Joint Computation Offloading and User Association in Multi-Task Mobile Edge Computing

Yueyue Dai, *Student Member, IEEE*, Du Xu, *Member, IEEE*, Sabita Maharjan, *Member, IEEE*, and Yan Zhang, *Senior Member, IEEE*

***Abstract*—Computation intensive and delay-sensitive applications impose severe requirements on mobile devices of providing required computation capacity and ensuring latency. Mobile edge computing (MEC) is a promising technology that can alleviate computation limitation of mobile users and prolong their lifetime through computation offloading. However, computation offloading in an MEC environment faces severe issues due to dense deployment of MEC servers. Moreover, a mobile user has multiple mutually dependent tasks, which make offloading policy design even more challenging. To address the above-mentioned problems in this paper, we first propose a novel two-tier computation offloading framework in heterogeneous networks. Then, we formulate joint computation offloading and user association problem for multi-task mobile edge computing system to minimize overall energy consumption. To solve the optimization problem, we develop an efficient computation offloading algorithm by jointly optimizing user association and computation offloading where computation resource allocation and transmission power allocation are also considered. Numerical results illustrate fast convergence of the proposed algorithm, and demonstrate the superior performance of our proposed algorithm compared to state of the art solutions.**

***Index Terms*—Mobile edge computing (MEC), computation off-loading, user association, resource allocation, optimization.**

I. Introduction

RECENT advancements in Internet of Things (IoT) and 5G wireless technologies have paved a path towards realizing new applications (e.g., surveillance, augmented/virtual reality, and e-Health care) through both machine-to-machine and machine-to-human interactions [1], [2]. However, the finite battery level and resource-constrained mobile devices, such as wearable devices, on-device sensors, and smart phones, cannot fully enable or support the computation-intensive and delay-sensitive services. To alleviate the computation limitations and prolong the lifetime of mobile devices, Mobile Edge Computing (MEC) is a promising paradigm that integrates cloud computing and mobile network to offer considerable computation resources at network edge. Exploiting MEC, the applications with stringent low-latency requirement can be processed in time by offloading from mobile devices to a proximate MEC server.

To achieve user proximity, MEC servers are deployed at both the Macro Base Station (MBS) and/or Small Base Station (SBS) which results in a dense deployment of MEC servers. Since MEC servers are densely distributed, mobile users face a user association dilemma. That is, each mobile user should determine whether or not to associate with a particular base station for offloading. The user association decision is of significant importance for offloading as it directly affects communication data rate. Different from user association schemes in conventional heterogeneous networks [3] and [4], the user association scheme in MEC should not only consider bandwidth, power, and interference but also should take both computation data size and delay requirement of applications into consideration.

Generally, an application is composed of several divisible and logically independent tasks. For example, an augmented reality application consists of five critical tasks, namely, video source, tracker, mapper, object recognizer, and renderer [5]. Among these tasks, the tracker, mapper, and objective recognizer are computation-intensive components which can be offloaded from IoT devices to MEC servers to process. Through offloading, both the energy consumption of IoT devices and the application processing time can be reduced. Moreover, there exists a certain dependency relationship among these tasks, such as sequential dependency, parallel dependency, and hybrid dependency [6]. The task dependency determines the execution order and execution time of these tasks. To ensure that an application with multiple tasks can be processed in time, it is necessary to take task dependency relationship into consideration for offloading policy design.

Computation offloading has recently attracted much attention. The authors in [7] proposed binary offloading where each task has to be computed as a whole either locally on the mobile device or remotely on the edge servers via offloading. Studies such as [8] and [9] introduced the idea of partial offloading where each task can be arbitrarily divided into two parts for local and edge computing. Since the efficiency of computation offloading for

Manuscript received July 14, 2018; revised October 2, 2018; accepted October 6, 2018. Date of publication October 18, 2018; date of current version December 14, 2018. This work was supported in part by the National key Research and Development Program of China (2016YFB0800105), in part by the projects 240079/F20 funded by the Research Council of Norway, in part by the 111 project (B14039), and in part by the scholarship from China Scholarship Council under Grant 201706070009. The review of this paper was coordinated by Prof. Y. Guo. *(Corresponding author: Yan Zhang.)*

Y. Dai and D. Xu are with the Key Laboratory of Optical Fiber Sensing and Communications (Ministry of Education), University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: daiyyue@gmail.com; xudu.uestc@gmail.com).

S. Maharjan is with the Simula Metropolitan Center for Digital Engineering, Norway, and also with the University of Oslo, 0316 Oslo, Norway (e-mail: sabita@simula.no).

Y. Zhang is with Department of Informatics, University of Oslo, Norway (e-mail: yanzhang@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2018.2876804

MEC highly depends on the wireless channel condition, various computation offloading policies were proposed. The authors in [10] and [11] proposed computation offloading schemes integrating wireless power transfer technology. The authors in [12] and [13] proposed offloading policies by jointly optimizing subcarrier allocation and computation resource. The authors in [14] proposed a dynamic offloading policy for MEC system powered by energy harvesting to minimize the execution cost where computation resource and transmission power are also considered. The authors in [15] proposed a computation offloading policy also incorporating interference management by jointly optimizing offloading, computation resource, and physical resource block allocation. However, there are some limitations in the commonly adopted assumptions in [7]–[15]. The aforementioned offloading policies assume that there is only one MEC server in the MEC system such that all mobile users have to determine whether or how much to offload to this MEC server. In reality, MEC servers are densely distributed. Therefore, for dense distribution of MEC servers, a novel computation offloading policy that takes user association into account, is necessary. Further, the above works assume that each mobile user has a single task thus making them inapplicable in the context of the multi-task offloading problem, especially with the consideration of task dependency relationship. A few works have focused on multi-task offloading. The work in [16] and [17] proposed online multi-task offloading schemes for MEC system. In [18], the authors proposed two multi-task offloading policies for Mobile Cloud Computing (MCC) taking sequential task dependency and parallel task dependency, respectively, into consideration. However, none of them considers how to schedule tasks in a distributed MEC system with dense distribution of MEC servers to reduce energy consumption or to improve performance.

To address the above challenges, in this paper, we integrate user association with computation offloading to minimize energy consumption of users and MEC servers in a multi-task scenario. Moreover, as task dependency has a noticeable impact in application completion time, we also incorporate task dependency into problem formulation. We consider a heterogeneous network, which consists of an MBS, multiple SBSs, MBS, SBSs, and multiple mobile users. Each base station is equipped with an MEC server and each mobile user has a delay-sensitive application which is composed of several separable tasks. By jointly optimizing computation offloading and user association, we develop an efficient computation offloading algorithm for minimizing energy consumption. Further, computation resource allocation and transmission power allocation are also taken into consideration. The key contributions of our work are as follows:

- We propose a two-tier computation offloading framework for multi-task offloading in heterogeneous networks, where both mobile users and SBSs can offload computation-intensive tasks to a particular MEC server or to an SBS and/or the MBS, respectively.
- In the two-tier offloading framework, we formulate the joint computation offloading and user association problem as an optimization problem for minimizing the total energy consumption of mobile users and MEC servers.

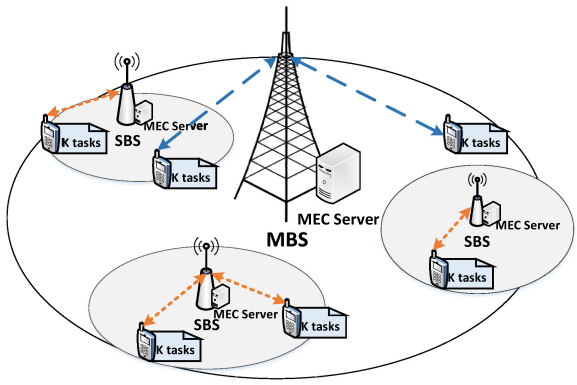


Fig. 1. The scenario with multiple users and tasks for mobile edge computing (MEC) under a 5G heterogeneous network.

- We propose an efficient computation offloading algorithm by jointly optimizing user association and computation offloading where computation resource allocation and transmission power allocation are also considered.

The rest of this paper is as follows. In Section II, we introduce the system model. In Section III, we formulate the joint computation offloading and user association problem and propose an efficient computation offloading algorithm to solve this problem. We evaluate the performance of the proposed algorithm and provide illustrative results in Section IV and conclude the paper in Section V.

II. System Model

In this section, we first introduce the system model, i.e., MEC system with multiple users and tasks. Then we present the communication model, computation model, and multi-task model in details.

A. Mobile Edge Computing System With Multiple Users and Tasks

We consider a 5G heterogeneous mobile edge computing system consisting of one MBS and M SBSs. Each base station is equipped with an MEC server, as shown in Fig. 1. The set of base stations is denoted as $\mathcal{B} = \{b_0, b_1, , \ldots, b_M\}$ where b_0 is the MBS. All M SBSs are connected to the MBS via wired links [19]. There are N mobile users, denoted as $\mathcal{U} = \{u_1, \ldots, u_N\}$. The SBSs and mobile users are randomly distributed within the coverage of the MBS. Each mobile user has a computation-intensive and delay-sensitive application to be executed within a certain execution deadline. Here, we focus on partition oriented applications [8], [20], such as the virus scan application and figure compression application. Such an application can be divided into K logically independent tasks, defined by the set $\mathcal{K} = \{1, 2, .., K\}$. Moreover, as the amount of data to be processed is known beforehand for this kind of applications, each logically independent task can be arbitrarily divided into several partitions to support parallelism. That is, partitions of each task can be processed locally and at the MEC side concurrently.

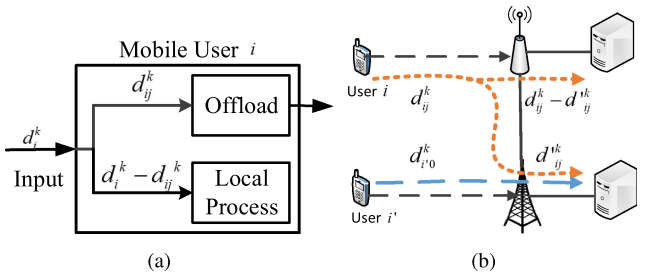


Fig. 2. (a) The mobile user i offloads parts of its task (i.e., d_{ij}^k) to base station j. (b) The offloading and computing routine on two kinds of users: user i is associated with SBS j and user i' is with the MBS.

There are K computational and logically independent tasks on each mobile user. For a single task, a mobile user partitions it into two parts. One part is offloaded to an MEC server via the base station it is associated to, and the other part is executed locally, as shown in Fig. 2(a). We assume the computation resource of each MEC server deployed on the SBS is limited. Thus, if a mobile user offloads parts of its task to an MEC server on an SBS but the computation resources of the MEC server is exhausted, the offloading partition should be further divided, and the SBS will offload the remaining part of the task to the MEC server on the MBS which has relatively higher amount of computation resources, as shown in Fig. 2(b). Therefore, the offloading in this scenario is two tier—by the mobile user and by the SBS. For K logically independent tasks of an application, the task dependency relationship has a direct impact on the application completion time. For simplicity, we analyze the typical multi-task dependency called, sequential dependency.

We consider that a mobile user is associated with one base station. Denote $x_{ij} \in \{0, 1\}$ as the association variable, i.e., $x_{ij} = 1$ if user $i \in \mathcal{U}$ is associated with base station $j \in \mathcal{B}$, and $x_{ij} = 0$ otherwise. Due to the difference of base stations on channel quality and computation resource, for a mobile user, selecting an appropriate base station to associate to, ensures that the computation tasks of this user can be completed within the delay constraint and the energy consumption can be optimized. On the contrary, the strict delay requirement of computation tasks influences the association decision.

In this paper, we focus on joint computation offloading and user association for multi-task to minimize overall energy consumption. Since the size of processed data or result is much smaller than the input data size, we only consider the energy consumption and latency constraints for offloading [7], [15], [13]. Next, we present the communication model, computation model, and multi-task model.

B. Communication Model

We employ Orthogonal Frequency Division Multiple Access (OFDMA) for communication between mobile users and base stations. We assume mobile users associated to the same base station are allocated orthogonal spectrum and the spectrum between the MBS and SBSs are also orthogonal [21]. Therefore, we only consider the inter-cell interference among SBSs [22].

The channel power gain H_{ij} between mobile user $i \in \mathcal{U}$ and base station $j \in \mathcal{B}$ is [3],

$$H_{ij} = G_{ij} F_{ij}, \tag{1}$$

where G_{ij} is the large-scale slow-fading component capturing effects of path loss and shadowing, and F_{ij} represents the small-scale Rayleigh fading. During one association period, G_{ij} is assumed to be a constant while F_{ij} fluctuates fast following an exponential distribution function with unit variance.

If mobile user i is associated with SBS j, the signal to interference plus noise ratio (SINR) of the uplink will be:

$$\gamma_{ij} = \frac{P_i H_{ij}}{\sigma^2 + \sum_{i'=1, i' \neq i}^{N} \sum_{j'=1, j' \neq j}^{M} P_{i'} H_{i'j'}} \quad \forall j \in \mathcal{B} \setminus \{b_0\}, \tag{2}$$

where P_i is the transmission power of mobile user i, σ^2 is the noise power, and $\sum_{i'=1, i' \neq i}^{N} \sum_{j'=1, j' \neq j}^{M} P_{i'} H_{i'j'}$ is the inter-cell interference among SBSs. As each SBS allocates bandwidth equally to its associated users, the achievable uplink data rate of mobile user i associated with SBS j can be expressed as:

$$R_{ij} = \frac{W_s}{\sum_{i=1}^{N} x_{ij}} \log(1 + \gamma_{ij}), \tag{3}$$

where W_s denotes the bandwidth of SBS j and $\sum_{i=1}^{N} x_{ij}$ is the amount of mobile users belonging to SBS j. From (2) and (3), we observe that if mobile users are associated with the same SBS to offload their tasks, they may suffer severe interference and low data rate.

Similarly, if mobile user i is associated with the MBS, the SINR is given by

$$\gamma_{i0} = \frac{P_i H_{i0}}{\sigma^2}. \tag{4}$$

The achievable rate of mobile user i associated with the MBS will then be:

$$R_{i0} = \frac{W_m}{\sum_{i=1}^{N} x_{i0}} \log(1 + \gamma_{i0}), \tag{5}$$

where W_m and $\sum_{i=1}^{N} x_{i0}$ are the bandwidth and the load of the MBS, respectively. If all users are associated with the MBS, due to the limitation of bandwidth, the uplink data rate of each users will be fairly low.

Therefore, the optimal user association policy is absolutely essential to trade-off the number of mobile users of the MBS and that of SBSs since it can improve communication data rate for transmitting offloading tasks and reduce energy consumption for offloading.

To obtain the highest data rate, a mobile user will be greedily associated with the base station with the maximum data rate. However, in the multi-user multi-task MEC system, since all computation tasks have a stringent delay constraint, a novel multi-user association scheme that takes both computation task size and delay requirement into consideration, is indeed needed. Specifically, a mobile user may choose a base station which provides the data rate that satisfies the delay constraint, instead of the base station with the maximum data rate.

C. Computation Model

We introduce the computation model in terms of energy consumption and task processing time. Consider that an application of mobile user i, composed of K computation tasks, is required to be accomplished within the execution deadline $T_i^{\max}$. Denote the k^{th} task of this application as $D_i^k \triangleq (d_i^k, c_i^k)$, where d_i^k denotes the data size of the k^{th} computation task, c_i^k is the number of CPU cycles for computing one bit of task D_i^k. Each task can be divided into several partitions for computation offloading. Next, we introduce the computation offloading process for task D_i^k.

The computation offloading process is two-tier as shown in Fig. 2. In the first tier, mobile user i offloads d_{ij}^k $(0 \leqslant d_{ij}^k \leqslant d_i^k)$ to base station j via wireless channel, and computes the rest $d_i^k - d_{ij}^k$ locally (Fig. 2(a)). In the second tier, SBS j further divides the fragment d_{ij}^k as $d_{ij}^k - {d'}_{ij}^k$ and ${d'}_{ij}^k$ $(0 \leqslant {d'}_{ij}^k \leqslant d_{ij}^k \leqslant d_i^k)$, and then offloads the latter to the MBS since it cannot process all the partitions by itself (Fig. 2(b)). Note that, if a mobile user is associated with the MBS, it directly offloads its task to the MBS without the second tier, like the user i' in Fig. 2(b). We next discuss the computation overhead of task D_i^k in terms of energy consumption and task processing time for local, SBS, and MBS computing.

1) Local Computing: For local computing, $d_i^k - d_{ij}^k$ partition of task D_i^k are executed locally on mobile user i. We denote f_i^k as the computation resource (i.e., CPU cycles per second) when processing task k, which is limited by the total computation resource F_i. When mobile user i is associated with base station j, the computation time $T_{ij}^L(k)$ of task D_i^k can be expressed as

$$T_{ij}^L(k) = \frac{(d_i^k - d_{ij}^k)c_i^k}{f_i^k}. \tag{6}$$

The energy consumption of each CPU cycle is $\varsigma(f_i^k)^2$, where ς is the effective switched capacitance depending on the chip architecture [8], [12]. We denote energy consumption for task D_i^k by $E_i^L(k)$, which can be defined as

$$E_{ij}^L(k) = \varsigma(d_i^k - d_{ij}^k)c_i^k(f_i^k)^2. \tag{7}$$

2) Offloading to SBS: As shown in Fig. 2(b), mobile user i offloads d_{ij}^k part of task D_i^k to SBS $j \in \mathcal{B} \setminus \{b_0\}$. If SBS j has sufficient computation resources, it will process this part alone, otherwise d_{ij}^k will be divided into two parts : ${d'}_{ij}^k$ and $d_{ij}^k - {d'}_{ij}^k$, and ${d'}_{ij}^k$ will be further offloaded to the MBS to process. We denote f_{ij}^k as the computation resource of SBS j assigned to task k for mobile user i, which is limited by the total computation resource F_j. Denote f_0 as the computation resource that the MBS assigns to task k, which is a pre-fixed value for each user [18]. r_0 represents the backhaul transmission rate. Therefore, the execution time $T_{ij}^{\text{SBS}}(k)$ of task D_i^k for mobile user i can be expressed as

$$T_{ij}^{\text{SBS}}(k) = \frac{d_{ij}^k}{R_{ij}} + \frac{(d_{ij}^k - {d'}_{ij}^k)c_i^k}{f_{ij}^k} + \frac{{d'}_{ij}^k}{r_0} + \frac{{d'}_{ij}^k c_i^k}{f_0}, \tag{8}$$

where the first term indicates the transmission time for offloading d_{ij}^k via wireless channel from user i to SBS j, the second term is the computation time that SBS j processes $d_{ij}^k - {d'}_{ij}^k$ of task k, the third term is the offloading time on wired line, and the forth term is the computation time for ${d'}_{ij}^k$ which is processed by the MBS.

The energy consumption $E_{ij}^{\text{SBS}}(k)$ is given by

$$E_{ij}^{\text{SBS}}(k) = \frac{P_i d_{ij}^k}{R_{ij}} + (d_{ij}^k - {d'}_{ij}^k)c_i^k e_j + \frac{\delta {d'}_{ij}^k}{r_0} + {d'}_{ij}^k c_i^k e_0, \tag{9}$$

where δ is the offloading power on wired line, e_j and e_0 are the energy consumption per CPU cycle of SBS j and the MBS, respectively. The first term indicates the transmission energy consumption for offloading d_{ij}^k of task k via wireless channel, the second term is the computation energy consumption of $d_{ij}^k - {d'}_{ij}^k$ which is processed by SBS j, the third term is the transmission energy consumption via wired line, and the fourth term is the computation energy consumption of ${d'}_{ij}^k$ processed by the MBS.

3) Offloading to MBS: If mobile user i is directly associated with the MBS, the offloading path of mobile user i is the same as that of mobile user i' in Fig. 2(b). In this case, the execution time $T_{i0}^{\text{MBS}}(k)$ can be written as

$$T_{i0}^{\text{MBS}}(k) = \frac{d_{i0}^k}{R_{i0}} + \frac{d_{i0}^k c_i^k}{f_0}, \tag{10}$$

where $\frac{d_{i0}^k}{R_{i0}}$ represents the transmission time of task k and $\frac{d_{i0}^k c_i^k}{f_0}$ is the computation time.

The energy consumption $E_{i0}^{\text{MBS}}(k)$ can now be expressed as

$$E_{i0}^{\text{MBS}}(k) = \frac{P_i d_{i0}^k}{R_{i0}} + d_{i0}^k c_i^k e_0, \tag{11}$$

where $\frac{P_i d_{i0}^k}{R_{i0}}$ and $d_{i0}^k c_i^k e_0$ are the transmission and computation energy consumption respectively.

D. Multi-Task Model

The application of mobile user i is composed of K tasks. Here, we consider sequential dependency between these K tasks such that each task is processed in order i.e., task k cannot be processed until all of $\{1, \ldots, k-1\}$ tasks have been executed.

Since each task can be processed locally and at the MEC concurrently, the latency of each task k is determined by the maximal value of $\sum_{j=0}^{M} x_{ij} T_{ij}^L(k)$ and $\sum_{j=1}^{M} x_{ij} T_{ij}^{\text{SBS}}(k) + x_{i0} T_{i0}^{\text{MBS}}(k)$. Thus, the completion time of the application can be written as

$$T_i^{Sq} = \sum_{k=1}^{K} \max\left(\sum_{j=0}^{M} x_{ij} T_{ij}^L(k), \sum_{j=1}^{M} x_{ij} T_{ij}^{\text{SBS}}(k) + x_{i0} T_{i0}^{\text{MBS}}(k) \right). \tag{12}$$

As the energy consumption of task k is $E_{ij}^L(k) + E_{ij}^{SBS}(k) + E_{i0}^{MBS}(k)$. The overall energy consumption for processing the

application of mobile user i can be expressed as

$$E_i^{Sq} = \sum_{j=0}^{M} x_{ij} E_{ij}^{L}(k) + \sum_{j=1}^{M} x_{ij} E_{ij}^{SBS}(k) + x_{i0} E_{i0}^{MBS}(k). \tag{13}$$

III. Joint Computation Offloading and User Association for Multi-Tasks

In this section, we first formulate the joint computation offloading and user association as an optimization problem and then decompose it into two sub-problems, i.e., user association and joint optimization of computation offloading. Next, we present the details about user association in Section III-A and present the details about joint optimization of computation offloading in Section III-B. Finally, we propose the joint algorithm for minimizing the overall energy consumption and analyze its complexity in Section III-C.

The objective of joint computation offloading and user association as an optimization problem is to minimize the overall energy consumption. Define $\mathbf{x} = \{x_{ij}\}$ as the vector of user association decision, $\mathbf{f} = \{f_i^k, f_{ij}^k\}$ as computation resource vector, $\mathbf{P} = \{P_i\}$ as transmission power vector, and $\mathbf{d} = \{d_{ij}^k, {d'}_{ij}^k\}$ as computation offloading vector. We formulate the optimization problem as follows:

$$\begin{aligned}
\min \quad & \mathbf{\Pi}(\mathbf{x}, \mathbf{f}, \mathbf{P}, \mathbf{d}) = \sum_{i=1}^{N} \left(\sum_{j=0}^{M} x_{ij} E_{ij}^{L} \right. \\
& \qquad \left. + \sum_{j=1}^{M} x_{ij} E_{ij}^{\text{SBS}} + x_{i0} E_{i0}^{\text{MBS}} \right) \\
\text{s.t.} \quad & T_i^{Sq} \leqslant T_i^{\max}, && \forall i \in \mathcal{U} && \text{(14a)} \\
& \sum_{j=0}^{M} x_{ij} = 1, && \forall i \in \mathcal{U} && \text{(14b)} \\
& \sum_{k=1}^{K} f_i^k \leqslant F_i, && \forall i \in \mathcal{U} && \text{(14c)} \\
& \sum_{i=1}^{N} \sum_{k=1}^{K} x_{ij} f_{ij}^k \leqslant F_j, && \forall j \in \mathcal{B} \setminus \{b_0\} && \text{(14d)} \\
& 0 \leqslant P_i \leqslant P_i^{\max}, && \forall i \in \mathcal{U} && \text{(14e)} \\
& 0 \leqslant {d'}_{ij}^k \leqslant d_{ij}^k \leqslant d_i^k, && \forall i \in \mathcal{U},\, j \in \mathcal{B},\, k \in \mathcal{K} && \text{(14f)} \\
& f_i^k \geqslant 0, && \forall i \in \mathcal{U},\, k \in \mathcal{K} && \text{(14g)} \\
& f_{ij}^k \geqslant 0, && \forall i \in \mathcal{U},\, j \in \mathcal{B} \setminus \{b_0\},\, k \in \mathcal{K} && \text{(14h)} \\
& x_{ij} \in \{0, 1\}, && \forall i \in \mathcal{U},\, j \in \mathcal{B} && \text{(14i)}
\end{aligned}$$

where $E_{ij}^{L} = \sum_{k=1}^{K} E_{ij}^{L}(k)$, $E_{ij}^{\text{SBS}} = \sum_{k=1}^{K} E_{ij}^{\text{SBS}}(k)$, and $E_{i0}^{\text{MBS}} = \sum_{k=1}^{K} E_{i0}^{\text{MBS}}(k)$. Since sequential dependency is considered, the first constraint (14a) guarantees that the application completion time on mobile user i is bounded by the maximum completion deadline $T_i^{\max}$. Constraints (14b) and (14i) ensure that each user is associated with one and only one base station. Constraints (14c), (14d), (14g), and(14h) ensure that the total computation resources assigned to K tasks do not exceed the overall computation capacity of mobile user i and SBS j, respectively. Constraint (14f) indicates the value range of offloading partitions d_{ij}^k and ${d'}_{ij}^k$. Constraint (14e) indicates that the transmission power of mobile user i cannot exceed the maximum transmission power $P_i^{\max}$. The key challenge in solving this problem is the integer constraint $x_{ij} \in \{0, 1\}$, which makes (14) a mixed-integer programming problem and this is in general non-convex and NP-hard [23].

It is quite challenging to optimally solve (14) due to its highly complex coupling among optimization variables and mixed combinatorial feature. Since user association decision $\mathbf{x}$ is an integer variable and the other three variables are continuous, we decompose (14) into two sub-problems:

1) user association: optimization of user association given $\mathbf{f}, \mathbf{P}$ and $\mathbf{d}$.
2) joint optimization of computation offloading: joint optimization of offloading partition, computation resource, and transmission power, for a particular user association decision.

A. User Association

The user association problem for a given offloading partition $\mathbf{d}$, computation resource $\mathbf{f}$, and transmission power $\mathbf{P}$ takes the form

$$\begin{aligned}
\min_{\mathbf{x}} \quad & \sum_{i=1}^{N} \left(\sum_{j=0}^{M} x_{ij} E_{ij}^{L} + \sum_{j=1}^{M} x_{ij} E_{ij}^{\text{SBS}} + x_{i0} E_{i0}^{\text{MBS}} \right) \\
\text{s.t.} \quad & \text{(14a)}, \text{(14b)}, \text{(14d)}, \text{(14i)}
\end{aligned} \tag{15}$$

Problem (15) is still challenging to solve because of nonlinear constraint (14a) and integer constraint (14i). Thus, we first transform (15) into an equivalent form as shown in Lemma 1 and then we propose a separable Semi-Definite Program (SDP) approach to obtain the binary association decision (i.e., the solution of (15)). The SDP approach consists of two steps. In the first step, we use Quadratically Constrained Quadratic Program (QCQP) transformation and semidefinite relaxtion to obtain a fractional solution. In the second step, we use the rounding technique of Shmoys and Tardos [24] to recover the integer value.

Lemma 1: The optimization problem (15) can be transformed into the following equivalent problem:

$$\begin{aligned}
\min_{\mathbf{x}} \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} (E_{ij}^{\text{SBS}} + E_{ij}^{L}) + \sum_{i=1}^{N} x_{i0} (E_{i0}^{\text{MBS}} + E_{i0}^{L}) \\
\text{s.t.} \quad & \sum_{k=1}^{K} t_i^k \leqslant T_i^{\max}, && \forall i \in \mathcal{U} && \text{(16a)} \\
& \sum_{j=0}^{M} x_{ij} T_{ij}^{L}(k) \leqslant t_i^k, && \forall i \in \mathcal{U},\, k \in \mathcal{K} && \text{(16b)}
\end{aligned}$$

$$\sum_{j=1}^{M} x_{ij} T_{ij}^{\mathrm{SBS}}(k) + x_{i0} T_{i0}^{\mathrm{MBS}}(k) \leqslant t_i^k, \quad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{16c}$$

$$x_{ij}(x_{ij}-1) = 0, \qquad \forall i \in \mathcal{U},\, j \in \mathcal{B} \tag{16d}$$

$$(14b), (14d)$$

where $t_i^k = \max\left(\sum_{j=0}^{M} x_{ij}\, T_{ij}^{L}(k), \sum_{j=1}^{M} x_{ij}\, T_{ij}^{\mathrm{SBS}}(k) + x_{i0} T_{i0}^{\mathrm{MBS}}(k)\right)$.

Proof: see Appendix A. ■

1) QCQP Transformation and Semidefinite Relaxtion: Problem (16) is still non-convex since constraint (16d) is a non-convex quadratic constraint. Here using QCQP transformation and semidefinite relaxtion, we can transform the non-convex problem (16) to a convex problem and we can obtain a lower bound by solving this convex problem.

We first transform (16) into an equivalent QCQP form. Define $\mathbf{w}_i \triangleq [x_{i0}, x_{i1}, \ldots, x_{iM}, t_i^1, \ldots, t_i^K]^{\mathrm{T}}$, for all i, and $\mathbf{e}_j$ as the $(M+K+1)\times 1$ standard unit vector with the j-th entry being 1. The optimization problem (16) can now be further transformed into the following equivalent separable QCQP:

$$\min_{\{\mathbf{w}_i\}} \quad \sum_{i=1}^{N} \mathbf{b}_i^{\mathrm{T}} \mathbf{w}_i$$

$$\text{s.t.} \quad (\mathbf{b}_i^{t_i})^{\mathrm{T}} \mathbf{w}_i \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \tag{17a}$$

$$(\mathbf{b}_{ik}^{L})^{\mathrm{T}} \mathbf{w}_i \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{17b}$$

$$(\mathbf{b}_{ik}^{\mathrm{BS}})^{\mathrm{T}} \mathbf{w}_i \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{17c}$$

$$(\mathbf{b}_i^{x})^{\mathrm{T}} \mathbf{w}_i = 1, \qquad \forall i \in \mathcal{U} \tag{17d}$$

$$\sum_{i=1}^{N} (\mathbf{b}_{ij}^{f})^{\mathrm{T}} \mathbf{w}_i \leqslant F_j, \qquad \forall j \in \mathcal{B} \setminus \{b_0\} \tag{17e}$$

$$\mathbf{w}_i^{\mathrm{T}} \mathrm{diag}(\mathbf{e}_j) \mathbf{w}_i - \mathbf{e}_j^{\mathrm{T}} \mathbf{w}_i = 0, \quad \forall i \in \mathcal{U},\, j \in \mathcal{B} \tag{17f}$$

$$\mathbf{w}_i \geqslant 0, \qquad \forall i \in \mathcal{U} \tag{17g}$$

where

$$\mathbf{b}_i = [E_{i0}^{\mathrm{MBS}} + E_{i0}^{L}, E_{i1}^{\mathrm{SBS}} + E_{i1}^{L}, \ldots, E_{iM}^{\mathrm{SBS}} + E_{iM}^{L}, \mathbf{0}_{1\times K}]^{\mathrm{T}}$$

$$\mathbf{b}_i^{t_i} = [\mathbf{0}_{1\times(M+1)}, \mathbf{1}_{1\times K}]^{\mathrm{T}},$$

$$\mathbf{b}_{ik}^{\mathrm{L}} = [T_{i0}^{L}(k), T_{i1}^{L}(k), \ldots, T_{iM}^{L}(k), \mathbf{0}_{1\times(k-1)}, -1, \mathbf{0}_{1\times(K-k)}]^{\mathrm{T}},$$

$$\mathbf{b}_{ik}^{\mathrm{BS}} = [T_{i0}^{\mathrm{MBS}}(k), T_{i1}^{\mathrm{SBS}}(k), \ldots, T_{iM}^{\mathrm{SBS}}(k), \mathbf{0}_{1\times(k-1)}, -1, \mathbf{0}_{1\times(K-k)}]^{\mathrm{T}},$$

$$\mathbf{b}_i^{x} = [\mathbf{1}_{1\times(M+1)}, \mathbf{0}_{1\times K}]^{\mathrm{T}},$$

$$\mathbf{b}_{ij}^{f} = \left[\mathbf{0}_{1\times j}, \sum_{k=1}^{K} f_{ij}^{k}, \mathbf{0}_{1\times(M-j+K)}\right]^{\mathrm{T}}.$$

Now we get the QCQP form (17) however it is a non-convex separable QCQP problem, which is still NP-hard. To find an approximate solution, we need to apply the separable semidefinite relaxation (SDR) [25]. Specifically, we define $\mathbf{Q}_i \triangleq [\mathbf{w}_i^{\mathrm{T}}, 1]^{\mathrm{T}}[\mathbf{w}_i^{\mathrm{T}}, 1]$ and release the rank constraint $\mathrm{rank}(\mathbf{Q}_i) = 1$, such that we can transform (17) into the following problem:

$$\min_{\{\mathbf{q}_i\}} \quad \sum_{i=1}^{N} \mathrm{Tr}(\mathbf{A}_i \mathbf{Q}_i)$$

$$\text{s.t.} \quad \mathrm{Tr}(\mathbf{A}_i^{t_i} \mathbf{Q}_i) \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \tag{18a}$$

$$\mathrm{Tr}(\mathbf{A}_{ik}^{L} \mathbf{Q}_i) \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{18b}$$

$$\mathrm{Tr}(\mathbf{A}_{ik}^{\mathrm{BS}} \mathbf{Q}_i) \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{18c}$$

$$\mathrm{Tr}(\mathbf{A}_i^{x} \mathbf{Q}_i) = 1, \qquad \forall i \in \mathcal{U} \tag{18d}$$

$$\sum_{i=1}^{N} \mathrm{Tr}(\mathbf{A}_{ij}^{f} \mathbf{Q}_i) \leqslant F_j, \qquad \forall j \in \mathcal{B} \setminus \{b_0\} \tag{18e}$$

$$\mathrm{Tr}(\mathbf{A}_{ij}^{e} \mathbf{Q}_i) = 0, \qquad \forall i \in \mathcal{U}, j \in \mathcal{B} \tag{18f}$$

$$\mathbf{Q}_i(M+K+2, M+K+2) = 1, \quad \forall i \in \mathcal{U}, j \in \mathcal{B} \tag{18g}$$

$$\mathbf{Q}_i \succeq 0, \qquad \forall i \in \mathcal{U} \tag{18h}$$

where

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{b}_i \\ \frac{1}{2}\mathbf{b}_i^{\mathrm{T}} & 0 \end{bmatrix}, \qquad \mathbf{A}_i^{t_i} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{b}_i^{t_i} \\ \frac{1}{2}(\mathbf{b}_i^{t_i})^{\mathrm{T}} & 0 \end{bmatrix},$$

$$\mathbf{A}_{ik}^{L} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{b}_{ik}^{L} \\ \frac{1}{2}(\mathbf{b}_{ik}^{L})^{\mathrm{T}} & 0 \end{bmatrix}, \quad \mathbf{A}_{ik}^{BS} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{b}_{ik}^{\mathrm{BS}} \\ \frac{1}{2}(\mathbf{b}_{ik}^{\mathrm{BS}})^{\mathrm{T}} & 0 \end{bmatrix},$$

$$\mathbf{A}_i^{x} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{b}_i^{x} \\ \frac{1}{2}(\mathbf{b}_i^{x})^{\mathrm{T}} & 0 \end{bmatrix}, \qquad \mathbf{A}_{ij}^{f} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{b}_{ij}^{f} \\ \frac{1}{2}(\mathbf{b}_{ij}^{f})^{\mathrm{T}} & 0 \end{bmatrix},$$

$$\mathbf{A}_{ij}^{e} = \begin{bmatrix} \mathrm{diag}(\mathbf{e}_j) & -\frac{1}{2}\mathbf{e}_j \\ -\frac{1}{2}\mathbf{e}_j^{\mathrm{T}} & 0 \end{bmatrix}, \quad \mathbf{0} \triangleq \mathbf{0}_{(M+K+1)\times(M+K+1)}.$$

Problem (18) is convex. The optimal solution $\{\mathbf{Q}_i^*\}$ to (18) can be obtained in polynomial time using standard SDP solver, such as Mosek [26]. Since problem (18) is a relaxation of (16), the optimal solution of problem (18) is the lower bound of the optimal solution of problem (16) if $\{\mathbf{Q}_i^*\}$ does not have rank 1. Therefore, it is necessary to recover a rank-1 solution from $\{\mathbf{Q}_i^*\}$ for the original problem. In the following, we use the rounding technique [24] to recover the integer user association decision $\mathbf{x}$.

2) Integer Association Decisions: The rounding technique consists of the following three steps: 1) obtain a fractional solution from $\{\mathbf{Q}_i^*\}$, 2) construct a weighted bipartite graph to establish the relationship between mobile users and base stations, 3) find an integer matching to obtain the integer solution.

Note that, only the upper-left $(M+1)\times(M+1)$ submatrix of $\mathbf{Q}_i^*$, denoted by $\tilde{\mathbf{Q}}_i^*$, is related to user association solution. In step 1, we define $\mathbf{z}_i \triangleq [z_{i0}, .., z_{iM}] = \mathrm{diag}(\tilde{\mathbf{Q}}_i^*)$ and $z_{ij} \in [0,1]$. Thus, $\mathbf{z}_i$ is the fractional association solution of mobile user i.

In step 2, we construct the weighted bipartite graph $\mathcal{G}(\mathcal{U}, \mathcal{V}, \mathcal{E})$ to establish the relationship between mobile users

Algorithm 1: Construct the Edges of Bipartite Graph $\mathcal{G}$.

1: Set $\mathcal{E} \leftarrow \varnothing$.
2: **if** $J_j \leqslant 1$ **then**
3: There is only one node v_{j1} corresponding to base station j.
4: **for** each $x'_{ij} > 0$ **do**
5: Add edge (u_i, v_{j1}) into $\mathcal{E}$ and set the weight of this edge as $e_{ij1} = x'_{ij}$.
6: **end for**
7: **else**
8: Find the minimum index i_s such that $\sum_{i=1}^{i_s} x'_{ij} \geqslant s$.
9: **if** $i = i_{s-1} + 1, \ldots, i_s - 1$, and $x'_{ij} > 0$ **then**
10: Add edge (u_i, v_{js}) into $\mathcal{E}$ with weight $e_{ijs} = x'_{ij}$.
11: **else if** $i = i_s$ **then**
12: Add edge (u_i, v_{js}) into $\mathcal{E}$ with weight $e_{ijs} = 1 - \sum_{i=1}^{i_s-1} x'_{ij}$. This ensures that the total weight of edges connecting v_{js} is at most 1.
13: **else**
14: Add edge $(u_i, v_{j(s+1)})$ into $\mathcal{E}$with weight $e_{ij(s+1)} = \sum_{i=1}^{i_s} x'_{ij} - s$.
15: **end if**
16: **end if**

and base stations. Set $\mathcal{U}$ represents the mobile users in the network. Set $\mathcal{V} = \{v_{js} : j = 0, 1, \ldots, M, s = 1, \ldots, J_j\}$, where $J_j = \lceil \sum_{i=1}^{N} z_{ij} \rceil$ implies base station j is associated with J_j mobile users. The nodes $\{v_{js:s=1,..,J_j}\}$ correspond to base station j. The most important procedure for constructing graph $\mathcal{G}$ is to set the edges and the edge weight between $\mathcal{U}$ and $\mathcal{V}$. The edges in $\mathcal{G}$ are constructed using Algorithm 1.

In step 3, we utilize the Hungarian algorithm [27] to find a complete max-weighted bipartite matching M_{match}. The M_{match} can be denoted as $\{(u_i, v_{js}, e_{ijs}) : u_i \in \mathcal{U}, v_{js} \in \mathcal{V}, e_{ijs} \in \mathcal{E}\}$ whose total edge weight is the maximum among all matchings. Moreover, since this is a complete matching, all mobile users could find the unique matching point v_{js} (i.e., this matching guarantees that each u_i has one and only one v_{js}).

According to the M_{match}, we obtain the integer user association decision. Specifically, we define $\mathbf{x}' \triangleq [\mathbf{x}'_1, \ldots, \mathbf{x}'_N]^{\mathrm{T}}$, where $\mathbf{x}'_i = [x'_{i0}, \ldots, x'_{iM}]$. If (u_i, v_{js}, e_{ijs}) is in the M_{match}, set $x'_{ij} = 1$; otherwise, $x'_{ij} = 0$. Since $\{z_{ij}\}$ specifies a fractional solution, the rounding result $\{x'_{ij}\}$ determines the integer association (i.e., the solution of (15)). Therefore, we can find an optimal integer solution $\mathbf{x}$ in polynomial time.

B. Joint Optimization of Computation Offloading

Substituting (7), (9), and (11) into (14), the joint optimization of offloading and resource problem for a given $\mathbf{x}$ takes the form

$$\min \quad \Theta(\mathbf{f}, \mathbf{P}, \mathbf{d}) = \sum_{i=1}^{N}\sum_{j=0}^{M}\sum_{k=1}^{K} x_{ij}\varsigma(d_i^k - d_{ij}^k)c_i^k(f_i^k)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K} x_{ij}\left[\frac{P_i d_{ij}^k}{R_{ij}} + (d_{ij}^k - d'^k_{ij})c_i^k e_j + \frac{\delta d'^k_{ij}}{r_0} + d'^k_{ij} c_i^k e_0\right] + \sum_{i=1}^{N}\sum_{k=1}^{K} x_{i0}\left(\frac{P_i d_{i0}^k}{R_{i0}} + d_{i0}^k c_i^k e_0\right)$$

$$\text{s.t.} \quad \text{(14a), (14c)–(14h)} \tag{19}$$

From problem (19), computation resource and transmission power are decoupled from each other in both the objective and the constraints. Thus, the optimization about computation resource and transmission power can be solved independently. However, offloading partitions are still highly coupled with computation resource and transmission power. To solve (19), we further decouple the offloading variables from computation resource variables and transmission power variables to develop low-complexity algorithms. Specifically, we respectively determine $\mathbf{f}$ and $\mathbf{P}$ under given $\mathbf{d}$ and $\mathbf{x}$. Then, obtain $\mathbf{d}$ under given $\mathbf{f}$, $\mathbf{P}$, and $\mathbf{x}$, and repeat this process until convergence.

1) Optimization of Computation Resource: Since $\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} x_{ij} [\frac{P_i d_{ij}^k}{R_{ij}} + (d_{ij}^k - d'^k_{ij})c_i^k e_j + \frac{\delta d'^k_{ij}}{r_0} + d'^k_{ij} c_i^k e_0] + \sum_{i=1}^{N}\sum_{k=1}^{K} x_{i0}(\frac{P_i d_{i0}^k}{R_{i0}} + d_{i0}^k c_i^k e_0)$ is constant for a given $\mathbf{d}$ and $\mathbf{x}$, we formulate the problem to optimize computation resource as

$$\min \quad \phi(\mathbf{f}) = \sum_{i=1}^{N}\sum_{j=0}^{M}\sum_{k=1}^{K} x_{ij}\varsigma(d_i^k - d_{ij}^k)c_i^k(f_i^k)^2$$

$$\text{s.t.} \quad \text{(14a), (14c), (14d), (14g), (14h)} \tag{20}$$

The domain of $\phi(\mathbf{f})$ is convex. The Hessian matrix of $\phi(\mathbf{f})$ is composed of elements either $\partial^2\phi/\partial(f_i^k)^2 > 0$ or 0. Hence, $\phi(\mathbf{f})$ is convex [28]. However, problem (20) is difficult to solve since the objective function is quadratic and constraint (14a) is non-linear.

To solve (20), we introduce two additional auxiliary variables, i.e., $\tau_i^k = \max(\sum_{j=0}^{M} x_{ij} T_{ij}^L(k), \sum_{j=1}^{M} x_{ij} T_{ij}^{\text{SBS}}(k) + x_{i0} T_{i0}^{\text{MBS}}(k))$ and $\eta_{ij}^k = 1/f_{ij}^k$. Thus, constraint (14a) can be transformed into constraints (21a), (21b), (21c), and (21d). Substituting (6) and (8) into (14a), (20) can be reformulated as

$$\min \quad \phi(\mathbf{f}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \varsigma\omega_{ik}(f_i^k)^2$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \tau_i^k \leqslant T_i^{\max}, \qquad \forall\, i \in \mathcal{U} \tag{21a}$$

$$w_{ik} - f_i^k \tau_i^k \leqslant 0, \qquad \forall\, i \in \mathcal{U},\, k \in \mathcal{K} \tag{21b}$$

$$\sum_{j=1}^{M} x_{ij}(d_{ij}^k - d'^k_{ij})c_i^k \eta_{ij}^k + \chi_{ik} \leqslant \tau_i^k, \quad \forall\, i \in \mathcal{U},\, k \in \mathcal{K} \tag{21c}$$

$$\eta_{ij}^k f_{ij}^k = 1, \qquad \forall\, i \in \mathcal{U},\, j \in \mathcal{B} \setminus \{b_0\},\, k \in \mathcal{K}$$

$$\text{(14c), (14d), (14g), (14h)} \tag{21d}$$

where $\chi_{ik} = \sum_{j=1}^{M} x_{ij}(\frac{d_{ij}^k}{R_{ij}} + \frac{d'^k_{ij}}{r_0} + \frac{d'^k_{ij} c_i^k}{f_0}) + x_{i0}\, T_{i0}^{\text{MBS}}(k)$, and $\omega_{ik} = \sum_{j=0}^{M} x_{ij}(d_i^k - d_{ij}^k)c_i^k$.

Define $\mathbf{v}_i \triangleq [f_i^1, \ldots f_i^K, \tau_i^1, \ldots, \tau_i^K, \mathfrak{F}_{i1}^{\mathrm{T}}, \ldots, \mathfrak{F}_{iK}^{\mathrm{T}}, \mathfrak{C}_{i1}^{\mathrm{T}}, \ldots, \mathfrak{C}_{iK}^{\mathrm{T}}]^{\mathrm{T}}$ where $\mathfrak{F}_{ik} \triangleq [f_{i1}^k, \ldots, f_{iM}^k]^{\mathrm{T}}$ and $\mathfrak{C}_{ik} \triangleq [\eta_{i1}^k, \ldots, \eta_{iM}^k]^{\mathrm{T}}$. Let $\mathbf{u}_{ik}$ be the $(2K + 2MK) \times 1$ unit vector with the k-th entry with $\varsigma\omega_{ik}$. The optimization problem (21) can be transformed into a QCQP problem as:

$$\min_{\{\mathbf{v}_i\}} \quad \sum_{i=1}^{N} \mathbf{v}_i^{\mathrm{T}} \mathrm{diag}(\mathbf{u}_{ik}) \mathbf{v}_i$$

$$\text{s.t.} \quad (\mathbf{c}_i^{t_i})^{\mathrm{T}} \mathbf{v}_i \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \tag{22a}$$

$$\mathbf{v}_i^{\mathrm{T}} \mathit{\Gamma}_i^k \mathbf{v}_i + \omega_{ik} \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{22b}$$

$$(\mathbf{c}_{ik}^{\mathrm{f}})^{\mathrm{T}} \mathbf{v}_i + \chi_{ik} \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{22c}$$

$$\mathbf{v}_i^{\mathrm{T}} \Gamma_{ijk}^{SBS} \mathbf{v}_i = 1, \forall i \in \mathcal{U},\, j \in \mathcal{B} \setminus \{b_0\},\, k \in \mathcal{K} \tag{22d}$$

$$(\mathbf{c}_i^L)^{\mathrm{T}} \mathbf{v}_i \leqslant F_i, \qquad \forall i \in \mathcal{U} \tag{22e}$$

$$\sum_{i=1}^{N} (\mathbf{c}_{ij}^{SBS})^{\mathrm{T}} \mathbf{v}_i \leqslant F_j, \qquad \forall j \in \mathcal{B} \setminus \{b_0\} \tag{22f}$$

$$\mathbf{v}_i \geqslant \mathbf{0}, \qquad \forall i \in \mathcal{U} \tag{22g}$$

where

$$\mathbf{c}_i^{t_i} = [\mathbf{0}_{1\times K}, \mathbf{1}_{1\times K}, \mathbf{0}_{1\times(2MK)}]^{\mathrm{T}},$$

$$\mathbb{F}_k = -0.5[\mathbf{0}_{2K\times(k-1)}, \mathbf{e}_{K+k}, \mathbf{0}_{2K\times(K-1)}, \mathbf{e}_k, \mathbf{0}_{2K\times(K-k)}],$$

$$\mathit{\Gamma}_i^k = \begin{bmatrix} \mathbb{F}_k & \mathbf{0}_{2K\times 2MK} \\ \mathbf{0}_{2MK\times 2K} & \mathbf{0}_{2MK\times 2MK} \end{bmatrix},$$

$$\mathbf{c}_{ik}^{\mathrm{f}} = [\mathbf{0}_{1\times(K+k-1)}, -1, \mathbf{0}_{1\times(K+MK+M(k-1)-k)}, x_{i1}(d_{i1}^k - d'^k_{i1})c_i^k, \ldots, x_{iM}(d_{iM}^k - d'^k_{iM})c_i^k, \mathbf{0}_{1\times M(K-k)}]^{\mathrm{T}},$$

$$\mathbb{F}_{jk}^{SBS} = 0.5[\mathbf{0}_{2MK\times(M(k-1)+j-1)}, \mathbf{e}_{MK+M(k-1)+j}, \mathbf{0}_{2MK\times(MK-1)}, \mathbf{e}_{M(k-1)+j}, \mathbf{0}_{2MK\times(MK-M(k-1)-j)}],$$

$$\mathit{\Gamma}_{ijk}^{SBS} = \begin{bmatrix} \mathbf{0}_{2K\times 2K} & \mathbf{0}_{2K\times 2MK} \\ \mathbf{0}_{2MK\times 2K} & \mathbb{F}_{jk}^{SBS} \end{bmatrix},$$

$$\mathbf{c}_i^L = [\mathbf{1}_{1\times K}, \mathbf{0}_{1\times(1+2M)K}]^{\mathrm{T}},$$

$$\mathbf{u}_{ij}^k = [\mathbf{0}_{1\times(j-1)}, x_{ij}, \mathbf{0}_{1\times(M-j)}],$$

$$\mathbf{c}_{ij}^{SBS} = [\mathbf{0}_{1\times 2K}, \mathbf{u}_{ij}^1, .., \mathbf{u}_{ij}^K, \mathbf{0}_{1\times MK}]^{\mathrm{T}}.$$

To solve problem (22), we further transform it into the following equivalent formulation by defining $\mathbf{V}_i \triangleq [\mathbf{v}_i^{\mathrm{T}}, 1]^{\mathrm{T}} [\mathbf{v}_i^{\mathrm{T}}, 1]$,

$$\min_{\{\mathbf{V}_i\}} \quad \sum_{i=1}^{N} \mathrm{Tr}(\mathbf{G}_i \mathbf{V}_i)$$

$$\text{s.t.} \quad \mathrm{Tr}(\mathbf{G}_i^{t_i} \mathbf{V}_i) \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \tag{23a}$$

$$\mathrm{Tr}(\mathbf{G}_{ik}^{\mathit{\Gamma}} \mathbf{V}_i) + \omega_{ik} \leqslant 0, \qquad \forall i \in \mathcal{U},\, k \in \mathcal{K} \tag{23b}$$

$$\mathrm{Tr}(\mathbf{G}_{ik}^{f} \mathbf{V}_i) + \chi_{ik} \leqslant 0, \qquad \forall 1 \in \mathcal{U},\, k \in \mathcal{K} \tag{23c}$$

$$\mathrm{Tr}(\mathbf{G}_{ijk}^{\mathit{\Gamma} SBS} \mathbf{V}_i) = 1, \forall i \in \mathcal{U},\, j \in \mathcal{B} \setminus \{b_0\},\, k \in \mathcal{K} \tag{23d}$$

$$\mathrm{Tr}(\mathbf{G}_i^L \mathbf{V}_i) \leqslant F_i, \qquad \forall i \in \mathcal{U} \tag{23e}$$

$$\sum_{i=1}^{N} \mathrm{Tr}(\mathbf{G}_{ij}^{SBS} \mathbf{V}_i) \leqslant F_j, \quad \forall j \in \mathcal{B} \setminus \{b_0\} \tag{23f}$$

$$\mathbf{V}_i(2K + 2MK + 1, 2K + 2MK + 1) = 1, \forall i \in \mathcal{U} \tag{23g}$$

$$\mathbf{V}_i \succeq 0, \qquad \forall i \in \mathcal{U} \tag{23h}$$

where

$$\mathbf{G}_i = \begin{bmatrix} \mathrm{diag}(\mathbf{u}_{ik}) & \mathbf{0}^c \\ \mathbf{0}^r & 0 \end{bmatrix}, \qquad \mathbf{G}_i^{t_i} = \begin{bmatrix} \mathbf{0}^h & \frac{1}{2}\mathbf{c}_i^{t_i} \\ \frac{1}{2}(\mathbf{c}_i^{t_i})^{\mathrm{T}} & 0 \end{bmatrix},$$

$$\mathbf{G}_{ik}^{\mathit{\Gamma}} = \begin{bmatrix} \mathit{\Gamma}_i^k & \mathbf{0}^c \\ \mathbf{0}^r & 0 \end{bmatrix}, \qquad \mathbf{G}_{ik}^{f} = \begin{bmatrix} \mathbf{0}^h & \frac{1}{2}\mathbf{c}_{ik}^{f} \\ \frac{1}{2}(\mathbf{c}_{ik}^{f})^{\mathrm{T}} & 0 \end{bmatrix},$$

$$\mathbf{G}_{ijk}^{\mathit{\Gamma} SBS} = \begin{bmatrix} \mathit{\Gamma}_{ijk}^{SBS} & \mathbf{0}^c \\ \mathbf{0}^r & 0 \end{bmatrix}, \quad \mathbf{G}_i^L = \begin{bmatrix} \mathbf{0}^h & \frac{1}{2}\mathbf{c}_i^L \\ \frac{1}{2}(\mathbf{c}_i^L)^{\mathrm{T}} & 0 \end{bmatrix},$$

$$\mathbf{G}_{ij}^{SBS} = \begin{bmatrix} \mathbf{0}^h & \frac{1}{2}\mathbf{c}_{ij}^{\mathrm{SBS}} \\ \frac{1}{2}(\mathbf{c}_{ij}^{\mathrm{SBS}})^{\mathrm{T}} & 0 \end{bmatrix}, \quad \mathbf{0}^c \triangleq \mathbf{0}_{(2K+2MK)\times 1},$$

$$\mathbf{0}^r \triangleq \mathbf{0}_{1\times(2K+2MK)}, \qquad \mathbf{0}^h \triangleq \mathbf{0}_{(2K+2MK)\times(2K+2MK)},$$

Problem (23) can be solved optimally and we can extract f from the optimal solution $\{\mathbf{V}_i\}$.

Remark 1: The objective of problem (20) is only related to the energy consumption of mobile users. That is, when solving the original problem (14) which is to minimize the system energy consumption, we always ensure that the energy consumption of mobile users is minimized.

2) Optimization of Transmission Power: Substituting (2)–(5) and (12) into (19), for each mobile user, the transmission power allocation problem for a given $\mathbf{d}$ and $\mathbf{x}$ takes the form

$$\min_{P_i} \quad \sum_{j=1}^{M}\sum_{k=1}^{K} \frac{x_{ij} P_i d_{ij}^k \sum_{i=1}^{N} x_{ij}}{W_s \log\left(1 + \frac{P_i H_{ij}}{\sigma^2 + \sum_{i'=1, i'\neq i}^{N} \sum_{j'=1, j'\neq j}^{M} P_{i'} H_{i'j'}}\right)} + \sum_{k=1}^{K} \frac{x_{i0} P_i d_{i0}^k \sum_{i=1}^{N} x_{i0}}{W_m \log\left(1 + \frac{P_i H_{i0}}{\sigma^2}\right)}$$

$$\text{s.t.} \quad \text{(14a), (14e)} \tag{24}$$

where $x_{ij}\varsigma(d_i^k - d_{ij}^k)c_i^k(f_i^k)^2$, $d_{i0}^k c_i^k e_0$, and $(d_{ij}^k - d'^k_{ij})c_i^k e_j + \frac{\delta d'^k_{ij}}{r_0} + d'^k_{ij} c_i^k e_0$ are omitted, as they are constant in problem (24).

Since there is only one $x_{ij} = 1$, problem (24) can be divided into the following two cases.

If mobile user i is associated with the MBS (i.e., $x_{i0} = 1$). (24) can be reformulated as:

$$\min \quad \Gamma_M(P_i) = \frac{P_i \sum_{k=1}^{K} d_{i0}^k \sum_{i=1}^{N} x_{i0}}{W_m \log(1 + \frac{P_i H_{i0}}{\sigma^2})}$$

$$\text{s.t.} \quad P_i^{\mathrm{low}} \leqslant P_i \leqslant P_i^{\max} \tag{25}$$

According to (14a), we have $\sum_{k=1}^{K} T_{i0}^{MBS}(k) \leqslant \sum_{k=1}^{K} \max\left(T_{i0}^{L}(k), T_{i0}^{\text{MBS}}(k)\right) \leqslant T_i^{\max}$. Thus, $P_i \geqslant \frac{\sigma^2}{H_{i0}}(2^{\Lambda_M} - 1)$ where $\Lambda_M = \frac{\sum_{k=1}^{K} d_{i0}^k \sum_{i=1}^{N} x_{i0}}{W_m\left(T_i^{\max} - \sum_{k=1}^{K} \frac{d_{i0}^k c_i^k}{f_0}\right)}$. As $0 \leqslant P_i \leqslant P_i^{\max}$, the transmission power of mobile user i should satisfy $P_i^{\text{low}} \leqslant P_i \leqslant P_i^{\max}$, where $P_i^{\text{low}} = \frac{\sigma^2}{H_{i0}}(2^{\Lambda_M} - 1)$.

Similarly, if mobile user i is associated with SBS j (i.e., $x_{ij} = 1$), (24) can be rewritten as:

$$\begin{aligned} \min\ \ & \Gamma_S(P_i) = \frac{P_i \sum_{k=1}^{K} d_{ij}^k \sum_{i=1}^{N} x_{ij}}{W_s \log(1 + \frac{P_i H_{ij}}{\sigma^2 + I_{ij}})} \\ \text{s.t.}\ \ & P_i^{\text{low}} \leqslant P_i \leqslant P_i^{\max} \end{aligned} \tag{26}$$

where $P_i^{\text{low}} = \frac{\sigma^2 + I_{ij}}{H_{ij}}(2^{\Lambda_S} - 1)$ and $\Lambda_S = \frac{\sum_{k=1}^{K} d_{ij}^k \sum_{i=1}^{N} x_{ij}}{W_s(T_i^{\max} - B_{ij})}$. $I_{ij} = \sum_{i'=1, i' \neq i}^{N} \sum_{j'=1, j' \neq j}^{M} P_{i'} H_{i'j'}$ is inter-cell interference and $B_{ij} = \sum_{k=1}^{K} (\frac{(d_{ij}^k - d'^k_{ij}) c_i^k}{f_{ij}^k} + \frac{d'^k_{ij}}{r_0} + \frac{d'^k_{ij} c_i^k}{f_0})$.

Since the first-derivation of $\Gamma_M(P_i)$ and $\Gamma_S(P_i)$ are respectively greater than 0, both of $\Gamma_M(P_i)$ and $\Gamma_S(P_i)$ are montonically increasing. Moreover, as $\Gamma_M(0) = 0$ and $\Gamma_S(0) = 0$, the optimal power P_i^* can be obtained at the lower bound point P_i^{low}, which can be expressed as

$$P_i^* = \begin{cases} \dfrac{\sigma^2}{H_{i0}}(2^{\Lambda_M} - 1), & \text{if} \quad x_{i0} = 1 \\ \dfrac{\sigma^2 + I_{ij}}{H_{ij}}(2^{\Lambda_S} - 1), & \text{if} \quad x_{ij} = 1, j \in \mathcal{B} \setminus \{b_0\} \end{cases} \tag{27}$$

Remark 2: Since the optimal power of each mobile user is always equal to the lower bound, the optimal power policy ensures the battery life of mobile users can be extended.

3) Optimization of Computation Offloading: In this subproblem, there are two offloading partition variables: d_{ij}^k and d'^k_{ij}, where d_{ij}^k represents the fraction of each task that needs to be offloaded to the associated base station and d'^k_{ij} represents the fraction of d_{ij}^k will be further be offloaded to the MBS. Given $\mathbf{x}$, $\mathbf{f}$, and $\mathbf{P}$, the offloading problem takes the form

$$\begin{aligned} \min_{\mathbf{d}} \quad & \sum_{i=1}^{N} \sum_{k=1}^{K} \left[\varsigma \left(d_i^k - \sum_{j=0}^{M} x_{ij} d_{ij}^k \right) c_i^k \left(f_i^k\right)^2 + x_{i0} \left(\frac{P_i d_{i0}^k}{R_{i0}} + d_{i0}^k c_i^k e_0 \right) \right] + \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} x_{ij} \left[\frac{P_i d_{ij}^k}{R_{ij}} \right. \\ & \left. + (d_{ij}^k - d'^k_{ij}) c_i^k e_j + \frac{\delta d'^k_{ij}}{r_0} + d'^k_{ij} c_i^k e_0 \right] \\ \text{s.t.} \quad & \text{(14a), (14f)} \end{aligned} \tag{28}$$

Lemma 2: Problem (28) is a convex problem.

Proof: See Appendix B. ■

Problem (28) is convex according to Lemma 2. Although convex, it is still difficult to solve (28) since constraint (14a) is non-linear.

The objective function of (28) can be rewritten as:

$$\begin{aligned} \mathfrak{D}\left(d_{ij}^k, d'^k_{ij}\right) &= \sum_{i=1}^{N} x_{i0} \sum_{k=1}^{K} d_{i0}^k \left(\frac{P_i}{R_{i0}} + c_i^k e_0 - \varsigma c_i^k (f_i^k)^2 \right) \\ &\quad + \sum_{i=1}^{N} x_{ij} \left[\sum_{j=1}^{M} d_{ij}^k \left(\frac{P_i}{R_{ij}} + c_i^k e_j - \varsigma c_i^k (f_i^k)^2 \right) \right. \\ &\quad \left. + \sum_{j=1}^{M} d'^k_{ij} \left(\frac{\delta}{r_0} + c_i^k e_0 - c_i^k e_j \right) \right] \\ &= \sum_{i=1}^{N} x_{i0} \mathfrak{D}_M(d_{i0}^k) + \sum_{i=1}^{N} x_{ij} \mathfrak{D}_S(d_{ij}^k, d'^k_{ij}). \end{aligned} \tag{29}$$

where $\mathfrak{D}_M(d_{i0}^k) = \sum_{k=1}^{K} d_{i0}^k (\frac{P_i}{R_{i0}} + c_i^k e_0 - \varsigma c_i^k (f_i^k)^2)$ and $\mathfrak{D}_S(d_{ij}^k, d'^k_{ij}) = \sum_{j=1}^{M} d_{ij}^k (\frac{P_i}{R_{ij}} + c_i^k e_j - \varsigma c_i^k (f_i^k)^2) + \sum_{j=1}^{M} d'^k_{ij} (\frac{\delta}{r_0} + c_i^k e_0 - c_i^k e_j)$.

From (29), we observe that the objective function $\mathfrak{D}(d_{ij}^k, d'^k_{ij})$ of problem (28) can be divided into two decoupled parts: $\mathfrak{D}_M(d_{i0}^k)$ and $\mathfrak{D}_S(d_{ij}^k, d'^k_{ij})$. Furthermore, there is only one $x_{ij} = 1$. Thus, based on the value of x_{ij}, we could simplify both objective function and constraints of problem (28) into the following two cases.

If mobile user i is associated with the MBS, we split (14a) into two constraints (30a) and (30b). Thus, for each mobile user, the computation offloading problem becomes:

$$\begin{aligned} \min\ \ & \mathfrak{D}_M(d_{i0}^k) = \sum_{k=1}^{K} d_{i0}^k \left(\frac{P_i}{R_{i0}} + c_i^k e_0 - \varsigma c_i^k (f_i^k)^2 \right) \\ \text{s.t.}\ \ & \sum_{k=1}^{K} T_{i0}^L(k) \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \qquad \text{(30a)} \\ & \sum_{k=1}^{K} T_{i0}^{MBS}(k) \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \qquad \text{(30b)} \\ & 0 \leqslant d_{i0}^k \leqslant d_i^k, \qquad \forall i \in \mathcal{U},\ k \in \mathcal{K} \qquad \text{(30c)} \end{aligned}$$

If mobile user i is associated with SBS j, we split (14a) into two constraints (31a) and (31b). The computation offloading problem of each mobile user can be rewritten as:

$$\begin{aligned} \min\ \ & \mathfrak{D}_S(d_{ij}^k, d'^k_{ij}) = \sum_{k=1}^{K} \left(d_{ij}^k \left(\frac{P_i}{R_{ij}} + c_i^k e_j - \varsigma c_i^k (f_i^k)^2 \right) \right. \\ & \left. + d'^k_{ij} \left(\frac{\delta}{r_0} + c_i^k e_0 - c_i^k e_j \right) \right) \\ \text{s.t.}\ \ & \sum_{k=1}^{K} T_{ij}^L(k) \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \qquad \text{(31a)} \\ & \sum_{k=1}^{K} T_{ij}^{SBS}(k) \leqslant T_i^{\max}, \qquad \forall i \in \mathcal{U} \\ & \text{(14f)} \qquad \text{(31b)} \end{aligned}$$

Algorithm 2: Computation Offloading by Jointly Optimize Computation Resource, Transmission Power, and Offloading Partition for a Given User Association.

Set $v_1 = 0$ and the maximum tolerance $\epsilon_1 > 0$.
repeat
  Solve problem (23) to find computation resource allocation $\mathbf{f}_{v_1}$ for each computation task based on $\mathbf{x}$ and $\mathbf{d}$;
  Calculate transmission power $\mathbf{P}_{v_1}$ of each mobile user using (27);
  for each mobile user **do**
    if $x_{i0} = 1$ **then**
      Solve problem (30) to obtain the offloading partition $\{d_{i0}^k, d_{i0}'^k\}$ for each computation task;
    else
      Solve problem (31) to obtain the offloading partition $\{d_{ij}^k, d_{ij}'^k\}$ for each computation task;
    end if
  end for
  Update $v_1 = v_1 + 1$;
Until $|\Theta_{v_1}(\mathbf{f}, \mathbf{P}, \mathbf{d}) - \Theta_{v_1-1}(\mathbf{f}, \mathbf{P}, \mathbf{d})| \leqslant \epsilon_1$.

It is easy to verify (30) and (31) are linear convex problems. Thus, we can obtain the optimal solution within polynomial time.

Combining computation resource allocation, transmission power allocation, and computation offloading, we propose Algorithm 2 to solve problem (19) for a given user association.

At each iteration of Algorithm 2, the computational complexity of solving convex problems (23), (30), and (31) are polynomial in the number of variables and constraints. Specifically, problem (23) is to allocate the computation resource with $a = (NK + NKM)$ decision variables, and $b = (3N + M + 2NK + NMK)$ constraints. The worst-case computational complexity required to solve (23) is $\mathcal{O}((a^2b + a^3)b^{1/2}\log(1/\varepsilon))$ given a solution accuracy $\varepsilon > 0$ using interior-point method [29]. Problem (30) is with K decision variables and $2N + NK$ constraints such that the complexity is $\mathcal{O}((K + 2N + NK)K^2\sqrt{2N + NK}$. Similarly, problem (31) is with $(2K)$ decision variables and $(2N + NK)$ linear constraints such that the complexity is $\mathcal{O}((2K + 2N + NK))4K^2\sqrt{2N + NK}$.

C. Joint Optimization of Computation Offloading and User Association

The proposed joint optimization of user association, resource allocation, and computation offloading for minimizing the overall energy consumption is summarized in Algorithm 3. It is an alternating optimization framework which requires solving two sub-problems (15) and (19) and repeating until convergence. Note that solving one instant of (15) and (19) alone already provides a better point, we use them as an alternating descent to achieve a much faster convergence. After a finite number of iterations, we can get a solution of problem (14) for a given error tolerance $\epsilon_2 > 0$.

Algorithm 3: Joint Optimization for User Association and Computation Offloading (J-UACO).

Initialize $x_{ij} = 1$, $P_i = P_i^{\max}$, $\{f_i^k, f_{ij}^k\} = \{\frac{F_i}{K}, \frac{F_j}{N}\}$, $\{d_{ij}^k, d_{ij}'^k\} = \{d_i^k, 0\}$ for all $i \in \mathcal{U}$, $j \in \mathcal{B}$, $k \in \mathcal{K}$;
Set $v_2 = 0$ and the maximum tolerance $\epsilon_2 > 0$.
repeat
  Based on $\mathbf{f}$, $\mathbf{d}$, and $\mathbf{P}$, solve user association problem (15) using SDP and rounding method;
  Allocate the computation resources $\mathbf{f}$ and transmission power $\mathbf{P}$, and compute the offloading value $\mathbf{d}$ by calling Algorithm 2 based on $\mathbf{x}$;
  Update $v_2 = v_2 + 1$;
Until $|\frac{\Pi_{v_2}(\mathbf{x},\mathbf{f},\mathbf{P},\mathbf{d}) - \Pi_{v_2-1}(\mathbf{x},\mathbf{f},\mathbf{P},\mathbf{d})}{\Pi_{v_2-1}(\mathbf{x},\mathbf{f},\mathbf{P},\mathbf{d})}| \leqslant \epsilon_2$

To solve problem (15) , at each iteration, we need to solve the SDP problem (18) and rounding. Problem (18) is with $a' = (N(M + 1))$ decision variables and $b' = (3N + 2NK + M + 2N(M + 1))$ constraints such that the worst-case complexity is $\mathcal{O}((a'^2b + a'^3)b'^{1/2}\log(1/\varepsilon'))$ given a solution accuracy $\varepsilon' > 0$. The complexity of rounding is polynomial in the number of nodes and edges, that is $\mathcal{O}(|\mathcal{V}||\mathcal{E}|)$.

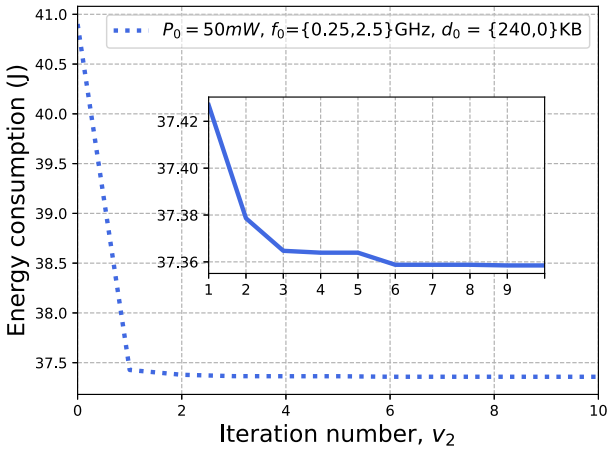
IV. Numerical Results

We consider a network topology of 1000 m × 1000 m consisting of one MBS , $M = 6$ SBSs, and $N = 40$ mobile users. The location of the MBS is fixed at the center of the network while both SBSs and mobile users are randomly distributed. The maximum transmission power of mobile user, P_i is set to 100 mW. The channel gain models presented in 3GPP standardization [30] are adopted here. Specifically, large scale fading of the channel between MBS and mobile user is modeled as $128.1 + 37.6\ \log_{10}(r) + \mu$, and that between SBS and mobile user is $140.7 + 36.7\ \log_{10}(r) + \mu$, where r is in km and μ follows log-normal distribution N (0, 8 dB). The Rayleigh fading model is adopted for small scale fading. The noise power is $\sigma^2 = 10^{-11}$ mW and the interference threshold is $I = -90$ dBm. The bandwidth of MBS and SBS are 10 MHz and 5 MHz.

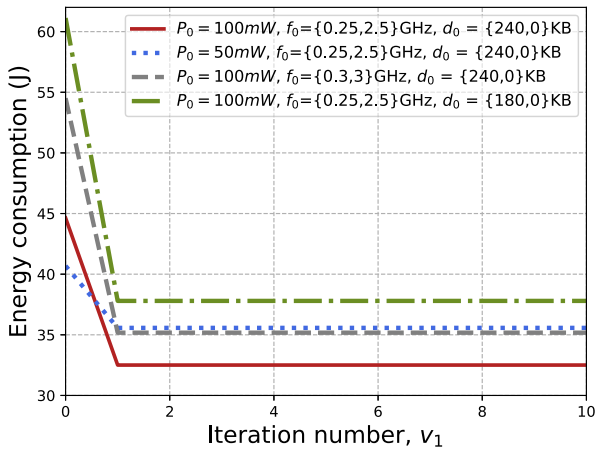
For the computation task, we consider the augmented reality application in [5], which is composed of 3 computation-intensive and separable tasks. Both the data size of each task and required number of CPU cycles per bit follow the uniform distribution with $d_i^{(k)} \sim U[200, 500]$ KB and $c_i^{(k)} \sim U[50, 100]$ cycles/bit. The CPU computation capabilities of mobile user and SBS are $F_i = 1$ GHz and $F_j = 20$ GHz, respectively. Due to the high computation resource of the MBS, the pre-determined f_0 is 5 GHz. In addition, $T_i^{\max} \sim U[5, 10]$ s, $e_j = e_0 = 1$ W/GHz [13], [31], $r_0 = 1$ Gbps, $\varsigma = 1 \times 10^{-25}$, and $\delta = 1$ mW.

To verify the performance of our proposed J-UACO algorithm, we introduce the two following benchmark polices,

- *Full offloading (FO):* All computation tasks are offloaded to the MEC server on the MBS for remote computing. In this policy, all mobile users are associated with the MBS.



(a) Inner loop



(b) Outer loop

Fig. 3. Convergence of the J-UACO under different initial points with $N = 30, K = 3$.

The feasible computation resource of each task can be obtained as $f_i^k = \frac{F_i d_i^k c_i^k}{\sum_{k=1}^{K} d_i^k c_i^k}$.

- *Partial offloading without user association (PO):* Given a feasible user association, each mobile user offloads its computation tasks to the associated base station (SBS or MBS). The computation resource, transmission power, and offloading are jointly optimized as described in Section III-B, as in [8].

A. Convergence

In this subsection, we evaluate the convergence of the proposed J-UACO algorithm under different initial points. Specifically, P_0 is the initial value of transmission power. The first value of f_0 is the initial value of f_i^k and the second is the initial value of f_{ij}^k. Similarly, the first value of d_0 is the initial value of d_{ij}^k and the second is the initial value of d'^k_{ij}.

Fig. 3(a) plots the convergence of the inner loop of the proposed J-UACO, i.e., Algorithm 2. We observe that it has a fast convergence rate and converges typically within 8 iterations. Fig. 3(b) displays the convergence of the outer loop of the proposed J-UACO, i.e., Algorithm 3. We observe that the proposed J-UACO converges within 2 iterations. Thus, J-UACO is cost efficient in terms of computational complexity.

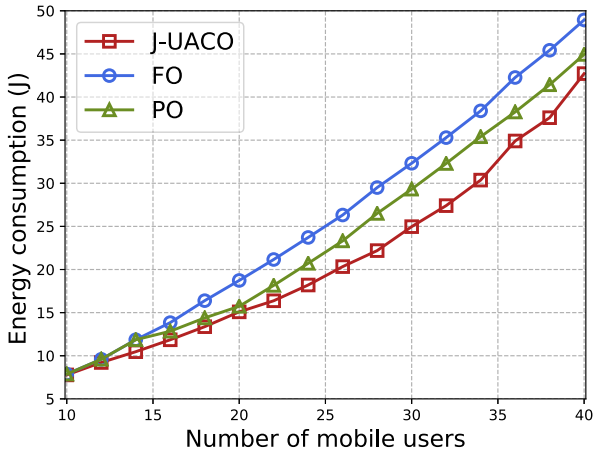


Fig. 4. Comparison of energy consumption of the number of mobile users under different policies with $K = 3$.

B. Performance Evaluation

In this subsection, we compare the proposed J-UACO algorithm with other two schemes, i.e., FO and PO. Fig. 4 plots the comparison of energy consumption with respect to the number of mobile users.

From Fig. 4, we can draw several observations. First, the energy consumption of the proposed J-UACO, as well as FO and PO, increases rapidly as the number of mobile users becomes large. The reason is that energy consumption is linearly related to the number of mobile users. Second, the performance of the proposed J-UACO significantly outperforms the two benchmark policies since J-UACO jointly optimizes user association, computation resource, transmission power, and offloading. More specifically, compared to J-UACO, the performance of PO is seriously affected by the initial deployment of user association owing to lack of user association optimization. Further, among the three policies, the energy consumption of FO is the worst. The reason is that FO always makes all mobile users offload their tasks to the MBS which leads to severe overloading of the MBS and increases energy consumption of the communication link. On the contrary, J-UACO and PO utilize small cells to share heavy load of the MBS to reduce the energy consumption.

C. Impact of Separable Tasks

In this subsection, we examine the impact of the number of separable tasks K on energy consumption. Fig. 5 denotes the energy consumption of J-UACO, FO, and PO when K varies from 2 to 5. Here, we set $N = 15$.

From the results, we can see that the energy consumption of the above three policies respectively increases with the increasement of K. This implies that the growth of K leads to more offloading requests, which results in the need and consumption of more communication resource and computation resource. Moreover, the performance of the proposed J-UACO

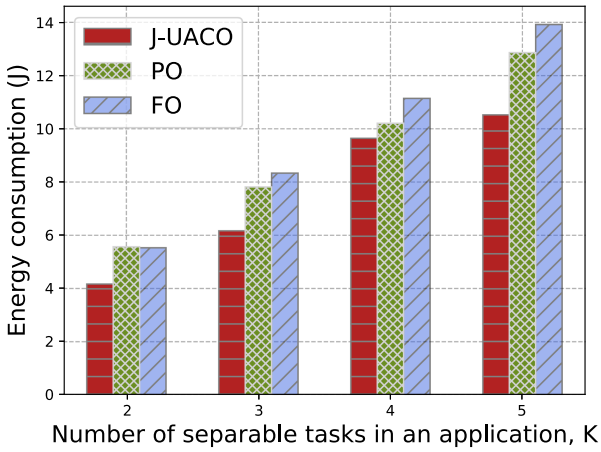


Fig. 5. Comparison of energy consumption of the number of separable tasks under different policies with $N = 15$.

is the best in the three policies since it jointly optimizes user association and offloading. The worst performance is experienced by FO, which offloads all tasks to the MBS creating a communication resource bottleneck.

V. Conclusion

In this paper, we proposed a two-tier computation offloading framework for multi-tasking in heterogeneous networks. The dependency relationship of these tasks is sequential. We formulated a joint computation offloading and user association problem for minimizing the energy consumption of mobile users and MEC servers. To solve this problem, we decomposed it in to two sub-problems, i.e., user association and computation offloading. The user-association sub problem determines whether a mobile user can associate with a particular base station. The sub-problem of computation offloading was to jointly optimize computation resource, transmission power, and offloading partition. Then an efficient joint optimization algorithm was developed by jointly optimizing user association and computation offloading where computation resource allocation and transmission power allocation are also considered. Numerical results demonstrated that the proposed computation offloading algorithm outperforms the benchmark policies under various system parameters.

Appendix A
Proof of Lemma 1

We introduce an additional auxiliary variable t_i^k, which is defined as

$$t_i^k = \max\left(\sum_{j=0}^{M} x_{ij} T_{ij}^{L}(k), \sum_{j=1}^{M} x_{ij} T_{ij}^{\mathrm{SBS}}(k) + x_{i0} T_{i0}^{\mathrm{MBS}}(k)\right) \tag{32}$$

Thus we have

$$\sum_{j=0}^{M} x_{ij} T_{ij}^{L}(k) \leqslant t_i^k,$$

$$\sum_{j=1}^{M} x_{ij} T_{ij}^{\mathrm{SBS}}(k) + x_{i0} T_{i0}^{\mathrm{MBS}}(k)) \leqslant t_i^k, \tag{33}$$

Constraint (14a) is equivalent to

$$\sum_{k=1}^{K} t_i^k \leqslant T_i^{\max},$$

$$\sum_{j=0}^{M} x_{ij} T_{ij}^{L}(k) \leqslant t_i^k,$$

$$\sum_{j=1}^{M} x_{ij} T_{ij}^{\mathrm{SBS}}(k) + x_{i0} T_{i0}^{\mathrm{MBS}}(k)) \leqslant t_i^k, \tag{34}$$

From the above transformations, we transform constraint (14a) from non-linear constraint to linear constraints.

As every integer x_{ij} satisfies $x_{ij} \in \{0, 1\}$, the integer constraint (14i) can be equivalently rewritten as a Boolean constraint $x_{ij}(x_{ij} - 1) = 0$. This transformation makes constraint (14i) become an equation-constraint form.

Appendix B
Proof of Lemma 2

Since $E_i^L(k)$, $E_{i0}^{MBS}(k)$ and $E_{ij}^{SBS}(k)$ are convex function w.r.t d_{ij}^k and d'^k_{ij}. Thus, the objective function, the summation of a set of convex functions, preserves the convexity. The $T_i^L(k)$, $T_{i0}^{MBS}(k)$ and $T_{ij}^{SBS}(k)$ are also convex function w.r.t d_{ij}^k and d'^k_{ij}. Since the maximum function does not change the convexity, (14a) is a convex constraint. Therefore, problem (28) is a convex problem.

References

[1] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[2] J. Ren, Y. Guo, D. Zhang, Q. Liu, and Y. Zhang, "Distributed and efficient object detection in edge computing: Challenges and solutions," *IEEE Netw.*, vol. 32, no. 6, pp. 137–143, Nov./Dec. 2018.

[3] Q. Han, B. Yang, G. Miao, C. Chen, X. Wang, and X. Guan, "Backhaul-aware user association and resource allocation for energy-constrained hetnets," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 580–593, Jan. 2017.

[4] D. Liu *et al.*, "User association in 5g networks: A survey and an outlook," *IEEE Commun. Surv. Tut.*, vol. 18, no. 2, pp. 1018–1044, Apr.–Jun. 2016.

[5] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 398–401, Jun. 2017.

[6] S. E. Mahmoodi, R. Uma, and K. Subbalakshmi, "Optimal joint scheduling and cloud offloading for mobile applications," *IEEE Trans. Cloud Comput.*, 2016, doi: 10.1109/TCC.2016.2560808.

[7] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[8] 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.

[9] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[10] 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.

[11] X. Hu, K.-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.
[12] Y. Yu, J. Zhang, and K. B. Letaief, "Joint subcarrier and cpu time allocation for mobile edge computing," in *Proc. IEEE Global Commun. Conf.*, 2016, pp. 1–6.
[13] K. Zhang *et al.*, "Energy-efficient offloading for mobile edge computing in 5g heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.
[14] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Comm.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.
[15] C. Wang, F. R. Yu, C. Liang, Q. Chen, and L. Tang, "Joint computation offloading and interference management in wireless cellular networks with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7432–7445, Aug. 2017.
[16] Y. Mao, J. Zhang, S. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994–6009, Sep. 2017.
[17] W. Chen, D. Wang, and K. Li, "Multi-user multi-task computation offloading in green mobile edge cloud computing," *IEEE Trans. Serv. Comput.*, 2018, doi: 10.1109/TSC.2018.2826544.
[18] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading and resource allocation for computation and communication in mobile cloud with computing access point," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2017, pp. 1–9.
[19] A. H. Jafari, D. López-Pérez, H. Song, H. Claussen, L. Ho, and J. Zhang, "Small cell backhaul: challenges and prospective solutions," *EURASIP J. Wireless Commun. Netw.*, vol. 206, pp. 1–18, 2015.
[20] O. Munoz, A. Pascual-Iserte, and J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4738–4755, Oct. 2015.
[21] W. Lu, Q. Fan, Z. Li, and H. Lu, "Power control based time-domain inter-cell interference coordination scheme in DSCNs," in *Proc. IEEE Int. Conf. Commun.*, 2016, pp. 1–6.
[22] T. Zahir, K. Arshad, A. Nakata, and K. Moessner, "Interference management in femtocells," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 1, pp. 293–311, Jan.–Mar. 2013.
[23] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.
[24] D. B. Shmoys and É. Tardos, "An approximation algorithm for the generalized assignment problem," *Math. Program.*, vol. 62, no. 1–3, pp. 461–474, 1993.
[25] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
[26] "Mosek optimization software." 2018. [Online]. Available: https://www.mosek.com/
[27] H. W. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Logistics*, vol. 2, no. 1/2, pp. 83–97, 1955.
[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[29] Y. Nesterov and A. Nemirovskii, *Interior-point Polynomial Algorithms in Convex Programming*. Philadelphia, PA, USA: SIAM, 1994, vol. 13.
[30] J. Ikuno, M. Wrulich, and M. Rupp, "3GPP tr 36.814 v9. 0.0-evolved universal terrestrial radio access (e-utra); further advancements for e-utra physical layer aspecs," Rep. 3GPP TR 36.814 V9.0.0, 2010.
[31] M. Wittmann, G. Hager, T. Zeiser, J. Treibig, and G. Wellein, "Chip-level and multi-node analysis of energy-optimized lattice Boltzmann CFD simulations," *Concurrency Comput., Practice Experience*, vol. 28, no. 7, pp. 2295–2315, 2016.

Yueyue Dai received the B.Sc. degree in communication and information engineering, in 2014, from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, where she is currently working toward the Ph.D. degree. She is currently a visiting Ph.D. student with the University of Oslo, Norway. Her current research interests include wireless network, mobile edge computing, Internet of Vehicles, blockchain, and deep reinforcement learning.

Du Xu received the B.S., M.S., and Ph.D. degrees from Southeast University, Nanjing, China, and the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1990, 1995, and 1998, respectively. He is a Professor with the UESTC. His research interests include network modeling and performance analysis, switching and routing, network virtualization and security. He presided over many advanced research projects, including NSFC, National 863 Plans and National key Research and Development Program of China.

Sabita Maharjan (M'09) received the Ph.D. degree in networks and distributed systems from the Simula Research Laboratory, Fornebu, Norway, and the University of Oslo, Oslo, Norway, in 2013. She is currently a Senior Research Scientist with the Simula Metropolitan Center for Digital Engineering, Norway, and an Associate Professor with the University of Oslo. Her current research interests include wireless networks, network security and resilience, smart grid communications, Internet of Things, machine-to-machine communication, software-defined wireless networking, and the Internet of Vehicles.

Yan Zhang received a Ph.D. degree from the School of Electrical & Electronics Engineering, Nanyang Technological University, Singapore. He is Full Professor with the Department of Informatics, University of Oslo, Norway. He is an Associate Technical Editor of the IEEE Communications Magazine, an Editor of the IEEE Network Magazine, an Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, an Editor of the IEEE Communications Surveys & Tutorials, an Editor of the IEEE Internet of Things Journal, an Editor of the IEEE Vehicular Technology Magazine, and an Associate Editor of the IEEE Access. He serves as chair positions in a number of conferences, including the IEEE GLOBECOM 2017, the IEEE VTC-Spring 2017, the IEEE PIMRC 2016, the IEEE CloudCom 2016, the IEEE ICCC 2016, the IEEE CCNC 2016, the IEEE SmartGridComm 2015, and the IEEE CloudCom 2015. He serves as TPC member for numerous international conferences including the IEEE INFOCOM, the IEEE ICC, the IEEE GLOBECOM, and the IEEE WCNC. His current research interests include next-generation wireless networks leading to 5G Beyond, green and secure cyber-physical systems (e.g., smart grid, healthcare, and transport). He is the IEEE VTS (Vehicular Technology Society) Distinguished Lecturer. He is also a Senior Member of the IEEE ComSoc, the IEEE CS, the IEEE PES, and the IEEE VT society.