Joint Computation and Communication Cooperation for Energy-Efficient Mobile Edge Computing

Xiaowen Cao, *Student Member, IEEE*, Feng Wang, *Member, IEEE*, Jie Xu, *Member, IEEE*, Rui Zhang, *Fellow, IEEE*, and Shuguang Cui, *Fellow, IEEE*

***Abstract*—This paper proposes a novel *user cooperation* approach in both computation and communication for mobile edge computing (MEC) systems to improve the energy efficiency for latency-constrained computation. We consider a basic three-node MEC system consisting of a user node, a helper node, and an access point (AP) node attached with an MEC server, in which the user has latency-constrained and computation-intensive tasks to be executed. We consider two different computation offloading models, namely, the partial and binary offloading, respectively. For partial offloading, the tasks at the user are divided into three parts that are executed at the user, helper, and AP, respectively; while for binary offloading, the tasks are executed as a whole only at one of three nodes. Under this setup, we focus on a particular time block and develop an efficient four-slot transmission protocol to enable the *joint computation and communication cooperation*. Besides the local task computing over the whole block, the user can offload some computation tasks to the helper in the first slot, and the helper cooperatively computes these tasks in the remaining time; while in the second and third slots, the helper works as a cooperative relay to help the user offload some other tasks to the AP for remote execution in the fourth slot. For both cases with partial and binary offloading, we jointly optimize the computation and communication resources allocation at both the user and the helper (i.e., the time and transmit power allocations for offloading, and the central process unit frequencies for computing), so as to minimize their total energy consumption while satisfying the user's computation latency constraint. Although the two problems are nonconvex in general, we develop efficient algorithms to solve them optimally. Numerical results show that the proposed joint computation and communication cooperation approach significantly improves the computation capacity and energy efficiency at the user and helper, as compared to other benchmark schemes without such a joint design.**

***Index Terms*—Computation offloading, joint computation and communication cooperation, mobile edge computing (MEC), resource allocation.**

I. Introduction

RECENT advancements in the fifth-generation (5G) cellular technologies have enabled various new applications, such as the augmented reality (AR), autonomous driving, and Internet of Things (IoT). These applications demand ultralow-latency communication, computation, and control among a large number of wireless devices (e.g., sensors and actuators) [2]. In practice, the real-time computation tasks to be executed can be quite intensive, but wireless devices are generally of small size and only have limited communication, computation, and storage resources (see [3]). Therefore, how to enhance their computation capabilities and reduce the computation latency is one crucial but challenging issue to be tackled for making these 5G applications a reality.

Conventionally, mobile cloud computing (MCC) has been widely adopted to enhance wireless devices' computation capabilities, by moving their computing and data storage to the remote centralized cloud [4]. However, as the cloud servers are normally distant from wireless devices, MCC may not be able to meet the stringent computation latency requirements for emerging 5G applications. To overcome such limitations, mobile edge computing (MEC) has been recently proposed as a new solution to provide cloud-like computing at the edge of wireless networks [e.g., access points (APs) and cellular base stations (BSs)], by deploying distributed MEC servers therein [3]–[9]. In MEC, wireless devices can offload computation-intensive and latency-critical tasks to APs/BSs in close proximity for remote execution, thus achieving much lower computation latency.

The computation offloading design in MEC systems critically relies on tractable computation task models. Two widely adopted task models in the MEC literature are *binary* and *partial* offloading, respectively, [6], [7]. In binary offloading, the computation tasks are not partitionable, and thus should be executed as a whole via either local computing at the user or offloading to the MEC server. This practically corresponds to highly integrated or relatively simple tasks, such

Manuscript received May 15, 2018; revised July 25, 2018; accepted October 6, 2018. Date of publication October 10, 2018; date of current version June 19, 2019. This work was supported in part by the Natural Science Foundation of China under Grant 61871137 and Grant 61629101, in part by the Natural Science Foundation under Grant DMS-1622433, Grant AST-1547436, and Grant ECCS-1659025, in part by the Shenzhen Fundamental Research Fund under Grant KQTD2015033114415450, Grant ZDSYS201707251409055, and Grant 2017ZT07X152, and in part by the Natural Science Foundation of Guangdong Province under Grant 2018A030310537. Part of this paper has been presented at the International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks Workshop on Edge and Fog Computing for Intelligent IoT Applications, Shanghai, China, May 7–11, 2018 [1]. *(Corresponding author: Feng Wang.)*

X. Cao, F. Wang, and J. Xu are with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: caoxwen@outlook.com; fengwang13@gdut.edu.cn; jiexu@gdut.edu.cn).

R. Zhang is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: elezhang@nus.edu.sg).

S. Cui is with the Shenzhen Research Institute of Big Data and School of Science and Engineering, Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616 USA (e-mail: robert.cui@gmail.com).

Digital Object Identifier 10.1109/JIOT.2018.2875246

as those in speech recognition and natural language translation. In contrast, for partial offloading, the computation tasks need to be partitioned into two or more independent parts, which can be executed in parallel by local computing and offloading. This corresponds to applications with multiple fine-grained procedures/components, in, e.g., AR applications [7]. Based on the binary and partial offloading models, the prior works (see [9]–[17]) investigated the joint computation and communication resources allocation to improve the performance of MEC. For example, Liu *et al.* [10] and Zhang *et al.* [11] considered a single-user MEC system with dynamic task arrivals and channel fading, in which the user jointly optimizes the local computing or offloading decisions to minimize the computation latency, subject to the computation and communication resource constraints. You *et al.* [12], Wang *et al.* [13], and Chen *et al.* [14] investigated the energy-efficient design in multiuser MEC systems with multiple users offloading their respective tasks to a single AP/BS for execution, in which the objective is to minimize the users' energy consumption while ensuring their computation latency requirements. Furthermore, [15]–[17] proposed wireless powered MEC systems by integrating the emerging wireless power transfer (WPT) technology into MEC for self-sustainable computing, where the AP employs WPT to power the users' local computing and offloading.

Despite the recent research progress, multiuser MEC designs still face several technical challenges. First, the computation resources at the MEC server and the communication resources at the AP should be shared among the actively-computing users. When the user number becomes large, the computation and communication resources allocated to each user are fundamentally limited, thus compromising the benefit of MEC. Next, due to the signal propagation loss over distances, far-apart users may spend much more communication resources than nearby users for offloading, which results in a near-far user fairness issue. Note that the 5G networks are expected to consist of massive wireless devices with certain computation and communication resources. Due to the burst nature of wireless traffic, each active device is highly likely to be surrounded by some idle devices with unused or additional resources. As such, in this paper we propose a novel *joint computation and communication cooperation* approach in multiuser MEC systems, such that the nearby users are enabled as helpers to share their computation and communication resources to help actively computing users, thereby improve the MEC computation performance.

In this paper, we consider a basic three-node MEC system with user cooperation, which consists of a user node, a helper node, and an AP node attached with an MEC server. Here, the helper node can be an IoT sensor, a smart phone, or a laptop, which is nearby and has certain computation and communication resources.[1] We focus on the user's latency-constrained computation within a given time block. To implement the joint computation and communication cooperation, the block is divided into four time slots. Specifically, in the first slot, the user offloads some computation tasks to the helper for remote execution. In the second and third slots, the helper acts as a decode-and-forward (DF) relay to help the user offload some other computation tasks to the AP, for remote execution at the MEC server in the fourth slot. Under this setup, we pursue an energy-efficient user cooperation MEC design for both the partial and binary offloading cases, by jointly optimizing the computation and communication resource allocations. The main results of this paper are summarized as follows.

1) First, for the partial offloading case, the user's computation tasks are partitioned into three parts for local computation, offloading to helper, and offloading to AP, respectively. Toward minimizing the total energy consumption at both the user and the helper subject to the user's computation latency constraint, we jointly optimize the task partition of the user, the central process unit (CPU) frequencies for local computing at both the user and the helper, as well as the time and transmit power allocations for offloading. The nonconvex problem of interests in general can be reformulated into a convex one. Leveraging the Lagrange duality method, we obtain the globally optimal solution in a semi-closed form.
2) Next, for the binary offloading case, the user should execute the nonpartitionable computation tasks by choosing one among three computation modes, i.e., the local computing at the user, computation cooperation (offloading to the helper), and communication cooperation (offloading to the AP). Solving the resultant latency-constrained energy minimization problem, we develop an efficient optimal algorithm, by first choosing the computation mode and then optimizing the corresponding joint computation and communication resources allocation.
3) Finally, extensive numerical results show that the proposed joint computation and communication cooperation approach achieves significant performance gains in terms of both the computation capacity and the system energy efficiency, compared with other benchmark schemes without such a joint design.

It is worth noting that there have been prior studies on communication cooperation (see [18]–[24]) or computation cooperation [25]–[29], respectively. On one hand, the cooperative communication via relaying has been extensively investigated in wireless communication systems to increase the communication rate and improve the communication reliability [20], [21], and applied in various other setups, such as the wireless powered communication [22] and the wireless powered MEC systems [24]. On the other hand, cooperative computation has emerged as a viable technique in MEC systems, which enables end users to exploit computation resources at nearby wireless devices (instead of APs or BSs). For example, in the so-called device-to-device (D2D) fogging [28] and peer-to-peer (P2P) cooperative computing [29], users with intensive computing tasks can offload all or part of the tasks to other idle users via D2D or P2P communications for execution. Similar computation task sharing among wireless devices has also been investigated in mobile social

[1] In general, the computation capability of the helper should be comparable or stronger than that of the user in order for the computation cooperation to be feasible.

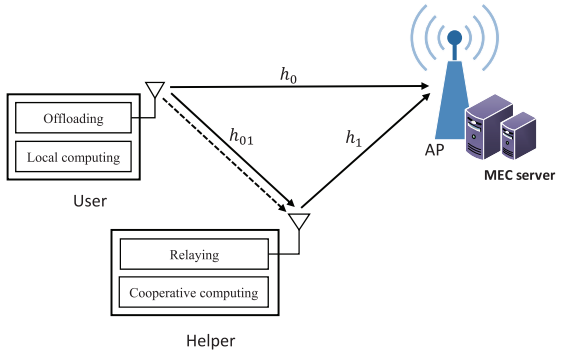

Fig. 1. Basic three-node MEC system with joint computation and communication cooperation. The dashed and solid lines indicate the tasks offloaded to the helper (for computation cooperation) and to the AP (via the helper's communication cooperation as a relay), respectively.

networks with crowdsensing [25] and in mobile wireless sensor networks [26], [27] for data fusion. However, different from these existing works with either communication or computation cooperation, this paper is the first to pursue a *joint computation and communication cooperation* approach, by unifying both of them for further improving the user's computation performance. Also note that this paper with user cooperation is different from the prior works on multiuser MEC systems [12]–[14], in which multiple users offload their own computation tasks to the AP/BS for execution, without user cooperation considered.

The remainder of this paper is organized as follows. Section II introduces the system model. Section III formulates the latency-constrained energy minimization problems under the partial and binary offloading models, respectively. Sections IV and V present the optimal solutions to the two problems of our interests, respectively. Section VI provides numerical results, followed by the conclusion in Section VII.

II. System Model

As shown in Fig. 1, we consider a basic three-node MEC system consisting of one user node, one helper node, and one AP node with an MEC server integrated, in which the three nodes are each equipped with one single antenna. We focus on a time block with duration $T > 0$, where the user needs to successfully execute computation tasks with $L > 0$ task input-bits within this block. By considering a latency-critical application, we assume that the block duration T is smaller than the channel coherence time, such that the channel power gain remains unchanged within the block of interest. Such an assumption has been commonly adopted in prior works [12]–[17]. It is further assumed that there is a central controller that is able to collect the global channel state information (CSI), and computation-related information for the three nodes; accordingly, the central controller can design and coordinate the computation and communication cooperation among the three nodes. This serves as a performance upper bound (or energy consumption lower bound) for practical cases when only partial CSI and computation-related information are known.

Specifically, without loss of generality, the L task input-bits can be divided into three parts intended for local computing, offloading to helper, and offloading to AP, respectively. Let $l_u \geq 0$, $l_h \geq 0$, and $l_a \geq 0$ denote the numbers of task input-bits for local computing at the user, offloading to the helper, and offloading to the AP, respectively. We then have

$$l_u + l_h + l_a = L. \tag{1}$$

Consider the two cases with partial offloading and binary offloading, respectively. In partial offloading, the computation task can be arbitrarily partitioned into subtasks. By assuming the number of subtasks are sufficiently large in this case, it is reasonable to approximate l_u, l_h, and l_a as real numbers between 0 and L subject to (1). In binary offloading, l_u, l_h, and l_a can only be set as 0 or L, and there is only one variable among them equal to L due to (1).

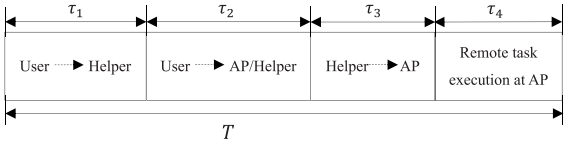

Fig. 2. MEC protocol with joint computation and communication cooperation.

A. MEC Protocol With Joint Computation and Communication Cooperation

As shown in Fig. 2, the duration-T block is generally divided into four slots for joint computation and communication cooperation. In the first slot with duration $\tau_1 \geq 0$, the user offloads the l_h task input-bits to the helper, and the helper can then execute them in the remaining time with duration $T - \tau_1$. In the second and third slots, the helper acts as a DF relay to help the user offload l_a task input-bits to the AP. In the second slot with duration $\tau_2 \geq 0$, the user transmits wireless signals containing the l_a task input-bits to both the AP and the helper simultaneously. After successfully decoding the received task input-bits, the helper forwards them to the AP in the third slot with duration $\tau_3 \geq 0$. After decoding the signals from the user and the helper, the MEC server can remotely execute the offloaded tasks in the fourth time slot with duration $\tau_4 \geq 0$.

As the computation results are normally of much smaller size than the input bits, the time for downloading the results to the user is negligible compared to the offloading time. Thus, we ignore the downloading time in this paper. In order to ensure the computation tasks to be successfully executed before the end of this block, we have the following time constraint:

$$\tau_1 + \tau_2 + \tau_3 + \tau_4 \leq T. \tag{2}$$

B. Computation Offloading

In this section, we discuss the computation offloading from the user to the helper and the AP, respectively.

1) Computation Offloading to Helper: In the first slot, the user offloads l_h task input-bits to the helper with transmit power $P_1 \geq 0$. Let $h_{01} > 0$ denote the channel power gain from the user to the helper and B the system bandwidth. Accordingly, the achievable data rate (in bits/s) for offloading

from the user to the helper is given by

$$r_{01}(P_1) = B\log_2\left(1 + \frac{P_1 h_{01}}{\Gamma\sigma_1^2}\right) \tag{3}$$

where σ_1^2 represents the power of additive white Gaussian noise at the helper, and $\Gamma \geq 1$ is a constant term accounting for the gap from the channel capacity due to a practical modulation and coding scheme. For simplicity, $\Gamma = 1$ is assumed throughout this paper. Consequently, we have the number l_h of task input-bits as

$$l_h = \tau_1 r_{01}(P_1). \tag{4}$$

Furthermore, let $P_{u,\max}$ denote the maximum transmit power at the user, and thus we have $0 \leq P_1 \leq P_{u,\max}$. For computation offloading, we consider the user's transmission energy as the dominant energy consumption and ignore the energy consumed by circuits in its radio-frequency chains, baseband signal processing, etc. Therefore, in the first slot, the energy consumption for the user to offload l_h task input-bits to the helper is given by

$$E_1^{\text{offl}} = \tau_1 P_1. \tag{5}$$

2) Computation Offloading to AP Assisted by Helper: In the second and third slots, the helper acts as a DF relay to help the user offload l_a task input-bits to the AP. Denote by $0 \leq P_2 \leq P_{u,\max}$ the user's transmit power in the second slot. In this case, the achievable data rate from the user to the helper is given by $r_{01}(P_2)$ with $r_{01}(\cdot)$ defined in (3). Denoting $h_0 > 0$ as the channel power gain from the user to the AP, the achievable data rate from the user to the AP is

$$r_0(P_2) = B\log_2\left(1 + \frac{P_2 h_0}{\sigma_0^2}\right) \tag{6}$$

where σ_0^2 is the noise power at the AP receiver.

After successfully decoding the received message, the helper forwards it to the AP in the third slot with the transmit power $0 \leq P_3 \leq P_{h,\max}$, where $P_{h,\max}$ denotes the maximum transmit power at the helper. Let $h_1 > 0$ denote the channel power gain from the helper to the AP. The achievable data rate from the helper to the AP is thus

$$r_1(P_3) = B\log_2\left(1 + \frac{P_3 h_1}{\sigma_0^2}\right). \tag{7}$$

By combining the second and third slots, the number of l_a task input-bits offloaded to the AP via a DF relay (the helper) should satisfy [19]–[21]

$$l_a = \min(\tau_2 r_0(P_2) + \tau_3 r_1(P_3),\ \tau_2 r_{01}(P_2)). \tag{8}$$

As in (5), we consider the user's and helper's transmission energy consumption for offloading as the dominant energy consumption in both the second and third slots. Therefore, we have

$$E_2^{\text{offl}} = \tau_2 P_2 \tag{9}$$

$$E_3^{\text{offl}} = \tau_3 P_3. \tag{10}$$

C. Computing at User, Helper, and AP

In the section, we explain the computing models at the user, the helper, and the AP, respectively.

1) Local Computing at User: The user executes the computation tasks with l_u task input-bits within the whole block. In practice, the number of CPU cycles for executing a computation task is highly dependent on various factors, such as the specific applications, the number of task input-bits, as well as the hardware (e.g., CPU and memory) architectures at the computing device [30]. To characterize the most essential computation and communication tradeoff and as commonly adopted in [10]–[16], we consider that the number of CPU cycles for this task is a linear function with respect to the number of task input-bits, where c_u denotes the number of CPU cycles for computing each one task input-bit at the user. Also, let $f_{u,n}$ denote the CPU frequency for the nth cycle, where $n \in \{1, \ldots, c_u l_u\}$. Note that the CPU frequency $f_{u,n}$ is constrained by a maximum value, denoted by $f_{u,\max}$, i.e.,

$$f_{u,n} \leq f_{u,\max} \qquad \forall n \in \{1, \ldots, c_u l_u\}. \tag{11}$$

As the local computing for the l_u task input-bits should be successfully accomplished before the end of the block, we have the following computation latency requirement:

$$\sum_{n=1}^{c_u l_u} \frac{1}{f_{u,n}} \leq T. \tag{12}$$

Accordingly, the user's energy consumption for local computing is [7], [15]

$$E_u^{\text{comp}} = \sum_{n=1}^{c_u l_u} \kappa_u f_{u,n}^2 \tag{13}$$

where κ_u denotes the effective capacitance coefficient that depends on the chip architecture at the user [30]. It has been shown in [15, Lemma 1] that to save the computation energy consumption with a computation latency requirement, it is optimal to set the CPU frequencies to be identical for different CPU cycles. By using this fact and letting the constraint in (12) be met with strict equality (for minimizing the computation energy consumption), we have

$$f_{u,1} = f_{u,2} = \ldots = f_{u,c_u l_u} = c_u l_u / T. \tag{14}$$

Substituting (14) into (13), the user's corresponding energy consumption for local computing is re-expressed as

$$E_u^{\text{comp}} = \frac{\kappa_u c_u^3 l_u^3}{T^2}. \tag{15}$$

Combining (14) with the maximum CPU frequency constraint in (11), it follows that:

$$c_u l_u \leq T f_{u,\max}. \tag{16}$$

2) Cooperative Computing at Helper: After receiving the offloaded l_h task input-bits in the first time slot, the helper executes the tasks during the remaining time with duration $(T - \tau_1)$. Let $f_{h,n}$ and $f_{h,\max}$ denote the CPU frequency for the nth CPU cycle and the maximum CPU frequency at the helper, respectively. Similarly as for the local computing at the user, it is optimal for helper to set the CPU frequency for the nth CPU cycle as $f_{h,n} = c_h l_h/(T - \tau_1)$, $n \in \{1, \ldots, c_h l_h\}$,

where c_h is the number of CPU cycles for computing one task-input bit at the helper. Accordingly, the energy consumption for cooperative computation at the helper is given by

$$E_h^{\text{comp}} = \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} \tag{17}$$

where κ_h is the effective capacitance coefficient of the helper.

Similarly as in (16), we have the constraint on the number of task input-bits as

$$c_h l_h \le (T-\tau_1) f_{h,\max} \tag{18}$$

where $f_{h,\max}$ denotes the maximum CPU frequency for the helper.

3) Remote Computing at AP (MEC Server): In the fourth slot, the MEC server at the AP executes the offloaded l_a task input-bits. In order to minimize the remote execution, the MEC server executes the offloaded tasks at its maximal CPU frequency, denoted by $f_{a,\max}$. Hence, the time duration τ_4 for the MEC server to execute the l_a offloaded bits is

$$\tau_4 = c_a l_a / f_{a,\max} \tag{19}$$

where c_a represents the number of CPU cycles required for computing one task-input bit at the AP. By substituting (19) into (2), the time allocation constraint is re-expressed as

$$\tau_1 + \tau_2 + \tau_3 + c_a l_a / f_{a,\max} \le T. \tag{20}$$

III. Problem Formulation

In this paper, we pursue an energy-efficient design for the three-node MEC system. As the AP normally has reliable power supply, we focus on the energy consumption at the wireless devices side (i.e., the user and helper) as the performance metric. In particular, we aim to minimize the total energy consumption at both the user and the helper (i.e., $\sum_{i=1}^{3} E_i^{\text{offl}} + E_u^{\text{comp}} + E_h^{\text{comp}}$), subject to the user's computation latency constraint T, by optimizing the task partition of the user, as well as the joint computation and communication resources allocation. The design variables include the time allocation vector of the slots $\boldsymbol{\tau} \triangleq [\tau_1, \tau_2, \tau_3]$, the user's task partition vector $\boldsymbol{l} \triangleq [l_u, l_h, l_a]$, and the transmit power allocation vector $\boldsymbol{P} \triangleq [P_1, P_2, P_3]$ for offloading of the user and helper.

In the case with partial offloading, the latency-constrained energy minimization problem is formulated as

$$\begin{aligned}
(\text{P1}): \min_{\boldsymbol{P},\boldsymbol{\tau},\boldsymbol{l}} \quad & \frac{\kappa_u c_u^3 l_u^3}{T^2} + \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} + \sum_{i=1}^{3} \tau_i P_i && \text{(21a)}\\
\text{s.t.} \quad & l_h \le \tau_1 r_{01}(P_1) && \text{(21b)}\\
& l_a \le \min(\tau_2 r_0(P_2) + \tau_3 r_1(P_3),\ \tau_2 r_{01}(P_2)) && \text{(21c)}\\
& 0 \le P_j \le P_{u,\max} \quad \forall j \in \{1,2\} && \text{(21d)}\\
& 0 \le P_3 \le P_{h,\max} && \text{(21e)}\\
& 0 \le \tau_i \le T \quad \forall i \in \{1,2,3\} && \text{(21f)}\\
& l_u \ge 0,\ l_h \ge 0,\ l_a \ge 0 && \text{(21g)}\\
& (1),\ (16),\ (18),\ \text{and}\ (20)
\end{aligned}$$

where (1) denotes the task partition constraint, (16) and (18) are the maximum CPU frequency constraints at the user and the helper, respectively, (20) denotes the time allocation constraint, (21b) and (21c) denote the constraints for the numbers of the offloaded bits from the user to the helper and to the AP, respectively. Note that in problem (P1), we replace the two equalities in (4) and (8) as two inequality constraints (21b) and (21c). It is immediate that constraints (21b) and (21c) should be met with strict equality at optimality of problem (P1). Also note that problem (P1) is nonconvex, due to the coupling of τ_i and P_i in the objective function (21a) and the constraints (21b) and (21c). Nonetheless, in Section IV we will transform (P1) into an equivalent convex problem and then present an efficiently algorithm to obtain the optimal solution of problem (P1) in a semi-closed form.

In the case with binary offloading, the latency-constrained energy minimization problem is formulated as

$$\begin{aligned}
(\text{P2}): \min_{\boldsymbol{P},\boldsymbol{\tau},\boldsymbol{l}} \quad & \frac{\kappa_u c_u^3 l_u^3}{T^2} + \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} + \sum_{i=1}^{3} \tau_i P_i && \text{(22a)}\\
\text{s.t.} \quad & l_u \in \{0, L\},\ l_h \in \{0, L\},\ l_a \in \{0, L\}\\
& (1),\ (16),\ (18),\ (20),\ \text{and}\ (21\text{b})\text{–}(21\text{f}). && \text{(22b)}
\end{aligned}$$

Note that problem (P2) is a mixed-integer nonlinear program [31] due to the involvement of integer variables l_u, l_h, and l_a. In Section V, we will develop an efficient algorithm to solve problem (P2) optimally by examining three computation modes, respectively.

A. Feasibility of (P1) and (P2)

Before solving problems (P1) and (P2), we first check their feasibility to determine whether the MEC system of interests can support the latency-constrained task execution or not. Let $L_{\max}^{(1)}$ and $L_{\max}^{(2)}$ denote the maximum numbers of task input-bits supported by the MEC system within the duration-T block under the partial and binary offloading cases, respectively. Evidently, if $L_{\max}^{(1)} \ge L$ (or $L_{\max}^{(2)} \ge L$), then problem (P1) [or (P2)] is feasible; otherwise, the corresponding problem is not feasible. Therefore, the feasibility checking procedures of problems (P1) and (P2) correspond to determining $L_{\max}^{(1)}$ and $L_{\max}^{(2)}$, respectively.

First, consider the partial offloading case. The maximum number $L_{\max}^{(1)}$ of task input-bits is attained when the three nodes fully use their available communication and computation resources. This corresponds to setting as $P_1 = P_2 = P_{u,\max}$, $P_3 = P_{h,\max}$, and letting the constraints (16), (18), (20), and (21c) be met with the strict equality in problem (P1). As a result, $L_{\max}^{(1)}$ is the optimal value of the following problem:

$$\begin{aligned}
L_{\max}^{(1)} \triangleq \max_{\boldsymbol{\tau},\boldsymbol{l}} \quad & l_u + l_h + l_a\\
\text{s.t.} \quad & \tau_1 + \tau_2 + \tau_3 + c_a l_a / f_{a,\max} = T\\
& l_h \le \tau_1 r_{01}(P_{u,\max}),\ l_a \le \tau_2 r_{01}(P_{u,\max})\\
& c_u l_u / T = f_{u,\max},\ c_h l_h / (T-\tau_1) = f_{h,\max}\\
& \tau_2 r_0(P_{u,\max}) + \tau_3 r_1(P_{h,\max}) = \tau_2 r_{01}(P_{u,\max})\\
& (21\text{f}) \text{ and } (21\text{g}). && \text{(23)}
\end{aligned}$$

Note that problem (23) is a linear program (LP) and can thus be efficiently solved via standard convex optimization techniques, such as the interior point method [33]. By comparing $L_{\max}^{(1)}$ versus L, the feasibility of problem (P1) is checked.

Next, consider the binary offloading case. The user's computation tasks can only be executed by one of the three computation modes, namely, the *local computing*, *computation cooperation* (offloading to helper), and *communication cooperation* (offloading to AP). For the three modes, the maximum numbers of supportable task input-bits can be readily obtained, as stated in the following.

1) For the local-computing mode, we have $l_h = l_a = 0$. With the maximum CPU frequency $f_{u,\max}$ and setting (16) to be tight, the maximum supportable number of task input-bits is given by
$$l_{u,\max}^{(2)} = \frac{Tf_{u,\max}}{c_u}. \tag{24}$$
2) For the computation-cooperation mode, we have $l_u = l_a = 0$. By setting the user's transmit power as $P_1 = P_{u,\max}$, and making the constraints (18) and (21b) be tight in problem (P2), the maximum number of task input-bits is thus,
$$l_{h,\max}^{(2)} = \tau_1^b r_{01}(P_{u,\max}) \tag{25}$$
where $\tau_1^b = Tf_{h,\max}/(c_h r_{01}(P_{u,\max}) + f_{h,\max})$ is the user's optimal time allocation for offloading.
3) For the communication-cooperation mode, we have $l_u = l_h = 0$, $P_2 = P_{u,\max}$, and $P_3 = P_{h,\max}$ in problem (P2). The maximum number of task input-bits $l_{a,\max}^{(2)}$ is obtained by solving the following LP:
$$\begin{aligned} l_{a,\max}^{(2)} = \max_{\tau_2,\tau_3,l_a} \quad & l_a \\ \text{s.t.} \quad & l_a \le \min\big(\tau_2 r_{01}(P_{u,\max})\tau_2 r_0(P_{u,\max}) \\ & \qquad + \tau_3 r_1(P_{h,\max})\big) \\ & \tau_2 + \tau_3 + c_a l_a / f_{a,\max} \le T \\ & 0 \le \tau_2 \le T,\ 0 \le \tau_3 \le T. \end{aligned} \tag{26}$$

Based on (24)–(26), the maximum number of supportable task input-bits for the binary offloading case is given by
$$L_{\max}^{(2)} = \max\Big(l_{u,\max}^{(2)}, l_{h,\max}^{(2)}, l_{a,\max}^{(2)}\Big). \tag{27}$$

By comparing $L_{\max}^{(2)}$ with L, the feasibility of problem (P2) is checked.

By comparing $L_{\max}^{(1)}$ and $L_{\max}^{(2)}$, we show that $L_{\max}^{(1)} \ge L_{\max}^{(2)}$. This is expected since that any feasible solution to problem (P2) is always feasible for problem (P1), but the reverse is generally not true. In other words, the partial offloading case can better utilize the distributed computation resources at different nodes, and thus achieves higher computation capacity than the binary offloading case.

IV. Optimal Solution to (P1)

In this section, we present an efficient algorithm for optimally solving problem (P1) in the partial offloading case.

Toward this end, we introduce an auxiliary variable vector $\boldsymbol{E} \triangleq [E_1, E_2, E_3]$ with $E_i = P_i\tau_i$ for all $i \in \{1,2,3\}$. Then it holds that $P_i = E_i/\tau_i$ if $\tau_i > 0$, and $P_i = 0$ if either $E_i = 0$ or $\tau_i = 0$ for any $i \in \{1,2,3\}$. By substituting $P_i = E_i/\tau_i$, $i \in \{1,2,3\}$, problem (P1) can be reformulated as

$$\begin{aligned} (\text{P1.1}): \min_{\boldsymbol{E},\boldsymbol{\tau},\boldsymbol{l}} \quad & \frac{\kappa_u c_u^3 l_u^3}{T^2} + \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} + \sum_{i=1}^{3} E_i && \text{(28a)} \\ \text{s.t.} \quad & l_h \le \tau_1 r_{01}\left(\frac{E_1}{\tau_1}\right) && \text{(28b)} \\ & l_a \le \tau_2 r_0\left(\frac{E_2}{\tau_2}\right) + \tau_3 r_1\left(\frac{E_3}{\tau_3}\right) && \text{(28c)} \\ & l_a \le \tau_2 r_{01}\left(\frac{E_2}{\tau_2}\right) && \text{(28d)} \\ & 0 \le E_j \le \tau_j P_{u,\max} \qquad \forall j \in \{1,2\} && \text{(28e)} \\ & 0 \le E_3 \le \tau_3 P_{h,\max} && \text{(28f)} \\ & (1),\ (16),\ (18),\ (20),\ (21\text{f}),\ \text{and}\ (21\text{g}) \end{aligned}$$

where both (28c) and (28d) follow from (21c).

Lemma 1: Problem (P1.1) is a convex problem.

Proof: The function $r_j(x)$ is a concave function with respect to $x \ge 0$ for any $j \in \{0, 1, 01\}$, and thus its perspective function $xr_j(y/x)$ is jointly concave with respect to $x > 0$ and $y \ge 0$ [33]. As a result, the set defined by constraints (28b)–(28d) becomes convex. The function l^3/τ^2 is a convex function with respect to $l \ge 0$ and $\tau > 0$, and hence the term $\kappa_h c_h^3 l_h^3/(T-\tau_1)^2$ in the objective function is jointly convex with respect to $l_h \ge 0$ and $0 \le \tau_1 < T$. Therefore, problem (P1.1) is convex. ■

As stated in Lemma 1, problem (P1.1) is convex and can thus be optimally solved by the standard interior point method [33]. Alternatively, to gain essential engineering insights, we next leverage the Lagrange duality method to obtain a well-structured optimal solution for problem (P1.1).

Let $\lambda_1 \ge 0$, $\lambda_2 \ge 0$, and $\lambda_3 \ge 0$ denote the dual variables associated with the constraints in (28b)–(28d), respectively, $\mu_1 \ge 0$ and $\mu_2 \in \mathbb{R}$ be the dual variables associated with the constraints in (20) and (1), respectively. Define $\boldsymbol{\lambda} \triangleq [\lambda_1, \lambda_2, \lambda_3]$ and $\boldsymbol{\mu} \triangleq [\mu_1, \mu_2]$. The partial Lagrangian of problem (P1.1) is given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{E},\boldsymbol{\tau},\boldsymbol{l},\boldsymbol{\lambda},\boldsymbol{\mu}) = & \sum_{i=1}^{3} E_i + \mu_1\tau_1 + \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} + (\lambda_1 - \mu_2) l_h \\ & - \lambda_1\tau_1 r_{01}\left(\frac{E_1}{\tau_1}\right) - \lambda_2\tau_2 r_0\left(\frac{E_2}{\tau_2}\right) \\ & - \lambda_3\tau_2 r_{01}\left(\frac{E_2}{\tau_2}\right) \\ & + \mu_1\tau_2 - \lambda_2\tau_3 r_1\left(\frac{E_3}{\tau_3}\right) + \mu_1\tau_3 \\ & + \frac{\kappa_u c_u^3 l_u^3}{T^2} - \mu_2 l_u \\ & + \big(\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2\big) l_a - \mu_1 T \\ & + \mu_2 L. \end{aligned}$$

Then the dual function of problem (P1.1) is given by

$$g(\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{\boldsymbol{E},\boldsymbol{\tau},\boldsymbol{l}} \quad \mathcal{L}(\boldsymbol{E},\boldsymbol{\tau},\boldsymbol{l},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{29}$$
$$\text{s.t.} \quad (16),\ (18),\ (21f),\ (21g),\ (28e),\ \text{and}\ (28f).$$

Lemma 2: In order for the dual function $g(\boldsymbol{\lambda},\boldsymbol{\mu})$ to be bounded from below, it must hold that $\lambda_1 - \mu_2 \geq 0$ and $\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2 \geq 0$.

Proof: See Appendix A. ∎

Based on Lemma 2, the dual problem of problem (P1.1) is given by

$$(\text{D1.1}): \max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \ g(\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{30a}$$
$$\text{s.t.} \ \ \mu_1 \geq 0,\ \lambda_i \geq 0 \quad \forall i \in \{1,2,3\} \tag{30b}$$
$$\lambda_1 - \mu_2 \geq 0 \tag{30c}$$
$$\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2 \geq 0. \tag{30d}$$

Denote $\mathcal{X}$ and $(\boldsymbol{\lambda}^{\text{opt1}}, \boldsymbol{\mu}^{\text{opt1}})$ as the feasible set and the optimal solution of $(\boldsymbol{\lambda},\boldsymbol{\mu})$ for problem (D1.1), respectively.

Since problem (P1.1) is convex and satisfies the Slater's condition, strong duality holds between problems (P1.1) and (D1.1) [33]. As a result, one can solve problem (P1.1) by equivalently solving its dual problem (D1.1). In the following, we first evaluate the dual function $g(\boldsymbol{\lambda},\boldsymbol{\mu})$ under any given $(\boldsymbol{\lambda},\boldsymbol{\mu}) \in \mathcal{X}$, and then obtain the optimal dual variables $(\boldsymbol{\lambda}^{\text{opt1}}, \boldsymbol{\mu}^{\text{opt1}})$ to maximize $g(\boldsymbol{\lambda},\boldsymbol{\mu})$. Denote $(\boldsymbol{E}^*, \boldsymbol{\tau}^*, \boldsymbol{l}^*)$ as the optimal solution to problem (29) under any given $(\boldsymbol{\lambda},\boldsymbol{\mu}) \in \mathcal{X}$, $(\boldsymbol{E}^{\text{opt1}}, \boldsymbol{\tau}^{\text{opt1}}, \boldsymbol{l}^{\text{opt1}})$ as the optimal primal solution to problem (P1.1).

A. Derivation of Dual Function $g(\boldsymbol{\lambda},\boldsymbol{\mu})$

First, we obtain $g(\boldsymbol{\lambda},\boldsymbol{\mu})$ by solving (29) under any given $(\boldsymbol{\lambda},\boldsymbol{\mu}) \in \mathcal{X}$. Equivalently, (29) can be decomposed into the following five subproblems:

$$\min_{E_1,\tau_1,l_h} \ E_1 + \mu_1\tau_1 - \lambda_1\tau_1 r_{01}\left(\frac{E_1}{\tau_1}\right) + \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} + (\lambda_1 - \mu_2) l_h$$
$$\text{s.t.} \ \ (18),\ 0 \leq E_1 \leq \tau_1 P_{u,\max},\ 0 \leq \tau_1 \leq T,\ l_h \geq 0 \tag{31}$$

$$\min_{E_2,\tau_2} \ E_2 + \mu_1\tau_2 - \lambda_2\tau_2 r_0\left(\frac{E_2}{\tau_2}\right) - \lambda_3\tau_2 r_{01}\left(\frac{E_2}{\tau_2}\right)$$
$$\text{s.t.} \ \ 0 \leq E_2 \leq \tau_2 P_{u,\max},\ 0 \leq \tau_2 \leq T \tag{32}$$

$$\min_{E_3,\tau_3} \ E_3 + \mu_1\tau_3 - \lambda_2\tau_3 r_1\left(\frac{E_3}{\tau_3}\right)$$
$$\text{s.t.} \ \ 0 \leq E_3 \leq \tau_3 P_{h,\max},\ 0 \leq \tau_3 \leq T \tag{33}$$

$$\min_{l_u \geq 0} \ \frac{\kappa_u c_u^3 l_u^3}{T^2} - \mu_2 l_u$$
$$\text{s.t.} \ \ c_u l_u \leq T f_{u,\max} \tag{34}$$

$$\min_{0 \leq l_a \leq L} \ \left(\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2\right) l_a. \tag{35}$$

The optimal solutions to problems (31)–(35) are presented in the following Lemmas 3–7, respectively. As these lemmas can be similarly proved via the Karush–Kuhn–Tucker (KKT) conditions [33], we only show the proof of Lemma 3 in Appendix B and omit the proofs of Lemmas 4–7 for brevity.

Lemma 3: Under given $(\boldsymbol{\lambda},\boldsymbol{\tau}) \in \mathcal{X}$, the optimal solution (E_1^*, τ_1^*, l_h^*) to problem (31) satisfies

$$E_1^* = P_1^* \tau_1^* \tag{36}$$
$$l_h^* = M_1^*\left(T - \tau_1^*\right) \tag{37}$$
$$\tau_1^* \begin{cases} = T, & \text{if } \rho_1 < 0 \\ \in [0,T], & \text{if } \rho_1 = 0 \\ = 0, & \text{if } \rho_1 > 0 \end{cases} \tag{38}$$

where $P_1^* = [(\lambda_1 B/\ln 2) - (\sigma_1^2/h_{01})]_0^{P_{u,\max}}$ with $[x]_b^a \triangleq \min\{a, \max\{x,b\}\}$, and

$$M_1^* \triangleq \begin{cases} \left[\sqrt{\frac{\mu_2-\lambda_1}{3\kappa_h c_h^3}}\right]_0^{\frac{f_{h,\max}}{c_h}}, & \text{if } \mu_2 - \lambda_1 \geq 0 \\ 0, & \text{if } \mu_2 - \lambda_1 < 0 \end{cases} \tag{39}$$

$$\rho_1 = \mu_1 - \lambda_1 r_{01}\left(P_1^*\right) + 2\kappa_h\left(c_h M_1^*\right)^3 + \frac{\lambda_1 B P_1^* h_{01}/\sigma_1^2}{\left(1 + P_1^* h_{01}/\sigma_1^2\right)\ln 2} - \alpha_1 P_{u,\max} + \frac{\beta_1 f_{h,\max}}{c_h} \tag{40}$$

$$\alpha_1 \triangleq \begin{cases} 0, & \text{if } P_1^* < P_{u,\max} \\ \frac{\lambda_1 B h_{01}/\sigma_1^2}{\ln 2\left(1+P_1^* h_{01}/\sigma_1^2\right)} - 1, & \text{if } P_1^* = P_{u,\max} \end{cases} \tag{41}$$

$$\beta_1 \triangleq \begin{cases} 0, & \text{if } M_1^* < \frac{f_{h,\max}}{c_h} \\ \mu_2 - \lambda_1 - 3\kappa_h c_h^3\left(M_1^*\right)^2, & \text{if } M_1^* = \frac{f_{h,\max}}{c_h}. \end{cases} \tag{42}$$

Proof: See Appendix B. ∎

Lemma 4: Under given $(\boldsymbol{\lambda},\boldsymbol{\tau}) \in \mathcal{X}$, the optimal solution (E_2^*, τ_2^*) to problem (32) satisfies

$$E_2^* = P_2^* \tau_2^* \tag{43}$$
$$\tau_2^* \begin{cases} = T, & \text{if } \rho_2 < 0 \\ \in [0,T], & \text{if } \rho_2 = 0 \\ = 0, & \text{if } \rho_2 > 0 \end{cases} \tag{44}$$

where $P_2^* = [(\sqrt{v^2 - 4uw} - v)/2u]_0^{P_{u,\max}}$ with $u = (\ln 2/B)(h_0/\sigma_0^2)(h_{01}/\sigma_1^2)$, $v = (\ln 2/B)(h_0/\sigma_0^2) + (h_{01}/\sigma_1^2) - (\lambda_2 + \lambda_3)(h_0/\sigma_0^2)(h_{01}/\sigma_1^2)$, $w = (\ln 2/B) - \lambda_2(h_0/\sigma_0^2) - \lambda_3(h_{01}/\sigma_1^2)$, $\rho_2 = \mu_1 - \lambda_2 r_0(P_2^*) + [(\lambda_2 B P_2^*(h_0/\sigma_0^2))/((1 + P_2^*(h_0/\sigma_0^2))\ln 2)] - \lambda_3 r_{01}(P_2^*) + [(\lambda_3 B P_2^*(h_{01}/\sigma_1^2))/((1 + P_2^*(h_{01}/\sigma_1^2))\ln 2)] - \alpha_2 P_{u,\max}$, and

$$\alpha_2 = \begin{cases} 0, & \text{if } P_2^* < P_{u,\max} \\ \frac{\lambda_3 B\frac{h_{01}}{\sigma_1^2}}{\left(1+P_2^*\frac{h_{01}}{\sigma_1^2}\right)\ln 2} + \frac{\lambda_2 B\frac{h_0}{\sigma_0^2}}{\left(1+P_2^*\frac{h_0}{\sigma_0^2}\right)\ln 2} - 1, & \text{if } P_2^* = P_{u,\max}. \end{cases}$$

Lemma 5: Under given $(\boldsymbol{\lambda},\boldsymbol{\tau}) \in \mathcal{X}$, the optimal solution (E_3^*, τ_3^*) to problem (33) satisfies

$$E_3^* = P_3^* \tau_3^* \tag{45}$$
$$\tau_3^* \begin{cases} = T, & \text{if } \rho_3 < 0 \\ \in [0,T], & \text{if } \rho_3 = 0 \\ = 0, & \text{if } \rho_3 > 0 \end{cases} \tag{46}$$

where $P_3^* = [(\lambda_2 B/\ln 2) - (\sigma_1^2/h_1)]_0^{P_{h,\max}}$ and $\rho_3 = \mu_1 + [(\lambda_2 B P_3^*(h_1/\sigma_1^2))/((1 + P_3^*(h_1/\sigma_1^2))\ln 2)] - \lambda_2 r_1(P_3^*) - \alpha_3 P_{h,\max}$ with

$$\alpha_3 = \begin{cases} 0, & \text{if } P_3^* < P_{h,\max} \\ \frac{\lambda_2 B h_1/\sigma_0^2}{(1+P_3^* h_1/\sigma_0^2)\ln 2} - 1, & \text{if } P_3^* = P_{h,\max}. \end{cases}$$

Lemma 6: For given $(\boldsymbol{\lambda}, \boldsymbol{\tau}) \in \mathcal{X}$, the optimal solution l_u^* to problem (34) is

$$l_u^* = \left[T\sqrt{\frac{\mu_2}{3\kappa_u c_u^3}} \right]_0^{\frac{Tf_{u,\max}}{cu}}. \tag{47}$$

Lemma 7: For given $(\boldsymbol{\lambda}, \boldsymbol{\tau}) \in \mathcal{X}$, the optimal solution l_a^* to problem (35) is

$$l_a^* \begin{cases} = 0, & \text{if } \lambda_2 + \lambda_3 + \mu_1 c_a / f_{a,\max} - \mu_2 > 0 \\ \in [0, L], & \text{if } \lambda_2 + \lambda_3 + \mu_1 c_a / f_{a,\max} - \mu_2 = 0 \\ = L, & \text{if } \lambda_2 + \lambda_3 + \mu_1 c_a / f_{a,\max} - \mu_2 < 0. \end{cases} \tag{48}$$

Remark 1: Note that in (38), (44), (46), and (48), if $\rho_i = 0$ (for any $i \in \{1, 2, 3\}$) or $\lambda_2 + \lambda_3 + \mu_1 c_a / f_{a,\max} - \mu_2 = 0$, then the optimal solution τ_i^* or l_a^* is nonunique in general. In this case, we choose $\tau_i^* = 0$ and $l_a^* = 0$ for the purpose of evaluating the dual function $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$. Such choices may not be feasible or optimal for problem (P1.1). To tackle this issue, we use an additional step in Section IV-C later to find the primal optimal τ_i^{opt1}'s and l_a^{opt1} for problem (P1.1).

By combining Lemmas 3–7, the dual function $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is evaluated for any given $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{X}$.

B. Obtaining $(\boldsymbol{\lambda}^{opt1}, \boldsymbol{\mu}^{opt1})$ to Maximize $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$

Next, we search over $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{X}$ to maximize $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ for solving problem (D1.1). Since the dual function $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is concave but nondifferentiable in general, one can use subgradient-based methods, such as the ellipsoid method [32], to obtain the optimal $\boldsymbol{\lambda}^{\text{opt1}}$ and $\boldsymbol{\mu}^{\text{opt1}}$ for (D1.1). For the objective function in (29), the subgradient with respect to $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is

$$\left[l_h^* - \tau_1^* r_{01}\left(\frac{E_1^*}{\tau_1^*}\right), l_a^* - \tau_2^* r_0\left(\frac{E_2^*}{\tau_2^*}\right) - \tau_3^* r_1\left(\frac{E_3^*}{\tau_3^*}\right) \right.$$
$$\left. l_a^* - \tau_2^* r_{01}\left(\frac{E_2^*}{\tau_2^*}\right), \sum_{i=1}^{3} \tau_i^* + \frac{l_a^* c_a}{f_{a,\max}} - T, L - l_u^* - l_h^* - l_a^* \right].$$

For the constraints $\mu_1 \geq 0$ and $\lambda_i \geq 0$, the subgradients are $\boldsymbol{e}_4$ and $\boldsymbol{e}_i$, $i \in \{1, 2, 3\}$, respectively, where $\boldsymbol{e}_i$ is the unit vector with one in the ith entry and zeros elsewhere in $\mathbb{R}^5$.

C. Optimal Primal Solution to (P1)

With $\boldsymbol{\lambda}^{\text{opt1}}$ and $\boldsymbol{\mu}^{\text{opt1}}$ obtained, it remains to determine the optimal solution to problem (P1.1) [and thus (P1)]. By replacing $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ in Lemmas 3–7 as $\boldsymbol{\lambda}^{\text{opt1}}$ and $\boldsymbol{\mu}^{\text{opt1}}$, we denote the corresponding P_i^*'s, l_u^*, and M_1^* as P_i^{opt1}'s, l_u^{opt1}, and M_1^{opt1}, respectively. Accordingly, $\boldsymbol{P}^{\text{opt1}} = [P_1^{\text{opt1}}, P_2^{\text{opt1}}, P_3^{\text{opt1}}]$ corresponds to the optimal solution of $\boldsymbol{P}$ to problem (P1), and l_u^{opt1} corresponds to the optimal solution of l_u to both problems (P1) and (P1.1). Nevertheless, due to the nonuniqueness of τ_i^*'s and l_a^*, we implement an additional step to construct the optimal solution of other variables to problem (P1). With $\boldsymbol{P}^{\text{opt1}}$, M_1^{opt1}, and l_u^{opt1}, the optimal solution must satisfy $l_h = M_1^{\text{opt1}}(T - \tau_1)$ and $E_i = P_i^{\text{opt1}} \tau_i$, $i \in \{1, 2, 3\}$. By substituting them in (P1) or (P1.1), we have the following LP to obtain $\boldsymbol{\tau}^{\text{opt1}}$ and l_a^{opt1}:

$$\min_{\boldsymbol{\tau}, l_a \geq 0} \ \kappa_h \left(c_h M_1^{\text{opt1}} \right)^3 (T - \tau_1) + \sum_{i=1}^{3} \tau_i P_i^{\text{opt1}}$$

TABLE I
ALGORITHM 1 FOR OPTIMALLY SOLVING PROBLEM (P1)

a) Initialization: Given an ellipsoid $\mathcal{E}((\boldsymbol{\lambda}, \boldsymbol{\mu}), \boldsymbol{A})$ containing $(\boldsymbol{\lambda}^{\text{opt1}}, \boldsymbol{\mu}^{\text{opt1}})$, where $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is the center point of $\mathcal{E}$ and the positive definite matrix $\boldsymbol{A}$ characterizes the size of $\mathcal{E}$.
b) **Repeat:**
   1) Obtain $\boldsymbol{P}^*, \boldsymbol{E}^*, \boldsymbol{\tau}^*$, and $\boldsymbol{l}^*$ with $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{X}$ according to Lemmas 4.3–4.7;
   2) Compute the subgradients of $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$, then update $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ using the ellipsoid method [32].
c) **Until** $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ converge with a prescribed accuracy.
d) **Set** $(\boldsymbol{\lambda}^{\text{opt1}}, \boldsymbol{\mu}^{\text{opt1}}) \leftarrow (\boldsymbol{\lambda}, \boldsymbol{\mu})$.
e) **Output**: Obtain $\boldsymbol{P}^{\text{opt1}}$ and l_u^{opt1} based on Lemmas 4.3–4.6 by replacing $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ as $\boldsymbol{\lambda}^{\text{opt1}}$ and $\boldsymbol{\mu}^{\text{opt1}}$, and then compute $\boldsymbol{\tau}^{\text{opt1}}$, l_h^{opt1}, and l_a^{opt1} by solving the LP in (49).

$$\begin{aligned} \text{s.t. } & M_1^{\text{opt1}}(T - \tau_1) \leq \tau_1 r_{01}\left(P_1^{\text{opt1}}\right) \\ & l_a \leq \tau_2 r_0\left(P_2^{\text{opt1}}\right) + \tau_3 r_1\left(P_3^{\text{opt1}}\right) \\ & l_a \leq \tau_2 r_{01}\left(P_2^{\text{opt1}}\right) \\ & M_1^{\text{opt1}}(T - \tau_1) + l_a + l_u^{\text{opt1}} = L \\ & 0 \leq \tau_i \leq T \quad \forall i \in \{1, 2, 3\}, \text{ and (20).} \end{aligned} \tag{49}$$

The LP in (49) can be efficiently solved by the interior-point method [33]. By combining $\boldsymbol{\tau}^{\text{opt1}}$, l_h^{opt1}, and l_a^{opt1}, together with $\boldsymbol{P}^{\text{opt1}}$ and l_u^{opt1}, the optimal solution to problem (P1) is finally found. In summary, we present Algorithm 1 for optimally solving problem (P1) under the partial offloading case in Table I.

Remark 2: Based on the optimal solution to (P1) in a semi-closed form, we have the following insights on the optimal joint computation and communication cooperation.

1) As for local computing, it follows from Lemma 6 that the number l_u^{opt1} of task input-bits for local computing generally increases as the block duration T becomes large. This shows that the user prefers locally computing more tasks when the computation latency becomes loose, as will be validated in numerical results later.
2) As for cooperative computation (i.e., offloading tasks to helper), it is evident that, based on Lemma 3, the offloading power P_1^{opt1} in the first slot increases as the corresponding channel power gain h_{01} becomes stronger. This is expected, since the marginal energy consumption for offloading from the user to the helper reduces in this case, and thus the user prefers offloading more tasks to the helper for cooperative computation.
3) As for cooperative communication (i.e., offloading to AP), it is observed from Lemmas 4 and 5 that the offloading power P_2^{opt1} in the second slot is dependent on both h_{01} and h_0, while P_3^{opt1} in the third slot increases as h_1 becomes large.

V. OPTIMAL SOLUTION TO (P2)

In this section, we develop an efficient algorithm to optimally solve problem (P2) in the binary offloading case.

Due to the constraints in (1) and (22b), there exist in total three computation modes for the user's task execution, i.e., the local computing mode (with $l_u = L$ and $l_h = l_a = 0$), the computation cooperation mode (with $l_h = L$ and $l_u = l_a = 0$), and the communication cooperation mode (with $l_a = L$ and $l_u = l_h = 0$). In the following, we first obtain the energy consumption under each of the three computation modes, and then choose the best mode with the minimum energy consumption as the optimal solution to problem (P2).

A. Computation Modes for Binary Offloading Case

1) Local Computing Mode: The local computing mode is feasible only when $l_{u,\max}^{(2)} \geq L$, with $l_{u,\max}^{(2)}$ given in (24). In this case, by substituting $l_u = L$ and $l_h = l_a = 0$ in (P2), we have the optimal transmit power and time allocation as $\boldsymbol{P} = \boldsymbol{0}$ and $\boldsymbol{\tau} = \boldsymbol{0}$, respectively. Therefore, the minimum energy consumption by the user in this mode is

$$E_u^{\text{opt2}} = \frac{\kappa_u c_u^3 L^3}{T^2}. \tag{50}$$

2) Computation Cooperation Mode: The computation cooperation mode is feasible only when $l_{h,\max}^{(2)} \geq L$, with $l_{h,\max}^{(2)}$ given in (25). Substituting $l_h = L$ and $l_u = l_a = 0$ into (P2), it then follows that $P_2 = P_3 = 0$ and $\tau_2 = \tau_3 = 0$. Consequently, problem (P2) is reduced as

$$\begin{aligned} \min_{P_1, \tau_1} \quad & \tau_1 P_1 + \frac{\kappa_h c_h^3 L^3}{(T-\tau_1)^2} && \text{(51a)} \\ \text{s.t.} \quad & L \leq \tau_1 r_{01}(P_1) && \text{(51b)} \\ & c_h L \leq (T-\tau_1) f_{h,\max} && \text{(51c)} \\ & 0 \leq \tau_1 \leq T, \ 0 \leq P_1 \leq P_{u,\max}. && \text{(51d)} \end{aligned}$$

Note that at the optimality of (51), the constraint (51b) must be tight. It thus follows that $P_1 = (2^{(L/B\tau_1)} - 1)(\sigma_1^2/h_{01})$. Accordingly, problem (51) is further reduced as the following univariable convex optimization problem:

$$\begin{aligned} \tau_1^{\text{opt2}} \triangleq \arg\min_{\tau_1} \quad & \left(2^{\frac{L}{B\tau_1}} - 1\right) \frac{\tau_1 \sigma_1^2}{h_{01}} + \frac{\kappa_h c_h^3 L^3}{(T-\tau_1)^2} \\ \text{s.t.} \quad & \left(2^{\frac{L}{B\tau_1}} - 1\right) \frac{\sigma_1^2}{h_{01}} \leq P_{u,\max} \\ & 0 \leq \tau_1 \leq T - \frac{c_h L}{f_{h,\max}} \end{aligned} \tag{52}$$

where the optimal solution τ_1^{opt2} to problem (52) can be efficiently found via a bisectional search procedure [33]. With τ_1^{opt2} obtained, the sum energy consumption at the user and the helper is given by

$$E_h^{\text{opt2}} = \left(2^{\frac{L}{B\tau_1^{\text{opt2}}}} - 1\right) \frac{\tau_1^{\text{opt2}} \sigma_1^2}{h_{01}} + \frac{\kappa_h c_h^3 L^3}{\left(T - \tau_1^{\text{opt2}}\right)^2}. \tag{53}$$

3) Communication Cooperation Mode: The communication cooperation mode is feasible only when $l_{a,\max}^{(2)} \geq L$ with $l_{a,\max}^{(2)}$ given in (26). With $l_a = L$ and $l_u = l_h = 0$, it follows that $P_1 = 0$ and $\tau_1 = 0$. Therefore, problem (P2) is re-expressed as

$$\begin{aligned} \min_{\tau_2, \tau_3, P_2, P_3} \quad & \tau_2 P_2 + \tau_3 P_3 \\ \text{s.t.} \quad & L \leq \min(\tau_2 r_0(P_2) + \tau_3 r_1(P_3), \ \tau_2 r_{01}(P_2)) \\ & \tau_2 + \tau_3 + L c_a / f_{a,\max} \leq T \\ & 0 \leq \tau_i \leq T \quad \forall i \in \{2, 3\} \\ & 0 \leq P_2 \leq P_{u,\max}, \ 0 \leq P_3 \leq P_{h,\max}. \end{aligned} \tag{54}$$

Similarly as for problem (P1), problem (54) can be optimally solved by Algorithm 1 by setting $l_u = 0$, $l_h = 0$, $l_a = L$, and $\tau_1 = 0$. We denote $(\tau_2^{\text{opt2}}, \tau_2^{\text{opt2}}, P_1^{\text{opt2}}, P_2^{\text{opt2}})$ as the optimal solution to problem (54). Therefore, we obtain the energy consumption for this mode as

$$E_a^{\text{opt2}} = \tau_2^{\text{opt2}} P_2^{\text{opt2}} + \tau_3^{\text{opt2}} P_3^{\text{opt2}}. \tag{55}$$

B. Computation Mode Selection

By comparing E_u^{opt2}, E_h^{opt2}, and E_a^{opt2}, we determine the optimal computation mode for problem (P2) as the one with the minimum energy consumption. Accordingly, the optimal l_a, l_u, and l_h are decided, and the corresponding computation and communication resources allocation for the selected computation mode becomes the optimal solution to problem (P2). As a result, problem (P2) in the binary offloading case is optimally solved.

VI. Numerical Results

In this section, numerical results are provided to evaluate the performance of the proposed joint computation and communication cooperation design for both partial and binary offloading cases, as compared to the following five benchmark schemes without such a joint design.

1) *Local Computing:* The user executes the computation tasks locally by itself. The energy consumption for local computing is obtained as E_u^{opt2} in (50).
2) *Computation Cooperation With Partial Offloading [28]:* The computation tasks are partitioned into two parts for the user's local computing and offloading to the helper, respectively. This corresponds to solving problem (P1) by setting $l_a = 0$ and $\tau_2 = \tau_3 = 0$.
3) *Communication Cooperation With Partial Offloading [12]:* The computation tasks are partitioned into two parts for the user's local computing and offloading to the AP, respectively. The offloading is assisted by the helper's communication cooperation as a DF relay. This corresponds to solving problem (P1) by setting $l_h = 0$ and $\tau_2 + \tau_3 = T$.
4) *Computation Cooperation With Binary Offloading:* The user offloads all the computation tasks to the helper for remote execution. The energy consumption corresponds to E_h^{opt2} in (53).
5) *Communication Cooperation With Binary Offloading:* The user can only offload its computation tasks to the AP assisted by the helper's communication cooperation. The energy consumption corresponds to E_a^{opt2} in (55) based on (54).

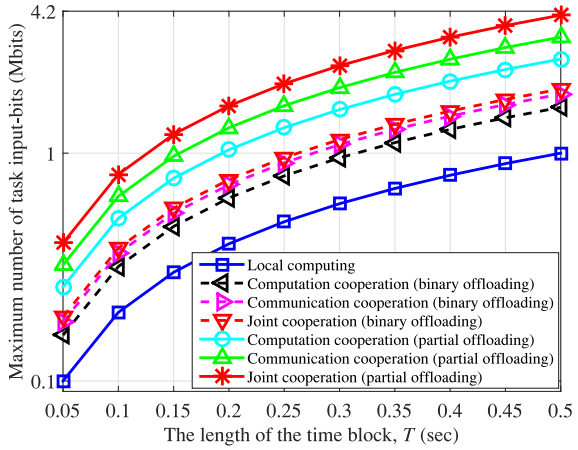

Fig. 3. Maximal number of task input-bits versus the block length.

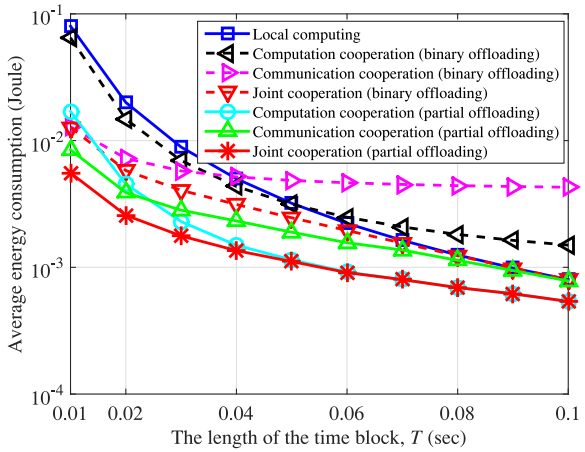

Fig. 4. Average energy consumption versus the block length.

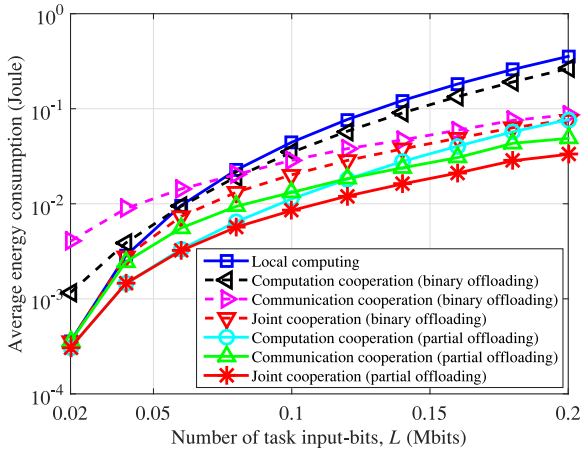

Fig. 5. Average energy consumption versus the number of task input-bits.

In the simulation, we consider that the user and the AP are located with a distance of 250 m and the helper is located on the line between them. Let D denote the distance between the user and the helper. The pathloss between any two nodes is denoted as $\beta_0(d/d_0)^{-\zeta}$, where $\beta_0 = -60$ dB corresponds to the path loss at the reference distance of $d_0 = 10$ m, d denotes the distance from the transmitter to the receiver, and $\zeta = 3$ is the pathloss exponent. Furthermore, we set $B = 1$ MHz, $\sigma_0^2 = \sigma_1^2 = -70$ dBm, $c_u = c_h = 10^3$ cycles/bit [15], $\kappa_u = 10^{-27}$ [12], $\kappa_h = 0.3 \times 10^{-27}$, $P_{u,\max} = P_{h,\max} = 40$ dBm, $f_{u,\max} = 2$ GHz, $f_{h,\max} = 3$ GHz, and $f_{a,\max} = 5$ GHz.

Fig. 3 shows the maximum number of task input-bits versus the block duration T, where $D = 20$ m. It is observed that for both partial and binary offloading cases in this setup, the computation cooperation and communication cooperation schemes achieve higher computation capacity than the local computing benchmark. Furthermore, for the binary or partial offloading, the communication cooperation scheme is observed to outperform the corresponding computation cooperation schemes, due to the higher computation capacities at the MEC server. In addition, our proposed joint cooperation design is observed to achieve the highest computation capacity by exploiting both benefits.

Fig. 4 shows the average energy consumption versus the block length T, where $L = 0.02$ Mb and $D = 120$ m. The proposed joint cooperation scheme is observed to achieve the minimum energy consumption for both the partial and binary offloading cases, respectively. In addition, we have the following observations.

1) The average energy consumption by all the schemes decreases as T increases. By comparing the partial and binary offloading, the computation and/or communication cooperation approach in the former case achieves more significant energy reduction than the corresponding one in the latter case. This indicates the benefit of task partitions in energy saving for MEC.
2) In the binary offloading case, when T is small (e.g., $T < 0.035$ s), the communication cooperation scheme achieved a lower energy consumption than the local computing and the computation cooperation schemes. When $0.035 \text{ s} < T \leq 0.05$ s, the computation cooperation scheme outperforms the other two schemes. As T becomes large (e.g., $T > 0.05$ s), the local computing scheme is beneficial.
3) In the partial offloading case, the communication cooperation scheme achieves a lower energy consumption than the computation cooperation scheme when T is small (e.g., $T < 0.025$ s), while the reverse is true when T becomes large. Also, the computation cooperation and communication cooperation schemes both outperform the local computing one. This is because that the two cooperation schemes additionally exploit computation resources at the helper and the AP, respectively.

Fig. 5 shows the average energy consumption versus the number of task input-bits L, where $T = 0.15$ s and $D = 120$ m. We have generally similar observations in Fig. 5 as in Fig. 4. Specifically, it is observed that at small L values (e.g., $L < 0.06$ Mb), the local computing is observed to achieve a similar performance as the proposed joint cooperation scheme with binary offloading, since in this case, the local computing is sufficient to execute the computation task input-bits. As L increases, the joint computation and communication cooperation is observed to achieve significant performance gain in terms of energy reduction.

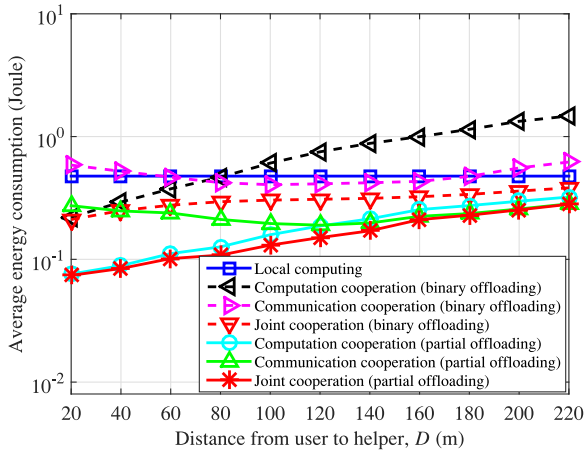

Fig. 6. Average energy consumption versus the distance between the user and the helper.

Fig. 6 shows the average energy consumption versus D in both the partial and binary offloading cases, where $T = 0.3$ s and $L = 0.5$ Mb. As expected, the average energy consumption by the local computing scheme remains unchanged for all D values. As D becomes larger, the average energy consumption by the communication cooperation scheme is observed to first decrease and then increase, while that by the computation cooperation scheme is observed to increase monotonically. This is because that as D increases, the channel gain between the user and the helper becomes smaller, while that between the helper and the AP becomes stronger; therefore, such a change benefits the offloading from the helper to the AP, but incurs increased offloading energy consumption from the user to the helper. Furthermore, the proposed joint cooperation scheme is observed to achieve significant gains over these benchmark schemes at all D values.

VII. Conclusion

In this paper, we proposed a novel joint computation and communication cooperation approach for improving the MEC performance, where nearby helper nodes share the computation and communication resources to help actively computing user nodes. By considering a basic three-node model under a four-slot cooperation protocol, we developed an energy-efficient design framework for both partial and binary offloading cases. We minimized the total energy consumption at both the user and the helper subject to the computation latency constraint, by jointly optimizing the task partition, as well as the computation and communication resources allocation. Based on convex optimization techniques, we presented an efficient algorithm to obtain the optimal solution in the partial offloading case. Computation mode selection was then applied for optimally solving the problem in the binary offloading case. Extensive numerical results demonstrated the merit of the proposed joint computation and communication cooperation scheme over alternative benchmarks. It is our hope that this paper sheds new light on how to optimally design multiresource user cooperation to improve the operation efficiency of MEC. Due to the space limitation, there have been various important issues that have not been addressed in this paper, which are discussed as follows to motivate future work.

1) Although this paper considered the basic setup with one user and one helper, our results are extendable to the more general case with multiple users and helpers. For example, in this case, we can employ a user-helper pairing procedure to pair each user with one helper, such that the helper can use the proposed joint communication and computation cooperation design to help the computation of the paired user. Furthermore, to fully utilize the computation and communication resources at multiple helpers, each user can offload the tasks to multiple helpers simultaneously for parallel execution, and multiple helpers can also cooperatively relay the user's tasks to the AP (e.g., via the collaborative beamforming). However, how to efficiently pair multiple users and helpers, and efficiently design the multiuser computation offloading and collaborative relaying are new and generally difficult problems worthy of further investigation.
2) This paper assumed that the user and helper have the common interest in improving the MEC system energy efficiency, such that the centralized resource allocation is employed for performance optimization. In practice, however, the helper can have self-interest. In this case, how to design incentive mechanisms (such as monetary and credit-based ones) and distributed algorithms to motivate the helper to participate in the joint cooperation is an interesting problem worth pursuing in the future.

Appendix A
Proof of Lemma 2

Note that for $\mathcal{L}(\boldsymbol{E}, \boldsymbol{\tau}, \boldsymbol{l}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ in (29), there exist two terms $(\lambda_1 - \mu_2)l_h$ and $(\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2)l_a$. Suppose $\lambda_1 - \mu_2 < 0$ (or $\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2 < 0$). Then $\mathcal{L}(\boldsymbol{E}, \boldsymbol{\tau}, \boldsymbol{l}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ becomes negative infinity when $l_h \to +\infty$ (or $l_a \to +\infty$). This implies that the dual function $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is unbounded from below in this case. Hence, it requires that $\lambda_1 - \mu_2 \geq 0$ and $\lambda_2 + \lambda_3 + \mu_1 c_a/f_{a,\max} - \mu_2 \geq 0$ to guarantee $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ to be bounded from below. Lemma 2 thus follows.

Appendix B
Proof of Lemma 3

As problem (31) is convex and satisfies the Slater's condition, strong duality holds between problem (31) and its dual problem. Therefore, one can solve this problem by applying the KKT conditions [33]. The Lagrangian of problem (31) is given by

$$\begin{aligned}\mathcal{L}_1 = {} & E_1 + \mu_1\tau_1 - \lambda_1\tau_1 r_{01}\left(\frac{E_1}{\tau_1}\right) - \mu_2 l_h + \lambda_1 l_h + \frac{\kappa_h c_h^3 l_h^3}{(T-\tau_1)^2} \\ & - a_1 E_1 + \alpha_1\left(E_1 - \tau_1 P_{u,\max}\right) - b_1\tau_1 + b_2(\tau_1 - T) \\ & - d_1 l_h + \beta_1\left(l_h - \frac{(T-\tau_1)f_{h,\max}}{c_h}\right)\end{aligned}$$

where a_1, α_1, b_1, b_2, d_1, and β_1 are the non-negative Lagrange multipliers associated with $E_1 \geq 0$, $E_1 \leq \tau_1 P_{u,\max}$, $\tau_1 \geq 0$, $\tau_1 \leq T$, $l_h \geq 0$, and $l_h \leq (T-\tau_1) f_{h,\max}/c_h$, respectively.

Based on the KKT conditions, it follows that:

$$\begin{aligned} a_1 E_1 &= 0 \\ \alpha_1 \left(E_1 - \tau_1 P_{u,\max}\right) &= 0 \\ b_2(\tau_1 - T) &= 0 \\ b_1 \tau_1 &= 0 \\ d_1 l_h &= 0 \end{aligned} \tag{56a}$$

$$\beta_1 \left(l_h - \frac{(T-\tau_1) f_{h,\max}}{c_h} \right) = 0 \tag{56b}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial E_1} &= 1 - \frac{\lambda_1 B \frac{h_{01}}{\sigma_1^2}}{\ln 2 \left(1 + \frac{E_1}{\tau_1} \frac{h_{01}}{\sigma_1^2}\right)} - a_1 + \alpha_1 \\ &= 0 \end{aligned} \tag{56c}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial \tau_1} &= \frac{2\kappa_h c_h^3 l_h^3}{(T-\tau_1)^3} - \lambda_1 B \log_2 \\ &\quad \times \left(1 + \frac{E_1}{\tau_1} \frac{h_{01}}{\sigma_1^2}\right) + \frac{\beta_1 f_{h,\max}}{c_h} \\ &\quad + \mu_1 + \frac{\lambda_1 B \frac{h_{01}}{\sigma_1^2} \frac{E_1}{\tau_1}}{\ln 2 \left(1 + \frac{E_1}{\tau_1} \frac{h_{01}}{\sigma_1^2}\right)} \\ &\quad - b_1 + b_2 + \alpha_1 P_{u,\max} \\ &= 0 \end{aligned} \tag{56d}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial l_h} &= \frac{3\kappa_h c_h^3 l_h^2}{(T-\tau_1)^2} - \mu_2 + \lambda_1 - d_1 + \beta_1 \\ &= 0 \end{aligned} \tag{56e}$$

where (56a) and (56b) denote the complementary slackness condition, (56c)–(56e) are the first-order derivative conditions of $\mathcal{L}_1$ with respect to E_1, τ_1, and l_h, respectively. Therefore, we have (36) based on (56c), and (37) holds due to (56e). Furthermore, based on (56c)–(56e) and with some manipulations, we have (41) and (42).

Furthermore, by substituting (36) and (37) into (56d) and assuming $\rho_1 = b_2 - b_1$, we have ρ_1 in (40). Hence, the optimal τ_1^* is given in (38). This lemma is thus proved.

REFERENCES

[1] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for mobile edge computing," in *Proc. WiOpt Workshops EFC-IoT*, Shanghai, China, May 2018, pp. 1–6.

[2] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Thing J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[3] W. Wang, Y. Chen, L. Wang, and Q. Zhang, "Sampleless Wi-Fi: Bringing low power to Wi-Fi communications," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1663–1672, Jun. 2017.

[4] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[5] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing: A key technology towards 5G," Sophia Antipolis, France, ETSI, White Paper, 2015. [Online]. Available: http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wP11_mec_a_key_technology_towards_5g.pdf

[6] 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.

[7] 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.

[8] V. Cozzolino, A. Y. Ding, J. Ott, and D. Kutscher, "Enabling fine-grained edge offloading for IoT," in *Proc. ACM SIGCOMM*, Los Angeles, CA, USA, Aug. 2017, pp. 124–126.

[9] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.

[10] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE ISIT*, Barcelona, Spain, Jul. 2016, pp. 1451–1455.

[11] Y. Zhang, D. Niyato, and P. Wang, "Offloading in mobile cloudlet systems with intermittent connectivity," *IEEE Trans. Mobile Comput.*, vol. 14, no. 12, pp. 2516–2529, Dec. 2015.

[12] 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.

[13] F. Wang, J. Xu, and Z. Ding, "Optimized multiuser computation offloading with multi-antenna NOMA," in *Proc. IEEE GLOBECOM Workshops NOMAT5G*, Singapore, Dec. 2017, pp. 1–7.

[14] M.-H. Chen, M. Dong, and B. Liang, "Joint offloading decision and resource allocation for mobile cloud with computing access point," in *Proc. IEEE ICASSP*, Shanghai, China, Mar. 2016, pp. 3516–3520.

[15] 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.

[16] 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. 1757–1770, May 2016.

[17] 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, Jun. 2018.

[18] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 74–80, Oct. 2004.

[19] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[20] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[21] Y. Liang and V. V. Veeravalli, "Gaussian orthogonal relay channels: Optimal resource allocation and capacity," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3284–3289, Sep. 2005.

[22] H. Ju and R. Zhang, "User cooperation in wireless powered communication networks," in *Proc. IEEE GLOBECOM*, Austin, TX, USA, Dec. 2014, pp. 1430–1435.

[23] S. Li *et al.*, "Location privacy preservation in collaborative spectrum sensing," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 729–737.

[24] 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.

[25] M. Xiao, J. Wu, L. Huang, Y. Wang, and C. Liu, "Multi-task assignment for crowdsensing in mobile social networks," in *Proc. IEEE INFOCOM*, Hong Kong, Apr./May 2015, pp. 2227–2235.

[26] Z. Sheng, C. Mahapatra, V. C. M. Leung, M. Chen, and P. K. Sahu, "Energy efficient cooperative computing in mobile wireless sensor networks," *IEEE Trans. Cloud Comput.*, vol. 6, no. 1, pp. 114–126, Jan./Mar. 2018.

[27] A. Mtibaa, A. Fahim, K. A. Harras, and M. H. Ammar, "Towards resource sharing in mobile device clouds: Power balancing across mobile devices," in *Proc. ACM SIGCOMM*, Hong Kong, Aug. 2013, pp. 51–56.

[28] L. Pu, X. Chen, J. Xu, and X. Fu, "D2D fogging: An energy-efficient and incentive-aware task offloading framework via network-assisted D2D collaboration," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3887–3901, Dec. 2016.

[29] C. You and K. Huang, "Exploiting non-causal CPU-state information for energy-efficient mobile cooperative computing," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4104–4117, Jun. 2018.

[30] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," *J. VLSI Signal Process. Syst.*, vol. 13, nos. 2–3, pp. 203–221, Aug./Sep. 1996.
[31] A. Ruszczyński, *Nonlinear Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2006.
[32] S. Boyd, *Ellipsoid Method*, Stanford Univ., Sacramento, CA, USA, 2008. [Online]. Available: https://web.stanford.edu/class/ee364b/lectures/ellipsoid_method_slides.pdf
[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, Mar. 2004.

Xiaowen Cao (S'18) received the B.Eng. degree from the Guangdong University of Technology, Guangzhou, China, in 2017, where she is currently pursuing the M.Sc. degree at the School of Information Engineering.

Her current research interests include mobile edge computing and unmanned aerial vehicle communications.

Feng Wang (M'16) received the B.Eng. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, and the M.Sc. and Ph.D. degrees from Fudan University, Shanghai, China, in 2012 and 2016, respectively.

He is currently an Assistant Professor with the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. From 2012 to 2013, he was a Research Fellow with the Department of Communication Technology, Sharp Laboratories of China, Shanghai, China. In 2017, he was a Post-Doctoral Research Fellow with the Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore, for eight months. His current research interests include signal processing for communications, energy harvesting wireless communications, and mobile edge computing.

Dr. Wang serves as a Reviewer for various IEEE journals, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE INTERNET OF THINGS JOURNAL.

Jie Xu (S'12–M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2007 and 2012, respectively.

From 2012 to 2014, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. From 2015 to 2016, he was a Post-Doctoral Research Fellow with the Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore. He is currently a Professor with the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His current research interests include energy efficiency and energy harvesting in wireless communications, wireless information and power transfer, wireless securities, UAV communications, and mobile edge computing.

Dr. Xu was a recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2017. He is currently an Editor of IEEE WIRELESS COMMUNICATIONS LETTERS, an Associate Editor of IEEE ACCESS, and a Guest Editor of *IEEE Wireless Communications*.

Rui Zhang (S'00–M'07–SM'15–F'17) received the B.Eng. (First Class Hons.) and M.Eng. degrees in electrical engineering from the National University of Singapore, Singapore, and the Ph.D. degree in electrical engineering from the Stanford University, Stanford, CA, USA.

From 2007 to 2010, he was a Research Scientist with the Institute for Infocomm Research, ASTAR, Singapore. In 2010, he has joined the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, where he is currently a Dean's Chair Associate Professor with the Faculty of Engineering. He has authored over 300 papers. His current research interests include wireless information and power transfer, drone communication, wireless eavesdropping and spoofing, energy-efficient and energy-harvesting-enabled wireless communication, multiuser MIMO, cognitive radio, and optimization methods.

Dr. Zhang was a recipient of the 6th IEEE Communications Society Asia–Pacific Region Best Young Researcher Award in 2011, the Young Researcher Award of National University of Singapore in 2015, and the IEEE Signal Processing Society Young Author Best Paper Award in 2017 for his co-authored paper. He was a co-recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2015, the IEEE Communications Society Asia–Pacific Region Best Paper Award in 2016, the IEEE Signal Processing Society Best Paper Award in 2016, the IEEE Communications Society Heinrich Hertz Prize Paper Award in 2017, the IEEE Signal Processing Society Donald G. Fink Overview Paper Award in 2017, and the IEEE Technical Committee on Green Communications and Computing Best Journal Paper Award in 2017. He has been listed as a Highly Cited Researcher (also known as the World's Most Influential Scientific Minds) by Thomson Reuters since 2015. He served for over 30 international conferences as the TPC Co-Chair or Organizing Committee member, and as the Guest Editor for three special issues of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He was an elected member of the IEEE Signal Processing Society SPCOM from 2012 to 2017 and SAM from 2013 to 2015 Technical Committees and served as the Vice Chair of the IEEE Communications Society Asia–Pacific Board Technical Affairs Committee from 2014 to 2015. He served as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2012 to 2016, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS: Green Communications and Networking Series from 2015 to 2016, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2013 to 2017. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He serves as a Steering Committee member of IEEE WIRELESS COMMUNICATIONS LETTERS.

Shuguang Cui (S'99–M'05–SM'12–F'14) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2005.

He has been an Assistant Professor, an Associate Professor, a Full Professor, and a Chair Professor in electrical and computer engineering with the University of Arizona, Tucson, AZ, USA, Texas A&M University, College Station, TX, USA, and University of California at Davis, Davis, CA, USA, respectively. He is currently a Professor with the Chinese University of Hong Kong Shenzhen, Shenzhen, China, and the Vice Director with the Shenzhen Research Institute of Big Data. His current research interests include data driven large-scale system control and resource management, large data set analysis, Internet of Things system design, energy harvesting-based communication system design, and cognitive network optimization.

Dr. Cui was a recipient of the IEEE Signal Processing Society 2012 Best Paper Award, the Highly Cited Researcher Award by Thomson Reuters, and the Worlds' Most Influential Scientific Minds by ScienceWatch in 2014. He has served as the General Co-Chair and the TPC Co-Chair for many IEEE conferences. He has also been serving as the Area Editor for *IEEE Signal Processing Magazine* and an Associate Editor for the IEEE TRANSACTIONS ON BIG DATA, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Series on Green Communications and Networking, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has been an elected member for IEEE Signal Processing Society SPCOM Technical Committee from 2009 to 2014 and the Elected Chair for IEEE ComSoc Wireless Technical Committee from 2017 to 2018. He is a member of the Steering Committee for the IEEE TRANSACTIONS ON BIG DATA and the Chair of the Steering Committee for the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He was a member of the IEEE ComSoc Emerging Technology Committee. He was elected as an IEEE ComSoc Distinguished Lecturer in 2014.