Digital Communications and Networks 6 (2020) 241–252

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

Joint resource allocation for hybrid NOMA-assisted MEC in 6G networks☆

Haodong Li [a,*], Fang Fang [b], Zhiguo Ding [a]

[a] Department of Electrical and Electronic Engineering, The University of Manchester, M13 9PL, UK
[b] Department of Engineering, Durham University, Durham DH1 3LE, UK

ARTICLE INFO

Keywords:
Non-orthogonal multiple access (NOMA)
Multi-access edge computing (MEC)
Resource allocation
User grouping
Task assignment

ABSTRACT

Multi-access Edge Computing (MEC) is an essential technology for expanding computing power of mobile devices, which can combine the Non-Orthogonal Multiple Access (NOMA) in the power domain to multiplex signals to improve spectral efficiency. We study the integration of the MEC with the NOMA to improve the computation service for the Beyond Fifth-Generation (B5G) and the Sixth-Generation (6G) wireless networks. This paper aims to minimize the energy consumption of a hybrid NOMA-assisted MEC system. In a hybrid NOMA system, a user can offload its task during a time slot shared with another user by the NOMA, and then upload the remaining data during an exclusive time duration served by Orthogonal Multiple Access (OMA). The original energy minimization problem is non-convex. To efficiently solve it, we first assume that the user grouping is given, and focuses on the one group case. Then, a multilevel programming method is proposed to solve the non-convex problem by decomposing it into three subproblems, i.e., power allocation, time slot scheduling, and offloading task assignment, which are solved optimally by carefully studying their convexity and monotonicity. The derived solution is optimal to the original problem by substituting the closed expressions obtained from those decomposed subproblems. Furthermore, we investigate the multi-user case, in which a close-to-optimal algorithm with low-complexity is proposed to form users into different groups with unique time slots. The simulation results verify the superior performance of the proposed scheme compared with some benchmarks, such as OMA and pure NOMA.

1. Introduction

With the paradigm of cloud computing, the new trend is emerging to implement cloud computing at the edge of a cellular network with an Access Point (AP), also known as Multi-access Edge Computing (MEC) [2,3]. The conventional cloud computing system, however, has a centralized data center to handle the computation requests [4]. Extra latency is introduced due to the long physical distance between users and the computation center. MEC is motivated by many emerging applications, such as virtual reality, the Internet of Things (IoT), and smart vehicles, which are all computationally intensive and latency-sensitive tasks [5,6]. Moreover, communication, computation, caching, and control (C4) are key features in the Beyond Fifth-Generation (B5G) and sixth-generation (6G) mobile networks. Therefore, MEC can improve the computation service for future mobile networks [7,8].

The main idea of MEC is to deploy the powerful computing equipment at the edge of cellular networks. Unlike the conventional cloud computing, this configuration intends to shorten the physical distance between the user and the server; therefore, the transmission latency can be reduced [9,10]. Here is an example regarding the IoT. With the number of IoT devices approaching 50 billion, a large amount of raw sensor data will be generated and needs to be processed and transmitted [11]. Nevertheless, most IoT devices have limited battery life and computation power, and then it is difficult to handle an intensive task within the deadline if only local computing is performed. The assistance of the MEC could enable IoT devices to offload their raw sensor data to the BS. The MEC server can allocate powerful computing resources to calculate the task, and the IoT device can download the outcomes within the delay constraint [12].

On the other hand, although the computing power at the Base Station (BS) has been enhanced, in order to further improve the performance of computing services, the application of Non-Orthogonal Multiple Access (NOMA) is another alternative method from a communication perspective [13,14]. It is predicted by the International Telecommunication Union (ITU) that the mobile data traffic will approach 5 zettabytes (ZB) per month by 2030 [15,16]. To support the explosive traffic volume in

☆ Part of this paper has been submitted to the IEEE International Conference on Communications, Dublin, Ireland, Jun. 2020 [1].

* Corresponding author.
E-mail addresses: haodong.li@manchester.ac.uk (H. Li), fang.fang@durham.ac.uk (F. Fang), zhiguo.ding@manchester.ac.uk (Z. Ding).

https://doi.org/10.1016/j.dcan.2020.05.005
Received 28 December 2019; Received in revised form 8 May 2020; Accepted 18 May 2020
Available online 12 July 2020

the future, NOMA has become one of the promising Multiple Access (MA) technology for 5G and 6G because of spectral efficiency gain [15,17]. For conventional Orthogonal Multiple Access (OMA), each user is served with a dedicated time or frequency resource block. Unlike OMA, NOMA can significantly improve the spectral efficiency because it allows multiple users to multiplex on the same frequency band with different power levels during transmission [18,19]. At the receiver, the application of Successive Interference Cancellation (SIC) can remove some interference from other users.

Due to the superiority of the NOMA in terms of spectral efficiency, it is promising to integrate NOMA with MEC. In a NOMA-assisted MEC system, multiple mobile users can offload their tasks simultaneously on the same frequency resource. For example, we assume that only a one-time slot is unoccupied at the moment, and two users will offload their tasks to the BS. If the OMA transmission is applied, only one user can transmit, and the other must wait. If the NOMA is applied, both users can transmit to the BS simultaneously, which can reduce the latency caused by the shortage of radio resources. Hence, the computing service in B5G and 6G can benefit from the combination of NOMA and MEC.

1.1. Related works

The integration of NOMA and MEC has already been investigated in academia. Several works in the literature attempt to overcome the technical challenges of NOMA-assisted MEC. Most papers focused on energy minimization problems [10,20–22] for NOMA-assisted MEC networks. Besides, delay minimization problems are studied in Ref. [23–25]. Among them, both uplink and downlink NOMA transmissions are considered to minimize energy consumption. Most papers focus on the uplink transmission that multiple users can offload their tasks to one BS at the same time. The authors in Refs. [10,26] proposed a downlink model in which a single user can offload its different task parts to multiple BSs. The author in Ref. [21] proposed a scheme to minimize the energy consumption for multi-antenna NOMA-assisted MEC. Moreover, [27] studies a half-duplex model that considers not only the time and energy consumption of offloading to the MEC server, but also the time and energy consumption of downloading results. The authors in Ref. [28] proposed a Device-to-Device (D2D)-assisted MEC, which enables collaboration among users to reduce the computational load of edge servers. The paper [29] studied an offloading scheme served by heterogeneous networks, which improves the offloading utilities and reduce the task backlog.

Most works focused on power and time resource allocation, while some papers studied the optimization with different offloading strategies [10,21]. Binary and partial offloading are two types of offloading strategy models [3]. If the binary offloading scheme is adopted, the entire task can be offloaded to the BS for remote computation or local calculation by the mobile device. While the partial offloading scheme enables each task to be offloaded partially, the remaining part can be computed locally. The offloading strategy can also break and distribute one task to multiple MEC servers, as studied in Ref. [30]. Moreover, a novel hybrid NOMA and OMA model, in which NOMA and OMA transmissions are applied in different time slots for offloading one task, is proposed in Ref. [20]. However, only a two-user case was studied in that paper.

Due to the SIC complexity of NOMA transmission, it is unrealistic to multiplex all the users within the same radio resource. Several papers investigated the user grouping to resolve this issue by allocating multiple users into different subchannels. The authors in Ref. [31] proposed an algorithm to assign two users in each group according to their channel condition and then perform the resource allocation. In Ref. [32], the author provided a crowdsensing scheme that distributes sensing tasks with resource allocation based on game theory. A low-complexity matching algorithm, which assigns multiple users to different subchannel to maximize the energy efficiency, was studied by the authors in Ref. [33]. Moreover, the author in Ref. [34] proposed a channel assignment algorithm based on many-to-many matching to maximize the sum rate, and also investigated the influence of user fairness on the matching.

1.2. Contributions

As mentioned before, many works focused on the resource allocation of the conventional pure NOMA-assisted MEC, and only a few research works investigated the MEC with hybrid NOMA and OMA protocol. In addition, most of the works attempted to perform user pairing to maximize the sum-rate according to the Channel State Information (CSI). In this work, we focus on the hybrid NOMA-assisted MEC system and adopt a partial offloading scheme by introducing a partial offloading strategy coefficient. We also attempt to design an efficient algorithm to form users into groups and to minimize overall energy consumption. Unlike the previous work, the user pairing is not determined solely by the CSI of each user. For instance, we also consider the delay tolerance of each user during matching. The main contributions of this paper are summarized as follows:

- We consider partial offloading scheme in an uplink hybrid NOMA-assisted MEC and formulate an energy minimization optimization problem with delay constraints for the hybrid NOMA-assisted MEC system. The formulated problem is non-convex; thus it is challenging to find the optimal solution within polynomial time. To reduce the complexity of solving this non-convex problem, we decomposed the original problem into three sub-problems to apply the multilevel programming method. The three levels include power allocation, time slot scheduling, and offloading task assignment, which are solved optimally by exploiting the convexity and monotonicity. Therefore, the optimal solution to the original problem is obtained.
- Some significant insights are obtained from the derived optimal solution, which indicates different offloading schemes, including OMA, pure NOMA, and hybrid NOMA.
- We further investigate the multi-user scenario, where user grouping is implemented by matching theory. The derived optimal solution from the two-user case can be utilized to minimize the energy consumption for a multi-user scenario, where the user grouping algorithm based on matching theory is proposed. During the match, two users from different groups can form a swap blocking pair if one user or one group can benefit from swapping the groups of those two users without causing any degradation to any users or groups. The proposed algorithm is more efficient than the exhaustive search for grouping multiple users into groups, and simulation results verify its close-to-optimal performance when compared with the exhaustive searching.

1.3. Organizations

The organization of this paper is as follows. In Section 2, the system model of the hybrid NOMA and OMA scheme for the MEC offloading transmission is introduced. Section 3 describes the formulated resource allocation problem and provides the optimal closed-form solutions. An efficient user grouping algorithm is proposed in Section 4. Simulation results are provided in Section 5, and Section 6 concludes this paper.

2. System model and problem formulation

In this paper, we consider a NOMA-assisted MEC offloading scenario, where an MEC server is deployed to the BS to serve $|\mathscr{K}| = K$ users. The total time consumption is denoted by t_k, during which user $k \in \mathscr{K}$ offloads its task to the MEC server and then obtains the outcome after the MEC server computation. This process includes three stages, as shown in Fig. 1. In this work, the time and energy of calculating and results downloading are omitted since those terms are relatively small due to the large amount of calculation and communication resources on the MEC server [20,23,24]. Therefore, the total time consumption of MEC

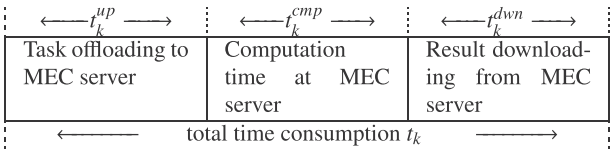

Fig. 1. Roundtrip time cost of MEC offloading.

offloading is approximated to $t_k \approx t_k^{up}$.

We further assume that all K users are divided into $|\phi| = \Phi$ groups, and each group φ for $\varphi = 1, \ldots, \Phi$ has an exclusive time slot to perform NOMA offloading. In order to reduce the decoding complexity, the number of users in each group is limited to 2, i.e., user m and user n who are served within the same time slot. The data amount of each task in group φ for user m and user n is denoted by $L_{m,\varphi}$ and $L_{n,\varphi}$ nats,[1] respectively. We also assume that in group φ, each user has a dedicated delay tolerance denoted as $\tau_{m,\varphi}$ and $\tau_{n,\varphi}$ nats,[1] respectively. Thus, the transmission time of each user satisfies the following equations: $t_{i,\varphi} \leq \tau_{i,\varphi}, \forall i \in \{m,n\}$, and. $\forall \varphi \in \{1, \ldots, \Phi\}$.

To further reduce the system complexity, the data amount for each user is assumed to be identical, and the index for L is thereby omitted. We define that in each group, the delay tolerance satisfies $\tau_{m,\varphi} \leq \tau_{n,\varphi}$. Because $\tau_{m,\varphi}$ is relatively shorter, it is interesting for user m to utilize its whole duration $\tau_{m,\varphi}$, which means that $t_{m,\varphi} = \tau_{m,\varphi}$.

In each group, the channel gain of a user in group u can be expressed as

$$H_{i,\varphi} = \tilde{h}_{i,\varphi} d_{i,\varphi}^{-\frac{\alpha}{2}}, \quad \forall i \in \{m,n\}, \forall \varphi, \tag{1}$$

where $\tilde{h}_{i,\varphi} \sim \mathcal{CN}(0,1)$ is the Rayleigh fading, $d_{i,\varphi}$ is the distance between the corresponded user and the BS, and α is the pass loss exponent. Given $z \sim \mathcal{CN}(0,\sigma^2)$ as the Addictive White Gaussian Noise (AWGN) with zero-mean and σ^2 variance, the channel gain could be normalized as

$$h_{i,\varphi} = \frac{|H_{i,\varphi}|^2}{\sigma^2}. \tag{2}$$

We adopt a block fading channel model in this paper, which indicates that the channel gain remains unchanged in each transmission block.

If the offloading transmission is served by OMA, every user m in group u will perform OMA transmission during $t_{m,\varphi}$. The achievable rate for user m is:

$$R_{m,\varphi} = B\ln\left(1 + P_{m,\varphi}|h_{m,\varphi}|^2\right), \tag{3}$$

where B denotes the system bandwidth, and $P_{m,\varphi}$ is the transmission power for user m. User m needs to transmit its task during $t_{m,\varphi}$, which indicates that

$$L = R_{m,\varphi} t_{m,\varphi}. \tag{4}$$

In the OMA transmission, user n has to wait during $t_{m,\varphi}$ time slot. Then, user n uploads its data in the duration $(\tau_{n,\varphi} - t_{m,\varphi})$, and the offloading has to satisfy the following equation:

$$L = (\tau_{n,\varphi} - t_{m,\varphi}) B\ln\left(1 + P_{n,\varphi}^{OMA}|h_{n,\varphi}|^2\right). \tag{5}$$

Furthermore, the proposed hybrid NOMA model enables user n to transmit its task during $t_{m,\varphi}$ to improve the spectral efficiency, which indicates that both users in the group offload their tasks to the BS by NOMA during $t_{m,\varphi}$. The time consumption for user m is denoted by its delay tolerance $\tau_{m,\varphi}$ from below. Since the priority is to serve user m within $\tau_{m,\varphi}$, user n has to be decoded first according to the principle of NOMA. User n will experience interference from user m, and user m has no interference from user n at the second stage of SIC. Hence, to guarantee the data rate of user m in this case that is the same as that in the OMA case after SIC, the achievable rate for user n has to satisfy:

$$R_{n,\varphi} \leq B\ln\left(1 + \frac{P_{n,\varphi}|h_{n,\varphi}|^2}{P_{m,\varphi}|h_{m,\varphi}|^2 + 1}\right), \tag{6}$$

where $P_{n,\varphi}$ is the transmission power of user n during $t_{m,\varphi}$ period. Due to the interference caused by user m, user n may not be able to complete the offloading during $t_{m,\varphi}$. An additional time slot $t_{r,\varphi}$ is scheduled to transmit the remaining part of user n, and the transmission time should not exceed the delay tolerance $\tau_{n,\varphi}$, i.e.,

$$t_{r,\varphi} \leq \tau_{n,\varphi} - \tau_{m,\varphi}. \tag{7}$$

During the second time slot, user n performs the OMA transmission, and the achievable rate is given by:

$$R_{r,\varphi} = B\ln\left(1 + P_{r,\varphi}|h_{n,\varphi}|^2\right), \tag{8}$$

where $P_{r,\varphi}$ denotes the transmission power of user n during the second time slot $t_{r,\varphi}$.

As shown in Fig. 2, the above configuration is called a hybrid NOMA scheme, in which two users by NOMA occupy the first time slot, and the second time slot is dedicated to the delay tolerable user.

Moreover, we adopt a partial offloading scheme in this model, and each task can be calculated either locally or remotely by the MEC server. An offloading strategy coefficient $\beta_\varphi \in [0,1]$ is introduced for user n in each group, which determines how much amount of data is offloaded to the MEC server, and the rest can be executed by the local device. Thus, the total amount of data offloading to the server for user n is $\beta_\varphi L$. The offloading transmission for user n has to satisfy that.

$$\tau_{m,\varphi} B\ln\left(1 + \frac{P_{n,\varphi}|h_{n,\varphi}|^2}{P_{m,\varphi}|h_{m,\varphi}|^2 + 1}\right)$$

$$+ t_{r,\varphi} B\ln\left(1 + P_{r,\varphi}|h_{n,\varphi}|^2\right) \geq \beta_\varphi L. \tag{9}$$

The total energy consumption of user n is composed of two parts, i.e., local computing energy consumption $E_{l,\varphi}$ and communication energy consumption $E_{o,\varphi}$. According to the models in Ref. [21,22], $E_{l,\varphi}$ is given as

$$E_{l,\varphi} = \frac{\kappa_0 [C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi} + t_{r,\varphi})^2}, \tag{10}$$

where κ_0 is the effective capacitance coefficient related to the Central Processing Unit (CPU) architecture of the mobile device, and C is the number of CPU cycles to process each nat. Thus, the total number of CPU cycles required for computing locally at the mobile device is $C(1-\beta_\varphi)L$, while those CPU cycles can be executed during $\tau_{m,\varphi} + t_{r,\varphi}$. The power for offloading is scheduled separately during these scheduled two-time slots according to equation (9), and thereby the offloading energy consumption $E_{o,\varphi}$ can be written as:

$$E_{o,\varphi} = t_{m,\varphi} P_{n,\varphi} + t_{r,\varphi} P_{r,\varphi}. \tag{11}$$

The total energy consumption for user n can be expressed as

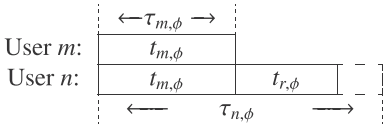

Fig. 2. System model.

[1] Nat is a unit of information, which is based on power of e and natural logarithms. One nat equals to $\frac{1}{\ln 2} \approx 1.443$ bits.

$$E_{n,\varphi} = E_{l,\varphi} + E_{o,\varphi}. \tag{12}$$

3. Joint resource allocation and computation task assignment

3.1. Problem formulation

This paper focuses on the total energy consumption minimization of the proposed NOMA-assisted MEC system. We assume that in each group, the resource allocation of user m is known, which means that $P_{m,\varphi}$ is given as a constant. The reason for this assumption is that the user m is served in priority, and user m consumes the whole time duration for offloading to minimize the energy consumption.

Therefore, we formulate the optimization problem to minimize the total energy consumption of user n, where power allocation, time slot scheduling, task assignment, and user grouping need to be jointly optimized. We first assume that the user grouping is given for the resource allocation. The implementation of user grouping can be found in section 4. Since each group is independent, it is promising to focus on a single group to reduce the complexity and then extend the solution to the multi-user case. Hence, the energy minimization problem can be formulated as follows:

$(P1):$

$$\min_{P_{n,\varphi},P_{r,\varphi}t_{r,\varphi},\beta_\varphi} \quad \frac{\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^2} + \tau_{m,\varphi}P_{n,\varphi} + t_{r,\varphi}P_{r,\varphi} \tag{13a}$$

$$\text{s.t.} \quad \tau_{m,\varphi}B\ln\left(1+\frac{P_{n,\varphi}|h_{n,\varphi}|^2}{P_{m,\varphi}|h_{m,\varphi}|^2+1}\right) + t_{r,\varphi}B\ln\left(1+P_{r,\varphi}|h_{n,\varphi}|^2\right) \geq \beta_\varphi L \tag{13b}$$

$$P_{n,\varphi} \geq 0, P_{r,\varphi} \geq 0 \tag{13c}$$

$$0 \leq t_{r,\varphi} \leq \tau_{n,\varphi} - \tau_{m,\varphi} \tag{13d}$$

$$0 \leq \beta_\varphi \leq 1, \tag{13e}$$

where (13b) is the transmission constraint, which ensures that user n could be able to complete the offloading within allocated time slots. (13c) is to make sure the power is positive. Following that, (13d) guarantees that the transmission is completed within its tolerance time, and (13e) sets the feasible range for coefficient β_φ.

The problem (P1) is non-convex and very difficult to be solved since variables are involved in multiplication and division operations. To solve this problem effectively, we divide this problem into three levels, and the optimal solutions to the problem (P1) are provided in the following subsections.

3.2. Uplink power allocation

At the first level, $t_{r,\varphi}$ and β_φ are assumed to be fixed, and then the problem is convex with respect to powers, which can be rearranged as follows:

$(P2):$

$$\min_{P_{n,\varphi},P_{r,\varphi}} \quad \frac{\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^2} + \tau_{m,\varphi}P_{n,\varphi} + t_{r,\varphi}P_{r,\varphi} \tag{14a}$$

$$\text{s.t.} \quad \tau_{m,\varphi}B\ln\left(1+\frac{P_{n,\varphi}|h_{n,\varphi}|^2}{P_{m,\varphi}|h_{m,\varphi}|^2+1}\right) + t_{r,\varphi}B\ln\left(1+P_{r,\varphi}|h_{n,\varphi}|^2\right) \geq \beta_\varphi L \tag{14b}$$

$$P_{n,\varphi} \geq 0, P_{r,\varphi} \geq 0. \tag{14c}$$

Due to the convexity of both the objective function and constraints, we can obtain the optimal power allocation for problem (P2) in closed-form, which is given in the following lemmas.

Lemma 1. *The optimal power allocation to (P2) is provided as functions of $t_{r,\varphi}$ and β_φ, which can be divided into the following cases:*

1. *Hybrid NOMA case: Given that $\tau_{n,\varphi} < 2\tau_{m,\varphi}$, the optimal power expressions for the hybrid NOMA scheme are:*

$$P^*_{n,\varphi} = |h_{n,\varphi}|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}} - e^{\frac{L}{B\tau_{m,\varphi}}}\right), \tag{15a}$$

$$P^*_{r,\varphi} = |h_{n,\varphi}|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}} - 1\right). \tag{15b}$$

   The reason for the condition $\tau_{n,\varphi} < 2\tau_{m,\varphi}$ is explained later in the Remark 1.

2. *Pure NOMA case: For $t_{r,\varphi} = 0$, user n only offloads its task to the BS using the same time slot with user m. The optimal power allocation can be expressed as*

$$P^{\text{PNOMA}*}_{n,\varphi} = |h_{n,\varphi}|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{B\tau_{m,\varphi}}} - e^{\frac{L}{B\tau_{m,\varphi}}}\right). \tag{16}$$

3. *OMA case: For $p_{n,\varphi} = 0$, user n does not offload its task to the BS until the second dedicate time slot $t_{r,\varphi}$. The optimal power* allocation *for this case is given as follows:*

$$P^{OMA*}_{r,\varphi} = |h_{n,\varphi}|^{-2}\left(e^{\frac{L\beta_\varphi}{Bt_{r,\varphi}}} - 1\right). \tag{17}$$

Proof. Please refer to Appendix A.

3.3. Time slot allocation for $t_{r,\varphi}$

To obtain the optimal time slot $t_{r,\varphi}$, the closed-form solutions of $P^*_{n,1}$ and $P^*_{n,2}$ can be substituted for the original problem (P1). Therefore, the optimization problem for $t_{r,\varphi}$ given by β can be written as

(P3):

$$g(t_{r,\varphi}) \triangleq \min_{t_{r,\varphi}} \frac{\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^2} + \tau_{m,\varphi}P_{n,\varphi} + t_{r,\varphi}P_{r,\varphi} \tag{18a}$$

$$\text{s.t.} \quad 0 \leq t_{r,\varphi} \leq \tau_{n,\varphi} - \tau_{m,\varphi}. \tag{18b}$$

The problem is convex with respect to $t_{r,\varphi}$, and for this particular problem, we can find the optimal solution through the monotonicity. The optimal time slot allocation is given in Lemma 2.

Lemma 2. *The optimal time slot allocation for user n during the OMA transmission is given as follows:*

$$t^*_{r,\varphi} = \tau_{n,\varphi} - \tau_{m,\varphi}. \tag{19}$$

Proof. *Please refer to* Appendix B.

3.4. Optimization of offloading strategy coefficient

Based on the expressions of optimal power and time allocations, the problem is to find the optimal offloading strategy coefficient to minimize the energy consumption, which is shown as follows:

$(P4)$:

$$\min_{\beta_\varphi} \quad \frac{\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^2}+\tau_{m,\varphi}P_{n,\varphi}+t_{r,\varphi}P_{r,\varphi} \tag{20a}$$

$$\text{s.t.} \quad 0\le\beta_\varphi\le 1. \tag{20b}$$

It is evident that the problem (P4) is convex, and we can find the Karush-Kuhn-Tucker (KKT) conditions to obtain the optimal solutions, which is given in the following lemma:

Lemma 3. *The optimal task offloading strategy coefficient is given as*

$$\beta_\varphi^*=1-\frac{2}{\theta_{2,\varphi}}\mathscr{W}\left(\theta_{1,\varphi}^{-\frac{1}{2}}\frac{\theta_{2,\varphi}}{2}e^{\theta_{2,\varphi}}\right), \tag{21}$$

where $\theta_{1,\varphi}=\frac{3\kappa_0BC^3L^2\left|h_{n,\varphi}\right|^2}{(\tau_{m,\varphi}+t_{r,\varphi})^2}$, $\theta_{2,\varphi}=\frac{L}{B(\tau_{m,\varphi}+t_{r,\varphi})}$. $\mathscr{W}(x)$ denotes the single-valued Lambert W function, which represents the inverse function of $f(\mu)=\mu e^{\mu}=x$ for given μ [35]. The solution $\mu=\mathscr{W}(x)$ can be obtained through Matlab.

Proof. Please refer to Appendix C

Remark 1. Given that $t_{r,\varphi}^*=\tau_{n,\varphi}-\tau_{m,\varphi}$, we have to ensure the optimal power allocation (15) we obtained from the hybrid NOMA is valid since it is possible that (15a) may be beyond the feasible range $P_{n,\varphi}\ge 0$ if inappropriate $t_{r,\varphi}$ and β_φ are given. It is important to note that in order to maintain the hybrid scheme and the feasibility of (15a), the offloading strategy coefficient has to satisfy the equation that

$$\beta_\varphi\ge\frac{\tau_{n,\varphi}}{\tau_{m,\varphi}}-1. \tag{22}$$

Furthermore, since the feasible range of β_φ is $[0,1]$, we can obtain that $\tau_{n,\varphi}$ has to satisfy:

$$\tau_{m,\varphi}\le\tau_{n,\varphi}, \tag{23a}$$

$$\tau_{n,\varphi}\le 2\tau_{m,\varphi}. \tag{23b}$$

It is obvious that the system intends to offload more data to the BS with the increase of $t_{r,\varphi}=\tau_{n,\varphi}-\tau_{m,\varphi}$.

4. Matching-based user grouping

The optimal resource allocation for minimizing the energy consumption is obtained in the previous sections. Subsequently, we adopt the user grouping to extend the proposed scheme to a multiple-user scenario. Conventionally, the optimal user grouping can be achieved by exploiting all possible user grouping, i.e., the exhaustive search method. However, to overcome the computational complexity issue of exhaustive search, a low complexity user grouping algorithm is proposed to form users into different groups to minimize the total energy consumption in the NOMA-assisted MEC system.

4.1. User grouping algorithm

To fulfill the user grouping algorithm, we can model a two-sided matching problem between the set of users $\mathscr{K}$ and the set of groups ϕ. In order to integrate the optimal resource allocation solutions with the matching algorithm, it is assumed that each two users are paired in a group, which means that $K=2\Phi$. Hence, we propose a two-to-one user grouping algorithm based on matching theory [31,33,36,37], which is defined as follows:

Definition 1. For two disjoint sets, i.e., $\mathscr{K}$ for users and ϕ for groups, a two-to-one matching Ψ represents the mapping relation from $\mathscr{K}$ to ϕ, which has to satisfy:

(a) $\Psi(k)\in\phi$, $\forall k\in\mathscr{K}$;
(b) $\Psi^{-1}(\mathscr{M}_\varphi)\subseteq\mathscr{K}$, $\forall\mathscr{M}_\varphi\in\phi$;
(c) $|\Psi(k)|=1$;
(d) $\Psi^{-1}(\mathscr{M}_\varphi)=2$;
(e) $\mathscr{M}_\varphi\in\Psi(k)\Leftrightarrow k=\Psi^{-1}(\mathscr{M}_\varphi)$.

In Definition 1, condition (a) indicates that each user is matched within one group, and condition (b) means that each group is matched with a subset of users. Conditions (c) and (d) represent that each user can only be paired with one group, and each group only contains two users. The last condition implies that the user k is matched with the group φ.

Users in each can be influenced by the mutual interference of the other user, and the deadlines of the other user could also influence the user's rate. Therefore, each user has its preference for pairing with other users, and each group has its preference for selecting among users. Note that for each user $k\in\mathscr{K}$, it prefers group $\mathscr{M}_\varphi$ to group $\mathscr{M}_{\overline{\varphi}}$:

$$(\mathscr{M}_\varphi,\Psi)\succ_k(\mathscr{M}_{\overline{\varphi}},\overline{\Psi})\Leftrightarrow E_k(\Psi)<E_k(\overline{\Psi}), \tag{24}$$

where $E_k(\Psi)$ is the energy consumption of user k when pairing with the group $\varphi=\Psi(k)$. Similarly, for the group $\mathscr{M}_\varphi$, its preference for the set of users can be expressed as

$$(k,\Psi)\succ_{\mathscr{M}_\varphi}(\overline{k},\overline{\Psi})\Leftrightarrow E_{\mathscr{M}_\varphi}(\Psi)<E_{\mathscr{M}_\varphi}(\overline{\Psi}), \tag{25}$$

where $E_{\mathscr{M}_\varphi}(\Psi)$ is the total energy consumption of group $\mathscr{M}_\varphi(\Psi)$. Based on the above relations, we can perform matching operations that allow users and groups to choose from each other.

For bilateral matching, two users in different groups may want to switch between their groups, and they formulate a swap-blocking pair in this case [38]. The definition of swap-blocking pair is given as:

Definition 2. For a given match Ψ and a pair of users $(k,\overline{k})$, if there exist two matches $\Psi(k)$ and $\Psi(\overline{k})$, then

$$\forall i\in\left\{k,\overline{k},\Psi^k,\Psi^{\overline{k}}\right\},\quad E_i\left(\Psi_{\overline{k}}^k\right)\le E_i(\Psi), \tag{26a}$$

$$\exists i\in\left\{k,\overline{k},\Psi^k,\Psi^{\overline{k}}\right\},\quad E_i\left(\Psi_{\overline{k}}^k\right)<E_i(\Psi), \tag{26b}$$

and the pair of users $(k,\overline{k})$ are defined as swap-blocking pair. $\Psi_{\overline{k}}^k$ implies the swapping operation between user k and $\overline{k}$.

The above definition means that if two users are willing to swap their groups, the swap operation will be approved if the energy consumption of each user or each group does not increase after swapping. At least the energy consumption for one user or one group would decrease. The matching process will repeatedly search the swap-blocking pairs, and it will eventually reach a stable status. The definition for the stable status is given as:

Definition 3. During a matching, if there is no more new swap-blocking pair to be identified, this matching is declared as two-sided exchange stable matching [38].

Hence, we propose an algorithm to describe the aforementioned two-to-one matching scheme, which is presented in Algorithm 1.

Algorithm 1. Two-to-One Matching-Based User Grouping

Algorithm 1 Two-to-One Matching-Based User Grouping

Initialization : Randomly allocate users into groups to obtain Ψ_{rd}.
Swap matching:
while There are new swap-blocking pairs in the previous round **do**
  for $i_{UE} = 1$ to K **do**
    Each user k sends the proposal to another user $\bar{k}$ in a different group.
    if $(k,\bar{k})$ is a swap-blocking pair that satisfies the two conditions in (26), **then**
      User k and $\bar{k}$ swap their groups.
    else
      The matching remains unchanged.
    end if
  end for
end while

4.2. Complexity analysis

In this subsection, we measure and compare the complexity of our proposed method and the exhaustive search. Assuming that a number of K users are paired with Φ groups, where $K = 2\Phi$. Based on this assumption, there are $\frac{K!}{2^{\Phi}}$ combinations, and therefore the complexity for the exhaustive search can be expressed as $O\left(\frac{K!}{2^{\Phi}}\right)$. By taking the natural logarithm, the complexity can be rewritten as $O(\ln(K!))$ where the lower order terms are omitted. It can be further reformulated by applying Stirling's approximation, which is $\ln(a!) = a\ln a - a + O\ln(a)$. Therefore, $O(\ln(K!))$ can be represented as $O(K\ln(K))$ by ignoring the lower order terms. Moreover, there is a while loop and a nested for loop in Algorithm 1, and the worst-case complexity is given as $O(K^2)$. By taking the natural logarithm, it is rewritten as $O(\ln(K))$. It is evident that $O(\ln(K)) < O(K\ln(K))$, and the complexity of Algorithm 1 is much lower than that of the exhaustive search.

5. Simulation results

In this part, we present the simulation results to evaluate the performance of our proposed NOMA-assisted MEC offloading protocol in terms of energy consumption, and compare it with the benchmarks listed as follows:

1. *OMA-based MEC:* The optimal energy consumption solution to the OMA case is given in (A.21). Users m and n are both served by TDMA in a different time slot, one following the other for offloading. In each group, user m consumes $\tau_{m,\varphi}$ to finish the task offloading, and user n has to complete the transmission within the remaining time $\tau_{n,\varphi} - \tau_{m,\varphi}$.
2. *Pure NOMA MEC:* In each group, user m and user n are served by NOMA, and both users can transmit during T_m simultaneously. The optimal solution to this case can be revealed from (A.7).
3. *Hybrid NOMA without task assignment:* This is the case same as our proposed method when the task assignment coefficient $\beta_{\varphi} = 1$.

Moreover, we compared our proposed user grouping algorithm with the optimal global solution based on the exhaustive search.

In the simulation, the users are randomly distributed in a disc area, with a minimum distance to the BS. The AWGN power is defined as $\sigma^2 = BN_0$. The rest of parameters are presented in Table .1.

As presented in Fig. 3, it reveals that the total energy consumption

Table 1
Simulation parameters.

Effective capacitance coefficient	10^{-28}
Number of CPU cycles required per bit	10^3
Transmission bandwidth B	20 MHz
Path loss exponent α	3.76
Noise spectral density N_0	-174 dBm/Hz
maximum cell radius	1000 m
minimum distance to BS	50 m

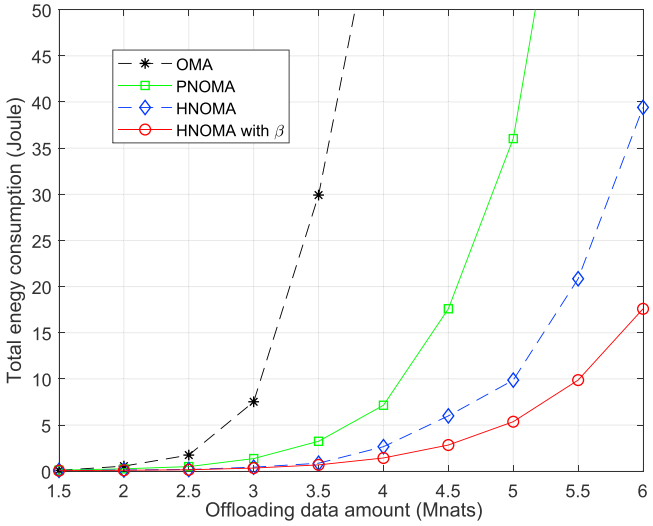

Fig. 3. The total energy consumption versus data amount for $K = 10$ users. The maximum delay tolerance is 0.08 s.

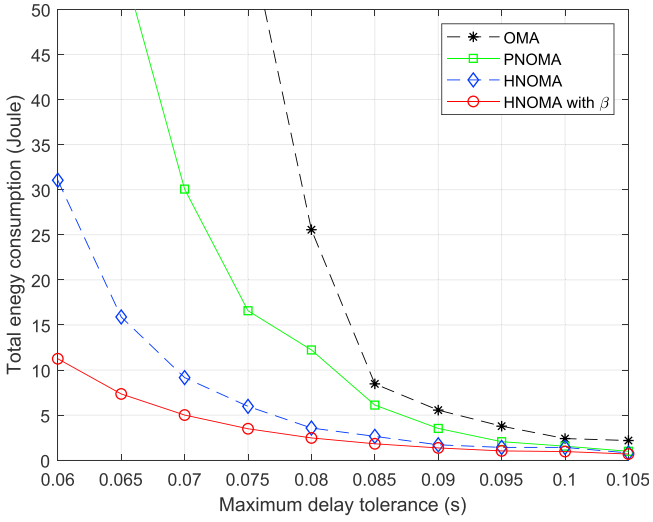

Fig. 4. The energy consumption versus maximum delay tolerance, where $L = 4$ Mnats and $K = 10$.

increases with the growth of offloading data amount for all kinds of schemes when $\varphi = 5$ groups. Specifically, the energy consumption for the OMA scheme is significantly higher than others since the user has to transmit all the data during a relatively short period of $t_{r,\varphi}$. The conventional NOMA scheme has lower energy consumption than OMA, and the offloading hybrid scheme provides better performance than the conventional scheme. Compared with those schemes, our proposed hybrid NOMA scheme adopts a partial offloading coefficient β_{φ}, which reflects that the solution we provided in this paper needs lower energy consumption for offloading, particularly when the data amount L

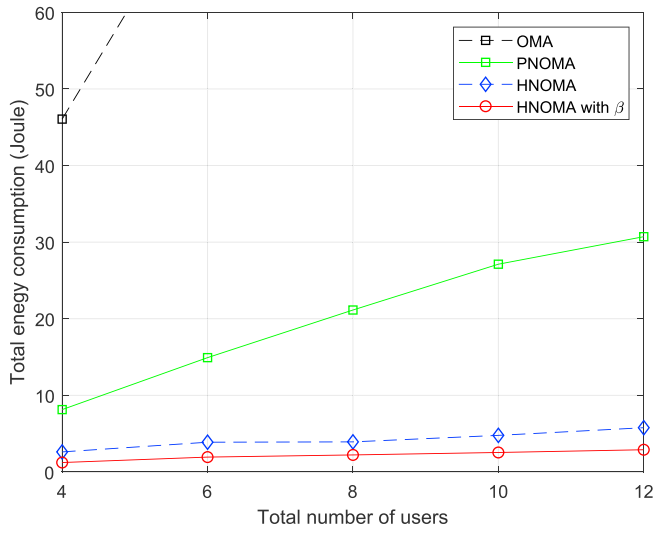

Fig. 5. The number of users versus total energy consumption. $L = 4$ Mnats and the maximum delay tolerance is 0.08 s.

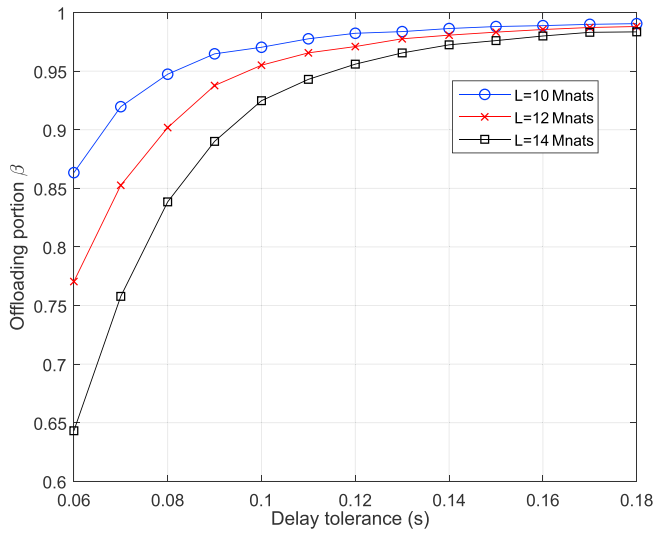

Fig. 6. The offloading portion versus delay tolerance τ_m, where $\tau_n = 1.25\,\tau_m$ and $L = [10, 12, 14]$ Mnats.

becomes large.

Fig. 4 illustrates the energy consumption versus the maximum delay tolerance when offloading 4 Mnats of data. The OMA scheme consumes much higher energy when transmitting such an amount of data. Similarly to our intuition, if more time is allowed for the offloading, less energy will be consumed for offloading. This figure also depicts our proposed partial offloading NOMA scheme achieves the lowest energy consumption when compared with those benchmarks.

Fig. 5 depicts the number of users in the total energy consumption. The total energy consumption increases if more users exist. Our proposed hybrid scheme has better performance than others, and the more users there are in the system, the more significant the difference.

In Fig. 6, we took one group as an example to depict the influence of the maximum delay tolerance on the offloading strategy coefficient β. As more time is given, the system tends to transmit more data to the MEC server as β increases since the offloading energy consumption is lower than that with tighter tolerance. Meanwhile, if the data amount increases, the system allocates more data to be executed on the mobile device because offloading within an extremely short time may consume more energy than local computing.

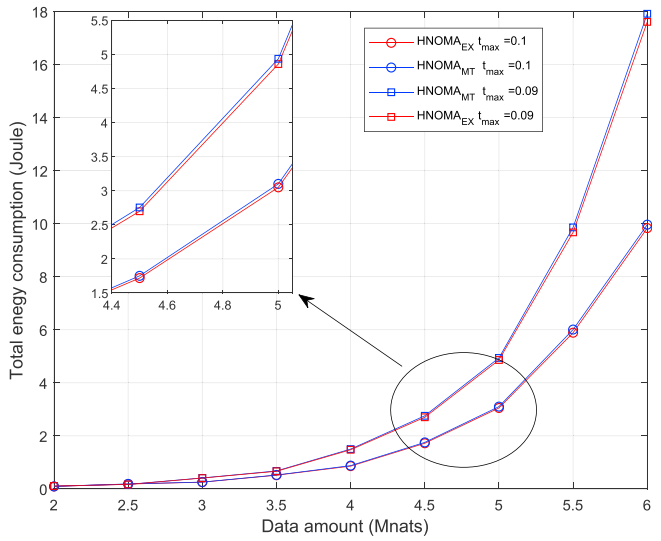

Fig. 7. The data amount versus the total energy consumption for $K = 10$ users.

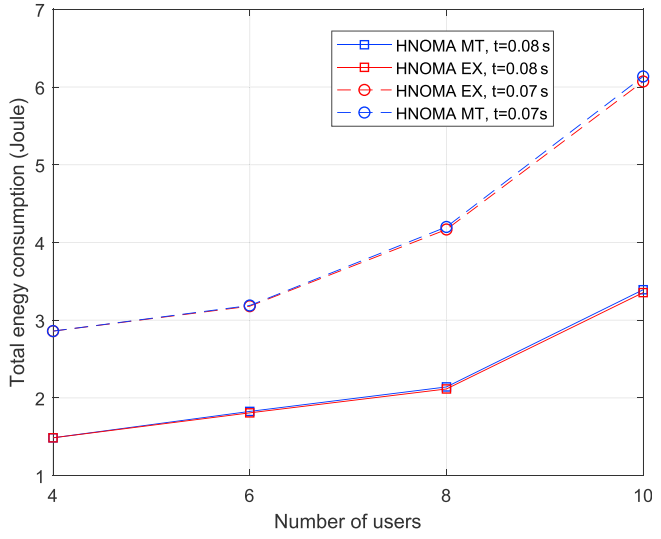

Fig. 8. Number of users versus total energy consumption, where $L = 4$ Mnats.

In Fig. 7, we have compared the proposed matching algorithm denoted as *MT*, with the exhaustive searching denoted as *EX*, respectively. The exhaustive searching provides the optimal results, and our algorithm is very close to the optimal results. Moreover, this figure also illustrates that the shorter maximum delay tolerance will result in more energy consumption when we set the maximum delay tolerance to 0.09s and 0.1s, respectively.

As can be revealed from Fig. 8, it compares the performance of our proposed algorithm to the optimal solution calculated by the exhaustive search. Due to the high computational cost of the exhaustive search, the maximum number of users is set to 10. The maximum delay tolerance is taken as 0.07s and 0.08s, respectively. With the increase in the number of users, the total energy consumption also increases. By comparing with the global optimal, our algorithm provides a very similar performance but with lower complexity.

6. Conclusion

In this paper, we have proposed a communication resource allocation scheme for the NOMA-assisted MEC, including power allocation, time slot allocation, task assignment, and user grouping. The optimization

problem has been formulated to minimize the system energy consumption under delay constraints. Assuming that user grouping is given initially, multi-level programming method is used to solve the original non-convex problem, and it is decomposed into three stages, including power allocation, time slot scheduling, and computation task assignment. The substitution of the closed-form solutions obtained from those sub-problems provides the optimal solution to the original problem we proposed. We also proposed an efficient user matching algorithm based on matching theory, and the simulation results show that the performance of our low-complexity algorithm is close to the optimal result calculated by the exhaustive search. By comparing our hybrid NOMA scheme with several benchmarks, the simulation results show that our proposed scheme can achieve superior performance in reducing system energy consumption.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the UK EPSRC under grant number EP/P 009719/2.

Appendix A. Proof of Lemma 1

By fixing $t_{r,\varphi}$ and β_{φ}, (P2) is obviously convex. Thus, we can obtain the optimal solution for $P_{n,\varphi}$ and $P_{r,\varphi}$ from the KKT conditions [39]. To find the KKT conditions, we need to find the Lagrangian function first. The Lagrangian function of Problem (P2) is given by:

$$\mathscr{L}(P_{n,\varphi}, P_{r,\varphi}, \boldsymbol{\lambda}) = \frac{\kappa_0 [C(1-\beta_{\varphi})L]^3}{(t_{m,\varphi}+t_{r,\varphi})^2} + t_{m,\varphi}P_{n,\varphi}$$

$$- \lambda_1 P_{n,\varphi} - \lambda_2 P_{r,\varphi} - \lambda_3 t_{m,\varphi} B \ln\left(1 + P_{n,\varphi}|h_n|^2 e^{-\frac{L}{Bt_{m,\varphi}}}\right)$$

$$- \lambda_3 t_{r,\varphi} B \ln\left(1 + P_{r,\varphi}|h_n|^2\right) + \lambda_3 \beta_{\varphi} L + t_{r,\varphi} P_{r,\varphi}, \tag{A1}$$

Where the Lagrangian multipliers are denoted as $\boldsymbol{\lambda} \triangleq [\lambda_1, \lambda_2, \lambda_3]$.

Therefore, the stationary condition can be written as follows:

$$\frac{\partial \mathscr{L}}{\partial P_{n,\varphi}} = t_{m,\varphi} - \lambda_1 - \lambda_3 \frac{t_{m,\varphi} B |h_{n,\varphi}|^2}{|h_{n,\varphi}|^2 P_{n,\varphi} + e^{\frac{L}{Bt_{m,\varphi}}}} = 0, \tag{A2a}$$

$$\frac{\partial \mathscr{L}}{\partial P_{r,\varphi}} = t_{r,\varphi} - \lambda_2 - \lambda_3 \frac{t_{r,\varphi} B |h_{n,\varphi}|^2}{P_{r,\varphi}|h_{n,\varphi}|^2 + 1} = 0. \tag{A2b}$$

Hence, the KKT conditions can be written as.

$$\beta_{\varphi} L - t_{m,\varphi} B \ln\left(1 + P_{n,\varphi}|h_{n,\varphi}|^2 e^{-\frac{L}{Bt_{m,\varphi}}}\right)$$

$$- t_{r,\varphi} B \ln\left(1 + P_{r,\varphi}|h_{n,\varphi}|^2\right) \le 0 \tag{A3a}$$

$$-P_{n,\varphi} \le 0, \quad -P_{r,\varphi} \le 0, \tag{A3b}$$

$$\lambda_j \ge 0, \quad \forall j \in \{1,2,3\} \tag{A3c}$$

$$\lambda_3 \beta_{\varphi} L - \lambda_3 t_{m,\varphi} B \ln\left(1 + P_{n,\varphi}|h_{n,\varphi}|^2 e^{-\frac{L}{Bt_{m,\varphi}}}\right)$$

$$- \lambda_3 t_{r,\varphi} B \ln\left(1 + P_{r,\varphi}|h_{n,\varphi}|^2\right) = 0 \tag{A3d}$$

$$\lambda_1 P_{n,\varphi} = 0, \quad \lambda_2 P_{r,\varphi} = 0 \tag{A3e}$$

$$t_{m,\varphi} - \lambda_1 - \lambda_3 \frac{t_{m,\varphi} B \left|h_{n,\varphi}\right|^2}{\left|h_{n,\varphi}\right|^2 P_{n,\varphi} + e^{\frac{L}{Bt_{m,\varphi}}}} = 0 \tag{A3f}$$

$$t_{r,\varphi} - \lambda_2 - \lambda_3 \frac{t_{r,\varphi} B \left|h_{n,\varphi}\right|^2}{P_{r,\varphi}\left|h_{n,\varphi}\right|^2 + 1} = 0 \tag{A3g}$$

In (A.3e), we assume that λ_1 and λ_2 should not be nonzero at the same time, which leads to $P_{n,\varphi} = P_{r,\varphi} = 0$, and user n transmits no data to the BS. Hence, we can divide this scheme into three scenarios, including $\lambda_1 = \lambda_2 = 0$, $\lambda_1 = 0, \lambda_2 \neq 0$ or $\lambda_2 = 0, \lambda_1 \neq 0$.

1. Hybrid NOMA ($\lambda_1 = \lambda_2 = 0$)

Based on this scheme, $P_{n,\varphi}$ and $P_{r,\varphi}$ should be nonzero, and λ_3 should be greater than zero because if $\lambda_3 = 0$, $t_{m,\varphi} = t_{r,\varphi} = 0$ according to equation (A.11) and (A.12), which cannot be true in this scenario. Thus, the simplified KKT conditions can be formulated as follows:

$$\lambda_3 \beta_\varphi L - \lambda_3 t_{m,\varphi} B \ln\left(1 + P_{n,\varphi}\left|h_{n,\varphi}\right|^2 e^{-\frac{L}{Bt_{m,\varphi}}}\right) - \lambda_3 t_{r,\varphi} B \ln\left(1 + P_{r,\varphi}\left|h_{n,\varphi}\right|^2\right) = 0 \tag{A4a}$$

$$t_{m,\varphi} - \lambda_3 \frac{t_{m,\varphi} B \left|h_{n,\varphi}\right|^2}{\left|h_{n,\varphi}\right|^2 P_{n,\varphi} + e^{\frac{L}{Bt_{m,\varphi}}}} = 0 \tag{A4b}$$

$$t_{r,\varphi} - \lambda_3 \frac{t_{r,\varphi} B \left|h_{n,\varphi}\right|^2}{P_{r,\varphi}\left|h_{n,\varphi}\right|^2 + 1} = 0 \tag{A4c}$$

The optimal solution can be obtained from (A.4) after some mathematical manipulations, and the closed-form solution is expressed as

$$P^*_{n,\varphi} = \left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{B(t_{m,\varphi}+t_{r,\varphi})}} - e^{\frac{L}{Bt_{m,\varphi}}}\right), \tag{A5a}$$

$$P^*_{r,\varphi} = \left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{B(t_{m,\varphi}+t_{r,\varphi})}} - 1\right). \tag{A5b}$$

2. Pure NOMA ($\lambda_1 = 0, \lambda_2 \neq 0$)

If $\lambda_2 \neq 0$, then we have $P_{r,\varphi} = 0$, which means that user n only utilizes the time slot $t_{n,\varphi}$ that is shared with user m to transmit its data, and thereby $t_{r,\varphi} = 0$. This case is named pure NOMA, and the optimal solution for $P_{n,\varphi}$ is:

$$P^{\mathrm{PNOMA}*}_{n,\varphi} = \left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{Bt_{m,\varphi}}} - e^{\frac{L}{Bt_{m,\varphi}}}\right). \tag{A6}$$

Given that $t_{n,\varphi} = 0$, the offloading energy consumption for pure NOMA scenario can be expressed as

$$E^{\mathrm{PNOMA}}_{o,\varphi} = t_{n,\varphi}\left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{Bt_{m,\varphi}}} - e^{\frac{L}{Bt_{m,\varphi}}}\right). \tag{A7}$$

3. OMA($\lambda_1 \neq 0, \lambda_2 = 0$)

In this case, we have $P_{n,\varphi} = 0$ and $P_{r,\varphi} \geq 0$, which illustrates that user n only occupies the second time slot $t_{r,\varphi}$ solely. Therefore, from KKT condition (3d), we can obtain the optimal solution in this case as

$$P^{OMA*}_{n,\varphi} = \left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L\beta_\varphi}{Bt_{r,\varphi}}} - 1\right). \tag{A8}$$

The energy consumption for the OMA case can be written as

$$E^{\mathrm{OMA}}_{o,\varphi} = t_{r,\varphi}\left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L\beta_\varphi}{Bt_{r,\varphi}}} - 1\right). \tag{A9}$$

Appendix B. Proof of Lemma 2

It is evident that the problem (P3) is convex for given β_φ. To obtain the optimal solution to the problem (P3), we can analyze its monotonicity. The derivative of $g(t_{r,\varphi})$ can be written as.

$$\frac{dg(t_{r,\varphi})}{dt_{r,\varphi}} = -\frac{2\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^3} - \left|h_{n,\varphi}\right|^{-2}$$

$$-\left|h_{n,\varphi}\right|^{-2} e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}} \frac{\tau_{m,\varphi}L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})^2}$$

$$-\left|h_{n,\varphi}\right|^{-2}\left(e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}} \frac{t_{r,\varphi}L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})^2} - 1\right) \tag{B1a}$$

$$= -\frac{2\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^3} - \left|h_{n,\varphi}\right|^{-2}$$

$$+\left|h_{n,\varphi}\right|^{-2} e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\left(1-\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}\right). \tag{B1b}$$

To find the monotonicity of the problem (P3), we have to determine whether the derivative of it in (B.1) is positive or negative. The first two terms in (B.1b) are negative, i.e., $-\frac{2\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^3} - \left|h_{n,\varphi}\right|^{-2} \le 0$. We have to determine the negativity of the following term:

$$-\left|h_{n,\varphi}\right|^{-2} e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\left(\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}-1\right), \tag{B2}$$

which can be rewritten as $\tilde{g}\left(\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}\right)$, where $\tilde{g}(x)$ is defined as

$$\tilde{g}(x) = \left|h_{n,\varphi}\right|^{-2} e^x (1-x). \tag{B3}$$

The derivative of (B.4) can be expressed as

$$\frac{d\tilde{g}(x)}{dx} = -\left|h_{n,\varphi}\right|^{-2} x e^x \le 0, \quad \forall x \ge 0, \tag{B4}$$

which is a monotonically decreasing function if $x \ge 0$. We can then obtain the following inequality relations:

$$\tilde{g}\left(\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}\right) \le \tilde{g}(0) = \left|h_{n,\varphi}\right|^{-2}. \tag{B5}$$

Therefore, the maximum value of $\tilde{g}(x)$ for $x \ge 0$ is $\left|h_{n,\varphi}\right|^{-2}$. Hence, $\frac{dg(t_{r,\varphi})}{dt_{r,\varphi}} \le 0$, which means that $g(t_{r,\varphi})$ is a monotonically decreasing function with respect to $t_{r,\varphi}$. Consequently, in order to minimize the energy consumption, $t_{r,\varphi}$ is preferably as large as possible, and the optimal solution is thereby $t^*_{r,\varphi} = \tau_{n,\varphi} - \tau_{m,\varphi}$.

Appendix C. Proof of Lemma 3

Since (P3.4) is convex in terms of β_φ, we can exploit the optimal task assignment coefficient β_φ. For $\lambda_4 \ge 0$ and $\lambda_5 \ge 0$, the Lagrangian function is given as.

$$\mathscr{L}(\beta_\varphi, \lambda_4, \lambda_5) = \frac{\kappa_0[C(1-\beta_\varphi)L]^3}{(\tau_{m,\varphi}+t_{r,\varphi})^2} + t_{m,\varphi}P_{n,\varphi}$$

$$+t_{r,\varphi}P_{r,\varphi} - \lambda_4\beta_\varphi + \lambda_5(\beta_\varphi - 1), \tag{C1}$$

where λ_4 and λ_5 are the Lagrangian multipliers. The stationary condition can be obtained as

$$\frac{\partial\mathscr{L}}{\partial\beta_\varphi} = \frac{-3\kappa_0(CL)^3(1-\beta_\varphi)^2}{(\tau_{m,\varphi}+t_{r,\varphi})^2}$$
$$+\left|h_{n,\varphi}\right|^{-2}\left(\frac{L}{B}e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\right) - \lambda_4 + \lambda_5 = 0. \tag{C2}$$

The KKT condition is given as follows:

$$-\beta_\varphi \le 0 \tag{C3a}$$

$$\beta_\varphi - 1 \le 0 \tag{C3b}$$

$$\lambda_4 \beta_\varphi = 0 \tag{C3c}$$

$$\lambda_5(\beta_\varphi - 1) = 0 \tag{C3d}$$

$$\frac{-3\kappa_0 (CL)^3 (1-\beta_\varphi)^2}{(\tau_{m,\varphi} + t_{r,\varphi})^2} + \left|h_{n,\varphi}\right|^{-2} \left(\frac{L}{B} e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\right) - \lambda_4 + \lambda_5 = 0 \tag{C3e}$$

If $\beta_\varphi = 0$, all the data will compute locally, and the total energy consumption only contains the local computing energy in (10), which is unnecessary for optimization. Moreover, if $\beta_\varphi = 1$, λ_4 has to be zero as condition (C.3d) is true. Then, (C.3d) becomes as

$$\left|h_{n,\varphi}\right|^{-2} \left(\frac{L}{B} e^{\frac{2L}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\right) + \lambda_5 = 0. \tag{C.4}$$

Since both terms in (C.9) are non-negative, and $\left|h_{n,\varphi}\right|^{-2} \left(\frac{L}{B} e^{\frac{2L}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\right) > 0$, which means that equation (C.9) cannot be true in this assumption. Therefore, we focus on the case that $0 < \beta < 1$, and $\lambda_4 = \lambda_5 = 0$. Based on this condition, we can obtain from (C.3e) the equation:

$$\frac{3\kappa_0 (CL)^3 (1-\beta_\varphi)^2}{(\tau_{m,\varphi} + t_{r,\varphi})^2} = \left|h_{n,\varphi}\right|^{-2} \left(\frac{L}{B} e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}}\right). \tag{C5}$$

It can be rearranged as

$$\frac{3\kappa_0 BC^3 L^2 (1-\beta_\varphi)^2 \left|h_{n,\varphi}\right|^2}{(\tau_{m,\varphi} + t_{r,\varphi})^2} = e^{\frac{L(\beta_\varphi+1)}{B(\tau_{m,\varphi}+t_{r,\varphi})}}. \tag{C6}$$

Define $\theta_{1,\varphi} \triangleq \frac{3\kappa_0 BC^3 L^2 \left|h_{n,\varphi}\right|^2}{(\tau_{m,\varphi}+t_{r,\varphi})^2}$, $\theta_{2,\varphi} \triangleq \frac{L}{B(\tau_{m,\varphi}+t_{r,\varphi})}$ and $\mu_\varphi \triangleq 1 - \beta_\varphi$. Equation (C.6) can be rewritten as

$$\theta_{1,\varphi} \mu_\varphi^2 e^{-\theta_{2,\varphi}} = e^{\theta_{2,\varphi}(1-\mu_\varphi)}. \tag{C.7}$$

Then, equation (C.7) is rearranged as

$$\frac{\theta_{2,\varphi}}{2} \mu_\varphi e^{\frac{\theta_{2,\varphi}}{2}\mu_\varphi} = \frac{\theta_{2,\varphi}}{2} \theta_{1,\varphi}^{-\frac{1}{2}} e^{\theta_{2,\varphi}}. \tag{C8}$$

By solving the equation above, we can obtain the following equation:

$$\mu_\varphi^* = \frac{2}{\theta_{2,\varphi}} \mathcal{W}\left(\theta_{1,\varphi}^{-\frac{1}{2}} \frac{\theta_{2,\varphi}}{2} e^{\theta_{2,\varphi}}\right). \tag{C9}$$

Therefore, the optimal solution for the task assignment ratio is:

$$\beta_\varphi^* = 1 - \mu_\varphi^* = 1 - \frac{2}{\theta_{2,\varphi}} \mathcal{W}\left(\theta_{1,\varphi}^{-\frac{1}{2}} \frac{\theta_{2,\varphi}}{2} e^{\theta_{2,\varphi}}\right), \tag{C10}$$

where $\frac{1}{W_0}$ denotes the single-valued Lambert W function.

References

[1] H. Li, F. Fang, Z. Ding, Joint resource allocation for NOMA-assisted MEC networks, in: IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020, pp. 1–6, https://doi.org/10.1109/ICCWorkshops49005.2020.9145154.

[2] N. Abbas, Y. Zhang, A. Taherkordi, T. Skeie, Mobile edge computing: a survey, IEEE Internet Things J 5 (1) (2018) 450–465, https://doi.org/10.1109/JIOT.2017.2750180.

[3] Y. Mao, C. You, J. Zhang, K. Huang, K.B. Letaief, A survey on mobile edge computing: the communication perspective, IEEE Commun. Surveys Tutorials 19 (4) (2017) 2322–2358, https://doi.org/10.1109/COMST.2017.2745201.

[4] M.A. Rodriguez, R. Buyya, Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds, IEEE Trans. on Cloud Comput. 2 (2) (2014) 222–235, https://doi.org/10.1109/TCC.2014.2314655.

[5] Z. Ding, P. Fan, H.V. Poor, Impact of non-orthogonal multiple access on the offloading of mobile edge computing, IEEE Trans. Commun. 67 (1) (2019) 375–390.

[6] Y. Wu, B. Shi, L.P. Qian, F. Hou, J. Cai, X. Shen, Energy-efficient multi-task multi-access computation offloading via NOMA transmission for IoTs, IEEE Trans Ind. Informat. (2019) 1, https://doi.org/10.1109/TII.2019.2944839.

[7] Y. Zhao, W. Wang, Y. Li, C. Colman Meixner, M. Tornatore, J. Zhang, Edge computing and networking: a survey on infrastructures and applications, IEEE Access 7 (2019) 101213–101230.

[8] E. Calvanese Strinati, S. Barbarossa, J.L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, C. Dehos, 6G: the next frontier: from holographic messaging to artificial intelligence using subterahertz and visible light communication, IEEE Veh. Technol. Mag. 14 (3) (2019) 42–50.
[9] B. Cao, L. Zhang, Y. Li, D. Feng, W. Cao, Intelligent offloading in multi-access edge computing: a state-of-the-art review and framework, IEEE Commun. Mag. 57 (3) (2019) 56–62, https://doi.org/10.1109/MCOM.2019.1800608.
[10] Q. Gu, G. Wang, J. Liu, R. Fan, D. Fan, Z. Zhong, Optimal offloading with non-orthogonal multiple access in mobile edge computing, in: IEEE Proc. of Global Commun. Conf, GLOBECOM), 2018, pp. 1–5, https://doi.org/10.1109/GLOCOM.2018.8647179.
[11] M. Chen, Y. Hao, Task offloading for mobile edge computing in software defined ultra-dense network, IEEE J. Sel. Area. Commun. 36 (3) (2018) 587–597.
[12] P. Wang, C. Yao, Z. Zheng, G. Sun, L. Song, Joint task assignment, transmission, and computing resource allocation in multilayer mobile edge computing systems, IEEE Internet Things J 6 (2) (2019) 2872–2884, https://doi.org/10.1109/JIOT.2018.2876198.
[13] Y. Wu, K. Ni, C. Zhang, L.P. Qian, D.H.K. Tsang, NOMA-assisted multi-access mobile edge computing: a joint optimization of computation offloading and time allocation, IEEE Trans. Veh. Technol. 67 (12) (2018) 12244–12258, https://doi.org/10.1109/TVT.2018.2875337.
[14] M. Vaezi, G. Amarasuriya, Y. Liu, A. Arafa, F. Fang, Z. Ding, Interplay between NOMA and Other Emerging Technologies: A Survey, arXiv E-Prints, 2019 arXiv:1903.10489arXiv:1903.10489.
[15] L. Zhu, Z. Xiao, X. Xia, D. Oliver Wu, Millimeter-wave communications with non-orthogonal multiple access for B5G/6G, IEEE Access 7 (2019) 116123–116132.
[16] F. Tariq, M. Khandaker, K.-K. Wong, M. Imran, M. Bennis, M. Debbah, A Speculative Study on 6G, 2019 arXiv e-prints, arXiv:1902.06700arXiv:1902.06700.
[17] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, L. Hanzo, Non-orthogonal multiple access for 5G and beyond, Proc. IEEE 105 (12) (2017) 2347–2381.
[18] Z. Ding, X. Lei, G.K. Karagiannidis, R. Schober, J. Yuan, V.K. Bhargava, A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends, IEEE J. Sel. Area. Commun. 35 (10) (2017) 2181–2195, https://doi.org/10.1109/JSAC.2017.2725519.
[19] F. Fang, Z. Ding, W. Liang, H. Zhang, Optimal energy efficient power allocation with user fairness for uplink MC-NOMA systems, IEEE Wireless Commun. Lett. 8 (4) (2019) 1133–1136, https://doi.org/10.1109/LWC.2019.2908912.
[20] Z. Ding, J. Xu, O.A. Dobre, H.V. Poor, Joint power and time allocation for NOMA-MEC offloading, IEEE Trans. Veh. Technol. 68 (6) (2019) 6207–6211, https://doi.org/10.1109/TVT.2019.2907253.
[21] F. Wang, J. Xu, Z. Ding, Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems, IEEE Trans. Commun. 67 (3) (2019) 2450–2463, https://doi.org/10.1109/TCOMM.2018.2881725.
[22] P. Mach, Z. Becvar, Mobile edge computing: a survey on architecture and computation offloading, IEEE Commun. Surveys Tutorials 19 (3) (2017) 1628–1656, https://doi.org/10.1109/COMST.2017.2682318.
[23] Z. Ding, D.W.K. Ng, R. Schober, H.V. Poor, Delay minimization for NOMA-MEC offloading, IEEE Signal Process. Lett. 25 (12) (2018) 1875–1879, https://doi.org/10.1109/LSP.2018.2876019.
[24] M. Zeng, N. Nguyen, O.A. Dobre, H.V. Poor, Delay minimization for NOMA-assisted MEC under power and energy constraints, IEEE Wireless Commun. Lett. (2019) 1, https://doi.org/10.1109/LWC.2019.2934453.
[25] Y. Wu, L.P. Qian, K. Ni, C. Zhang, X. Shen, Delay-minimization non-orthogonal multiple access enabled multi-user mobile edge computation offloading, IEEE J. Sel. Topics Signal Process 13 (3) (2019) 392–407, https://doi.org/10.1109/JSTSP.2019.2893057.
[26] F. Fang, K. Wang, Z. Ding, Optimal task assignment and power allocation for downlink NOMA MEC networks, in: IEEE Proc. of Global Commun. Conf, GLOBECOM), 2019, pp. 1–6 (accepted).
[27] Y. Pan, M. Chen, Z. Yang, N. Huang, M. Shikh-Bahaei, Energy-efficient NOMA-based mobile edge computing offloading, IEEE Commun. Lett. 23 (2) (2019) 310–313.
[28] X. Diao, J. Zheng, Y. Wu, Y. Cai, Joint computing resource, power, and channel allocations for d2d-assisted and noma-based mobile edge computing, IEEE Access 7 (2019) 9243–9257, https://doi.org/10.1109/ACCESS.2018.2890559.
[29] Y. Li, S. Xia, M. Zheng, B. Cao, Q. Liu, Lyapunov optimization based trade-off policy for mobile cloud offloading in heterogeneous wireless networks, IEEE Trans. on Cloud Comput (2019) 1, https://doi.org/10.1109/TCC.2019.2938504.
[30] J. Chen, S. Chen, Q. Wang, B. Cao, G. Feng, J. Hu, iRAF, A deep reinforcement learning approach for collaborative mobile edge computing iot networks, IEEE Internet Things J 6 (4) (2019) 7011–7024, https://doi.org/10.1109/JIOT.2019.2913162.
[31] J. Zhu, J. Wang, Y. Huang, S. He, X. You, L. Yang, On optimal power allocation for downlink non-orthogonal multiple access systems, IEEE J. Sel. Area. Commun. 35 (12) (2017) 2744–2757, https://doi.org/10.1109/JSAC.2017.2725618.
[32] B. Cao, S. Xia, J. Han, Y. Li, A distributed game methodology for crowdsensing in uncertain wireless scenario, IEEE Trans. Mobile Comput. 19 (1) (2020) 15–28, https://doi.org/10.1109/TMC.2019.2892953.
[33] F. Fang, H. Zhang, J. Cheng, V.C.M. Leung, Energy-efficient resource allocation for downlink non-orthogonal multiple access network, IEEE Trans. Commun. 64 (9) (2016) 3722–3732, https://doi.org/10.1109/TCOMM.2016.2594759.
[34] B. Di, L. Song, Y. Li, Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks, IEEE Trans. Wireless Commun. 15 (11) (2016) 7686–7698.
[35] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, D. Knuth, On the lambert w function, Adv. Comput. Math. 5 (1996) 329–359, https://doi.org/10.1007/BF02124750.
[36] H. Sasaki, M. Toda, Two-sided matching problems with externalities, J. Econ. Theor. 70 (1) (1996) 93–108.
[37] A. Roth, M. Sotomayor, Two-sided matching, in: first ed., in: R. Aumann, S. Hart (Eds.), Handbook of Game Theory with Economic Applications, vol. 1, 1992, pp. 485–541. Ch. 16.
[38] Y. Gu, W. Saad, M. Bennis, M. Debbah, Z. Han, Matching theory for future wireless networks: fundamentals and applications, IEEE Commun. Mag. 53 (5) (2015) 52–59.
[39] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, NY, USA, 2004.