

Multiuser Resource Allocation for Mobile-Edge Computation Offloading

Changsheng You and Kaibin Huang

Department of Electrical and Electronic Engineering

The University of Hong Kong

Email: csyoun@eee.hku.hk, huangkb@eee.hku.hk

Abstract—*Mobile-edge computation offloading (MECO) offloads intensive mobile computation to clouds located at the edges of cellular networks. Thereby, MECO is envisioned as a promising technique for prolonging the battery lives and enhancing the computation capacities of mobiles. In this paper, we consider resource allocation in a MECO system comprising multiple users that time share a single edge cloud and have different computation loads. The optimal resource allocation is formulated as a convex optimization problem for minimizing the weighted sum mobile energy consumption under constraint on computation latency and for both the cases of infinite and finite edge cloud computation capacities. The optimal policy is proved to have a threshold-based structure with respect to a derived offloading priority function, which yields priorities for users according to their channel gains and local computing energy consumption. As a result, users with priorities above and below a given threshold perform complete and minimum offloading, respectively. Computing the threshold requires iterative computation. To reduce the complexity, a sub-optimal resource-allocation algorithm is proposed and shown by simulation to have close-to-optimal performance.*

I. INTRODUCTION

The realization of Internet of Things (IoT) will connect tens of billions of resource-limited mobiles, e.g., mobile devices, sensors and wearable computing devices, to Internet via cellular networks. The finite battery lives and limited computation capacities of mobiles pose challenges for designing IoT. One promising solution is to leverage mobile-edge computing [1] and offload intensive mobile computation to nearby clouds at the edges of cellular networks, called *edge clouds*, with short latency, referred to as *mobile-edge computation offloading (MECO)*. In this paper, we consider a MECO system with a single edge cloud serving multiple users and investigate the energy-efficient resource allocation.

Mobile computation offloading (MCO) (or mobile cloud computing) has been extensively studied in computer science, including system architectures [2], virtual machine migration [3] and server consolidation [4]. It is commonly assumed that the implementation of MCO relies on a network architecture with a central cloud (e.g., a data center). This architecture has the drawbacks of high overhead and long backhaul latency [5] and will soon encounter the performance bottleneck of finite backhaul capacity in view of exponential mobile traffic growth. These issues can be overcome by MECO based on a network architecture supporting distributed mobile-edge computing.

Energy efficient MECO requires the joint design of MCO and wireless communication techniques. Recent years have seen research progress on this topic. For a single-user MECO system, the optimal offloading decision policy was derived

in [6] by comparing the energy consumption of optimized local computing (with variable CPU cycles) and offloading (with variable transmission rates). This framework was further developed in [7] and [8] to enable adaptive offloading powered by wireless energy transfer and energy harvesting, respectively. In [9], also for a single-user MECO system, dynamic offloading was integrated with adaptive LTE/WiFi link selection. Moreover, resource allocation for MECO has been studied for various types of multiuser systems [10]–[12]. In [10], considering a multi-cell MECO system, the radio and computation resources were jointly allocated to minimize the mobile energy consumption under offloading latency constraints. With the coexistence of central and edge clouds, the optimal user scheduling for offloading to different clouds was studied in [11]. In addition, the distributed offloading for multiuser MECO was designed in [12] using game theory for both energy-and-latency minimization. Prior work on MECO resource allocation focuses on complex algorithmic designs and yields little insight into the optimal policy structures. In contrast, for a multiuser MECO system based on time-division multiple access (TDMA), the optimal resource-allocation policy is shown in current work to have a simple threshold-based structure with respect to a derived offloading priority function.

Resource allocation has been widely studied for various types of multiuser communication systems, e.g., TDMA (see e.g., [13]), orthogonal frequency-division multiple access (OFDMA) (see e.g., [14]) and code-division multiple access (CDMA) (see e.g., [15]). Note that all of them only focus on the radio resource allocation. In contrast, for newly proposed MECO systems, both the computation and radio resource allocation at edge clouds need to be jointly optimized for the maximum mobile energy savings, which makes the algorithmic design more complex.

This paper considers a multiuser MECO system based on TDMA. Consider both the cases of infinite and finite cloud computation capacities. The optimal resource-allocation policy is derived by solving a convex optimization problem that minimizes the weighted sum mobile energy consumption. Note that the consideration of MECO simplifies the problem formulation since the long backhaul latency and heavy overhead in central clouds can be neglected. To solve the problem, an *offloading priority function* is derived that yields priorities for users and depends on their channel gains and local computing energy consumption. Based on this, the optimal policy is proved to have an insightful threshold-based structure that determines complete or minimum offloading for users with priorities above

or below a given threshold, respectively. Moreover, to reduce the complexity for computing the threshold, a simple sub-optimal resource-allocation algorithm is designed and shown to have close-to-optimal performance by simulation.

II. SYSTEM MODEL

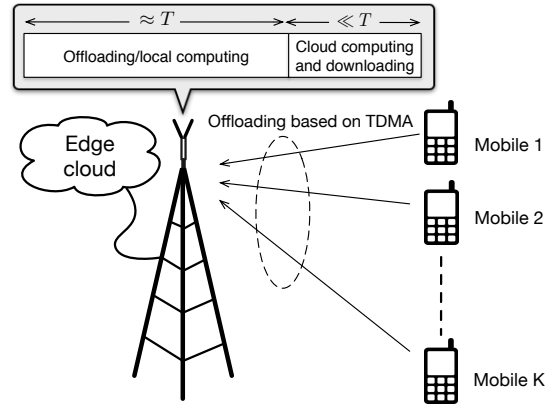
Consider a multiuser MECO system shown in Fig. 1(a) that comprises K single-antenna mobiles, indexed as $1, 2, \dots, K$, and one single-antenna base station (BS) that is the gateway of an edge cloud. Time is divided into slots each with a duration of T seconds. As shown in Fig. 1(a), each slot comprises two sequential phases for 1) mobile offloading or local computing and 2) cloud computing and downloading of computation results from the edge cloud to mobiles. Cloud computing has small latency; the downloading does not consume mobile energy and furthermore is much faster than offloading due to relative smaller sizes of computation results. For these reasons, the second phase is assumed to have a negligible duration compared with the first phase and not considered in resource allocation. Considering an arbitrary slot, the BS schedules a subset of users for complete/partial offloading based on TDMA. The user with partial or no offloading computes a fraction of or all input data, respectively, using a local CPU. Moreover, the BS is assumed to have perfect knowledge of multiuser channel gains, local computing energy per bit and sizes of input data at all users. Using these information, the BS selects offloading users, determines the offloaded data sizes and allocates fractions of the slot to offloading users with the criterion of minimum weighted sum mobile energy consumption. In addition, channels are assumed to remain constant within each slot.

The model of local computing is described as follows. Assume that the CPU frequency is fixed at each user and may vary over users. Consider an arbitrary time slot. Following the model in [12], let C_k denote the number of CPU cycles required for computing 1-bit of input data at the k -th mobile, and P_k the energy consumption per cycle for local computing at this user. Then the product $C_k P_k$ gives computing energy per bit. As shown in Fig. 1(b), mobile k is required to compute R_k -bit input data within the slot, out of which ℓ_k -bit is offloaded and $(R_k - \ell_k)$ -bit is computed locally. Then the total energy consumption for local computing at mobile k , denoted as $E_{\text{loc},k}$, is given by $E_{\text{loc},k} = (R_k - \ell_k)C_k P_k$. Let F_k denote the computation capacity of mobile k that is measured by the number of CPU cycles per second. Under the computation latency constraint, $C_k(R_k - \ell_k) \leq F_k T$. As a result, the offloaded data at mobile k has the minimum size of $\ell_k \geq m_k^+$ with $m_k = R_k - \frac{F_k T}{C_k}$, where the function $(x)^+ = \max\{x, 0\}$.

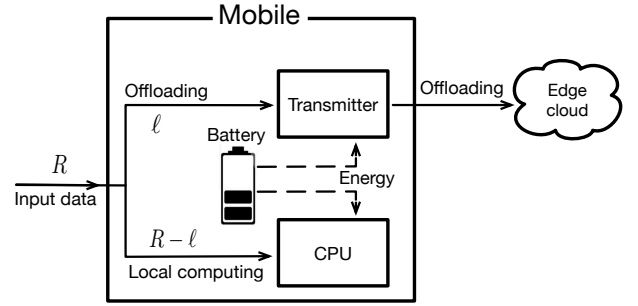
Next, the energy consumption for offloading is modeled. Let h_k denote the channel gain and p_k the transmission power for mobile k . Then the achievable rate, denoted by r_k , is given as:

$$r_k = B \log \left(1 + \frac{p_k h_k^2}{N_0} \right) \quad (1)$$

where N_0 is the variance of complex white Gaussian channel noise. The fraction of slot allocated to mobile k for offloading is denoted as t_k with $t_k \geq 0$, where $t_k = 0$ corresponds to no offloading. For the case of offloading ($t_k > 0$), the transmission



(a) Multiuser MECO system.



(b) Mobile computation offloading.

Figure 1. (a) Multiuser MECO system and (b) mobile computation offloading.

rate is fixed as $r_k = \ell_k / t_k$ since this is the most energy-efficient transmission policy under a deadline constraint. Define a function $f(x) = N_0(2^{\frac{x}{B}} - 1)$. It follows from (1) that the energy consumption for offloading at mobile k is

$$E_{\text{off},k} = p_k t_k = \frac{t_k}{h_k^2} f\left(\frac{\ell_k}{t_k}\right). \quad (2)$$

Note that if either $\ell_k = 0$ or $t_k = 0$, $E_{\text{off},k}$ is equal to zero.

Last, consider the edge cloud. It is assumed that the edge cloud has finite computation capacity, denoted as F , measured as the maximum CPU cycles allowed for computing the sum offloaded data in each slot: $\sum_{k=1}^K C_k \ell_k \leq F$. This constraint ensures low latency for cloud computing.

III. MULTIUSER MECO: PROBLEM FORMULATION

In this section, resource allocation for multiuser MECO is formulated as an optimization problem. The objective is to minimize the weighted sum mobile energy consumption: $\sum_{k=1}^K \beta_k (E_{\text{off},k} + E_{\text{loc},k})$, where the positive weight factors $\{\beta_k\}$ account for fairness among mobiles. Under the constraints on time-sharing, cloud computation capacity and computation latency, the resource allocation problem is formulated as follows:

$$\begin{aligned} \min_{\{\ell_k, t_k\}} \quad & \sum_{k=1}^K \beta_k \left[\frac{t_k}{h_k^2} f\left(\frac{\ell_k}{t_k}\right) + (R_k - \ell_k) C_k P_k \right] \\ \text{s.t.} \quad & \sum_{k=1}^K t_k \leq T, \quad \sum_{k=1}^K C_k \ell_k \leq F, \\ & t_k \geq 0, \quad m_k^+ \leq \ell_k \leq R_k, \quad \forall k. \end{aligned} \quad (\text{P1})$$

Several basic characteristics of Problem P1 are given in the following two lemmas.

Lemma 1. Problem P1 is a convex optimization problem.

Proof: See Appendix A. \square

Lemma 2. The feasibility condition for Problem P1 is: $\sum_{k=1}^K m_k^+ C_k \leq F$.

Proof: See Appendix B. \square

Lemma 2 shows that whether the cloud computation capacity constraint is satisfied determines the feasibility of this optimization problem, while the time-sharing constraint can always be satisfied and only affects the mobile energy consumption.

Assume that Problem P1 is feasible. The direct solution of Problem P1 using the dual-decomposition approach (the Lagrange method) requires iterative computation and yields no insight into the structure of the optimal policy. To address these issues, we adopt a two-stage solution approach that requires first solving Problem P2 below that relaxes Problem P1 by removing the constraint on the cloud computation capacity:

$$\begin{aligned} \min_{\{\ell_k, t_k\}} \quad & \sum_{k=1}^K \beta_k \left[\frac{t_k}{h_k^2} f\left(\frac{\ell_k}{t_k}\right) + (R_k - \ell_k) C_k P_k \right] \\ \text{s.t.} \quad & \sum_{k=1}^K t_k \leq T, \\ & t_k \geq 0, \quad m_k^+ \leq \ell_k \leq R_k, \quad \forall k. \end{aligned} \quad (\text{P2})$$

If the solution for Problem P2 violates the constraint on cloud computation capacity, Problem P1 is then incrementally solved building on the solution for Problem P2. This approach allows the optimal policy to be shown to have the said threshold-based structure and also facilitates the design of low-complexity close-to-optimal resource-allocation algorithm. It is interesting to note that Problem P2 corresponds to the case where the edge cloud has infinite computation capacity. The detailed procedures for solving Problems P1 and P2 are presented in the subsequent two sections.

IV. MULTIUSER MECO: INFINITE CLOUD CAPACITY

In this section, by solving Problem P2 using the Lagrange method, we derive a threshold-based policy for the optimal resource allocation. Moreover, the policy is simplified for several special cases.

To solve Problem P2, the Lagrange function is defined as

$$L = \sum_{k=1}^K \beta_k \left[\frac{t_k}{h_k^2} f\left(\frac{\ell_k}{t_k}\right) + (R_k - \ell_k) C_k P_k \right] + \lambda \left(\sum_{k=1}^K t_k - T \right)$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with the time-sharing constraint. For ease of notation, define a function $g(x) = f(x) - x f'(x)$. Let $\{\ell_k^{*(2)}, t_k^{*(2)}\}$ denote the solution for Problem P2 that always exists according to Lemma 2. Then applying KKT conditions leads to the following necessary and sufficient conditions:

$$\frac{\partial L}{\partial \ell_k^{*(2)}} = \frac{\beta_k f'\left(\frac{\ell_k^{*(2)}}{t_k^{*(2)}}\right)}{h_k^2} - \beta_k C_k P_k \begin{cases} > 0, & \ell_k^{*(2)} = m_k^+ \\ = 0, & \ell_k^{*(2)} \in (m_k^+, R_k), \forall k., \\ < 0, & \ell_k^{*(2)} = R_k \end{cases} \quad (3a)$$

$$\frac{\partial L}{\partial t_k^{*(2)}} = \frac{\beta_k g\left(\frac{\ell_k^{*(2)}}{t_k^{*(2)}}\right)}{h_k^2} + \lambda^* \begin{cases} > 0, & t_k^{*(2)} = 0 \\ = 0, & t_k^{*(2)} > 0 \end{cases}, \forall k., \quad (3b)$$

$$\sum_{k=1}^K t_k^{*(2)} \leq T, \quad \lambda^* \left(\sum_{k=1}^K t_k^{*(2)} - T \right) = 0. \quad (3c)$$

Based on these conditions, the optimal policy for resource allocation is characterized in the following sub-sections.

A. Offloading Priority Function

Define an (mobile) *offloading priority function*, which is essential for the optimal resource allocation, as follows:

$$\varphi(\beta_k, C_k, P_k, h_k) = \begin{cases} \frac{\beta_k N_0}{h_k^2} (v_k \ln v_k - v_k + 1), & v_k \geq 1 \\ 0, & v_k < 1 \end{cases}, \quad (4)$$

with the constant v_k defined as

$$v_k = \frac{B C_k P_k h_k^2}{N_0 \ln 2}. \quad (5)$$

This function is derived by solving a useful equation as shown in the following lemma.

Lemma 3. Given $v_k \geq 1$, the offloading priority function $\varphi(\beta_k, C_k, P_k, h_k)$ in (4) is the root of the equation with respect to x :

$$f'^{-1}(C_k P_k h_k^2) = g^{-1}\left(\frac{-h_k^2 x}{\beta_k}\right).$$

Proof: See Appendix C. \square

The function generates an offloading priority value, $\varphi_k = \varphi(\beta_k, C_k, P_k, h_k)$, for mobile k depending on corresponding variables quantifying fairness, local computing and channel. The amount of offloaded data by a mobile grows with an increasing offloading priority as shown in the next sub-section. It is useful to understand the effects of parameters on the offloading priority that are characterized as follows.

Lemma 4. Given $v \geq 1$, $\varphi(\beta, C, P, h)$ is a *monotone increasing function* for β , C , P and h .

Lemma 4 can be easily proved by deriving the first derivatives of φ with respect to each parameter. Moreover, it is consistent with the intuition that, to reduce energy consumption by offloading, the BS should schedule those mobiles having high computing energy consumption per bit (i.e., large C and P) or good channels (i.e., large h).

Remark 1 (Effects of parameters on the offloading priority). It can be observed from (4) and (5) that the offloading priority scales with local computing energy per bit CP approximately as $(CP) \ln(CP)$ and with the channel gain h approximately as $\ln h$. The former scaling is much faster than the latter. This shows that the computing energy per bit is dominant over the channel on determining whether to offload.

B. Optimal Resource-Allocation Policy

Based on conditions in (3a)-(3c) and Lemma 3, the main result of this section is derived, given in the following theorem.

Theorem 1 (Optimal Resource-Allocation Policy). Consider the case of infinite cloud computation capacity. The optimal policy solving Problem P2 has the following structure.

- 1) If $v_k \leq 1$ and the minimum offloaded data size $m_k^+ = 0$ for all k , none of these users performs offloading, i.e.,

$$\ell_k^{*(2)} = t_k^{*(2)} = 0 \quad \forall k.$$

- 2) If there exists mobile k such that $v_k > 1$ or $m_k^+ > 0$, for $k = 1, 2, \dots, K$,

$$\ell_k^{*(2)} \begin{cases} = m_k^+, & \varphi_k < \lambda^* \\ \in [m_k^+, R_k], & \varphi_k = \lambda^* \\ = R_k, & \varphi_k > \lambda^* \end{cases}$$

and

$$t_k^{*(2)} = \frac{\ln 2}{B \left[W_0 \left(\frac{\lambda^* h_k^2 / \beta_k - N_0}{N_0 e} \right) + 1 \right]} \times \ell_k^{*(2)}$$

where $W_0(x)$ is the Lambert function and λ^* is the optimal value of the Lagrange multiplier. Furthermore, the time-sharing constraint is active: $\sum_{k=1}^K t_k^{*(2)} = T$.

Proof: See Appendix D. \square

Theorem 1 reveals that the optimal resource-allocation policy has a threshold-based structure when offloading saves energy. In other words, since the exact case of $\varphi_k = \lambda^*$ rarely occurs in practice, the optimal policy makes a *binary offloading decision* for each mobile. Specifically, if the corresponding offloading priority exceeds a given threshold, the mobile should offload all input data to the edge cloud; otherwise, the mobile should offload only the minimum amount of data under the computation latency constraint. This result is consistent with the intuition that the greedy method can lead to the optimal resource allocation.

Remark 2 (Offloading or not?). For a conventional TDMA communication system, continuous transmission by at least one mobile is always advantageous under the criterion of minimum sum energy consumption [13]. However, this does not always hold for a TDMA MECO system where no offloading for all users may be preferred as shown in Theorem 1. There are two cases where offloading is necessary. First, there exists at least one mobile whose input data size is too large such that complete local computing fails to meet the latency constraint. Second, some mobile has a sufficient high value for the product $C_k P_k h_k^2$, indicating that energy savings can be achieved by offloading because of high channel gain or large local computing energy consumption.

Remark 3 (Offloading rate). It can be observed from Theorem 1 that the offloading rate, defined as $\ell_k^{*(2)} / t_k^{*(2)}$ for mobile k , is determined only by the channel gain and fairness weight factor while other factors, namely C_k and P_k , affect the offloading decision. The rate increases with a growing channel gain and vice versa since a large channel gain supports a

higher transmission rate or reduces transmission power, making offloading desirable for reducing mobile energy consumption.

Remark 4 (Algorithm computation complexity). The traditional method for solving Problem P2 is the block-coordinate descending which performs iterative optimization of the two sets of variables, $\{\ell_k\}$ and $\{t_k\}$, resulting in high computation complexity. In contrast, by exploiting the threshold-based structure of the optimal resource-allocation policy in Theorem 1, the proposed solution approach, described in Algorithm 1, needs to perform only a *one-dimension* search for λ^* , reducing the computation complexity significantly. To facilitate the search, next lemma gives the range of λ^* , which can be easily proved from Theorem 1 and omitted for simplicity.

Lemma 5. When there is at least one offloading mobile, the optimal Lagrange multiplier λ^* satisfies:

$$0 \leq \lambda^* \leq \lambda_{\max} = \max_k \varphi_k.$$

Algorithm 1 Optimal Algorithm for Problem P2.

• **Step 1** [Initialize]:

Let $\lambda_\ell = 0$ and $\lambda_h = \lambda_{\max}$. According to Theorem 1, obtain $T_\ell = \sum_{k=1}^K t_{k,\ell}^{*(2)}$ and $T_h = \sum_{k=1}^K t_{k,h}^{*(2)}$, where $\{t_{k,\ell}^{*(2)}\}$ and $\{t_{k,h}^{*(2)}\}$ are the allocated fractions of slot for the cases of λ_ℓ and λ_h , respectively.

• **Step 2** [Bisection search]:

While $T_\ell \neq T$ and $T_h \neq T$, update $\{\lambda_\ell, \lambda_h\}$ as follows.

(1) Define $\lambda_m = (\lambda_\ell + \lambda_h)/2$ and compute T_m .

(2) If $T_m = T$, then $\lambda^* = \lambda_m$ and the optimal policy can be determined. Otherwise, if $T_m < T$, let $\lambda_h = \lambda_m$ and if $T_m > T$, let $\lambda_\ell = \lambda_m$.

C. Special Cases

The optimal resource-allocation policies for several special cases considering equal weight factors are discussed as follows.

1) *Uniform channels and local computing:* Consider the simplest case where $\{h_k, C_k, P_k\}$ are identical for all k . Then all mobiles have uniform offloading priorities. In this case, for optimal resource allocation, different mobiles can offload arbitrary data sizes so long as the sum offloaded data size satisfies the following constraint:

$$\sum_{k=1}^K \ell_k^{*(2)} \leq TB \log_2 \left(\frac{BCPh^2}{N_0 \ln 2} \right).$$

2) *Uniform channels:* Consider the case of $h_1 = h_2 \dots = h_K$. The offloading priority for each mobile, say mobile k , is only affected by the corresponding local-computing parameters P_k and C_k . Without loss of generality, assume that $P_1 C_1 \leq P_2 C_2 \dots \leq P_K C_K$. Then the optimal resource-allocation policy is given in the following corollary of Theorem 1.

Corollary 1. Assume infinite cloud computation capacity, $h_1 = h_2 \dots = h_K$ and $P_1 C_1 \leq P_2 C_2 \dots \leq P_K C_K$. Let k_t denote the index such that $\varphi_k < \lambda^*$ for all $k < k_t$ and $\varphi_k > \lambda^*$ for all $k \geq k_t$. The optimal resource-allocation policy is given

as follows:

$$\ell_k^{*(2)} = \begin{cases} R_k, & k \geq k_t \\ m_k^+, & \text{otherwise} \end{cases},$$

and

$$t_k^{*(2)} = \frac{\ln 2}{B \left[W_0 \left(\frac{\lambda^* h^2 / \beta - N_0}{N_0 e} \right) + 1 \right]} \times \ell_k^{*(2)}.$$

The result shows that the optimal resource-allocation policy follows a *greedy* approach that selects mobiles in a descending order of energy consumption per bit for complete offloading until the time-sharing duration is fully utilized.

3) *Uniform local computing*: Consider the case of $C_1 P_1 = C_2 P_2 \cdots = C_K P_K$. Similar to the previous case, the optimal resource-allocation policy can be shown to follow the greedy approach that selects mobiles for complete offloading in the descending order of channel gain.

V. MULTIUSER MECO: FINITE CLOUD CAPACITY

In this section, we consider the case of finite cloud computation capacity and analyze the optimal resource-allocation policy for solving Problem P1. The policy is shown to also have a threshold-based structure as the infinite-capacity counterpart derived in the preceding section. Both the optimal and sub-optimal algorithms are presented for policy computation.

A. Optimal Resource-Allocation Policy

To solve the convex Problem P1, the corresponding Lagrange function can be written as

$$\begin{aligned} \tilde{L} = & \sum_{k=1}^K \beta_k \left[\frac{t_k}{h_k^2} f\left(\frac{\ell_k}{t_k}\right) + (R_k - \ell_k) C_k P_k \right] \\ & + \lambda \left(\sum_{k=1}^K t_k - T \right) + \mu \left(\sum_{k=1}^K C_k \ell_k - F \right). \end{aligned}$$

where $\mu \geq 0$ is the Lagrange multiplier corresponding to the cloud computation capacity constraint. Using the above Lagrange function, it is straightforward to show that the corresponding KKT conditions can be modified from their infinite-capacity counterparts in (3a)-(3c) by replacing P_k with $\tilde{P}_k = P_k - \mu$, called the *effective computation energy per cycle*. The resultant *effective offloading priority function*, denoted as $\tilde{\varphi}_k$, can be modified accordingly from that in (4) as

$$\tilde{\varphi}(\beta_k, C_k, P_k, h_k, \mu) = \begin{cases} \frac{\beta_k N_0}{h_k^2} (\tilde{v}_k \ln \tilde{v}_k - \tilde{v}_k + 1), & \tilde{v}_k \geq 1 \\ 0, & \tilde{v}_k < 1 \end{cases}, \quad (6)$$

where $\tilde{v}_k = \frac{BC_k(P_k - \mu)h_k^2}{N_0 \ln 2}$. Based on above discussion, the main result of this section follows as shown below.

Theorem 2 (Optimal Resource-Allocation Policy). Consider the case of finite cloud computation capacity. The optimal policy solving Problem P1 has the same structure as that in Theorem 1 and is expressed in terms of the priority function $\tilde{\varphi}_k$ in (6) and the optimized Lagrange multipliers $\{\lambda^*, \mu^*\}$.

Computing the threshold for the optimal resource-allocation policy requires a *two-dimension search* over the Lagrange multipliers $\{\lambda^*, \mu^*\}$, using Algorithm 2. For an efficient search, it is useful to limit the range of λ^* and μ^* as follows.

Lemma 6. When there is at least one offloading mobile, the optimal Lagrange multipliers $\{\lambda^*, \mu^*\}$ satisfy:

$$0 \leq \lambda^* \leq \lambda_{\max},$$

$$0 \leq \mu^* \leq \mu_{\max} = \max_k \left\{ P_k - \frac{N_0 \ln 2}{BC_k h_k^2} \right\}$$

where λ_{\max} has been defined in Lemma 5.

Proof: See Appendix E □

Note that $\mu^* = 0$ corresponds to the case of infinite cloud computation capacity and $\mu^* = \mu_{\max}$ to the case where offloading yields no energy savings for any mobile.

Algorithm 2 Optimal Algorithm for Solving Problem P1.

- **Step 1**[Check solution for Problem P2]:
Perform Algorithm 1. If $\sum_{k=1}^K \ell_k^{*(2)} \leq F$, the optimal policy is given in Theorem 1. Otherwise, go to Step 2.
- **Step 2** [Initialize]:
Let $\mu_\ell = 0$ and $\mu_h = \mu_{\max}$. Based on Theorem 2, obtain $F_\ell = \sum_{k=1}^K C_k \ell_{k,\ell}^*$ and $F_h = \sum_{k=1}^K C_k \ell_{k,h}^*$, where $\{\ell_{k,\ell}^*\}$ and $\{\ell_{k,h}^*\}$ are the offloaded data sizes for μ_ℓ and μ_h , respectively, involving the one-dimension search for λ^* .
- **Step 3** [Bisection search]:
While $F_\ell \neq F$ and $F_h \neq F$, update $\{\mu_\ell, \mu_h\}$ as follows.
(1) Define $\mu_m = (\mu_\ell + \mu_h)/2$ and compute F_m .
(2) If $F_m = F$, then $\mu^* = \mu_m$ and the optimal policy can be determined. Otherwise, if $F_m < F$, let $\mu_h = \mu_m$ and if $F_m > F$, let $\mu_\ell = \mu_m$.

B. Sub-Optimal Resource-Allocation Policy

To reduce the computation complexity of Algorithm 2 due to the two-dimension search, one simple sub-optimal policy is designed using Algorithm 3. The key idea is to decouple the computation and radio resource allocation. In Step 2, based on the *approximated* offloading priority in (4) for the case of infinite cloud computation capacity, we allocate the computation resource to mobiles with high offloading priorities. Step 3 optimizes the corresponding fractions of slot given offloaded data. This sub-optimal algorithm has low complexity requiring only a one-dimension search. Moreover, its performance is shown by simulation to be close-to-optimal in the sequel.

VI. SIMULATION RESULTS

The simulation settings are as follows unless specified otherwise. The MECO system comprises $K = 30$ mobiles with equal fairness weight factors, namely that $\beta_k = 1$ for all k such that the weighted sum mobile energy consumption represents the total mobile energy consumption. The time slot $T = 100$ ms and channels are modeled as independent Rayleigh fading with average power loss set as 10^{-6} . In addition, the variance of complex white Gaussian channel noise is $N_0 = 10^{-9}$ W and the bandwidth $B = 10$ Mhz. Consider mobile k . The CPU computation capacity F_k is uniformly

Algorithm 3 Sub-optimal Algorithm for Solving Problem P1.

- **Step 1:** Perform Algorithm 1. If $\sum_{k=1}^K \ell_k^{*(2)} \leq F$, Theorem 1 gives the optimal policy. Otherwise, go to Step 2.
 - **Step 2:** Based on offloading priorities in (4), offload the data from mobiles in the descending order of offloading priority until the cloud computation capacity is fully occupied, i.e., $\sum_{k=1}^K C_k \ell_k^* = F$.
 - **Step 3:** With $\{\ell_k^*\}$ derived in Step 2, search for λ^* such that $t_k^* = \frac{\ell_k^* \ln 2}{B[W_0(\frac{\lambda^* h_k^2 / \beta_k - N_0}{N_0 e}) + 1]}$ satisfying $\sum_{k=1}^K t_k^* = T$.
-

selected from the set $\{0.1, 0.2, \dots, 1.0\}$ Ghz and the local computing energy per cycle P_k follows a uniform distribution in the range $(0, 20 \times 10^{-11})$ J/cycle. For the computing task, both the data size and required number of CPU cycles per bit follow the uniform distribution with $R_k \in [100, 500]$ KB and $C_k \in [500, 1500]$ cycles/bit. All random variables are independent for different mobiles, modeling heterogeneous mobile computing capabilities. Last, the cloud computation capacity is set as $F = 6 \times 10^9$ cycles per slot.

For performance comparison, a baseline *equal resource-allocation* policy is considered, which allocates equal offloading duration for mobiles satisfying $v_k > 1$ and based on this, the offloaded data sizes are optimized.

Fig. 2 shows the curves of total mobile energy consumption versus the time slot duration T . Several observations can be made. First, the total mobile energy consumption reduces as the slot duration grows. Next, the sub-optimal policy computed using Algorithm 3 is found to have close-to-optimal performance and yields total mobile energy consumption less than half of that for the equal resource-allocation policy. The energy reduction is more significant for a shorter slot duration since without the optimization on fractions of slot, the offloading energy of baseline policy grows exponentially with the decrease of allocated time fractions.

The curves of total mobile energy consumption versus the cloud computation capacity are displayed in Fig. 3. It can be observed that the performance of the sub-optimal policy approaches to that of the optimal one when the cloud computation capacity increases and achieves substantial energy savings gains over the equal resource-allocation policy. Furthermore, the total mobile energy consumption is invariant after the cloud computation capacity exceeds some threshold (about 6×10^9). This suggests that there exists some critical value for the cloud computation capacity, above which increasing the capacity yields no reduction on the total mobile energy consumption.

VII. CONCLUSION

Consider a multiuser MECO system based on TDMA. This work shows that the optimal energy-efficient resource-allocation policy for clouds with infinite or finite computation capacities, is featured with a threshold-based structure. Specifically, the BS makes a binary offloading decision for each mobile, where users with priorities above or below a given threshold will perform complete or minimum offloading. Moreover, a simple sub-optimal algorithm is proposed to reduce the complexity for computing the threshold.

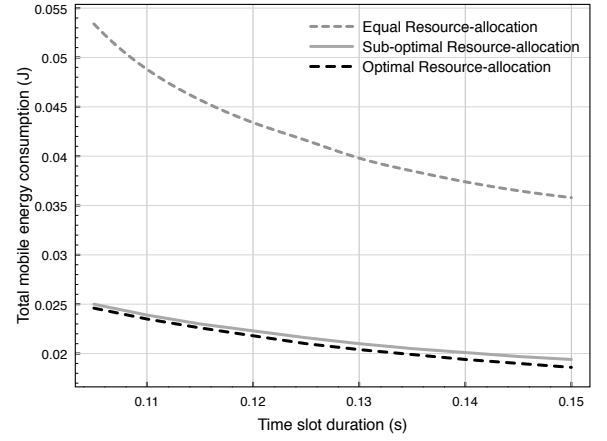


Figure 2. Total mobile energy consumption vs. time slot duration.

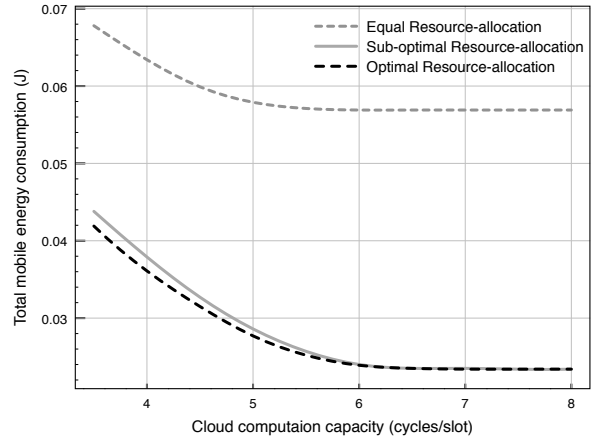


Figure 3. Total mobile energy consumption vs. cloud computation capacity.

APPENDIX

A. Proof of Lemma 1

Since $f(x)$ is a convex function, its perspective function, defined as $t_k f(\frac{\ell_k}{t_k})$, is still convex. Thus, the objective function, the summation of a set of convex functions, preserves the convexity. Combining it with the linear convex constraints leads to the desired result. ■

B. Proof of Lemma 2

Whether Problem P1 is feasible depends on the following two key constraints: $\sum_{k=1}^K C_k \ell_k \leq F$ and $m_k^+ \leq \ell_k \leq R_k$. Assume $m_k^+ \leq \ell_k \leq R_k$ is satisfied. Then it has

$$\sum_{k=1}^K C_k m_k^+ \leq \sum_{k=1}^K C_k \ell_k \leq \sum_{k=1}^K C_k R_k.$$

Thus, only when $\sum_{k=1}^K C_k m_k^+ \leq F$, Problem P1 is feasible. ■

C. Proof of Lemma 3

First, we derive a general result that is the root of equation: $f'^{-1}(p) = g^{-1}(y)$ with respect to y as follows.

According to the definitions of $f(x)$ and $g(x)$, it has

$$f'(x) = \frac{N_0 \ln 2}{B} 2^{\frac{x}{B}} \quad \text{and} \quad f'^{-1}(y) = B \log_2 \left(\frac{By}{N_0 \ln 2} \right). \quad (7)$$

Thus, the solution for the general equation is

$$\begin{aligned} y &= g(f'^{-1}(p)) = f(f'^{-1}(p)) - f'^{-1}(p) \times f'(f'^{-1}(p)) \\ &= f(f'^{-1}(p)) - f'^{-1}(p) \times p \\ &= \frac{Bp}{\ln 2} - N_0 - pB \log_2 \left(\frac{Bp}{N_0 \ln 2} \right). \end{aligned} \quad (8)$$

Note that to ensure $\ell_k^{*(2)} \geq 0$ in Problem P1, we need $f'^{-1}(C_k P_k h_k^2) \geq 0$, which is equivalent to $v_k \geq 1$ derived from (7). Then, by substituting $p = C_k P_k h_k^2$ and $y = \frac{-h_k^2 x}{\beta_k}$ to (8) and making arithmetic operations gives the desired result. ■

D. Proof of Theorem 1

First, to prove this theorem, we need the following lemmas.

Lemma 7. The function $g^{-1}(y)$ can be expressed as

$$g^{-1}(y) = \frac{B \left[W_0 \left(\frac{y+N_0}{-N_0 e} \right) + 1 \right]}{\ln 2}. \quad (9)$$

Proof: Since $g^{-1}(y)$ denotes the root of equation $g(x) = y$ for $x \geq 0$, it has

$$\begin{aligned} y &= g(x) = \left[N_0 - \frac{(\ln 2) N_0 x}{B} \right] \times 2^{\frac{x}{B}} - N_0 \\ &= (-N_0 e) \times \left[\left(\frac{x \ln 2}{B} - 1 \right) e^{\frac{x \ln 2}{B} - 1} \right] - N_0. \end{aligned}$$

Thus, based on the definition for Lambert function, we have $\frac{x \ln 2}{B} - 1 = W_0 \left(\frac{y+N_0}{-N_0 e} \right)$. Then the desired result follows. □

Lemma 8. The function $g^{-1}(y)$ is a monotone decreasing function for $y < 0$.

Proof: From (9), for $y \leq 0$, it has $\frac{y+N_0}{-N_0 e} \geq -1/e$. Since the single-valued Lambert function $W_0(x)$ is monotone increasing for $x \geq -1/e$, we can easily obtain the desired result. □

Then, consider case 1) in Theorem 1. Note that for mobile k , if $m_k^+ = 0$ and $v_k \leq 1$, it results in $\ell_k^{*(2)} = 0$ derived from (3a). Thus, if these two conditions are satisfied for all k , it leads to $\ell_k^{*(2)} = t_k^{*(2)} = 0$. For case 2), if there exists mobile k such that $v_k > 1$ or $m_k^+ > 0$, it ensures $\ell_k^{*(2)} > 0$. And the time-sharing constraint should be active since remaining time can be used for offloading so as to reduce the transmission energy. Moreover, consider each user $k = 1, 2, \dots, K$. If $v_k \geq 1$, from (3a) and (3b), $\{\ell_k^{*(2)}, t_k^{*(2)}\}$ should satisfy the following:

$$\frac{\ell_k^{*(2)}}{t_k^{*(2)}} = \min \left\{ \max \left[\frac{m_k^+}{t_k^{*(2)}}, f'^{-1}(C_k P_k h_k^2) \right], \frac{R_k}{t_k^{*(2)}} \right\} \quad (10a)$$

$$= \max \left\{ \frac{m_k^+}{t_k^{*(2)}}, \min \left[f'^{-1}(C_k P_k h_k^2), \frac{R_k}{t_k^{*(2)}} \right] \right\} \quad (10b)$$

$$= g^{-1} \left(\frac{-h_k^2 \lambda^*}{\beta_k} \right). \quad (10c)$$

Using Lemma 3 and Lemma 8, we have the following:

- 1) If $\varphi_k > \lambda^* \geq 0$, it has $-h_k^2 \varphi_k < -h_k^2 \lambda^* \leq 0$. Then, from (10a), it gives

$$\begin{aligned} \max \left[\frac{m_k^+}{t_k^{*(2)}}, f'^{-1}(C_k P_k h_k^2) \right] &\geq f'^{-1}(C_k P_k h_k^2) \\ &= g^{-1} \left(\frac{-h_k^2 \varphi_k}{\beta_k} \right) > g^{-1} \left(\frac{-h_k^2 \lambda^*}{\beta_k} \right). \end{aligned} \quad (11)$$

From (10a), (10c) and (11), it follows that $\ell_k^{*(2)} = R_k$.

- 2) If $\varphi_k = \lambda^*$, it has $f'^{-1}(C_k P_k h_k^2) = g^{-1} \left(\frac{-h_k^2 \lambda^*}{\beta_k} \right)$.
- 3) If $0 \leq \varphi_k < \lambda^*$, it has $-h_k^2 \varphi_k > -h_k^2 \lambda^*$. Combining it with (10b) leads to

$$\begin{aligned} \min \left[f'^{-1}(C_k P_k h_k^2), \frac{R_k}{t_k^{*(2)}} \right] &\leq f'^{-1}(C_k P_k h_k^2) \\ &= g^{-1} \left(\frac{-h_k^2 \varphi_k}{\beta_k} \right) < g^{-1} \left(\frac{-h_k^2 \lambda^*}{\beta_k} \right). \end{aligned} \quad (12)$$

From (10b), (10c) and (12), it follows that $\ell_k^{*(2)} = m_k^+$.

Furthermore, if $v_k < 1$, it has $\ell_k^{*(2)} = m_k^+$. Note that this case can be included in the scenario of $\varphi_k < \lambda^*$ with the definition of φ_k in (4).

Last, from (10c), it follows that

$$t_k^{*(2)} = \frac{\ell_k^{*(2)}}{g^{-1} \left(\frac{-h_k^2 \lambda^*}{\beta_k} \right)} \stackrel{(a)}{=} \frac{\ell_k^{*(2)} \ln 2}{B \left[W_0 \left(\frac{\lambda^* h_k^2 / \beta_k - N_0}{-N_0 e} \right) + 1 \right]}$$

where (a) is derived using Lemma 7, completing the proof. ■

E. Proof of Lemma 6

If there exists offloading mobile k , it must satisfy $\lambda^* \leq \tilde{\varphi}_k$ and $1 \leq \tilde{v}_k$. Thus, considering all mobiles, it follows $\lambda^* \leq \max_k \{\tilde{\varphi}_k\} = \lambda_{\max}$ and $1 \leq \max_k \left\{ \frac{BC_k(P_k - \mu^*)h_k^2}{N_0 \ln 2} \right\}$. The latter condition is equivalent to $\mu^* \leq \mu_{\max}$, completing the proof. ■

REFERENCES

- [1] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal, et al., "Mobile-edge computing introductory technical white paper," *White Paper, Mobile-edge Computing (MEC) industry initiative*, 2014.
- [2] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *J. Wireless Commun. and Mobile Computing*, vol. 13, no. 18, pp. 1587–1611, 2013.
- [3] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, pp. 1107–1117, Sep. 2013.
- [4] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proc. HotPower*, vol. 10, pp. 1–5, 2008.
- [5] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. IEEE Intl. Conf. Intel. Sys and Cont*, 2016.
- [6] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, 2013.
- [7] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer (extended version)," [Online]. Available: <http://arxiv.org/abs/1507.04094>.
- [8] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *submitted to IEEE J. Select. Areas Commun.*, 2016.
- [9] X. Xiang, C. Lin, and X. Chen, "Energy-efficient link selection and transmission scheduling in mobile cloud computing," *IEEE Wireless Commun. Letters*, vol. 3, pp. 153–156, Jan. 2014.
- [10] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal and Info. Processing over Networks*, Jun. 2015.

- [11] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, "A cooperative scheduling scheme of local cloud and internet cloud for delay-aware mobile cloud computing," *Proc. IEEE Globecom*, pp. 1–6, 2015.
- [12] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE Trans. Networking*, vol. PP, pp. 1–1, Oct. 2015.
- [13] X. Wang and G. B. Giannakis, "Power-efficient resource allocation for time-division multiple access over fading channels," *IEEE Trans. Info. Theory*, vol. 54, pp. 1225–1240, Mar. 2008.
- [14] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1747–1758, Oct. 1999.
- [15] S.-J. Oh, D. Zhang, and K. M. Wasserman, "Optimal resource allocation in multiservice CDMA networks," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 811–821, Jul. 2003.