Optimizing Computation Efficiency for NOMA-Assisted Mobile Edge Computing with User Cooperation

Binbin Su, Qiang Ni, *Senior Member*, *IEEE*, Wenjuan Yu, *Member*, *IEEE*, and Haris Pervaiz, *Member*, *IEEE*

***Abstract*—In this paper, we investigate the application of user cooperation (UC) and non-orthogonal multiple access (NOMA) schemes for a wireless powered mobile edge computing (MEC) system under the non-linear energy harvesting model, in which two single-antenna mobile users first harvest energy from a multi-antenna access point (AP) integrated with an MEC server. Then, during the computation offloading phase, both mobile users simultaneously offload tasks to the MEC server with the harvested energy, by performing NOMA protocol. To better enhance the system performance, UC scheme is carried out, where the near user acts as a relay to help the far user offload computation tasks to the AP. To obtain energy efficient MEC design, our objective is to maximize the computation efficiency (i.e., the total computation bits divided by the consumed energy) by jointly designing the energy beamforming, time and power allocations, which yields a challenging nonconvex optimization problem. To deal with it, the original problem is first transformed into a more tractable formulation by applying the semidefinite relaxation (SDR) technique and then solved by utilizing the sequential convex approximation (SCA) method. Numerical results demonstrate that UC has a great impact when two users are close, while NOMA makes effect when two users are relatively far. Combining both NOMA and UC, the proposed scheme, named NOMA-UC MEC, yields better system performance than the benchmark schemes.**

***Index Terms*-Mobile edge computing, wireless powered, user cooperation, non-orthogonal multiple access, computation efficiency.**

I. Introduction

The bursting computation-intensive applications prevalent in the Internet of Things (IoT) systems as well as the increasing amount of latency-critical tasks in future-generation networks pose significant challenges in real-time communication system design [1]. To address the requirements of the growing demand for massive computing and overcome the resource limitations (i.e., small size and low power budget) of mobile devices, mobile edge computing (MEC) has been proposed as a potential solution to enhance mobile users' computational capability and realize low-latency communications [2]. Different from conventional cloud computing, where cloud server is deployed far from mobile devices leading to high transmission cost and long latency, the cloud-like server is integrated with the access point (AP) on the edge of MEC networks [3]. By leveraging MEC in proximity, resource-limited mobile users are enabled to offload computation tasks to the more powerful MEC server for remote execution, which brings the benefit of improved computation capacity and reduced latency.

In addition, to overcome the insufficient power supply of batteries and prolong the sustainable operation for mobile users, wireless power transfer (WPT) has emerged as a promising solution via energizing mobile devices remotely [4]. Specifically, WPT is used to charge the battery of energy-harvesting devices by adopting the dedicated radio frequency (RF) energy transmitters at the AP. Moreover, multi-antenna transmitters can be further employed to improve the energy harvesting efficiency by properly designing energy beamforming. The integration of WPT and MEC is envisioned to significantly improve the computation performance. For example, the joint computation offloading and computing resource allocation has been investigated in [5] to minimize the system energy consumption for a wireless powered multi-user MEC system. The authors in [6] maximized the sum computation rate for wireless powered MEC under binary offloading by jointly optimizing the computing mode selection and transmission time allocation. A wireless powered cooperative MEC system has been presented in [7], where nearby devices are exploited as MEC servers. It is noted that all the works mentioned above mainly concentrate on the linear energy harvesting model, which is inaccurate in practice due to the existence of non-linear elements such as diode-connected transistors in RF circuits [8]. Henceforth, a more practical non-linear energy harvesting model [9] should be further taken into account.

Moreover, considering the features and characteristics of the wireless powered MEC model, it suffers from serious "doubly near-far" effect, caused by the double distance-dependent signal attenuation in both the downlink WPT transmission and the uplink computation offloading phase. Consequently, serious unfairness arises among users. On one hand, user cooperation (UC) can be regarded as an effective way to improve the capacity and guarantee user fairness, which enables the near user to act as a relay to transmit the signal of the far user [10]. Specifically, a wireless powered cooperation-assisted MEC system has been investigated in [11], where UC is utilized to improve system performance. On the other hand, non-orthogonal multiple access (NOMA), compared with conventional orthogonal multiple access (OMA) schemes, has

Binbin Su is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K, and also with the National Key Laboratory of Science and Technology on Vessel Integrated Power System, Naval University of Engineering, Wuhan, 430033, China. (Email: subinbinbin_sdu@yahoo.com).

Qiang Ni, Wenjuan Yu, and Haris Pervaiz are with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K. (Email: {q.ni, w.yu8, h.b.pervaiz}@lancaster.ac.uk).

been shown to gain more benefits, i.e., increasing the system capacity and improving the performance of weak users [12]. The authors in [13] minimized the overall delay of mobile users for the NOMA-assisted MEC system. However, all the above researches only considered the case of single-antenna AP. Besides, the application of NOMA was not studied in [11], while there is lack of UC application in [14], [15]. Therefore, to further enhance system performance, it is of vital importance to investigate the multi-antenna NOMA-assisted MEC system with UC.

Note that energy efficient communications have drawn tremendous attention due to the fact that the ever-increasing energy consumption of the information and communication technologies (ICT) contributes more and more to the greenhouse gas emissions [16]. However, most previous works on MEC systems focus on either maximizing the sum computation rates [6], [17], or minimizing the consumed energy [11], [18], which cannot achieve good tradeoff between the energy consumption and the compassable computation bits. Therefore, to better reveal the system efficiency from the perspective of the computation bits per Joule, the computation efficiency measurement metric [19], [20], defined as the ratio of the system computation bits to the consumed energy, is adopted in this paper. Motivated by the above observations, we aim to maximize the computation efficiency of the proposed wireless powered NOMA-assisted MEC system with UC by jointly optimizing the computation and computing resource allocations.

The main contributions of this paper are summarized as follows:

- With the introduction of NOMA and UC, a novel wireless powered MEC system is proposed to overcome the "doubly near-far" effect, where a practical non-linear energy harvesting model is considered. To provide an energy efficient design, a new measurement metric, namely computation efficiency, is adopted. The objective is to maximize the system computation efficiency while satisfying the quality of service (QoS) computation requirements, by jointly optimizing the energy beamforming, power and time allocations.
- Due to the incorporation of the multi-antenna AP and non-linear energy harvesting model, the formulated problem is nonconvex. To solve the intractable formulation, semidefinite relaxation (SDR) technique is firstly employed to linearize the energy beamforming terms, and rank-one optimality is proved to demonstrate SDR tightness. Then the reformulated problem is further converted into convex approximations with the aid of sequential convex approximation (SCA).
- Numerical results verify the theoretical analysis and demonstrate that the partial offloading scheme achieves the best system performance. In addition, the proposed design, i.e., NOMA-UC MEC, outperforms the benchmark schemes.

The remainder of this paper is organized as follows. In Section II, we present the system model of the wireless powered NOMA-assisted MEC with UC and formulate the computation efficiency optimization problem. A solution approach based on SDR and SCA is developed in Section III. Simulation results are presented in Section IV and finally the paper is concluded in Section V.

II. System model and problem formulation

Consider a wireless powered MEC system, which is composed of a N_t-antenna AP integrated with a MEC server and K single-antenna users, namely as $\mathcal{U} = \{U_1, U_2, \cdots, U_K\}$. Hybrid NOMA technique is applied to pair users into J NOMA clusters, where each cluster can have S_j number of users represented by, $\mathbf{J} = \{S_j\}, \forall j \in \{1, 2, \cdots, J\}$ such that $\sum_{j=1}^{J} |S_j| = K$ and $S_j \geq 2$. In practice, each cluster should have no more than 3 to 4 users. A special case for a wireless powered MEC system of a N_t-antenna AP integrated with a MEC server and K single-antenna users with J clusters such that one of the j^{th} cluster with $S_j = 2$ users as shown in Fig. 1. Without loss of generality, U_1 is assumed to be far from the AP and U_2 is close to the AP. Let d_1, d_2, and d_{12} represent the distance between U_1 and the AP, U_2 and the AP, and that from U_1 to U_2, respectively. Particularly, $d_1 \geq d_{12}$ is assumed to guarantee that U_2 has an advantage in decoding U_1's message than the AP.

The system is assumed to be divided into resource blocks, and the time duration of each block is T seconds. T is chosen to be no more than the user latency requirement and the channel coherence time, hence the channels remain unchanged during one block. It is assumed that perfect channel state information (CSI) is available at the AP [1]. For a given block, two processes, namely the WPT phase and the computation offloading phase, will be performed. During the WPT phase, the AP broadcasts wireless energy via downlink transmission and the received signals at both users can be expressed as

$$y_j = \boldsymbol{g}_j^H \boldsymbol{w} x + n_j, \ j = \{1, 2, \cdots, S_j\}, \tag{1}$$

where $\boldsymbol{g}_j \in \mathbb{C}^{N_t \times 1}$ is the channel gain from AP to U_j, $j = \{1, 2, \cdots, S_j\}$, $\boldsymbol{w} \in \mathbb{C}^{N_t \times 1}$ denotes the RF energy beamforming vector, x is the RF energy signal with normalized transmit power, i.e., $\mathbb{E}[\|x\|^2] = 1$, and n_i is the additive white Gaussian noise (AWGN) following $n_i \sim \mathcal{CN}(0, \sigma^2)$.

The received RF power at the receiver can be denoted as

$$P_j(\boldsymbol{w}) = |\boldsymbol{g}_j^H \boldsymbol{w}|^2, j = \{1, 2, \cdots, S_j\}. \tag{2}$$

For the considered non-linear energy harvesting model, according to [8], [9], the harvested energy at the users during the WPT phase occupying the time period t_0 can be expressed as

$$E_j = t_0 \left[\frac{\Psi_j}{X_j} - Y_j \right], \ j = \{1, 2, \cdots, S_j\}, \tag{3}$$

with

$$\Psi_j = \frac{Q_j}{1 + \exp(-a_j(P_j(\boldsymbol{w}) - b_j))}, \tag{4a}$$

[1]For time division duplexing (TDD) mode, by sending a beacon signal at the beginning of a time slot, the BS can synchronize the uplink transmissions. This beacon signal can be used as a pilot signal to estimate the CSI. Though the estimation of CSI may be imperfect, the perfect CSI can serve as the upper bound on MEC design for imperfect CSI scenarios. Perfect CSI has been used as a common assumption in many research studies on MEC design. On the other hand, for the imperfect CSI scenarios, robust optimization techniques [21], [22] can be applied to deal with the channel uncertainties, which is a potential research topic in future studies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3056770, IEEE Transactions on Green Communications and Networking

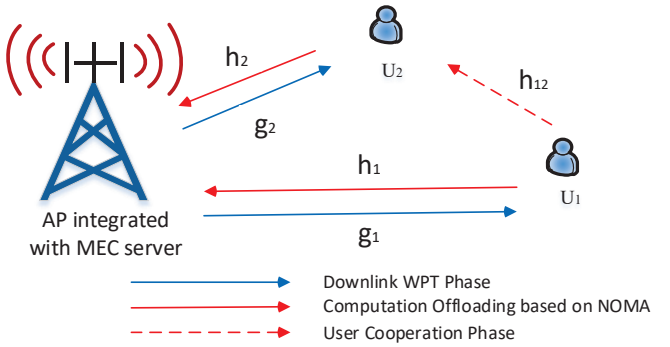

Fig. 1. NOMA-assisted wireless powered MEC with user cooperation.

$$X_j = \frac{\exp(a_j b_j)}{1 + a_j b_j}, \tag{4b}$$

$$Y_j = \frac{Q_j}{\exp(a_j b_j)}, \tag{4c}$$

where Q_j, a_j and b_j are constants capturing the non-linear properties of the energy harvesting system. Specifically, Q_j denotes the maximum output power of the energy harvesting circuits, while a_j and b_j reflect the hardware phenomena, i.e., the capacitance, the resistance and the circuit sensitivity.

A. Computation Offloading Phase

The partial offloading case is considered, where the computation task of each user is divided into two parts for remote execution at the AP and local computing, respectively. Practically, partial offloading is suitable for the scenario of complex tasks composed of multiple parallel segments. Moreover, compared with full offloading, partial offloading is more beneficial to reduce the latency since it takes advantage of parallelism between users and the AP [23]. To reduce system complexity, we focus on one pair of two NOMA users, i.e., U_1 and U_2, which are served within the same resource block [2]. The time allocation structure of the computation offloading phase with $S_j = 2$ is illustrated in Fig. 2. To exploit the benefit of UC while maintaining the advantages of NOMA, three slots are included for UC-enabled uplink NOMA transmission. The offloaded information of both users is divided into two segments, where the two segments are transmitted to the AP directly in the first and the third slot for User 2. For User 1, the first segment is transmitted collaboratively to the AP in the first and second slots, and the second segment is transmitted directly to the AP in the third slot. Specifically, during the subsequent period t_1, due to the application of NOMA protocol, U_1 and U_2 offload some input-bits simultaneously with power p_{11} and p_{20}. Then, by assuming that U_2 works in full duplex mode, both the AP and U_2 can decode the signal of U_1, while the AP also needs to decode U_2's information. For information decoding at the AP, the user with the better channel gain is firstly decoded for uplink NOMA, i.e., the AP first detects U_2's message by treating the message of U_1 as noise, and then removes it with SIC to further decode U_1's information. The remaining time is divided into two parts, given as t_{21} and t_{22}. UC is applied during the second period t_{21}, i.e., U_2 acts as a DF relay to forward the signal of U_1 to the AP with power p_{21}. In the third slot t_{22}, U_1 and U_2 offload their own input-bits to the AP with power p_{12} and p_{22}.

Combing the observation from both t_1 and t_{22}, by regarding the signal of U_1 as noise, the offloaded data size of U_2 can be characterized as

$$l_2^{off} \leq t_1 \text{Blog}_2(1+\frac{p_{20}|\boldsymbol{h}_2|^2}{I_1+\sigma^2})+t_{22}\text{Blog}_2(1+\frac{p_{22}|\boldsymbol{h}_2|^2}{I_2+\sigma^2}), \tag{5}$$

where $\boldsymbol{h}_i, i = \{1, 2\}$, denotes the uplink channel gain from the users to the MEC server. Then, $I_1 = p_{11}|\boldsymbol{h}_1|^2$ and $I_2 = p_{12}|\boldsymbol{h}_1|^2$ represent the interference caused by U_1 during t_1 and t_{22}.

For uplink NOMA transmission, to guarantee the correct SIC decoding in a given order and allocate non-trivial data rate to U_2, the following inequality should be satisfied [25]:

$$p_{20}|\boldsymbol{h}_2|^2 \geq p_{11}|\boldsymbol{h}_1|^2, \tag{6a}$$

$$p_{22}|\boldsymbol{h}_2|^2 \geq p_{12}|\boldsymbol{h}_1|^2. \tag{6b}$$

After removing the signal of U_2, the offloaded data size of U_1 can be given as

$$l_1^{off} = l_{1,1} + l_{1,2}, \tag{7}$$

where $l_{1,1}$ represents the offloaded data size of U_1 via the help of UC scheme. Based on [26], $l_{1,1}$ can be expressed as $l_{1,1} \leq \min\{l_{1,direct}, l_{1,relay}\}$, where $l_{1,direct}$ and $l_{1,relay}$ are the offloaded data size of U_1 at the AP and U_2, which are given as $l_{1,direct} = t_1\text{Blog}_2(1+\frac{p_{11}|\boldsymbol{h}_1|^2}{\sigma^2})+t_{21}\text{Blog}_2(1+\frac{p_{21}|\boldsymbol{h}_2|^2}{\sigma^2})$, $l_{1,relay} = t_1\text{Blog}_2(1+\frac{p_{11}|h_{12}|^2}{\sigma^2})$. Moreover, $l_{1,2}$ denotes the offloaded data size during the period t_{22}, i.e., $l_{1,2} \leq t_{22}\text{Blog}_2(1+\frac{p_{12}|\boldsymbol{h}_1|^2}{\sigma^2})$.

Different from [27], we assume that the time consumption of two processes, i.e., task execution at the MEC server and MEC server transmitting computed results back to users, are negligible [11], [15]. The reason is that, the the MEC-integrated AP generally provides sufficient computation and communication capabilities, and the output computed results are much smaller compared with that of the input data sizes. Furthermore, U_1's information decoding time at U_2 is also ignored, as it is much smaller compared with the computation offloading time. Therefore, the system latency constraint including the WPT and computation offloading can be given as

$$t_0 + t_1 + t_{21} + t_{22} \leq T. \tag{8}$$

During this phase, the consumed energy of U_1 and U_2 can be respectively denoted as

$$E_1^{off} = p_{11}t_1 + p_{12}t_{22}, \tag{9a}$$

$$E_2^{off} = p_{20}t_1 + p_{21}t_{21} + p_{22}t_{22}. \tag{9b}$$

[2]The considered two-user model can be easily extended to multiple NOMA pairs after performing user pairing [24], where each pair is composed of a near user and a far user. By allocating orthogonal frequency bands to different NOMA pairs, each NOMA pair can be managed independently, which is exactly the focus of our paper. The simplified two-user case can help build the insightful understanding of the cooperative NOMA-assisted MEC system. Moreover, to reduce system complexity, it is of practical interest to focus on one pair where two users are served in a resource block, since it is unrealistic for a large number of users to perform NOMA in an interference-limited NOMA system.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3056770, IEEE Transactions on Green Communications and Networking

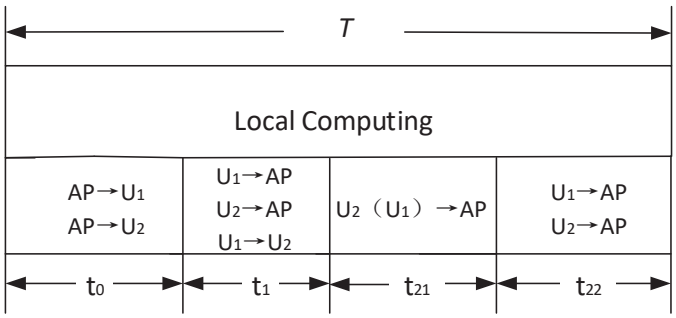

Fig. 2.Time allocation structure for the wireless powered NOMA-assisted MEC with user cooperation.

B. Local Computing

Furthermore, during the whole block duration T, $l_i^{loc}, i = \{1,2\}$, input-bits are executed by local computing at the users. Similar to [28], [29], identical CPU frequency $f_i = \frac{C_i l_i^{loc}}{T}$ is adopted for CPU cycle, where $C_i, i = \{1,2\}$, denotes the number of required CPU cycles to compute one input-bit locally. f_i is constrained by a maximum CPU frequency $f_{\max}$, which can be equivalently expressed as

$$C_i l_i^{loc} \leq T f_{\max}. \tag{10}$$

Accordingly, the consumed energy for local computing can be given as

$$E_i^{loc} = \frac{\kappa_i C_i^3 (l_i^{loc})^3}{T^2},\ i = \{1,2\}, \tag{11}$$

where κ_i is a constant denoting the effective capacitance coefficient and the value is dependent on the chirp architecture [30].

Due to the fact that the consumed energy at the users cannot exceed the total harvested energy obtained from WPT, we have that

$$E_i^{loc} + E_i^{off} \leq E_i,\ i = \{1,2\}. \tag{12}$$

The computation efficiency is defined as a ratio of the total calculated data bits to the system energy consumption, which can be given as

$$\eta = \frac{\sum_{i=1}^{2} l_i^{loc} + l_i^{off}}{t_0 |\boldsymbol{w}|^2}. \tag{13}$$

Finally, with the aim of obtaining an energy efficient design, the computation efficiency maximization problem is formulated as follows

$$\begin{aligned}
(\mathbf{P1}) \quad & \max_{\boldsymbol{w},\boldsymbol{t},\boldsymbol{p},\boldsymbol{l}} \quad \eta, && (14a)\\
& s.t. \quad t_0 + t_1 + t_{21} + t_{22} \leq T, && (14b)\\
& E_i^{loc} + E_i^{off} \leq E_i, i = 1,2, && (14c)\\
& l_i^{loc} + l_i^{off} \geq L_i, i = 1,2, && (14d)\\
& |\boldsymbol{w}|^2 \leq P_{\max}, && (14e)\\
& \boldsymbol{t} \geq 0, \boldsymbol{p} \geq 0, \boldsymbol{l} \geq 0 && (14f)\\
& (5),\ (6a),\ (6b),\ (7),\ (9a),\ (9b), && (14g)
\end{aligned}$$

where $\mathbf{t} = [t_0, t_1, t_{21}, t_{22}]$, $\mathbf{p} = [p_{11}, p_{12}, p_{20}, p_{21}, p_{22}]$, and $\boldsymbol{l} = [l_1^{off}, l_2^{off}, l_1^{loc}, l_2^{loc}]$ denote the time allocation vector, the power allocation vector and the calculated data size sets for computation offloading and local computing, respectively. Further, constraint (14d) denotes the minimum required computing data bits for user i, $i = \{1,2\}$. The maximum available power at the AP is limited by (14e).

III. Solution Approach

Note that (**P1**) is a nonconvex problem, which cannot be solved directly. The challenge is twofold, i.e., 1) the objective is a fractional function involving the energy beamfoming vector, 2) the expressions of U_2's offloading data size and the adopted non-linear energy harvesting model are complicated. In this section, the optimal time allocation condition is first provided. Then, we relax the problem by leveraging the SDR approach. For the relaxed problem, the objective function, the energy-limited constraints and $U_2's$ offloading bits l_2^{off} are further converted into convex approximations with the application of SCA.

A. SCA-based Approach

Firstly, to solve (**P1**), the optimal time utilization is obtained with the following lemma.

Lemma 1. *The maximum computation efficiency of (**P1**) can be achieved with* $t_0 + t_1 + t_{21} + t_{22} = T$.

$Proof$: The proof is provided in Appendix A.

To deal with the beamforming vector $\boldsymbol{w}$, SDR technique is applied to transform (**P1**) into a more tractable form. Specifically, $\boldsymbol{w}$ is replaced by the semidefinite positive matrix, i.e., $\boldsymbol{W} = \boldsymbol{w}\boldsymbol{w}^H$. The constraint (14e) can be then reformulated as

$$\mathrm{Tr}(\boldsymbol{W}) \leq P_{\max}, \boldsymbol{W} \succeq 0, \mathrm{rank}(\boldsymbol{W}) \leq 1. \tag{15}$$

Then, by introducing some slack variables $\boldsymbol{\zeta} = [\zeta_1, \zeta_2]$, $\boldsymbol{\tau} = [\tau_1, \tau_2]$, and several substitution variables, i.e., $\boldsymbol{E} = [E_{11}, E_{12}, E_{20}, E_{21}, E_{22}]$, where $E_{11} = t_1 p_{11}$, $E_{12} = t_{22} p_{12}$, $E_{20} = t_1 p_{20}$, $E_{21} = t_{21} p_{21}$, and $E_{22} = t_{22} p_{22}$, (14c) can be decoupled into the following constraints

$$E_{11} + E_{12} + \frac{\kappa_1 C_1^3 (l_1^{loc})^3}{T^2} \leq \zeta_1, \tag{16a}$$

$$(\frac{\zeta_1}{t_0} + Y_1) X_1 \leq \frac{M_1}{1 + \exp(-a_1(\tau_1 - b_1))}, \tag{16b}$$

$$\tau_1 \leq \mathrm{Tr}(\boldsymbol{G}_1 \boldsymbol{W}), \tag{16c}$$

$$E_{20} + E_{21} + E_{22} + \frac{\kappa_2 C_2^3 (l_2^{loc})^3}{T^2} \leq \zeta_2, \tag{16d}$$

$$(\frac{\zeta_2}{t_0} + Y_2) X_2 \leq \frac{M_2}{1 + \exp(-a_2(\tau_2 - b_2))}, \tag{16e}$$

$$\tau_2 \leq \mathrm{Tr}(\boldsymbol{G}_2 \boldsymbol{W}), \tag{16f}$$

where $\boldsymbol{G}_i \triangleq \boldsymbol{g}_i \boldsymbol{g}_i^H$, $i = \{1,2\}$. Note that after the reformulations, (16b) and (16e) are still nonconvex constraints.

Further, (6a) and (6b) can be reformulated as

$$E_{20} |\boldsymbol{h}_2|^2 \geq E_{11} |\boldsymbol{h}_1|^2, \tag{17a}$$

$$E_{22} |\boldsymbol{h}_2|^2 \geq E_{12} |\boldsymbol{h}_1|^2. \tag{17b}$$

By further introducing two slack variables and applying the epigraph reformulation, (16b) can be reformulated as

$$(\upsilon_1 + Y_1) X_1 \leq \frac{M_1}{1 + \exp(-a_1(\tau_1 - b_1))}, \tag{18a}$$

$$\begin{bmatrix} \upsilon_1 & \omega_1 \\ \omega_1 & t_0 \end{bmatrix} \succeq 0, \tag{18b}$$

$$\omega_1^2 \geq \zeta_1, \tag{18c}$$

where (18a) is a convex function, (18b) is a convex linear matrix inequality (LMI), and (18c) is nonconvex.

Moreover, SCA can be adopted to obtain the convex approximation of (18c). The key idea of SCA is to sequentially optimize (18c) by establishing a convex trust region around the original nonconvex spatial points. Though the results may depend on the initial points, it has been verified that SCA often works well in pratical applications [31]. Due to the convex feature of ω_1^2, the lower bound approximation can be derived by performing the first-order Taylor approximation:

$$\omega_1^2 \geq 2\omega_1^{(n)}\omega_1 - (\omega_1^{(n)})^2, \tag{19}$$

where $\omega_1^{(n)}$ denotes the value of ω_1 during the n-th iteration.

Hence, (18c) is transformed into the following inequality:

$$2\omega_1^{(n)}\omega_1 - (\omega_1^{(n)})^2 \geq \zeta_1, \tag{20}$$

Similarly, (16e) can be approximated as

$$(\upsilon_2 + Y_2)X_2 \leq \frac{M_2}{1+\exp(-a_2(\tau_2 - b_2))}, \tag{21a}$$

$$\begin{bmatrix} \upsilon_2 & \omega_2 \\ \omega_2 & t_0 \end{bmatrix} \succeq 0, \tag{21b}$$

$$2\omega_2^{(n)}\omega_2 - (\omega_2^{(n)})^2 \geq \zeta_2. \tag{21c}$$

Then, by introducing auxiliary variables μ and β, the objective function can be approximated as follows

$$\max_{\boldsymbol{w},\boldsymbol{t},\boldsymbol{p},\boldsymbol{l}} \quad \mu, \tag{22a}$$

$$s.t. \quad \sum_{i=1}^{2} l_i^{loc} + l_i^{off} \geq \sqrt{\mu\beta}, \tag{22b}$$

$$t_0 \mathrm{Tr}(\boldsymbol{W}) \leq \sqrt{\beta}, \tag{22c}$$

where the equivalence is guaranteed when (22b) and (22c) hold with equality at optimum.

It is noted that $\sqrt{\mu\beta}$ is a joint concave function with respect to μ and β, which can be approximated by its upper bound as below

$$\sqrt{\mu\beta} \triangleq g(\mu,\beta) \leq g^{'}(\mu,\beta,\mu^{(n)},\beta^{(n)}), \tag{23a}$$

$$\begin{aligned} g^{'}(\mu,\beta,\mu^{(n)},\beta^{(n)}) = \sqrt{\mu^{(n)}\beta^{(n)}} &+ 0.5\sqrt{\frac{\mu^{(n)}}{\beta^{(n)}}}(\beta - \beta^{(n)}) \\ &+ 0.5\sqrt{\frac{\beta^{(n)}}{\mu^{(n)}}}(\mu - \mu^{(n)}), \end{aligned} \tag{23b}$$

where $\mu^{(n)}$ and $\beta^{(n)}$ denote the value of variables μ and β at the n-th iteration, and $g^{'}(\mu,\beta,\mu^{(n)},\beta^{(n)})$ represents the first-order Taylor approximation around $(\mu^{(n)},\beta^{(n)})$.

Accordingly, (22b) can be reformulated as

$$\sum_{i=1}^{2} l_i^{loc} + l_i^{off} \geq g^{'}(\mu,\beta,\mu^{(n)},\beta^{(n)}). \tag{24}$$

For (22c), arithmetic geometric mean (AGM) method [32] can be applied to get the approximation as

$$(\nu^{(n)} t_0)^2 + (\mathrm{Tr}(\boldsymbol{W})/\nu^{(n)})^2 \leq 2\sqrt{\beta}, \tag{25}$$

where $\nu^{(n)}$ can be updated as below during the n-th iteration

$$\nu^{(n)} = \sqrt{\mathrm{Tr}(\boldsymbol{W})^{(n-1)}/t_0^{(n-1)}}. \tag{26}$$

Moreover, by substituting $\boldsymbol{E}$ into (5), it can be reformulated as below

$$\begin{cases} l_1^{off} = l_{1,1} + l_{1,2}, \\ l_{1,1} \leq t_1 \mathrm{Blog}_2(1 + \frac{E_1|\boldsymbol{h}_1|^2}{t_1\sigma^2}) + t_{21}\mathrm{Blog}_2(1 + \frac{E_{21}|\boldsymbol{h}_2|^2}{t_{21}\sigma^2}), \\ l_{1,1} \leq t_1 \mathrm{Blog}_2(1 + \frac{E_1|h_{12}|^2}{t_1\sigma^2}), \\ l_{1,2} \leq t_{22}\mathrm{Blog}_2(1 + \frac{E_{12}|h_1|^2}{t_{22}\sigma^2}). \end{cases} \tag{27}$$

As a function $f^{'}(x) = \log_2(1+\frac{x}{\sigma^2})$ is concave, its perspective function $tf^{'}(\frac{x}{t}) = \log_2(1+\frac{x}{t\sigma^2})$ is also concave. This indicates that the constraints in (27) are all convex.

Then, by introducing two slack variables $l_{2,1}^{off}$ and $l_{2,2}^{off}$, the offloaded data size of U_2 can be recast as

$$l_2^{off} \leq l_{2,1}^{off} + l_{2,2}^{off}, \tag{28a}$$

$$l_{2,1}^{off} \leq t_1 \mathrm{Blog}_2(1 + \frac{E_{20}|\boldsymbol{h}_2|^2}{E_{11}|\boldsymbol{h}_1| + \sigma^2 t_1}), \tag{28b}$$

$$l_{2,2}^{off} \leq t_{22} \mathrm{Blog}_2(1 + \frac{E_{22}|\boldsymbol{h}_2|^2}{E_{12}|\boldsymbol{h}_1| + \sigma^2 t_{22}}), \tag{28c}$$

where the optimality can be guaranteed when (28b) and (28c) hold with equality. To further transform (28b), it can be firstly rewritten as $l_{2,1}^{off} \leq m_1(\boldsymbol{E},t_1) - z_1(\boldsymbol{E},t_1)$, where $m_1(\boldsymbol{E},t_1)$ and $z_1(\boldsymbol{E},t_1)$ are defined as

$$m_1(\boldsymbol{E},t_1) = t_1 \mathrm{Blog}_2(1 + \frac{E_{11}|\boldsymbol{h}_1|^2 + E_{20}|\boldsymbol{h}_2|^2}{\sigma^2 t_1}), \tag{29a}$$

$$z_1(\boldsymbol{E},t_1) = t_1 \mathrm{Blog}_2(1 + \frac{E_{11}|\boldsymbol{h}_1|^2}{\sigma^2 t_1}). \tag{29b}$$

It is worth noting that both $m_1(\boldsymbol{E},t_1)$ and $z_1(\boldsymbol{E},t_1)$ are joint concave functions with respect to $\boldsymbol{E}$ and t_1. Therefore, we can see that $m_1(\boldsymbol{E},t_1) - z_1(\boldsymbol{E},t_1)$ is a difference of convex (DC) programming function [33], which can be converted into convex expression with the aid of SCA. As $z_1(\boldsymbol{E},t_1)$ is a concave function, an upper bound can be given by using its first-Taylor expansion as below:

$$\begin{aligned} z_1(\boldsymbol{E},t_1) \leq z_1(\boldsymbol{E}^{(n)},t_1^{(n)}) &+ \nabla z_1(\boldsymbol{E}^{(n)})(E_{11} - E_{11}^{(n)}) \\ &+ \nabla z_1(t_1^{(n)})(t_1 - t_1^{(n)}), \end{aligned} \tag{30}$$

where $E_{11}^{(n)}$ and $t_1^{(n)}$ represent the values of E_{11} and t_1 at the n-th iteration. $\nabla z_1(\boldsymbol{E}^{(n)})$ and $\nabla z_1(t_1^{(n)})$ denote the gradients of $z_1(\boldsymbol{E},t_1)$ over E_{11} and t_1, which are expressed as

$$\begin{cases} \nabla z_1(\boldsymbol{E}^{(n)}) = \frac{\mathrm{B}t_1^{(n)}|\boldsymbol{h}_1|^2}{(E_{11}^{(n)}|\boldsymbol{h}_1|^2 + \sigma^2 t_1^{(n)})\ln 2}, \\ \nabla z_1(t_1^{(n)}) = \mathrm{Blog}_2(1 + \frac{E_{11}^{(n)}|\boldsymbol{h}_1|^2}{\sigma^2 t_1^{(n)}}) - \frac{\mathrm{B}E_{11}^{(n)}|\boldsymbol{h}_1|^2}{(E_{11}^{(n)}|\boldsymbol{h}_1|^2 + \sigma^2 t_1^{(n)})\ln 2}. \end{cases} \tag{31}$$

As a result, (28b) can be reformulated as

$$\begin{aligned} l_{2,1}^{off} \leq m_1(\boldsymbol{E},t_1) - z_1(\boldsymbol{E}^{(n)},t_1^{(n)}) - \\ \nabla z_1(\boldsymbol{E}^{(n)})(E_{11} - E_{11}^{(n)}) - \nabla z_1(t_1^{(n)})(t_1 - t_1^{(n)}). \end{aligned} \tag{32}$$

Furthermore, following a similar procedure, (28c) can be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3056770, IEEE Transactions on Green Communications and Networking

then recast as

$$l_{2,2}^{off} \leq m_2(\boldsymbol{E}, t_{22}) - z_2(\boldsymbol{E}^{(n)}, t_{22}^{(n)}) - \nabla z_2(\boldsymbol{E}^{(n)})(E_{12} - E_{12}^{(n)}) - \nabla z_2(t_{22}^{(n)})(t_{22} - t_{22}^{(n)}), \tag{33}$$

where $m_2(\boldsymbol{E}, t_{22})$, $z_2(\boldsymbol{E}, t_{22})$, $\nabla z_2(\boldsymbol{E}^{(n)})$, and $\nabla z_2(t_{22}^{(n)})$ are defined as

$$\begin{cases} m_2(\boldsymbol{E}, t_{22}) = t_{22}\mathrm{Blog}_2(1 + \frac{E_{12}|\boldsymbol{h}_1|^2 + E_{22}|\boldsymbol{h}_2|^2}{\sigma^2 t_{22}}), \\ z_2(\boldsymbol{E}, t_{22}) = t_{22}\mathrm{Blog}_2(1 + \frac{E_{12}|\boldsymbol{h}_1|^2}{\sigma^2 t_{22}}), \\ \nabla z_2(\boldsymbol{E}^{(n)}) = \frac{\mathrm{B}t_{22}^{(n)}|\boldsymbol{h}_1|^2}{(E_{12}^{(n)}|\boldsymbol{h}_1|^2 + \sigma^2 t_{22}^{(n)})\ln 2}, \\ \nabla z_2(t_{22}^{(n)}) = \mathrm{Blog}_2(1 + \frac{E_{12}^{(n)}|\boldsymbol{h}_1|^2}{\sigma^2 t_{22}^{(n)}}) - \frac{\mathrm{B}E_{12}^{(n)}|\boldsymbol{h}_1|^2}{(E_{12}^{(n)}|\boldsymbol{h}_1|^2 + \sigma^2 t_{22}^{(n)})\ln 2}. \end{cases} \tag{34}$$

Finally, the original problem (**P1**) can be transformed into a convex formulation by dropping the rank-one constraint. During the n-th iteration, we need to solve the following convex optimization problem:

$$\begin{aligned} (\mathbf{P2}) \quad & \max_{\boldsymbol{W}, \boldsymbol{t}, \boldsymbol{E}, \boldsymbol{\zeta}, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\upsilon}, \mu, \beta} \quad \mu, && \text{(35a)} \\ & s.t. \quad (16a),\ (16c),\ (16d),\ (16f), && \text{(35b)} \\ & \quad (17a),\ (17b),\ (18a),\ (18b), && \text{(35c)} \\ & \quad (20),\ (21a),\ (21b),\ (21c),\ (24), && \text{(35d)} \\ & \quad (25),\ (27),\ (28a),\ (32),\ (33), && \text{(35e)} \\ & \quad \mathrm{Tr}(\boldsymbol{W}) \leq P_{\max}, && \text{(35f)} \\ & \quad t_0 + t_1 + t_{21} + t_{22} = T, && \text{(35g)} \\ & \quad \boldsymbol{t} \succeq 0, \boldsymbol{E} \succeq 0, \boldsymbol{W} \succeq 0, \boldsymbol{l} \succeq 0, && \text{(35h)} \end{aligned}$$

where $\boldsymbol{\zeta}, \boldsymbol{\tau}, \boldsymbol{\omega}$ and $\boldsymbol{\upsilon}$ represent the corresponding sets of the introduced slack variables.

Therefore, we provide Algorithm 1 to outline the detailed process to solve (**P2**).

Note that the nonconvex rank-one constraint, i.e., $\mathrm{rank}(\boldsymbol{W}) \leq 1$ is dropped for (**P2**). To demonstrate the equivalence between (**P2**) and (**P1**), we provide the following theorem.

Theorem 1. *An optimal solution $\boldsymbol{W}^*$ to* (**P2**) *always exists, whenever the problem is feasible.*

Proof: The proof is provided in Appendix B.

Moreover, to prove the convergence of the proposed Algorithm 1, we have the following theorem.

Theorem 2. *Algorithm 1 produces a non-decreasing sequence of the objective values, i.e., $\mu^{(n+1)} \geq \mu^{(n)}$, which indicates the convergence of Algorithm 1.*

Proof: The proof is provided in Appendix C.

Theorem 3. *The proposed Algorithm 1 continuously converges to a Karush-Kuhn-Tucker (KKT) point of problem (**P1**) whenever problem (**P2**) is feasible.*

Proof: The proof is provided in Appendix D.

B. Complexity Analysis

Note that the computational complexity of Algorithm 1 consists of two loops: the outer iteration loop and the inner loop to solve (**P2**). Specifically, denote the maximum iteration number of Algorithm 1 as L_{max}, while the complexity of the interior point method to solve (**P2**) is proportional to $O(r^{3.5}\delta)$ [34], where r represents the number of variables, and δ accounts for the number of bits needed to denote the entries in the optimization problem. In summary, the whole complexity is $O(L_{max}r^{3.5}\delta)$, where r is the total number of variables $(\boldsymbol{W}, \boldsymbol{t}, \boldsymbol{E}, \boldsymbol{\zeta}, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\upsilon}, \mu, \beta)$ to solve (**P2**).

Algorithm 1 Computation Efficiency Maximization Algorithm

1: Initialize energy allocation $\boldsymbol{E}^{(0)}$ and time allocation $\boldsymbol{t}^{(0)}$, set iteration number $n = 0$, $\mu^0 = 0$, $\mu^1 = 1$, and the precision tolerance $\epsilon = 10^{-3}$.
2: **while** $|\mu^{n+1} - \mu^n| \geq \epsilon$
3: Update the n-th iteration $\boldsymbol{E}^{(n)}$ and $\boldsymbol{t}^{(n)}$ by solving (**P2**);
4: Update $\mu^n = \mu^{n-1}$;
5: Update $n = n + 1$;
6: **end while**
7: Output: the optimal energy beamforming vector $\boldsymbol{W}^{(n)}$, energy allocation $\boldsymbol{E}^{(n)}$, and time allocation $\boldsymbol{t}^{(n)}$.

IV. Numerical Results

Numerical results are provided to estimate the performance of the proposed scheme. The parameters are set as below, unless otherwise stated. It is assumed that the AP is situated at the edge of the network with a coordinate of (0, 5 m). The two users are randomly distributed in a 8 m×10 m coverage region. The bandwidth is set as $\mathrm{B} = 1$ MHz, the capacitance coefficient $\kappa_i = 10^{-28}$, the maximum CPU frequency $f_{\max} =$ 2 GHz, and the noise power $\sigma^2 = 10^{-9}$ W [35]. The number of CPU cycles required to compute one input-bit at user i are given as $C_i = 1,000, \{i = 1, 2\}$ [7], [11]. Without loss of generality, the channel reciprocity is assumed to hold for the downlink and uplink, i.e., $\boldsymbol{h}_i = \boldsymbol{g}_i, \{i = 1, 2\}$, and the channel coefficient is modeled as $\boldsymbol{h}_i = 10^{-1.5}\tilde{\boldsymbol{h}}_i d_i^{-\frac{\alpha}{2}}, i = \{1, 2\}$, where $\alpha = 3$ denotes the path loss exponent, and $\tilde{\boldsymbol{h}}$ follows the Rayleigh fading distribution. Furthermore, we set $L_1 = L_2 = L$, which indicates that two users have the same computation rate requirement. For the non-linear EH model, the parameters are set as $M_1 = M_2 = 24$ mW, $a_1 = a_2 = 150$ and $b_1 = b_2 = 0.024$.

For simplicity, the proposed scheme is referred to as "NOMA-UC MEC" in the following figures. To provide a comprehensive study, we also simulate the baseline schemes, which are described as follows:

- UC-MEC represents the wireless powered UC-enabled MEC scheme.
- For NOMA-MEC, it denotes the wireless powered MEC scheme, with NOMA protocol applied.
- MEC denotes the conventional MEC scheme based on TDMA protocol.
- With regards to the local computing scheme, the users execute the computation task by itself only, which corresponds to the condition of $l_1^{off} = 0$ and $l_2^{off} = 0$
- For offloading only scheme, the computation tasks are fully computed at the MEC server integrated with the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3056770, IEEE Transactions on Green Communications and Networking

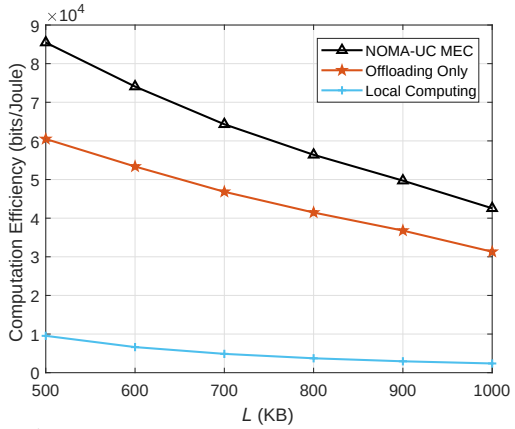

Fig. 3. Maximum computation efficiency vs. L.

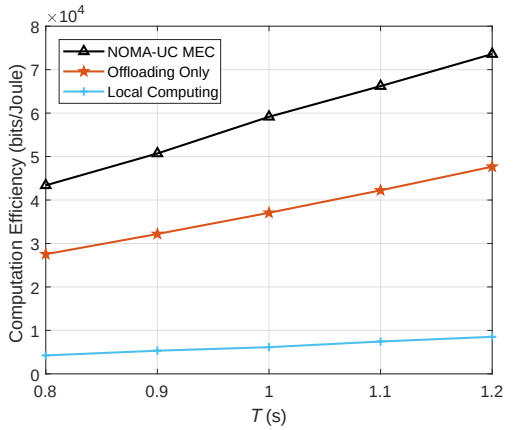

Fig. 4. Maximum computation efficiency vs. T.

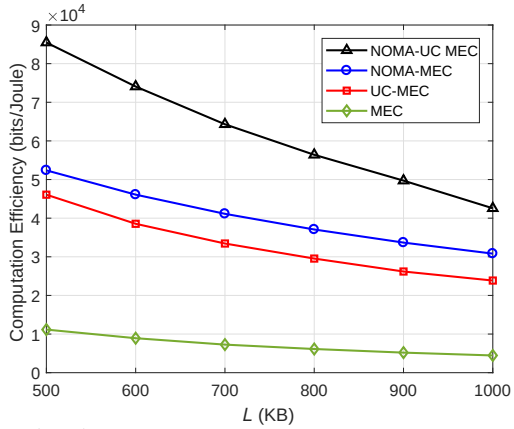

Fig. 5. Maximum computation efficiency vs. L.

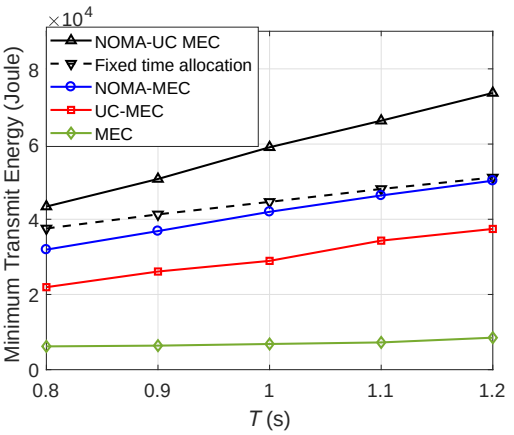

Fig. 6. Maximum computation efficiency vs. T.

AP.

- For the fixed time allocation scheme, the system model is the same as the proposed NOMA-UC MEC, except that the time allocations are fixed constants.

In Fig. 3, we present the relationship between the maximum computation efficiency and the computation data size requirement. To show the effectiveness of the partial offloading, the results of offloading only and local computing schemes are provided for comparison. As can be seen from Fig. 3, the computation efficiency decreases with larger required data bits for all three schemes, which implies that the growth rate of the required energy to compute runs faster than the data size. In addition, it is obvious that the proposed NOMA-UC MEC scheme is superior to the baseline schemes. Specifically, the local computing scheme yields the worst performance, indicating that the application of MEC greatly contributes to performance improvement.

Fig. 4 illustrates the influence of the block slot duration on the computation efficiency. It is noted that the computation efficiency increases monotonically with the block slot duration for all the schemes, and the proposed NOMA-UC MEC scheme outperforms the benchmark schemes. For example, when the time duration is 1s, the achievable computation efficiency for NOMA-UC MEC is about 6×10^4 bits/Joule, while for offloading only and local computing are 3.8×10^4 bits/Joule and 0.7×10^4 bits/Joule, respectively.

To show the effect of NOMA and UC application in MEC design, four schemes, namely the proposed NOMA-UC MEC, NOMA-MEC, UC-MEC, and MEC schemes are presented in Fig. 5. The computation efficiency performs decreasing trends with the increase of required data bits for all schemes, while the proposed NOMA-UC MEC scheme produces the best performance. In addition, compared with the MEC scheme, both NOMA-UC and UC-MEC achieve higher computation efficiency, proving the benefit of applying NOMA and UC in dealing with the doubly near-far effect in wireless powered MEC systems.

The comparison between the computation efficiency performance and the block slot duration T is presented in Fig. 6. We see that the trend for all the five curves is similar, and the proposed NOMA-UC MEC scheme produces the best performance. This indicates NOMA-UC MEC can enhance the system computation efficiency. Besides, the proposed 'NOMA-UC MEC' gains better system performance than the 'Fixed time allocation' scheme, which reveals the merit of the joint resource allocation optimization for system performance enhancement. Moreover, the performance of both NOMA-MEC and UC-MEC is superior to that of MEC, proving the advantage of applying NOMA and UC into the MEC design.

To evaluate the impact of users' locations, we assume that the AP and two users are placed in a line, i.e., $d_1 = 8$ m, $d_2 = \phi d_1$, and $d_{12} = (1-\phi)d_1$, where $\phi \in [0.5, 0.75]$. When the distance between the AP and U_2 becomes larger, the two users get closer. As can be observed from Fig. 7, the plain MEC scheme still yields the worst performance. Moreover,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3056770, IEEE Transactions on Green Communications and Networking

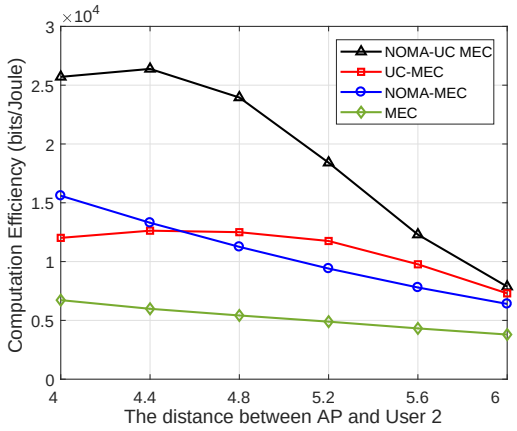

Fig. 7. Maximum computation efficiency vs. the distance between the AP and U_2.

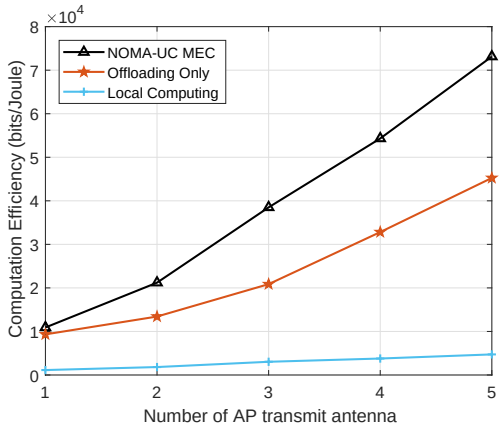

Fig. 8. Maximum computation efficiency vs. number of AP transmit antenna.

both NOMA-UC MEC and UC-MEC show a similar trend, where the computation efficiency first increases and then decreases when the distance between the AP and U_2 becomes larger. The reason is that, with UC applied to the two schemes, it can make a great impact with better channel gain between the two users. When the distance between the AP and U_2 becomes larger, U_2's channel gain degradation dominates the effect of UC, and thus the computation efficiency decreases. Moreover, NOMA-MEC outperforms UC-MEC when U_2 is closer to the AP, while UC-MEC gains better performance when the two users get closer. This is because the channel gain of U_2 decreases with the distance, resulting in a monotonically decreasing performance of NOMA-MEC. However, when the distance between the AP and U_2 becomes larger, two users get closer to each other, thus the channel degradation can be better compensated for the UC-MEC scheme.

Fig. 8 shows the relationship between the achieved computation efficiency and the number of transmit antennas N_t equipped at the AP with $d_1 = 8$ m, $d_2 = 4.8$ m, and $d_{12} = 3.2$ m. Due to the additional degrees of freedom introduced by the increasing number of transmit antennas, the computation efficiency can be improved for all the schemes. Particularly, compared with NOMA-UC MEC and offloading only scheme, the performance improvement of local computing is limited. The reason is that, the channel gain can be improved in both the downlink WPT transmission and uplink communications for NOMA-UC MEC and offloading only scheme, whereas only downlink WPT transmission is affected for local computing.

V. Conclusions

We investigated the application of NOMA and UC in a wireless powered MEC system under the non-linear energy harvesting model, in which the joint optimization problem of energy beamforming, time and power allocations was formulated to maximize the system computation efficiency. To solve the challenging nonconvex problem, SDR technique was first applied to transform the original problem into a more tractable expression. Then, the transformed problem was reformulated with variables substitutions, which can be finally solved by applying the SCA method. Numerical results demonstrated the superiority of applying NOMA and UC in wireless powered MEC design.

Based on the challenges and limitations of NOMA-assisted mobile edge computing with user cooperation, possible future extensions are listed as below. Firstly, perfect decoding at the AP is assumed for theoretical analysis. Practically, incorrect decoding may happen in NOMA scenarios due to imperfect SIC. Therefore, it would be interesting to investigate the impact of imperfect SIC in future studies. Secondly, it is also challenging to extend current work to other practical setups, i.e., imperfect CSI, users with multi-antenna, etc.

Appendix A
Proof of Lemma 1

Lemma 1 can be proved by contradiction approach. Suppose that $\{\boldsymbol{w}^*, \boldsymbol{t}^*, \boldsymbol{p}^*, \boldsymbol{l}^*\}$ is the optimal solution to (**P1**) corresponding to the maximum objective η^*, and the time allocation satisfies $t_0^* + t_1^* + t_{21}^* + t_{22}^* < T$. Based on the expression of (14a), with fixed t_0^*, η can be further improved as increasing $\{t_1 + t_{21} + t_{22}\}$ results in larger computation bits in the numerator, contradicting that the solution is optimal. Therefore, the maximum computation efficiency can be achieved with $t_0 + t_1 + t_{21} + t_{22} = T$.

Appendix B
Proof of Theorem 1

Assume that (**P2**) is feasible and it is also dual feasible. As can be observed from (**P2**), there are three linear constraints (16c, 16f and 35f) concerned with $\boldsymbol{W}^*$. According to [36, Theorem 3.2], we have that

$$\mathrm{rank}^2(\mathrm{W}^*) \leq 3. \tag{36}$$

If (**P2**) is feasible, we can infer that $\boldsymbol{W}^* > \mathbf{0}$, according to (16f) and (35f). Moreover, considering inequality constraint (36), we can further infer that $\mathrm{rank}(\boldsymbol{W}^*) = 1$. Hence, the relaxation is tight, and one can deduce that an optimal solution $\boldsymbol{W}^*$ always exists for problem (**P2**).

Furthermore, it is worth noting that (**P2**) is a convex optimization problem, hence the interior point method can be used to derive the global optimal solution $(\boldsymbol{W}^*, \boldsymbol{t}^*, \boldsymbol{E}^*)$. If $\mathrm{rank}(\boldsymbol{W}^*) = 1$, we can get that $\boldsymbol{W}^* = \frac{\boldsymbol{W}^*}{t_0}$, and the optimal energy beamforming vector $\boldsymbol{w}$ can be computed from $\boldsymbol{W}^*$ by applying eigen-decomposition. Otherwise, Gaussian randomization [36] can be used to attain a suboptimal solution.

APPENDIX C
PROOF OF THEOREM 2

To reveal that Algorithm 1 converges, we need to demonstrate that the sequence of the objective values obtained from Algorithm 1 is non-decreasing for each iteration, i.e., $\mu^{(n+1)} \geq \mu^{(n)}$.

Denote $\boldsymbol{W}^*, \boldsymbol{t}^*, \boldsymbol{E}^*, \mu^*, \beta^*$ as the optimal solution to (**P2**) during the n-th iteration. Note that during the problem transformation, constraints (19), (21c), (24), (30), and (33) are approximated with SCA. Take (24) as an example, denoting that $l = \sum_{i=1}^{2} l_i^{loc} + l_i^{off}$, we have that

$$l \geq \sqrt{\mu^{(n)}\beta^{(n)}} + 0.5\sqrt{\frac{\mu^{(n)}}{\beta^{(n)}}}(\beta^* - \beta^{(n)}) + 0.5\sqrt{\frac{\beta^{(n)}}{\mu^{(n)}}}(\mu^* - \mu^{(n)}). \tag{37}$$

The variables in iteration $n+1$ are updated accordingly, i.e., $\mu^{(n+1)} = \mu^{(*)}$, $\beta^{(n+1)} = \beta^{(*)}$, while (22b) can still be satisfied. By substituting the updated parameters into (24) during iteration $n+1$, we have

$$\begin{aligned}
&\sqrt{\mu^{(*)}\beta^{(*)}} + 0.5\sqrt{\frac{\mu^{(*)}}{\beta^{(*)}}}(\beta^* - \beta^{(*)}) + 0.5\sqrt{\frac{\beta^{(*)}}{\mu^{(*)}}}(\mu^* - \mu^{(*)}) \\
&= \sqrt{\mu^{(*)}\beta^{(*)}} && \text{(38a)} \\
&\leq \sqrt{\mu^{(n)}\beta^{(n)}} + 0.5\sqrt{\frac{\mu^{(n)}}{\beta^{(n)}}}(\beta^* - \beta^{(n)}) \\
&\quad + 0.5\sqrt{\frac{\beta^{(n)}}{\mu^{(n)}}}(\mu^* - \mu^{(n)}) && \text{(38b)} \\
&\leq l, && \text{(38c)}
\end{aligned}$$

where (38a) is derived by replacing (μ^{n+1}, β^{n+1}) with the obtained optimal solution (μ^*, β^*), (38b) is the upper-bounded approximation of (38a), and inequality (38c) is deduced with the aid of (37). Similar steps can be applied to prove the convergence of (19), (21c), (30) and (33), where the detailed process is omitted here.

In summary, it is proved that the solution derived during iteration n is a feasible point of the $n+1$-th iteration for (**P2**). Based on the above analysis, and considering that the objective of (**P2**) is a concave function, $\mu^{(n+1)} \geq \mu^{(n)}$ is proved. The proof is completed.

APPENDIX D
PROOF OF THEOREM 3

Moreover, denote χ^n as the optimal solutions to (**P2**) during the n-th iteration of Algorithm 1, due to the convergence feature of Algorithm 1 introduced by Theorem 2, $\chi^n \to \chi^*$ holds when $n \to \infty$, where χ^* denotes the optimal solution to (**P2**). In addition, we note that (**P2**) is obtained from (**P1**) by performing SCA, while the bound approximation introduced by the SCA produces the same function value and gradient value around the original spatial point during any iterations. In conclusion, the proposed Algorithm 1 can continuously coverage to a KKT point. The proof is completed.

REFERENCES

[1] S. Li *et al.*, "Energy-efficient resource allocation for industrial cyber-physical IoT systems in 5G era," *IEEE Trans. Ind. Inform.*, vol. 14, no. 6, pp. 2618–2628, June 2018.

[2] Y. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing:A key technology towards 5G," *White Paper, ETSI, Sophia Antipolis, France*, vol. 11, no. 11, pp. 1–16, 2015.

[3] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, May 2017.

[4] K. Huang and V. K. Lau, "Enabling wireless power transfer in cellular networks: Architecture, modeling and deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 902–912, Feb. 2014.

[5] 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.

[6] S. Bi and Y. J. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, June 2018.

[7] D. Wu, F. Wang, X. Cao, and J. Xu, "Wireless powered user cooperative computation in mobile edge computing systems," in *Proc. IEEE Globecom Workshops*, Dec, 2018, pp. 1–7.

[8] 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.

[9] Y. Lu, K. Xiong, P. Fan, Z. Zhong, and K. B. Letaief, "Robust transmit beamforming with artificial redundant signals for secure SWIPT system under non-linear EH model," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2218–2232, Apr. 2018.

[10] H. Ju and R. Zhang, "User cooperation in wireless powered communication networks," in *IEEE Global Commun. Conf.*, 2014, pp. 1430–1435.

[11] X. Hu, K. Wong, and K. Yang, "Wireless powered cooperation-assisted mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2375–2388, Apr. 2018.

[12] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, May 2018.

[13] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12 244–12 258, Dec. 2018.

[14] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

[15] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1875–1879, Dec. 2018.

[16] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[17] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in UAV-enabled wireless-powered mobile-edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1927–1941, Sep. 2018.

[18] 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.

[19] H. Sun, F. Zhou, and R. Q. Hu, "Joint offloading and computation energy efficiency maximization in a mobile edge computing system," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3052–3056, Jan. 2019.

[20] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation efficiency in a wireless-powered mobile edge computing network with NOMA," in *IEEE International Conference on Communications (ICC), Shanghai, May, 2019*, pp. 1–7.

[21] Y. Hao, Q. Ni, H. Li, and S. Hou, "Robust multi-objective optimization for EE-SE tradeoff in D2D communications underlaying heterogeneous networks," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4936–4949, Oct. 2018.

[22] B. Su, Q. Ni, and W. Yu, "Robust transmit beamforming for SWIPT-enabled cooperative NOMA with channel uncertainties," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4381–4392, Feb. 2019.

[23] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016.

[24] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[25] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.
[26] 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.
[27] J. Guo, Z. Song, Y. Cui, Z. Liu, and Y. Ji, "Energy-efficient resource allocation for multi-user mobile edge computing," in *Proc. IEEE Global Commun. Conf., Singapore*, Dec. 2017, pp. 1–7.
[28] 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, Aug. 2017.
[29] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4188–4200, June 2019.
[30] 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, Mar. 2017.
[31] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[32] H. Kosaki, "Arithmetic–geometric mean and related inequalities for operators," *J. Funct. Anal.*, vol. 156, no. 2, pp. 429–451, 1998.
[33] Y. Liu and Y. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 583–596, Nov. 2014.
[34] I. J. Lustig, R. E. Marsten, and D. F. Shanno, "Interior point methods for linear programming: Computational state of the art," *ORSA Journal on Computing*, vol. 6, no. 1, pp. 1–14, 1994.
[35] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE Trans. Nets.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2015.
[36] Y. Huang and D. P. Palomar, "Rank-constrained separable semidefinite programming with applications to optimal beamforming," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 664–678, Sep. 2010.