Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems

Feng Wang, *Member, IEEE*, Jie Xu, *Member, IEEE*, Xin Wang, *Senior Member, IEEE*, and Shuguang Cui, *Fellow, IEEE*

***Abstract*— Mobile-edge computing (MEC) and wireless power transfer (WPT) have been recognized as promising techniques in the Internet of Things era to provide massive low-power wireless devices with enhanced computation capability and sustainable energy supply. In this paper, we propose a unified MEC-WPT design by considering a wireless powered multiuser MEC system, where a multiantenna access point (AP) (integrated with an MEC server) broadcasts wireless power to charge multiple users and each user node relies on the harvested energy to execute computation tasks. With MEC, these users can execute their respective tasks locally by themselves or offload all or part of them to the AP based on a time-division multiple access protocol. Building on the proposed model, we develop an innovative framework to improve the MEC performance, by jointly optimizing the energy transmit beamforming at the AP, the central processing unit frequencies and the numbers of offloaded bits at the users, as well as the time allocation among users. Under this framework, we address a practical scenario where latency-limited computation is required. In this case, we develop an optimal resource allocation scheme that minimizes the AP's total energy consumption subject to the users' individual computation latency constraints. Leveraging the state-of-the-art optimization techniques, we derive the optimal solution in a semiclosed form. Numerical results demonstrate the merits of the proposed design over alternative benchmark schemes.**

***Index Terms*— Mobile-edge computing, wireless power transfer, computation offloading, energy beamforming, convex optimization.**

Manuscript received May 26, 2017; revised October 16, 2017; accepted December 12, 2017. Date of publication December 22, 2017; date of current version March 8, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0403402, in part by the National Natural Science Foundation of China under Grant 61671154, in part by DoD under Grant HDTRA1-13-1-0029, in part by NSFC under Grant 61328102/61629101, in part by the Shenzhen Fundamental Research Fund under Grant KQTD2015033114415450, and in part by NSF under Grant DMS-1622433, Grant AST-1547436, Grant ECCS-1508051/1659025, and Grant CNS-1343155. This paper was presented in part at the IEEE International Conference on Communications, Paris, France, May 21–25, 2017 [1]. The associate editor coordinating the review of this paper and approving it for publication was T. Melodia. *(Corresponding author: Jie Xu.)*

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

X. Wang is with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Department of Communication Science and Engineering, Shanghai Institute for Advanced Communication and Data Science, Fudan University, Shanghai 200433, China (e-mail: xwang11@fudan.edu.cn).

S. Cui is with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA, and also with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: sgcui@ucdavis.edu).

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/TWC.2017.2785305

I. Introduction

THE recent advancement of Internet of Things (IoT) has motivated various new applications (e.g., autonomous driving, virtual reality, augmented reality, and tele-surgery) to provide real-time machine-to-machine and machine-to-human interactions [2]. These emerging latency-sensitive applications critically rely on the real-time communication and computation of massive wireless devices. For example, smart wireless sensors in IoT networks may need to perceive the physical environment and then use the built-in computation resources to preprocess the sensed data in real time before sending it to the access point (AP) [2]. As extensive existing works focus on improving their communication performance, how to provide these devices with enhanced computation capability is a crucial yet challenging task to be tackled, especially when they are of small size and low power. To resolve this issue, mobile-edge computing (MEC) has emerged as a promising technique by providing cloud-like computing at the edge of mobile networks via integrated MEC servers at wireless APs and base stations (BSs) [3], [4]. Leveraging MEC, resource-limited wireless devices can offload their computation tasks to APs/BSs; then the integrated MEC servers can compute these tasks remotely. In general, the computation offloading can be implemented in two ways, namely *binary* and *partial* offloading [3]. In the binary offloading case, the computation task is not partitionable and should be offloaded as a whole. In the partial offloading case, the task can be partitioned into two parts, and only one of them is offloaded. The MEC technique facilitates the real-time implementation of computation-extensive tasks at massive low-power devices, and thus has attracted growing research interests in both academia and industry [3]–[7].

On the other hand, how to provide sustainable and cost-effective energy supply to massive computation-heavy devices is another challenge facing IoT. Radio-frequency (RF) signal based wireless power transfer (WPT) provides a viable solution by deploying dedicated energy transmitters to broadcast energy wirelessly [18]. Recently, emerging wireless powered communication networks (WPCNs) and simultaneous wireless information and power transfer (SWIPT) paradigms have been proposed to achieve ubiquitous wireless communications in a self-sustainable way [12]–[15], where WPT and wireless communications are combined into a joint design. In order to improve the WPT efficiency from the energy transmitter to one or more energy receivers, *transmit energy beamforming*

has been proposed as a promising solution by deploying multiple antennas at energy transmitters [9]. By properly adjusting the transmit beamforming vectors, energy transmitters can concentrate the radiative energy towards the intended receivers for efficient WPT. Motivated by these approaches, it is expected that the transmit energy beamforming-enabled WPT can also play an important role in facilitating self-sustainable computing for a large number of IoT devices.

To explore benefits of both MEC and WPT in ubiquitous computing, this paper develops a joint MEC-WPT design by considering a wireless powered *multiuser* MEC system that consists of a multi-antenna AP and multiple single-antenna users. The AP employs energy transmit beamforming to simultaneously charge the users, and each user relies on its harvested energy to execute the respective computation task. Suppose that partial offloading is allowed such that each user can arbitrarily partition the computation task into two *independent* parts for local computing and offloading, respectively. Furthermore, we assume that the downlink WPT and the uplink wireless communication (for computation offloading) are operated simultaneously over orthogonal frequency bands.[1] In addition, a time division multiple access (TDMA) protocol is employed to coordinate computation offloading, where different users offload their respective tasks to the AP over orthogonal time slots. The main results of this paper are summarized as follows.

- To improve the performance of such a wireless powered multiuser MEC system, we develop an innovative design framework by jointly optimizing the energy transmit beamforming at the AP, the central processing unit (CPU) frequencies[2] and the numbers of offloaded bits at the users, as well as the offloading time allocation among users. Note that the number of offloaded bits at each user corresponds to the multiplication of the offloading rate and allocated offloading time in this block.
- Targeting an energy-efficient wireless-powered MEC design, we minimize the AP's total energy consumption subject to the users' individual computation latency constraints. Leveraging the state-of-the-art optimization techniques, we obtain the optimal solution in a semi-closed form. It is revealed that at the optimal solution, the number of locally computed bits at each user should be strictly positive; i.e., it is always beneficial for each user to leave certain bits for local computing. It is also shown that the optimal offloading rate (and equivalent transmit power) at each user critically depends on the channel power gain and the circuit power.
- Extensive numerical results are provided to gauge the performance of the proposed designs with joint WPT, local computing, and offloading (i.e., task partition per user and offloading time allocation among the users) optimization, over benchmark schemes without such a joint consideration. It is shown that the proposed design can significantly reduce the energy consumption of the wireless powered MEC systems.

[1]The wireless energy harvesting in the downlink and the information transmission (or offloading) in the uplink can be performed simultaneously over orthogonal frequency bands in one single antenna with a duplexer, as commonly used in conventional frequency-division-duplexing (FDD) wireless communication transceivers.

[2]The term *CPU* generally refers to the processing unit and control unit at each user that takes charge of the local computing of computation tasks. The CPU frequency, i.e., the frequency of the CPU's clock pulses, determines the rate at which a CPU executes instructions.

A. Related Works

Transmit energy beamforming enabled WPT has been extensively studied in the literature (see, e.g., [9]–[24] and references therein). By considering a linear energy harvesting (EH) model, various prior works have investigated the optimal design of energy beamforming under different setups with SWIPT, e.g., in two-user multi-input multi-output (MIMO) systems [9], secrecy communications systems [10], multi-input single-output (MISO) interference channels [11], and multi-user MISO downlink channels [14]–[16]. Furthermore, some recent works investigated the transmit power allocation [23] and the transmit waveform optimization [24] for WPT by taking into account the nonlinear nature of the rectifier in EH [20], [21]. In addition, the benefit of energy beamforming crucially relies on the channel state information (CSI) known at the transmitter. The reverse-link channel training [17] and the energy measurement and feedback methods [18], [19] were proposed in WPT systems for the energy transmitter to practically learn the CSI to users. Furthermore, Lee and Zhang [22] developed a distributed energy beamforming system for multiple energy transmitters to charge multiple energy receivers simultaneously, with the help of the energy measurement and feedback.

On the other hand, several existing works [25]–[33] investigated the energy-efficient multiuser MEC design, where each user is powered by fixed energy sources such as battery, and the objective is to minimize the energy consumption at the users via joint computing and offloading optimization at the demand side. For example, Liu *et al.* [25] provided an overview on the applications and challenges of computation offloading. Liu *et al.* [26] and Huang *et al.* [27] investigated the dynamic offloading for MEC systems based on the techniques of Markov decision process and Lyapunov optimization, respectively. Munoz *et al.* [28] and Wang *et al.* [29] considered the joint computation and communication resource allocation in single-user MEC systems, and such designs were extended to multiuser MEC systems in [30]–[33]. Different from these prior works that studied WPT and MEC separately, this paper pursues a joint MEC-WPT design in a wireless powered multiuser MEC system, by jointly optimizing the WPT supply at the AP, as well as the local computing and offloading demands at the users.

It is worth noting that a prior work [34] considered the wireless powered *single-user* MEC system with binary offloading, where the user aims to maximize the probability of successful computation, by deciding whether a task should be fully offloaded or not, subject to the computation latency constraint. By contrast, this paper considers a more general case with more than one user, and allows for more flexible partial offloading to improve the system performance in terms of the energy efficiency (i.e., minimizing the total energy

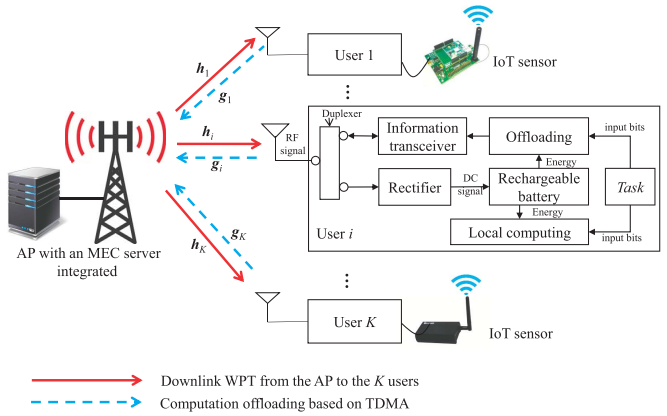

Fig. 1. A wireless powered multiuser MEC system with WPT in the downlink and computation offloading in the uplink.

consumption at the AP including the radiated energy for WPT and the energy for computing the offloaded tasks).

The remainder of the paper is organized as follows. Section II presents the system model. Section III formulates the computation latency constrained energy consumption minimization problem, and develops an efficient algorithm to obtain a well-structured optimal solution. Section IV provides numerical results to demonstrate the merits of the proposed design. Finally, Section V concludes this paper.

Notations: Boldface letters refer to vectors (lower case) or matrices (upper case). For a square matrix $\boldsymbol{S}$, $\mathrm{tr}(\boldsymbol{S})$ denotes its trace, while $\boldsymbol{S} \succeq \boldsymbol{0}$ means that $\boldsymbol{S}$ is positive semidefinite. For an arbitrary-size matrix $\boldsymbol{M}$, $\mathrm{rank}(\boldsymbol{M})$, $\boldsymbol{M}^{\dagger}$, and $\boldsymbol{M}^{H}$ denote its rank, transpose, and conjugate transpose, respectively. $\boldsymbol{I}$ and $\boldsymbol{0}$ denote an identity matrix and an all-zero vector/matrix, respectively, with appropriate dimensions. $\mathbb{C}^{x\times y}$ denotes the space of $x \times y$ complex matrices; $\mathbb{R}$ denotes the set of real numbers. $\mathbb{E}[\cdot]$ denotes the statistical expectation. $\|\boldsymbol{x}\|$ denotes the Euclidean norm of a vector $\boldsymbol{x}$, $|z|$ denotes the magnitude of a complex number z, and $[x]^{+} \triangleq \max(x, 0)$.

II. System Model

As shown in Fig. 1, we consider a wireless powered multiuser MEC system consisting of an N-antenna AP (integrated with an MEC server) and a set $\mathcal{K} \triangleq \{1, \ldots, K\}$ of single-antenna users. In this system, the AP employs RF signal based energy transmit beamforming to charge the K users. Each user $i \in \mathcal{K}$ utilizes the harvested energy to execute its computation task through local computing and offloading. Suppose that the downlink WPT from the AP to the users and the uplink computation offloading are operated simultaneously over orthogonal frequency bands, and the uplink for computation offloading and the downlink for computation result downloading are operated over the same frequency band. Assume a block-based model, and we focus on one particular block with length T. Here, T is chosen to be no larger than the latency of the MEC application and also no larger than the channel coherence time, such that the wireless channels remain unchanged during this block. For simplifying the analysis and better capturing the AP's transmission energy for computation offloading, we assume that the AP perfectly knows the CSI from/to the K users,[3] as well as their computation requirements. In accordance with such information, the AP coordinates the downlink WPT, the computation offloading, and the local computing for the K users.

A. Energy Transmit Beamforming From AP to Users

Let $\boldsymbol{s} \in \mathbb{C}^{N\times 1}$ denote the energy-bearing transmit signal by the AP, which is assumed to be a random signal with its power spectral density satisfying certain regulations on RF radiation [18]. Let $\boldsymbol{Q} \triangleq \mathbb{E}[\boldsymbol{s}\boldsymbol{s}^{H}] \succeq \boldsymbol{0}$ denote the energy transmit covariance matrix and $\mathbb{E}[\|\boldsymbol{s}\|^2] = \mathrm{tr}(\boldsymbol{Q})$ the transmit power at the AP. In general, the AP can use multiple energy beams to deliver the wireless energy, i.e., $\boldsymbol{Q}$ can be of any rank. Supposing $r = \mathrm{rank}(\boldsymbol{Q}) \leq N$, then a total of r energy beams can be obtained via the eigenvalue decomposition (EVD) of $\boldsymbol{Q}$ [18]. Let $\boldsymbol{h}_i \in \mathbb{C}^{N\times 1}$ denote the channel vector from the AP to user $i \in \mathcal{K}$, and define $\boldsymbol{H}_i \triangleq \boldsymbol{h}_i\boldsymbol{h}_i^{H}$, $\forall i \in \mathcal{K}$. Accordingly, the received RF power at each user $i \in \mathcal{K}$ is given by $|\boldsymbol{h}_i^{H}\boldsymbol{s}|^2$. In order to harvest such energy, each user i first converts the received RF signal into a direct current (DC) signal by a rectifier and then stores the energy of the DC signal in its chargeable battery (cf. Fig. 1). Note that the harvested DC power is generally nonlinear with respect to the received RF power [20], due to the nonlinear devices such as the diodes and diode-connected transistors. Moreover, the nonlinear RF-to-DC conversion greatly depends on the input power level and the transmit waveform. In the literature, there have been a handful of recent works on analytic nonlinear EH models, which characterize such nonlinear relations between the harvested DC power and the input RF power [23] or transmit waveform [24]. However, there still lacks a generic EH model that captures all practical issues [21]. Therefore, for simplicity, we assume that the input RF power is within the linear regime of the rectifier, and consider a linear EH model which has been commonly adopted in the WPT literature [9]–[12], [14]–[22]. Accordingly, the harvested energy amount by user i over this time block is

$$E_i = T\zeta\mathbb{E}\left[\left|\boldsymbol{h}_i^{H}\boldsymbol{s}\right|^2\right] = T\zeta\mathrm{tr}(\boldsymbol{Q}\boldsymbol{H}_i), \tag{1}$$

where $0 < \zeta \leq 1$ denotes the constant EH efficiency per user. The harvested energy E_i is used by user i for both computation offloading and local computing.

B. Energy Consumption at Users for Computation

For each user $i \in \mathcal{K}$, the computation task with $R_i > 0$ computation input bits is partitioned into two parts with $\ell_i \geq 0$ and $q_i \geq 0$ bits, which are offloaded to the MEC server at the AP or locally computed, respectively.[4] We assume that such a partition does not incur additional computation input bits, i.e., $R_i = \ell_i + q_i$, $\forall i \in \mathcal{K}$.

[3] When the CSI at the AP is not perfect (e.g., subject to some CSI estimation errors), the WPT and MEC performance may degrade. In this case, robust optimization techniques (see, e.g., [10], [15]) may be applied to obtain the energy beamforming vectors. However, the imperfect CSI scenario is out of scope of this paper.

[4] Each input bit can be treated as the smallest task unit, which includes the needed program codes and input parameters.

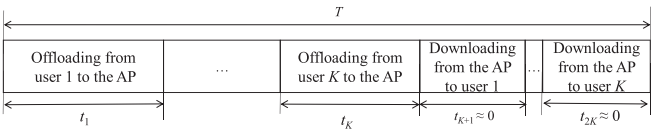

Fig. 2. The TDMA protocol for multiuser computation offloading.

1) Computation Offloading From Users to the AP: In order for the K users to offload their respective bits to the AP for computation, we adopt a TDMA protocol without interference as shown in Fig. 2, where the block is divided into $2K$ time slots each with a length of t_i, $\forall i \in \{1, \ldots, 2K\}$. In the first K time slots, the K users offload their computation bits to the AP one by one. After the MEC server executes the computation tasks on behalf of these users, the AP sends the computation results to the K users in the last K time slots. Due to the sufficient CPU capability at the MEC server, the computation time consumed at the MEC server are relatively small and negligible. Therefore, we assume that the users can download the computation results immediately after the first K offloading time slots. Furthermore, as the AP is usually with high transmit power and the computed results are usually of small size, we ignore the downloading time, i.e., $t_i \approx 0$, $\forall i \in \{K+1, \ldots, 2K\}$, and also ignore the energy consumption for transmitting and receiving the computation results in this paper.

For computation offloading in time slot i, let $\boldsymbol{g}_i \in \mathbb{C}^{N\times 1}$ denote the uplink channel vector from user i to the AP and p_i the transmit power for offloading. Assume further that the AP employs the maximum ratio combining (MRC) receiver to decode the information. The achievable offloading rate (in bits/sec) for user i is given by

$$r_i = B \log_2 \left(1 + \frac{p_i \tilde{g}_i}{\Gamma \sigma^2}\right), \quad \forall i \in \mathcal{K}, \tag{2}$$

where B denotes the spectrum bandwidth, $\tilde{g}_i \triangleq \|\boldsymbol{g}_i\|^2$ denotes the effective channel power gain from user i to the AP, σ^2 is the noise power at the receiver of the AP, and $\Gamma \geq 1$ is a constant accounting for the gap from the channel capacity due to a practical coding and modulation scheme. For simplicity, $\Gamma = 1$ is assumed throughout this paper. As a result, the number of offloaded bits ℓ_i by user i to the AP can be expressed as

$$\ell_i = r_i t_i, \quad \forall i \in \mathcal{K}. \tag{3}$$

Computation offloading incurs energy consumption at both the K users and the AP. Per user $i \in \mathcal{K}$, in addition to the transmit power p_i, a constant circuit power $p_{\mathrm{c},i}$ (by the digital-to-analog converter (DAC), filter, etc.) is consumed. The offloading energy consumption at user i is then $E_{\mathrm{offl},i} = (p_i + p_{\mathrm{c},i})t_i$. With simple manipulations based on (2) and (3), the transmit power p_i can be expressed as $p_i = \frac{1}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right)$, where $\beta(x) \triangleq \sigma^2(2^{\frac{x}{B}} - 1)$ is a monotonically increasing and convex function with respect to x.[5] Hence, the offloading energy consumption at user i is

$$E_{\mathrm{offl},i} = \frac{t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) + p_{c,i}t_i. \tag{4}$$

Remark 1: Note that in practice, in order for the AP to acquire the CSI (to the K users) for the energy beamforming design, each user needs to consume a certain amount of energy (e.g., for CSI feedback), and there generally exists a tradeoff between such energy consumption at the users versus the CSI accuracy at the AP. However, with the technical advancements, the user's feedback overhead for CSI acquisition could be made very small. Specifically, there are generally three types of CSI acquisition methods in the literature, namely the channel estimation and feedback [8], reverse-link training based on the channel reciprocity [17], and energy measurement and feedback [18], [19]. In the energy measurement and feedback method [18], each user only needs to measure its harvested energy level over each block and send one feedback bit to the AP per block; based on the feedback bits, the AP can sequentially improve the accuracy of CSI estimation; such a one-bit feedback is negligible when compared to the user reverse-link traffic for task offloading. Thus it is practically reasonable to ignore the feedback overhead and energy consumption at each user.

As for the AP, the energy is mainly consumed for executing the offloaded computation tasks and transmitting the computation results back to the users [4]. As the AP and its integrated MEC server generally have sufficient communication and computation capacities,[6] it can adopt a large transmit power (accordingly high communication rate) and a high constant CPU frequency to minimize the latency. In this case, the AP's energy consumption is generally proportional to the totally offloaded bits $\sum_{i=1}^{K} \ell_i$ from the K users. Therefore, we adopt a simplified linear energy consumption model for the computation at the AP as

$$E_{\mathrm{MEC}} = \alpha \sum_{i=1}^{K} \ell_i, \tag{5}$$

where α denotes the energy consumption per offloaded bit at the AP. In practice, α depends on the transceiver structure of the AP, the chip structure of the MEC server, and its operated CPU frequencies, etc. [4].

2) Local Computing at Users: Consider next the local computing for executing q_i input bits at each user $i \in \mathcal{K}$. Let C_i denote the number of CPU cycles required for computing one input bit at user i. Then the total number of CPU cycles required for the q_i bits is $C_i q_i$. By applying dynamic voltage and frequency scaling (DVFS) techniques [3], [4], user i can control the energy consumption for local task execution by adjusting the CPU frequency $f_{i,n}$ for each cycle n, where $f_{i,n} \in (0, f_i^{\max}]$, $n \in \{1, \ldots, C_i q_i\}$, and $f_i^{\max}$ denotes user

[5]Note that to avoid the issue of dividing by zero, we define $\beta\left(\frac{\ell_i}{t_i}\right) = 0$ when either $\ell_i = 0$ or $t_i = 0$ holds.

[6]In the case when the MEC server's computing capacity is limited, the computation offloading protocol needs to be redesigned, by taking into account the computation time at the MEC server as well as the computation resource sharing among these different users. Under such a scenario, how to jointly design the WPT and MEC optimally is out of the scope of this paper. It is an interesting direction to pursue in the future work.

i's maximum CPU frequency.[7] With $f_{i,n}$'s, the execution time for local computing at user i is $\sum_{n=1}^{C_i q_i} \frac{1}{f_{i,n}}$. As each user $i \in \mathcal{K}$ needs to accomplish the task execution within a block, the execution time cannot exceed the block length T, i.e.,

$$\sum_{n=1}^{C_i q_i} \frac{1}{f_{i,n}} \leq T, \quad \forall i \in \mathcal{K}. \tag{6}$$

Under the assumption of a low CPU voltage that normally holds for low-power devices, the consumed energy for local computing at user $i \in \mathcal{K}$ could be expressed as [35]

$$E_{\mathrm{loc},i} = \sum_{n=1}^{C_i q_i} \kappa_i f_{i,n}^2, \tag{7}$$

where κ_i is the effective capacitance coefficient that depends on the chip architecture at user i.

C. Energy Harvesting Constraints at Users

As each user $i \in \mathcal{K}$ is powered by the WPT from the AP to achieve self-sustainable operation, the so-called energy harvesting constraint needs to be imposed such that the totally consumed energy at the user cannot exceed the harvested energy E_i in (1) per block. By combining the computation offloading energy in (4) and the local computation energy in (7), the total energy consumed by user i within the block is $E_{\mathrm{offl},i} + E_{\mathrm{loc},i}$. Therefore, we must have per user $i \in \mathcal{K}$[8]:

$$E_{\mathrm{loc},i} + E_{\mathrm{offl},i} \leq E_i. \tag{8}$$

III. Computation Latency Constrained Energy Minimization

A. Problem Formulation

Under the above setup, we pursue an energy-efficient MEC-WPT design by considering a computation latency constrained energy minimization problem. Suppose that each user $i \in \mathcal{K}$ has a computation task with $R_i > 0$ input bits, which needs to be successfully executed before the end of the block. In this case, the sum of the number of offloaded bits ℓ_i and the number of locally computed bits q_i should be equal to R_i, i.e., we have $q_i = R_i - \ell_i$, $\forall i \in \mathcal{K}$.

We aim to minimize the energy consumption at the AP (including the energy consumption $\sum_{i=1}^K \alpha \ell_i$ in (5) for computation and $T\mathrm{tr}(\boldsymbol{Q})$ for WPT) while ensuring the successful execution of the K users' computation tasks per time block. To this end, we jointly optimize the energy transmit covariance matrix $\boldsymbol{Q}$ at the AP, the local CPU frequencies $\{f_{i,1}, \ldots, f_{i,C_i(R_i-\ell_i)}\}$, and the numbers of offloaded bits ℓ_i's at the users, as well as the time allocation t_i's among different users. Let $\boldsymbol{t} \triangleq [t_1, \ldots, t_K]^\dagger$, $\boldsymbol{\ell} \triangleq [\ell_1, \ldots, \ell_K]^\dagger$, and $\boldsymbol{f} \triangleq [f_{1,1}, \ldots, f_{K,C_K(R_K-\ell_K)}]^\dagger$. Mathematically, the latency-constrained energy minimization problem is formulated as

$$(\mathcal{P}1): \min_{\boldsymbol{Q}\succeq\boldsymbol{0},\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{f}} \quad T\mathrm{tr}(\boldsymbol{Q}) + \sum_{i=1}^K \alpha \ell_i \tag{9a}$$

$$\text{s.t.} \quad \sum_{n=1}^{C_i(R_i-\ell_i)} \frac{1}{f_{i,n}} \leq T, \quad \forall i \in \mathcal{K} \tag{9b}$$

$$\sum_{n=1}^{C_i(R_i-\ell_i)} \kappa_i f_{i,n}^2 + \frac{t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) + p_{c,i}t_i - T\zeta\mathrm{tr}(\boldsymbol{Q}\boldsymbol{H}_i) \leq 0, \quad \forall i \in \mathcal{K} \tag{9c}$$

$$\sum_{i=1}^K t_i \leq T, \quad t_i \geq 0, \ 0 \leq \ell_i \leq R_i, \ \forall i \in \mathcal{K} \tag{9d}$$

$$0 \leq f_{i,n} \leq f_i^{\max}, \ \forall n, \ \forall i \in \mathcal{K}. \tag{9e}$$

Here, the constraints in (9b) and (9c) represent the K users' individual local computing latency and energy harvesting constraints, respectively. Note that due to the non-convex nature of (9b) and (9c), problem $(\mathcal{P}1)$ is non-convex in the current form. However, we can transform it into a convex form and find the well-structured optimal solution, as will be shown in the next subsection.

B. Optimal Solution to Problem $(\mathcal{P}1)$

In this subsection, we provide the optimal solution to the computation latency constrained energy minimization problem $(\mathcal{P}1)$. To cope with the non-convex constraints in (9b) and (9c), we first establish the following lemma.

Lemma 1: Given the number of offloaded bits $\boldsymbol{\ell}$, the optimal solution of the local CPU frequencies $f_{i,n}$'s to problem $(\mathcal{P}1)$ should satisfy

$$f_{i,1} = \ldots = f_{i,C_i(R_i-\ell_i)} = C_i(R_i - \ell_i)/T, \quad \forall i \in \mathcal{K}. \tag{10}$$

Proof: See Appendix A. ∎

Lemma 1 indicates that at each user $i \in \mathcal{K}$, the local CPU frequencies for different CPU cycles are identical at the optimality. Hence, problem $(\mathcal{P}1)$ can be equivalently reformulated as

$$(\mathcal{P}1.1): \min_{\boldsymbol{Q}\succeq\boldsymbol{0},\boldsymbol{t},\boldsymbol{\ell}} \quad T\mathrm{tr}(\boldsymbol{Q}) + \sum_{i=1}^K \alpha \ell_i \tag{11a}$$

$$\text{s.t.} \quad \sum_{i=1}^K t_i \leq T \tag{11b}$$

$$\frac{\kappa_i C_i^3 (R_i - \ell_i)^3}{T^2} + \frac{t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) + p_{c,i}t_i - T\zeta\mathrm{tr}(\boldsymbol{Q}\boldsymbol{H}_i) \leq 0 \quad \forall i \in \mathcal{K} \tag{11c}$$

$$0 \leq \ell_i \leq R_i, \quad t_i \geq 0, \ \forall i \in \mathcal{K}. \tag{11d}$$

As $\beta(x)$ is convex as a function of $x \geq 0$, its perspective function $\frac{t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right)$ is jointly convex with respect to $t_i \geq 0$ and $\ell_i \geq 0$ [37]. As a result, the energy harvesting constraints

[7] Note that in practice, each CPU frequency $f_{i,n}$ can only be an integer chosen from a finite set. However, such an integer constraint may make the design problem a mixed-integer one that is NP-hard in general. To avoid this, we model $f_{i,n}$ as continuous variables to provide a performance upper-bound for the practical cases with discrete CPU frequencies.

[8] Note that in (8) we consider that the totally consumed energy should not exceed the totally harvested one, instead of the "energy causality" in conventional energy harvesting communications (see, e.g., [13]). This consideration implies that at the beginning of the block each user has sufficiently large energy storage, such that the stored energy will never be used up at any time within each block and the energy storage level will be refilled via energy harvesting by the end of each block.

in (11c) become convex. Furthermore, since the objective function in (11a) is affine and the other constraints are all convex, problem ($\mathcal{P}1.1$) is convex and can thus be optimally solved by standard convex optimization techniques. Nevertheless, to gain engineering insights, we derive its optimal solution in a semi-closed form by leveraging the Lagrange duality method [37].

Let $\mu \geq 0$ and $\lambda_i \geq 0$ denote the dual variables associated with the time-allocation constraint in (11b) and the i-th energy harvesting constraint in (11c), $\forall i \in \mathcal{K}$, respectively. Then the partial Lagrangian of ($\mathcal{P}1.1$) is expressed as

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{Q}, \boldsymbol{t}, \boldsymbol{\ell}, \boldsymbol{\lambda}, \mu) &= T\mathrm{tr}\left(\left(\boldsymbol{I} - \sum_{i=1}^{K} \zeta \lambda_i \boldsymbol{H}_i\right)\boldsymbol{Q}\right) - \mu T \\ &\quad + \sum_{i=1}^{K}\left(\alpha \ell_i + \frac{\lambda_i \kappa_i C_i^3 (R_i - \ell_i)^3}{T^2} \right. \\ &\quad \left. + \frac{\lambda_i t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) + \lambda_i p_{c,i} t_i + \mu t_i\right), \end{aligned} \tag{12}$$

where $\boldsymbol{\lambda} \triangleq [\lambda_1, \ldots, \lambda_K]^\dagger$. Accordingly, the dual function is given by

$$\begin{aligned}\Phi(\boldsymbol{\lambda}, \mu) = \min_{\boldsymbol{Q}\succeq \boldsymbol{0},\, \boldsymbol{t},\, \boldsymbol{\ell}} \quad & \mathcal{L}(\boldsymbol{Q}, \boldsymbol{t}, \boldsymbol{\ell}, \boldsymbol{\lambda}, \mu) \\ \text{s.t.} \quad & 0 \leq \ell_i \leq R_i, \quad t_i \geq 0, \ \forall i \in \mathcal{K}. \end{aligned} \tag{13}$$

Consequently, the dual problem of ($\mathcal{P}1.1$) is

$$(\mathcal{D}1.1): \max_{\boldsymbol{\lambda},\, \mu} \quad \Phi(\boldsymbol{\lambda}, \mu) \tag{14a}$$

$$\text{s.t.} \quad \boldsymbol{F}(\boldsymbol{\lambda}) \triangleq \boldsymbol{I} - \sum_{i=1}^{K} \zeta \lambda_i \boldsymbol{H}_i \succeq \boldsymbol{0} \tag{14b}$$

$$\mu \geq 0, \quad \lambda_i \geq 0, \ \forall i \in \mathcal{K}. \tag{14c}$$

Note that the constraint $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \boldsymbol{0}$ is necessary to ensure the dual function $\Phi(\boldsymbol{\lambda}, \mu)$ to be bounded from below (as proved in Appendix B). We denote the feasible set of $(\boldsymbol{\lambda}, \mu)$ characterized by (14b) and (14c) as $\mathcal{S}$.

Since problem ($\mathcal{P}1.1$) is convex and satisfies the Slater's condition, strong duality holds between ($\mathcal{P}1.1$) and its dual problem ($\mathcal{D}1.1$) [37]. As a result, we can solve ($\mathcal{P}1.1$) by equivalently solving ($\mathcal{D}1.1$). In the following, we first obtain the dual function $\Phi(\boldsymbol{\lambda}, \mu)$ for any given $(\boldsymbol{\lambda}, \mu) \in \mathcal{S}$, and then find the optimal dual variables $\boldsymbol{\lambda}$ and μ to maximize $\Phi(\boldsymbol{\lambda}, \mu)$ using the ellipsoid method [38]. For convenience of presentation, let $(\boldsymbol{Q}^*, \boldsymbol{t}^*, \boldsymbol{\ell}^*)$ denote the optimal solution to problem (13) for given $\boldsymbol{\lambda}$ and μ, $(\boldsymbol{Q}^{\mathrm{opt}}, \boldsymbol{t}^{\mathrm{opt}}, \boldsymbol{\ell}^{\mathrm{opt}})$ denote the optimal primary solution to ($\mathcal{P}1.1$), and $(\boldsymbol{\lambda}^{\mathrm{opt}}, \mu^{\mathrm{opt}})$ denote the optimal dual solution to ($\mathcal{D}1.1$).

1) Evaluating the Dual Function $\Phi(\boldsymbol{\lambda}, \mu)$: First, we obtain the dual function $\Phi(\boldsymbol{\lambda}, \mu)$ in (13) for any given $(\boldsymbol{\lambda}, \mu) \in \mathcal{S}$. To this end, problem (13) can be decomposed into $(K+1)$ subproblems as follows, one for optimizing $\boldsymbol{Q}$ and the other K for jointly optimizing t_i's and ℓ_i's.

$$\min_{\boldsymbol{Q}} \quad \mathrm{tr}(\boldsymbol{Q}\boldsymbol{F}(\boldsymbol{\lambda})) \qquad \text{s.t.} \quad \boldsymbol{Q} \succeq \boldsymbol{0}. \tag{15}$$

$$\begin{aligned}\min_{t_i, \ell_i} \quad & \alpha \ell_i + \frac{\lambda_i \kappa_i C_i^3 (R_i - \ell_i)^3}{T^2} + \frac{\lambda_i t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) \\ & + \lambda_i p_{c,i} t_i + \mu t_i \end{aligned} \tag{16a}$$

$$\text{s.t.} \quad 0 \leq \ell_i \leq R_i, \quad t_i \geq 0, \tag{16b}$$

where each subproblem i in (16a) is for the user $i \in \mathcal{K}$. Under the condition of $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \boldsymbol{0}$, it is evident that the optimal value of problem (16a) is zero and its optimal solution $\boldsymbol{Q}^*$ can be any positive semidefinite matrix in the null space of $\boldsymbol{F}(\boldsymbol{\lambda})$. Without loss of optimality, we simply set $\boldsymbol{Q}^* = \boldsymbol{0}$ for the purpose of obtaining the dual function $\Phi(\boldsymbol{\lambda}, \mu)$ and computing the optimal dual solution.[9] Note that $\boldsymbol{Q}^* = \boldsymbol{0}$ is generally not the optimal primary solution to ($\mathcal{P}1.1$). As a result, after finding the optimal dual solution $(\boldsymbol{\lambda}^{\mathrm{opt}}, \mu^{\mathrm{opt}})$, we need to use an additional step to retrieve the optimal primary solution of $\boldsymbol{Q}^{\mathrm{opt}}$ to ($\mathcal{P}1.1$), as will be shown in Section IV-C.

For the i-th subproblem in (16a), it is convex and satisfies the Slater's condition. Based on the Karush-Kuhn-Tucker (KKT) conditions [37], one can obtain the optimal solution (t_i^*, ℓ_i^*) to (16a) in a semi-closed form, as stated in the following lemma.

Lemma 2: For any given $(\boldsymbol{\lambda}, \mu) \in \mathcal{S}$, the optimal solution (t^*, ℓ_i^*) to problem (16a) can be obtained as follows.

- If $\lambda_i = 0$, we have $\ell_i^* = 0$ and $t_i^* = 0$;
- If $\lambda_i > 0$, we have

$$\ell_i^* = \left[R_i - \sqrt{\frac{T^2}{3\kappa_i C_i^3}\left(\frac{\alpha}{\lambda_i} + \frac{\sigma^2 \ln 2}{B\tilde{g}_i} 2^{\frac{r_i^*}{B}}\right)}\right]^+ \tag{17}$$

$$t_i^* = \ell_i^* / r_i^*, \tag{18}$$

where $r_i^* \triangleq \frac{B}{\ln 2}\left(W_0\left(\frac{\tilde{g}_i}{\sigma^2 e}\left(\frac{\mu}{\lambda_i} + p_{c,i}\right) - \frac{1}{e}\right) + 1\right)$ denotes the offloading rate of user i, $W_0(x)$ is the principal branch of the Lambert W function defined as the solution for $W_0(x)e^{W_0(x)} = x$ [36], and e is the base of the natural logarithm.

Proof: See Appendix C. ∎

By combining Lemma 2 and $\boldsymbol{Q}^* = \boldsymbol{0}$, the dual function $\Phi(\boldsymbol{\lambda}, \mu)$ can be evaluated for any given $(\boldsymbol{\lambda}, \mu) \in \mathcal{S}$.

2) Obtaining the Optimal $\boldsymbol{\lambda}^{\mathrm{opt}}$ and μ^{opt} to Maximize $\Phi(\boldsymbol{\lambda}, \mu)$: Having obtained $(\boldsymbol{Q}^*, \boldsymbol{\ell}^*, \boldsymbol{t}^*)$ for given $\boldsymbol{\lambda}$ and μ, we can next solve the dual problem ($\mathcal{D}1.1$) to maximize $\Phi(\boldsymbol{\lambda}, \mu)$. Note that the dual function $\Phi(\boldsymbol{\lambda}, \mu)$ is concave but non-differentiable in general [37]. Hence, we use subgradient based methods, e.g., the ellipsoid method [38], to obtain the optimal $\boldsymbol{\lambda}^{\mathrm{opt}}$ and μ^{opt} for problem ($\mathcal{D}1.1$). The basic idea of the ellipsoid method is to find a series of ellipsoids to localize the optimal dual solution $\boldsymbol{\lambda}^{\mathrm{opt}}$ and μ^{opt} [38]. To start with, we choose a given $(\boldsymbol{\lambda}, \mu) \in \mathcal{S}$ as the center of the initial ellipsoid and set its volume to be sufficiently large to contain $(\boldsymbol{\lambda}^{\mathrm{opt}}, \mu^{\mathrm{opt}})$. Then, at each iteration, we update the dual variables $(\boldsymbol{\lambda}, \mu)$ based on the subgradients of both the

[9] Note that $\boldsymbol{Q}^* = \boldsymbol{0}$ is not a unique optimal solution to problem (16a) when $\boldsymbol{F}(\boldsymbol{\lambda})$ is rank-deficient, i.e., $\mathrm{rank}(\boldsymbol{F}(\boldsymbol{\lambda})) < N$.

objective function and the constraint functions, and accordingly construct a new ellipsoid with reduced volume. When the volume of the ellipsoid is reduced below a certain threshold, the iteration will terminate and the center of the ellipsoid is chosen to be the optimal dual solution $(\boldsymbol{\lambda}^{\rm opt}, \mu^{\rm opt})$. More details can be referred to in [38].

To implement the ellipsoid method, it remains to determine the subgradients of both the objective function in (14a) and the constraint functions in (14b) and (14c). For the objective function $\Phi(\boldsymbol{\lambda}, \mu)$ in (14a), one subgradient is given by [38]

$$\left[\frac{\kappa_1 C_1^3 (R_1-\ell_1^*)^3}{T^2}+\frac{t_1^*}{\tilde{g}_1}\beta\left(\frac{\ell_1^*}{t_1^*}\right)+p_{{\rm c},1}t_1^*,\ldots,\right.$$
$$\left.\frac{\kappa_K C_K^3 (R_K-\ell_K^*)^3}{T^2}+\frac{t_K^*}{\tilde{g}_K}\beta\left(\frac{\ell_K^*}{t_K^*}\right)+p_{{\rm c},K}t_K^*,\sum_{i=1}^K t_i^*-T\right]^\dagger. \tag{19a}$$

As for the constraint $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \mathbf{0}$ in (14b), we have the following lemma.

Lemma 3: Let $\boldsymbol{v} \in \mathbb{C}^{N\times 1}$ be the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{F}(\boldsymbol{\lambda})$, i.e., $\boldsymbol{v} = \arg\min_{\|\boldsymbol{\xi}\|=1} \boldsymbol{\xi}^H \boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{\xi}$. Then the constraint $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \mathbf{0}$ is equivalent to the constraint of $\boldsymbol{v}^H \boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{v} \geq 0$, and the subgradient of $\boldsymbol{v}^H \boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{v}$ at the given $\boldsymbol{\lambda}$ and μ is

$$\left[\zeta \boldsymbol{v}^H \boldsymbol{H}_1 \boldsymbol{v},\ldots,\zeta \boldsymbol{v}^H \boldsymbol{H}_K \boldsymbol{v}, 0\right]^\dagger. \tag{20}$$

Proof: See Appendix D. ∎

Furthermore, the subgradient of $\lambda_i \geq 0$ in (14c) is given by the elementary vector $\boldsymbol{e}_i \in \mathbb{R}^{(K+1)\times 1}$ (i.e., $\boldsymbol{e}_i$ is of all zero entries except for the i-th entry being one), $\forall i \in \mathcal{K}$, while that of $\mu \geq 0$ is $\boldsymbol{e}_{K+1}$. By using this together with (19) and (20), the ellipsoid method can be applied to efficiently update $\boldsymbol{\lambda}$ and μ towards the optimal $\boldsymbol{\lambda}^{\rm opt}$ and $\mu^{\rm opt}$ for $(\mathcal{D}1.1)$.

3) Finding the Optimal Primary Solution to $(\mathcal{P}1)$*:* With $\boldsymbol{\lambda}^{\rm opt}$ and $\mu^{\rm opt}$ obtained, it remains to determine the optimal primary solution to $(\mathcal{P}1.1)$ (or equivalently $(\mathcal{P}1)$). Specifically, by replacing $\boldsymbol{\lambda}$ and μ with $\boldsymbol{\lambda}^{\rm opt}$ and $\mu^{\rm opt}$ in Lemma 2, one can obtain the optimal $(\boldsymbol{t}^{\rm opt}, \boldsymbol{\ell}^{\rm opt})$ for $(\mathcal{P}1)$ in a semi-closed form. Furthermore, by substituting $\boldsymbol{\ell}^{\rm opt}$ in Lemma 1, one can then obtain the optimal local CPU frequencies $\{f_{i,n}^{\rm opt}\}$ for the K users. However, we cannot directly obtain the optimal energy transmit covariance matrix $\boldsymbol{Q}^{\rm opt}$ for $(\mathcal{P}1)$ from the solution to problem (16a), since its solution is non-unique in general. Therefore, we adopt an additional step to obtain $\boldsymbol{Q}^{\rm opt}$ by solving a semidefinite program (SDP), which corresponds to solving problem $(\mathcal{P}1.1)$ for $\boldsymbol{Q}$ under the given $(\boldsymbol{t}^{\rm opt}, \boldsymbol{\ell}^{\rm opt})$.

We can then readily establish the following proposition.

Proposition 1: The optimal solution $(\{f_{i,n}^{\rm opt}\}, \boldsymbol{Q}^{\rm opt}, \boldsymbol{t}^{\rm opt}, \boldsymbol{\ell}^{\rm opt})$ for problem $(\mathcal{P}1)$ is given by

$$\ell_i^{\rm opt} = \begin{cases} \left[R_i - \sqrt{\frac{T^2}{3\kappa_i C_i^3}\left(\frac{\alpha}{\lambda_i^{\rm opt}}+\frac{\sigma^2 \ln 2}{B\tilde{g}_i}2^{\frac{r_i^{\rm opt}}{B}}\right)}\right]^+, \\ \qquad \text{if } \lambda_i^{\rm opt} > 0, \\ 0, \\ \qquad \text{if } \lambda_i^{\rm opt} = 0, \end{cases} \tag{21}$$

$$t_i^{\rm opt} = \begin{cases} \ell_i^{\rm opt}/r_i^{\rm opt}, & \text{if } \lambda_i^{\rm opt} > 0, \\ 0, & \text{if } \lambda_i^{\rm opt} = 0, \end{cases} \tag{22}$$

$$f_{i,1}^{\rm opt} = \ldots = f_{i,C_i(R_i-\ell_i^{\rm opt})}^{\rm opt} = C_i(R_i-\ell_i^{\rm opt})/T, \quad \forall i \in \mathcal{K}, \tag{23}$$

and

$$\boldsymbol{Q}^{\rm opt} = \arg\min_{\boldsymbol{Q}\succeq\mathbf{0}} \quad T\,{\rm tr}(\boldsymbol{Q})$$
$$\text{s.t.} \quad \frac{\kappa_i C_i^3 (R_i-\ell_i^{\rm opt})^3}{T^2}+\frac{t_i^{\rm opt}}{\tilde{g}_i}\beta\left(r_i^{\rm opt}\right)+p_{c,i}t_i^{\rm opt}$$
$$-T\zeta\,{\rm tr}(\boldsymbol{Q}\boldsymbol{H}_i) \leq 0, \quad \forall i \in \mathcal{K}, \tag{24}$$

where

$$r_i^{\rm opt} \triangleq \frac{B}{\ln 2}\left(W_0\left(\frac{\tilde{g}_i}{\sigma^2 e}\left(\frac{\mu^{\rm opt}}{\lambda_i^{\rm opt}}+p_{c,i}\right)-\frac{1}{e}\right)+1\right) \tag{25}$$

corresponds to the optimal offloading rate for user i, $\forall i \in \mathcal{K}$.

Proposition 1 can be verified by simply combining Lemmas 1 and 2; hence, we omit its detailed proof for conciseness. Note that (24) is an instance of SDP, which can thus be efficiently solved by off-the-shelf solvers, e.g., CVX [39].

Summarizing, we present Algorithm 1 to solve the computation latency constrained energy minimization problem $(\mathcal{P}1)$.

Algorithm 1 for Solving the Energy Minimization Problem $(\mathcal{P}1)$

1: **Initialization:** Given an ellipsoid $\mathcal{E}((\boldsymbol{\lambda}, \mu), \boldsymbol{A})$ containing $(\boldsymbol{\lambda}^{\rm opt}, \mu^{\rm opt})$, where $(\boldsymbol{\lambda}, \mu)$ is the center of $\mathcal{E}$ and $\boldsymbol{A} \succ \mathbf{0}$ characterizes the volume of $\mathcal{E}$.
2: **Repeat:**
- For each user $i \in \mathcal{K}$, obtain (t_i^*, ℓ_i^*) by Lemma 2 under given λ_i and μ;
- Compute the subgradients of the objective function and the constraints of $(\mathcal{D}1.1)$ as in Section III-B.2;
- Update $\boldsymbol{\lambda}$ and μ using the ellipsoid method [38];

3: **Until** $\boldsymbol{\lambda}$ and μ converge within a prescribed accuracy.
4: **Set** $(\boldsymbol{\lambda}^{\rm opt}, \mu^{\rm opt}) \leftarrow (\boldsymbol{\lambda}, \mu)$.
5: **Output**: Obtain $(\boldsymbol{t}^{\rm opt}, \boldsymbol{\ell}^{\rm opt})$, $\{f_{i,n}^{\rm opt}\}$, and compute $\boldsymbol{Q}^{\rm opt}$ by (24).

Remark 2: Proposition 1 shows that the optimal joint computing and offloading design has the following interesting properties to minimize the energy consumption at the AP.

1) First, if the energy harvesting constraint is not tight for user i (i.e., user i harvests sufficient wireless energy), then no computation offloading is required and user i should compute all the tasks locally (i.e., $\ell_i^{\rm opt} = 0$). This can be explained based on the complementary slackness condition [37], i.e.,

$$\lambda_i^{\rm opt}\left(\frac{\kappa_i C_i^3 (R_i-\ell_i^{\rm opt})^3}{T^2}+\frac{t_i^{\rm opt}}{\tilde{g}_i}\beta\left(r_i^{\rm opt}\right)+p_{c,i}t_i^{\rm opt}\right.$$
$$\left.-T\zeta\,{\rm tr}\left(\boldsymbol{Q}^{\rm opt}\boldsymbol{H}_i\right)\right) = 0, \quad \forall i \in \mathcal{K}. \tag{26}$$

In this case, if the energy harvesting constraint is not tight for user i, then based on (26) we have $\lambda_i^{\rm opt} = 0$,

and accordingly $\ell_i^{\text{opt}} = 0$ holds from (21). This property is intuitive: when the user has sufficient energy to accomplish the tasks locally, there is no need to employ computation offloading that incurs additional energy consumption for the MEC server's computation at the AP.

2) Next, it is always beneficial to leave some bits for local computing at each user $i \in \mathcal{K}$, i.e., $\ell_i^{\text{opt}} < R_i$ always holds (see (21)). In other words, offloading all the bits to the AP is always suboptimal. This is because when $\ell_i^{\text{opt}} \to R_i$, the marginal energy consumption of local computing is almost zero, and thus it is beneficial to leave some bits for local computing in this case.
3) Furthermore, it is observed that for each user i, more stringent the energy harvesting constraint is (or the associated dual variable λ_i^{opt} is larger), more bits should be offloaded to the AP with a smaller offloading rate r_i^{opt}. This property follows based on (21) and (25), in which a larger λ_i^{opt} admits a larger ℓ_i^{opt} and a smaller r_i^{opt}.
4) Finally, the number of offloaded bits ℓ_i^{opt} and the offloading rate r_i^{opt} for each user i are affected by the channel gain $\tilde{g}_i$, the block length T, the circuit power $p_{\text{c},i}$, and the MEC energy consumption α per offloaded bit in the following way: 1) when the channel condition becomes better (i.e., $\tilde{g}_i$ becomes larger), both ℓ_i^{opt} and r_i^{opt} increase, and thus user i is likely to offload more bits with a higher offloading rate; 2) a higher circuit power $p_{c,i}$ at the user leads to a higher offloading rate r_i^{opt}; 3) when T or α increases, ℓ_i^{opt} reduces and thus fewer bits are offloaded to the AP.

IV. Numerical Results

In this section, numerical results are provided to gauge the performance of the proposed design with joint WPT, offloading, and computing optimization, as compared to the following four benchmark schemes.

1) *Local computing only:* each user $i \in \mathcal{K}$ accomplishes its computation task by only local computing. This scheme corresponds to solving problem ($\mathcal{P}1$) by setting $\ell_i = 0$, $\forall i \in \mathcal{K}$.
2) *Computation offloading only:* each user $i \in \mathcal{K}$ accomplishes its computation task by fully offloading the computation bits to the AP. This scheme corresponds to solving ($\mathcal{P}1$) by setting $f_{i,n} = 0$, $\forall n$, $\forall i \in \mathcal{K}$, as well as $\ell_i = R_i$ for ($\mathcal{P}1$), $\forall i \in \mathcal{K}$.
3) *Joint design with isotropic WPT:* the N-antenna AP radiates the RF energy isotropically or omni-directionally over all directions by setting $\boldsymbol{Q} = p\boldsymbol{I}$, where $p \geq 0$ denotes the transmit power at each antenna. This scheme corresponds to solving problem ($\mathcal{P}1$) by replacing $\boldsymbol{Q}$ as $p\boldsymbol{I}$ with p being another optimization variable.
4) *Separate MEC-WPT design:* this scheme separately designs the computation offloading for MEC and the energy beamforming for WPT [21], [32]. First, the K users minimize their sum-energy consumption subject to the users' individual computation latency constraints [32]. Then, under the constraints of energy demand at the K users, the AP designs the transmit energy beamforming with minimum energy consumption [21].

In the simulations, the EH efficiency is set as $\zeta = 0.3$. The system parameters are set as (unless stated otherwise): the number of the AP antennas $N = 4$, $C_i = 10^3$ cycles/bit, $\kappa_i = 10^{-28}, \forall i \in \mathcal{K}$ [32], the circuit power $p_{c,i} = 10^{-4}$ Watt (W), the energy consumption per offloaded bit by the MEC server $\alpha = 10^{-4}$ Joule/bit, the receiver noise power $\sigma^2 = 10^{-9}$ W, and the spectrum bandwidth for offloading $B = 2$ MHz. By considering a Rayleigh fading channel model, the wireless channel from the AP to each user $i \in \mathcal{K}$ is set as

$$\boldsymbol{h}_i = \theta_0 d_i^{-3} \bar{\boldsymbol{h}}_i, \quad \boldsymbol{g}_i = \theta_0 d_i^{-3} \bar{\boldsymbol{g}}_i, \tag{27}$$

where $\bar{\boldsymbol{h}}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ and $\bar{\boldsymbol{g}}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, $\forall i \in \mathcal{K}$, is an independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian (CSCG) random vector with zero mean and covariance $\boldsymbol{I}$; $\theta_0 = 6.25 \times 10^{-4}$ (i.e., -32 dB) corresponds to the channel power gain at a reference distance of one meter; d_i denotes the distance from the AP to user $i \in \mathcal{K}$; and the path-loss exponent is assumed to be 3. The numerical results are obtained by averaging over 500 randomized channel realizations. Note that the simulation parameters are specifically chosen, but our approaches can be also applied to other system setups.

A. Case With Homogeneous Users

First, we consider the case with homogeneous users, where the distances from the AP to all the users are identical with $d_i = 5$ meters, $\forall i \in \mathcal{K}$. The corresponding average power loss is set to be 5×10^{-6} (i.e., -53 dB). Additionally, the numbers of computation bits at all users are set to be identical, i.e., $R = R_i$, $\forall i \in \mathcal{K}$. Figs. 3–6 show the average energy consumption at the AP under different system parameters. It is observed that the proposed joint design achieves the lowest average energy consumption at the AP among all the five schemes. The joint design with isotropic WPT achieves a suboptimal performance due to the loss of multi-antenna energy beamforming gain. The suboptimal performance of the separate-design scheme implies the necessity of unified demand-supply optimization in wireless powered MEC systems.

Fig. 3 shows the average energy consumption at the AP versus the time block length T, where $R = 10$ kbits and $K = 10$. First, with a small value of T (e.g., $T = 0.05$ sec), the benchmark schemes but the local-computing-only scheme are observed to achieve a near optimal performance close to that with the proposed joint design, while as T increases, the energy consumption with the local-computing-only scheme significantly decreases, approaching that with the proposed joint design. It is also observed that the energy consumption with the full-offloading-only scheme remains almost unchanged when $T \geq 0.1$ sec. This is due to the fact that in this case, the optimal offloading time for all users is fixed to be around 0.1 sec for saving the circuit energy consumption; hence, increasing T cannot further improve the energy efficiency in this case. By contrast, the energy consumption with

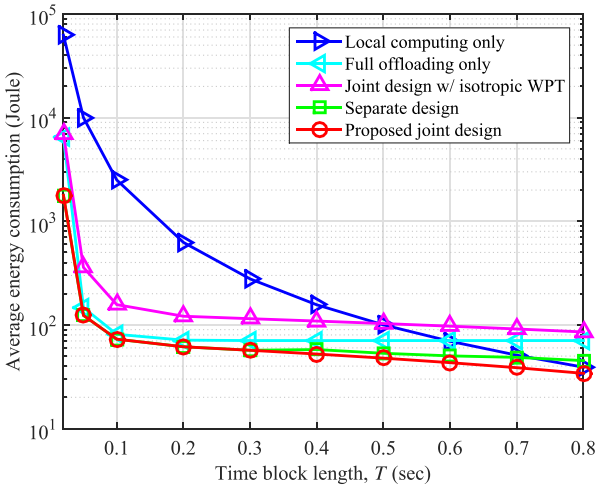

Fig. 3. The average energy consumption at the AP versus the block length T.

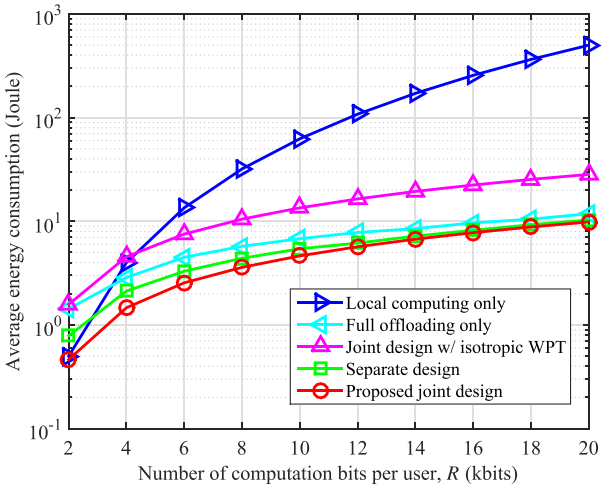

Fig. 5. The average energy consumption at the AP versus the number of computation bits R at each user.

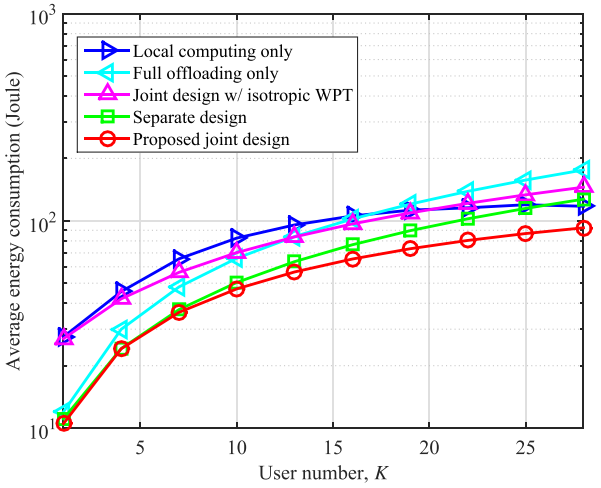

Fig. 4. The average energy consumption at the AP versus the user number K.

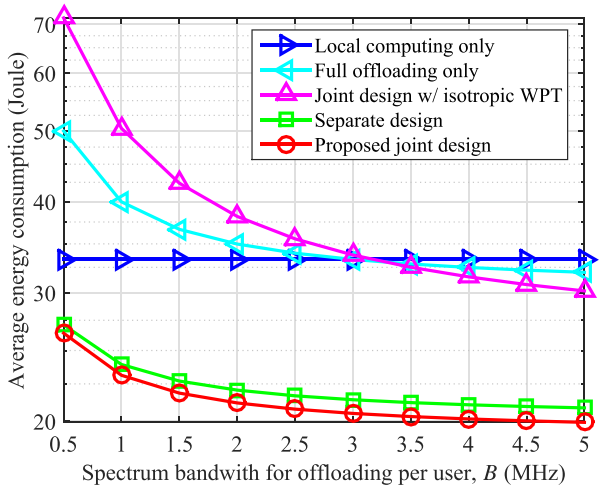

Fig. 6. The average energy consumption at the AP versus the spectrum bandwidth B for offloading.

the local-computing-only scheme decreases monotonically as T increases. This is because as T increases, one can always lower down the CPU frequency to save energy for local computing. Finally, it is seen in Fig. 3 both the separate-design and the equal-offloading-time-allocation schemes achieve a very similar performance in the interested time block regime.

Fig. 4 depicts the average energy consumption versus the user number K, where $R = 10$ kbits and $T = 0.5$ sec. It is shown that the gain achieved by the proposed joint design becomes more significant as the user number K becomes large. The full-loading-only scheme outperforms the local-computing-only scheme, but with a decreasing gain as K increases. This is because in the full-offloading-only scheme, all users share the finite time block and the offloading energy consumption would increase drastically when K becomes large. It is also observed that the performance of the equal-offloading-time-allocation scheme becomes closer to that of the proposed joint design with larger $K \geq 15$. This indicates that an equal offloading time is desirable for a large number of the users in order to minimize the energy consumption at the AP.

Fig. 5 shows the average energy consumption at the AP versus the number of computation bits R at each user, where $K = 2$ and $T = 0.05$ sec. It is shown that the average energy consumption by all the six schemes increases as R becomes large, and the full-offloading-only scheme outperforms the local-computing-only one, especially when R becomes large. This indicates that with large R values, it is desirable to offload more computation bits to the AP in order to reduce the energy consumption. Furthermore, the full-offloading-only scheme is observed to achieve a near optimal performance close to that with the proposed joint design when R becomes large. This is because the energy consumption per bit for offloading is significantly smaller than that for local computing in the large R case. It is also observed that the separate-design scheme outperforms all the other benchmark schemes in this setup.

Fig. 6 shows the average energy consumption versus the spectrum bandwidth B for offloading, where $K = 6$, $T = 0.5$ sec, and $R = 50$ kbits. As expected, Fig. 6 shows that the energy consumption by the four schemes with offloading

TABLE I

OFFLOADED BITS AND RESIDUAL ENERGY AT USERS FOR THE PROPOSED OPTIMAL JOINT DESIGN UNDER DIFFERENT d_2

d_2 (meters)	2	3	4	5	6	7	8
Near user's offloaded bits ℓ_1^{opt} (kbits)	0.165	0.142	0.124	0.092	0.068	0.028	0.006
Far user's offloaded bits ℓ_2^{opt} (kbits)	1.798	6.974	11.817	13.682	13.585	12.972	12.162
Near user's residual energy ($\times 10^{-5}$ Joule)	0.007	0.026	0.062	0.531	3.276	9.218	21.105
Far user's residual energy ($\times 10^{-5}$ Joule)	0.003	0	0	0	0	0	0

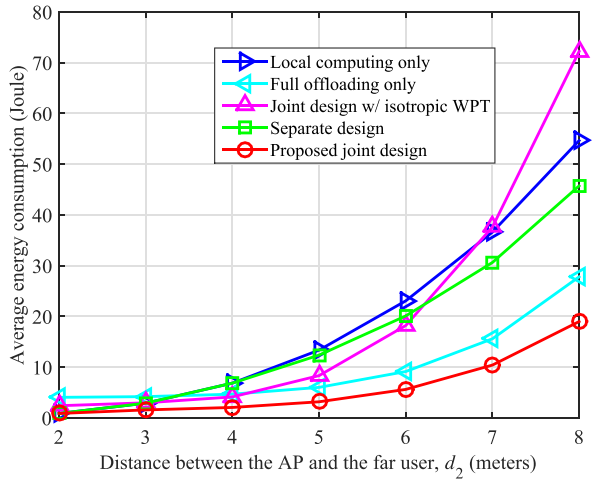

Fig. 7. The average energy consumption at the AP versus the distance d_2 from the AP to the far user.

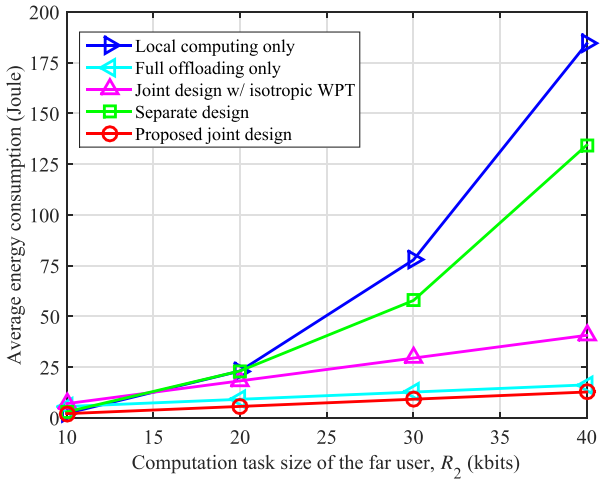

Fig. 8. The average energy consumption at the AP versus the computation task size R_2 kbits.

decreases as B increases, and the one by the local-computing-only scheme remains unchanged. This indicates that a large value of B not only implies a high offloading rate, but also helps save the energy consumption in computation offloading. It is also observed that at small B values (e.g., $B \leq 3$ MHz), the local-computing-only scheme outperforms the full-offloading-only scheme, but it does not hold for large B cases. This indicates that offloading becomes a better option than local-computing as B increases.

B. Case With Heterogeneous Users

Next, we evaluate the performance of the wireless powered MEC system in the case with heterogeneous users. For the purpose of illustration, we focus on the scenario with only $K = 2$ users. It is assumed that the distances from the AP to the two users (namely near and far users) are $d_1 = 2$ meters and $d_2 \geq 2$ meters, respectively. The time block length is set as $T = 0.2$ sec.

Fig. 7 shows the average energy consumption versus the distance d_2 from the AP to the far user, where the computation task sizes for both users are set as $R_1 = R_2 = 20$ kbits. It is observed that as d_2 increases, the energy consumption by the six schemes increases significantly, and the proposed joint design achieves the lowest energy consumption among them. The local-computing-only scheme is observed to outperform the full-offloading-only scheme when $d_2 < 3$ meters, but performs inferior to the full-offloading-only scheme when $d_2 > 3$ meters. Furthermore, the proposed joint design is observed to achieve a significant performance gain over the separate-design one when $d_2 > 4$ meters.

To provide more insights, Table I demonstrates the numbers of offloaded bits ℓ_i^{opt}s at both users and their residual energy (i.e., $E_i - E_{\text{loc},i} - E_{\text{offl},i}$) under different values of d_2 for the proposed joint design. It is observed that the far user prefers offloading significantly more bits than the near user, especially at a larger d_2. As d_2 increases, the number of offloaded bits ℓ_1^{opt} by the near user decreases significantly, while that by the far user (i.e., ℓ_2^{opt}) increases. This result is generally consistent with the first property in Remark 2. Furthermore, it is observed that the residual energy at the near user increases dramatically as d_2 increases, while the far user always uses up all its energy when $d_2 > d_1$. This shows that as d_2 increases, the energy consumption increase at the AP in Fig. 7 is mainly to satisfy the energy requirement at the far user. In this case, the near user will harvest a lot of energy.

Furthermore, we consider the case when the near and far users have distinct computation task sizes. Fig. 8 depicts the average energy consumption versus the computation task size R_2 at the far user, where $R_1 = 20$ kbits and $d_2 = 6$ meters. It is observed that the energy consumption by the six schemes increases as R_2 increases, and both the local-computing-only and the separate-design schemes lead to much higher energy consumption than the other four schemes when $R_2 > 20$ kbits. This is due to the fact that in the local-computing-only and the separate-design schemes, the far user cannot explore the benefit of task offloading for energy saving. Among the five benchmark schemes, the full-offloading-only scheme achieves the best performance close to the optimal proposed one. Table II presents the numbers of offloaded bits at both users

TABLE II
OFFLOADED BITS AND RESIDUAL ENERGY AT USERS FOR THE PROPOSED OPTIMAL JOINT DESIGN UNDER DIFFERENT R_2

R_2 (kbits)	10	20	30	40
Near user's offloaded bits $\ell_1^{\rm opt}$ (kbits)	0.825	0.432	0.239	0.282
Far user's offloaded bits $\ell_2^{\rm opt}$ (kbits)	3.586	13.62	23.791	33.264
Near user's residual energy ($\times 10^{-5}$ Joule)	0.426	3.317	6.42	9.545
Far user's residual energy ($\times 10^{-5}$ Joule)	0	0	0	0

and their residual energy for the proposed joint design. It is observed that as R_2 increases, the number of offloaded bits $\ell_1^{\rm opt}$ by the near user decreases, while $\ell_2^{\rm opt}$ by the far user increases significantly. Similarly as in Table I, $\ell_2^{\rm opt}$ is observed to be significantly larger than $\ell_1^{\rm opt}$. It is also observed that with R_2 increasing, the residual energy at the near user becomes more significant, while that at the far user is zero.

Tables I and II show that when the users are heterogeneous in locations and/or task sizes, even the optimal joint design still leads to unbalanced energy demand and supply at these users. In particular, the AP needs to use a large transmit power to satisfy the high energy demand of users that are far apart and/or have heavy computation tasks. At the same time, the nearby users with light computation tasks can accordingly harvest more energy and are likely to have energy surplus. To better balance the energy demand and surplus, it can be viable to enable user cooperation between near and far users, which is an interesting research direction worth pursuing in future work.

V. CONCLUSION

We developed a unified MEC-WPT design framework with joint energy beamforming, offloading, and computing optimization in emerging wireless powered multiuser MEC systems. In particular, we proposed an efficient wireless powered multiuser MEC design by considering the latency-constrained computation, for which the AP minimizes the total energy consumption subject to the users' individual computation latency constrains. Leveraging the Lagrange duality method, we obtained the optimal solution in a semi-closed form. Numerical results demonstrated the merits of the proposed joint design over alternative benchmark schemes. The proposed unified MEC-WPT design can pave the way to facilitate ubiquitous computing for IoT devices in a self-sustainable way.

APPENDICES

A. Proof of Lemma 1

First, consider the case when there exists some user i with $\ell_i = R_i$, i.e., user i offloads all of its computation task bits to the AP. As user i does not perform local computing in this case, the local CPU frequency of user i is evidently zero.

We next consider the nontrivial case of $0 \le \ell_i < R_i$, $\forall i \in \mathcal{K}$. Define $f_i \triangleq \frac{\sum_{n=1}^{C_i(R_i-\ell_i)} f_{i,n}}{C_i(R_i-\ell_i)}$, $\forall i \in \mathcal{K}$. Since that both $1/x$ and x^2 are convex functions with respect to $x > 0$, based on Jensen's inequality [37], it follows that

$$C_i(R_i-\ell_i)/f_i \le \sum_{n=1}^{C_i(R_i-\ell_i)} 1/f_{i,n} \tag{28a}$$

$$C_i(R_i-\ell_i)\kappa_i f_i^2 \le \sum_{n=1}^{C_i(R_i-\ell_i)} \kappa_i f_{i,n}^2, \tag{28b}$$

where both the equalities hold if and only if

$$f_{i,1} = \ldots = f_{i,C_i(R_i-\ell_i)}, \quad \forall i \in \mathcal{K}. \tag{29}$$

As a result, the optimality of problem $(\mathcal{P}1)$ is achieved when (29) holds. Therefore, by replacing $f_i \triangleq f_{i,n}$, $\forall n$, problem $(\mathcal{P}1)$ is equivalently expressed as

$$\min_{\boldsymbol{Q}\succeq\mathbf{0},\boldsymbol{t},\boldsymbol{\ell},\{f_i\}} \quad T\,\mathrm{tr}(\boldsymbol{Q}) + \sum_{i=1}^{K} \alpha \ell_i \tag{30a}$$

$$\text{s.t.} \quad C_i(R_i-\ell_i)/f_i \le T, \quad \forall i \in \mathcal{K} \tag{30b}$$

$$\kappa_i C_i(R_i-\ell_i) f_i^2 + \frac{t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) + p_{c,i}t_i - T\zeta\,\mathrm{tr}(\boldsymbol{Q}\boldsymbol{H}_i) \le 0, \quad \forall i \in \mathcal{K} \tag{30c}$$

$$\sum_{i=1}^{K} t_i \le T, \quad t_i \ge 0, \ 0 \le \ell_i \le R_i, \ \forall i \in \mathcal{K}. \tag{30d}$$

For a given $(\boldsymbol{t}, \boldsymbol{\ell})$, it is evident that the optimal f_i's for (30) (equivalent $(\mathcal{P}1)$) should be as small as possible by (30c). Since f_i is bounded below by $C_i(R_i-\ell_i)/T$ in (30b), it follows that the optimal f_i's are

$$f_i = C_i(R_i-\ell_i)/T, \quad \forall i \in \mathcal{K}. \tag{31}$$

It then readily follows that, at optimum of $(\mathcal{P}1)$, $f_{i,1} = \ldots = f_{i,C_i(R_i-\ell_i)} = C_i(R_i-\ell_i)/T$, $\forall i \in \mathcal{K}$.

B. Proof of $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \mathbf{0}$

$\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \mathbf{0}$ can be verified by contradiction. Assume that $\boldsymbol{F}(\boldsymbol{\lambda})$ is not positive semidefinite. Denote by $\boldsymbol{\xi}$ one eigenvector corresponding to the negative eigenvalue of $\boldsymbol{F}(\boldsymbol{\lambda})$. By setting $\boldsymbol{Q} = \tau\boldsymbol{\xi}\boldsymbol{\xi}^H \succeq \mathbf{0}$ with τ going to infinity (which is feasible for (13)), it follows that

$$\lim_{\tau\to+\infty} \mathrm{tr}(\boldsymbol{Q}\boldsymbol{F}(\boldsymbol{\lambda})) = \lim_{\tau\to+\infty} \tau\boldsymbol{\xi}^H\boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{\xi} = -\infty, \tag{32}$$

which in turn implies that the objective value in (13) is unbounded below over $\boldsymbol{Q} \succeq \mathbf{0}$. Therefore, in order for the dual function value $\Phi(\boldsymbol{\lambda}, \mu)$ to be bounded below, we need $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \mathbf{0}$.

C. Proof of Lemma 2

Given $(\boldsymbol{\lambda}, \mu) \in \mathcal{S}$, we solve problem (16a) for each user $i \in \mathcal{K}$. When $\lambda_i = 0$, the objective function in (16a) becomes $\alpha\ell_i + \mu t_i$. It is evident that $t_i^* = 0$ and $\ell_i^* = 0$ are optimal for (16a).

For $\lambda_i > 0$, the Lagrangian of (16a) is given by

$$\mathcal{L}_i = \alpha\ell_i + \frac{\lambda_i \kappa_i C_i^3 (R_i - \ell_i)^3}{T^2} + \frac{\lambda_i t_i}{\tilde{g}_i}\beta\left(\frac{\ell_i}{t_i}\right) + \lambda_i p_{c,i} t_i + \mu t_i + \gamma_i(\ell_i - R_i) - \nu_i \ell_i - \eta_i t_i, \quad (33)$$

where γ_i, ν_i, and η_i are the non-negative Lagrangian multipliers associated with $\ell_i \le R_i$, $\ell_i \ge 0$, and $t_i \ge 0$, respectively. Based on the KKT conditions [37], the necessary and sufficient conditions for the optimal primal-dual point $(t_i^*, \ell_i^*, \gamma_i^*, \nu_i^*, \eta_i^*)$ are

$$t_i^* \ge 0, \quad 0 \le \ell_i^* \le R_i \quad (34a)$$

$$\gamma_i^* \ge 0, \quad \nu_i^* \ge 0, \quad \eta_i^* \ge 0 \quad (34b)$$

$$\gamma_i^*(\ell_i^* - R_i) = 0, \quad \nu_i^* \ell_i^* = 0, \ \eta_i^* t_i^* = 0 \quad (34c)$$

$$\frac{\lambda_i}{\tilde{g}_i}\left(\beta\left(\frac{\ell_i^*}{t_i^*}\right) - \frac{\ell_i^*}{t_i^*}\beta'\left(\frac{\ell_i^*}{t_i^*}\right)\right) + \lambda_i p_{c,i} + \mu - \eta_i^* = 0 \quad (34d)$$

$$\alpha - \frac{3\lambda_i \kappa_i C_i^3 (R_i - \ell_i^*)^2}{T^2} + \frac{\lambda_i}{\tilde{g}_i}\beta'\left(\frac{\ell_i^*}{t_i^*}\right) + \gamma_i^* - \nu_i^* = 0, \quad (34e)$$

where $\beta'(x) \triangleq \frac{\sigma^2 \ln 2}{B} 2^{\frac{x}{B}}$ is the first-order derivative of $\beta(x)$ with respect to x. Note that (34c) denotes the complementary slackness condition, while the left-hand-side terms of (34d) and (34e) are the first-order derivatives of $\mathcal{L}_i$ with respect to t_i^* and ℓ_i^*, respectively. For the function $y = \beta(x) - x\beta'(x)$ of $x > 0$, its inverse function can be shown to be [36]

$$x = \frac{B}{\ln 2}\left(W_0\left(-\frac{y}{\sigma^2 e} - \frac{1}{e}\right) + 1\right). \quad (35)$$

Let $r_i^* \triangleq \ell_i^*/t_i^*$. From (34b) and (34d), we have $\beta(r_i^*) - r_i^*\beta'(r_i^*) = -\tilde{g}_i\left(\frac{\mu}{\lambda_i} + p_{c,i}\right)$. Based on (35), it follows that

$$r_i^* = \frac{B}{\ln 2}\left(W_0\left(\frac{\tilde{g}_i}{\sigma^2 e}\left(\frac{\mu}{\lambda_i} + p_{c,i}\right) - \frac{1}{e}\right) + 1\right). \quad (36)$$

Since $W_0(x)$ is a monotonically increasing function of $x \ge -\frac{1}{e}$ and $W_0(-\frac{1}{e}) = -1$ [36], it follows that $r_i^* > 0$ with non-zero $p_{c,i}$. From (34c) and (34e), it is immediate that

$$\ell_i^* = \left[R_i - \sqrt{\frac{1}{3\kappa_i C_i^3}\left(\frac{\alpha}{\lambda_i} + \frac{\sigma^2 \ln 2}{B\tilde{g}_i} 2^{\frac{r_i^*}{B}}\right)}\right]^+. \quad (37)$$

With (36) and (37), the optimal t_i^* is then obtained as $t_i^* = \ell_i^*/r_i^*$.

D. Proof of Lemma 3

The positive semidefinite constraint $\boldsymbol{F}(\boldsymbol{\lambda}) \succeq \mathbf{0}$ can be equivalently expressed as a scalar inequality constraint as [37]

$$\pi(\boldsymbol{\lambda}) \triangleq \min_{\|\boldsymbol{\xi}\|=1} \boldsymbol{\xi}^H \boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{\xi} \ge 0. \quad (38)$$

Given a query point $\boldsymbol{\lambda}_1 \triangleq [\lambda_{1,1}, \ldots, \lambda_{1,K}]^\dagger$, one can find the normalized eigenvector $\boldsymbol{v}_1$ of $\boldsymbol{F}(\boldsymbol{\lambda}_1)$ corresponding to the smallest eigenvalue of $\boldsymbol{F}(\boldsymbol{\lambda}_1)$ (i.e., $\pi(\boldsymbol{\lambda}_1)$). Consequently, we can determine the value of the scalar constraint at a query point as $\pi(\boldsymbol{\lambda}_1) = \boldsymbol{v}_1^H \boldsymbol{F}(\boldsymbol{\lambda}_1)\boldsymbol{v}_1$. To obtain a subgradient, we have

$$\begin{aligned}\pi(\boldsymbol{\lambda}) - \pi(\boldsymbol{\lambda}_1) &= \min_{\|\boldsymbol{\xi}\|=1} \boldsymbol{\xi}^H \boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{\xi} - \boldsymbol{v}_1^H \boldsymbol{F}(\boldsymbol{\lambda}_1)\boldsymbol{v}_1 \\ &\le \boldsymbol{v}_1^H \left(\boldsymbol{F}(\boldsymbol{\lambda}) - \boldsymbol{F}(\boldsymbol{\lambda}_1)\right)\boldsymbol{v}_1 \quad (39a) \\ &= \sum_{i=1}^{K} \left(\lambda_{1,i} - \lambda_i\right)\zeta \boldsymbol{v}_1^H \boldsymbol{H}_i \boldsymbol{v}_1, \quad (39b)\end{aligned}$$

where the last equality follows from the affine structure of $\boldsymbol{F}(\cdot)$ in (14b). By the weak subgradient calculus [37], the subgradient of $\boldsymbol{F}(\boldsymbol{\lambda})$ at the given $\boldsymbol{\lambda}$ and μ is then

$$[\zeta \boldsymbol{v}^H \boldsymbol{H}_1 \boldsymbol{v}, \ldots, \zeta \boldsymbol{v}^H \boldsymbol{H}_K \boldsymbol{v}, 0]^\dagger, \quad (40)$$

where $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{F}(\boldsymbol{\lambda})$, and the last zero entry follows from the fact that $\pi(\boldsymbol{\lambda})$ is independent of μ.

REFERENCES

[1] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," in *Proc. IEEE ICC*, Paris, France, May 2017, 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] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Survey Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017. [Online]. Available: https://arxiv.org/abs/1701.01090

[4] S. Barbarossa, S. Sardellitti, and P. Di 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.

[5] E. Cuervo *et al.*, "MAUI: Making smartphones last longer with code offload," in *Proc. ACM MobiSys*, San Francisco, CA, USA, Jun. 2010, pp. 49–62.

[6] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 945–953.

[7] Y. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing: A key technology towards 5G," ETSI, Sophia Antipolis, France, White Paper 11, 2015. [Online]. Available: http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp11_mec_a_key_technology_towards_5g.pdf

[8] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[9] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, May 2013.

[10] D. W. K. Ng, E. S. Lo, and R. Schober, "Robust beamforming for secure communication in systems with wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4599–4615, Aug. 2014.

[11] S. Timotheou, I. Krikidis, G. Zheng, and B. Ottersten, "Beamforming for MISO interference channels with QoS and RF energy transfer," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2646–2658, May 2014.

[12] S. Bi, C. K. Ho, and R. Zhang, "Wireless powered communication: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 117–125, Apr. 2015.

[13] H. Li, J. Xu, R. Zhang, and S. Cui, "A general utility optimization framework for energy-harvesting-based wireless communications," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 79–85, Apr. 2015.

[14] J. Xu, L. Liu, and R. Zhang, "Multiuser MISO beamforming for simultaneous wireless information and power transfer," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4798–4810, Sep. 2014.

[15] F. Wang, T. Peng, Y. Huang, and X. Wang, "Robust transceiver optimization for power-splitting based downlink MISO SWIPT systems," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1492–1496, Sep. 2015.

[16] F. Wang, C. Xu, Y. Huang, X. Wang, and X.-Q. Gao, "REEL-BF design: Achieving the SDP bound for downlink beamforming with arbitrary shaping constraints," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2672–2685, May 2017.

[17] Y. Zeng and R. Zhang, "Optimized training design for wireless energy transfer," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 536–550, Feb. 2015.

[18] J. Xu and R. Zhang, "Energy beamforming with one-bit feedback," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5370–5381, Oct. 2014.

[19] J. Xu and R. Zhang, "A general design framework for MIMO wireless energy transfer with limited feedback," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2475–2488, May 2016.

[20] C. R. Valenta and G. D. Durgin, "Harvesting wireless power: Survey of energy-harvester conversion efficiency in far-field, wireless power transfer systems," *IEEE Microw. Mag.*, vol. 15, no. 4, pp. 108–120, Jun. 2014.

[21] Y. Zeng, B. Clerckx, and R. Zhang, "Communications and signals design for wireless power transmission," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2264–2290, May 2017.

[22] S. Lee and R. Zhang, "Distributed wireless power transfer with energy feedback," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1685–1699, Apr. 2017.

[23] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082–2085, Dec. 2015.

[24] B. Clerckx and E. Bayguzina, "Waveform design for wireless power transfer," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6313–6328, Dec. 2016.

[25] F. Liu *et al.*, "Gearing resource-poor mobile devices with powerful clouds: Architectures, challenges, and applications," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 14–22, Jun. 2013.

[26] 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.

[27] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 1991–1995, Jun. 2012.

[28] 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.

[29] 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, Aug. 2016.

[30] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for multi-user multi-task mobile cloud," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[31] 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.

[32] 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.

[33] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[34] 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. 1757–1771, May 2016.

[35] 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, 1996.

[36] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert *W* function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, Mar. 2004.

[38] S. Boyd, "Ellipsoid method," Stanford Univ., Stanford, CA, USA, Tech. Rep., May 2014. [Online]. Available: http://stanford.edu/class/ee364b/lectures/ellipsoid_method_notes.pdf

[39] M. Grant, S. Boyd, and Y. Ye. (2009). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://cvxr.com/cvx/

Feng Wang (M'16) received the B.Eng. degree from the Nanjing University of Posts and Telecommunications, China, in 2009, and the M.Sc. and Ph.D. degrees from Fudan University, China, in 2012 and 2016, respectively. From 2012 to 2013, he was a Research Fellow with the Department of Communication Technology, Sharp Laboratories of China. In 2017, he joined the Engineering Systems and Design Pillar, Singapore University of Technology and Design, as a Post-Doctoral Research Fellow. He is currently an Assistant Professor with the School of Information Engineering, Guangdong University of Technology, China. His research interests include signal processing for communications, wireless power transfer, and edge computing.

Jie Xu (S'12–M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of 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. From 2015 to 2016, he was a Post-Doctoral Research Fellow with the Engineering Systems and Design Pillar, Singapore University of Technology and Design. He is currently with the School of Information Engineering, Guangdong University of Technology, China. His research interests include energy efficiency and energy harvesting in wireless communications, wireless information and power transfer, wireless securities, UAV communications, and mobile-edge computing. He was a recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2017. He is currently an Editor of the IEEE WIRELESS COMMUNICATIONS LETTERS, an Associate Editor of the IEEE ACCESS, and a Guest Editor of the IEEE WIRELESS COMMUNICATIONS.

Xin Wang (SM'09) received the B.Sc. and M.Sc. degrees in electrical engineering from Fudan University, Shanghai, China, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering from Auburn University, Auburn, AL, USA, in 2004.

From 2004 to 2006, he was a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA. In 2006, he joined the Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA, as an Assistant Professor, and then an Associate Professor since 2010. He is currently a Distinguished Professor with the Department of Communication Science and Engineering, Fudan University. His research interests include stochastic network optimization, energy-efficient communications, cross-layer design, and signal processing for communications. He served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

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 was an Assistant, Associate, and Full Professor in electrical and computer engineering with The University of Arizona and Texas A&M University. He is currently a Child Family Endowed Chair Professor of electrical and computer engineering with the University of California, Davis, CA, USA. His current research interests focus on 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. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the Worlds' Most Influential Scientific Minds by ScienceWatch in 2014. He was a recipient of the IEEE Signal Processing Society 2012 Best Paper Award. He has served as the general co-chair and TPC co-chair for many IEEE conferences. He has also been serving as an Area Editor for the IEEE *Signal Processing Magazine* and an Associate Editor for the IEEE TRANSACTIONS ON BIG DATA, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE JSAC Series on Green Communications and Networking, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was an Elected Member of the IEEE Signal Processing Society SPCOM Technical Committee from 2009 to 2014 and the Elected Chair of the IEEE ComSoc Wireless Technical Committee from 2017 to 2018. He was elected as the IEEE ComSoc Distinguished Lecturer in 2014. He is a member of the Steering Committee for the IEEE TRANSACTIONS ON BIG DATA and the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is also a member of the IEEE ComSoc Emerging Technology Committee.