Energy-Efficient Joint Offloading and Resource Allocation Strategy in MEC Server Systems with Delay Constrained

Mingxiong Zhao Junjie Yu Wentao Li

***Abstract*—Mobile edge computing (MEC) is an emerging paradigm that mobile devices can offload the computationintensive or latency-critical tasks to MEC servers,so as to save energy and extend battery life. Unlike the cloud server, MEC server is a small-scale data center deployed at a wireless access point, thus it is highly sensitive to both radio and computing resource. In this paper, we consider an Orthogonal Frequency-Division Multiplexing Access (OFDMA) based multi-user and Single-MEC-server system, where the task offloading strategies and resource allocation strategy are jointly investigated under delay constrained. Aiming at minimizing the total energy consumption, we propose the joint offloading and resource allocation strategy for latencycritical applications. Through the three-layer iterative optimization approach,the original NP-hard problem is decoupled into three simple sub-questions. Simulation results show that the proposed algorithm achieves excellent performance in energy saving and successful offloading probability (SOP) in comparison with conventional schemes.**

I. Introduction

The future 5G cellular network aims to support various low-latency mobile applications, such as vehicleto-everything (V2X) communications, internet of things (IoT), and intelligent wearable devices. The traditional cloud computing technology, which aims to deal with the problem of insufficient computational capacity and limited storage of mobile devices, cannot function effectively to provide millisecond-scale latency services with the explosive growth in the number of mobile devices. To meet with such a critical challenge, the novel technology of mobile edge computing (MEC) has emerged and called lots of attention in recent years. With the help of edge cloud, the base stations (BSs) in cellular networks can provide cloud-computing capabilities [1], [2]. In this way, the millisecond-scale latency requirement could be achieved by virtue of the closer distance between the cloud server and users. Moreover, the device energy consumption could also be effectively reduced by offloading computationally intensive tasks to the proximate MEC server for computing, which is known as mobile edge computation offloading (MECO).

To further improve energy efficiency and shorten endto-end latency of the MEC system, the joint allocation of communication and computation resources among MEC servers and mobile devices is worth investigation. Many recent studies have been focusing on the joint communication and computation resource allocation for energy-efficiency improvement [3], [4] and end-to-end latency reduction [5], [6]. Under the computation latency constraint, the authors in [3] have proposed to dynamically select local computing or cloud computing for minimizing the total energy consumption, and the authors in [4] have investigated the optimal resource allocation and offloading decision policy to reduce the weighted-sum mobile energy consumption. On the other hand, latencyoptimal task assignment in a single-user MECO system under resource constraints has been studied in [5], and a stochastic task arrival mode to solve the energy-latency tradeoff problem has been proposed in [6].

Most existing works mentioned above have focused on improving energy efficiency under certain latency constraint or minimizing the end-to-end latency in a singleuser scenario. Although the paper [7] has proposed a latency-optimal joint communication and computing resource allocation in the multi-user MECO system, it considered the time division multiple access (TDMA) scenario rather than the orthogonal frequency division multiple access (OFDMA) scenario, which is the main channel access technique for the 5G network.

Therefore, we are motivated to investigate the latencyminimization problem in an OFDMA-based MECO system in this paper. In such a system, we assume that each mobile device has raw data to be processed, such as image recognition, video reconstruction, and virus detection. Part of data can be computed locally at the mobile device while the other part is sent to the edge cloud for computing. After the computation is finished, the latter part is then sent back to the mobile device. Therefore, it is important to optimally allocate both communication and computation resources to minimize the end-to-end latency of each user.

Aiming at that, we formulate a resource allocation problem to minimize the maximal delay of each mobile device to guarantee certain fairness among users. The problem is hard to solve due to the binary variable of the channel allocation indicator. Therefore, we first relax the channel allocation indicators into continuous variables, and propose a lower-bound algorithm by solving a maxmin problem. Then, to reduce the computational complexity, we develop a heuristic algorithm by separating subcarrier assignment and power allocation. Finally, we use simulation results to show the performance of our proposed algorithms.

We organize the rest of this paper as follows. We will introduce the system scenario and problem formulation in Section II. The lower-bound algorithm and heuristic algorithm will be developed in Section III and Section IV, respectively.

Section V presents the simulation results and we conclude the paper in Section VI.

II. System Model

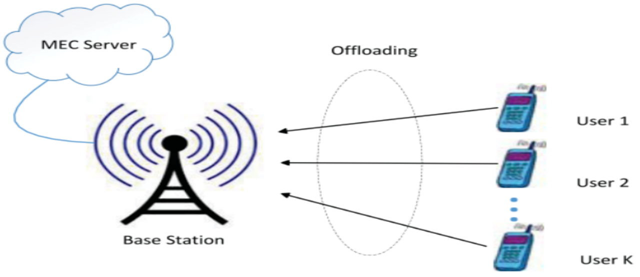

Fig. 1. Multi-user MECO system model.

We consider a single MEC server OFDMA system with K users denoted $\mathcal{K} = \{1, 2, \cdots, K\}$. each user has a task,each task is described by a three-field notation $< R_k, c_k, T_k >$, where R_K indicates the amount of data to be processed and the computation workload/intensity c_k in CPU cycles per bit)and T represents the task deadline. The edge server is connected by mobile users through wireless channels. There are N subcarriers in system,each has bandwidth of B Hz. User k's upload power p_k. We assume that the data can partitioned into two parts with one can be processed at the MEC server and then sent back to the mobile device after the work is done,and the proportion of data R_k that is processed at the edge server as $\lambda_k \in [0, 1]$. The computing capacity of the MEC server is F. The calculation frequency of the assigned user k is denoted f_k^M. Then we can break down the detailed procedures of partial computation offloading into three steps:

- Step 1:Mobile device k processes (1-λ_k)R_k bits of the raw data locally.
- Step 2:Mobile device k transmits the remaining $\lambda_k R_k$ bits to the edge server for processing.
- Step 3:Finally,the edge server processes $\lambda_k R_k$ bits of the raw data,and MEC sends the result to the mobile device.

III. Delay and Power Analysis

A. Delay Analysis

The time for downlink can be considered as negligible since it is much shorter than uplink and computational delay. We use $t_{k,l}, t_{k,u}, t_{k,s}$ to denote the delay of each step,respectively. Let $p_{k,n}$ represent power allocated on subcarrier n. Denoting the instantaneous channel power gain between user k and the BS on subcarrier n as $h_{k,n}$,the channel capacity can be expressed as

$$r_{k,n} = B \log_2 \left(1 + \frac{p_{k,n} h_{k,n}}{\sigma^2} \right) \tag{1}$$

where σ^2 represents the variance of the additive white Gaussian noise. The overall data rate of user k can be expressed as

$$r_k = \sum_{n=1}^{N} x_{k,n} r_{k,n} \tag{2}$$

where $\boldsymbol{x} = \{x_{k,n}\}$,and $x_{k,n}$ is the channel allocation indicator,i.e. $x_{k,n} = 1$ means user k uses subcarrier n,and $x_{k,n} = 0$ otherwise. Therefore,the time consumption for each process can be expressed as

$$\begin{aligned} t_{k,l} &= \frac{c_k \left(1 - \lambda_k\right) R_k}{f_k} \\ t_{k,u} &= \frac{\lambda_k R_k}{r_k} \\ t_{k,s} &= \frac{\lambda_k R_k c_k}{f_k^M} \end{aligned} \tag{3}$$

Since local computation and the communication between mobile devices and the edge cloud can happen simultaneously,the total time consumption for user k is decided by the longer process and can be expressed as

$$t_k = \max\{t_{k,l}, t_{k,u} + t_{k,s}\} \tag{4}$$

B. Power Analysis

Also, the energy consumption is presented separately with respect to as below.

1) Local Computing: According to ['1'],given the running frequency f_k,the energy consumption on user k during each CPU circle is $k_0 f_k^2$,where k_0 is a constant related to CPU of mobile user. Thus,we can get the energy consumption of task k for local computing as

$$E_{k,l} = k_0 c_k \left(1 - \lambda_k\right) R_k f_k^2 \tag{5}$$

2) Transmission: The energy consumption for transmitting data $\lambda_k R_k$ to MEC server is given by

$$E_{k,u} = \sum_{n=1}^{N} x_{k,n} p_{k,n} \frac{\lambda_k R_k}{r_k} \tag{6}$$

3) MEC Computing : The energy consumed to offload data to the MEC server can be denoted by

$$E_{k,s} = k_m \lambda_k c_k R_k {f_k^M}^2 \tag{7}$$

where k_m is also constant and related to the CPU of MEC. Then,the total energy energy consumption for task of user k is given by

$$E_k = E_{k,l} + E_{k,u} + E_{k,s} \tag{8}$$

IV. Problem Formulation

Under the constraints of task deadline T_k, user upload power and MEC computing capacity,the problem of minimizing the energy consumption of the total MEC system can be write as

$$\mathbf{P1}: \min_{\boldsymbol{X},\boldsymbol{\lambda},\boldsymbol{f}} \sum_{k=1}^{K} E_k \tag{9a}$$

$$\text{s.t.} \sum_{k=1}^{K} f_k^M \leq F, \tag{9b}$$

$$0 \leq f_k^M, \tag{9c}$$

$$t_k \leq T_k, \forall k \tag{9d}$$

$$\sum_{k=1}^{K} x_{k,n} \leq 1, \forall n \tag{9e}$$

$$x_{k,n} \in \{0,1\}, \forall k, n \tag{9f}$$

$$\sum_{n=1}^{N} p_{k,n} x_{k,n} \leq p_k^{max}, \tag{9g}$$

$$0 \leq \lambda_k \leq 1, \forall k \tag{9h}$$

where $\boldsymbol{X} = \{x_{k,n}\}$, $\boldsymbol{\lambda} = \{\lambda_k\}$, $\boldsymbol{f} = \{f_k^M\}$.i Constranint (9b) shows that the sum of the computing resources allocated to the user cannot exceed the computing power of the MEC. (9c) indicates that the computing resources assigned to the user are non-negative. (9d) show delay constraint. (9e) and (9f) enforce that each subcarrier can only be used by one user to avoid the multi-user interference. (9g) indicates the peak transmit power constraint on each user (9h) shows range of unloading ratio.

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

A. *Iterative optimization approach*

We adopt three-layer iterative optimization approach to solve this P1. Firstly,given the subcarrier allocation strategy $\boldsymbol{X}$,the optimal computing resource allocation $\boldsymbol{f}$ and task offload ratio $\boldsymbol{\lambda}$ are solved by the shortest time required by the user. Then,according to the optimal task offload ratio $\boldsymbol{\lambda}$ and computing resource allocation $\boldsymbol{f}$,we can get the optimal the subcarrier allocation strategy $\boldsymbol{X}$. Finally,the optimal computing resource allocation $\boldsymbol{f}$ is solved with given $\boldsymbol{X}$ and $\boldsymbol{\lambda}$. 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 $\mathcal{T}$ as all sets of possible $\boldsymbol{X}$ that satisfy (9e) and (9f),$\mathcal{R}$ as all sets of possible $\boldsymbol{\lambda}$ that satisfy $0 \leq \lambda_k \leq 1.$, and $\mathcal{F}$ as all sets of possible $\boldsymbol{\lambda}$ that satisfy (9h).

1) To solve task offload ratio: To solve $\boldsymbol{\lambda}$ with given $\boldsymbol{X}$ and $\boldsymbol{f}$,in order to achieve the lowest energy consumption with minimum time consumption, according to (9d) in P1, when $t_{k,l} = t_{k,u} + t_{k,m}$, the time consumption is the smallest, so we can get

$$\lambda_k = \frac{f_k^M r_k}{c_k f_k^M r_k + f_k r_k c_k + f_k f_k^M} \tag{10}$$

Then,the optimal $\boldsymbol{\lambda}^*$ can be obtained.

2) To solve subcarrier allocation strategy: According to $t_{k,l} = t_{k,u} + t_{k,m}$ and (4),the (9d) in P1 can be transformed into $t_{k,l} \leq T_k$,then we can get P2

$$\mathbf{P2}: \min_{\boldsymbol{X}\in\mathcal{T},\boldsymbol{\lambda}\in\mathcal{R},\boldsymbol{f}\in\mathcal{F}} \sum_{k=1}^{K} E_k \tag{11a}$$

$$\text{s.t. } t_{k,l} \leq T_k, \forall k \tag{11b}$$

$$\sum_{n=1}^{N} p_{k,n} x_{k,n} \leq p_k^{max}, \forall k \tag{11c}$$

The Lagrangian function for P2 is given by

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{X}, \lambda, \boldsymbol{f}, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \\
&\sum_{k=1}^{K}[k_0 c_k (1-\lambda_k) R_k f_k^2 + k_m c_k \lambda_k R_k (f_k^M)^2 \\
&+ \sum_{n=1}^{N} x_{k,n} p_{k,n} \frac{\lambda_k R_k}{\sum_{n=1}^{N} x_{k,n} r_{k,n}} \\
&+ \beta_k \frac{c_k (1-\lambda_k) R_k}{f_k} + \zeta_k \sum_{n=1}^{N} x_{k,n} p_{k,n}] - [\sum_{k=1}^{K} (\beta_k T_k + \zeta_k P_k^{\max})] \\
&= \sum_{k=1}^{K} \sum_{n=1}^{N} [w(\lambda_k, f_k^M) + \zeta_k x_{k,n} p_{k,n}] - [\sum_{k=1}^{K} (\beta_k T_k + \zeta_k P_k^{\max})]
\end{aligned} \tag{12}$$

where the function $w(\lambda_k, f_k^M)$ is expressed as

$$\begin{aligned}
&w(\lambda_k, f_k^M) = \\
&\frac{k_0 c_k (1-\lambda_k) R_k f_k^2}{N} + \frac{k_m c_k \lambda_k R_k (f_k^M)^2}{N} + \\
&\frac{\beta_k c_k (1-\lambda_k) R_k}{N f_k} + \frac{p_{k,n} \lambda_k R_k}{r_{k,n} N}
\end{aligned} \tag{13}$$

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

$$g(\boldsymbol{\beta}, \boldsymbol{\zeta}) = \min_{X\in\mathcal{T},\boldsymbol{\lambda}\in\mathcal{R},\boldsymbol{f}} \mathcal{L}(\boldsymbol{X}, \lambda, \boldsymbol{f}, \boldsymbol{\beta}, \boldsymbol{\zeta}) \tag{14}$$

The dual problem is thus given by $\max_{\boldsymbol{\beta},\boldsymbol{\zeta}\succeq 0} g(\alpha, \boldsymbol{\beta}, \boldsymbol{\zeta})$.For the minmization problem in (11a),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

$$\mathcal{L} = \sum_{n=1}^{N} \mathcal{L}_n - [\sum_{k=1}^{K} (\beta_k T_k + \zeta_k P_k^{\max})] \tag{15}$$

where

$$\mathcal{L}_n = \sum_{k=1}^{K} w(\lambda_k, f_k^M) + \zeta_k p_{k,n} \tag{16}$$

Thus,the subproblem is given by

$$\min_{X_n\in\mathcal{T}} \mathcal{L}_n\left(\boldsymbol{\lambda}, X_n, \boldsymbol{f}\right) \tag{17}$$

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

$$x_{k,n}^{*} = \left\{ \begin{array}{l} 1,\ \text{if } k = k^{*} = \operatorname{argmin}_k \mathcal{L}_n \\ 0,\ \text{otherwise} \end{array} \right. \tag{18}$$

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 $t_k = t_{k,l} = t_{k,u} + t_{k,s}$,So we can get subproblem P3 of P1

$$\mathbf{P3}: \ \min_{\boldsymbol{f}\in\mathcal{F}} \sum{}_{k=1}^{K} E_{k,s} \tag{19a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} f_k^M \le F \tag{19b}$$

$$t_{k,u} + t_{k,s} \le T_k, \forall k \tag{19c}$$

The Lagrangian function for P3 is given by

$$\begin{aligned} \mathcal{G}(\boldsymbol{f}, \gamma, \boldsymbol{\xi}) &= \sum_{k=1}^{K} E_{k,s} + \gamma(\sum_{k=1}^{K} f_k^M - F) \\ &- \sum_{k=1}^{K} \xi_k (f_k^M + \frac{\lambda_k R_k c_k r_k}{\lambda_k R_k - T_k r_k}) \\ &= \sum_{k=1}^{K} k_m c_k \lambda_k R_k (f_k^M)^2 + \gamma(\sum_{k=1}^{K} f_k^M - F) \\ &- \sum_{k=1}^{K} \xi_k (f_k^M + \frac{\lambda_k R_k c_k r_k}{\lambda_k R_k - T_k r_k}) \end{aligned} \tag{20}$$

where γ,$\boldsymbol{\xi} = [\xi_1, \xi_2, ..., \xi_k]^T$ are 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

$$\frac{\partial \mathcal{G}}{\partial f_k^{M*}} = 2k_m \lambda_k R_k f_k^M + \gamma - \xi_k = 0 \tag{21}$$

and we can get

$$f_k^M = [\frac{\xi_k - \gamma}{2k_m \lambda_k R_k}]^{+}, \forall k \in k \tag{22}$$

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

$$\mathcal{Q}(\gamma, \boldsymbol{\xi}) = \min_{\boldsymbol{f}\in\mathcal{F}} \mathcal{G}(\boldsymbol{f}, \gamma, \boldsymbol{\xi}) \tag{23}$$

The dual problem is thus given by $\max_{\gamma\ge 0, \boldsymbol{\xi}\succeq 0} \mathcal{Q}(\lambda, \boldsymbol{\xi})$. With the fixed $\boldsymbol{X}$ and $\boldsymbol{f}$,the optimal λ_{k*} can be obtianed by(8), 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 $\gamma \ge 0, \boldsymbol{\xi} \succeq 0$.

We can easily prove that the dual problem is a convex one. Thus,we can use the subgradient method to solve it. The detailed Algorithm in the following.

Algorithm 1 Proposed Iterative Algorithm

```
initialize:
  • Set $\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \gamma, \boldsymbol{\xi}, \mathcal{I}_{max}, \epsilon$.
  • Set i = 0
repeat
    repeat
        Solve task offload ratio $\boldsymbol{\lambda}$ according to (10)
        Determine subcarrier allocation $\boldsymbol{X}$ according to (17) and compute L according to (12).
        Allocate computing resource $\boldsymbol{f}$ accoring to (21).
    until Lagrangian function converges.
    Update $\boldsymbol{\beta}, \boldsymbol{\zeta}, \gamma, \boldsymbol{\xi}$ using the subgradient method.
    i = i + 1
    if $\|\beta(i+1) - \beta(i)\| \le \epsilon$ and $\|\zeta(i+1) - \zeta(i)\| \le \epsilon$ and $\|\gamma(i+1) - \gamma(i)\| \le \epsilon$ and $\|\xi(i+1) - \xi(i)\| \le \epsilon$ then
        break.
until $i > \mathcal{I}_{max}$
```