IEEE Transactions on Wireless Communications

Resource Allocation for Hybrid NOMA MEC Offloading

Journal:	*IEEE Transactions on Wireless Communications*
Manuscript ID	Paper-TW-Aug-19-1018.R1
Manuscript Type:	Original Transactions Paper
Date Submitted by the Author:	02-Jan-2020
Complete List of Authors:	Zhu, Jianyue Wang, Jiaheng; Southeast University, National Mobile Communications Research Laboratory Huang, Yongming; School of Information Science and Engineering, Southeast University Fang, Fang; The University of Manchester, Navaie, Keivan; Lancaster University, School of Computing and Communications Ding, Zhiguo; The University of Manchester
Keyword:	

Resource Allocation for Hybrid NOMA MEC Offloading

Jianyue Zhu, Jiaheng Wang, Yongming Huang, Fang Fang, Keivan Navaie, Zhiguo Ding

Abstract

Non-orthogonal multiple access (NOMA) and mobile edge computing (MEC) have been recognized as promising technologies for the beyond fifth generation networks to achieve significant capacity improvement and delay reduction. In this paper, the technologies of hybrid NOMA and MEC are integrated. In the hybrid NOMA MEC system, multiple users are classified into different groups and each group is allocated a dedicated time slot. In each group, a user first offloads its task by sharing a time slot with another user, and then solely offloads during a time interval. To reduce the delay and save the energy consumption, we consider jointly optimizing the power and time allocation in each group as well as the user grouping. As the main contribution, the optimal power and time allocation is characterized in closed form. In addition, by incorporating the matching algorithm with the optimal power and time allocation, we propose a low complexity method to efficiently optimize user grouping. Simulation results demonstrate that the proposed resource allocation method in the hybrid NOMA MEC systems not only yields better performance than the conventional OMA scheme but also achieves quite close performance as global optimal solution.

Index Terms

Non-orthogonal multiple access, mobile edge computing, time delay, energy consumption, user grouping, resource allocation

This paper in part has been accepted in IEEE Global Communication Conference, Waikoloa, HI, USA, Dec. 2019. [1]

J. Zhu, J. Wang, and Y. Huang are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. (email: {zhujy, jhwang, huangym}@seu.edu.cn).

F. Fang and Z. Ding are with the School of Electrical and Electronic Engineering, Manchester University, Manchester, UK (email: {fang.fang, zhiguo.ding}@manchester.ac.uk).

K. Navaie is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, United Kingdom (email: k.navaie@lancaster.ac.uk).

I. Introduction

With the development of Internet-of-Things (IoT) and wireless networks, the beyond fifth generation (B5G) communication systems impose an explosive demand of data traffic. In order to offer significant improvements of network capacity, the B5G wireless networks require spectral efficient multiple access techniques [2]. Recently, it is shown that nonorthogonal multiple access (NOMA) can support overloaded transmission and improve the spectral efficiency. Therefore, the technique of NOMA has been recognized as one of the key technologies in the upcoming B5G wireless networks [3].

Conventionally, the orthogonal multiple access (OMA) schemes are not able to support large wireless network capacity because orthogonal resources are allocated to different users [4]. However, in NOMA systems, one resource (e.g., frequency, time, code, or spatial) unit channel can be allocated to multiple users at the same time [5], which leads to better spectral efficiency than the OMA scheme [6], [7]. In [8] and [9], the authors discussed the application of NOMA in multiple-input multiple-output (MIMO) systems. In addition, NOMA has also been proposed to be incorporated into other technologies such as visible light communication [10], wireless caching [11], and millimeter wave communication [12].

Recently, there has yielded a variety of computation-hungry applications, e.g., virtual reality, [13], which makes mobile networks computationally constrained [14]. Nevertheless, most mobile users have limited computation and power resources, i.e., if the mobile users complete intensive tasks locally, the batteries will be drained quickly and the users might not be able to complete the tasks within their deadlines. To address this issue, mobile edge computing (MEC) is introduced as one of the key emerging technologies for B5G networks [15], [16]. The main idea of MEC is to employ more resourceful computing facilities at the edge of mobile networks. Then the users are able to offload their computationally intensive tasks to the MEC. In the literature, there are many works focusing on the technique of MEC. For instance, in [17], in order to improve the energy efficiency for latency-constrained computation, the authors proposed a user scheduling scheme to achieve a better performance in terms of the reliability and latency for task offloading. In [18], the authors proposed a user scheduling scheme to achieve a better performance in terms of the reliability and latency for task offloading.

Integrating MEC and NOMA, it is shown in [19] and [20] that we can not only avoid sever delay but also reduce energy consumption. Moreover, in [21], the authors studied the application

of uplink NOMA and downlink NOMA in MEC systems. The authors developed analytical results to depict that the use of NOMA can efficiently reduce the delay and energy consumption for MEC offloading. Therefore, the combination of NOMA and MEC is another important communication technique in future wireless networks, which has received much attention recently. In [22], the authors considered an MEC system exploiting the NOMA for both task uploading and result downloading, where the transmit powers, transmission time allocation, and task offloading were optimized to minimize total energy consumption. Furthermore, [23] minimized the overall delay of the users by jointly optimizing the users' offloaded workload and the NOMA transmission time. Multi-antenna NOMA was also applied in multiuser MEC systems in [24], where the authors considered both cases with partial and binary offloading.

Note that, a lot of resources, e.g., time and power, are needed for the process of offloading. Hence, the optimization of resource for offloading is a key problem in NOMA MEC systems, which has attracted a lot of interests such as [19], [20], [22]–[24]. Most of the existing works, e.g., [19], [20], [22]–[24], only considered two offloading strategies, which are respectively OMA and pure NOMA. Here, pure NOMA means both users share the same time to offload all the task. Actually, there is a third strategy, which has been termed as hybrid NOMA in [25], [26]. In the hybrid NOMA scheme, a user firstly offloads parts of its task by sharing a time slot allocated to another user, and then solely offloads the remaining task during a time interval. The hybrid NOMA MEC not only outperforms OMA in terms of delay but also achieves lower energy consumption than NOMA. Practically, by using the hybrid NOMA MEC offloading scheme, the resources of time and energy can be saved for the users with different deadlines.

In addition, it is worth pointing out that both energy consumption and delay are important performance measures in communication systems. In order to achieve a tradeoff between energy consumption and delay, we investigate the resource allocation for minimizing the weighted sum of energy consumption and delay (WSED) in hybrid NOMA MEC systems. Moreover, we apply the hybrid NOMA scheme to multiple users case, where multiple users are classified into different groups and each group is allocated a dedicated time slot. However, the existing works [25], [26] just considered limited number of users, i.e., two users case was studied.

Overall, in this paper, we focus on the resource allocation for minimizing the WSED for multiple users in hybrid NOMA MEC systems, which is actually the joint optimization of power, time, and user grouping. The contributions in this paper are summarized as follows:

- We consider minimizing the WSED with rate and deadline constraints in hybrid NOMA

MEC systems, where multiple users transmit through multiple groups and each group occupies a dedicated time slot. This is the original work in the literature.

- With given user grouping, we analyze the performance of three strategies, i.e., OMA, pure NOMA, and hybrid NOMA, where closed-form solutions for the optimal power and time allocation are characterized.
- The obtained closed-form solutions provide significant quantitative insights on the properties of hybrid NOMA MEC offloading. For instance, it is proved that hybrid NOMA MEC can be superior to OMA MEC in the cases where users have demanding delay requirements for their task offloading. But if the user has a delay tolerant task, OMA MEC is preferred.
- By using the closed-form solutions, we further provide an efficient algorithm via matching to deal with the user grouping. The proposed closed-form time and power allocation even reduces the complexity of the exhaustive search for multiple users through multiple groups.

The rest of the paper is organized as follows. Section II introduces the hybrid NOMA MEC system model and the formulated optimization problem for minimizing WSED. In Section III, we investigate the optimal power and time allocation. In Section IV, we propose an efficient user grouping algorithm. The simulation results of the proposed resource allocation are evaluated in section V. In Section VI, we conclude the paper.

II. System Model and Problem Formulation

A. System Model

We consider an MEC offloading scenario, wherein the base station (BS) equipped with an MEC server serves N users with different delay and task requirements. The MEC server can serves users in different groups and each group occupies a dedicated time slot. It is also assumed that each time slot can be simultaneously occupied by multiple users and hence, these N users are divided into L pairs. Let $N_l \in \{N_1, \cdots, N_L\}$ be the number of users in group l for $l = 1, \cdots, L$ and $\text{UE}_{n,l}$ denotes user n in group l for $l = 1, \cdots, N_l$. Since the computational capabilities of these users are limited, the users are assumed to offload their tasks, which is computationally intensive, timely, and inseparable, to the server.

Let $M_{n,l}$ and $D_{n,l}$ for $n = 1, \cdots, N_l$ respectively denote the number of bits contained in $\text{UE}_{n,l}$'s task and the computation deadline of $\text{UE}_{n,l}$'s task. Without loss of generality, assume that $M_{n,l} = M$, for $n = 1, \cdots, N_l,\ l = 1, \cdots, L$, and in each group, the users are ordered

according to their computation deadlines, i.e., $D_{1,l} \leq D_{2,l} \leq \cdots \leq D_{N_l,l}$. Hence, $\mathrm{UE}_{1,l}$ has the most demanding deadline and $\mathrm{UE}_{N_l,l}$ has the least demanding deadline.

In NOMA systems, using SIC at the receiver causes additional complexity, which is proportional to the number of users performing NOMA [27], [28]. Thus, in practice, it is often assumed that two users are paired to perform NOMA and this assumption is implemented in LTE-A [29]. In this paper, we also focus on this typical situation. In each group l, the MEC server schedules only two users, i.e., $\mathrm{UE}_{1,l}$ and $\mathrm{UE}_{2,l}$, to be served at the same time slot.

In order to better illustrate the benefit of NOMA, we should first introduce OMA MEC. In OMA MEC systems, each user is allocated a dedicated time slot for offloading. In each group l, according to our assumption that $D_{1,l} \leq D_{2,l}$, $\mathrm{UE}_{1,l}$ is served first. Therefore:

$$D_{1,l} B \ln\left(1 + p_{1,l}^{OMA} |h_{1,l}|^2\right) = M, \tag{1}$$

$$t_l B \ln\left(1 + p_{2,l}^{OMA} |h_{2,l}|^2\right) = M, \tag{2}$$

where t_l, satisfying $0 \leq t_l \leq D_{2,l} - D_{1,l}$, is the time interval solely occupied by $\mathrm{UE}_{2,l}$ and $p_{n,l}^{OMA}$, $n = 1, 2$ denotes the transmit power of $\mathrm{UE}_{n,l}$. In addition, B is the bandwidth, $h_{n,l} = g_{n,l} d_{n,l}^{-\nu} / \sigma_{n,l}^2$ is the channel to noise ratio from the BS to $\mathrm{UE}_{n,l}$, where $g_{n,l}$ follows a Rayleigh distribution, $d_{n,l}$ is the distance between $\mathrm{UE}_{n,l}$ and the BS, ν is the path-loss exponent, and $\sigma_{n,l}^2$ is the variance of the additive white Gaussian noise (AWGN).

In NOMA MEC systems, in group l, the NOMA principle allows two users to simultaneously offload their tasks to the server during $D_{1,l}$ to the server. Here, it is worth pointing out that $\mathrm{UE}_{1,l}$ achieves the same performance as in OMA if the message of $\mathrm{UE}_{2,l}$ is decoded first. This is because, by exploiting SIC, the message of $\mathrm{UE}_{1,l}$ can be decoded by removing $\mathrm{UE}_{2,l}$'s message, which also implies the data rate of $\mathrm{UE}_{2,l}$ during $D_{1,l}$ is constrained as

$$R_{2,l} \leq B \ln\left(1 + \frac{p_{2,l}^1 |h_{2,l}|^2}{p_{1,l}^{OMA} |h_{1,l}|^2 + 1}\right), \tag{3}$$

where $p_{2,l}^1$ denotes the power used by $\mathrm{UE}_{2,l}$ during $D_{1,l}$. Actually, (3) is to ensure that the implementation of NOMA is transparent to $\mathrm{UE}_{1,l}$ [26].

In this paper, we will consider hybrid NOMA MEC in that [21] has pointed that in group l, $\mathrm{UE}_{2,l}$ needs to consume more energy in NOMA than in OMA if $\mathrm{UE}_{2,l}$ completely relies on $D_{1,l}$. The time sharing scheme for hybrid NOMA strategy is shown in Fig. 1, wherein $\mathrm{UE}_{2,l}$ shares $D_{1,l}$ with $\mathrm{UE}_{1,l}$ and then continuously transmits for another time interval after $D_{1,l}$, which is denoted by t_l. In addition, the power used by $\mathrm{UE}_{2,l}$ during t_l is denoted by $p_{2,l}^2$.

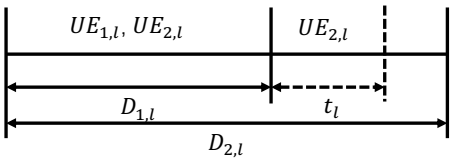

Figure 1. Time sharing scheme for hybrid NOMA strategy.

B. Problem Formulation

In this paper, we investigate the energy consumption and delay of the hybrid NOMA MEC systems. Similar to [26], the time cost for the server to send the outcomes of the task to the users and compute the tasks is omitted, which is negligibly small compared to the considered offloading costs. In addition, considering the server is not energy constrained, the energy consumption at the server is also ignored. Therefore, we consider only the energy consumption and delay of users for the task offloading process for each group l, $l = 1, \cdots, L$, which are respectively given by

$$E_l = D_{1,l} p_{1,l}^{OMA} + D_{1,l} p_{2,l}^{1} + t_l p_{2,l}^{2}, \tag{4}$$

$$D_l = D_{1,l} + t_l. \tag{5}$$

Note that both energy consumption and delay are necessary to be considered in the process of offloading. Similar to [30], [31], in each group l, the non-negative weight factors α_l and β_l are introduced to tradeoff the energy consumption and delay. Therefore, in group l, the weighted sum of energy consumption and delay (WSED) in hybrid NOMA MEC systems is given by

$$C_l = \alpha_l E_l + \beta_l D_l, \tag{6}$$

where α_l and β_l are two weight factors which indicate the weights of energy consumption and delay. For $l = 1, \cdots, L$, we set $0 \leq \alpha_l, \beta_l \leq 1$ and $\alpha_l + \beta_l = 1$. In order to meet the specific demands of users, different users are allowed to choose different weight factors. For example, if a user is in a low battery state, to save more energy, it would choose a larger α_l, i.e., put more weight on the energy consumption. Similarly, for cases when a user is running a delay sensitive application, to reduce the latency, the user would choose a larger β_l, i.e., put more emphasise on the time delay.

Note that $D_{1,l}$ and $p_{1,l}^{OMA}$ are both constants, we can optimize the resource allocation, i.e., the joint optimization of power and time, by simplifying C_l as

$$C_l = \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2 \right) + \beta_l t_l. \tag{7}$$

In fact, the simplified WSED in (7) represents $\text{UE}_{2,l}$'s performance, which is because $\text{UE}_{1,l}$ experiences the same performance as OMA. Therefore, in this paper, in each group l, we focus on the performance of $\text{UE}_{2,l}$, and the resource allocation problem for minimizing WSED in hybrid NOMA MEC systems is formulated as the following

$$\min \sum_{l=1}^{L} \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2 \right) + \beta_l t_l, \tag{8}$$

$$\text{s.t. } D_{1,l} B \ln \left(1 + \frac{|h_{2,l}|^2 p_{2,l}^1}{|h_{1,l}|^2 p_{1,l}^{OMA} + 1} \right) + t_l B \ln \left(1 + |h_{2,l}|^2 p_{2,l}^2 \right) \geq M, l = 1, \cdots, L, \tag{9}$$

$$0 \leq t_l \leq D_{2,l} - D_{1,l}, l = 1, \cdots, L, \tag{10}$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0, l = 1, \cdots, L, \tag{11}$$

where constraint (9) indicates the rate constraint to guarantee that $\text{UE}_{2,l}$'s M bits are offloaded before $D_{1,l} + t_l$ and (10) is the deadline constraint of $\text{UE}_{2,l}$, i.e., $t_l + D_{1,l} \leq D_{2,l}$.

The resource allocation problem in hybrid NOMA MEC systems is a joint optimization of power allocation, time allocation, and user grouping, which is a difficult mixed integer problem. To solve this problem efficiently, we will treat user grouping, power and time allocation separately. Specifically, assuming the user grouping is given, we first find the optimal power and time allocation for the users in each group, which is even characterized in closed form. Then, using the proposed optimal power and time allocation, we exploit the matching theory to optimize the user grouping. The proposed solution will improve the system performance and dramatically simplify the resource allocation.

III. Optimal Power and Time Allocation

In this section, assuming the user grouping is given, we focus on the optimization problem of WSED minimization, which is given by

$$\min_{\left\{p_{2,l}^1, p_{2,l}^2, t_l\right\}_{l=1}^L} \sum_{l=1}^{L} \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2\right) + \beta_l t_l, \tag{12}$$

$$\text{s.t. } D_{1,l} B \ln\left(1 + \frac{\left|h_{2,l}\right|^2 p_{2,l}^1}{\left|h_{1,l}\right|^2 p_{1,l}^{OMA} + 1}\right),$$

$$+ t_l B \ln\left(1 + \left|h_{2,l}\right|^2 p_{2,l}^2\right) \geq M, l = 1, \cdots, L, \tag{13}$$

$$0 \leq t_l \leq D_{2,l} - D_{1,l}, l = 1, \cdots, L, \tag{14}$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0, l = 1, \cdots, L, \tag{15}$$

which is a nonconvex problem. Note that problem (12) can be decoupled into a series of subproblems and for each group, we have

$$\min_{p_{2,l}^1, p_{2,l}^2, t_l} \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2\right) + \beta_l t_l, \tag{16}$$

$$\text{s.t. } D_{1,l} B \ln\left(1 + \frac{\left|h_{2,l}\right|^2 p_{2,l}^1}{\left|h_{1,l}\right|^2 p_{1,l}^{OMA} + 1}\right),$$

$$+ t_l B \ln\left(1 + \left|h_{2,l}\right|^2 p_{2,l}^2\right) \geq M, \tag{17}$$

$$0 \leq t_l \leq D_{2,l} - D_{1,l}, \tag{18}$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0, \tag{19}$$

which is also a nonconvex problem. The non-convexity lies on the objective function and constraint (17). In the following, we first achieve the optimal power allocation of users in each group, which can be expressed as functions of the time interval of each group. Then, we further optimize the time intervals and thus obtain the optimal power and time allocation, which can be characterized in a closed form.

A. Optimal Power Allocation for Minimizing WSED

In this subsection, we first optimize the power by fixing the time. Note that for each group, we have $D_{1,l} B \ln\left(1 + p_{1,l}^{OMA} \left|h_{1,l}\right|^2\right) = M$ and hence the WSED minimization problem is given

by

$$\begin{aligned} \min_{p_{2,l}^1, p_{2,l}^2} \quad & \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2 \right), \\ \text{s.t.} \quad & D_{1,l} B \ln \left(1 + |h_{2,l}|^2 p_{2,l}^1 e^{\frac{-M}{D_{1,l} B}} \right) \\ & + t_l B \ln \left(1 + |h_{2,l}|^2 p_{2,l}^2 \right) \geq M, \\ & p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0, \end{aligned} \tag{20}$$

Since t_l is fixed, problem (20) aims to minimize the energy consumption of these two users. In addition, one can easily find that, in problem (20), both the objective function and constraints are convex and hence the optimal solution can be easily obtained by using the standard optimization tools, e.g., CVX. Furthermore, by exploiting the convex problem (20), the closed-form optimal power allocation for problem (8) is provided in the following Theorem.

Theorem 1. *The optimal solution to problem* (20) *is obtained in the following with three cases:*

$$\textit{NOMA}: t_l = 0 \Rightarrow \begin{cases} p_{2,l}^{1*} & = |h_{2,l}|^{-2} \left(e^{\frac{2M}{BD_{1,l}}} - e^{\frac{M}{BD_{1,l}}} \right), \\ p_{2,l}^{2*} & = 0, \end{cases} \tag{21}$$

$$\textit{Hybrid NOMA}: 0 < t_l < D_{1,l} \Rightarrow \begin{cases} p_{2,l}^{1*} & = |h_{2,l}|^{-2} \left(e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - e^{\frac{M}{BD_{1,l}}} \right), \\ p_{2,l}^{2*} & = |h_{2,l}|^{-2} \left(e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1 \right), \end{cases} \tag{22}$$

$$\textit{OMA}: t_l \geq D_{1,l} \Rightarrow \begin{cases} p_{2,l}^{1*} & = 0, \\ p_{2,l}^{2*} & = |h_{2,l}|^{-2} \left(e^{\frac{M}{Bt_l}} - 1 \right). \end{cases} \tag{23}$$

Proof. See Appendix A. □

Remark 1. In Theorem 1, by fixing t_l, the optimal power allocation for problem (8) is characterized in three cases. The first case is the pure NOMA in which $t_l = 0$ indicates that the two users offload their tasks during the same time $D_{1,l}$. The second case is hybrid NOMA case, which is because $p_{2,l}^1$ and $p_{2,l}^2$ are both non-zero. In the third case, we have $p_{2,l}^1 = 0$ and $p_{2,l}^2 > 0$, which is OMA.

Corollary 1. *For problem* (20)*, in the hybrid NOMA case, we always have* $p_{2,l}^{2*} > p_{2,l}^{1*}$.

Proof. See Appendix B. □

According to Corollary 1, in the hybrid NOMA case, $\mathrm{UE}_{2,l}$ is allocated with more power during t_l than $D_{1,l}$, which is in line with our expectation. Actually, in the hybrid NOMA case, $\mathrm{UE}_{2,l}$ experiences no interference during t_l while it is interfered by $\mathrm{UE}_{1,l}$ during $D_{1,l}$. Therefore, $\mathrm{UE}_{2,l}$ allocates a higher power during t_l to achieve a lower energy consumption**.**

Corollary 2. *For problem (20), in each group l, $E_l^{NOMA} \leq E_l^{OMA}$ if and only if $D_{2,l} < 2D_{1,l}$.*

Proof. See Appendix C. □

From Corollary 2, given $D_{2,l} < 2D_{1,l}$, the NOMA scheme achieves a higher performance. In the following Proposition, we will further investigate the superiority of NOMA over OMA.

Proposition 1. *For problem (20), given $D_{2,l} < 2D_{1,l}$, we have*

$$\Delta\left(t_l\right) = E^{OMA} - E^{H-NOMA} = \frac{\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - \left(D_{1,l}e^{\frac{M}{BD_{1,l}}} + t_l e^{\frac{M}{Bt_l}}\right)}{\left|h_{2,l}\right|^2}, \tag{24}$$

which is a monotonically non-increasing function and satisfies $\Delta\left(t_l\right)_{\max} = \Delta\left(D_{2,l} - D_{1,l}\right) < 0$.

Proof. See Appendix D. □

Proposition 1 suggests that for $D_{2,l} < 2D_{1,l}$, the largest gap between hybrid NOMA MEC and OMA MEC is achieved at $t_l = D_{2,l} - D_{1,l}$. Therefore, the optimal strategy is that $\mathrm{UE}_{2,l}$ shall consume all its time until its deadline. ”

Corollary 3. *For problem (16), given $D_{2,l} < 2D_{1,l}$, $\alpha_l = 1$, and $\beta_l = 0$, the optimal time solution is $t_l^* = D_{2,l} - D_{1,l}$.”*

Proof. See Appendix B. □

Remark 2. Given $\alpha_l = 1$, $\beta_l = 0$, WSED minimization is equivalent to minimizing energy consumption. Therefore, to save energy, $\mathrm{UE}_{2,l}$ will consume all its time, i.e., $t_l = D_{2,l} - D_{1,l}$. Compared to the pure NOMA scheme, i.e., $t_l = 0$, the hybrid NOMA expectedly induces less energy consumption.

Note that, in this subsection, we have optimized the power by fixing time t_l, and hence only the energy consumption is optimized. In the following subsection, we further study the optimization of time to achieve the minimum WSED.

B. Optimal Time Allocation for Minimizing WSED

In this subsection, in group l, we focus on optimizing the time requested for $\mathrm{UE}_{2,l}$ to transmit solely. According to Theorem 1, the optimal power allocation is given in three cases. In the following, when further optimizing the time, the cases of pure NOMA and hybrid NOMA are considered together, which are termed as NOMA. Hence, we will respectively characterize the optimal t_l^* in the cases of NOMA and OMA.

1) Optimal Time Allocation for NOMA MEC: From Theorem 1, the case of NOMA corresponds to the condition of $0 \le t_l < D_{1,l}$. Since $0 \le t_l \le D_{2,l} - D_{1,l}$, we have $D_{1,l} > D_{2,l} - D_{1,l}$, i.e., $D_{2,l} < 2D_{1,l}$. Then, given $D_{2,l} < 2D_{1,l}$, by using the optimal powers given in Theorem 1, in each group, the corresponding time optimization problem is:

$$\min_{t_l} C\left(t_l\right), \tag{25}$$

$$\text{s.t. } 0 \le t_l \le D_{2,l} - D_{1,l}, \tag{26}$$

where

$$C\left(t_l\right) = \frac{\alpha_l}{|h_{2,l}|^2}\left(D_{1,l}e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} + t_l\left(e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1\right)\right) + \beta_l t_l. \tag{27}$$

We first show that the convexity of problem (25) in the following Proposition.

Proposition 2. *Problem* (25) *is convex.*

Proof. See Appendix F.

Therefore, the optimal solution to problem (25) can be easily obtained by using standard convex tools, such as interior method. The optimal solution is also given in closed form in the following Proposition. □

Proposition 3. *Given $D_{2,l} < 2D_{1,l}$, the optimal solution to problem* (25) *is:*

$$t_l^* = \begin{cases} 0, & \Omega < 0, \\ \Omega, & 0 \le \Omega \le D_{2,l} - D_{1,l}, \\ D_{2,l} - D_{1,l}, & \Omega > D_{2,l} - D_{1,l}, \end{cases} \tag{28}$$

where

$$\Omega = \frac{2M}{B\left(W_0\left(\frac{\beta_l\left|h_{2,l}\right|^2 - \alpha_l}{e\alpha_l}\right) + 1\right)} - D_{1,l}, \tag{29}$$

and $W_0\left(\frac{\beta_l\left|h_{2,l}\right|^2-\alpha_l}{e\alpha_l}\right)$ *denotes the single-valued Lambert function (satisfying* $W_0\left(\frac{\beta_l\left|h_{2,l}\right|^2-\alpha_l}{e\alpha_l}\right) \geq -1$ *).*

Proof. See Appendix G. □

Remark 3. From Proposition 3, the optimal time allocated to $\mathrm{UE}_{2,l}$ to transmit solely is closely connected with the weight factors of the energy consumption and delay, i.e., α_l and β_l. In the following, we obtain the conditions of the weights for the three cases proposed in Proposition 3.

Corollary 4. *For problem* (25)*, the weight conditions for the three cases are given respectively by C1* : $t_l^* = 0$*, C2* : $t_l^* = \Omega$*, C3* : $t_l^* = D_{2,l} - D_{1,l}$*, where*

$$C1 : W_0\left(\frac{|h_{2,l}|^2}{e}\frac{\beta_l}{\alpha_l} - \frac{1}{e}\right) > \frac{2M}{BD_{1,l}} - 1, \tag{30}$$

$$C2 : \frac{2M}{BD_{2,l}} - 1 \leq W_0\left(\frac{|h_{2,l}|^2}{e}\frac{\beta_l}{\alpha_l} - \frac{1}{e}\right) \leq \frac{2M}{BD_{1,l}} - 1, \tag{31}$$

$$C3 : W_0\left(\frac{|h_{2,l}|^2}{e}\frac{\beta_l}{\alpha_l} - \frac{1}{e}\right) < \frac{2M}{BD_{2,l}} - 1. \tag{32}$$

Proof. The conditions are obtained using Proposition 3 and straight forward calculus. □

Remark 4. From Corollary 4, one can easily find that the value of $W_0\left(\frac{\left|h_{2,l}\right|^2}{e}\frac{\beta_l}{\alpha_l} - \frac{1}{e}\right)$ affects t_l^* and a higher value of $W_0\left(\frac{\left|h_{2,l}\right|^2}{e}\frac{\beta_l}{\alpha_l} - \frac{1}{e}\right)$ is more likely to induce a lower t_l^*. Note that $W_0\left(\frac{\left|h_{2,l}\right|^2}{e}\frac{\beta_l}{\alpha_l} - \frac{1}{e}\right)$ is a monotonically increasing function of $\frac{\beta_l}{\alpha_l}$. Hence, a higher value of $\frac{\beta_l}{\alpha_l}$, i.e., more weight is given to the delay, induces a lower t_l^*, and a lower value of $\frac{\beta_l}{\alpha_l}$, i.e., more weight is occupied by energy consumption, induces, as expected, a larger t_l^*.

Corollary 5. *Given* $D_{2,l} < 2D_{1,l}$ *and C2, or* $D_{2,l} < 2D_{1,l}$ *and C3 , for* (8)*, the hybrid NOMA scheme always yields the best performance in terms of WSED.*

Proof. This can be easily shown using Proposition 3 and Corollary 4. □

Remark 5. As can be seen in Corollary 5, conditions C2 and C3 may correspond to a scenario where $\mathrm{UE}_{2,l}$ is in a low battery state and it would choose a larger α_l. In this case, if $\mathrm{UE}_{2,l}$ completely relies on $D_{1,l}$, $\mathrm{UE}_{2,l}$ may need to consume more energy. Therefore, the hybrid NOMA

scheme is preferred. Moreover, if the weight of energy consumption is large enough, e.g., to save energy, $\mathrm{UE}_{2,l}$ chooses to finish offloading its task at its deadline, i.e., $t_l = D_{2,l} - D_{1,l}$.

Furthermore, if the channel of $\mathrm{UE}_{2,l}$ is weak or the deadline of $\mathrm{UE}_{1,l}$, i.e., $D_{1,l}$, is small, condition C2 and C3 might also be easily satisfied. In other words, $\mathrm{UE}_{2,l}$ is a cell edge user or $\mathrm{UE}_{1,l}$ is running a delay sensitive application, the hybrid NOMA scheme is thus preferred.

2) Optimal Time Allocation for OMA MEC: If $t_l \geq D_{1,l}$, from the optimal power allocation proposed in Theorem 1, the optimal powers are obtained in the OMA case and the corresponding time optimization problem is:

$$\min_{t_l} \frac{\alpha_l}{|h_{2,l}|^2} t_l \left(e^{\frac{M}{Bt_l}} - 1 \right) + \beta_l t_l, \tag{33}$$

$$\text{s.t. } D_{1,l} \leq t_l \leq D_{2,l} - D_{1,l}. \tag{34}$$

One can easily find that problem (33) is feasible if and only if $D_{2,l} \geq 2D_{1,l}$, i.e., the optimal t_l can be found if and only if $D_{2,l} \geq 2D_{1,l}$. This conclusion is consistent with Corollary **??** that OMA performs better than NOMA if and only if $D_{2,l} \geq 2D_{1,l}$. In other words, if $\mathrm{UE}_{2,l}$ has less demanding delay requirements, the conventional OMA scheme induces the minimum WSED.

Hence, we assume $D_{2,l} \geq 2D_{1,l}$ and focus on solving problem (33), whose optimal solution is characterized in the following Proposition.

Proposition 4. *Given* $D_{2,l} \geq 2D_{1,l}$, the optimal solution to (33) is given by:

$$t_l^* = \begin{cases} D_{1,l}, & \Lambda < D_{1,l}, \\ \Lambda, & D_{1,l} \leq \Lambda \leq D_{2,l} - D_{1,l} \\ D_{2,l} - D_{1,l}, & \Lambda > D_{2,l} - D_{1,l}, \end{cases}, \tag{35}$$

where

$$\Lambda = \frac{M}{B\left(W_0\left(\frac{|h_{2,l}|^2}{e} \frac{\beta_l}{\alpha_l} - \frac{1}{e} \right) + 1 \right)}. \tag{36}$$

Proof. See Appendix H. □

In Proposition 4, the condition $D_{2,l} \geq 2D_{1,l}$ indicates that OMA MEC yields a better performance than NOMA MEC and hence the proposed optimal time solution is for the OMA MEC case. Furthermore, in order to achieve the minimum weighted sum of energy consumption and

delay, the optimal time solution is closely connected with the value of $\frac{\beta_l}{\alpha_l}$. Specifically, if α_l is large and β_l is small, the optimal value of t_l will be large. This is in line with our expectation because more weight is given to the energy consumption. Conversely, in the case when α_l is small and β_l is large, i.e., the system focus more on the delay minimization, the optimal value of t_l will be small.

IV. User Grouping via Matching

In the previous sections, the optimal resources of power and time allocation for minimizing WSED are characterized in closed form. Then, the optimal user grouping can be found by, e.g., checking all possible user-group matchings. However, considering the complexity of the exhaustive search, in this section, we study the optimization of user grouping in hybrid NOMA MEC systems. Enlightened by the optimal power and time allocation, we propose an algorithm with low complexity to optimize the user grouping.

A. Design of User Grouping Algorithm

We consider user grouping as a two-sided matching process between the set of N users and the set of L groups, where $N = 2L$ since each group is shared by two users. Actually, each group is defined by a subchannel and two users are allocated on each subchannel. Let $\mathbb{L}$ and $\mathbb{N}$ respectively denote the sets of groups and users, which are two disjoint sets of players. By allocating UE_n in $\mathbb{N}$ to a group l in $\mathbb{L}$, the user grouping problem is defined as follows.

Definition 1. A two-to-one matching Φ is a mapping from all the subsets of users $\mathbb{N}$ into the groups set $\mathbb{L}$, satisfying the following properties for $\mathrm{UE}_n \in \mathbb{N}$ and $C_l \in \mathbb{L}$

(a) $\Phi(\mathrm{UE}_n) \in \mathbb{L}$;

(b) $\Phi(C_l) \subseteq \mathbb{N}$;

(c) $|\Phi(\mathrm{UE}_n)| = 1, |\Phi(C_l)| = 2$;

(d) $C_l \in \Phi(\mathrm{UE}_n) \Longleftrightarrow \mathrm{UE}_n = \Phi(C_l)$.

In Definition 1, property (a) and property (b) respectively indicate that each user only matches with one group and each group can be matched with a subset of users, property (c) means that only two users can be assigned to each group, and property (d) states that UE_n and C_l are matched with each other.

Remark 6. According to Definition 1, the optimization of user grouping is formulated as a two-to-one matching problem. Considering the co-channel interference between the users in the same group, each user's rate is partially decided by another user sharing the same group. Therefore, the WSED of each user depends on the user in the same group and the user grouping problem is a matching with externalities [32]–[34].

Then, we establish the preference list of users and groups. For any $\mathrm{UE}_n \in \mathbb{N}$ and two different groups $C_l \in \mathbb{L}$ and $C_{l'} \in \mathbb{L}$, UE_n prefers group C_l rather than $C_{l'}$ can be expressed as

$$(C_l, \Phi) \succ_{\mathrm{UE}_n} (C_{l'}, \Phi') \iff \mathrm{WSED}_{\mathrm{UE}_n}(\Phi) < \mathrm{WSED}_{\mathrm{UE}_n}\left(\Phi'\right), \tag{37}$$

where $\mathrm{WSED}_{\mathrm{UE}_n}(\Phi)$ is the WSED of UE_n with the group $L_k = \Phi(\mathrm{UE}_n)$. In terms of groups, $C_l \in \mathbb{L}$ prefers to match with UE_n rather than $\mathrm{UE}_{n'}$ is described as

$$(\mathrm{UE}_n, \Phi) \succ_{C_l} \left(\mathrm{UE}_{n'}, \Phi'\right) \iff \mathrm{WSED}_{C_l}(\Phi) < \mathrm{WSED}_{C_l}\left(\Phi'\right), \tag{38}$$

where $\mathrm{WSED}_{C_l}(\Phi)$ is the total WSED of the users matched with group C_l.

Considering the externalities, the stable matching is difficult to obtain [35]. The reason is, with externalities, the reactions of the users not in the group may affect the blocking possibility of the group. In order to guarantee all the users are well matched, in the following, we will propose the user grouping algorithm, which can achieve the solution with stability and low complexity. The concepts of two-sided exchange matching and two-sided exchange stability [34] are exploited in the matching process.

In a model of two-sided exchange matching, every two users in different groups can exchange their matched groups, which is defined as the swap operation. Specifically, a swap matching Φ_n^m means UE_n switches to UE_m's group and UE_m is assigned to UE_n's group while keeping other users' assignment the same. The definition of swap matching is mathematically described as follows.

Definition 2. A swap matching is denoted by $\Phi_n^m = \{\Phi \setminus \{(l, \mathrm{UE}_m), (l', \mathrm{UE}_n)\} \cup \{(l, \mathrm{UE}_n), (l', \mathrm{UE}_m)\}\}$, where $\mathrm{UE}_m \in \Phi(l)$, $\mathrm{UE}_n \in \Phi(l')$, $\mathrm{UE}_m \in \Phi_n^m(l')$, and $\mathrm{UE}_n \in \Phi_n^m(l)$.

With any swap operation of UE_m and UE_n, where $C_l = \Phi(\mathrm{UE}_m)$, $C_{l'} = \Phi(\mathrm{UE}_n)$, the original matching Φ is transformed to Φ_n^m. However, in a swap operation, considering their own interests, the players might not be approved by other players. In the following Definition, we introduce

the concept of swap-blocking pair and then we evaluate the conditions under which the swap operations can be approved.

Definition 3. Given a matching Φ and a pair $(\mathrm{UE}_m, \mathrm{UE}_n)$ with $\mathrm{UE}_m, \mathrm{UE}_n$ matched in Φ, if there exist $\Phi(\mathrm{UE}_m)$ and $\Phi(\mathrm{UE}_n)$ such that:
(a) $\forall i \in \{\mathrm{UE}_m, \mathrm{UE}_n, \Phi(\mathrm{UE}_m), \Phi(\mathrm{UE}_n)\}, WSED_i(\Phi_n^m) \leq WSED_t(\Phi)$;
(b) $\exists i \in \{\mathrm{UE}_m, \mathrm{UE}_n, \Phi(\mathrm{UE}_m), \Phi(\mathrm{UE}_n)\}, WSED_i(\Phi_n^m) < WSED_t(\Phi)$;
then swap matching Φ_n^m is approved, and $(\mathrm{UE}_m, \mathrm{UE}_n)$ is called a swap-blocking pair in Φ.

The Definition 3 indicates that a swap matching will be approved only when the WSED of any player does not increase, and at least one player's WSED decreases. Using the above definitions, the users' behaviors in a matching are described as follows. A potential swap blocking pair might be formed by choosing every two users in the system. Then, the BS checks whether these two users can benefit from each other by exchange their groups without hurting the interests of corresponding groups. After multiple swap operations, the externalities of the matching games's will be well handled. The matching process then reaches a stable status, which is also defined as a two-sided exchange stable matching as follows.

Definition 4. Φ is a two-sided exchange stable matching (2ES) if Φ is not blocked by any swap blocking pair $(\mathrm{UE}_m, \mathrm{UE}_n)$.

Based on the Definition 4, a matching based user grouping algorithm is proposed in Algorithm 1. At the beginning, we randomly assign users into groups and obtain an initial matching Φ_{init}. At each round, some user searches for another user in a different group and exchange their groups. The WSED can be updated in each group by using the proposed optimal power and time allocation. Then, if the swap operation is approved, the swap-blocking pair is formed and the matching is accepted. The swap matching phase is repeated until there is no users wants to exchange with another user.

In addition,

Remark 7. Note that the optimal resource allocation characterized in this work can be exploited not only with the proposed matching algorithm (i.e., Algorithm 1) but also with any other grouping algorithms. Furthermore, in the simulation results, we will show that the proposed low-complexity resource optimization method achieves a quite close performance as the globally optimal solution found by exhaustive search.

Algorithm 1 Matching Based User Grouping Algorithm

1: **Initialization**

Obtain Φ_{init} by Randomly matching users and groups

2: **Swap matching**

(1): **Repeat**

(2): Each UE_n searches another UE_m, where $\Phi(\mathrm{UE}_n) \neq \Phi(\mathrm{UE}_m)$.

(3): **If** $(\mathrm{UE}_n, \mathrm{UE}_m)$ is a swap-blocking pair

(4): The matching Φ_n^m is approved.

(5): UE_n and UE_m exchange the groups.

(6): Set $\Phi = \Phi_n^m$.

(7): **Else**

(8): UE_n keeps its match.

(9): **Until** there is no swap-blocking pair in a new round.

3: **End of algorithm**

B. Properties Analysis

In this subsection, the properties in terms of effectiveness, stability, convergence, and complexity are analyzed.

1) Effectiveness: : In the following Lemma, we will prove that the proposed user grouping algorithm greatly improves the performance.

Lemma 1. *The WSED of the system decreases after each swap operation.*

Proof. Suppose a swap operation from Φ to Φ_n^m. According to the proposed algorithm, a swap operation occurs and one user has searched another user for the exchange operation, which is approved by the two users and their groups. Hence, a swap-blocking pair has been successively formed. Based on the preference relations in (37) and (38), the WSED of each related player is not increased during the exchange operation. Note that the WSED of the unrelated player is unchanged. Therefore, we have

$$\sum_{i\in\mathbb{N}} WSED_i(\Phi) > \sum_{i\in\mathbb{N}} WSED_i(\Phi_n^m). \tag{39}$$

From (39), we conclude that the system WSED decreases after each successful swap operation.

□

2) Convergence:: In the following Proposition, we prove that the convergence of the proposed algorithm can be guaranteed.

Proposition 5. *Given any initial matching, the user grouping algorithm can always converge to a stable matching.*

Proof. In the user grouping algorithm, the number of users is limited, which implying the number of potential swap operations is finite. Moreover, from Lemma 1, we know that the system WSED decreases after each successful swap operation. Since the WSED has an lower bound, the swap operation stop when the lower bound has been achieved. Therefore, the proposed algorithm can always converge to a final state. □

3) Stability:: Using the definition of 2ES, the stability of the user grouping algorithm is proved as follows.

Proposition 6. *The final matching generated by the user grouping algorithm is 2ES.*

Proof. Assume the final matching Φ_{final} is not 2ES. According to Definition 4, there exists at least one swap blocking pair which can further reduce the WSED by performing swap operation. However, Φ_{final} is the final matching, which causes conflict. Therefore, the proposed algorithm reaches a 2ES matching. □

4) Complexity:: The complexity of the proposed matching based user grouping algorithm depends on the the number of cycles in the swap operation. Considering the worst case of the user grouping algorithm, the complexity is illustrated in the following.

Proposition 7. *Given a number of cycles C, the computational complexity of the user grouping algorithm is given as $\mathcal{O}\left(CN^2\right)$ in the worst case.*

Proof. According to the proposed algorithm, each user needs to search $N-2$ users to perform swap operation. In the worst case, all users search other in a complete cycle and hence at most $N(N-2)$ times of calculations are performed in each cycle. Practically, the number of swap operations can be reduced, which is because the user can successfully exchange with another user and the user assigned to the same group can be skipped. Given a number of cycles C, the computational complexity of the user grouping algorithm in the worst case is approximately $\mathcal{O}\left(CN^2\right)$. □

Table I

TABLE OF PARAMETERS

AWGN spectral density	$N_0 = -174$dBm
Path loss exponent	$v = 3$
Bandwidth	1MHz
Cell radius	100m

V. SIMULATION RESULTS

In this section, the performance of the proposed optimal power, time allocation and user grouping, i.e., the hybrid NOMA resource allocation, is evaluated. In simulations, the BS is located at the cell center and the users are randomly distributed in a circular. Each channel coefficient follows an i.i.d. Gaussian distribution as $g \sim \mathcal{CN}(0, \sigma^2)$, where the noise power is $\sigma^2 = BN_0$. The parameters are shown in Table I.

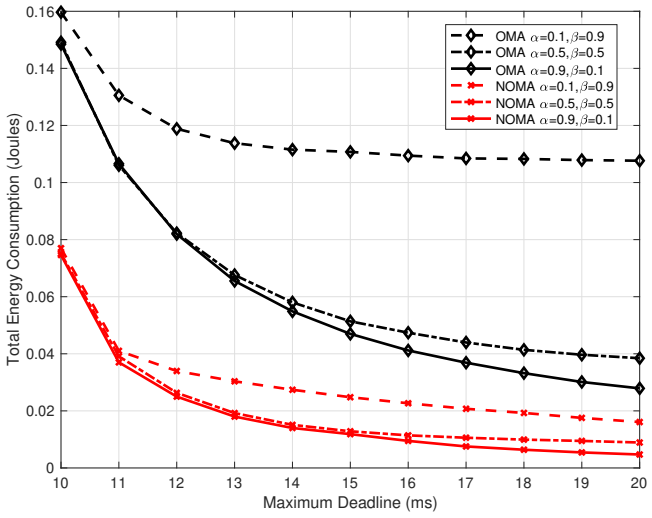

Figure 2. Energy consumption versus maximum deadline with different weights.

Fig. 2 and Fig. 3 respectively depict the total energy consumption and delay versus the maximum deadline respectively using the resource allocation in hybrid NOMA MEC systems and in OMA MEC systems with different weights. In these two figures, the maximum deadline is written as $\max\{D_{2,l}\}_{l=1}^{L}$ and the total delay is given by $D = \sum_{l=1}^{L}(D_{1,l} + t_l)$. The weights of energy consumption and delay are respectively taken as $\alpha_l = \alpha = [0.1, 0.5, 0.9]$

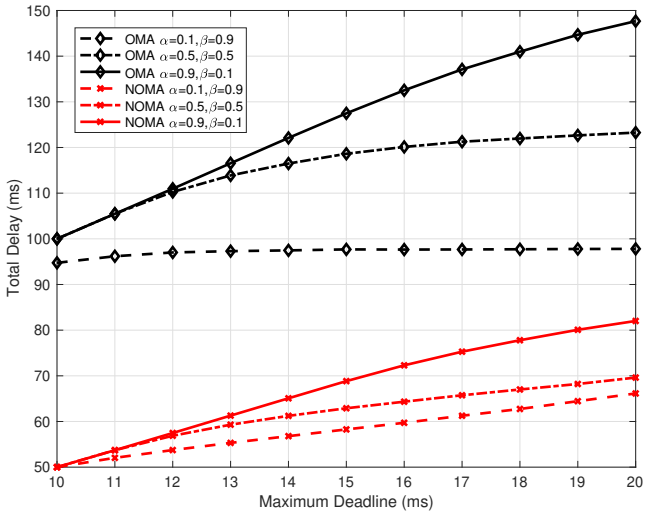

Figure 3. Total delay versus maximum deadline with different weights.

and $\beta_l = \beta = [0.9, 0.5, 0.1]$ for $l = 1, \cdots, L$. In Fig. 2 and Fig. 3, the number of bits in the task is $M = 100$Kbits and the number of users is $N = 10$. The resource allocation, i.e., the joint optimization of power and time, in OMA MEC systems is also optimized similar to NOMA MEC, i.e., we first obtain the optimal power by fixing time delay and then optimize the time. It is seen that NOMA outperforms OMA in terms of energy consumption and delay. In addition, one can observe that both the schemes of hybrid NOMA MEC and OMA MEC achieve a smaller energy consumption with a larger allocated weight α. This is because, a larger α means that, compared to delay minimization, the system puts more effort to minimize the energy consumption. Similarly, it also can be easily found that given a lager value of β, these two schemes achieves a smaller delay, which is because a lager value of β implies that the system puts more efforts to minimize the delay.

In Fig. 4 and Fig. 5, we respectively evaluate the energy consumption and delay versus the number of bits in the task with different weights. In these two figures, the maximum deadline is taken as $D_{\max} = 15$ms and the number of users is $N = 10$. One can see similar phenomenon as Fig. 2 and Fig. 3 that a higher α induces lower energy consumption and larger delay and the energy consumption becomes larger and delay becomes lower with a higher β. In addition, the proposed hybrid NOMA MEC scheme outperforms the conventional OMA MEC scheme. Moreover, the performance gap becomes larger when the number of bits in the task increases.

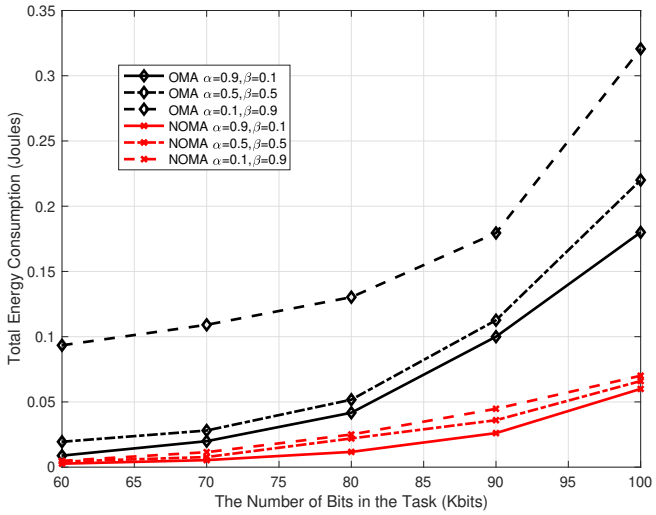

Figure 4. Energy consumption versus the number of bits in the task with different weights.

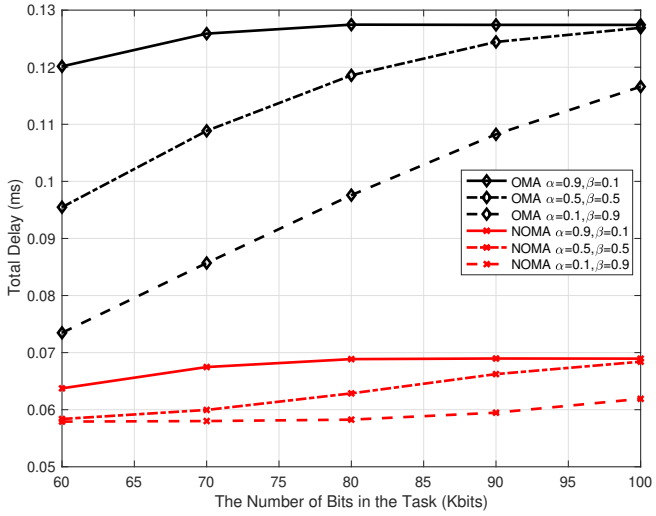

Figure 5. Total delay versus the number of bits in the task with different weights

Fig. 6 displays the total cost, i.e., the total weighted sum of energy consumption and delay, in this hybrid NOMA MEC system versus the number of users with different maximum deadlines. In this figure, the number of bits in the task is $M = 80$Kbits and the weights are taken as $\alpha_l = \beta_l = 0.5$ for $l = 1, \cdots, L$. As expected, the proposed hybrid NOMA MEC scheme outperforms the conventional OMA scheme. Furthermore, with the increasing of the number of

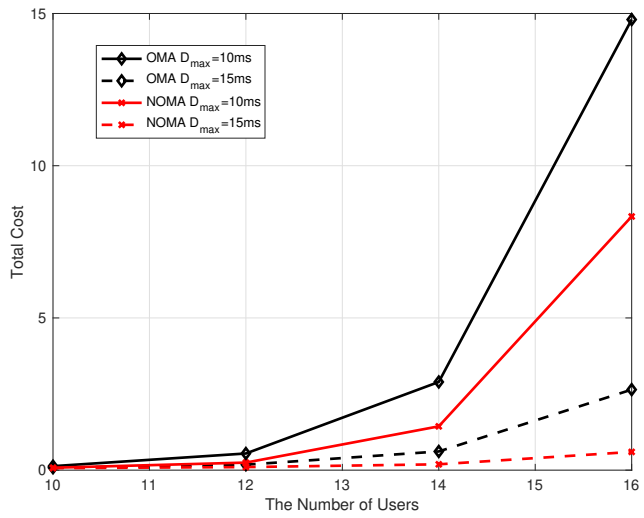

Figure 6. Total cost versus the number of users with different maximum deadline.

the users, the performance gap becomes larger. In addition, it is found that a higher maximum deadline induces a lower cost. This is because with a higher maximum deadline, the energy consumption will decrease.

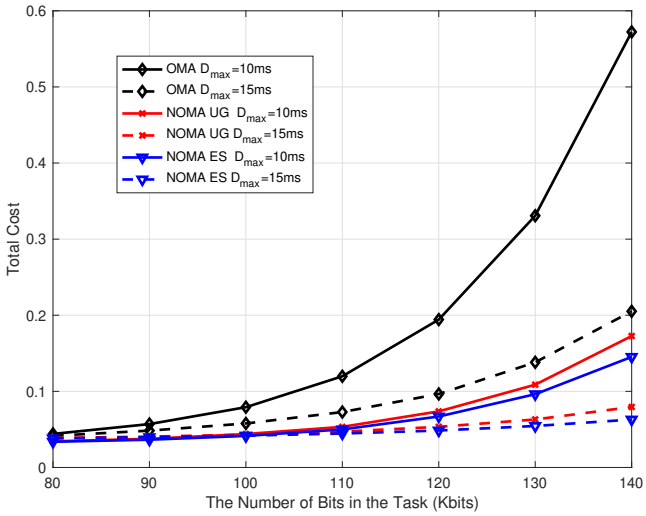

Figure 7. Total cost versus the number of bits in the task with different maximum deadline.

In Fig. 7, with different maximum deadline, we compare the total cost using the proposed user grouping (US) algorithm with the method of exhaustive search (ES) in hybrid NOMA MEC

systems, and in OMA MEC systems. Considering the high complexity of ES, the number of users is set as $N = 6$. The weights are taken as $\alpha_l = \beta_l = 0.5$ for $l = 1, \cdots, L$. One can see the similar phenomenon as in Fig. 2 to Fig. 6 that hybrid NOMA MEC performs better than OMA MEC. Furthermore, the performance achieved using the proposed methods is very close to the globally optimal value. Therefore, with low complexity, the proposed resource allocation achieves near-optimal performance.

VI. Conclusion

In this paper, we studied the resource allocation, i.e., the joint power, time allocation and user grouping, in hybrid NOMA MEC systems to minimize the WSED. Three strategies, i.e., pure NOMA, hybrid NOMA and OMA, were considered and the corresponding optimal power and time allocation solutions in closed forms were characterized. We also showed that all the three strategies might possibly happen to the users when taking different values of weight factors, deadlines, and channel gains. In addition, using the proposed closed-form power and time allocation, we proposed an efficient user grouping algorithm to solve the resource allocation problem for multiple users in hybrid NOMA MEC systems. The simulation results showed that the proposed resource optimization method for hybrid NOMA MEC over performed OMA MEC in terms of energy consumption and delay.

Appendix

A. Proof of Theorem 1

1) The Case of $t_l = 0$: Firstly, we consider the special condition of $t_l = 0$, i.e., the pure NOMA. In this case, $p_{2,l}^{2*} = 0$, and problem (20) is reduced to

$$\min_{p_{2,l}^1} \ \alpha_l D_{1,l} p_{2,l}^1, \tag{40}$$

$$\text{s.t. } D_{1,l} B \ln\left(1 + |h_{2,l}|^2 p_{2,l}^1 e^{\frac{-M}{D_{1,l}B}}\right) - M \geq 0, \tag{41}$$

$$p_{2,l}^1 \geq 0, \tag{42}$$

Note that (40) is a monotonically increasing function of $p_{2,l}^1$ and from constraint (41), the lower bound of $p_{2,l}^1$ is obtained as

$$p_{2,l}^{1*} = \frac{e^{\frac{2M}{BD_{1,l}}} - e^{\frac{M}{BD_{1,l}}}}{|h_{2,l}|^2}. \tag{43}$$

2) The Case of $0 < t_l < D_{1,l}$*:* Since problem (20) is a convex problem, we can exploit the Lagrangian of problem (20) to find the optimal solution, which is given by

$$L = \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2 \right) - \lambda_1 \Xi - \lambda_2 p_{2,l}^1 - \lambda_3 p_{2,l}^2, \tag{44}$$

where

$$\Xi = D_{1,l} B \ln \left(1 + \left| h_{2,l} \right|^2 p_{2,l}^1 e^{\frac{-M}{D_{1,l} B}} \right) + t_l B \ln \left(1 + \left| h_{2,l} \right|^2 p_{2,l}^2 \right) - M, \tag{45}$$

$\lambda_i \geq 0$ for $i = 1, 2, 3$ are the Lagrangian multipliers. Then the Karush-Kuhn-Tucker (KKT) conditions are given as follows

$$\frac{\partial L}{\partial p_{2,l}^1} = \alpha_l D_{1,l} - \lambda_1 \frac{D_{1,l} B \left| h_{2,l} \right|^2 e^{\frac{-M}{D_{1,l} B}}}{1 + \left| h_{2,l} \right|^2 p_{2,l}^1 e^{\frac{-M}{D_{1,l} B}}} - \lambda_2 = 0, \tag{46}$$

$$\frac{\partial L}{\partial p_{2,l}^2} = \alpha_l t_l - \lambda_1 \frac{t_l B \left| h_{2,l} \right|^2}{1 + \left| h_{2,l} \right|^2 p_{2,l}^2} - \lambda_3 = 0, \tag{47}$$

$$\lambda_1 \Xi = 0, \tag{48}$$

$$\lambda_2 p_{2,l}^1 = 0, \tag{49}$$

$$\lambda_3 p_{2,l}^2 = 0. \tag{50}$$

Here, it is worth pointing out that $\lambda_2 > 0,\ \lambda_3 > 0$ is not possible. This is because, if $\lambda_2 > 0,\ \lambda_3 > 0$, from (49) and (50), we will have $p_{2,l}^1 = 0,\ p_{2,l}^2 = 0$, which can not satisfy the constraint $\Xi \geq 0$. In addition, from (46) or (47), the impossible case $\lambda_2 > 0,\ \lambda_3 > 0$ implies $\lambda_1 > 0$. Therefore, we focus on three cases, i.e., $\lambda_1 > 0,\ \lambda_2 = 0,\ \lambda_3 = 0$, $\lambda_1 > 0,\ \lambda_2 > 0,\ \lambda_3 = 0$, and $\lambda_1 > 0,\ \lambda_2 = 0,\ \lambda_3 > 0$.

In the case of $\lambda_1 > 0,\ \lambda_2 = 0,\ \lambda_3 = 0$, we easily have $p_{2,l}^1 > 0,\ p_{2,l}^2 > 0$ from (49) and (50), and hence this case can be termed as hybrid NOMA. Using (46) and (47), we have

$$p_{2,l}^1 = \frac{\lambda_1 B \left| h_{2,l} \right|^2 - \alpha_l e^{\frac{M}{D_{1,l} B}}}{\alpha_l \left| h_{2,l} \right|^2}, \tag{51}$$

$$p_{2,l}^2 = \frac{\lambda_1 B \left| h_{2,l} \right|^2 - \alpha_l}{\alpha_l \left| h_{2,l} \right|^2}, \tag{52}$$

where the Lagrangian multiplier, λ_1, is obtained from (48) with $\lambda_1 > 0$: $\Xi = 0$, hence

$$\lambda_1 = e^{\frac{\frac{M}{B} - D_{1,l} \ln \left(\frac{B \left| h_{2,l} \right|^2 e^{\frac{-M}{D_{1,l} B}}}{\alpha_l} \right) - t_l \ln \left(\frac{B \left| h_{2,l} \right|^2}{\alpha_l} \right)}{D_{1,l} + t_l}} = \frac{\alpha_l e^{\frac{2M}{B \left(D_{1,l} + t_l \right)}}}{B \left| h_{2,l} \right|^2}. \tag{53}$$

By taking (53) into (51) and (52), the optimal solution to problem (20) is

$$p_{2,l}^{1*}=\frac{e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-e^{\frac{M}{BD_{1,l}}}}{\left|h_{2,l}\right|^2},\ p_{2,l}^{2*}=\frac{e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-1}{\left|h_{2,l}\right|^2}. \tag{54}$$

Here, given $0<t_l<D_{1,l}$, the optimal solution in (54) satisfies the constraint $p_{2,l}^{i*}\geq 0$ for $i=1,2$. Note that, by fixing t_l, the optimization problem (20) aims to minimize the energy consumption of users. Therefore, for hybrid NOMA, the minimum energy consumption in each group is

$$E^{H-NOMA*}=\frac{\alpha_l}{\left|h_{2,l}\right|^2}e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}\left(D_{1,l}+t_l\right)-\frac{\alpha_l}{\left|h_{2,l}\right|^2}\left(D_{1,l}e^{\frac{M}{BD_{1,l}}}+t_l\right). \tag{55}$$

In the case of $\lambda_1>0,\ \lambda_2>0,\ \lambda_3=0$, it is easy to have

$$p_{2,l}^{1*}=0,\ p_{2,l}^{2*}=\frac{e^{\frac{M}{Bt_l}}-1}{\left|h_{2,l}\right|^2}, \tag{56}$$

from (48), (49), and (50), which is the OMA scheme. Therefore, the minimum energy consumption in OMA scheme in each group is

$$E^{OMA*}=\frac{\alpha_l t_l}{\left|h_{2,l}\right|^2}\left(e^{\frac{M}{Bt_l}}-1\right). \tag{57}$$

Finally, in the case of $\lambda_1>0,\ \lambda_2=0,\ \lambda_3>0$, we easily obtain $p_{2,l}^{2*}=0$ from (50) and

$$p_{2,l}^{1*}=\frac{e^{\frac{2M}{BD_{1,l}}}-e^{\frac{M}{BD_{1,l}}}}{\left|h_{2,l}\right|^2}, \tag{58}$$

from (48). This case corresponds to an extreme situation where all the power of $\mathrm{UE}_{2,l}$ is allocated to $D_{1,l}$, which is termed as pure NOMA. Then, in the pure NOMA scheme, the minimum energy consumption in each group is

$$E^{NOMA*}=\frac{\alpha_l D_{1,l}}{\left|h_{2,l}\right|^2}e^{\frac{M}{BD_{1,l}}}\left(e^{\frac{M}{BD_{1,l}}}-1\right). \tag{59}$$

Although there exist three cases where $0<t_l<D_{1,l}$, we prove that only hybrid NOMA achieves the minimum energy consumption. On one hand, we compare the hybrid NOMA with OMA. From (55) and (57), we have

$$\begin{aligned}E^{H-NOMA*}-E^{OMA*}&=\frac{\alpha_l}{\left|h_{2,l}\right|^2}\left(\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-\left(D_{1,l}e^{\frac{M}{BD_{1,l}}}+t_l\right)\right)-\frac{\alpha_l t_l}{\left|h_{2,l}\right|^2}\left(e^{\frac{M}{Bt_l}}-1\right),\\&=\frac{\alpha_l}{\left|h_{2,l}\right|^2}\left(\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-\left(D_{1,l}e^{\frac{M}{BD_{1,l}}}+t_l e^{\frac{M}{Bt_l}}\right)\right).\end{aligned} \tag{60}$$

In order to identify whether (60) is positive or negative, we define

$$f\left(t_l\right)=\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-\left(D_{1,l}e^{\frac{M}{BD_{1,l}}}+t_l e^{\frac{M}{Bt_l}}\right), \tag{61}$$

and the derivative of $f\left(t_l\right)$ is given by

$$\frac{df(t_l)}{dt_l}=\left(1-\frac{2M}{B\left(D_{1,l}+t_l\right)}\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-\left(1-\frac{M}{Bt_l}\right)e^{\frac{M}{Bt_l}}. \tag{62}$$

We then define another function $g\left(x\right)$ as

$$g\left(x\right)=\left(1-x\right)e^{x}, \tag{63}$$

which is a monotonically non-increasing function of $x\geq 0$ in that we have

$$\frac{dg\left(x\right)}{dx}=-xe^{x}\leq 0. \tag{64}$$

Hence, using the monotonically non-increasing function $g\left(x\right)$, we have

$$\frac{df(t_l)}{dt_l}=g\left(\frac{2M}{B\left(D_{1,l}+t_l\right)}\right)-g\left(\frac{M}{Bt_l}\right), \tag{65}$$

which is positive for $\frac{2M}{B\left(D_{1,l}+t_l\right)}<\frac{M}{Bt_l}$, i.e., $t_l<D_{1,l}$, and negative for $\frac{2M}{B\left(D_{1,l}+t_l\right)}>\frac{M}{Bt_l}$, i.e., $t_l>D_{1,l}$. Therefore,

$$f\left(t_l\right)_{\max}=f\left(D_{1,l}\right)=0. \tag{66}$$

Given $0<t_l<D_{1,l}$, we always have

$$E^{H-NOMA*}<E^{OMA*}. \tag{67}$$

On the other hand, the pure NOMA case is compared to the hybrid NOMA case. From (55) and (59), we have

$$E^{H-NOMA*}-E^{NOMA*}=\frac{\alpha_l}{\left|h_{2,l}\right|^2}\left(\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-\left(D_{1,l}e^{\frac{2M}{BD_{1,l}}}+t_l\right)\right). \tag{68}$$

Let

$$w(t_l)=\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-\left(D_{1,l}e^{\frac{2M}{BD_{1,l}}}+t_l\right), \tag{69}$$

and the derivative of $w\left(t_l\right)$ is

$$\frac{dw(t_l)}{dt_l}=\left(1-\frac{2M}{B\left(D_{1,l}+t_l\right)}\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}}-1=g\left(\frac{2M}{B\left(D_{1,l}+t_l\right)}\right)-g\left(0\right), \tag{70}$$

Here, function $g\left(x\right)$ has been defined in (63), which is monotonically non-increasing. Therefore, since $\frac{2M}{B\left(D_{1,l}+t_l\right)} \geq 0$, we have

$$g\left(\frac{2M}{B\left(D_{1,l}+t_l\right)}\right) - g\left(0\right) \leq 0, \tag{71}$$

implying $\frac{dw(t_l)}{dt_l} \leq 0$, therefore $w(t_l)$ is also a monotonically non-increasing function of t_l. Hence, the maximum value of $w\left(t_l\right)$ is given by

$$w\left(t_l\right)_{\max} = w\left(0\right) = 0. \tag{72}$$

Since $0 < t_l < D_{1,l}$, we have

$$E^{H-NOMA*} < E^{NOMA*}. \tag{73}$$

Combining (67) and (73), the hybrid NOMA case, i.e., the case of $\lambda_1 > 0,\ \lambda_2 = 0,\ \lambda_3 = 0$, yields the minimum energy consumption. Therefore, by fixing time t_l, given the condition of $0 < t_l < D_{1,l}$, the optimal power allocation is obtained in the scheme of hybrid NOMA.

3) The Case of $t_l \geq D_{1,l}$: Note that the optimal $p_{2,l}^{i*},\ i = 1, 2$ obtained in (54) can satisfy the constraint $p_{2,l}^{i*} > 0,\ i = 1, 2$ only if $t_l < D_{1,l}$. However, if $t_l \geq D_{1,l}$, the optimal solution is not achieved in the hybrid NOMA case. The optimal solution is achieved in the case of OMA or pure NOMA. In the following, it is proved that if $t_l \geq D_{1,l}$, the optimal solution is achieved in the OMA case. According to (57) and (59), the energy consumption gap between OMA and pure NOMA is given by

$$E^{OMA*} - E^{NOMA*} = \frac{\alpha_l t_l}{\left|h_{2,l}\right|^2}\left(e^{\frac{M}{Bt_l}} - 1\right) - \frac{\alpha_l D_{1,l}}{\left|h_{2,l}\right|^2} e^{\frac{M}{BD_{1,l}}}\left(e^{\frac{M}{BD_{1,l}}} - 1\right). \tag{74}$$

We define

$$u(t_l) = t_l\left(e^{\frac{M}{Bt_l}} - 1\right), \tag{75}$$

where the derivative of function $u(t_l)$ is

$$\frac{du(t_l)}{dt_l} = e^{\frac{M}{Bt_l}}\left(1 - \frac{M}{Bt_l}\right) - 1 = g\left(\frac{M}{Bt_l}\right) - g\left(0\right), \tag{76}$$

where the monotonically non-increasing function $g\left(x\right)$ has been defined in (63). Here, since $\frac{M}{Bt_l} \geq 0$, we have $\frac{du(t_l)}{dt_l} \leq 0$ from (76), which implies function $u(t_l)$ is monotonically non-increasing. With the condition of $t_l > D_{1,l}$, we have

$$u(t_l) < u(D_{1,l}) = D_{1,l}\left(e^{\frac{M}{BD_{1,l}}} - 1\right). \tag{77}$$

Using (77) and (74), it is easy to show that

$$E^{OMA*} - E^{NOMA*} < \frac{\alpha_l D_{1,l}}{|h_{2,l}|^2}\left(e^{\frac{M}{BD_{1,l}}} - 1\right) - \frac{\alpha_l D_{1,l}}{|h_{2,l}|^2} e^{\frac{M}{BD_{1,l}}}\left(e^{\frac{M}{BD_{1,l}}} - 1\right),$$
$$= -\frac{\alpha_l D_{1,l}}{|h_{2,l}|^2}\left(e^{\frac{M}{BD_{1,l}}} - 1\right)^2 \leq 0, \tag{78}$$

implying $E^{OMA*} < E^{NOMA*}$. Therefore, in the condition of $t_l \geq D_{1,l}$, the optimal solution is achieved in the case of OMA, i.e.,

$$p_{2,l}^{1*} = 0,\ p_{2,l}^{2*} = \frac{e^{\frac{M}{Bt_l}} - 1}{|h_{2,l}|^2}. \tag{79}$$

B. Proof of Corollary 1

According to Theorem 1, in the hybrid NOMA case, the gap between the powers consumed during $D_{1,l}$ and t_l is given by

$$p_{2,l}^{1*} - p_{2,l}^{2*} = \frac{e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - e^{\frac{M}{BD_{1,l}}}}{|h_{2,l}|^2} - \frac{e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1}{|h_{2,l}|^2} = \frac{1 - e^{\frac{M}{BD_{1,l}}}}{|h_{2,l}|^2} \leq 0, \tag{80}$$

implying $p_{2,l}^{1*} \leq p_{2,l}^{2*}$.

C. Proof of Corollary 2

First, we prove the sufficient condition. According to Theorem 1, NOMA outperforms OMA when $0 \leq t_l < D_{1,l}$. In addition, since $0 \leq t_l \leq D_{2,l} - D_{1,l}$, we have $D_{2,l} - D_{1,l} < D_{1,l}$, i.e., $D_{2,l} < 2D_{1,l}$.

We then prove the necessary condition. Given $D_{2,l} < 2D_{1,l}$, i.e., $D_{2,l} - D_{1,l} < D_{1,l}$, from the constraint $0 \leq t_l \leq D_{2,l} - D_{1,l}$, we obtain $0 \leq t_l < D_{1,l}$ and hence NOMA outperforms OMA according to Theorem 1.

D. Proof of Proposition 1

Since $0 \leq t_l \leq D_{2,l} - D_{1,l}$ and $D_{2,l} < 2D_{1,l}$, we easily have $0 \leq t_l < D_{1,l}$. Therefore, by using (21) and (22) in Theorem 1, the energy consumption gap between NOMA MEC and OMA MEC is

$$\Delta = \frac{\alpha_l}{|h_{2,l}|^2}\left(\left(D_{1,l} + t_l\right) e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - \left(D_{1,l} e^{\frac{M}{BD_{1,l}}} + t_l\right)\right) - \frac{\alpha_l t_l}{|h_{2,l}|^2}\left(e^{\frac{M}{Bt_l}} - 1\right),$$
$$= \frac{\alpha_l}{|h_{2,l}|^2}\left(\left(D_{1,l} + t_l\right) e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - \left(D_{1,l} e^{\frac{M}{BD_{1,l}}} + t_l e^{\frac{M}{Bt_l}}\right)\right), \tag{81}$$

which has been proved to be a monotonically non-decreasing function for $t_l < D_{1,l}$ in the proof of Theorem 1 in (60). Given $0 \le t_l \le D_{2,l} - D_{1,l} < D_{1,l}$, we have $\Delta(t_l)_{\max} = \Delta(D_{2,l} - D_{1,l}) < \Delta(D_{1,l}) = 0$. This completes the proof.

E. Proof of Corollary 3

Since $0 \le t_l \le D_{2,l} - D_{1,l}$ and $D_{2,l} < 2D_{1,l}$, we easily have $0 \le t_l < D_{1,l}$. Therefore, from Theorem 1, with $\alpha_l = 1,\ \beta_l = 0$, the energy consumption is given by

$$E_l(t_l) = \frac{1}{|h_{2,l}|^2}\left((D_{1,l}+t_l)\, e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - \left(D_{1,l} e^{\frac{M}{BD_{1,l}}} + t_l\right)\right). \tag{82}$$

The derivative of $E_l(t_l)$ is

$$\begin{aligned}\frac{dE_l}{dt_l} &= \frac{1}{|h_{2,l}|^2}\left(\left(1 - \frac{2M}{B(D_{1,l}+t_l)}\right) e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1\right), \\ &= \frac{1}{|h_{2,l}|^2}\left(g\left(\frac{2M}{B(D_{1,l}+t_l)}\right) - g(0)\right),\end{aligned} \tag{83}$$

where $g(x)$ is a monotonically non-increasing function of $x \ge 0$, which has been defined in (63). Since $\frac{2M}{B\left(D_{1,l}+t_l\right)} \ge 0$, we have $g\left(\frac{2M}{B\left(D_{1,l}+t_l\right)}\right) - g(0) \le 0$ and $\frac{dE_l}{dt_l} \le 0$. Therefore, $E_l(t_l)$ is a monotonically non-increasing function, and hence $E_l(t_l)_{\min} = E_l(D_{2,l} - D_{1,l})$.

F. Proof of Proposition 2

In problem (25), the constraint (26) is linear, thus we focus on investigating the convexity of the objective function $C(t_l)$. The second order derivative of function $C(t_l)$ is

$$\frac{d^2C(t_l)}{dt_l^2} = \frac{4\alpha_l M}{|h_{2,l}|^2 B(D_{1,l}+t_l)^3} e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} \ge 0, \tag{84}$$

implying function $C(t_l)$ is convex, which completes the proof.

G. Proof of Proposition 3

By setting the derivative of the objective function $C(t_l)$ to zero, we have

$$\frac{dC(t_l)}{dt_l} = \frac{\alpha_l}{|h_{2,l}|^2}\left(\left(1 - \frac{2M}{B(D_{1,l}+t_l)}\right) e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1\right) + \beta_l = 0, \tag{85}$$

thus

$$\left(\frac{2M}{B(D_{1,l}+t_l)} - 1\right) e^{\frac{2M}{B\left(D_{1,l}+t_l\right)} - 1} = \left(\frac{\beta_l |h_{2,l}|^2 - \alpha_l}{e\alpha_l}\right). \tag{86}$$

Hence, by using the Lambert function, the unique root of (85) is given as $t_l = \Omega$. Here, since $\frac{2M}{B\left(D_{1,l}+t_l\right)} - 1 \geq -1$, the Lambert function can be denoted by a single-valued function $W_0\left(\frac{\beta_l\left|h_{2,l}\right|^2-\alpha_l}{e\alpha_l}\right)$. Note that t_l ranges from zero to $D_{2,l} - D_{1,l}$, we should consider three cases to characterize the minimum point of function $C\left(t_l\right)$. Firstly, if $\Omega < 0$, we have

$$C_{\min}\left(t_l\right) = C\left(0\right), \tag{87}$$

which is the pure NOMA case and $\text{UE}_{2,l}$ can offload its task within $D_{1,l}$. Secondly, if $0 \leq \Omega \leq D_{2,l} - D_{1,l}$, we have

$$C_{\min}\left(t_l\right) = C\left(\Omega\right), \tag{88}$$

implying $\text{UE}_{2,l}$ can offload its task before its deadline. Finally, if $\Omega > D_{2,l} - D_{1,l}$, we have

$$C_{\min}\left(t_l\right) = C\left(D_{2,l} - D_{1,l}\right), \tag{89}$$

which means $\text{UE}_{2,l}$ also offloads its task before its deadline.

H. Proof of Proposition 4

Firstly, define

$$G\left(t_l\right) = \frac{\alpha_l}{\left|h_{2,l}\right|^2} t_l \left(e^{\frac{M}{Bt_l}} - 1\right) + \beta_l t_l, \tag{90}$$

which is the objective function of (33). By setting the derivative of $G\left(t_l\right)$ to zero, we have

$$\frac{dG\left(t_l\right)}{dt_l} = \frac{\alpha_l}{\left|h_{2,l}\right|^2} \left(e^{\frac{M}{Bt_l}} - 1 - \frac{M}{Bt_l} e^{\frac{M}{Bt_l}}\right) + \beta_l = 0, \tag{91}$$

leading to a unique root $t_l = \Lambda$ by using the Lambert function. In addition, we have

$$\frac{d^2G\left(t_l\right)}{dt_l^2} = \frac{\alpha_l}{\left|h_{2,l}\right|^2} \frac{M}{Bt_l^3} e^{\frac{M}{Bt_l}} \geq 0, \tag{92}$$

indicating that Λ is a maximizer. However, since the constraint (34), i.e., $D_{1,l} \leq t_l \leq D_{2,l} - D_{1,l}$, we should consider three cases to identify the optimal point. If $\mathit{\Lambda} > D_{2,l} - D_{1,l}$, we have

$$G_{\min}\left(t_l\right) = G\left(D_{1,l}\right). \tag{93}$$

In cases where $D_{1,l} \leq \mathit{\Lambda} \leq D_{2,l} - D_{1,l}$, the optimal function value is given by

$$G_{\min}\left(t_l\right) = G\left(\mathit{\Lambda}\right). \tag{94}$$

Finally, when $\mathit{\Lambda} < D_{1,}$, the corresponding optimal function value is

$$G_{\min}\left(t_l\right) = G\left(D_{2,l} - D_{1,l}\right). \tag{95}$$

REFERENCES

[1] J. Zhu, J. Wang, Y. Huang, F. Fang, K. Navaie, and Z. Ding, "Optimal resource allocation in hybrid NOMA MEC systems," *has been accepted in Proc. IEEE Global Commun. Conf.*, Dec. 2019.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, pp. 1065–1082, Jun. 2014.

[3] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *available on arXiv.org*, 2016.

[4] P. Wang, J. Xiao, and P. Li, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, pp. 4–11, Sept. 2006.

[5] M. Vaezi , Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Springer, 2019.

[6] 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 Commun. Mag.*, vol. 53, pp. 74–81, Sept. 2015.

[7] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, pp. 2744–2757, Dec. 2017.

[8] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 537–552, Jan. 2016.

[9] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 4438–4454, Jun. 2016.

[10] X. Zhang, Q. Gao, C. Gong, and Z. Xu, "User grouping and power allocation for NOMA visible light communication multi-cell networks," *IEEE Commun. Lett.*, vol. 21, pp. 777–780, A. 2016.

[11] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "On the application of NOMA to wireless caching," in *Int. Conf. on Commun.*, (Kansas City, MO), pp. 1–7, May 2018.

[12] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Feb. 2017.

[13] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, pp. 854–864, Dec. 2016.

[14] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuto.*, vol. 19, pp. 2322–2358, Oct. 2017.

[15] A. U. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A survey of mobile cloud computing application models," *IEEE Commun. Surveys Tut.*, vol. 16, pp. 393–413, Jan. 2014.

[16] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. S. Quek, "Offloading in mobile edge computing: Task allocation and computational frequency scaling," *IEEE Trans. on Commun.*, vol. 65, pp. 3571–3584, Aug. 2017.

[17] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet of Things Journal*, vol. 6, pp. 4188–4200, Jun. 2019.

[18] J. Liu and Q. Zhang, "Offloading schemes in mobile edge computing for ultra-reliable low latency communications," *IEEE Access*, vol. 6, pp. 12825–12837, Feb. 2018.

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

[20] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet of Things Journal*, vol. 5, pp. 1299–1306, Apr. 2018.

[21] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. on Commun.*, vol. 67, pp. 375–390, Jan. 2019.

[22] 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.

[23] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. K. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. on Veh. Technol.*, vol. 67, pp. 12244–12258, Dec. 2018.

[24] F. Wang, J. Xu, and Z. Ding, "Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems," *IEEE Trans. on Commun*, pp. 1–1, Nov. 2018.

[25] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Processing Lett.*, vol. 25, pp. 1875–1879, Dec. 2018.

[26] Z. Ding, J. Xu, O. A. Dobre, and V. Poor, "Joint power and time allocation for NOMA-MEC offloading," *IEEE Trans. on Veh. Technol.*, pp. 1–1, Mar. 2019.

[27] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 6010–6023, Aug. 2016.

[28] F. Fang, H. Zhang, J. Cheng, and V. C. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, pp. 3722–3732, May. 2016.

[29] 3rd Generation Partnership Project (3GPP), "Study on downlink multiuser superposition transmission for LTE," Mar. 2015.

[30] J. Xu and S. Ren, "Online learning for offloading and autoscaling in renewable-powered mobile edge computing," in *IEEE Global Commun. Conf.*, pp. 1–6, Dec. 2016.

[31] L. Tianze, W. Muqing, Z. Min, and L. Wenxing, "An overhead-optimizing task scheduling strategy for ad-hoc based mobile edge computing," *IEEE Access*, vol. 5, pp. 5609–5622, Mar. 2017.

[32] H. Sasaki and M. Toda, "Two-sided matching problems with externalities," *Journal of Economic Theory*, vol. 70, no. 1, pp. 93–108, 1996.

[33] K. Bando, "Many-to-one matching markets with externalities among firms," *Journal of Mathematical Economics*, vol. 48, no. 1, pp. 14–20, 2012.

[34] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *International Symposium on Algorithmic Game Theory*, pp. 117–129, Springer, 2011.

[35] A. E. Roth and M. Sotomayor, "Two-sided matching," *Handbook of game theory with economic applications*, vol. 1, pp. 485–541, 1992.

Response to the Reviewers' Comments for Paper "Resource Allocation for Hybrid NOMA MEC Offloading"

Jianyue Zhu, Jiaheng Wang, Yongming Huang, Fang Fang, Keivan Navaie, Zhiguo Ding

We would like to thank the Editor for professionally and expeditiously handling the review of our manuscript. We also thank the reviewers for their positive, insightful, and constructive comments, which have helped us improve the quality of the paper dramatically. We have carefully addressed the reviewers' concerns and properly revised the paper according to the reviewers' suggestions. The main changes include:

1) We have rewritten the abstract to better illustrate the contribution.
2) We have improved the illustration of the concept of hybrid NOMA and MEC and its main application in future in Section I.
3) We have clarified the issues on the assumptions in Section II.
4) We have explained the optimality of the obtained solutions in Section III.
5) We have described the rationale behind Proposition 4 in Section III.
6) We have rewritten Theorem 1, Corollaries 1, 2, 3, and Proposition 1 to make them easier to understand in Section III.
7) We have rewritten the texts after Corollaries 1, 2, and Proposition 1 to make them better matched with the corresponding theoretical results in Section III.
8) We have added a subsection (i.e., Section IV-B) to theoretically analysis the properties of the proposed user grouping algorithm in Section IV.
9) We have added a table to summarize the detailed parameter-settings in Section V.
10) We have fixed the typos pointed out by the reviewers.

The main changes are marked in blue in the revised paper. The detailed response to the reviewers' comments is provided in the following.

RESPONSE TO EDITOR

Comment: *The reviewer(s) have suggested some major revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript. When revising your manuscript, please pay attention to the following three aspects:1, better motivate the blending of the hybrid NOMA and MEC technique and also the formulated problem; 2, clarify the issues on some important assumptions made in the paper, including the assumption that the most demanding task-deadline should be served first and the assumption on the delay; 3, provide some theoretical analysis results and more discussions on Algorithm 1 such as its scalability, global solution, convergence, and complexity.*

Response: We sincerely thank you for sacrificing time to handle the review process of our paper and your helpful suggestion.

First, to address the reviewers' concerns on the motivation of the hybrid NOMA MEC scheme and the formulated problem, we provide the following explanation.

The reason why we integrated hybrid NOMA and MEC is given as follows:

1) In NOMA MEC systems, two users are allowed to simultaneously offload their tasks to the server. Therefore, compared to OMA, i.e., TDMA, the delay in NOMA MEC systems is reduced.
2) Moreover, [21] has pointed that in group l, $\text{UE}_{2,l}$ has to consume more energy in NOMA than in OMA if $\text{UE}_{2,l}$ completely relies on $D_{1,l}$. Hence, the hybrid NOMA, i.e., $\text{UE}_{2,l}$ first shares $D_{1,l}$ with $\text{UE}_{1,l}$ and then is allocated with another time interval to complete the task offloading, was studied.
3) In practice, hybrid NOMA MEC can often be used when users have various deadlines. In a hybrid NOMA MEC system, both of the time and energy consumption can be greatly saved by reusing the offloading time of the users who have urgent tasks.

In addition, the motivation of the formulated problem is as follows:

1) In the literature, it has been proved that the integration of NOMA and MEC can not only avoid sever delay but also reduce energy consumption [19][20]. Hence, in this paper, we optimized both of the delay and energy consumption, i.e., the weighted sum of energy consumption and delay.
2) In each group l, the non-negative weight factors α_l and β_l are respectively introduced to tradeoff the energy consumption and delay. In order to meet the specific demands of

users, different users are allowed to choose different weight factors. Practically, if a user is in a low battery state, it would choose a larger α_l, i.e., put more weight on the energy consumption, to save more energy. Similarly, in case that a user is running a delay sensitive application, the user would choose a larger β_l, i.e., put more emphasise on the time delay, to reduce the latency.

3) The formulated problem is meaningful to the practical systems and exploits the full advantage of the combination of NOMA and MEC.

Second, thank you for mentioning the issue of the assumptions in this paper, which stimulates us to describe the assumptions more carefully. On the one hand, the assumption that the most demanding task-deadline should be served first is actually for the OMA scheme. In the OMA MEC system, each user occupies a dedicated time slot and the most demanding task-deadline should be served first. Differently, in the proposed hybrid NOMA MEC system, all the users are allowed to simultaneously offload their tasks to the server. Therefore, this assumption is not for our work. In the revised paper, we have removed this assumption for hybrid NOMA MEC in the second paragraph in Section II-A as follows:

"Hence, $\mathrm{UE}_{1,l}$ has the most demanding deadline and $\mathrm{UE}_{N_l,l}$ has the least demanding deadline."

On the other hand, the assumption on the delay, i.e., $D_{1,l} \leq D_{2,l} \leq \cdots \leq D_{N_l,l}$, $l = 1, \cdots, L$, is closely related to the proposed hybrid NOMA scheme. In our paper, in each group l, the users are ordered according to their deadlines. Specifically, for $N_l = 2$, $l = 1, \cdots, L$, i.e., each group has two users, $\mathrm{UE}_{1,l}$ has a more demanding deadline and $\mathrm{UE}_{2,l}$ has a less demanding deadline. After using successive interference cancelation (SIC), the message of $\mathrm{UE}_{1,l}$ can be decoded by removing $\mathrm{UE}_{2,l}$'s message while $\mathrm{UE}_{2,l}$ treats $\mathrm{UE}_{1,l}$'s message as interference. Moreover, given $D_{1,l} \leq D_{2,l}$, $\mathrm{UE}_{2,l}$ first offloads its task by sharing a time slot with $\mathrm{UE}_{1,l}$, and then solely offloads during a time interval. Practically, there always exist tasks with various deadlines and the proposed hybrid NOMA MEC can be used to save energy consumption and delay.

In addition, in this paper, the time for the edge-server to send the outcomes of the task and compute the task is omitted, as it is negligibly small than the offloading costs. In the revised paper, we have stated this in the first paragraph in Section II-B as follows:

"Similar to [25][26], the time cost for the server to send the outcomes of the task to the users and compute the tasks is omitted, which is negligibly small compared to the considered offloading costs. "

We sincerely thank the reviewers for their helpful comments. To address the reviewers'

concerns on the properties of proposed algorithm, we have added a subsection, i.e., Section IV-B, to theoretically analyze the properties of the proposed user grouping algorithm. The stability of the user grouping algorithm is proved in the following Proposition:

"**Proposition 6.** The final matching generated by the user grouping algorithm is 2ES."

Note that the global optimal solution needs exhaustive search with high complexity. In this paper, the proposed algorithm solution greatly saves the costs with low complexity. The following lemma has been added to prove that the proposed user grouping algorithm indeed improves the performance.

"**Lemma 1:** The WSED of the system decreases after each swap operation."

To show that the convergence of the proposed algorithm, in the revised paper, we have added the following Proposition:

"**Proposition 5.** Given any initial matching, the user grouping algorithm can always converge to a stable matching."

The complexity of the user grouping algorithm depends on the number of cycles in the swap operation. Considering the worst case of the user grouping algorithm, the complexity is illustrated in the following Proposition:

"**Proposition 7.** Given a number of cycles C, the computational complexity of the user grouping algorithm is given as $\mathcal{O}\left(CN^2\right)$ in the worst case."

To make this response letter not too long, the proofs of the added Lemma and Propositions are not shown here and we would like to refer you to Section IV-B of the revised version for details.

Finally, we would like to thank you again for your positive decision and constructive comments, which have helped us improve the quality of the paper.

RESPONSE TO REVIEWER 1

Comment: *The paper considers the special case of MEC offloading, when a mix of OMA and NOMA can be applied, that is, for a time period transmitters transmit at the same time, and this period is followed by another one, when only one of the transmitters is transmitting. The overall objective is to minimize transmission power under delay constraints. This is mixed with a problem formulation where the decrease of both the power and the transmission time is of interest. The paper is well in line with the MEC and NOMA research in recent years.*

While the topic is of interest, there are several weaknesses that decrease the value of the presented work.

Response: Thank you very much for acknowledging our contributions and also for your insightful comments, which helped us improve the paper quality. We have revised our paper following your suggestions.

Comment: 1. *I do not feel that the problem formulation with the weighted sum of energy consumption and delay is motivated in this paper. I would suggest to skip that part, or motivate it seriously (note that the maximum transmission delay is an input parameter for the optimization). In this case, however, the remaining contribution is not that much (mainly the proof of equation 25).*

Response: Thank you for the comment. It has been proved that integrating MEC and NOMA can not only avoid sever delay but also reduce energy consumption [19][20]. Hence, in this paper, we consider optimizing both of the energy consumption and delay, i.e., the weighted sum of energy consumption and delay. Specifically, in each group l, the non-negative weight factors α_l and β_l are respectively introduced to tradeoff the energy consumption and delay. In order to meet the specific demands of users, different users are allowed to choose different weight factors. For example, if a user is in a low battery state, it would choose a larger α_l, i.e., put more weight on the energy consumption, to save more energy. Similarly, in case that a user is running a delay sensitive application, the user would choose a larger β_l, i.e., put more emphasise on the time delay, to reduce the latency. Therefore, the problem formulation is meaningful to the practical systems and exploits the full advantage of the combination of NOMA and MEC.

Comment: 2. *The key element of the problem formulation is that the transmission of one of the messages in the NOMA scheme is not optimized at all. It is assumed, that this transmission uses the entire time until its deadline. The authors should show that this is the optimal choice.*

Response: Thank you for the comment. In this paper, we considered the hybrid NOMA MEC strategy where $\mathrm{UE}_{2,l}$ shares $D_{1,l}$ with $\mathrm{UE}_{1,l}$ and then continuously transmits for another time interval $t_l \geq 0$ after $D_{1,l}$. It is worthy pointing out that $\mathrm{UE}_{1,l}$ achieves the same performance as in OMA because the message of $\mathrm{UE}_{2,l}$ is decoded first. By exploiting successive interference cancelation (SIC), the message of $\mathrm{UE}_{1,l}$ can be decoded by removing $\mathrm{UE}_{2,l}$'s message and hence we obtain

$$D_{1,l}B\ln\left(1+p_{1,l}^{OMA}\left|h_{1,l}\right|^2\right)=M,$$

implying

$$p_{1,l}^{OMA}=\frac{e^{\frac{M}{D_{1,l}B}}-1}{\left|h_{1,l}\right|^2},$$

which is a constant. Therefore, we did not optimize the transmission of $\mathrm{UE}_{1,l}$'s message in the hybrid NOMA MEC system.

Furthermore, in the following, we will prove that, finishing $\mathrm{UE}_{1,l}$'s transmission at its deadline is the optimal choice. Let k denote $\mathrm{UE}_{1,l}$'s transmission time satisfying $0 \leq k \leq D_{1,l}$. Then, we have

$$kB\ln\left(1+p_{1,l}^{OMA}\left|h_{1,l}\right|^2\right)=M,$$

and hence

$$p_{1,l}^{OMA}=\frac{e^{\frac{M}{kB}}-1}{\left|h_{1,l}\right|^2}. \tag{1}$$

which is decreasing with k and achieves the minimum value when $k = D_{1,l}$. In addition, in group l, the hybrid NOMA MEC scheme implies the delay is $D_l = D_{1,l} + t \geq D_{1,l}$. Therefore, to achieve the best performance, $\mathrm{UE}_{1,l}$ shall finish its transmission at its deadline.

Comment: 3. *The authors should elaborate a bit on hybrid NOMA, that requires that a part of a message is decoded, before the entire message arrives (message of UE_2 in the paper). Is it always possible or does it require special coding schemes?*

Response: Thank you for the comment. The scale of task is generally much larger than that of a packet. In fact, the task of $\mathrm{UE}_{2,l}$ contains a lot of small packets and each small packet can be independently encoded and decoded. Thus, in the proposed hybrid NOMA, in each group l, some small packets in $\mathrm{UE}_{2,l}$'s task are first decoded within $D_{1,l}$ and the remaining packets are then been decoded within the time interval t_l.

Comment: 4. *The authors state that the problem of optimal power and time allocation is a non-convex problem. It needs to be proved. Moreover, the statement then that "However in the following we are able to find the optimal power and time allocation in closed form" needs justification. Specifically, the results of subsections A and B need to be coupled to get the final solution.*

Response: Thank you for the comment. The non-convexity of the formulated problem (16) lies on the objective function and constraint (17). Firstly, the objective function is given as

$$C_l\left(p_{2,l}^1, p_{2,l}^2, t_l\right) = \alpha_l\left(D_{1,l}p_{2,l}^1 + t_l p_{2,l}^2\right) + \beta_l t_l,$$

whose Hessian matrix is

$$H_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \alpha_l \\ 0 & \alpha_l & 0 \end{bmatrix}.$$

Obviously, H_1 is not a positive semidefinite matrix and hence function $C_l\left(p_{2,l}^1, p_{2,l}^2, t_l\right)$ is not convex. On the other hand, the constraint (17) is given as

$$D_{1,l}B\ln\left(1+\frac{\left|h_{2,l}\right|^2 p_{2,l}^1}{\left|h_{1,l}\right|^2 p_{1,l}^{OMA}+1}\right) + t_l B\ln\left(1+\left|h_{2,l}\right|^2 p_{2,l}^2\right) - M \geq 0,$$

which is also not convex and the reason is given as follows. Denote

$$y\left(p_{2,l}^1, p_{2,l}^2, t_l\right) = D_{1,l}B\ln\left(1+\frac{\left|h_{2,l}\right|^2 p_{2,l}^1}{\left|h_{1,l}\right|^2 p_{1,l}^{OMA}+1}\right) + t_l B\ln\left(1+\left|h_{2,l}\right|^2 p_{2,l}^2\right) - M.$$

Then the Hessian matrix of $y\left(p_{2,l}^1, p_{2,l}^2, t_l\right)$ is given by

$$H_2 = \begin{bmatrix} \frac{-D_{1,l}B\left|h_{2,l}\right|^4}{\left(\left|h_{1,l}\right|^2 p_{1,l}^{OMA}+\left|h_{2,l}\right|^2 p_{2,l}^1+1\right)^2} & 0 & 0 \\ \frac{-t_l B\left|h_{2,l}\right|^4}{\left(1+\left|h_{2,l}\right|^2 p_{2,l}^2\right)^2} & 0 & \frac{B\left|h_{2,l}\right|^2}{1+\left|h_{2,l}\right|^2 p_{2,l}^2} \\ 0 & 0 & \frac{B\left|h_{2,l}\right|^2}{1+\left|h_{2,l}\right|^2 p_{2,l}^2} \end{bmatrix}.$$

One of H_2's eigenvalues is

$$\lambda = \frac{-D_{1,l}B\left|h_{2,l}\right|^4}{\left(\left|h_{1,l}\right|^2 p_{1,l}^{OMA}+\left|h_{2,l}\right|^2 p_{2,l}^1+1\right)^2} < 0,$$

and hence H_2 is not a positive semidefinite matrix. Therefore the function $y\left(p_{2,l}^1, p_{2,l}^2, t_l\right)$ is not convex, i.e. the constraint (17) is not convex.

In addition, we would like to emphasize that the obtained closed form solution is optimal for problem (16). Here, we illustrate the optimality of our solution. First of all, the original optimization problem (16) is given as

$$\min_{p_{2,l}^1, p_{2,l}^2, t_l} \ \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2 \right) + \beta_l t_l, \tag{2}$$

$$\text{s.t. } D_{1,l} B \ln \left(1 + \frac{|h_{2,l}|^2 p_{2,l}^1}{|h_{1,l}|^2 p_{1,l}^{OMA} + 1} \right),$$

$$+ t_l B \ln \left(1 + |h_{2,l}|^2 p_{2,l}^2 \right) \geq M, \tag{3}$$

$$0 \leq t_l \leq D_{2,l} - D_{1,l}, \tag{4}$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0. \tag{5}$$

Then, problem (2) can be decomposed into two subproblem, one is the power optimization problem for given t_l:

$$\min_{p_{2,l}^1, p_{2,l}^2} \ \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2 \right), \tag{6}$$

$$\text{s.t. } D_{1,l} B \ln \left(1 + |h_{2,l}|^2 p_{2,l}^1 e^{\frac{-M}{D_{1,l} B}} \right)$$

$$+ t_l B \ln \left(1 + |h_{2,l}|^2 p_{2,l}^2 \right) \geq M,$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0,$$

which is convex and the optimal power solutions are characterized in Theorem 1. The other is the time optimization problem:

$$\min_{t_l} \ C\left(t_l\right), \tag{7}$$

$$\text{s.t. } 0 \leq t_l \leq D_{2,l} - D_{1,l}, \tag{8}$$

where $C\left(t_l\right)$ is the optimal objective value of (6) with given t_l and given by

$$C\left(t_l\right) = \frac{\alpha_l}{|h_{2,l}|^2} \left(D_{1,l} e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} + t_l \left(e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1 \right) \right) + \beta_l t_l. \tag{9}$$

There is not any loss of optimality in the above process. Furthermore, it can be verified that $C\left(t_l\right)$ is convex. Therefore, (7) is a convex problem, whose solution can be efficiently found via standard convex optimization tools, e.g., CVX. Nevertheless, in our paper, we analytically characterized the optimal solution in Proposition 3. Therefore, our solution, i.e., the power and time allocation, is optimal.

Follow your advice, in the revised paper, to clarify the non-convexity of the original problem and the optimality of the proposed solutions, we have added the content below problem (16) in the first paragraph in Section III as follows:

"The non-convexity lies on the objective function and constraint (17). In the following, we first achieve the optimal power allocation of users in each group, which can be expressed as functions of the time interval of each group. Then, we further optimize the time intervals and thus obtain the optimal power and time allocation, which can be characterized in a closed form. "

Comment: 5. *The authors propose the grouping of the users based on matching, formulating the problem as matching with externalities. The theoretical content of this part is very weak and should be improved significantly:*

- Why is this a good approach to form the pairs? In which sense will the result be good?

- The authors write that "Considering the externalities, the stable matching is difficult to obtain." Difficult in which sense? If it is very difficult, why to follow this approach? The difficulty level should be defined formally, and the decision to use this method should be motivated.

- As I understand, there are conditions for the matching with externalities to stabilize. Are these conditions met in this case?

- Finally, I am not sure that the steps of the matching process needs to be described, since it is given in the related literature.

Response: Thank you for the comment. Firstly, the user grouping problem is actually an integer programming. The global optimal solution can only be found with exhaustive method, whose complexity increases exponentially with the number of users. Hence, in this paper, we consider the user grouping as a two-side many to one matching. The groups and users act as two sets of players and interact with each other to maximize the weighted sum of energy consumption and delay. By using the matching theory, we provided a mathematically tractable and low-complexity solution for the combinational problem. In addition, in the revised paper, we have added a lemma to illustrate that the system performance is improved by using the proposed algorithm, which is given by

"Lemma 1. The WSED of the system decreases after each swap operation."

To make this response letter not too long, the proofs of this added Lemma and the following Proposition are not shown here and we would like to refer you to Section IV-B of the revised

version for details.

Secondly, in the presence of externalities, a deviating group need to consider the reactions of the other users because they may affect the blocking possibility of the group. Therefore, the final solution is difficult to obtain. The reason why we chose the matching method is the proposed algorithm is able to achieve a two-side exchange stability with limited number of iterations. Furthermore, the system weighted sum of energy consumption and delay decreases after each swap operation and hence the solution greatly improves the system performance. In the revised paper, we have described the difficulty and motivation of the proposed method in the sixth paragraph in Section IV-A:

"Considering the externalities, the stable matching is difficult to obtain. The reason is, with externalities, the reactions of the users not in the group may affect the blocking possibility of the group. In order to guarantee all the users are well matched, in the following, we will propose the user grouping algorithm, which can achieve the solution with stability and low complexity. The concepts of two-sided exchange matching and two-sided exchange stability [34] are exploited in the matching process."

Then, regarding the stability of the proposed algorithm, in the revised paper, we have proved that the proposed user grouping algorithm can always achieve the stability. Proposition 6 has been added to illustrate the stability of the proposed algorithm in Section IV-B as follows:

"**Proposition 6.** The final matching generated by the user grouping algorithm is 2ES."

Finally, in this paper, we briefly described the steps of the matching process to make this work integral. Although this matching theory has been widely used in the literature, we used this method for different objective, i.e., minimizing the weighted sum of energy consumption and delay. Moreover, the provided optimal power and time solutions are integrated with the user grouping algorithm. Therefore, I think it is necessary to briefly describe the steps of the matching process.

Comment: *6. The part on Theorem 1 and related Corollaries and Propositions should be rewritten. It is very hard to follow, and feels to be unnecessarily complicated. Among others, no intuitive explanations should be given after formal proofs, they do not feel well justified (e.g., the texts after Corollaries 1 and 2 and Proposition 1). I would also move the proofs to the main body of the paper.*

Response: Thank you for the advice. Regrading to Theorem 1, since the optimal power

solutions, i.e., $p_{2,l}^{1*}$ and $p_{2,l}^{2*}$, depends on the value of t_l, the description of Theorem 1 is a bit complex. We are vey sorry about the complex form but we have tried our best to clarify the Theorem. In the revised paper, we have updated this Theorem in a compact form.

Moreover, we have rewritten the following Corollaries, Proposition and the explanations after Corollary 1, Corollary 2, and Proposition 1 in Section III-A:

"**Corollary 1.** For problem (20), in the hybrid NOMA case, we always have $p_{2,l}^{2*} > p_{2,l}^{1*}$.

According to Corollary 1, in the hybrid NOMA case, $\mathrm{UE}_{2,l}$ is allocated with more power during t_l than during $D_{1,l}$, which is in line with our expectation. Actually, in the hybrid NOMA case, $\mathrm{UE}_{2,l}$ experiences no interference during t_l while it is interfered by $\mathrm{UE}_{1,l}$ during $D_{1,l}$. Therefore, $\mathrm{UE}_{2,l}$ allocates a higher power during t_l to have a lower energy consumption**.**

Corollary 2. For problem (20), in each group l, $E_l^{NOMA} \leq E_l^{OMA}$ if and only if $D_{2,l} < 2D_{1,l}$.

From Corollary 2, given $D_{2,l} < 2D_{1,l}$, the NOMA scheme achieves a higher performance. In the following Proposition, we will further investigate the superiority of NOMA over OMA.

Proposition 1. For problem (20), given $D_{2,l} < 2D_{1,l}$, we have

$$\Delta\left(t_l\right) = E^{OMA} - E^{H-NOMA} = \frac{\left(D_{1,l}+t_l\right)e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - \left(D_{1,l}e^{\frac{M}{BD_{1,l}}} + t_l e^{\frac{M}{Bt_l}}\right)}{\left|h_{2,l}\right|^2}, \tag{10}$$

which is a monotonically non-increasing function and satisfies $\Delta\left(t_l\right)_{\max} = \Delta\left(D_{2,l} - D_{1,l}\right) < 0$.

Proposition 1 suggests that for $D_{2,l} < 2D_{1,l}$, the largest gap between hybrid NOMA MEC and OMA MEC is achieved at $t_l = D_{2,l} - D_{1,l}$. Therefore, the optimal strategy is $\mathrm{UE}_{2,l}$ shall consume all its time until its deadline. "

Corollary 3. For problem (16), given $D_{2,l} < 2D_{1,l}$, $\alpha_l = 1$, and $\beta_l = 0$, the optimal time solution is $t_l^* = D_{2,l} - D_{1,l}$."

In addition, considering the content of the proofs is too much, especially the proof of Theorem 1, we think it would better put these proofs in the Appendix.

Comment: *7. Proof of Corollary one seem to use that* $\frac{M}{BD_{1,l}} < 1$. *Is it always true?*

Response: Thank you for the query. We kindly disagree with your opinion. We always have $\frac{M}{BD_{1,l}} \geq 0$ and hence $e^{\frac{M}{BD_{1,l}}} \geq 1$. Therefore, in the proof of Corollary 1, we have

$$p_{2,l}^{1*} - p_{2,l}^{2*} = \frac{e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - e^{\frac{M}{BD_{1,l}}}}{\left|h_{2,l}\right|^2} - \frac{e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1}{\left|h_{2,l}\right|^2} = \frac{1 - e^{\frac{M}{BD_{1,l}}}}{\left|h_{2,l}\right|^2} \leq 0,$$

implying $p_{2,l}^{1*} \leq p_{2,l}^{2*}$.

Comment: *8. Page 5: is the exact path loss model (Rayleigh distribution) is relevant?*

Response: Thank you for the query. In this work, similar to [27][28], the channels are assumed to be independent and identically distributed.

Comment: *9. Page 4 and 5: in the model the channel coefficient seem to depend on the group (as well as on the node). Does it mean that a group is actually defined by a sub-channel? This should be clearly stated. This would make the section on user grouping easier to understand as well.*

Response: Thank you for the advice. We agree with your opinion. In this paper, each group is actually defined by a sub-channel. In the revised paper, we have clearly stated this in the first paragraph in Section IV-A:

"Actually, each group is defined by a subchannel and two users are allocated on each sub-channel."

Comment: *10. Page 14, definition 1: notation needs to be defined.*

Response: Thank you for reminding us. Φ denotes a two-to-one matching and is a mapping from all the subsets of users $\mathbb{N}$ into the groups set $\mathbb{L}$. In the revised paper, the Definition 1 in Section IV-A has been updated as follows.

"Definition 1. A two-to-one matching Φ is a mapping from all the subsets of users $\mathbb{N}$ into the groups set $\mathbb{L}$, satisfying the following properties for $\mathrm{UE}_n \in \mathbb{N}$ and $C_l \in \mathbb{L}$

(a) $\Phi(\mathrm{UE}_n) \in \mathbb{L}$;

(b) $\Phi(C_l) \subseteq \mathbb{N}$;

(c) $|\Phi(\mathrm{UE}_n)| = 1, |\Phi(C_l)| = 2$;

(d) $C_l \in \Phi(\mathrm{UE}_n) \Longleftrightarrow \mathrm{UE}_n = \Phi(C_l)$."

Comment: *11. Page 18 and on: the various terms of delay are not explained. What is D_max? What is Maximum deadline? What is Total Delay? How can the Total Delay be more than the Maximum Deadline?*

Response: Thank you for the comment. In this paper, the maximum deadline is $\max\{D_{2,l}\}_{l=1}^{L}$. Total deadline is given by $D = \sum_{l=1}^{L}(D_{1,l} + t_l)$, which is the sum deadline of the users through all groups and hence can be more than the maximum deadline. In the revised paper, we have

added the explanations of the maximum deadline and total delay in the second paragraph in Section V as follows:

"In these two figures, the maximum deadline is written as $\max\{D_{2,l}\}_{l=1}^{L}$ and the total delay is given by $D = \sum_{l=1}^{L}(D_{1,l} + t_l)$."

Comment: *12. Results: It would be nice if the numerical results would demonstrate the theoretical results of the paper, e.g. the results of Theorem 1.*

Response: Thank you for the advice. The following figure displays the numerical and theoretical results of the paper, i.e., the optimal power and time allocation proposed in Theorem 1 and Proposition 3. In this figure, the numerical solution is computed via the interior point method. One can observe that the theoretical solutions perfectly match the numerical ones, indicating the accuracy of the theoretical solutions.

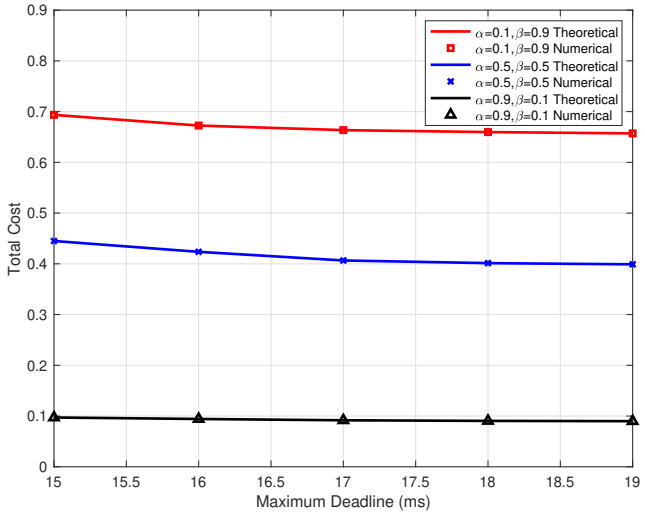

Figure 1. Total cost versus the maximum deadline.

However, due to the limited space, we are very sorry that we cannot put this figure in the revised paper.

RESPONSE TO REVIEWER 2

Comment: *The paper investigates the technologies of hybrid NOMA and MEC, in which a user may firstly offload parts of its task by sharing a time slot with another user, and then solely offloads the remaining task during a time interval. This paper focuses on the downlink hybrid NOMA MEC systems, where multiple users are classified into different groups and each group is allocated a dedicated time slot. In order to achieve a tradeoff between energy consumption and delay, the paper introduces the weight factors and the resources of power, time, and user grouping are optimized to minimize the weighted sum of energy consumption and delay. In particular, the paper characterizes the optimal power and time allocation in closed form. By incorporating the matching algorithm with the optimal power and time allocation, the paper proposes an efficient method to optimize user grouping. Overall, the topic of this paper is very interesting and timely. The proposed modeling of hybrid NOMA and MEC is interesting, and the design of solution methodology for finding the solution is novel. The whole paper is also well organized and presented. The reviewer just has some comments that may help improve the quality of this paper.*

Response: We sincerely thank the reviewer for sacrificing time to review our paper and acknowledging our contributions. In the following, we will carefully address your concerns.

Comment: *1. In the modeling part, the paper assumes that the UE with the most demanding task-deadline should be served first. The reviewer is curious whether this is a must-assumption for the following analysis in this paper or not. What if the other orders are used? In addition, the paper mainly focuses on the transmission delay in MEC, while the computation delay at the edge-server is not considered. This point should be clearly stated in the system model.*

Response: Thank you for the comment. Actually, the assumption that the user with the most demanding deadline should be served first is used in the conventional OMA scheme. Differently, in the NOMA MEC system, all the users are allowed to simultaneously offload their tasks to the server. In the revised paper, we have removed this assumption for hybrid NOMA MEC in the second paragraph in Section II-A as follows:

"Hence, $\mathrm{UE}_{1,l}$ has the most demanding deadline and $\mathrm{UE}_{N_l,l}$ has the least demanding deadline."

In addition, in this paper, similar to [25][26], the time for the edge-server to compute is omitted, as it is negligibly small compared to the considered offloading costs. Following your advice, we have stated this in the first paragraph in Section II-B as follows:

"Similar to [25][26], the time cost for the server to send the outcomes of the task to the users and compute the tasks is omitted, which is negligibly small compared to the considered offloading costs. "

Comment: *2. Above eq. (7), the paper states that "$D_{1,l}$ and $p_{1,l}^{OMA}$ are both constants, ... ". Please explain why are they both constant? To minimize the total consumption and delay, I feel that we can also jointly tune the values of $D_{1,l}$ and $p_{1,l}^{OMA}$. Specifically, in order to minimize the total energy consumption and delay, $UE_{1,l}$ probably can finish its transmission earlier than $D_{1,l}$.*

Response: Thank you for the comment. $D_{1,l}$ is the given deadline of $\text{UE}_{1,l}$ and hence is a constant. In addition, $\text{UE}_{1,l}$ achieves the same performance as in OMA because the message of $\text{UE}_{2,l}$ is decoded first. By exploiting SIC, the message of $\text{UE}_{1,l}$ can be decoded by removing $\text{UE}_{2,l}$'s message and thus we have

$$D_{1,l}B\ln\left(1+p_{1,l}^{OMA}\left|h_{1,l}\right|^2\right)=M,$$

implying

$$p_{1,l}^{OMA}=\frac{e^{\frac{M}{D_{1,l}B}}-1}{\left|h_{1,l}\right|^2},$$

which is also a constant.

Furthermore, in the following, we will prove that, finishing $\text{UE}_{1,l}$'s transmission at its deadline is the optimal choice. Let k denote $\text{UE}_{1,l}$'s transmission time satisfying $0\leq k\leq D_{1,l}$. Therefore, we obtain

$$kB\ln\left(1+p_{1,l}^{OMA}\left|h_{1,l}\right|^2\right)=M,$$

and hence

$$p_{1,l}^{OMA}=\frac{e^{\frac{M}{kB}}-1}{\left|h_{1,l}\right|^2}. \tag{11}$$

Note that, from (11), $p_{1,l}^{OMA}$ is decreasing with k and achieves the minimum value when $k=D_{1,l}$. In addition, in group l, the hybrid NOMA MEC scheme implies the delay is $D_l=D_{1,l}+t\geq D_{1,l}$. Therefore, to achieve the minimum weighted sum of energy consumption and delay, $\text{UE}_{1,l}$ shall finish its transmission at its deadline.

Comment: *3. Proposition 4 is an interesting result in this paper. Could this paper provide any explanations on the rationale behind Proposition 4?*

Response: Thank you for the advice. Firstly, given $D_{2,l} \geq 2D_{1,l}$, which corresponds to a scenario in which $\text{UE}_{2,l}$ has less demanding delay requirements, OMA MEC will yield a better performance than NOMA MEC. Therefore, this proposition provides the optimal time allocation in the OMA MEC case. Furthermore, according to Proposition 4, the values of α_l and β_l will affect the optimal time allocation. In the revised paper, to better illustrate the rationale behind Proposition 4, we have rewritten the last paragraph in Section III-B as follows:

"In Proposition 4, the condition $D_{2,l} \geq 2D_{1,l}$ indicates that OMA MEC yields a better performance than NOMA MEC and the proposed optimal time solution is for the OMA MEC case. Furthermore, in order to achieve the minimum WSED, the optimal time solution is closely connected with the value of $\frac{\beta_l}{\alpha_l}$. Specifically, if α_l is large and β_l is small, the optimal value of t_l will be large. This is in line with our expectation because more weight is given to the energy consumption. Conversely, in the case when α_l is small and β_l is large, i.e., the system focuses more on the delay minimization, the optimal value of t_l will be small."

Comment: *4. Algorithm 1 is an important contribution in this paper for finding the grouping of UEs for matching. However, Algorithm 1 is somewhat intuitive, and just for the purpose of reaching a stable matching solution. It will be more interesting for the paper to discuss how Algorithm 1can help to achieve the matching solution that can also minimize the total cost function of all users as in (8).*

Response: Thank you for the comment. Actually, by using the proposed algorithm, the matching solution is not only stable but also minimizes the total cost function of all users. Note that the global optimal solution can only be found with exhaustive search with high complexity. The proposed algorithm solution can greatly save the cost with low complexity. In the revised paper, we have added a Lemma which proves that after each swap operation, the system cost decreases. The Lemma is given in Section IV-B as follows:

"Lemma 1. The WSED of the system decreases after each swap operation.

To make this response letter not too long, the proof of this added Lemma is not shown here and we would like to refer you to Section IV-B of the revised version for details. Moreover, in Section IV-B, we have also provided some theoretical analysis results and more discussions on Algorithm 1 such as its scalability, convergence, and complexity.

Comment: *5. For the sake of clear presentation, it is better to use a table to summarize all used notations in this paper, since many notations are used in this paper. Similar suggestion*

holds for the detailed parameter-settings in the section of numerical results.

Response: Thank you for the advice. Following your advice, we have added Table I to summarize the parameter-settings, which is given as follows:

Table I

TABLE OF PARAMETERS

AWGN spectral density	$N_0 = -174$dBm
Path loss exponent	$v = 3$
Bandwidth	1MHz
Cell radius	100m

In addition, we are very sorry that we cannot add the table to summarize all used notations due to the limited space. However, we have clearly denoted all the notations in the revised paper.

Comment: *6. A minor typo: "... assume the user grouping is given, we first find ..." should read "... assuming that the user grouping is given, we first find ...".*

Response: Thank for reminding this. We apologized for the grammar mistake. In the revised paper, the typos have been corrected and we have carefully checked the typo and grammatical errors through the paper draft.

RESPONSE TO REVIEWER 3

Comment: *This paper investigates the resource allocation for the downlink hybrid NOMA MEC systems, where multiple users are classified into different groups and each group is allocated a dedicated time slot. The closed form expression for the optimal power and time allocation is derived. An efficient method to optimize user grouping is proposed. Simulation results show that the proposed resource allocation method can achieve quite close performance as global optimal value. Overall, this paper studies an interesting topic, which is timely and novel. Nevertheless, the reviewer has the following concerns, which suggests a major revision for this paper.*

Response: We sincerely thank the reviewer for sacrificing time to review our paper and acknowledging our contributions. Your concerns are carefully addressed in the following response.

Comment: *1. The abstract needs to be further polished, especially to make the contribution more clear, as well as reducing the trivial discussions about state-of-art.*

Response: Thank you for the advice. Following your advice, the abstract has been rewritten as follows:

"Non-orthogonal multiple access (NOMA) and mobile edge computing (MEC) have been recognized as promising technologies for the beyond fifth generation networks to achieve significant capacity improvement and delay reduction. In this paper, the technologies of hybrid NOMA and MEC are integrated. In the hybrid NOMA MEC system, multiple users are classified into different groups and each group is allocated a dedicated time slot. In each group, a user first offloads its task by sharing a time slot with another user, and then solely offloads during a time interval. To reduce the delay and save the energy consumption, we consider jointly optimizing the power and time allocation in each group as well as the user grouping. As the main contribution, the optimal power and time allocation is characterized in closed form. In addition, by incorporating the matching algorithm with the optimal power and time allocation, we propose a low complexity method to efficiently optimize user grouping. Simulation results demonstrate that the proposed resource allocation method in the hybrid NOMA MEC systems not only yields better performance than the conventional OMA scheme but also achieves quite close performance as global optimal solution."

Comment: *2. In Introduction section, the authors should provide more discussions on why blending the concept of hybrid NOMA and MEC technique, and what is the main application*

in future.

Response: Thank you for the advice. In this paper, the reason why we focused on the hybrid NOMA MEC scheme is explained as follows. Firstly, compared to OMA, NOMA allows two users to simultaneously offload their tasks to the server during $D_{1,l}$ and hence the delay is smaller. Nevertheless, [21] has pointed that in group l, $\mathrm{UE}_{2,l}$ needs to consume more energy in NOMA than in OMA if $\mathrm{UE}_{2,l}$ completely relies on $D_{1,l}$. Therefore, the hybrid NOMA MEC scheme was studied in this work.

Practically, the hybrid NOMA MEC can often be used when users have different deadlines. In the hybrid NOMA MEC system, the time and energy resources are saved a lot by reusing the transmission time of the users who have urgent tasks.

In the revised paper, we have added the reason why we studied the hybrid NOMA MEC scheme and the main application in future in the fifth paragraph in Section I as:

"The hybrid NOMA MEC not only outperforms OMA in terms of delay but also achieves lower energy consumption than NOMA. Practically, by using the hybrid NOMA MEC offloading scheme, the resources of time and energy can be saved for the users with various deadlines. "

Comment: *3. According to Problem (12), this paper actually studies a user pairing problem instead of user grouping. The authors should clarify this issue. Moreover, if it is user pairing, how about the scalability of the proposed user pairing algorithms. Please give more discussions.*

Response: Thank you for the comment. Actually, from the formulated problem (12), the user grouping problem can be equivalently transformed into a user pairing problem. For the hybrid NOMA MEC system, we focus on dividing all users into small groups and two users are in each group. Therefore, the users can be each paired or can be divided into groups. However, in the proposed user grouping algorithm, we exploit two-to-one matching rather than one-to-one matching. Hence, in this paper, we used the concept of user grouping rather than user pairing.

In addition, regrading the scalability, in the revised paper, we have added the content which depicts the scalability of the proposed user grouping algorithm. The added proposition is given in Section IV-B:

"**Proposition 6.** The final matching generated by the user grouping algorithm is 2ES. "

To make this response letter not too long, the proofs of this added proposition and the following propositions are not shown here and we would like to refer you to Section IV-B of the revised

version for details. Thus, the proposed user pairing algorithm can always achieve a two-side exchange-stable matching, i.e., the scalability is guaranteed.

Comment: *4. The optimal solutions given in Section III are not the global optimal one, since either time variable or power and user grouping variables are fixed. Please clarify this issue in the paper.*

Response: Thank you for the comment. We kindly disagree with your opinion. In each group l, the proposed power and time solution is optimal and the reason is given as follows. The original optimization problem (16) is written as

$$\min_{p_{2,l}^1, p_{2,l}^2, t_l} \ \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2\right) + \beta_l t_l, \tag{12}$$

$$\text{s.t. } D_{1,l} B \ln\left(1 + \frac{|h_{2,l}|^2 p_{2,l}^1}{|h_{1,l}|^2 p_{1,l}^{OMA} + 1}\right),$$

$$+ t_l B \ln\left(1 + |h_{2,l}|^2 p_{2,l}^2\right) \geq M, \tag{13}$$

$$0 \leq t_l \leq D_{2,l} - D_{1,l}, \tag{14}$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0. \tag{15}$$

Problem (12) can be decomposed into two subproblem, one is the power optimization problem for given t_l:

$$\min_{p_{2,l}^1, p_{2,l}^2} \ \alpha_l \left(D_{1,l} p_{2,l}^1 + t_l p_{2,l}^2\right), \tag{16}$$

$$\text{s.t. } D_{1,l} B \ln\left(1 + |h_{2,l}|^2 p_{2,l}^1 e^{\frac{-M}{D_{1,l} B}}\right)$$

$$+ t_l B \ln\left(1 + |h_{2,l}|^2 p_{2,l}^2\right) \geq M,$$

$$p_{2,l}^1 \geq 0, p_{2,l}^2 \geq 0,$$

which is convex and the optimal power solutions are characterized in Theorem 1. The other is the time optimization problem:

$$\min_{t_l} \ C\left(t_l\right), \tag{17}$$

$$\text{s.t. } 0 \leq t_l \leq D_{2,l} - D_{1,l}, \tag{18}$$

where $C\left(t_l\right)$ is the optimal objective value of (16) with given t_l and is given by

$$C\left(t_l\right) = \frac{\alpha_l}{|h_{2,l}|^2} \left(D_{1,l} e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} + t_l \left(e^{\frac{2M}{B\left(D_{1,l}+t_l\right)}} - 1\right)\right) + \beta_l t_l. \tag{19}$$

There is not any loss of optimality in the above process. Moreover, $C(t_l)$ is a convex function and hence (17) is a convex problem, whose optimal solution in closed form was characterized in Proposition 3. Therefore, the proposed power and time allocation is optimal.

Comment: *5. Please analyze the convergence and complexity of the proposed matching algorithm.*

Response: Thank you for the comment. In the revised paper, we have added the analysis of convergence and complexity of the proposed algorithm in Section IV-B. The analysis of convergence is given by

"**Proposition 5.** Given any initial matching, the user grouping algorithm can always converge to a stable matching."

and the illustration of complexity is given by

"**Proposition 7.** Given a number of cycles C, the computational complexity of the user grouping algorithm is given as $\mathcal{O}(CN^2)$ in the worst case."

Comment: *6. This paper investigates the difference between OMA, pure NOMA and hybrid NOMA used in MEC system, and proposed several resource allocation algorithms for power, time and user grouping. One suggests that the authors should consider the fairness issue of the algorithm.*

Response: Thank you for the comment. In this paper, for every two users in the same group, we did not consider the weights of the users and the reason is given as follows. Note that $\mathrm{UE}_{1,l}$ in group l achieves the same performance as in OMA and its allocated power and time are respectively given by

$$p_{1,l}^{OMA} = \frac{e^{\frac{M}{D_{1,l}B}} - 1}{|h_{1,l}|^2},$$

and $D_{1,l}$, which are both constants. Therefore, we just need to optimize the power and time allocation for $\mathrm{UE}_{2,l}$ in each group l. Therefore, in the proposed algorithm, we did not consider the fairness issue.

Comment: *7. The authors need to carefully check the grammar mistakes and typos. For instance, in page 2, line 34, there is a mistake that "catching", which should be "caching".*

Response: Thank you for reminding this. We apologize for the grammar mistakes and typos. In the revised paper, we have corrected the typos and also proofread the manuscript more carefully to avoid typos and grammar mistakes.