldsymbol{f}, \gamma, \boldsymbol{\xi}) = \\ &\sum_{k=1}^{K} E_{k,s} + \gamma \left(\sum_{k=1}^{K} f_{k,m} - F \right) \\ &- \sum_{k=1}^{K} \xi_k \left(f_{k,m} + \frac{\lambda_k R_k c_k r_k}{\lambda_k R_k - T_k r_k} \right) \\ &= \sum_{k=1}^{K} k_m c_k \lambda_k R_k f_{k,m}^2 + \gamma \left(\sum_{k=1}^{K} f_{k,m} - F \right) \\ &- \sum_{k=1}^{K} \xi_k \left(f_{k,m} + \frac{\lambda_k R_k c_k r_k}{\lambda_k R_k - T_k r_k} \right) \end{split} \tag{28}$$

where γ, $\boldsymbol{\xi} = [\xi_1, \xi_2, ..., \xi_k]^T$ is the nonnegative Lagrange multipliers. Observing P3, it is easy to get P3 is a convex optimization problem. So, applying the Karush-Kuhn-Tucker (KKT) conditions, each f_k^{M*} has to satisfy

$$\frac{\partial \mathcal{G}}{\partial f_k^{M*}} = 2k_m \lambda_k R_k f_{k,m} + \gamma - \xi_k = \begin{cases} =0, & f_k^{M*} > 0 \\ <0, & f_k^{M*} = 0 \end{cases} \tag{29}$$

and we can get

$$f_{k,m} = \left[\frac{\xi_k - \gamma}{2k_m \lambda_k R_k} \right]^+, \forall k \in k \tag{30}$$

where $x^+ = \max(0, x)$. The dual function is then defined as

$$\mathcal{Q}(\gamma, \boldsymbol{\xi}) = \min_{\boldsymbol{f} \in \mathcal{F}} \mathcal{G}(\boldsymbol{f}, \gamma, \boldsymbol{\xi}) \tag{31}$$

Thus,the dual problem of (28) is given by

$$\begin{align} &\max \mathcal{Q}(\lambda, \boldsymbol{\xi}) \tag{32a} \\ &\text{s.t.} \gamma \geq 0, \boldsymbol{\xi} \succeq 0 \tag{32b} \end{align}$$

With the fixed $\boldsymbol{X}$ and $\boldsymbol{f}$, the optimal λ_{k*} can be obtianed by (13), and with the fixed $\boldsymbol{\lambda}, \boldsymbol{f}$ the optimal $\boldsymbol{X}$ can be obtained, Then with the fixed $\boldsymbol{X}, \boldsymbol{\lambda}$, the optimal $\boldsymbol{f}$ can be obtained. Thus, the above process can be iterated until the optimal value of the objective function ceases to increase.

B. Lagrange Multipliers Update

Then, we have determined the optimal $\boldsymbol{\lambda}^*, \boldsymbol{X^*}, \boldsymbol{f^*}$ for given $\boldsymbol{\beta}, \boldsymbol{\zeta}$ and $\gamma, \boldsymbol{\xi}$, we can discuss how to update Lagrange Multipliers in the following. The dual problem can be expressed as $\max g(\boldsymbol{\beta}, \boldsymbol{\zeta})$ and $\max \mathcal{Q}(\gamma, \boldsymbol{\xi})$,where $\boldsymbol{\beta} \succeq 0, \boldsymbol{\zeta} \succeq 0$ and $\gamma \geq 0, \boldsymbol{\xi} \succeq 0$.

We can easily prove that the dual problem is a convex one. Thus, a subgradient of $g(\boldsymbol{\beta}, \boldsymbol{\zeta})$is given by

$$\begin{split} \Delta \beta_k = \frac{c_k (1 - \lambda_k R_k)}{f_k} - T_k \\ \Delta \zeta_k = \sum_{n=1}^{N} x_{k,n} p_{k,n} - p_k^{max} \end{split} \tag{33}$$

And the subgradient of $\mathcal{Q}(\gamma, \boldsymbol{\xi})$ is given by

$$\begin{split} \Delta \gamma = \sum_{k=1}^{K} f_{k,m} - F \\ \Delta \xi_k = \frac{\lambda_k R_k}{r_k} + \frac{\lambda_k R_k c_k}{f_{k,m}} - T_k \end{split} \tag{34}$$

Thus,the sunbgradient projection method for (18a) and (29a) is respectively as

$$\begin{aligned} \beta_k(z+1) &= [\beta_k(z) - \alpha_k \Delta\beta_k]^+ \\ \zeta_k(z+1) &= [\zeta_k(z) - \phi_k \Delta\zeta_k]^+ \end{aligned} \tag{35}$$

and

$$\begin{aligned} \gamma(z+1) &= [\gamma(z) - \rho_k \Delta\gamma]^+ \\ \zeta_k(z+1) &= [\zeta_k(z) - \eta_k \Delta\zeta_k]^+ \end{aligned} \tag{36}$$

where $z \geq 0$ is the iteration index,$[\alpha_1, \alpha_2, ...\alpha_k]$, $[\zeta_1, \zeta_2, ...\zeta_k]$, and ρ, $[\eta_1, \eta_2, ...\eta_k]$ are properly small positive step-sizes.

C. Algorithm

The whole procedure to solve P1 is summarized in Algorithm 1 in the following.

Algorithm 1 Proposed Iterative Algorithm

```
initialize:
  • Set $\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \gamma, \boldsymbol{\xi}, \mathcal{Z}_{max}, \epsilon$.
  • Set z = 0
repeat
    repeat
        Solve task offload ratio $\boldsymbol{\lambda}$ according to (13).
        Determine subcarrier allocation $\boldsymbol{X}$ according to (23) and compute L according to (16).
        Allocate computing resource $\boldsymbol{f}$ accoring to (27).
    until $\sum_{k=1}^{K} E_k$ converges.
    Update $\boldsymbol{\beta}, \boldsymbol{\zeta}$, and $\gamma, \boldsymbol{\xi}$ from (32) and (33).
    z = z + 1
    if $\|\boldsymbol{\beta}(z+1) - \boldsymbol{\beta}(z)\| \leq \epsilon$ and $\|\boldsymbol{\zeta}(z+1) - \boldsymbol{\zeta}(z)\| \leq \epsilon$ and $\|\gamma(z+1) - \gamma(z)\| \leq \epsilon$ and $\|\boldsymbol{\xi}(z+1) - \boldsymbol{\xi}(z)\| \leq \epsilon$ then
        break.
until $z > \mathcal{Z}_{max}$
```

References

[1] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 2961–2991, 2018.

[2] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing-A key technology towards 5G. ETSI White Paper," *ETSI White Pap.*, vol. 11, no. 11, pp. 1–16. 2015.

[3] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing with Energy Harvesting Devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, 2016, ISSN: 07338716. DOI: 10.1109/JSAC.2016.2611964.

[4] X. Yang, X. Yu, H. Huang, and H. Zhu, "Energy efficiency based joint computation offloading and resource allocation in multi-access mec systems," *IEEE Access*, vol. 7, pp. 117 054–117 062, 2019. DOI: 10 . 1109 / ACCESS.2019.2936435.

[5] X. Hu, K. K. Wong, and K. Yang, "Wireless powered cooperation-assisted mobile edge computing," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2375–2388, 2018.

[6] C. Kang, Y. Teng, W. Sun, L. An, and X. Wang, "Energy-efficient joint offloading and wireless resource allocation strategy in multi-mec server systems," *Proc. IEEE ICC*, pp. 1–6, 2018.

[7] W. Zhang, Y. Wen, K. Guan, K. Dan, 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, 2013.

[8] Y. Wang, S. Min, X. Wang, W. Liang, 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.

[9] S. Bi, J. Ying, and Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 4177–4190, 2018.