Asynchronous Protocol Designs for Energy Efficient Mobile Edge Computing Systems

Subin Eom, Hoon Lee, *Member, IEEE*, Junhee Park, and Inkyu Lee, *Fellow, IEEE*

***Abstract*—This paper investigates an asynchronous offloading protocol for mobile edge computing (MEC) systems where the computational tasks of mobile users are partially offloaded to a MEC server at a base station (BS). In the proposed asynchronous approach, upload, computation, and download of the tasks are performed in an asymmetric and non-orthogonal manner. A joint optimization problem of the transmit power, the offloading size, and the time-frequency resources is addressed to minimize the energy consumption of the overall system. By applying convex optimization techniques, the optimal solution for the asynchronous MEC offloading problem can be obtained. Numerical results demonstrate the effectiveness of the proposed asynchronous protocol over the conventional synchronous approach.**

***Index Terms*—Mobile edge computing, task offloading, energy efficient, resource allocation, convex optimization.**

I. Introduction

Recently, mobile edge computing (MEC) has been regarded as a promising solution for overcoming low computing power and short battery lifetime of internet-of-things (IoT) devices through cooperation with a computing server located at a network edge [1], [2]. The MEC protocol is carried out sequentially as the uplink, computing, and downlink stages. First, in the uplink stage, instead of processing computationally intensive applications directly at the device itself, full or part of their computational tasks is transferred to the MEC server. Then, the MEC server handles the task data in the computing stage, and the device receives the computation results in the downlink stage. Thus, the data computing load at the device is alleviated, which results in reduced energy consumption.

For the offloading of the tasks through wireless channels, the MEC systems have been investigated to identify proper communication protocols [3], [4]. To minimize the energy consumption, power and time allocation was studied for the two-user case with the task offloading stage in [3]. In [4], a synchronous protocol was adopted where users' data are simultaneously offloaded, processed, and downloaded. By jointly optimizing the transmission time allocation, offloading partition, and power control, the total energy consumption was minimized. However, as uplink/downlink and computation are orthogonally performed in the synchronous protocol, resource utilization may not be efficient.

Manuscript received July 22, 2020; revised October 7, 2020 and December 1, 2020; accepted December 8, 2020. Date of publication December 11, 2020; date of current version February 12, 2021. This work was supported in part by the National Research Foundation (NRF) through the Ministry of Science, ICT, and Future Planning (MSIP), Korea Government under Grant 2017R1A2B3012316 and in part by the NRF through the Ministry of Science, ICT (MSIT), Korea Government under Grant 2019R1F1A1060648. The review of this article was coordinated by Dr. K. Bian. *(Corresponding author: Inkyu Lee.)*

Subin Eom, Junhee Park, and Inkyu Lee are with the School of Electrical Engineering, Korea University, Seoul 02841, Korea (e-mail: esb777@korea.ac.kr; pjh0585@korea.ac.kr; inkyu@korea.ac.kr).

Hoon Lee is with the Department of Information and Communications Engineering, Pukyong National University, Busan 48513, Korea (e-mail: hlee@pknu.ac.kr).

Digital Object Identifier 10.1109/TVT.2020.3044073

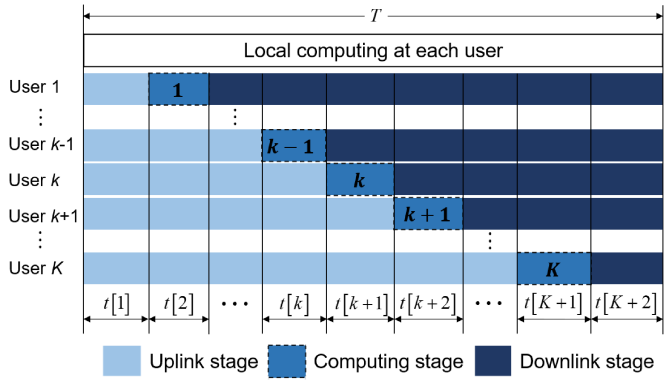

Fig. 1. Frame structure of an asynchronous MEC protocol.

In this paper, we propose a new asynchronous design based MEC offloading protocol where upload, computation, and download of the tasks are performed in an asymmetric and non-orthogonal manner. In detail, to fully exploit communication and computation resources, the offloaded task data from users are computed sequentially at the MEC server in a base station (BS), while the users who are not involved in the computing stage can upload or download the data. Considering the green communication for the MEC system [5], the total energy consumption minimization problem is addressed by jointly optimizing the transmit power, the bandwidth allocation, the time duration, and task offloading partition under the proposed asynchronous protocol. In order to deal with the problem of interests, the original non-convex problem is turned into an equivalent convex problem. Then, the optimal solution is obtained via convex optimization techniques, such as the Lagrange duality method and the ellipsoid method [6]. Numerical results demonstrate the performance improvement over the conventional synchronous protocol [4].

II. System Model and Problem Formulation

We consider a MEC offloading system where users offload their computation tasks to a MEC server installed in a BS. The task offloading is performed for a group of K users whose task are arrived within a similar time. User k ($k = 1, \ldots, K$) needs to finish the computation of the task $\mathcal{S}_k$ of length L_k bits within the maximum allowable latency T. Since the computing power of mobile devices is generally limited, a partial offloading strategy has been adopted [2], [4] where user k offloads $l_k \in [0, L_k]$ bits of its task $\mathcal{S}_k$ to the MEC server. Then, the remaining $L_k - l_k$ bits are locally computed at user k. To this end, after the BS receives the offloading task of length l_k bits through the uplink channels, the MEC server performs the computation of the offloaded bits. Then, the BS transmits the computation results to the users in the downlink.

In conventional synchronous protocols, these uplink, computing, and downlink stages are assumed to be perfectly synchronized for all users. However, since computation and uplink/downlink are performed separately at the MEC server, the BS can still receive and transmit the data during the computing state. Thus, the performance of the synchronous protocol in [4] can be further improved by allowing asynchronous transmission and reception. Motivated by this, we propose a new frame structure for the asynchronous MEC protocol in Fig. 1.

The details of the proposed asynchronous protocol are discussed in the following.

A. Proposed Asynchronous Offloading Protocol

Unlike the synchronous scheme where all users are aligned at the same stage [4], in the proposed asynchronous system, the duration of each stage can be different for each user. To model this feature, we divide the total system latency T into $K+2$ time slots as shown in Fig. 1. Denoting $t[n]$ as the duration of time slot n ($n = 1, \ldots, K+2$), the total latency constraint is given by

$$\sum_{n=1}^{K+2} t[n] \leq T. \tag{1}$$

Without loss of generality, it is assumed that the computation at the MEC server is sequentially carried out from user 1 to user K.[1] In the uplink stage, user k offloads l_k bits to the BS during time slots $n = 1, \ldots, k$. Then, at time slot $k+1$, the offloaded data from user k is computed at the MEC server. As communication resources can be utilized concurrently, the BS at time slot $k+1$ receives the offloaded data from users $k+1, \ldots, K$, while users $1, 2, \ldots, k-1$ download their computation results from the BS. After the computing stage at the BS, the computation result can be downloaded to user k during time slots $k+2, \ldots, K+2$.

1) Offloading Model: As the proposed asynchronous framework carefully assigns either uplink, computation, and downlink stages, multi-user interference can be successfully mitigated through the optimized time durations regardless of multiple access schemes. Due to the simplicity of the decoding strategy, frequency division duplexing and frequency division multiple access schemes are employed for the upload and download of the tasks. So, all the uplink and downlink transmissions of the users are carried out in orthogonal frequency resources with the bandwidth optimization. Let $\alpha_k[n]$ and $\beta_k[n]$ be the bandwidth ratio for user k at time slot n assigned to the uplink and the downlink, respectively. Then, we have

$$\sum_{k=n}^{K} \alpha_k[n] + \sum_{k=1}^{n-2} \beta_k[n] = 1, \forall n. \tag{2}$$

Let h_k be the channel gain between the BS and user, which is assumed to be constant during the latency T. Then, in the uplink stage, the number of the offloaded bits $I_{U,k}[n]$ from user k at time slot n ($n \leq k$) is written by

$$I_{U,k}[n] = t[n] W \alpha_k[n] \log_2 \left(1 + \frac{p_k[n]\gamma_k}{\alpha_k[n]}\right), \tag{3}$$

where W stands for the total bandwidth, $p_k[n]$ is the uplink transmit power of user k at time slot n, and $\gamma_k \triangleq h_k/(W\sigma^2)$ represents the effective signal-to-noise ratio (SNR) with σ^2 being the noise variance. It is noted that $p_k[n] = 0$ for $n > k$ from the asynchronous structure in Fig. 1. To upload total l_k bits to the MEC server during the uplink stage, each user needs to satisfy

$$\sum_{n=1}^{k} I_{U,k}[n] \geq l_k, \forall k. \tag{4}$$

After user k completes offloading the task data, the MEC server executes the computation of l_k bits at time slot $k+1$. Let C_k (cycle/bit) be the number of the central processing unit (CPU) cycles for processing the input task $\mathcal{S}_k$ of user k. Then, the CPU cycle frequency $f_{\mathrm{S},k}$ (cycle/sec) required for calculating l_k bits at the MEC server is expressed as $f_{\mathrm{S},k} = C_k l_k / t[k+1]$. Due to practical limitations of the CPU, its cycle frequency $f_{\mathrm{S},k}$ should be designed not to exceed its maximum bound $f_{\mathrm{S,max}}$, which can be written by

$$C_k l_k \leq t[k+1] f_{\mathrm{S,max}}, \forall k. \tag{5}$$

The energy consumption $E_{\mathrm{S},k}$ for the computation of l_k bits at the MEC server can be characterized by the dynamic voltage and frequency scaling (DVFS) model as $E_{\mathrm{S},k} = \kappa_{\mathrm{S}}(C_k l_k)^3/t^2[k+1]$ [1], [2], [4], where κ_{S} is the effective capacitance coefficient at the MEC server. After processing at the MEC server, the amount of the computation results is assumed to equal $\eta_k l_k$, where η_k indicates the reduction ratio of the data size after computation which depends on the task type of user k.

In the downlink stage, the computation result is forwarded to user k through the downlink channels during time slots $k+2, k+3, \ldots, K+2$. Similar to (3), the number of bits $I_{D,k}[n]$ downloaded from the BS to user k at time slot n ($n \geq k+2$) can be expressed as

$$I_{D,k}[n] = t[n] W \beta_k[n] \log_2 \left(1 + \frac{q_k[n]\gamma_k}{\beta_k[n]}\right), \tag{6}$$

where $q_k[n]$ accounts for the downlink power of the BS for user k at time slot n, and is 0 for $n < k+2$. Since the length of the computation results is $\eta_k l_k$, we have the constraint on the downlink transmission of user k as

$$\sum_{n=k+2}^{K+2} I_{D,k}[n] \geq \eta_k l_k, \forall k. \tag{7}$$

2) Local Computing Model: While the offloaded l_k data bits are being processed at the MEC server, each user needs to locally address the computation of the remaining $L_k - l_k$ within the total allowed latency T. Similar to (5), the CPU cycle frequency $f_{\mathrm{L},k} = C_k(L_k - l_k)/T$ of the local computation at user k has the maximum frequency constraint as

$$C_k(L_k - l_k) \leq T f_{\mathrm{L,max}}, \forall k, \tag{8}$$

where $f_{\mathrm{L,max}}$ denotes the maximum frequency at each user. The energy consumption of user k for the local computation is given by $E_{\mathrm{L},k} = \kappa_{\mathrm{L}}(C_k(L_k - l_k))^3/T^2$, where κ_{L} represents the effective capacitance coefficient at the user.

B. Problem Formulation

We aim to minimize the weighted sum energy consumption by jointly optimizing the transmit power $\mathbf{P} \triangleq \{p_k[n], q_k[n], \forall k, n\}$, the bandwidth allocation $\mathbf{W} \triangleq \{\alpha_k[n], \beta_k[n], \forall k, n\}$, the time duration $\mathbf{T} \triangleq \{t[n], \forall n\}$, and the offloading partition $\mathbf{O} \triangleq \{l_k, \forall k\}$. The overall energy consumptions ζ_{user} and ζ_{BS} at the users and the BS, which include both the computation and communication processes, are respectively obtained as

$$\zeta_{\mathrm{user}} = \sum_{k=1}^{K} E_{\mathrm{L},k} + \sum_{k=1}^{K}\sum_{n=1}^{k} t[n] p_k[n], \tag{9}$$

$$\zeta_{\mathrm{BS}} = \sum_{k=1}^{K} E_{\mathrm{S},k} + \sum_{k=1}^{K}\sum_{n=k+2}^{K+2} t[n] q_k[n]. \tag{10}$$

[1]The impact of the user ordering on the offloading performance will be examined in Section IV.

Denoting P_U and P_D as the peak power budget at the user and the BS, respectively, the problem is formulated as

$$(P1): \min_{\mathbf{P},\mathbf{W},\mathbf{T},\mathbf{O}} w_1\zeta_{\text{user}} + w_2\zeta_{\text{BS}} \tag{11a}$$

$$\text{s.t. } l_{k,\min} \le l_k \le L_k, \forall k, \tag{11b}$$

$$p_k[n] \le P_U, \forall k, n, \ \sum_{k=1}^{K} q_k[n] \le P_D, \forall n, \tag{11c}$$

$$(1), (2), (4), (5), (7),$$

where w_1 and w_2 indicate non-negative weights for total energy consumption, and $l_{k,\min} = \max(0, L_k - Tf_{\text{L,max}}/C_k)$ from (8). For objective (11a) and constraints (4) and (7), since the multiplication of affine and concave function is neither convex nor concave, (P1) is generally a non-convex problem and thus, it is difficult to handle the problem (P1). In the following section, we investigate the globally optimal approach to solve (P1).

III. Proposed Algorithm

We propose an optimal algorithm for solving the non-convex formulation (P1). To this end, we first recast (P1) as an equivalent convex problem by employing the change of the variables as

$$E_{U,k}[n] = t[n]p_k[n], \ E_{D,k}[n] = t[n]q_k[n], \tag{12}$$

$$A_k[n] = t[n]\alpha_k[n], \ B_k[n] = t[n]\beta_k[n]. \tag{13}$$

From (12), the energy consumptions in (9) and (10) are turned into

$$\xi_{\text{user}} = \sum_{k=1}^{K} E_{\text{L},k} + \sum_{k=1}^{K}\sum_{n=1}^{k} E_{U,k}[n], \tag{14}$$

$$\xi_{\text{BS}} = \sum_{k=1}^{K} E_{\text{S},k} + \sum_{k=1}^{K}\sum_{n=k+2}^{K+2} E_{D,k}[n]. \tag{15}$$

In addition, by applying (13), we can rewrite $I_{U,k}[n]$ in (3) and $I_{D,k}[n]$ in (6) as

$$I_{U,k}[n] = WA_k[n]\log_2\left(1 + \frac{E_{U,k}[n]\gamma_k}{A_k[n]}\right), \tag{16}$$

$$I_{D,k}[n] = WB_k[n]\log_2\left(1 + \frac{E_{D,k}[n]\gamma_k}{B_k[n]}\right). \tag{17}$$

Then, the equivalent reformulation of (P1) is given as

$$(P1.1): \min_{\mathbf{E}\succeq\mathbf{0},\mathbf{A}\succeq\mathbf{0},\mathbf{T}\succeq\mathbf{0},\mathbf{O}} w_1\xi_{\text{user}} + w_2\xi_{\text{BS}} \tag{18a}$$

$$\text{s.t.} \quad E_{U,k}[n] \le t[n]P_U, \forall k, n, \tag{18b}$$

$$\sum_{k=1}^{K} E_{D,k}[n] \le t[n]P_D, \forall n, \tag{18c}$$

$$\sum_{k=n}^{K} A_k[n] + \sum_{k=1}^{n-2} B_k[n] \le t[n], \forall n, \tag{18d}$$

$$(1), (4), (5), (7), (11b),$$

where $\mathbf{E} \triangleq \{E_{U,k}[n], E_{D,k}[n], \forall k, n\}$ and $\mathbf{A} \triangleq \{A_k[n], B_k[n], \forall k, n\}$. Problem (P1.1) now becomes convex since (14) and (15) are convex and (16) and (17) are jointly concave functions of $\{E_{U,k}[n], A_k[n]\}$ and $\{E_{D,k}[n], B_k[n]\}$, respectively. Also, due to the fact that the Slater's condition is satisfied, the strong duality holds for (P1.1). Thus, the globally optimal solution can be attained by the Lagrange duality method.

The Lagrangian of (P1.1) is expressed as

$$\mathcal{L}(\mathbf{E}, \mathbf{A}, \mathbf{T}, \mathbf{O}, \boldsymbol{\lambda}, \mu, \nu) = \sum_{k=1}^{K}\sum_{n=1}^{k} F_{U,k}[n] + \sum_{k=1}^{K} F_{C,k} + \sum_{k=1}^{K}\sum_{n=k+2}^{K+2} F_{D,k}[n] + \rho_1 t[1] + \rho_{K+2} t[K+2] - \mu_T T, \tag{19}$$

where

$$F_{U,k}[n] \triangleq (w_1 + \lambda_{U,k}[n])E_{U,k}[n] - \nu_{U,k}I_{U,k}[n] + \mu_A[n]A_k[n],$$

$$F_{C,k} \triangleq w_1 E_{\text{L},k} + w_2 E_{\text{S},k} + Y_k t[k+1] + Z_k l_k,$$

$$Y_k \triangleq \begin{cases} J_k - P_U\left(\sum_{m=k+1}^{K}\lambda_{U,m}[k+1]\right), & \text{if } k = 1 \\ J_k - P_U\left(\sum_{m=k+1}^{K}\lambda_{U,m}[k+1]\right) - P_D\lambda_D[k+1], & \\ \qquad\qquad \text{if } k = 2, \ldots, K-1 & \\ J_k - P_D\lambda_D[k+1], & \text{if } k = K, \end{cases}$$

$$J_k \triangleq -\mu_A[k+1] + \mu_T - \frac{\mu_{O,k} f_{S,\max}}{C_k},$$

$$Z_k \triangleq \mu_{O,k} + \nu_{U,k} + \eta_k \nu_{D,k},$$

$$F_{D,k}[n] \triangleq (w_2 + \lambda_D[n])E_{D,k}[n] - \nu_{D,k}I_{D,k}[n] + \mu_A[n]B_k[n],$$

$$\rho_1 \triangleq P_U\sum_{k=1}^{K}\lambda_{U,k}[1] - \mu_A[1] + \mu_T,$$

$$\rho_{K+2} \triangleq -P_D\lambda_D[K+2] - \mu_A[K+2] + \mu_T.$$

In (19), $\boldsymbol{\lambda} \triangleq \{\lambda_{U,k}[n], \forall k, \forall n \le k, \ \lambda_D[n], \ 3 \le n \le K+2\}$, $\boldsymbol{\mu} \triangleq \{\mu_A[n], \forall n, \mu_T, \mu_{O,k}, \forall k\}$, and $\boldsymbol{\nu} \triangleq \{\nu_{U,k}, \nu_{D,k}, \forall k\}$ are the non-negative Lagrange multipliers corresponding to the constraints in (18b)-(18 d), (1), (5), (4), and (7), respectively.

Then, the dual function $g(\boldsymbol{\lambda}, \mu, \nu)$ can be written by

$$g(\boldsymbol{\lambda}, \mu, \nu) = \min_{\mathbf{E}\succeq 0, \mathbf{A}\succeq 0, \mathbf{T}\succeq 0, \mathbf{O}} \mathcal{L}(\mathbf{E}, \mathbf{A}, \mathbf{T}, \mathbf{O}, \boldsymbol{\lambda}, \mu, \nu) \tag{20}$$

$$\text{s.t.} \qquad (11b).$$

From (19), we have $\rho_1 \ge 0$, $\rho_{K+2} \ge 0$, and $Y_k \ge 0, \forall k$ to ensure $g(\boldsymbol{\lambda}, \mu, \nu) > -\infty$. It is revealed from (19) that the Lagrangian is given by a linear combination of four different terms $F_{U,k}[n]$, $F_{C,k}$, $F_{D,k}[n]$, and $\rho_n t[n]$, which can be individually addressed with given dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$.

Based on this observation, by fixing the dual variables, we decouple (20) into $K^2 + 2K + 2$ subproblems as

$$(SP-1): \min_{l_{k,\min} \le l_k \le L_k, t[k+1] \ge 0} F_{C,k}, \quad \forall k, \tag{21}$$

$$(SP-2): \min_{t[n] \ge 0} \rho_n t[n], \text{ for } n = 1 \text{ and } K+2, \tag{22}$$

$$(SP-3): \min_{E_{U,k}[n] \ge 0, A_k[n] \ge 0} F_{U,k}[n], \quad \forall k, n, \tag{23}$$

$$(SP-4): \min_{E_{D,k}[n] \ge 0, B_k[n] \ge 0} F_{D,k}[n], \quad \forall k, n. \tag{24}$$

Thanks to the convexity, the optimal solutions for these subproblems are readily obtained by examining the Karush-Kuhn-Tucker (KKT)

conditions. First, from the zero gradient condition, the optimal l_k^* and $t^*[k+1]$ for $(SP-1)$ are computed as

$$l_k^* = \left[L_k - T\sqrt{\frac{3C_k(\omega_2\kappa_{\rm S}Y_k^2/4)^{1/3}+Z_k}{3w_1\kappa_{\rm L}C_k^3}} \right]_{l_{k,\min}}^{L_k}, \quad (25)$$

$$t^*[k+1] = l_k^* \left(\frac{Y_k}{2w_2\kappa_{\rm S}C_k^3} \right)^{-1/3}, \quad (26)$$

where $[x]_y^z \triangleq \max\{y, \min\{x,z\}\}$ means that x has the lower and the upper limits as y and z, respectively.

For the linear program (LP) $(SP-2)$, the optimal solution $t^*[n]$ for $n=1$ and $K+2$ can be obtained as

$$t^*[n] = \begin{cases} \tau[n], & \text{if } \rho_{\rm n}=0, \\ 0, & \text{if } \rho_{\rm n}>0, \end{cases} \quad (27)$$

where $\tau[n]$ can be an arbitrary non-negative real numbers. Substituting $t^*[n]$ into the objective value of (SP-2) always results in zero for any given $\tau[n]$ and ρ_n. This means that the choice of the optimal solution $t^*[n]$ does not affect the dual function. For simplicity, we set $t^*[n]=0$ for $n=1$ and $K+2$ to obtain the dual variables.

Next, the optimal solutions $\{E_{U,k}^*[n], A_k^*[n]\}$ and $\{E_{D,k}^*[n], B_k^*[n]\}$ for $(SP-3)$ and $(SP-4)$ are respectively obtained as

$$E_{U,k}^*[n] = A_k^*[n]Q_{U,k}[n], \quad (28)$$

$$E_{D,k}^*[n] = B_k^*[n]Q_{D,k}[n], \quad (29)$$

where

$$Q_{U,k}[n] \triangleq \left[\frac{\nu_{U,k}W}{\ln 2(\mathrm{w}_1+\lambda_{\mathrm{U,k}}[\mathrm{n}])} - \frac{1}{\gamma_k} \right]^+,$$

$$Q_{D,k}[n] \triangleq \left[\frac{\nu_{D,k}W}{\ln 2(\mathrm{w}_2+\lambda_{\mathrm{D}}[\mathrm{n}])} - \frac{1}{\gamma_k} \right]^+,$$

with $[x]^+ \triangleq \max\{x,0\}$. Similar to the $(SP-2)$ case, $A_k^*[n]$ and $B_k^*[n]$ have no impact on the dual function, and thus are set to zero.

By using the analytical solutions (25)-(29), the dual function $g(\boldsymbol{\lambda},\mu,\nu)$ is readily evaluated for arbitrary given dual variables. The dual problem of (P1.1) is formulated as

$$\max_{\boldsymbol{\lambda}\succeq 0,\boldsymbol{\mu}\succeq 0,\boldsymbol{\nu}\succeq 0} \quad g(\boldsymbol{\lambda},\mu,\nu) \quad (30a)$$

$$\text{s.t.} \quad \rho_1\geq 0,\ \rho_{K+2}\geq 0,\ Y_k\geq 0, \forall k, \quad (30b)$$

$$V_{U,k}[n]\geq 0, \forall k,n, \quad (30c)$$

$$V_{D,k}[n]\geq 0, \forall k,n, \quad (30d)$$

where (30b) is the bound condition from (20), and (30c) and (30 d) are the KKT condition from $(SP-3)$ and $(SP-4)$, respectively, with

$$V_{m,k}[n] \triangleq -\nu_{m,k}W\log_2(1+Q_{m,k}[n]\gamma_k) + \frac{\nu_{m,k}W\gamma_k Q_{m,k}[n]}{\ln 2(1+Q_{m,k}[n]\gamma_k)} + \mu_A[n], \text{ for } m\in\{U,D\}.$$

The dual problem can be optimally addressed via the subgradient method, e.g., the ellipsoid method [7]. Let Δx be the subgradient the objective in (30) with respect to x. Then, we have

$$\Delta\lambda_{U,k}[n] = E_{U,k}[n]-t[n]P_U,\ \forall k, \forall n\leq k,$$

$$\Delta\lambda_D[n] = \sum_{k=1}^{n-2} E_{D,k}[n]-t[n]P_D,\ \text{for } 3\leq \mathrm{n}\leq \mathrm{K}+2,$$

Algorithm 1: Algorithm for Solving (P1).

Initialize $\{\boldsymbol{\lambda},\mu,\nu\}$.
Repeat
Obtain $\{\mathbf{E}^*,\mathbf{A}^*,\mathbf{T}^*,\mathbf{O}^*\}$ from (25)-(29).
Update $\{\boldsymbol{\lambda},\mu,\nu\}$ via the constrained ellipsoid method.
Until Convergence.
Obtain $\{\mathbf{A}^*,t^*[1],t^*[K+2]\}$ by solving problem (31) with given $\{\boldsymbol{\lambda}^*,\mu^*,\nu^*\}$.
Obtain $p_k^*[n]=E_{U,k}^*[n]/t^*[n]$, $q_k^*[n]=E_{D,k}^*[n]/t^*[n]$, $\alpha_k^*[n]=A_k^*[n]/t^*[n]$, and $\beta_k^*[n]=B_k^*[n]/t^*[n]$, $\forall k,n$.

$$\Delta\mu_A[n] = \sum_{k=n}^{K} A_k[n] + \sum_{k=1}^{n-2} B_k[n] - t[n],\ \forall n,$$

$$\Delta\mu_T = \sum_{n=1}^{K+2} t[n]-T,\ \ \Delta\mu_{O,k} = l_k - \frac{t[k+1]f_{\rm S,max}}{C_k},\ \forall k,$$

$$\Delta\nu_{U,k} = l_k - \sum_{n=1}^{k} I_{U,k}[n],\ \forall k,$$

$$\Delta\nu_{D,k} = \eta_k l_k - \sum_{n=k+2}^{K+2} I_{D,k}[n],\ \forall k.$$

With the optimal dual variable $\{\boldsymbol{\lambda}^*,\mu^*,\nu^*\}$ at hands, the optimal primal solutions l_k^* and $t^*[k+1]$ can be obtained as (25) and (26), respectively. However, the optimal primal solutions $\mathbf{E}^*$, $\mathbf{A}^*$, $t^*[1]$, and $t^*[K+2]$ should be redefined when $\rho_1=0$, $\rho_{K+2}=0$, $V_{U,k}[n]=0$, or $V_{D,k}[n]=0$ based on the KKT condition. To this end, the corresponding problem is constructed by substituting $E_{U,k}^*[n]$ and $E_{D,k}^*[n]$ in (28) and (29) into (P1.1) as

$$\min_{\mathbf{A}\succeq 0,t[1]\geq 0,t[K+2]\geq 0} \quad w_1\sum_{k=1}^{K}\sum_{n=1}^{k} Q_{U,k}[n]A_k[n] + w_2\sum_{k=1}^{K}\sum_{n=k+2}^{K+2} Q_{D,k}[n]B_k[n] \quad (31a)$$

$$\text{s.t. } Q_{U,k}[n]A_k[n]\leq t[n]P_U, \forall k,n, \quad (31b)$$

$$\sum_{k=1}^{n-2} Q_{D,k}[n]B_k[n]\leq t[n]P_D, \forall n, \quad (31c)$$

$$\sum_{n=1}^{k} W\log_2(1+Q_{U,k}[n]\gamma_k)A_k[n]\geq l_k, \forall k, \quad (31d)$$

$$\sum_{n=k+2}^{K+2} W\log_2(1+Q_{D,k}[n]\gamma_k)B_k[n]\geq \eta_k l_k, \forall k, \quad (31e)$$

$$(1), (18b).$$

Since (31) is a convex LP, the optimal solution $\{\mathbf{A}^*,t^*[1],t^*[K+2]\}$ can be easily attained by the interior-point method [6]. One optimal procedure for solving (P1) is summarized in Algorithm 1.

The computations of Algorithm 1 is dominated by the ellipsoid method that iteratively updates $N_{dual}=0.5K^2+5.5\mathrm{K}+3$ dual variables, requiring the complexity $O(N_{dual}^2)=O(K^4)$ at each iteration. Since the ellipsoid takes $O(N_{dual}^2)$ steps for the convergence, the complexity of Algorithm 1 is given by $O(K^8)$ [7]. Notice that the number of the primal variables $N_{primal}=2K^2+6K+2$ of (P1.1) is four times larger than N_{dual}. Therefore, the proposed algorithm, which

TABLE I
ENERGY CONSUMPTION (JOULES) OF VARIOUS ORDERING SCHEMES WITH $T = 50$ MSEC AND $L_k = 200$ KBITS

K	Optimal	Proposed	Descend	Ascend	Random
4	0.5025	0.5038	0.5179	0.6519	0.5740
5	0.8272	0.8295	0.8588	1.1021	0.9699
6	1.2401	1.2436	1.2924	1.6486	1.4605

alternatively addresses the dual formulation, has lower computational complexity compared to the approach tackling (P1.1) directly. The overall procedure of the protocol is carried out as follows. A BS receives information from users and calculates the optimal resource allocation based on Algorithm 1. After the BS feeds it back to the users, the MEC system can be operated as in Fig. 1.

IV. NUMERICAL RESULTS

We present numerical results demonstrating the performance of the asynchronous protocol in the MEC systems with the total bandwidth $W = 10$ MHz and $\sigma^2 = -174$ dBm/Hz. The uplink and downlink power constraints are set to $P_U = 35$ dBm and $P_D = 40$ dBm, respectively. The channel gain h_k is generated as $h_k = Gd_k^{-\theta}$ [2] where $G = -60$ dB represents the reference pathloss at 1 m, d_k stands for the distance from the BS to user k who is uniformly distributed over $[50\text{ m}, 200\text{ m}]$, and $\theta = 3.5$ is the path loss exponent. The maximum CPU cycles and the effective capacitance coefficients are fixed as $f_{\mathrm{L,max}} = 2 \times 10^9$, $f_{\mathrm{S,max}} = 5 \times 10^{10}$, $\kappa_{\mathrm{L}} = 10^{-27}$, and $\kappa_{\mathrm{S}} = 10^{-29}$ [4]. In addition, we apply $\eta_k = 0.5$, $\forall k$, $C_k = 10^3$, $\forall k$ [2], $w_1 = 1$, and $w_2 = 0.2$ [4]. As a benchmark scheme, we consider the conventional synchronous MEC offloading protocol [4], which can be formulated from the proposed asynchronous protocol where all offloaded data are processed simultaneously at the MEC server with the CPU cycle $f_{\mathrm{S}} = \sum_{k=1}^{K} C_k l_k / \sum_{n=2}^{K+1} t[n]$ and the communication resources are not utilized ($p_k[n] = q_k[n] = 0, \forall k$) during the time slots $2, \ldots, K+1$.

Table I investigates the impact of the user ordering on the energy consumption performance of the proposed asynchronous method. We examine the following sorting strategies.

- *Optimal:* The best order minimizing the energy consumption is numerically identified using exhaustive search among $K!$ candidates.
- *Proposed:* The users are sorted such that $h_1 \geq h_K \geq h_2 \geq h_3 \geq \ldots \geq h_{K-1}$.
- *Descend/ascend:* The users are sorted in the descending/ascending order in terms of the channel gains $\{h_k\}$.
- *Random:* No ordering is considered and the task offloading is performed with an arbitrary priority.

From the table, we can see that the proposed ordering achieves near-optimal performance and is superior to other methods. Let us explain the insight behind the proposed ordering. In the proposed asynchronous scheme, users $k = 2, \ldots, K-1$ are assigned at least two time slots for the upload and download task. In contrast, users 1 and K only have a single time slot $t[1]$ and $t[K+2]$ for upload and download, respectively. Therefore, their transmission must be completed during one time slot. Thus, when users with favorable channels are allocated as the first and the last users, the transmission time $t[1]$ and $t[K+2]$ can be decrease. As the time for computing $t[2], \ldots, t[K+1]$ increases, the computing energy consumed at the server can be reduced. As a result, the performance based on the proposed ordering is almost the same as that of the optimal scheme with much reduced complexity.

Fig. 2 shows the energy consumption performance with respect to the number of users K for different latency T. For all K cases, the

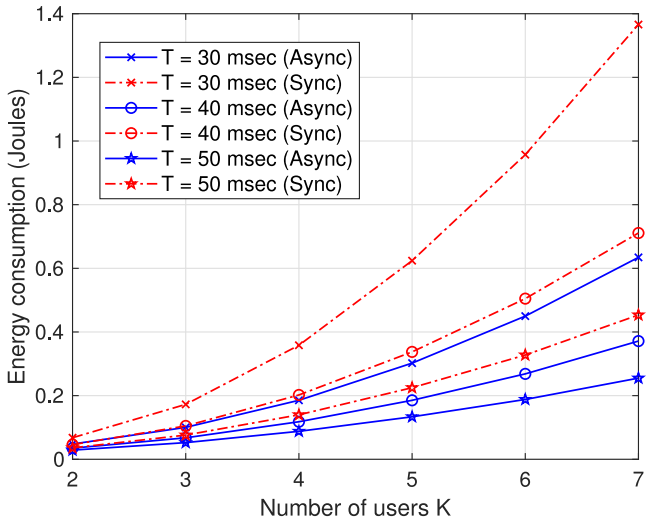

Fig. 2. Energy consumption with respect to the number of users K with $L_k = 100$ kbits.

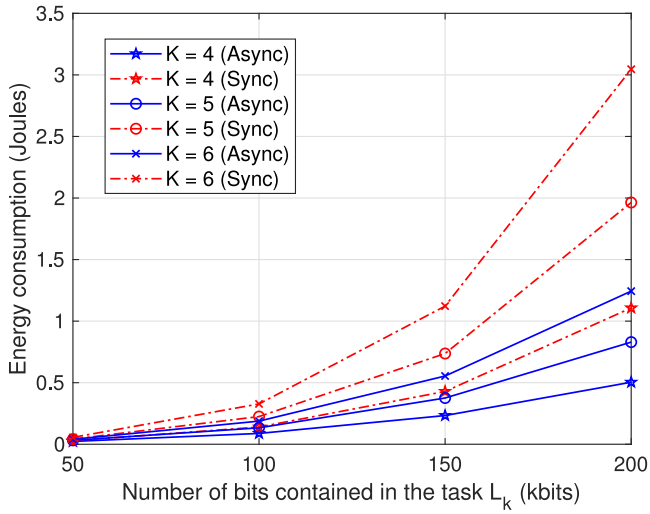

Fig. 3. Energy consumption with respect to the number of bits contained in the task L_k with $T = 50$ msec.

proposed asynchronous scheme performs better than the conventional synchronous scheme. In the synchronous approach, the task data from all K users are computed simultaneously at the MEC server, and upload and download are executed only when the MEC server is not computing, which results in inefficient usage of energy. On the other hand, since the limited wireless resources are carefully optimized in an asynchronous manner, the proposed scheme can conserve energy. For similar reasons, the proposed method presents less energy consumption regardless of T. For example, for $K = 7$ and $T = 30$ msec, we observe that the proposed method saves energy consumption by 54% over the synchronous method.

The energy consumption with different task size L_k is presented in Fig. 3. For all L_k, the proposed asynchronous method can significantly reduce the energy consumption compared to the conventional synchronous scheme. In the proposed method, as the transmission is carried out in an asynchronous way, relatively more time can be allocated to the uplink and downlink stages compared to the synchronous scheme. It increases the degree of freedom for task offloading partitions and the tasks can be offloaded in a more energy-efficient way. This can be seen that the asynchronous method can handle the task of a larger size under

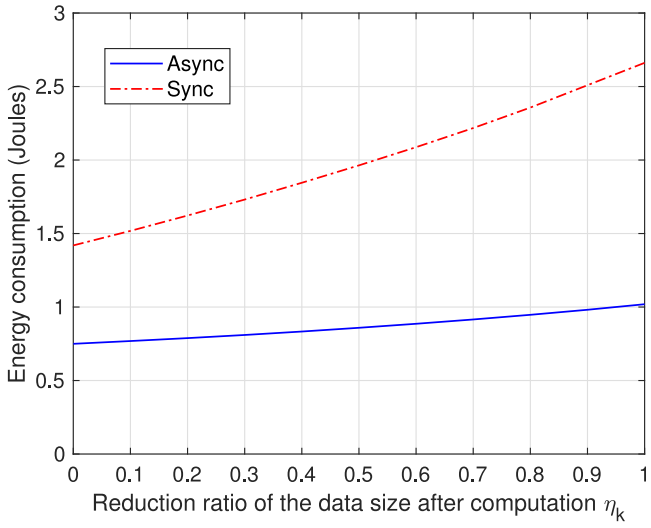

Fig. 4. Energy consumption with respect to the reduction ratio of the data size after computation η_k.

the same energy allowance. In a system of $K = 6$ users with 1 Joules energy capacity, the asynchronous method can process approximately 40 kbits more than the synchronous method. Fig. 4 plots the energy consumption performance for the reduction ratio of the data size after computation η_k with $K = 5$, $T = 50$ msec, and $L_k = 200$ kbits. Regardless of η_k, the proposed asynchronous method outperforms the conventional synchronous method. Both in the asynchronous and synchronous protocols, the energy consumption monotonically increases as η_k grows, since it requires more energy for the download of the computation results. In addition, as the time for downloading grows, the time allocated to uploading and server computing is reduced, which results in the increased energy consumption. From the numerical results, we can conclude that the proposed asynchronous scheme enables energy-efficient MEC offloading compared to the synchronous scheme through efficient resource utilization.

V. Conclusion

We have proposed the asynchronous MEC protocol where the upload, computation, and download of the tasks are performed in an asynchronous manner. The total energy consumption minimization problem has been addressed by optimizing the transmit power, offloading size, and time-frequency allocation. With convex optimization techniques, the optimal solution for the problem can be achieved. Numerical results have demonstrated the effectiveness of the asynchronous operations in the MEC systems. As future works, research on asynchronous MEC protocol with parallel computing and latency minimization problem could be an important research direction. In addition, the proposed MEC protocol can be applied to unmanned aerial vehicle aided systems [8], relay network [9], wireless powered communication network [10], or learning framework [11].

References

[1] 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, pp. 2322–2358, 2017.

[2] 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, pp. 4188–4200, Jun. 2019.

[3] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA-MEC offloading," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 6207–6211, Jun. 2019.

[4] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Commun. Lett.*, vol. 23, pp. 310–313, Feb. 2019.

[5] M. M. Mowla, I. Ahmad, D. Habibi, and Q. V. Phung, "A green communication model for 5G systems," *IEEE Trans. Green Commun. Netw.*, vol. 1, pp. 264–280, Sep. 2017.

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

[7] S. Boyd, "Ellipsoid method," Stanford Univ., Stanford, CA, USA. 2017. [Online]. Available: https://stanford.edu/class/ee364b/ lectures/ ellipsoid _ method_notes.pdf

[8] S. Eom, H. Lee, J. Park, and I. Lee, "UAV-aided wireless communication designs with propulsion energy limitations," *IEEE Trans. Veh. Technol.*, vol. 69, pp. 651–662, Jan. 2020.

[9] D. Darsena, G. Gelli, and F. Verde, "Design and performance analysis of multiple-relay cooperative MIMO networks," *J. Commun. Netw*, vol. 21, pp. 25–32, Feb. 2019.

[10] H. Lee, K.-J. Lee, H.-B. Kong, and I. Lee, "Sum-rate maximization for multiuser MIMO wireless powered communication networks," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 9420–9424, Nov. 2016.

[11] J. Moon, O. Simeone, S.-H. Park, and I. Lee, "Online reinforcement learning of X-Haul content delivery mode in fog radio access networks," *IEEE Signal Process. Lett.*, vol. 26, pp. 1451–1455, Oct. 2019.