Energy-efficient Resource Allocation in Mobile Edge Computing with Multiple Relays

Xiang Li, Rongfei Fan, Han Hu, and Ning Zhang

Abstract—In recent years, mobile edge computing (MEC) has attracted tremendous research thanks to its advantage in handling computation intensive latency critical tasks. To overtake the bad channel condition in the process of task offloading, multiple-relay assisted MEC system is considered in this paper. In specific, three cases including TDMA scenario, FDMA scenario in decode-and-forward (DF) mode and amplify-and-forward (AF) mode are investigated. The target is to minimize the overall energy consumption of mobile user and relays by jointly optimizing offloading data amount, transmit power and slot duration (in TDMA, or bandwidth allocation in FDMA, or amplitude gain in AF). In the scenario of TDMA, we show the associated problem is convex and solve it in a easier way through the manner of bi-level optimization. In the upper level, the optimal data amount for offloading is acquired, which corresponds to a simpler convex optimization problem, while in the lower level, the optimal solution of the rest of variable is found via KKT conditions. In the scenario of FDMA, the associated optimization problem is non-convex. Global optimal solution is found with the help of bi-level optimization and monotonic programming. For AF mode, bi-level optimization is also utilized in which neither of the two levels is convex. To this end, geometric programming and successive convex approximation (SCA) is used to find the convergent solution of the lower level while monotonic programming is adopted in the upper level. Numerical results proves the effectiveness of the proposed strategies under various scenarios investigated in this paper.

I. INTRODUCTION

Recent years have witnessed the rapidly growing demand for complex computation in a wide range of emerging mobile applications, such as image recognition and virtual reality [1]. These applications bring about challenges for mobile devices on the issue of low battery life and long latency. To tackle this problem, mobile edge computing (MEC) is widely perceived as a promising technology [2]. In the MEC system, the base station (BS) or access point (AP) has abundant computation capacity compared with mobile devices. Therefore, the mobile devices can offload computation task to the BS or the AP, which is called edge server in the following context. On one hand, with the help of edge server, many computation-intensive tasks become available on the mobile device. On the other hand, providing computation resource to the edge server is of vital importance, to reduce cost functions formulated by energy consumption and complete latency. To be specific, the cost function is mainly affected by partition for local computing and offloading [3], frequency of CPU for local computation [4] and transmit power of mobile device in the offloading process [5], [6]. When the task is offloaded to the edge server, the duration of data offloading and edge computing need to be optimized, in order to reduce the total latency [7].

In general, research on MEC considers two different offloading modes: binary and partial offloading. In binary offloading mode [4], [8], [9], the task requires to be locally computed or offloaded as a whole. In partial offloading [5], [10], [11], however, the task is partitioned and executed in different place. The latter takes advantage of parallel computation between the mobile device and the edge server, thus can reduce the complete latency in most cases. Furthermore, with the application of virtual machines (VMs) on the edge server [10], partial offloading is more flexible when complicated communication scenario is considered in the system.

In the category of partial offloading, plenty of research have been devoted to reveal the optimal radio-and-computation resource allocation. As an example of early research in this field, reference [5] studied the minimization of energy consumption and latency when one mobile device offload its task to the BS. Typically, The proportion of offloading, frequency of local CPU and with transmit power are optimized. To provide computation resource to multiple mobile devices under one edge server, later research extend the communication model into multilink. In specific, to guarantee successful decoding, different communication protocols when offloading the task is considered by various research. In [12], users work under time-division multiple access (TDMA) or orthogonal frequency-division multiple access (OFDMA). In TDMA scenario, the transmit slot of different users are optimized along with offloaded data amount. In OFDMA scenario, bandwidth for the users are allocated. In [13], task uploading and result downloading are implemented with nonorthogonal multiple access (NOMA). Thus, decoding order and the overall transmit duration are optimized to reduce complete latency. Unlike the previous two research, [7] proposed a wireless powered MEC system, in which duration of downlink energy transfer is studied, with users’ slot allocation in uploading the data under TDMA protocol.

Despite abundant research on the multiuser MEC system,
few of them has taken into account the cooperation between mobile devices under the same BS. In extreme situations that the channel is not ideal due to long distance or deep fading, it is helpful to provide extra computation or communication resources with idle mobile devices. To this end, [14] considers one mobile user assisted by one helper, on both communication and computation. The user first upload the task to the helper for cooperative computation, then broadcast the task for edge computation to the helper and AP. After decoding the latter task, the helper send it to the AP. With this two-hop structure, the helper can enhance the service on the account of the channel condition between the helper and the AP. In addition, reference [15] designs a slotted harvest-then-offload structure, where a single helper who has task to offload as well, assists the uploading of one mobile user to the AP. After downlink energy transfer, the mobile user broadcast the task for offloading, followed by retransmission by the helper. Finally, uploading of the helper is carried out.

In the above cooperative MEC systems, only single assist device is considered. Thanks to the density of mobile devices in practical 5G system, aided communication of multiple relays are frequent, and can provide higher channel capacity. To this end, we investigate a partial offloading scenario in which the mobile device is aided by multiple relays working under decode-and-forward (DF) mode and amplify-and-forward (AF) mode. The problem of energy consumption minimization subject to the latency requirement and channel capacity is studied.

In DF mode, we consider two cases separately:

- **DF-TDMA**: the mobile device transmits to the relays in different time slots. In this case, a convex problem is formulated with respect to offloaded data amount, allocated time slot for different relays and transmit power of the mobile device and relays. To get more insight of the problem structure and reduce computation complexity, a bilevel optimization method is utilized. In the upper level, the optimal data amount for offloading is acquired, while in the lower level, other variables are optimized. The lower level problem is convex, and transformed into a linear programming with KKT conditions. The upper level problem is as well proved to be a single variable convex problem.

  With the above methods, global optimal is obtained with lower complexity, compared with directly using traditional numerical methods.

- **DF-FDMA**: the mobile device transmits to the relays simultaneously in different subbands. In this case, a nonconvex problem is formulated with respect to offloaded data amount, overall transmit duration, allocated bandwidth for different relays and transmit power of the mobile device and relays. Utilizing bilevel method, in the upper level, the optimal data amount for offloading is acquired, while in the lower level, other variables are optimized. The lower level problem is convex, and transformed into a linear programming with KKT conditions. In the upper level, we form the problem into a monotonic programming, and apply Polyblock Algorithm to solve it. With the above methods, global optimal is obtained.

The rest of the paper is organized as follows. System model is presented in Section II. For three different relay modes, optimization problems are formulated in Section III. Whereafter, we solve the problems respectively in Section IV. In Section V, numerical results are followed, which prove the effectiveness of our proposed algorithms. Finally, conclusion is presented in Section VI.

**II. SYSTEM MODEL**

Consider a MEC system with one mobile device and one BS, as shown in Fig. 1. The mobile device has one computation task to complete, which is denoted as $T$. Specifically, the task $T$, can be described by a three-tuple $(T, D, L)$. In this tuple, $T$ indicates the maximal delay the mobile device can tolerate. $D$ is the total amount of input data to process in order to complete the task $T$, which is in unit of nat for simplicity of presentation in the following. $L$ is the number of CPU cycles required for computing unit data nat. Hence $LD$ represents the total amount of CPU cycles to process for completing task $T$.

The task $T$ is separable, which means that it can be calculated at two or multiple sites simultaneously. To complete task $T$ in an energy-efficient way, the mobile device can offload part of the data for edge computing in the BS, while the rest of data is left for local computing. Suppose the amount of data to offload to the BS is $D_e$ and the amount of data for local computing is $D_l$, then there is

$$D = D_e + D_l. \quad (1)$$
For the data offloaded to the BS, after the uplink transmission, the BS will first process it and then return computational results to the mobile device. Since the computational results are always of small data size and the BS (who is generally rich in power supply) can easily achieve high data rate, the time for feeding back the computational results to the mobile device is omitted [9].

Due to blockage or deep fading, there is no direct link between the mobile device and BS, data offloading is assisted by \( N \) relay nodes. These relay nodes are denoted as \( R_n \) for \( n \in \mathcal{N} \) where \( \mathcal{N} \triangleq \{ 1, 2, ..., N \} \). Denote the channel gain between the mobile device and relay \( R_n \) as \( h_n \), and the channel gain between relay \( R_n \) and BS as \( g_n \). Both \( h_n \) and \( g_n \) for \( n \in \mathcal{N} \) are block faded, which means that these channel gains are stable within the duration of one fading block and varies randomly and independently in different fading blocks. Note that the value of \( h_n \) and \( g_n \) for \( n \in \mathcal{N} \) can be measured at the beginning of one fading block, which generally leads to negligible time overhead. Suppose the system bandwidth is \( W \), which is no larger than coherence bandwidth. Hence both \( h_n \) and \( g_n \) for \( n \in \mathcal{N} \) are stable within the system bandwidth \( W \).

For data relaying between the mobile device and BS, the relays work in a half-duplex manner. In the first phase, the mobile device transmit the data for edge computing to the relays, whereas in the second phase, the relays transmit to the BS. In the first and the second phase, the transmit duration are the same and denoted as \( t \). Specifically, three relaying modes are investigated.

- **DF-TDMA** In this mode, DF is utilized on every relay node, which indicates that every relay node first decodes the received signal from the mobile device and then forward the decoded information to the BS. To be interference free, these \( N \) relay nodes are orthogonal in time while occupying the common bandwidth \( W \), as shown in Fig. 2. For brevity, this mode is also called as TDMA in the following. For ease of implementation, for each relay \( R_n \) and \( n \in \mathcal{N} \), its transmit duration in the first and second phase are the same, which is denoted as \( t \). Therefore, \( \sum_{n=1}^{N} t_n = t \). For \( n \in \mathcal{N} \), denote the transmit power from mobile device to relay \( R_n \) as \( P_n \) and the transmit power from relay \( R_n \) to BS as \( Q_n \), respectively. The relays work on a fixed bandwidth \( W \), and the power spectral density (PSD) of background noise is \( \delta^2 \). Applying Shannon capacity, the amount of data transmitted through relay \( R_n \) is

\[
D_n^F = \min \left( t_n W \ln \left( 1 + \frac{P_n h_n}{\delta^2 W} \right) \right),
\]

\[
t_n W \ln \left( 1 + \frac{Q_n g_n}{\delta^2 W} \right) .
\]

- **DF-FDMA** In this mode, DF is also utilized on every relay node. To be interference free, these \( N \) relay nodes work on different subbands while transmitting with common duration \( t \), as shown in Fig. 3. For brevity, this mode is also called as FDMA in the following. Specifically, the bandwidth taken up by these \( N \) relays should be no larger than system bandwidth \( W \). In other words, \( \sum_{n=1}^{N} w_i \leq W \). For \( n \in \mathcal{N} \), similar to TDMA mode, denote the transmit power from mobile device to relay \( R_n \) as \( P_n \) and the transmit power from Relay \( R_n \) to BS as \( Q_n \), respectively. Relay \( R_n \) is assigned a subband \( w_n \), and the PSD of background noise is also \( \delta^2 \). Applying Shannon capacity, the amount of data transmitted through relay \( R_n \) is

\[
D_n^F = \min \left( t_w W \ln \left( 1 + \frac{P_n h_n}{\delta^2 W} \right) \right),
\]

\[
t_w W \ln \left( 1 + \frac{Q_n g_n}{\delta^2 w_n} \right) .
\]

- **AF** In this mode, every relay node will directly amplify the received signal transmitted from the mobile device. Similar to [16], in the first phase, the mobile device transmit the offloaded data to the relays, whereas in the second phase, the relays amplify the received signal and transmit to the BS. Due to blockage and long distance between mobile device and BS, the direct link in between is neglected. Denoting the amplitude gain of relay \( R_n \) as \( \beta_n \), let the unit power transmit signal in the mobile device be \( s \) and transmit power be \( P \), the received signal at relay \( R_n \) is

\[
m_n = \sqrt{h_n P s + N_0}.
\]

In this expression, the second term, \( N_0 \), refers to the noise at the receiver of the relay \( R_n \), \( n \in \mathcal{N} \). After receiving the signal from mobile device, the relay nodes amplify and transmit it directly, with amplitude gain \( \beta_n \), \( n \in \mathcal{N} \).
Similar to DF modes, denote PDF of the noise as $\delta^2$ and the bandwidth as $W$. The transmitted signal from relay $R_n$ is $\beta_n \left( \sqrt{k_n} P_s N_0 \right)$, whose power is apparently $\beta_n^2 \left( P h_n + \delta^2 W \right)$. Therefore, the received signal at the BS is
\[
 r = \sum_{n=1}^{N} \sqrt{h_n g_n} \beta_n \sqrt{P_s} + \sum_{n=1}^{N} \sqrt{g_n} \beta_n N_0 + N_0, \tag{5}
\]
in which the last term is the noise at the receiver of the BS. In (5), the first term is the signal, while the second and third term is additional noise due to the first and second phase of AF transmission. Result from the first term in (5), the signal power is represented by
\[
P_s^A = P \left( \sum_{n=1}^{N} \sqrt{h_n g_n} \beta_n \right)^2. \tag{6}
\]
Since noise at the receiver of the relays and the BS are independent and identically distributed, the total noise power at decoder of the BS can be expressed as
\[
P_n^A = \delta^2 W + \delta^2 W \sum_{n=1}^{N} g_n \beta_n^2, \tag{7}
\]
in which the first term refers to noise at the receiver of BS, and the second term refers to noise at the receiver of the relays amplified after going through the relays. With the above signal and noise power, the amount of data transmitted through all the relays in AF mode is
\[
D^A = tW \ln \left(1 + \frac{P_s^A}{P_n^A}\right). \tag{8}
\]
In the following, computation model of our work is presented.

- **Computing at Local** Since local computing should be accomplished before deadline, letting $f_{l,m}$ be the frequency of the $m$-th cycle at local CPU, we have:
\[
L(D-d) \sum_{m=1}^{\frac{T}{f_{l,m}}} \frac{1}{f_{l,m}} \leq T. \tag{9}
\]
Normally, the computation energy consumption of CPU per cycle is proportional to the CPU frequency square $[4]$, which can be expressed as:
\[
E_c = \sum_{m=1}^{L(D-d)} \kappa f_{m,l}^2, \tag{10}
\]
where $\kappa$ is decided by chip structure. Previous research $[5]$ has proved that, with given latency constraint, setting the CPU frequencies identical for each cycle achieves optimal energy consumption, i.e. $f_{n,l} = L(D-d)/T$. Thus, referring to (10), the total energy consumption is
\[
E_c = \frac{\kappa L^3(D-d)^3}{T^2}. \tag{11}
\]

- **Computing at Base Station** After receiving the offloaded task, the BS starts to compute for mobile device with a fixed CPU frequency. Assume the computation capacity of the BS, i.e. maximum CPU frequency is $f_{\text{max}}$, and $2t$ is total duration of transmission in the first and second phases. The edge CPU frequency is upper bounded by $f_{\text{max}}$, i.e.
\[
\frac{L_d}{T - 2t} \leq f_{\text{max}}, \tag{12}
\]
Note that the above requirement implies $L_d \leq f_{\text{max}} T$, which serve as an upper bound of the offloaded data. In fact, CPU capacity $f_{\text{max}}$ of the BS is usually large, thus we assume the inequality always holds.

### III. Problem Formulation

In this section, for TDMA, FDMA and AF, optimization problems are formulated on the basis of system models in Section II. Our objective is to minimize the overall energy consumption of the mobile device and relays by adjusting the amount of offloaded data, transmit duration, transmit power and resource allocation of DF relays (or amplitude gain of AF relays), while respecting the latency requirements.

For TDMA case, the transmit duration satisfies $\sum_{n=1}^{N} t_n = t$. Substitute the equation into (12), the latency constraint for data offloading is $2 \sum_{n=1}^{N} t_n \leq T - \frac{L_d}{f_{\text{max}}}$. Jointly considering this with (2) and (11), the problem is formulated as:

**Problem 1:**
\[
\begin{align*}
\min_{d, \{t_n \mid n \in \mathcal{N}\}, \{P_n \mid n \in \mathcal{N}\}, \{Q_n \mid n \in \mathcal{N}\}} & \sum_{n=1}^{N} (P_n t_n + Q_n t_n) + \kappa L^3(D-d)^3 \frac{1}{T^2} \\
\text{s.t.} & \sum_{n=1}^{N} D_n^T \leq \frac{L_d}{f_{\text{max}}} \tag{13a} \\
& 2 \sum_{n=1}^{N} t_n \leq T - \frac{L_d}{f_{\text{max}}} \tag{13b} \\
& 0 \leq d \leq D, t_n \geq 0, P_n \geq 0, Q_n \geq 0, \forall n \in \mathcal{N}. \tag{13c}
\end{align*}
\]
In this problem, $P_n t_n$ and $Q_n t_n$ are energy consumption of the mobile device and the relay, respectively, for the transmission through relay $R_n, \forall n \in \mathcal{N}$. The last term in the objective function, as in (11), is energy consumption of local computation.

For FDMA case, the relays should meet system bandwidth constraint $\sum_{n=1}^{N} w_n \leq W$. According to (5), (11) and (12), the problem is:
Problem 2:
\[
\min_{d,t, \{P_n|n \in N\}, \{Q_n|n \in N\}, \{w_n|n \in N\}} \sum_{n=1}^{N} (P_n t + Q_n t) + \frac{\kappa L^3(D - d)^3}{T^2} \\
\text{s.t.} \quad d \leq \sum_{n=1}^{N} D_n^F 
\]
(14a)
\[
2t \leq T - \frac{L d}{f_{\text{max}}} 
\]
(14b)
\[
\sum_{n=1}^{N} w_i \leq W 
\]
(14c)
\[
0 \leq d \leq D, t \geq 0, \\
P_n \geq 0, Q_n \geq 0, w_n \geq 0, \forall n \in N. 
\]
(14d)

Similar to Problem 1, \(P_n t\) and \(Q_n t\) are energy consumption of the mobile device and the relay, respectively, for the transmission through relay \(R_n\), \(\forall n \in N\).

In AF scenario, the data offloading process is no longer split into different resource blocks, and the amplitude gain for relays remain to be solved. Based on equation (6), (7), (8), (11) and (12), the problem is derived as:

Problem 3:
\[
\min_{d,t,P, \{\beta_n|n \in N\}} \sum_{n=1}^{N} \beta_n^2 \left( P h_n + \delta^2 \right) t + \frac{\kappa L^3(D - d)^3}{T^2} \\
\text{s.t.} \quad d \leq t W \ln \left( 1 + \frac{P \left( \sum_{n=1}^{N} \sqrt{h_n g_n^2 \beta_n^2} \right)^2}{\delta^2 \left( 1 + \sum_{n=1}^{N} g_n \beta_n^2 \right)} \right) 
\]
(15a)
\[
2t \leq T - \frac{L d}{f_{\text{max}}} 
\]
(15b)
\[
0 \leq d \leq D, t \geq 0, P_n \geq 0, \beta_n \geq 0, \forall n \in N. 
\]
(15c)

In Problem 3, the objective function consists of three terms, in which the first is transmit energy consumption of the mobile device, the second is the energy consumption of relays when amplifying the received signal and noise, and the third is energy consumption of local computation.

IV. Optimal Solution

In Section III, optimization problem concerning the resource allocation in TDMA mode, FDMA mode and AF mode are introduced. In order to conserve energy while finishing the task on time, in this section, Problem 1, Problem 2 and Problem 5 for three different modes are solved, successively.

A. Solution for TDMA case

Observing (2), for the optimal value of \(P_n\) and \(Q_n\) in Problem 1
\[
P_n^* h_n = Q_n^* g_n. 
\]
(16)

Note that if \(P_n h_n \leq Q_n g_n\), i.e. \(t_n W \ln(1 + \frac{P_n h_n}{W}) \leq t_n W \ln(1 + \frac{Q_n g_n}{W})\), the relay \(R_n\) can achieve the same throughput with a smaller transmit power \(Q_n^* g_n\) that satisfies \(P_n^* h_n = Q_n^* g_n\), which brings about smaller objective value. The case for \(P_n^* h_n \geq Q_n^* g_n\) is similar.

Substitute the equality of (16) into the objective function and constraint (13a) in Problem 1, the problem is transformed into:

Problem 4:
\[
\min_{d,t, \{P_n|n \in N\}, \{E_n|n \in N\}} \sum_{n=1}^{N} P_n t_n \left( 1 + \frac{h_n}{g_n} \right) + \frac{\kappa L^3(D - d)^3}{T^2} \\
\text{s.t.} \quad d \leq \sum_{n=1}^{N} t_n W \ln \left( 1 + \frac{P_n h_n}{g_n \delta^2 W} \right) 
\]
(17a)
\[
2 \sum_{n=1}^{N} t_n \leq T - \frac{L d}{f_{\text{max}}} 
\]
(17b)
\[
0 \leq d \leq D, t_n \geq 0, P_n \geq 0, \forall n \in N. 
\]
(17c)

Problem 4 is nonconvex due to variable coupling in the objective function and constraint (17a). Replace \(P_n\) with new variable \(E_n = P_n t_n\). Problem 4 can be converted into the following form:

Problem 5:
\[
\min_{d, \{t_n|n \in N\}, \{E_n|n \in N\}} \sum_{n=1}^{N} E_n \left( 1 + \frac{h_n}{g_n} \right) + \frac{\kappa L^3(D - d)^3}{T^2} \\
\text{s.t.} \quad d \leq \sum_{n=1}^{N} t_n W \ln \left( 1 + \frac{E_n h_n}{t_n \delta^2 W} \right) 
\]
(18a)
\[
2 \sum_{n=1}^{N} t_n \leq T - \frac{L d}{f_{\text{max}}} 
\]
(18b)
\[
0 \leq d \leq D, t_n \geq 0, E_n \geq 0, \forall n \in N. 
\]
(18c)

In Problem 5, the objective function is linear. The right-hand side of constraint (18a) is perspective function of a concave function \(18\) with \(E_n\) over \(t_n\). Problem 5 becomes convex and can be solved by traditional numerical optimization methods such as interior point methods. However, traditional methods converge to optimal solution by iteration and provide little insight of the problem structure. In our work, we will use bilevel optimization and Karush-Kuhn-Tucker(KKT) conditions to analyze the properties of Problem 5. On the basis of these properties, a fast algorithm is proposed.

First, a two-level structure is formed, by defining function of \(d\) in Problem 6

Problem 6:
\[
U(d) = \min_{\{E_n|n \in N\}} \sum_{n=1}^{N} E_n \left( 1 + \frac{h_n}{g_n} \right) \\
\text{s.t.} \quad d \leq \sum_{n=1}^{N} t_n W \ln \left( 1 + \frac{E_n h_n}{t_n \delta^2 W} \right) 
\]
(19a)
\[
2 \sum_{n=1}^{N} t_n \leq T - \frac{L d}{f_{\text{max}}} 
\]
(19b)
\[
E_n \geq 0, t_n \geq 0, \forall n \in N. 
\]
(19c)

1To make sure the equivalence before and after variable substitution, we emphasize that \(P_n = 0\) when \(E_n = 0\) or \(t_n = 0\). For FDMA case, the definition is similar.
In Problem 5, given $d$ fixed, the transmit energy consumption and duration is jointly optimized to obtain function $U(d)$. Thus Problem 5 is equivalent to Problem 7 in the following.

**Problem 7:**

$$\min_d \quad U(d) + \frac{\kappa L^3(D - d)^3}{T^2}$$

s.t. \quad 0 \leq d \leq D. \quad (20a)$$

The lower level Problem 6 is convex and satisfies Slater’s conditions. The KKT conditions, which are sufficient and necessary for the optimal solution, are listed below:

$$1 + \frac{h_n}{g_n} - \frac{\mu n h_n}{t_n \delta^2 W} + E_n h_n^- - \zeta_n = 0, \quad \forall n \in \mathcal{N} \quad (21a)$$

$$2\lambda - \mu W \left( \ln \left( 1 + \frac{E_n h_n}{t_n \delta^2 W} \right) - \frac{E_n h_n}{t_n \delta^2 W} \right) - \eta_n = 0, \quad \forall n \in \mathcal{N} \quad (21b)$$

$$\mu \left( d - \sum_{n=1}^{N} t_n W \ln \left( 1 + \frac{E_n h_n}{t_n \delta^2 W} \right) \right) = 0 \quad (21c)$$

$$\lambda \left( 2 \sum_{n=1}^{N} t_n + \frac{Ld}{f_{\max}} - T \right) = 0 \quad (21d)$$

$$\zeta_n E_n = 0, \quad \forall n \in \mathcal{N} \quad (21e)$$

$$\eta_n t_n = 0, \quad \forall n \in \mathcal{N} \quad (21f)$$

$$d \leq \sum_{n=1}^{N} t_n W \ln \left( 1 + \frac{E_n h_n}{t_n \delta^2 W} \right) \quad \forall n \in \mathcal{N} \quad (21g)$$

in which $\mu, \lambda, \zeta_n$ and $\eta_n$ are non-negative Lagrange multipliers associated with constraints (21b), (21h), (21i) and (21j). (21a), (21b) are gradient vanishing condition of Lagrangian with respect to $E_n, t_n$, respectively.

Letting $\text{SNR}_n = \frac{E_n h_n}{\delta^2 W t_n}$, we have the Lemma 1 to characterize the signal-to-noise ratio for the relay nodes.

**Lemma 1:** For $i, j \in \mathcal{N}$ that satisfies $E_i > 0, E_j > 0, t_i > 0$ and $t_j > 0$, \[ \begin{align*} \text{SNR}_i = \text{SNR}_j = \text{SNR}. \end{align*} \] (22)

**Proof:** For $E_n > 0$ and $t_n > 0$, $\zeta_n$ and $\eta_n$ are zero, it can be derived from (21b) that

$$\ln \left( 1 + \text{SNR}_n \right) - \frac{\text{SNR}_n}{1 + \text{SNR}_n} = \frac{2\lambda}{\mu W}, \quad \forall n \in \mathcal{N}. \quad (23)$$

Looking into function $\theta(x) = \ln(1 + x) - \frac{x}{1 + x}$, it is strictly increasing for $x > 0$ and have inverse function. Therefore, we have $\text{SNR}_n = \theta^{-1}\left( \frac{2\lambda}{\mu W} \right)$, in which $\lambda$ and $\mu$ are global Lagrangian multipliers and remains the same for $\forall n \in \mathcal{N}$.

Letting $\text{SNR} = \theta^{-1}\left( \frac{2\lambda}{\mu W} \right)$ completes the proof.

With the aid of Lemma 1 the following equation is derived.

$$\text{SNR}_n = \frac{E_n h_n}{t_n \delta^2 W} = \frac{\sum_{n=1}^{N} E_n h_n}{\delta^2 W \sum_{n=1}^{N} t_n}. \quad (24)$$

**Lemma 2:** For optimal solution of Problem 6 the equality of constraint (19a) and (19b) hold.

**Proof:** For the equality of constraints (19a), we prove by contradiction. Note that the right-hand side of (19a) is monotonically increasing for both $E_n$ and $t_n, \forall n \in \mathcal{N}$, if $\{E_n^t\}$ and $\{t_n^t\}$ are optimal solution for the problem and lead to $d < \sum_{n=1}^{N} t_n^t W \ln \left( 1 + \frac{E_n^t h_n}{\delta^2 W t_n^t} \right)$, one can always reduce objective value by randomly choosing $i \in \mathcal{N}$ and replacing $E_i^t$ with $E_i^\dagger$ that satisfies $d = \sum_{n=1, n \neq i}^{N} t_n^t W \ln \left( 1 + \frac{E_i^\dagger h_n}{\delta^2 W t_n^t} \right) + t_i W \ln \left( 1 + \frac{E_i^\dagger h_n}{\delta^2 W t_i^t} \right)$. This contradicts to the optimality of $\{E_n^t\}$ and $\{t_n^t\}$, thus the proof is complete.

Due to the equality of (19a), we assume the optimal solution $\{E_i^t\}$ and $\{t_i^t\}$ leads to $d = \sum_{n=1}^{N} t_n^t W \ln \left( 1 + \frac{E_i^t h_n}{\delta^2 W t_i^t} \right)$ and $2 \sum_{n=1}^{N} t_n^t < T - \frac{Ld}{f_{\max}}$. Here, we can randomly choose $i \in \mathcal{N}$ and replace $t_i^t$ with $t_i^\dagger$ that satisfies $2 \left( \sum_{n=1, n \neq i}^{N} t_n^t + t_i^t \right) = T - \frac{Ld}{f_{\max}}$. It is obvious that $t_i^t > t_i^\dagger$, therefore $E_i^t < E_i^\dagger$, which cause the decrease of objective value and contradict to the optimality of $\{E_i^t\}$ and $\{t_i^t\}$. Therefore, equality of (19b) is proven.

This completes the proof.

Substitute (24) into the equality of constraints (19a) and (19b). Problem 6 becomes:

**Problem 8:**

$$\begin{align*} &\min_{\{E_n \mid n \in \mathcal{N}\}} \sum_{n=1}^{N} E_n \left( 1 + \frac{h_n}{g_n} \right) \\
&\text{s.t.} \quad \sum_{n=1}^{N} E_n h_n \geq \frac{1}{2} \delta^2 W \left( T - \frac{Ld}{f_{\max}} \right) \left( e^{\frac{2d}{T - \frac{Ld}{f_{\max}}}} - 1 \right) \quad (25a) \\
&\quad E_n \geq 0, \quad \forall n \in \mathcal{N}. \quad (25b) \end{align*}$$

This is a linear programming with respect to $\{E_n\}$ and can be solved by numerical methods.

Next, we look into the upper level Problem 7 of $d$.

**Lemma 3:** Problem 7 is a convex problem.

**Proof:** To prove the convexity of Problem 7 we need to assure that $U(d)$ is a convex function of $d$. In constraint (25a), we define the right-hand side as a function:

$$\sigma(d) = \frac{1}{2} \delta^2 W \left( T - \frac{Ld}{f_{\max}} \right) \left( e^{\frac{2d}{T - \frac{Ld}{f_{\max}}}} - 1 \right) \quad (26)$$

whose second order derivative is

$$\sigma''(d) = \frac{2\delta^2 f_{\max} T^2}{W (f_{\max} T - Ld)^3} e^{\frac{2d}{f_{\max} T - Ld}} \quad (27)$$

Since $f_{\max} T \geq Ld$, (27) is positive. Therefore, $\sigma(d)$ is a convex function of $d$.

Suppose $E_n^t$ and $E_n^\dagger$ are optimal solution in Problem 8 for given $d^t$ and $d^\dagger$, i.e. $U(d^t) = \sum_{n=1}^{N} E_n^t \left( 1 + \frac{h_n}{g_n} \right)$ and $U(d^\dagger) = \sum_{n=1}^{N} E_n^\dagger \left( 1 + \frac{h_n}{g_n} \right)$. To satisfy constraint (25a), \[ \sum_{n=1}^{N} E_n^\dagger h_n \geq \sigma(d^\dagger) \text{ and } \sum_{n=1}^{N} E_n^t h_n \geq \sigma(d^t) \].
For $\epsilon \in [0, 1]$, we have
\[
\epsilon \sum_{n=1}^{N} E_n^d h_n + (1 - \epsilon) \sum_{n=1}^{N} E_n^t h_n \\
\geq \epsilon \sigma(d^d) + (1 - \epsilon) \sigma(d^t) \\
\geq \sigma(e d^d + (1 - \epsilon) d^t)
\] (28)
which means $\{E_n\} = \{\epsilon E_n^d + (1 - \epsilon) E_n^t\}$ is a feasible solution for $d = e d^d + (1 - \epsilon) d^t$, and its objective value
\[
\epsilon E_n^d (1 + \frac{h_n}{g_n}) + (1 - \epsilon) E_n^t (1 + \frac{h_n}{g_n}) = \epsilon U(d^d) + (1 - \epsilon) U(d^t) \\
\geq U(\epsilon d^d + (1 - \epsilon) d^t)
\] (29)
Thus, $U(d)$ is a convex function, and Problem 7 is a convex problem. The proof is complete.

In TDMA mode, after variable substitution, the optimization problem is transformed into convex with respect to $d$, $E_n$ and $t_n$ in Problem 5 and can be solved by traditional methods. However, by using equality constraints in Lemma 1 and 2 the problem is reduced into optimization of $d$ and $E_n$ in Problem 7 and 8 respectively. Consider interior point methods [18] to solve convex problem and linear programming problem. The computation complexity is $O((2n)^{3.5})$ for directly solving Problem 5. In Problem 8 $t_n$ is no longer an optimization variable and the computation complexity is reduced to $O(n^{3.5})$. Therefore, for TDMA case, by mathematical analysis of Problem 5 we develop faster algorithm than traditional methods.

B. Solution for FDMA case

Observing (3), for the optimal value of $P_n$ and $Q_n$ in Problem 2

\[
P_n^* h_n = Q_n^* g_n.
\] (30)
This is similar to the case of TDMA and the explanation is left out here. After substitution of (30), Problem 2 is transformed into:

Problem 9:

\[
\min_{d,t,\{P_n|n \in \mathcal{N} \}} \sum_{n=1}^{N} P_n t_n \left( 1 + \frac{h_n}{g_n} \right) + \frac{\kappa L^3(D-d)^3}{T^2}
\]
\[\text{s.t.} \quad d \leq \sum_{n=1}^{N} t w_n \ln \left( 1 + \frac{P_n h_n}{\delta^2 w_n} \right) \quad (31a)
\]
\[2t \leq T - \frac{Ld}{f_{max}} \quad (31b)
\]
\[\sum_{n=1}^{N} w_i \leq W \quad (31c)
\]
\[0 \leq d \leq D, t \geq 0, P_n \geq 0, w_n \geq 0, \forall n \in \mathcal{N}. \quad (31d)
\]

In Problem 9, the objective function and constraint (31a) are nonconvex due to variable coupling. In the following, by denoting $E_n = P_n t_n$, we have the Lemma 4 to simplify this problem.

Lemma 4: For optimal solution of Problem 9, the equality of constraint (31a), (31b), and (31c) hold.

Proof: Substitute $E_n = P_n t_n$ into Problem 9 (31a) becomes
\[d \leq \sum_{n=1}^{N} t w_n \ln \left( 1 + \frac{E_n h_n}{\delta^2 w_n} \right), \quad \text{and the objective function is} \sum_{n=1}^{N} E_n \left( 1 + \frac{h_n}{g_n} \right) + \frac{\kappa L^3(D-d)^3}{T^2}. \]
Next we prove the lemma by contradiction. Suppose $d^i$, $t^i$, $\{E_n^i\}$ and $\{w_n^i\}$ are optimal solution and $d^i < \sum_{n=1}^{N} t^i w_n^i \ln \left( 1 + \frac{E_n^i h_n}{\delta^2 w_n^i} \right)$, then we can randomly choose $i \in \mathcal{N}$ and replace $E_n^i$ with $E_n^\dagger$ that satisfies $d = \sum_{n=1,n\neq i}^{N} t^i w_n^i \ln \left( 1 + \frac{E_n^i h_n}{\delta^2 w_n^i} \right) + t^i w_n^i \ln \left( 1 + \frac{E_n^\dagger h_n}{\delta^2 w_n^\dagger} \right)$, which is a feasible solution and obviously results in smaller objective value. This contradicts to the assumption of optimal solution. Thus equality of (31a) is proven.

Next we prove the equality of (31b). In equation $d = \sum_{n=1}^{N} t w_n \ln \left( 1 + \frac{E_n h_n}{\delta^2 w_n} \right)$, the right-hand side is a monotonically increasing function with respect to $E_n$, $w_n$ and $t$. For given $d$ and $\{w_n\}$, to reduce the objective value implies to reduce $\{E_n\}$, and thus to make $t$ as large as possible, which, leads to the equality of constraint (31b). Proof for equality of (31c) is similar to that of (31b) and omitted here.

This completes the proof.

Based on Lemma 4 substitute the equation of (31b) into Problem 9. Given $d$ fixed, the lower level Problem 10 is derived:

Problem 10:

\[
V(d) = \max_{\{P_n|n \in \mathcal{N}\}} \min_{\{w_n|n \in \mathcal{N}\}} \sum_{n=1}^{N} P_n \left( 1 + \frac{h_n}{g_n} \right) \quad (32a)
\]
\[\text{s.t.} \quad \frac{2d}{T - \frac{Ld}{f_{max}}} \leq \sum_{n=1}^{N} w_i \ln \left( 1 + \frac{P_n h_n}{\delta^2 w_n} \right) \quad (32b)
\]
\[\sum_{n=1}^{N} w_i \leq W \quad (32c)
\]
\[P_n \geq 0, w_n \geq 0, \forall n \in \mathcal{N}. \quad (32d)
\]

In Problem 10 the transmit power and allocated bandwidth is jointly optimized to obtain function $V(d)$, thus Problem 9 is equivalent to Problem 11 in the following.

Problem 11:

\[
\min_d \quad \frac{T - \frac{Ld}{f_{max}}}{2} V(d) + \frac{\kappa L^3(D-d)^3}{T^2}
\]
\[\text{s.t.} \quad 0 \leq d \leq D. \quad (33a)
\]
KKT conditions are given in the following:
\[
1 + \frac{h_n}{g_n} - \frac{\mu h_n w_n}{\delta^2 w_n + P_n h_n} - \zeta_n = 0, \quad \forall n \in \mathcal{N}
\]
\[
\lambda - \mu \left( \ln \left( 1 + \frac{P_n h_n}{\delta^2 w_n} \right) - \frac{P_n h_n}{\delta^2 w_n} \right) - \eta_n = 0,
\]
\[
\forall n \in \mathcal{N}
\]
\[
\mu \left( \frac{2d}{T - \frac{f_{\text{max}}}{2}} - \sum_{n=1}^{N} w_n \ln \left( 1 + \frac{P_n h_n}{\delta^2 w_n} \right) \right) = 0
\]
\[
\lambda \sum_{i=1}^{N} w_i - W = 0
\]
\[
\zeta_n P_n = 0, \quad \forall n \in \mathcal{N}
\]
\[
\eta_n w_n = 0, \quad \forall n \in \mathcal{N}
\]
\[
\sum_{n=1}^{N} w_i \leq W
\]
\[
P_n = 0, \quad \forall n \in \mathcal{N}
\]
\[
w_n = 0, \quad \forall n \in \mathcal{N}
\]
in which \(\mu, \lambda, \zeta_n\) and \(\eta_n\) are non-negative Lagrangian multipliers associated with (34a), (34b), (34c) and (34d), respectively. (34a), (34b) are gradient vanishing requirements for \(P_i\) and \(w_i\).

Denoting \(\text{SNR}_n = \frac{P_n h_n}{\delta^2 w_n}\), Lemma 5 is derived to characterize the signal-to-noise ratio for the relay nodes.

**Lemma 5:** For \(i, j \in \mathcal{N}\) that satisfies \(P_i > 0\), \(P_j > 0\), \(w_i > 0\) and \(w_j > 0\),
\[
\text{SNR}_i = \text{SNR}_j = \text{SNR}
\]

**Proof:** We use the same monotonic increasing function \(\theta(x)\) as in proof of Lemma 1. For \(P_i > 0\) and \(w_i > 0\), \(\zeta_i = 0\) and \(\eta_i = 0\), it can be drawn from (34b) that \(\theta(\text{SNR}_n) = \frac{1}{\mu}\), in which \(\lambda\) and \(\mu\) are global Lagrangian multipliers. Therefore for \(\forall n \in \mathcal{N}\) that satisfies \(P_n > 0\) and \(w_n > 0\), \(\text{SNR}_n\) are the same.

Letting \(\text{SNR} = \theta^{-1}(\frac{1}{\mu})\) completes the proof.

With the above lemmas, \(\text{SNR}_n = \frac{P_n h_n}{\delta^2 w_n} = \frac{\sum_{i=1}^{N} P_i h_i}{\sum_{i=1}^{N} \delta^2 W}\).

After substitution of this equation, Problem 10 becomes:

**Problem 12:**
\[
\min_{\{P_n | n \in \mathcal{N}\}} \sum_{n=1}^{N} P_n \left( 1 + \frac{h_n}{g_n} \right)
\]
\[
\text{s.t.} \quad \sum_{n=1}^{N} P_n h_n \geq \delta^2 W \left( e^{\frac{2d}{T - \frac{f_{\text{max}}}{2}}} - 1 \right)
\]
\[
P_n \geq 0, \quad \forall n \in \mathcal{N}
\]
which is linear programming and can be solved by existing methods.

To solve the upper level Problem 11, the following lemma is derived.

**Lemma 6:** \(V(d)\) is a monotonic increasing function.

**Proof:** Define the right-hand side of constraint (36a) as a function
\[
\phi(d) = \delta^2 W \left( e^{\frac{2d}{T - \frac{f_{\text{max}}}{2}}} - 1 \right)
\]
Its first-order derivative with respect to \(d\) is
\[
\phi'(d) = \frac{2\delta^2 f_{\text{max}} T}{(T - \frac{f_{\text{max}}}{2})^2} e^{\frac{2d}{T - \frac{f_{\text{max}}}{2}}}
\]
which is positive for \(Ld \leq f_{\text{max}} T\). Then \(\phi(d)\) is monotonic increasing. In Problem 12 increasing \(d\) shrinks the feasible region of \(P_n\), thus increase the optimal value \(V(d)\). The proof is complete.

Note that the objective function of Problem 11 is equivalent to difference between two functions, i.e. \(G(d) - H(d)\), where \(G(d) = \frac{f_{\text{max}}}{2} V(d)\) and \(H(d) = \frac{f_{\text{max}}}{2} V(d) - \frac{\kappa L (D - d)^2}{T^2}\). Due to strict increasing property of \(V(d)\), the function \(G(d)\) and \(H(d)\) are monotonic increasing.

By introducing a new variable \(\omega\), Problem 11 is equivalent to

**Problem 13:**
\[
\max_{d, \omega} \quad H(d) + \omega
\]
\[
\text{s.t.} \quad \omega + G(d) \leq G(D)
\]
\[
0 \leq d \leq D
\]
\[
\omega \geq 0
\]

The reason for the equivalence is as follows: first, minimizing the objective function in Problem 11 is equivalent to maximizing \(H(d) - G(d) + G(D)\), and the maximal objective value happens only when \(\omega = G(D) - G(d)\).

Problem 13 is in the form of monotonic programming [19], and can be solved by Polyblock Algorithm, which is shown below.

**Algorithm 1 Polyblock Algorithm**

1. Choose a small value \(\nu\).
2. Define a two-dimension point set \(\mathcal{P} = \{p_1, p_2, \ldots, p_{|\mathcal{P}|}\}\), and initialize the set with point \(p_0 = (d_0, \omega_0)\), where \(d_0 = D, \omega_0 = G(D) - G(0)\).
3. Initialize the best point as \(p_{\text{opt}} = \emptyset\), and the best value \(v_{\text{opt}} = -\infty\).
4. while \(\mathcal{P} = \emptyset\) do
5. for \(i = 1, 2, \ldots, |\mathcal{P}|\) do
6. Calculate \(l_i\) that satisfies \(G(l_i p_i(1)) + l_i p_i(2) = G(D)\). Set \(\pi_i = (l_i p_i(1), l_i p_i(2))\).
7. Find \(i^* = \arg \max_{1 \leq i \leq |\mathcal{P}|} H(\pi_i(1)) + \pi_i(2)\). Set the maximal value as \(v_{e^*}\).
8. If \(v_{e^*} \geq v_{\text{opt}}\), let \(p_{\text{opt}} := p_{e^*}\); \(v_{\text{opt}} := v_{e^*}\).
9. For \(p_i \in \mathcal{P}\), if \(H(p_i(1)) + p_i(2) < v_{e^*} + \nu\), then remove \(p_i\).
10. Generate new points \(s_1 = (\pi_{e^*}(1), p_{e^*}(2))\) and \(s_2 = (p_{e^*}(1), \pi_{e^*}(2))\).
11. Add these two points into \(\mathcal{P}\).
12. Output the best point \(p_{\text{opt}}\) and best value \(v_{\text{opt}}\).
C. Solution for AF case

For AF mode, we look into the nonconvex Problem 3. Similar to TDMA and FDMA cases, the following lemma is derived.

**Lemma 7:** For optimal solution of Problem 3, equality of constraint (15a) and (15b) hold.

**Proof:** The proof for Lemma 7 is similar to that of Lemma 2 and Lemma 4 and thus omitted here. \( \square \)

After substitution of \( t = \frac{T_d d}{Tu} \) and some manipulation, Problem 3 is transformed into the following form:

**Problem 14:**

\[
\begin{align*}
& \min_{d, P, \{\beta_n | n \in \mathcal{N}\}} \left( T - \frac{L_d}{T_{su}} \left( P + \sum_{n=1}^{N} \beta_n^2 (P h_n + \delta^2 W) \right) + \frac{\kappa L^3 (D - d)^3}{T^2} \right) \\
& \text{s.t.} \quad \frac{P \left( \sum_{n=1}^{N} \sqrt{h_n g_n \beta_n} \right)^2}{\delta^2 W \left( 1 + \sum_{n=1}^{N} g_n \beta_n^2 \right)} \geq e^{\frac{2d}{T_{su}}} - 1 \\
& \quad 0 \leq d \leq D, P_n \geq 0, \beta_n \geq 0, \forall n \in \mathcal{N}.
\end{align*}
\]

This is still a nonconvex problem. To make Problem 14 tractable, we split Problem 14 into two levels. In the lower level, the amount of offloaded data \( d \) is fixed. The rest variables \( P \) and \( \{\beta_n | n \in \mathcal{N}\} \) are optimized to obtain function \( X(d) \) in Problem 15.

**Problem 15:**

\[
\begin{align*}
& \min_{P, \{\beta_n | n \in \mathcal{N}\}} \left( P + \sum_{n=1}^{N} \beta_n^2 (P h_n + \delta^2 W) \right) \\
& \text{s.t.} \quad \frac{P \left( \sum_{n=1}^{N} \sqrt{h_n g_n \beta_n} \right)^2}{\delta^2 W \left( 1 + \sum_{n=1}^{N} g_n \beta_n^2 \right)} \geq e^{\frac{2d}{T_{su}}} - 1 \\
& \quad P_n \geq 0, \beta_n \geq 0, \forall n \in \mathcal{N}.
\end{align*}
\]

In the upper level Problem 16, variable \( d \) is optimized:

**Problem 16:**

\[
\begin{align*}
& \min_{d} \left( T - \frac{L_d}{T_{su}} X(d) + \frac{\kappa L^3 (D - d)^3}{T^2} \right) \\
& \text{s.t.} \quad 0 \leq d \leq D
\end{align*}
\]

Defining \( \psi(d) = e^{\frac{2d}{T_{su}}} - 1 \) and introducing slack variable \( \varepsilon \), Problem 15 is equivalent to:

**Problem 17:**

\[
\begin{align*}
& \min_{P, \varepsilon, \{\beta_n | n \in \mathcal{N}\}} \left( P + \sum_{n=1}^{N} \beta_n^2 (P h_n + \delta^2 W) \right) \\
& \text{s.t.} \quad \frac{\delta^2 W \left( 1 + \sum_{n=1}^{N} g_n \beta_n^2 \right)}{\varepsilon} \leq 1 \\
& \quad P \left( \sum_{n=1}^{N} \sqrt{h_n g_n \beta_n} \right)^2 \geq \varepsilon \\
& \quad P \geq 0, \beta_n \geq 0, \varepsilon \geq 0, \forall n \in \mathcal{N}.
\end{align*}
\]

The objective function and constraints (43a) and (43b) in Problem 17 with respect to \( P \) and \( \beta_n \) are in posynomial form. Therefore, we utilize variable substitution \( P = e^q, \varepsilon = e^s \), \( \beta_n = e^{\alpha_n} \), and further take logarithm of the objective function and constraints. Problem 17 is transformed into:

**Problem 18:**

\[
\begin{align*}
& \min_{q, s, \{\alpha_n | n \in \mathcal{N}\}} \ln \left( e^q + \sum_{n=1}^{N} \sqrt{h_n g_n e^{\alpha_n}} \right) + q - s \\
& \text{s.t.} \quad \ln \left( \sum_{n=1}^{N} g_n e^{2 \alpha_n} + 1 \right) - s \leq - \ln \psi(d) - \ln \delta^2 W \\
& \quad 2 \ln \left( \sum_{n=1}^{N} \sqrt{h_n g_n e^{\alpha_n}} \right) + q - s \geq 0 \\
& \quad q \geq 1, s \geq 1, \alpha_n \geq 1, \forall n \in \mathcal{N}.
\end{align*}
\]

Notice that the relationship between the minimum of Problem 17 and that of Problem 18 is, \( Y(d) = \ln X(d) \). Due to strict monotonicity of the logarithm function, Problem 17 and 18 are equivalent.

In Problem 18, the objective function and the left-hand side of constraint (44a) and (44b) are log-sum-exp functions [18], which is convex since its second order derivative is positive definite. However, constraint (44b) is in the form of a convex function larger than 0. To solve the problem, we introduce successive convex approximation (SCA). Before going into details of the algorithm, the iterative problem is defined as follows:
In Problem 19, \((q', s', \{\alpha_n'\})\) is a given and fixed point in the feasible region of Problem 18. Note that Problem 19 is a convex problem and can be solved by traditional methods. Define the objective function of Problem 18 as \(y(q, s, \{\alpha_n\})\), and that of Problem 19 as \(y'(q, s, \{\alpha_n\})\), the following lemma is expected.

**Lemma 8:** \(y'(q, s, \{\alpha_n\}) \geq y(q, s, \{\alpha_n\})\), i.e., the objective function of Problem 18 serves as a global lower bound for that of Problem 19.

**Proof:** The left-hand side of (44b) is jointly convex with respect to \((q, s, \{\alpha_n\})\). Using first-order condition of convex functions [18], for any feasible points \((q', s', \{\alpha_n'\})\) of Problem 18 we have

\[
\frac{2\sqrt{h_n} e^{\alpha_n}}{\sum_{n=1}^{N} \sqrt{h_n} e^{\alpha_n}} \alpha_n - \alpha_n' + (q - q') - (s - s') + 2 \ln \left( \frac{\sum_{n=1}^{N} \sqrt{h_n} e^{\alpha_n}}{\sum_{n=1}^{N} \sqrt{h_n} e^{\alpha_n}} \right) + q - s
\]

Thus, (45b) is a sufficient but not necessary condition to (44b), and Problem 19 have smaller feasible region than Problem 18. Therefore, Problem 18 may yields lower objective value.

This completes the proof.

The process of solving Problem 18 is shown in Algorithm 2.

**Algorithm 2** Successive convex approximation for Problem 18

1. Choose a small value \(\nu\).

2. Randomly choose a feasible point \((q^0, s^0, \{\alpha_n^0\})\) of Problem 18 denote its objective value as \(Y^0\), let \(Y^0 = Y^0\).

3. Given the fixed point \((q^i, s^i, \{\alpha_n^i\})\), solve Problem 19. Let \((q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\) be the optimal solution, \(Y^i = y(q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\) is the optimal value.

4. if \(Y^i - Y^i - 1 < \nu\) then

5. Quit. Claim the optimal solution is \((q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\).

6. else

7. Set \((q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\) the new fixed point and let \(i := i + 1\), go back to 3.

**Lemma 9:** For arbitrary \((q^0, s^0, \{\alpha_n^0\})\) in the feasible region of Problem 18, Algorithm 2 generates a sequence of improved points, which converges to a stationary point.

**Proof:** Note that \((q', s', \{\alpha_n'\})\) and \((q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\) are feasible points for both Problem 18 and 19. We have

\[
y(q', s', \{\alpha_n'\}) = y(q', s', \{\alpha_n'\}) \geq y(q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\}) \geq y(q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\}) \geq y(q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})
\]

in which the first equality is due to the definition of Problem 19 and the second inequality is because of the optimality of \((q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\) in Problem 19 with fixed point \((q^i, s^i, \{\alpha_n^i\})\). (47) indicates that, \((q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})\) yields lower objective value than \((q', s', \{\alpha_n'\})\) in Problem 18. From Cauchy's theorem and limited feasible region, there is a convergent subsequence \((q^i, s^i, \{\alpha_n^i\})\) with limit point \((q^*, s^*, \{\alpha_n^*\})\), which satisfies

\[
\lim_{v \to \infty} (y(q^{i'}, s^{i'}, \{\alpha_n^{i'}\}) - y(q^*, s^*, \{\alpha_n^*\})) = 0\] (48)

For certain \(i\), there must exists a \(v\) that \(i_v \leq i \leq i_{v+1}\) so that

\[
y(q^{i'}, s^{i'}, \{\alpha_n^{i'}\}) \geq y(q^{i'}, s^{i'}, \{\alpha_n^{i'}\}) \geq y(q^{i+1}, s^{i+1}, \{\alpha_n^{i+1}\})
\]

When \(i\) goes to infinity,

\[
0 \geq \lim_{v \to \infty} (y(q^{i'}, s^{i'}, \{\alpha_n^{i'}\}) - y(q^*, s^*, \{\alpha_n^*\})) \geq \lim_{v \to \infty} (y(q^{i'}, s^{i'}, \{\alpha_n^{i'}\}) - y(q^*, s^*, \{\alpha_n^*\})) \geq \lim_{v \to \infty} (y(q^{i'+1}, s^{i'+1}, \{\alpha_n^{i'+1}\}) - y(q^*, s^*, \{\alpha_n^*\})) = 0
\]

(50)

Therefore, the sequence \((q^i, s^i, \{\alpha_n^i\})\) is also convergent, with its limit point \(\lim_{i \to \infty} (q^i, s^i, \{\alpha_n^i\}) = (q^*, s^*, \{\alpha_n^*\})\). Based on [20] Theorem 1, the limit point \((q^*, s^*, \{\alpha_n^*\})\) is a stationary point.

This completes the proof.

By analyzing the function \(\psi(d)\) defined in Problem 17, the following lemma can be expected.

**Lemma 10:** \(X(d)\) is a monotonic increasing function.
Proof: The first-order derivative of $\psi(d)$ is

$$\psi'(d) = \frac{2f_{\text{max}}^2 T}{W(f_{\text{max}} T - Ld)^2} \exp\left(\frac{2d_{\text{max}}}{f_{\text{max}} T - Ld}\right)$$

(51)

which is positive for $Ld \leq f_{\text{max}} T$. Then $\psi(d)$ is monotonic increasing. In Problem 15 increasing $d$ shrinks the feasible region, thus increase the optimal value $X(d)$. The proof is complete.

Due to the monotonic property of $X(d)$, the objective function of Problem 16 can be recognized as difference of two monotonic functions with respect to $d$, which are $\frac{f_{\text{max}}}{L} X(d)$ and $\frac{d_{\text{max}}}{L} X(d) - \frac{\kappa L^3 (D-d)^2}{L}$. Similar to the case for FDMA, we use Polyblock Algorithm [19] to solve Problem 16. The detailed algorithm is omitted.

V. NUMERICAL RESULTS

In this section, numerical results of our proposed algorithm is given out and analyzed for TDMA mode, FDMA mode and AF mode. In the simulation, bandwidth of the whole system is set as $W = 1$ MHz. For all the relays, the distances from the mobile device to the relays and from the relays to the BS range from 100 to 500 meters. Rayleigh block-fading channel is considered, in which the channel gain is affected by free space path loss and Rayleigh distribution. The free space path loss is acquired with the formula (in dB)

$$PL = 32.4 + 20 \times \log \text{Distance} + 20 \times \log \text{Bandwidth}$$

The channel gains under Rayleigh channel obey exponential distribution with mean of 0.5. The spectral noise power density is -140dBW/Hz. Similar to [5], the energy consumption coefficient for local computing $\kappa = 10^{-25}$. The maximum CPU frequency of edge server in the BS $f_{\text{max}}$ ranges from 2GHz to 6GHz. To guarantee quality of service, computation task of size around $8 \times 10^4$ nats is supposed to be finished in about 0.01 second. Finally, similar to [5], $L = 50$ cycles/nat.

A. Verification of convergence and optimality

In this subsection, the convergence of SCA to solve Problem 17 is verified, followed by one-dimension search of Problem 16 to make sure that monotonic programming in upper problem leads to optimal solution.

Fig.4 illustrates the efficiency of SCA in Problem 17. When $D = 8 \times 10^4$ nats and $d = 6 \times 10^4$ nats, lower objective value of Problem 17 approaches a fixed number about 1.201W with deviation no more than $10^{-5}$ after 15 iterations, in which each iteration contains a solution of convex problem with interior point methods.

Next, we utilizes one-dimension search to draw the curve of objective value in Problem 16. Having $D = 8 \times 10^4$ nats, step length is set as 100 nats, the objective function is unimodal with respect to $d$. Observing Fig.5 by offloading task in the size of $d = 5.318 \times 10^4$ nats, mobile device and relays achieve the lowest total energy consumption. Fig.5 together with Fig.6 and Fig.7 confirms the optimality of monotonic programming in our proposed methods.

B. Comparison between different cases

In this subsection, optimal data offloading and relative energy consumption is depicted as functions of task size $D$, latency requirement $T$ and edge computing capacity $f_{\text{max}}$, respectively. To make comparison, we merge the curves for above three cases into same figures. Furthermore, a suboptimal case of TDMA and FDMA, in which the relays equally split the communication resource despite channel condition, is considered as a benchmark scheme.

Given $T = 0.01$ second and $f_{\text{max}} = 5$ GHz, Fig.6 shows the optimal offloading data amount versus total data amount $D$. Due to limited communication capacity, the optimal offloading data amount varies sublinearly with $D$. Fig.7 expresses minimum energy consumption as a function of total data amount. Further, combining Fig.6 and Fig.7 for $D = 8 \times 10^4$ nats in AF case, the optimal offloading amount and energy consumption matches the marked point of one-dimensional search in Fig.5 which confirms the optimality of our proposed method for
Fig. 6. Optimal offloading data amount for different $D$ in the range from $6 \times 10^4$ to $10^5$ nats.

Fig. 8. Optimal offloading data amount for different $T$ in the range from 0.08 to 0.12 second.

Fig. 7. Optimal energy consumption for different $D$ in the range from $6 \times 10^4$ to $10^5$ nats.

Fig. 9. Optimal energy consumption for different $T$ in the range from 0.08 to 0.12 second.

Problem 14.

With $D = 8 \times 10^4$ nats and $f_{\text{max}} = 5$ GHz, Fig. 8 plots the optimal offloading data amount versus required latency $T$. The curve is nearly linear for all four cases. This is reasonable since both computation and communication resource are directly proportional to time. In Fig. 9, it can be seen that the energy consumption reduces greatly when the latency is prolonged. Therefore, it is rather energy consuming for real-time task to reduce its latency.

Setting $D = 8 \times 10^4$ nats and $T = 0.01$ second, Fig. 10 and Fig. 11 explores the optimal offloading data amount and relative energy consumption for different edge computation capacities $f_{\text{max}}$. It can be noticed that the computation duration in the BS defined by $Ld/f_{\text{max}}$ is around $2.5 \times 10^{-3}$ second. Comparing this with the overall latency, major latency is due to the transmission process. Therefore, optimal offloading amount is mainly affected by channel condition.

Remarks:

- In the above figures, curves for TDMA and FDMA are precisely the same. Despite the difference between Problem 1 and Problem 2, the allocation of communication resources for either TDMA or FDMA is reduced into selection of relay nodes that have higher channel gain given offloading data $d$. As for upper level problem, despite distinct in convexity, the final results are the same.
- Combining the above figures, AF slightly outperforms DF-TDMA and DF-FDMA by allowing larger amount of data be offloaded. This is because the channel parameter is similar for different relays, and using amplify-and-forward can enhance the signal-to-noise ratio to some degree. In decode-and-amplify cases, as vertex in linear programming, only the best relay is selected often.
geometric programming and SCA are utilized to obtain a convergent solution. Effectiveness of the proposed strategies are verified with numerical results. This research could provide helpful insight on optimal resource allocation under different working mode for relay assisted MEC system.

REFERENCES

[1] W. Shi, J. Cao, Q. Zhang, Y. Li and L. Xu, “Edge Computing: Vision and Challenges,” IEEE Internet Things J., vol. 3, no. 5, pp. 637-646, Oct. 2016.

[2] P. Mach and Z. Becvar, “Mobile Edge Computing: A Survey on Architecture and Computation Offloading,” IEEE Commun. Surveys Tuts., vol. 19, no. 3, pp. 1628-1656, 3rd. Quart 2017.

[3] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, “A Survey on Mobile Edge Computing: The Communication Perspective,” IEEE Commun. Surveys Tuts., vol. 19, no. 4, pp. 2322-2358, 4th. Quart 2017.

[4] C. You, K. Huang and H. Chae, “Energy efficient mobile cloud computing powered by wireless energy transfer,” IEEE J. Sel. Areas Commun., vol. 34, no. 5, pp. 1757â-˘A¸S1771, May 2016.

[5] Y. Wang, M. Sheng, X. Wang, L. Wang and J. Li, “Mobile-Edge Computing: Partial Computation Offloading Using Dynamic Voltage Scaling,” IEEE Trans. Commun., vol. 64, no. 10, pp. 4268-4282, Oct. 2016.

[6] J. Ren, G. Yu, Y. Cai and Y. He, “Latency Optimization for Resource Allocation in Mobile-Edge Computation Offloading,” IEEE Trans. Wireless Commun., vol. 17, no. 8, pp. 5506-5519, Aug. 2018.

[7] F. Wang, J. Xu, X. Wang and S. Cui, “Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems,” IEEE Trans. Wireless Commun., vol. 17, no. 3, pp. 1784-1797, Mar. 2018.

[8] T. Q. Dinh, J. Tang, Q. D. La and T. Q. S. Quek, “Offloading in Mobile Edge Computing: Task Allocation and Computational Frequency Scaling,” IEEE Trans. Commun., vol. 65, no. 8, pp. 3571-3584, Aug. 2017.

[9] S. Bi and Y. J. Zhang, “Computation Rate Maximization for Wireless Powered Mobile-Edge Computing With Binary Computation Offloading,” IEEE Trans. Wireless Commun., vol. 17, no. 6, pp. 4177-4190, June 2018.

[10] Z. Liang, Y. Liu, T. Lok and K. Huang, “Multiuser Computation Offloading and Downloading for Edge Computing With Virtualization,” IEEE Trans. Wireless Commun., vol. 18, no. 9, pp. 4298-4311, Sept. 2019.

[11] W. Zhang, Y. Ren and D. O. Wu, “Collaborative task execution in mobile cloud computing under a stochastic wireless channel,” IEEE Trans. Wireless Commun., vol. 14, no. 1, pp. 81AÂ’s93, Jan. 2015.

[12] C. You, K. Huang, H. Chae and B. Kim, “Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading,” IEEE Trans. Wireless Commun., vol. 16, no. 3, pp. 1397-1411, Mar. 2017.

[13] Y. Wu, L. P. Qian, K. Ni, C. Zhang and X. Shen, “Delay-Minimization Nonorthogonal Multiple Access Enabled Multi-User Mobile Edge Computation Offloading,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 3, pp. 392-407, June 2019.

[14] X. Cao, F. Wang, J. Xu, R. Zhang and S. Cui, “Joint computation and communication cooperation for mobile edge computing,” in Proc. 16th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw., 2018.

[15] X. Hu, K. Wong and K. Yang, “Wireless Powered Cooperation-Assisted Mobile Edge Computing,” IEEE Trans. Wireless Commun., vol. 17, no. 4, pp. 2375-2388, Apr. 2018.

[16] Y. J. Zhang, L. Qian and J. Huang, “Monotonic Optimization in Communication and Networking Systems,” Found. Trends Netw., vol. 7, no. 1, pp. 1-75, Oct. 2013.

[17] B. R. Marks and G. P. Wright, “A General Inner Approximation Algorithm for Nonconvex Mathematical Programs,” Oper. Res., vol. 26, no. 4, pp. 681-683, July-Aug. 1978.

[18] H. H. M. Tam, H. D. Tuan, D. T. Ngo, T. Q. Duong and H. V. Poor, “Joint Load Balancing and Interference Management for Small-Cell Heterogeneous Networks With Limited Backhaul Capacity,” IEEE Trans. Wireless Commun., vol. 16, no. 2, pp. 872-884, Feb. 2017.

Fig. 10. Optimal offloading data amount for different $f_{\text{max}}$ in the range from 2 to 6 GHz.

Fig. 11. Optimal energy consumption for different $f_{\text{max}}$ in the range from 2 to 6 GHz.

VI. CONCLUSIONS

In this paper, we have considered different relay modes to minimize the energy consumption in a multiple-relay assisted MEC system. In this paper, we have investigated a MEC system aided by multiple relays working in DF-TDMA, DF-FDMA, and AF modes respectively. For DF-TDMA mode, the amount of offloaded data, slot duration and transmit power of the mobile user and different relays are jointly optimized. To solve the associated problem optimally in an easy way, a method of bi-level optimization is utilized. For DF-FDMA mode, bandwidth allocation, instead of slot duration is optimized. In this nonconvex problem, bi-level optimization and monotonic programming is used to find the global optimal solution. For AF mode, the amount of offloaded data, transmit duration, transmit power of the mobile user and amplitude gain of the relays are optimized. To solve this nonconvex problem, geometric programming and SCA are utilized to obtain a convergent solution. Effectiveness of the proposed strategies are verified with numerical results. This research could provide helpful insight on optimal resource allocation under different working mode for relay assisted MEC system.

In this paper, we have considered different relay modes to minimize the energy consumption in a multiple-relay assisted MEC system. In this paper, we have investigated a MEC system aided by multiple relays working in DF-TDMA, DF-FDMA, and AF modes respectively. For DF-TDMA mode, the amount of offloaded data, slot duration and transmit power of the mobile user and different relays are jointly optimized. To solve the associated problem optimally in an easy way, a method of bi-level optimization is utilized. For DF-FDMA mode, bandwidth allocation, instead of slot duration is optimized. In this nonconvex problem, bi-level optimization and monotonic programming is used to find the global optimal solution. For AF mode, the amount of offloaded data, transmit duration, transmit power of the mobile user and amplitude gain of the relays are optimized. To solve this nonconvex problem, geometric programming and SCA are utilized to obtain a convergent solution. Effectiveness of the proposed strategies are verified with numerical results. This research could provide helpful insight on optimal resource allocation under different working mode for relay assisted MEC system.