Energy-efficient Resource Allocation for NOMA-assisted Mobile Edge Computing

Ming Zeng, Viktoria Fodor
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology
{mzeng,vfodor}@kth.se

***Abstract*—In this paper we evaluate the effect of increased wireless spectral efficiency on the performance of mobile edge computing. Specifically, we study the energy minimization of computation offloading for a multi-carrier non-orthogonal multiple access (NOMA) assisted mobile edge computing (MEC) system. A joint radio-and-computational resource allocation problem is formulated, in which three different resources should be appropriately allocated, including subcarriers, transmission power and computational resources. The formulated resource allocation problem belongs to mixed integer nonlinear programming (MILNP) and is NP-hard. We propose therefore a heuristic solution consisting of two steps, NOMA clustering and subcarrier allocation, and joint computational resource and power allocation. Our numerical results show that NOMA based MEC significantly outperforms its OMA counterpart, especially in scenarios with strict delay limits, where both the transmission and the computational resources become scarce.**

I. Introduction

Mobile edge computing (MEC) is envisioned as a promising technology for enhancing the computation capacities and prolonging the lifespan of mobile devices. As it enables mobile devices to offload computation-intensive tasks to servers in close proximity, MEC also avoids the large delay introduced by centralized cloud computing [1]. For multiuser MEC systems, how to share the time and frequency resources among the devices is of significance, as it directly affects the energy consumption and the delay of computation offloading. Various resource allocation schemes based on orthogonal multiple access (OMA) have been proposed to enhance the performance of multiuser MEC systems. For instance, TDMA is employed in [2], [3], while FDMA is adopted by [4], [5]. In addition, both TDMA and OFDMA are considered in [6]. Non-orthogonal code division multiple access (CDMA) is employed for multiuser computation offloading by treating the interference among users as noise in [7]. These works however do not address whether advanced wireless communication techniques can improve the edge computing performance.

Non-orthogonal multiple access (NOMA) is one of the popular recent solutions to increase the spectral efficiency and system capacity of mobile networks [8]. Different from OMA, multiple users are multiplexed over the same resource block in NOMA by applying superposition coding at the transmitter, and successive interference cancellation (SIC) at the receiver. The motivation of adopting NOMA comes from the finding that NOMA is able to achieve the capacity of single-input single-output (SISO) (Gaussian) broadcast channel [9]. So far, extensive studies show that NOMA can achieve superior spectral efficiency [10], [11] and energy efficiency than OMA [12], [13].

In this paper, we aim at evaluating whether a NOMA-assisted MEC system can deliver superior performance to its OMA based counterpart. The resource allocation for NOMA-assisted MEC, including the allocation of transmission resource blocks, transmission power and computational resources, inherits the high complexity of NOMA resource allocation when only transmission is considered. Therefore, we propose a heuristic based solution, considering bipartite-graph matching based resource block allocation in the first step and a greedy joint transmission power and computational resource allocation in the second step. We present numerical results, comparing the proposed solution to its OMA counterpart, and demonstrate that increasing the spectral efficiency remains important even for edge computing, if the transmission and computational resources are scarce.

NOMA-assisted MEC has been addressed recently in [14] and [15]. In [14] weighted sum energy consumption minimization is considered, however, by treating the computation time at the server as zero. In [15] the resource allocation problem is formulated as in our work, considering the objective of minimizing the transmission energy consumption of all users under individual delay requirements. However, the solutions proposed for the allocation of the three different resources, i.e., the subcarriers, transmission power and the computational resource are weakly motivated, and the performance of the proposed solution is not evaluated in detail. We will use this prior work as baseline solution in our study.

The rest of the paper is organized as follows. The system model is introduced in Section II, while the problem formulation is given in Section III. The proposed resource allocation scheme is presented in Section IV, whereas numerical results are shown in Section V. Conclusions are finally drawn in Section VI.

978-1-5386-6009-6/18/$31.00 ©2018 IEEE

II. System Model

We consider a single-cell scenario, in which one eNB equipped with a cloudlet serves multiple edge computing users. The set of users are denoted as $\mathcal{U}, |\mathcal{U}| = U$. Each user $u \in \mathcal{U}$ has a computational task, characterized by the data input size L_u, i.e., the number of bits needed to be transmitted from the user to the eNB, and the workload W_u, i.e., the number of CPU cycles required to complete the execution of the task. The computation starts when all data is received by the eNB. This scenario corresponds to processing of individual sets of data, like for visual analysis and large scale graph or matrix operations.

A. Communication Resources

The overall transmission bandwidth is B Hz, which is equally divided into a set of frequency resource blocks (RBs) $\mathcal{R}_f, |\mathcal{R}_f| = M$, each $m \in \mathcal{R}_f$ with B/M Hz. Each RB can accommodate multiple users by employing NOMA, while it is assumed that each user can access only one RB. Whether user u is multiplexed on RB m is reflected by the binary variable $x_{u,m}$, i.e., $x_{u,m} = 1$ if user u is multiplexed on RB m, and $x_{u,m} = 0$ otherwise. Moreover, $\sum_{m=1}^{M} x_{u,m} = 1$. Users sharing the same RB form a cluster.

The channel gain of user u on RB m is $h_{u,m}$, and is considered to be constant for the time of transmitting L_u bits, corresponding to a low mobility scenario. Within a cluster, the interference from the users with better channel gains can be removed by employing SIC, and thus, an increase in the signal-to-interference-plus noise ratio (SINR) is achieved under NOMA.

To express the achievable rate of user u on RM m, let us assume that the users' channels are arranged in a descending order on each RB, and the position of user u on RB m in the sorted sequence is denoted by $b_m(u)$. Denote the power for user u on RB m as $p_{u,m}$, and assume to have an additive white Gaussian noise, with zero mean and variance σ^2. According to NOMA, the achievable rate of user u on RB m is given by

$$r_{u,m} = \frac{B}{M} \log_2 \left(1 + \frac{p_{u,m} h_{u,m}}{\sum\limits_{k \in \mathcal{U}: b_m(k) > b_m(u)} x_{k,m} p_{k,m} h_{k,m} + \sigma^2} \right) \tag{1}$$

Therefore, the achievable rate of user u can be expressed as $r_u = \sum_{m=1}^{M} x_{u,m} r_{u,m}$. Accordingly, the uplink transmission time of task u is $T_u = \frac{L_u}{r_u}$ and the energy consumption of user u is $E_u = T_u p_u$, where $p_u = \sum_{m=1}^{M} p_{u,m}$.

B. Computing Resources

The total computing capacity of the cloudlet is C CPU cycles per second, which is equally divided into a set of computational RBs $\mathcal{R}_c, |\mathcal{R}_c| = N$, each $n \in \mathcal{R}_c$ with C/N CPU cycles per second. In practice, a computational RB can be a core or one virtual machine.

We denote the number of computational RBs allocated to user u as q_u. Accordingly, the computational time is $Q_u = \frac{NW_u}{q_u C}$, where $\sum_{u=1}^{U} q_u \leq N$.

III. Problem Formulation

We consider the problem of total transmission energy minimization, under the constraint on the completion times of the computational tasks. That is, for each user u, the sum of the transmission and computational times should not violate a certain maximum delay D_u, i.e., $T_u + Q_u \leq D_u$. This delay constraint can be turned into the following rate requirement

$$r_u \geq \frac{L_u}{D_u - Q_u} \Leftrightarrow R_u \geq \frac{ML_u}{B(D_u - Q_u)}, \tag{2}$$

where $D_u - Q_u > 0$, and $R_u = r_u \times \frac{M}{B}$ represents the achievable data rate over a unit bandwidth.

Then, the optimal resource allocation problem can be formulated as

$$\begin{aligned}
\text{P1}: & \underset{\vec{P}, \vec{X}, \vec{q}}{\text{minimize}} \sum_{u \in \mathcal{U}} E_u \\
\text{s.t.} \quad & \text{C1}: R_u \geq \frac{ML_u}{B(D_u - Q_u)}, \forall u \in \mathcal{U} \\
& \text{C2}: \sum_{m \in \mathcal{R}_f} p_{u,m} \leq P_u^{\max}, \forall u \in \mathcal{U} \\
& \text{C3}: x_{u,m} \in \{0,1\}, \forall u \in \mathcal{U}, \forall m \in \mathcal{R}_f \\
& \text{C4}: \sum_{m \in \mathcal{R}_f} x_{u,m} = 1, \forall u \in \mathcal{U} \\
& \text{C5}: \sum_{u \in \mathcal{U}} x_{u,m} = U_m, \forall m \in \mathcal{R}_f \\
& \text{C6}: q_u \in \mathcal{N}, \forall u \in \mathcal{U} \\
& \text{C7}: \sum_{u \in \mathcal{U}} q_u \leq N,
\end{aligned}$$

where $\vec{P}, \vec{X}, \vec{q}$ are the vectors of allocated powers $p_{u,m}$, RB assignments $x_{u,m}$ and computational unit assignments q_u, respectively. Inequality constraints C1 reflects the minimum data rate requirement for each user. Constraints C2 limits the transmit power for each user to a maximum transmit power. Constraint C3 restricts $x_{u,m}$ to binary choices. Constraints C4 and C5 ensures that each user can access only one RB and one RB can accommodate up to U_m users, respectively. Constraints C6 and C7 ensure that the number of allocated computational RBs is an integer number and that the total number of the allocated computational RBs does not exceed the available ones. Note that for simplicity in problem P1 we assume that the communication and computing resources of a user cannot be reused by others even after the user completes the transmission or the computation. As a

consequence, a solution to P1 slightly overestimates the required transmission energy.

It can be observed that the formulated problem P1 belongs to mixed integer nonlinear programming (MILNP) [16], and as shown in [10], it is NP-hard, even if the computational requirements are considered to be zero. To find a sub-optimal solution to P1, we propose a tractable heuristic in two steps, including user clustering, and joint power and computational RB allocation.

IV. NOMA-assisted MEC Resource Allocation

A. User Clustering

To define a reasonable heuristic solution and ensure user fairness, first we restrict the possible solutions to those, where U_m, the number of users assigned to an RB is balanced, that is, $U_m = \{\lceil \frac{U}{M} \rceil - 1, \lceil \frac{U}{M} \rceil\}, \forall m \in \mathcal{R}_f$, and $\sum_{m=1}^{M} U_m = U$, where $\lceil \cdot \rceil$ is the ceiling function.

Then, we assign users to RBs in a way that leads to sum rate maximization in an OMA system with equal transmission resource allocation within a cluster [17]. The assignment is based on maximum weight bipartite graph matching [18] as follows. One set of the nodes of the bipartite graph is formed by the set of U users, while the other set is formed by the set of M transmission RBs. Edges represent the transmission channels, with a weight of $h_{u,m}$ for user u and RB m. On this basis, we run the maximum weight bipartite graph matching to pair M users to the RBs. Following this, we remove the paired users and repeat the above pairing process until all users are assigned with one RB. See Algorithm 1, lines 1 to 9.

Once the user clustering is done, the variables $x_{u,m}$ are known and can be removed from the problem formulation. Then, the problem can be simplified as

$$\begin{aligned} \text{P2}: \underset{\vec{P},\vec{q}}{\text{minimize}} \ & \sum_{u\in\mathcal{U}} E_u \\ \text{s.t.} \ \ \text{C1}: \ & R_u \geq \frac{ML_u}{B(D_u - Q_u)}, \forall u \in \mathcal{U} \\ \text{C2}: \ & \sum_{m\in\mathcal{R}_f} p_{u,m} \leq P_u^{\max}, \forall u \in \mathcal{U} \\ \text{C3}: \ & q_u \in \mathcal{N}, \forall u \in \mathcal{U} \\ \text{C4}: \ & \sum_{u\in\mathcal{U}} q_u \leq N. \end{aligned}$$

B. Joint Power and Computational RBs Allocation

As users in different clusters are assigned to different frequency bands, there exists no inter-cluster interference. However, due to the commonly shared computational RBs, the energy minimization in different clusters is still interdependent. This coupling among the clusters makes the problem P2 difficult to analyse. To address this, we first consider the power allocation under a given computational RBs allocation, and derive general guidelines. Then, based on these guidelines, we propose a joint power and computational RBs allocation algorithm.

Algorithm 1 NOMA assisted MEC user clustering and resource allocation algorithm.

1: **User clustering:**
2: **while** $(U \geq M)$
3: run the maximum weight matching algorithm;
4: allocate the RBs to the paired M users;
5: removed the paired users;
6: $U \leftarrow U - M$;
7: **end while**
8: run the maximum weight matching algorithm;
9: fix the remaining $M - U$ users to the paired RBs;
10: **Computational RBs allocation:**
11: Initialization: $q_u^{\text{Int}} \leftarrow \left\lceil \frac{q_u}{\left(D_u - \frac{ML_u}{B\log_2(1+\frac{P_u^{\max}}{\sigma^2})}\right)C} \right\rceil$;
12: $q_u \leftarrow q_u^{\text{Int}}$;
13: $\hat{q} \leftarrow N - \sum_{u=1}^{U} q_u^{\text{Int}}$;
14: **while** $(\hat{q} \neq 0)$
15: $R_{u,m} = \frac{ML_u}{B(D_u - NW_u/(q_u C))}$;
16: $\hat{R}_{u,m} = \frac{ML_u}{B(D_u - NW_u/((q_u+1)C))}$;
17: $\Delta_{u^\star} \leftarrow \max \left\{ \sum_{u\in U_m} \frac{\sigma^2(2^{R_{u,m}}-1)2^{\sum_{l=u+1}^{U_m} R_{l,m}}}{R_{u,m}h_{u,m}} - \sum_{u\in U_m} \frac{\sigma^2(2^{\hat{R}_{u,m}}-1)2^{\sum_{l=u+1}^{U_m} R_{l,m}}}{\hat{R}_{u,m}h_{u,m}} \right\} | \forall u \in U$.
18: $q_{u^\star} \leftarrow q_{u^\star} + 1$.
19: $\hat{q} \leftarrow \hat{q} - 1$.
20: **end while**
21: **Power allocation:**
22: $Q_u \leftarrow \frac{NW_u}{q_u C}, \forall u \in U$;
23: $R_{u,m}^{\min} \leftarrow \frac{L_u}{D_u - Q_u}, \forall u \in U$;
24: $p_{u,m} \leftarrow \frac{\sigma^2(2^{R_{u,m}^{\min}}-1)2^{\sum_{l=u+1}^{U_m} R_{l,m}^{\min}}}{h_{u,m}}$.

Specifically, when q_u is given, the power allocation problem is independent across clusters. Therefore, the energy minimization in each cluster is equivalent to the energy minimization of the whole system. Without loss of generality, we consider cluster m, and formulate the corresponding problem as

$$\begin{aligned} \text{P3}: \underset{\vec{P}_m}{\text{minimize}} \ & \sum_{u\in U_m} E_{u,m} \\ \text{s.t.} \ \ \text{C1}: \ & R_{u,m} \geq \frac{ML_u}{B(D_u - Q_u)}, \forall u \in U_m \\ \text{C2}: \ & p_{u,m} \leq P_u^{\max}, \forall u \in U_m \end{aligned}$$

where $\vec{P}_m = [p_{1,m}, \cdots, p_{U_m,m}]$ and $E_{u,m} = \frac{ML_u p_{u,m}}{BR_{u,m}}$. Besides, the data rate of user u on RB m is given by

$$R_{u,m} = \log_2\left(1 + \frac{p_{u,m}h_{u,m}}{\sum_{l=u+1}^{U_m} p_{l,m}h_{l,m} + \sigma^2}\right). \quad (3)$$

Theorem 1: To minimize the energy consumption, power should be allocated such that $R_{u,m} = R_{u,m}^{\min}$, $\forall u, m$, where $R_{u,m}^{\min}$ is the minimum rate that still fulfills the delay requirement, i.e., $R_{u,m}^{\min} = \frac{ML_u}{B(D_u - Q_u)}$.

Proof: P3 is clearly non-convex. We reformulate it by changing the variables from the power values to data rates according to (3), and obtain

$$p_{u,m} = \frac{\sigma^2(2^{R_{u,m}} - 1)2^{\sum_{l=u+1}^{U_m} R_{l,m}}}{h_{u,m}}. \tag{4}$$

Substituting the above equation into P3, the optimization problem becomes

$$\text{P4}: \underset{R_{u,m}}{\text{minimize}} \sum_{u \in U_m} \frac{(2^{R_{u,m}} - 1)2^{\sum_{l=u+1}^{U_m} R_{l,m}} \sigma^2 ML_u}{R_{u,m} h_{u,m} B} \tag{5}$$

$$\text{s.t.} \quad \text{C1}: R_{u,m} \geq R_{u,m}^{\min}, \forall u \in U_m$$

$$\text{C2}: \frac{\sigma^2(2^{R_{u,m}} - 1)2^{\sum_{l=u+1}^{U_m} R_{l,m}}}{h_{u,m}} \leq P_u^{\max}.$$

Focusing on $R_{u,m}$, the objective function has the following form:

$$f(R_{u,m}) = \frac{(2^{R_{u,m}} - 1)a}{R_{u,m}} + b \times 2^{R_{u,m}} + c \tag{6}$$

where

$$a = \frac{2^{\sum_{l=u+1}^{U_m} R_{l,m}} \sigma^2 ML_u}{h_{u,m} B},$$

$$b = \sum_{l=1}^{u-1} \frac{(2^{R_{l,m}} - 1)2^{\sum_{k=l+1, k \neq u}^{U_m} R_{k,m}} \sigma^2 ML_l}{R_{l,m} h_{l,m} B},$$

$$c = \sum_{l=u+1}^{U_m} \frac{(2^{R_{l,m}} - 1)2^{\sum_{k=l+1}^{U_m} R_{k,m}} \sigma^2 ML_l}{R_{l,m} h_{l,m} B}.$$

We observe that a, b and c are positive constants when we consider only the variable $R_{u,m}$, i.e., when other variables are fixed and given.

We would like to see how $f(R_{u,m})$ depends on the rate $R_{u,m}$. For the convenience of the notation, let us first replace $R_{u,m}$ with x ($x > 0$). After some mathematical manipulations, we obtain

$$\frac{\partial f(x)}{\partial x} = \frac{(x \cdot 2^x \cdot \ln 2 - 2^x + 1)a}{x^2} + b \cdot 2^x \cdot \ln 2.$$

Both $2^x > 0$ and $x^2 > 0$. Let us consider $g(x) = x \cdot 2^x \cdot \ln 2 - 2^x + 1$, with derivative

$$\begin{aligned} \frac{\partial g(x)}{\partial x} &= x \cdot 2^x \cdot \ln 2 \cdot \ln 2 + 2^x \cdot \ln 2 - 2^x \cdot \ln 2 \\ &= x \cdot 2^x \cdot \ln 2 \cdot \ln 2 > 0. \end{aligned}$$

Thus, $g(x)$ increases with x. Since $g(0) = 0$, we can conclude that $g(x) > 0$, and consequently $\frac{\partial f(x)}{\partial x} > 0$, i.e., $\frac{\partial f(R_{u,m})}{\partial R_{u,m}} > 0$.

Owing to $\frac{\partial f(R_{u,m})}{\partial R_{u,m}} > 0$, the objective function increases with $R_{u,m}$. Thus, to minimize the energy consumption, the minimum $R_{u,m}$ should be used for all the users. Considering the constraint C1, we have $R_{u,m} = R_{u,m}^{\min}, \forall u \in U_m$. Note that this minimum rate is also most likely to satisfy the power constraint C2. ■

Remark: Note that $R_{u,m} = R_{u,m}^{\min}$ if and only if $D_u = T_u + Q_u$. Therefore, the system will operate at the delay limit. Furthermore, as $R_{u,m}^{\min}$ declines with D_u, a more loose delay requirement yields a lower data rate, and further, a lower energy consumption.

Based on Theorem 1, we peopose a greedy computational RB allocation scheme, that allocates the computational RBs to the users one by one, and for each RB allocation, selects the user such that the energy consumption decrease is maximized.

Specifically, the greedy computational RB allocation scheme works as follows.

First, we assign an initial number of computational RBs to each user to satisfy the delay constraint $Q_u + \frac{L_u}{R_u} \leq D_u$, which can be further expressed as

$$Q_u \leq D_u - \frac{L_u}{R_u^{\max}} \leq D_u - \frac{ML_u}{B \log_2\left(1 + \frac{P_u^{\max} h_{u,m}}{\sigma^2}\right)}. \tag{7}$$

Replacing Q_u with $\frac{NW_u}{q_u C}$, we obtain

$$q_u^{\text{Int}} = \left\lceil \frac{NW_u}{\left(D_u - \frac{ML_u}{B \log_2(1 + \frac{P_u^{\max}}{\sigma^2})}\right) C} \right\rceil, \forall u \in \mathcal{U}. \tag{8}$$

Second, the remaining RBs are allocated one by one to the users, such that the decrease of the total energy consumption is maximized in each step. Allocating a computational RB to user u in cluster m affects the energy consumption of that cluster. Specifically, increasing the allocated RBs from q_u to $q_u + 1$ changes the rate requirement from $R_{u,m} = R_{u,m}^{\min} = \frac{ML_u}{B(D_u - NW_u/(Cq_u))}$ to $\hat{R}_{u,m} = \frac{ML_u}{B(D_u - NW_u/(C(q_u+1)))}$. Accordingly, based on (5), the energy decrease, considering all the users in cluster m becomes

$$\begin{aligned} \Delta_u = &\sum_{u \in U_m} \frac{(2^{R_{u,m}} - 1)2^{\sum_{l=u+1}^{U_m} R_{l,m}} \sigma^2 ML_u}{R_{u,m} h_{u,m} B} \\ &- \sum_{u \in U_m} \frac{(2^{\hat{R}_{u,m}} - 1)2^{\sum_{l=u+1}^{U_m} R_{l,m}} \sigma^2 ML_u}{\hat{R}_{u,m} h_{u,m} B}. \end{aligned} \tag{9}$$

Then, the extra RB is allocated to the user with the maximum Δ_u. This process is repeated until there are no computational RBs left. The greedy allocation is summarized in Algorithm 1 (lines 10 to 20). It is possible that there exists no feasible solution for the proposed solution. This happens when either the required initial computational RBs exceeds N (line 12) or the required power exceeds the maximum power constraint (line 24).

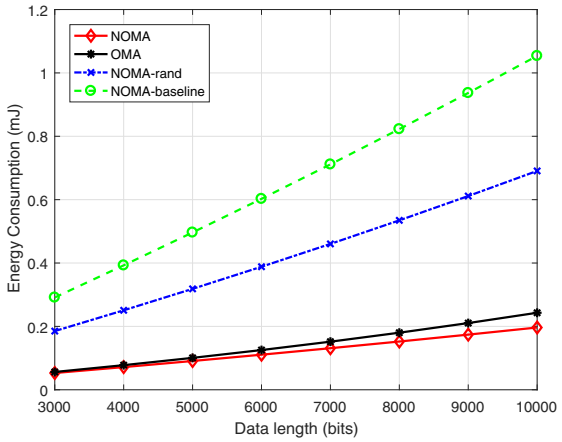

Fig. 1: Energy consumption versus data length for all four considered algorithms: $D_u = 500$ ms and $N = 30$.

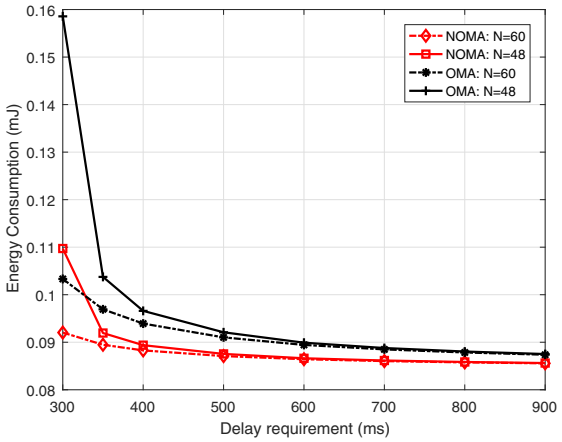

Fig. 2: Energy consumption versus delay requirement: $L_u = 5000$ bits.

TABLE I: Simulation Parameters.

Parameter	Value
Maximum transmit power	20 [dBm]
Channel bandwidth	180 [KHz]
Thermal noise density	-174 [dBm/Hz]
Path-loss model	$126 + 38 \log_{10}(d)$, d in km
Cell radius	1 [km]
Computing need per user W	1 [Giga cycle]
Computing power per process unit	1 [Giga cycle/s]
Number of processing unit N	30
Data input size per user L	5000 [bits]
Delay requirement D	500 [ms]
Number of clusters M	4
Number of users U	12

V. Performance Evaluation

In this section we evaluate the performance of the proposed NOMA-assisted MEC. The default simulation parameters are listed in Table I.

As for baseline algorithms, OMA with equal degrees of freedom is considered. We adopt the same user clustering and joint transmission power and computational RB allocation algorithms as proposed for NOMA. We also adopt two NOMA schemes. NOMA-baseline is the algorithm proposed in [15], while NOMA-rand follows the joint computing RB and power allocation, but performs random user clustering and RB assignment. In the simulations, the users are randomly placed within the cell radius following a uniform distribution. The obtained results are averaged over 10^4 times of random trials.

Fig. 1 shows the energy consumption versus data size L_u for all four considered algorithms. It can be seen that the energy consumption grows almost linearly with the data size. This is because when the data length lies within 3000 to 10000 bits, the required data rate is less than 0.1 bit/s/Hz. Therefore, we have

$$\sum_{u\in U_m} E_{u,m} = \sum_{u\in U_m} \frac{(2^{R_{u,m}}-1)2^{\sum_{l=u+1}^{U_m} R_{l,m}}\sigma^2 M L_u}{R_{u,m}h_{u,m}B} =$$

$$\sum_{u\in U_m} \frac{(2^{\sum_{l=u}^{U_m} R_{l,m}}-2^{\sum_{l=u+1}^{U_m} R_{l,m}})\sigma^2 M L_u}{R_{u,m}h_{u,m}B} \approx \sum_{u\in U_m} \frac{\ln 2(\sum_{l=u}^{U_m} R_{l,m}-\sum_{l=u+1}^{U_m} R_{l,m})\sigma^2 M L_u}{R_{u,m}h_{u,m}B} = \sum_{u\in U_m} \frac{\sigma^2 M \ln 2 L_u}{h_{u,m}B}$$

where the approximation utilizes $2^x \approx x \ln 2 + 1$, when $x \to 0$. Clearly, the energy consumption increases with L_u linearly. Similar explanation can be given for the other baseline algorithms. Among the four considered algorithms, our NOMA scheme achieves the lowest energy consumption, followed by OMA and NOMA-rand. The algorithm proposed in [15] is the worst. These results show that all the resources need to be allocated carefully in NOMA to outperform the OMA based solution.

In the following, we focus on the comparison of the proposed NOMA and OMA.

Fig. 2 shows the energy consumption versus delay requirement for NOMA and OMA for different number of processing units N. As expected, the energy consumption for both NOMA and OMA declines as the delay requirement loosens. Moreover, under any delay requirement, NOMA achieves lower energy consumption than OMA for both N values. The energy consumption, as well as the NOMA gain increases sharply as the delay limit gets strict.

In general, we can look at the relation between the energy consumption and delay requirement. We have

$$\sum_{u\in U_m} E_{u,m} = \sum_{u\in U_m} \frac{(2^{R_{u,m}}-1)2^{\sum_{l=u+1}^{U_m} R_{l,m}}\sigma^2 M L_u}{R_{u,m}h_{u,m}B} \approx \sum_{u\in U_m} \frac{2^{\sum_{l=u+1}^{U_m} R_{l,m}}\sigma^2 M L_u \ln 2}{h_{u,m}B},$$

where the approximation utilizes $\frac{2^{R_{u,m}}-1}{R_{u,m}} \approx \ln 2$ when $R_{u,m} \to 0$. Then, if the the computing RB allocation remains fixed, we have $R_{u,m} = \frac{L_u M}{B(D_u - Q_u)} \propto K D_u^{-1}$. Substituting this into the equation before we obtain

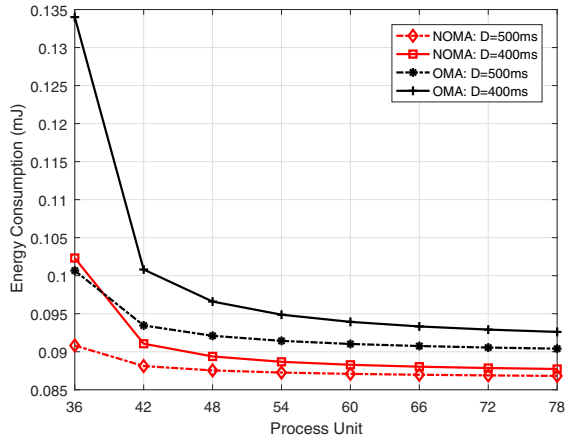

Fig. 3: Energy consumption versus number of computing RBs: $L_u = 5000$ bits.

$\sum_{u\in U_m} E_{u,m} \approx \sum_{u\in U_m} \frac{\ln 2\sigma^2 M L_u 2^{\sum_{l=u+1}^{U_m} KD_l^{-1}}}{h_{u,m}B}$. That is, the relation between the energy consumption and delay requirement follows $e^{1/x}$. For large delay limit D, the above equation can be further approximated as $\sum_{u\in U_m} E_{u,m} \approx \sum_{u\in U_m} \frac{\sigma^2 M L_u \ln 2(\ln 2\sum_{l=u+1}^{U_m} KD_l^{-1}+1)}{h_{u,m}B}$, which fits the results after $D \geq 600$ ms.

Finally, Fig. 3 shows how the energy consumption varies as the number of computing RB increases. Two scenarios with different delay requirements are considered. NOMA outperforms OMA in terms of energy consumption for both D values, with increasing gain at low number of processing units. Adding processing units has diminishing gain, since the processing time Q_u is inversely proportional to q_u.

To summarize, the energy consumption gap between NOMA and OMA increases when the users' required data rates are large. This happens when there is a large amount of data to transmit, or when the computing capacity is small, or when the delay limit is strict. The gain of applying NOMA can be up to 10-30% for the considered scenario.

VI. Conclusion

In this paper we evaluated whether advanced transmission techniques proposed for 5G, like NOMA, can improve the performance of mobile edge computing. The NOMA-assisted MEC resource allocation for energy minimization has high computational complexity. To address it, we first propose a heuristic method for subcarrier allocation based on maximum weight bipartite graph matching. Then, the joint power and computational resource allocation is tackled, in which optimal power allocation solution is ensured. Our numerical results demonstrate that i) to ensure the superiority of NOMA over OMA, the joint resource allocation should be carefully designed, otherwise, NOMA can be even worse than OMA, as the NOMA-baseline algorithm; ii) the performance gain between NOMA and OMA is large when high transmission rate is required, which happens under low delay constraint and heavy transmission and computational needs.

References

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.

[2] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," in *IEEE ICC*, May 2017.

[3] Y. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *[Online]. Available: https://arxiv.org/pdf/1704.00163.pdf*.

[4] M. H. Chen, M. Dong, and B. Liang, "Joint offloading decision and resource allocation for mobile cloud with computing access point," in *IEEE ICASSP*, March 2016.

[5] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994–6009, Sept 2017.

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

[7] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, October 2016.

[8] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[9] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, Jan 1972.

[10] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[11] M. Zeng and V. Fodor, "Sum-rate maximization under QoS constraint in MIMO-NOMA systems," in *IEEE WCNC*, Apr 2018.

[12] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[13] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[14] F. Wang, J. Xu, and Z. Ding, "Optimized multiuser computation offloading with multi-antenna noma," in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2017.

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

[16] Belotti, Pietro, et al., *Mixed-integer nonlinear optimization*. Acta Numerica, 2013, vol. 22.

[17] D. Yuan, J. Joung, C. K. Ho, and S. Sun, "On tractability aspects of optimal resource allocation in ofdma systems," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 863–873, Feb 2013.

[18] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 1-2, pp. 83–97, 1955.