Energy-efficient Resource Allocation and Channel Assignment for NOMA-based Mobile Edge Computing

Lina Liu*, Bo Sun*, Xiaoqi Tan†, Yu Sing Xiao*, Danny H.K. Tsang*
*Dept. of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
†Dept. of Electrical and Computer Engineering, University of Toronto
Email: *{lliuaw, bsunaa, ysxiao, eetsang}@ust.hk, †xiaoqi.tan@utoronto.ca

***Abstract*— In this paper, we study resource allocation (including power and computation resources) and channel assignment in an uplink Non-orthogonal Multiple Access (NOMA)-based Mobile Edge Computing (MEC) system. Our objective is to minimize the total energy consumption of all users. The problem, however, is a non-convex combinatorial optimization problem. We first investigate the hidden convexity by reformulating the resource allocation problem when the channel assignment is given, and propose an efficient algorithm to allocate the resources by dual decomposition methods. Furthermore, we design a heuristic algorithm to decide the channel assignment leveraging the structural property in the reformulation. Extensive simulations verify that NOMA has great advantages over Orthogonal Multiple Access (OMA) in multi-user latency-intensive MEC systems.**

***Index Terms*—Mobile Edge Computing, Non-orthogonal Multiple Access, Resource Allocation, Channel Assignment.**

I. Introduction

The rapid development of Internet of Things (IoT) is now putting forward stringent requirements for IoT devices in terms of low latency and massive connectivity [1]. Typically characterized by the limited CPU capacity and battery lifetime [2], end devices tend to offload the computationally intensive tasks to the cloud. As a critical offloading technology, Mobile Edge Computing (MEC), which is named as Multi-access Edge Computing in the latest ETSI standard, provides an IT service environment and cloud-computing capabilities at the edge of the mobile network [3]. The close proximity to users effectively reduces latency. Meanwhile, power domain Non-orthogonal Multiple Access (NOMA) allows multiple users to transmit in the same frequency resource block simultaneously, and adopts Successive Interference Cancellation (SIC) to decode the superimposed signals at the receiver. Compared to conventional Orthogonal Multiple Access (OMA), NOMA achieves improved spectral efficiency, high system throughput and massive connectivity [4]. Therefore, we propose a multi-user NOMA-based MEC system by combining the two technologies. The influence of NOMA on reducing users' energy consumption in MEC systems is especially studied.

A. Related Works

MEC systems based on OMA schemes have been investigated with different optimization objectives and in multifarious situations [2] [5] [6]. Meanwhile, NOMA has attracted huge research interests both in uplink and downlink scenarios [7] [8]. Recently, introducing NOMA into MEC systems has gained increasing attention. We classify related literatures into two main streams according to their optimization goals: 1) Delay Minimization: In [9], computation offloading with opportunistic channel sharing was studied to minimize the latency by power control. The offloaded workloads and the transmission time were jointly optimized in [10] to minimize the overall delay of the users for finishing the required computation. 2) Energy Minimization: Joint power and time allocation scheme was illustrated in [11], where offloading time was a controllable variable and was optimized to reduce energy consumption. Groups of NOMA users were considered to transmit in sequence in [12], and time allocation among different groups was investigated. Focusing on finding out the optimal task partition, users transmission power and decoding order, authors in [13] designed a multi-group system, in which NOMA-grouped users transmitted concurrently in different channels. In a similar multi-group system, [14] discretized computation and communication resources into blocks and optimized their allocation along with user clustering and power allocation. Apart from that, the impact of NOMA on both latency and energy consumption was studied in [15], where uplink and downlink situations were discussed with fixed resource allocation. All literatures show that NOMA can help improve the performance of MEC systems.

B. Main Contributions

In this paper, we devise a NOMA-based MEC system, aiming to minimize users' energy consumption. Unlike [9]-[12] [15], we consider an MEC system with multiple groups of NOMA users who can transmit concurrently in different sub-channels. Compared to [13]- [14], our major contributions are as follows: 1) We take the computational execution time into account and form a latency-guarantee problem, considering that the execution time has a significant impact on satisfying the latency requirement when the edge cloud computation resource is shared by multiple users. 2) We identify the hidden convexity of the resource allocation problem given channel assignments by reformulation, and design a heuristic algorithm

978-1-5386-7646-2/19/$31.00 ©2019 IEEE

to assign subchannels to NOMA users in an iterative manner by leveraging the convexity property of the resource allocation problem. 3) We evaluate the performance of the NOMA-based MEC system via extensive simulations and show the potential of NOMA technology in reducing energy consumption in a multi-user latency-intensive MEC system.

II. Energy Minimization Problem in the NOMA-based MEC Sytem

A. System Model

In the proposed NOMA-based MEC system, a Base Station (BS) coordinates the offloading of K users' computation tasks through N subchannels. The BS is equipped with an edge cloud server, which is shared by K users to execute their offloaded computation workloads. The offloading is carried out over a frequency band and the total bandwidth B is equally divided into N orthogonal subchannels. Each subchannel holds two users who transmit concurrently by adopting power-domain NOMA technology[1]. We consider each subchannel has two distinct positions corresponding to the two users with distinct channel gains in the same subchannel. Thus, N subchannels have $M = 2N$ positions. Denote by $\mathcal{M} := \{1, \cdots, 2N\}$ the set of all subchannel positions. Furthermore, let $m = 2n-1$ and $m = 2n$ represent the subchannel positions with large and small channel gains of subchannel n, respectively. We denote by $\mathcal{M}^l := \{m \mid m = 2n-1, n = 1, \cdots, N\}$ and $\mathcal{M}^s := \{m \mid m = 2n, n = 1, \cdots, N\}$ the subsets of subchannel positions with large and small channel gains. Assume quasi-static channel situation, where users have constant channel gains over one time slot τ (e.g., several hundred milliseconds) and can have time-varying channel gains from slot to slot. We consider the case that $K = 2N$. Let $\mathcal{K} := \{1, \cdots, 2N\}$ denote the set of users. Each user has a computation task to offload within each time slot over one subchannel. The task should complete its execution in each time slot. The BS and users all have a single antenna for transmitting and receiving.

Consider two users $k, j \in \mathcal{K}$ transmit over subchannel n with large and small channel gains, respectively. The received signal over subchannel n at the receiver is: $y_n = \sqrt{p_{k,2n-1}} h_{k,2n-1} s_{k,2n-1} + \sqrt{p_{j,2n}} h_{j,2n} s_{j,2n} + z_n$, where $s_{k,2n-1}$, $s_{j,2n}$ are the modulated symbols of users k, j, $h_{k,2n-1}$, $h_{j,2n}$ are relative large and small uplink channel gains, $p_{k,2n-1}, p_{j,2n}$ are the transmission power and $z_n \sim \mathcal{N}(0, \sigma^2)$ is the additive white Gaussian noise. Suppose the BS knows perfect Channel State Information (CSI) and applies SIC to decode the superimposed signals on each subchannel. The signal of user k will be decoded first by treating that of user j as interference. User j suffers no interference from user k after subtracting the decoded signal of user k.

B. Problem Formulation

We denote the channel assignment decision variables by $\boldsymbol{x} := \{x_{k,m} \mid k \in \mathcal{K}, m \in \mathcal{M}\}$. Specifically, $x_{k,m} = 1$ represents that user k is assigned to the subchannel position m and $x_{k,m} = 0$ otherwise. We require that each subchannel position has one user and each user is assigned to one subchannel position. In addition, the user in subchannel position $m \in \mathcal{M}^l$ should have no smaller channel gain than the user who shares the same subchannel for data transmission. Thus, the feasible set of $x_{k,m}$ can be represented as

$$\begin{aligned}\mathcal{X} = \Bigg\{ & \boldsymbol{x} \mid x_{k,m} \in \{0,1\}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}; \\ & \textstyle\sum_{k=1}^{K} x_{k,m} = 1, \forall m \in \mathcal{M}; \sum_{m=1}^{M} x_{k,m} = 1, \forall k \in \mathcal{K}; \\ & \sum_{k \in \mathcal{K}} x_{k,m} |h_{k,m}|^2 \geq \sum_{k \in \mathcal{K}} x_{k,m+1} |h_{k,m+1}|^2, \forall m \in \mathcal{M}^l \Bigg\}.\end{aligned}$$

The data rate then can be expressed as

$$\begin{aligned} r_k = & \frac{B}{N} \sum_{m \in \mathcal{M}^l} x_{k,m} \log_2 \left(1 + \frac{p_{k,m} |h_{k,m}|^2}{\sigma^2 + i_m}\right) \\ & + \frac{B}{N} \sum_{m \in \mathcal{M}^s} x_{k,m} \log_2 \left(1 + \frac{p_{k,m} |h_{k,m}|^2}{\sigma^2}\right), \end{aligned} \tag{1}$$

where $i_m = \sum_{j \in \mathcal{K}} x_{j,m+1} p_{j,m+1} |h_{j,m+1}|^2$. Let d_k (bit) denote the data size of user k to be offloaded. The offloading time of user k can be represented as

$$t_k^{\mathrm{o}} = \frac{d_k}{r_k}. \tag{2}$$

Energy consumption for user k to offload is

$$e_k = \sum_{m=1}^{M} x_{k,m} p_{k,m} t_k^{\mathrm{o}}. \tag{3}$$

Let C_k denote the CPU cycles required by each bit of user k for computation and describe the execution time of user k as

$$t_k^{\mathrm{e}} = \frac{d_k C_k}{f_k}, \tag{4}$$

where f_k (cycle/s) is the computation resource allocated to user k at the BS. The feasible set of f_k can be represented as

$$\mathcal{F} := \left\{ \boldsymbol{f} \mid f_k \geq 0, \forall k \in \mathcal{K}; \textstyle\sum_{k=1}^{K} f_k \leq F \right\},$$

where $\boldsymbol{f} := \{f_k \mid k \in \mathcal{K}\}$ and F is the total computation resource available at the BS.

Following the work [2], the latency for users to download processed data from the BS is negligible due to the fact that the processed data has a much smaller size compared to the offloaded raw data and the BS has more power to transmit with a higher data rate. Thus, the total latency includes time for task offloading and computation execution, which is

$$t_k = t_k^{\mathrm{o}} + t_k^{\mathrm{e}}. \tag{5}$$

We intend to minimize the total energy consumption of all users, which comes from task offloading. Therefore, we

[1]The number of users who occupy the same subchannel is constrained to be 2 in order to control the complexity of the SIC [12].

formulate a weighted sum energy minimization problem with weighting factors $\boldsymbol{w} := [w_1, ..., w_K]^T \in \mathbb{R}_+^K$ as follows

$$\begin{aligned}
(\text{P1}) \quad \min_{\boldsymbol{x},\boldsymbol{p},\boldsymbol{f}} \quad & E = \sum\nolimits_{k=1}^{K} w_k e_k, \\
\text{s.t.} \quad & t_k^{\text{o}} + t_k^{\text{e}} \le \tau, \forall k \in \mathcal{K}, && \text{(6a)} \\
& \sum\nolimits_{m=1}^{M} x_{k,m} p_{k,m} \le P_k, \forall k \in \mathcal{K}, && \text{(6b)} \\
& \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{p} \in \mathcal{P}, \quad \boldsymbol{f} \in \mathcal{F}, && \text{(6c)} \\
& \text{constraints} \quad (1) - (5),
\end{aligned}$$

where $\boldsymbol{p} := \{p_{k,m} \mid k \in \mathcal{K}, m \in \mathcal{M}\}$, $\mathcal{P} := \{\boldsymbol{p} \mid p_{k,m} \ge 0, \forall k \in \mathcal{K}\}$ and P_k is the maximal transmission power of user k. Constraint (6a) is to meet the latency requirement and (6b) is to satisfy the power budget P_k.

Problem (P1) is a mixed integer non-convex problem, which is difficult to find a systematic and computationally efficient solution approach. In the sequel, we design an algorithm to solve the problem by settling resource allocation and channel assignment subproblems in an iterative way.

III. Efficient Algorithm Design for P1

The difficulty of solving (P1) as a whole drives us to decompose the problem into two steps. The first step addresses the power and computation resource allocation subproblem given a channel assignment. The second step deals with the channel assignment subproblem taking advantage of the solution obtained in the first step.

A. Power and Computation Resource Allocation

We first solve the power and computation resource allocation subproblem given a channel assignment. In this case, each subchannel position is occupied by a user, and hence we can use the index of subchannel position m to represent either the subchannel position or user. Particularly, user $m = 2n-1$ offloads with a large channel gain over subchannel n and user $m = 2n$ offloads with a small channel gain over subchannel n. Then according to (1), the data rates of the two users in subchannel n are expressed by

$$r_m = \begin{cases} \frac{B}{N} \log_2 \left(1 + \frac{p_{2n-1}|h_{2n-1}|^2}{\sigma^2 + p_{2n}|h_{2n}|^2}\right), & m = 2n-1, \\ \frac{B}{N} \log_2 \left(1 + \frac{p_{2n}|h_{2n}|^2}{\sigma^2}\right), & m = 2n, \end{cases} \tag{7}$$

where $p_{2n-1}(p_{2n})$ and $h_{2n-1}(h_{2n})$ represent the transmission power and channel gain of user $2n-1(2n)$ in subchannel n. The offloading time, execution time and energy consumption of user m are expressed as

$$t_m^{\text{o}} = \frac{d_m}{r_m}, t_m^{\text{e}} = \frac{d_m C_m}{f_m}, e_m = p_m t_m^{\text{o}}, m \in \{2n-1, 2n\}. \tag{8}$$

Notice that the two users $2n-1$ and $2n$ have different energy consumption expressions in terms of power allocation. For a better structural interpretation, we consider the two users in subchannel n as a group with total energy consumption

$$E_n(p_{2n-1}, p_{2n}) = w_{2n-1} p_{2n-1} \frac{d_{2n-1}}{r_{2n-1}} + w_{2n} p_{2n} \frac{d_{2n}}{r_{2n}}, \tag{9}$$

where r_{2n-1} and r_{2n} are functions of p_{2n-1} and p_{2n} as shown in (7). Minimizing the total energy consumption of all users is equivalent to minimizing that of all groups in all subchannels. The resource allocation problem can then be written as

$$\begin{aligned}
(\text{P2}) \quad \min_{p_m, f_m} \quad & E = \sum\nolimits_{n=1}^{N} E_n(p_{2n-1}, p_{2n}), \\
\text{s.t.} \quad & t_m^{\text{o}} + t_m^{\text{e}} \le \tau, \forall m \in \mathcal{M}, && \text{(10a)} \\
& 0 < p_m \le P_m, f_m > 0, \forall m \in \mathcal{M}, && \text{(10b)} \\
& \sum\nolimits_{m=1}^{M} f_m \le F, && \text{(10c)} \\
& \text{constraints} \quad (7) - (9).
\end{aligned}$$

We now derive the following lemma.

Lemma 1. *Constraints* (10a) *are binding for all* m *under the optimal power and computation resource allocation in (P2).*

Proof. The main idea is to prove that the energy consumption of each user is monotonically increasing with its transmission power, which is shown by analyzing the first order derivative of energy consumption with respect to power allocation. Therefore, users prefer to transmit with minimal transmission power, leading to the longest allowable transmission time. Finally, under the optimal conditions, the latency requirements will be binding regardless of the computation resource allocation. The complete proof is available at [16]. ∎

Under the binding conditions, the power allocation decisions p_{2n-1} and p_{2n}, as well as the transmission rates r_{2n-1} and r_{2n}, can be uniquely determined by computation resource allocation decisions. To show this, we rewrite the power allocation decisions as functions of transmission rates according to (7)

$$p_m = \begin{cases} \frac{\sigma^2}{|h_{2n-1}|^2} e^{a r_{2n}} (e^{a r_{2n-1}} - 1), & m = 2n-1, \\ \frac{\sigma^2}{|h_{2n}|^2} (e^{a r_{2n}} - 1), & m = 2n, \end{cases} \tag{11}$$

where $a := \frac{N \ln 2}{B}$. The relationship between the transmission rates and the computation resource allocation decisions can be denoted as

$$r_m = \frac{d_m}{\tau_m - d_m C_m / f_m}. \tag{12}$$

By slightly abusing the notations, we redefine the energy consumption of the two users in channel n as $E_n(p_{2n-1}, p_{2n}) := E_n(f_{2n-1}, f_{2n})$ and reformulate (P2) as

$$\begin{aligned}
(\text{P3}) \quad \min_{f_{2n-1}, f_{2n}} \quad & E = \sum\nolimits_{n=1}^{N} E_n(f_{2n-1}, f_{2n}), \\
\text{s.t.} \quad & \sum\nolimits_{n=1}^{N} (f_{2n-1} + f_{2n}) \le F, && \text{(13a)} \\
& (f_{2n-1}, f_{2n}) \in \mathcal{F}_n, && \text{(13b)}
\end{aligned}$$

where $\mathcal{F}_n := \{(f_{2n-1}, f_{2n}) \mid f_{2n-1}, f_{2n} > 0; r_{2n-1}(f_{2n-1}), r_{2n}(f_{2n}) > 0; 0 < p_{2n-1}(f_{2n-1}, f_{2n}) \le P_{2n-1}, 0 < p_{2n}(f_{2n}) \le P_{2n}\}$ defines the feasible set of (f_{2n-1}, f_{2n}). Note that only the computation resource allocation is optimized in (P3), and this facilitates the development of the following theorem.

Theorem 1. *(P3) is a convex problem.*

Proof. The main effort is to prove that the constraints of (P3) define a convex set of (f_{2n-1}, f_{2n}) and the objective is convex. The convexity of the feasible region is proved by showing that the feasible set of (f_{2n-1}, f_{2n}) is an intersection of the convex sets derived from all constraints. Furthermore, the convexity of the objective is proved by showing that the Hessian matrix with respect to (f_{2n-1}, f_{2n}) is positive semidefinite. This is done by taking advantage of the positive semidefiniteness of the Hessian matrix of energy consumption with respect to (r_{2n-1}, r_{2n}) and the relationship between the data rate the computation resource allocation. The detailed proof can be found in [16]. ∎

Since there is a coupling constraint (13a), we apply dual decomposition methods to solve the computation resource allocation problem [17]. Based on the convexity of (P3), strong duality holds between the primal and dual problems. Therefore, given a channel assignment, we can efficiently obtain the optimal computation resource allocation. The corresponding power allocation decisions can also be derived.

B. Channel Assignment

In this subsection, we propose a heuristic algorithm to solve the channel assignment problem given the optimal computation resource allocation f_m^*. Note that f_m^* is the computation resource allocated to subchannel position m, no matter which user is assigned to this position. Therefore, different channel assignment decisions correspond to different power allocation decisions in the same subchannel position. We need to decide the channel assignment and the corresponding power allocation together given the computation resource allocation. In the following part of this subsection, we use $r_{k,m}$, $e_{k,m}$ to further specify the transmission data rate and the energy consumption of user k in subchannel position m.

Given the computation resource allocation f_m^*, we can represent the transmission data rate $r_{k,m}$ as

$$r_{k,m} = \frac{d_k}{\tau - d_k C_k / f_m^*}. \tag{14}$$

Based on (8) and (11), when user k is assigned to subchannel position m, the energy consumption can be expressed as

$$e_{k,m} = \begin{cases} \zeta_{k,m} e^{a \sum_{j \in \mathcal{K}_{k,m}} x_{j,m+1} r_{j,m+1}}, & m \in \mathcal{M}^l, \\ \zeta_{k,m}, & m \in \mathcal{M}^s. \end{cases} \tag{15}$$

$\mathcal{K}_{k,m} := \left\{ j \mid |h_{j,m+1}|^2 \le |h_{k,m}|^2, \forall j \in \mathcal{K} \right\}$ denotes all the users whose channel gain in subchannel position $m+1$ is no more than $h_{k,m}$, and $\zeta_{k,m}$ is defined as

$$\zeta_{k,m} = w_k \frac{\sigma^2 d_k}{|h_{k,m}|^2} \frac{e^{a r_{k,m}} - 1}{r_{k,m}}. \tag{16}$$

In (15), $\sum_{j \in \mathcal{K}_{k,m}} x_{j,m+1} r_{j,m+1}$ represents the transmission data rate of user j who shares the same subchannel with user k but has a small channel gain. Thus, we propose to approximate $\sum_{j \in \mathcal{K}_{k,m}} x_{j,m+1} r_{j,m+1}$ by $\sum_{j \in \mathcal{K}_{k,m}} r_{j,m+1} / |\mathcal{K}_{k,m}|$, which is the average transmission rate of all other users who have no larger channel gains on subchannel position $m+1$ than user k. The purpose of performing the approximation is to eliminate the dependence of energy consumption between the subchannel positions m and $m+1$. Define

$$\eta_{k,m} := \begin{cases} e^{a \sum_{j \in \mathcal{K}_{k,m}} r_{j,m+1} / |\mathcal{K}_{k,m}|}, & m \in \mathcal{M}^l, \mathcal{K}_{k,m} \neq \emptyset, \\ L_\eta, & m \in \mathcal{M}^l, \mathcal{K}_{k,m} = \emptyset, \\ 1, & m \in \mathcal{M}^s. \end{cases} \tag{17}$$

Note that if $|h_{k,m}|^2 < |h_{j,m+1}|^2, \forall j \in \mathcal{K}$, we have $|\mathcal{K}_{k,m}| = 0$. If user k is assigned to subchannel position $m \in \mathcal{M}^l$, user k should have no smaller channel gain than the user on the same subchannel. Thus, this is an infeasible situation. Under this condition, we set $\eta_{k,m} = L_\eta$ as a large enough number to prevent assigning user k to subchannel position m. Given the current optimal computation resource allocation f_m^*, we update the channel assignment by solving the following problem

$$\text{(P4)} \quad \min_{\boldsymbol{x}} \quad \tilde{E} = \sum\nolimits_{m=1}^{M} \sum\nolimits_{k=1}^{K} x_{k,m} \zeta_{k,m} \eta_{k,m}, \quad s.t. \quad x \in \mathcal{X}. \tag{18a}$$

The objective $\tilde{E}$ of (P4) is an approximation of the total energy consumption of all users. We relax the constraint $\sum_{k \in \mathcal{K}} x_{k,m} |h_{k,m}|^2 \ge \sum_{k \in \mathcal{K}} x_{k,m+1} |h_{k,m+1}|^2, \forall m \in \mathcal{M}^l$ in $\mathcal{X}$ since we have considered the channel gain requirements in approximation in (17). Note that the constraint coefficients of the relaxed problem (P4) form a totally unimodular matrix [18]. We can efficiently derive the new assignment by further relaxing constraint $x_{k,m} \in \{0, 1\}$ to $0 \le x_{k,m} \le 1$ and solving the resulting linear program.

We summarize the overall algorithm to compute the energy-efficient resource allocation and channel assignment in Algorithm 1, where steps (7)-(11) are to compensate for the relaxation of constraint $\sum_{k \in \mathcal{K}} x_{k,m} |h_{k,m}|^2 \ge \sum_{k \in \mathcal{K}} x_{k,m+1} |h_{k,m+1}|^2, \forall m \in \mathcal{M}^l$ when solving (P4). In addition, steps (12)-(14) are designed to avoid repeatedly looping through the same sequence of channel assignments. In other words, if the newly determined channel assignment has already been considered previously, the algorithm will randomly choose another channel assignment that is not considered before .

IV. Simulation Results and Performance Analysis

In this section, we show the performance of our proposed algorithm and analyze the influence of the NOMA scheme on users' energy consumption in MEC systems.

In the simulation, each user has a random data size $d_k \in [2, 5] \times 10^5$ bits and a random distance to the BS $g_k \in [200, 500]$ m. Both the data size and distance are uniformly distributed. Without loss of generality, all P_k is set to be 30dBm, w_k is 1 and C_k is 10^3 cycles/bit. The BS has the total computation resource $F = 100 \times 10^9$ cycles/s [2]. The noise power spectrum density (PSD) at the BS receiver is $N_0 = -174$ dBm/Hz. Other system settings are as follows unless expressly stated: the length of time slot $\tau = 0.3$ s and the total bandwidth $B = 10$ MHz. We consider the path loss

Algorithm 1 Overall Algorithm

Input: Number of maximum iterations I, system parameters K, N, B, τ, base station information F, user information $w_k, d_k, C_k, P_k, \forall k \in \mathcal{K}$, channel information $h_{k,m}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}$.
Output: The optimal channel assignment $\boldsymbol{x}^*$, computation resource allocation $\boldsymbol{f}^*$, power allocation $\boldsymbol{p}^*$, and minimal energy consumption E^*.
Initialization: Randomly choose a channel assignment $\boldsymbol{x}^{(0)} \in \mathcal{X}$; define the set of channel assignments already considered as $\mathcal{X}_r = \boldsymbol{x}^{(0)}$; set E^* to some arbitrarily large value.
for $i = 1 : I$ **do**
  Calculate the optimal energy consumption $E^{(i)}$, computation resource allocation $\boldsymbol{f}^{(i)}$ and power resource allocation $\boldsymbol{p}^{(i)}$ by solving (P3), given the channel assignment $\boldsymbol{x}^{(i-1)}$;
  Based on $\boldsymbol{f}^{(i)}$, update the channel assignment to $\boldsymbol{x}^{(i)}$ by solving (P4);
  for $m \in \mathcal{M}^l$ **do**
    if $\sum_{k\in\mathcal{K}} x^{(i)}_{k,m}|h_{k,m}|^2 < \sum_{k\in\mathcal{K}} x^{(i)}_{k,m+1}|h_{k,m+1}|^2$ **then**
      find k, j that satisfy $x^{(i)}_{k,m} = 1, x^{(i)}_{j,m+1} = 1$ and set $x^{(i)}_{k,m} = 0, x^{(i)}_{k,m+1} = 1, x^{(i)}_{j,m} = 1, x^{(i)}_{j,m+1} = 0$;
    end if
  end for
  if $\boldsymbol{x}^{(i)} \in \mathcal{X}_r$ **then**
    Randomly choose a channel assignment decision $\boldsymbol{x}^{(i)} \in \mathcal{X} \backslash \mathcal{X}_r$
  end if
  $\mathcal{X}_r = \mathcal{X}_r \bigcup \boldsymbol{x}^{(i)}$
  if $E^{(i)} < E^*$ **then**
    $\boldsymbol{x}^* = \boldsymbol{x}^{(i-1)}, \boldsymbol{f}^* = \boldsymbol{f}^{(i)}, \boldsymbol{p}^* = \boldsymbol{p}^{(i)}, E^* = E^{(i)}$;
  end if
end for

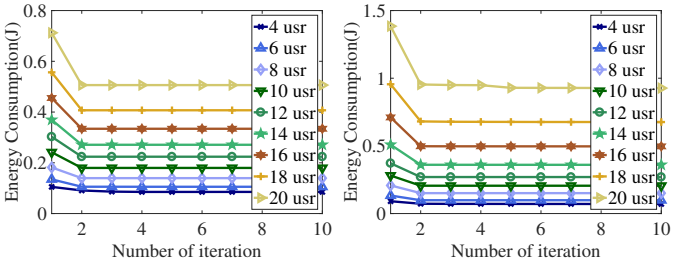
as $\sqrt{G_0\left(\frac{g_k}{d_0}\right)^{-\theta}}$, where $G_0 = -40$dB corresponds to the path loss at a reference distance of $d_0 = 1$m, and $\theta = 3.7$ is the path loss exponent [13].

We denote by NOMA-H our NOMA-based MEC system with the optimal power and computation resource allocation, and the heuristic channel assignment. Apart from NOMA-H, we consider the following three schemes for comparison: 1) NOMA-R: Users adopt NOMA technology to transmit data but are randomly assigned to subchannels with assignment decision $\boldsymbol{x} \in \mathcal{X}$. After assignment, users' power and computation resource allocation are optimized. 2) FDM-R: The total bandwidth is divided into N orthogonal subchannels and each subchannel can only hold one user, where $K = N$ is assumed for simplicity. Each user is randomly assigned to one subchannel and then the power and computation resource allocation are optimized. 3) FDM-H: As in FDM-R, each user transmits over one subchannel individually. For FDM-based MEC systems, the total energy consumption is

$$E = \sum_{n=1}^{N} E_n = \sum_{n=1}^{N}\sum_{k\in\mathcal{K}} x_{k,n} w_k \frac{\sigma^2 d_k}{|h_{k,n}|^2} \frac{e^{ar_k} - 1}{r_k} \tag{19}$$

We follow the method in Algorithm 1 to decide the optimal power and computation resource allocation and the heuristic channel assignment. All the optimization processes involved in three schemes are similar to those in Section III and thus are omitted for brevity. Related NOMA-based and FDM-based systems have the same settings including the total bandwidth, the total computation resource, the length of time slot, users' data sizes and distances to the BS except for channel states. We set the maximum number of iteration $I = 10$ and $L_\eta = 100$. For each setting, we run 100 times with different Rayleigh fading channels and take the average.

First, to demonstrate the performance of our heuristic channel assignment algorithm, we show the dynamics of users' total energy consumption of NOMA-H and FDM-H in Fig. 1. It can be seen that our algorithm efficiently finds a good



(a) NOMA-H System (b) FDM-H System

Fig. 1. Dynamics of Users' Energy Consumption

solution. When brute force in channel assignment is tractable ($K \leq 8$ in NOMA-based systems and $K \leq 6$ in FDM-based systems), we compare the results of our NOMA-H and FDM-H algorithms with those of brute force and observe that our algorithm exactly finds the optimal solutions as brute force. (The brute force of NOMA-based and FDM-based systems are named as NOMA-B and FDM-B, respectively. The energy consumption of both schemes is shown in Fig. 2.) When brute force is not viable, we can still see a considerable energy reduction achieved by the channel assignment. Consider $K = 20$ as an example, NOMA-H (i.e., $I = 10$) consumes 71% energy of NOMA-R (i.e., $I = 1$). Similarly, FDM-H consumes 67% energy of FDM-R.

Next, we evaluate the impact of NOMA on energy-efficient offloading in MEC systems by comparing the four schemes. The energy consumption curves are plotted in Fig. 2. Several important observations can be made: 1) When an MEC system has a small number of users (e.g., $K \leq 6$), an FDM-based system consumes less energy than a NOMA-based system. That is because users in the FDM-based systems can still have enough bandwidth to ensure transmission data rate, while users in the NOMA-based systems have to deal with interference within each subchannel. 2) When an MEC system holds more users (e.g., $K \geq 8$), a NOMA-based system

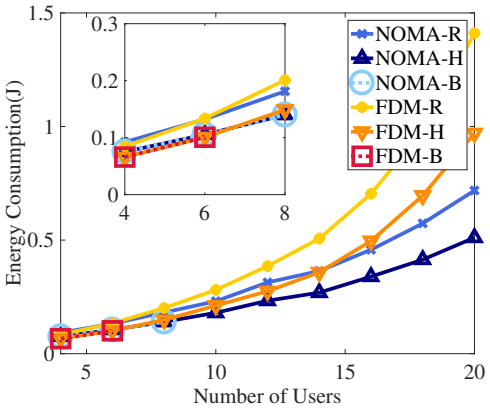


Fig. 2. Energy Consumption of Different Schemes

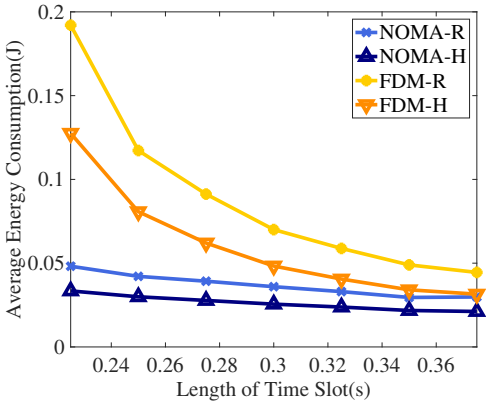


Fig. 3. Average Energy Consumption for Each User when $K = 20$

consumes less energy than an FDM-based system and the energy consumption gap becomes larger with the increase of the number of users. This indicates that under resource-scarce cases, NOMA's more efficient use of bandwidth begins to have positive effect on reducing users' energy consumption. 3) When an MEC system has even more users (e.g., $K \geq 14$), NOMA-R outperforms the FDM-H in terms of users' energy consumption. It shows significant advantages of adopting NOMA scheme to do energy-efficient computation offloading when there are many users in the MEC system.

Finally, we change τ from 0.225s to 0.375s and plot the average energy consumption for each user when $K = 20$ in Fig. 3. When the length of time slot becomes smaller and data sizes keep unchanged, it is equivalent to having more stringent latency requirements. We can observe that NOMA can achieve more energy reduction under latency-intensive cases compared to FDM. Also, NOMA-based MEC systems are less sensitive to the latency requirement compared to FDM-based MEC systems in terms of energy consumption.

V. Conclusion

Energy-efficient computation offloading in NOMA-based MEC systems has been investigated in this paper. We have formulated a non-convex combinatorial problem to jointly optimize the resource allocation (including power and computation resources) and channel assignment with the aim of minimizing the total energy consumption of all users. An algorithm has been devised to solve the problem by dealing with resource allocation and channel assignment subproblems in an iterative way. The algorithm has been demonstrated to be efficient with only a small number of iterations required to produce good performance. Meanwhile, numerical results have shown that compared to FDM-based MEC systems, NOMA-based MEC systems have great advantages in reducing energy consumption when there are multiple users and strict latency requirements.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2017.

[3] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing-A key technology towards 5G," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[4] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[5] Y. Wu, L. P. Qian, H. Mao, X. Yang, H. Zhou, X. Tan, and D. H. K. Tsang, "Secrecy-driven resource management for vehicular computation offloading networks," *IEEE Network*, vol. 32, no. 3, pp. 84–91, 2018.

[6] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2049–2063, 2018.

[7] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory," in *Global Communications Conference (GLOBECOM), 2015 IEEE*. IEEE, 2015, pp. 1–6.

[8] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7244–7257, 2016.

[9] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *arXiv preprint arXiv:1807.06810*, 2018.

[10] Y. Wu, K. Ni, C. Zhang, L. Qian, and D. H. K. Tsang, "NOMA assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Transactions on Vehicular Technology*.

[11] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA-MEC offloading," *arXiv preprint arXiv:1807.06306*, 2018.

[12] Z. Yang, J. Hou, and M. Shikh-Bahaei, "Energy efficient resource allocation for mobile-edge computation networks with NOMA," *arXiv preprint arXiv:1809.01084*, 2018.

[13] F. Wang, J. Xu, and Z. Ding, "Optimized multiuser computation offloading with multi-antenna NOMA," in *Globecom Workshops (GC Wkshps), 2017 IEEE*. IEEE, 2017, pp. 1–7.

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

[15] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *arXiv preprint arXiv:1804.06712*, 2018.

[16] L. Liu, B. Sun, X. Tan, Y. Xiao, and D. H. K. Tsang, "Energy-efficient resource allocation and channel assignment for NOMA-based MEC: Complete Proofs of Lemma 1 and Theorem 1," http://c2e.ece.ust.hk/linaliu/proofs.pdf.

[17] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[18] R. Meyer, "A class of nonlinear integer programs solvable by a single linear program," *SIAM Journal on Control and Optimization*, vol. 15, no. 6, pp. 935–946, 1977.