EDGE COMPUTING AWARE NOMA FOR 5G NETWORKS

ABBAS KIANI
NIRWAN ANSARI

TR-ANL-2017-007
December 13, 2017

ADVANCED NETWORKING LABORATORY
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
NEW JERSEY INSTITUTE OF TECHNOLOGY
Abstract

With the fast development of Internet of things (IoT), the fifth generation (5G) wireless networks need to provide massive connectivity of IoT devices and meet the demand for low latency. To satisfy these requirements, Non-Orthogonal Multiple Access (NOMA) has been recognized as a promising solution for 5G networks to significantly improve the network capacity. In parallel with the development of NOMA techniques, Mobile Edge Computing (MEC) is becoming one of the key emerging technologies to reduce the latency and improve the Quality of Service (QoS) for 5G networks. In order to capture the potential gains of NOMA in the context of MEC, this paper proposes an edge computing aware NOMA technique which can enjoy the benefits of uplink NOMA in reducing MEC users’ uplink energy consumption. To this end, we formulate a NOMA based optimization framework which minimizes the energy consumption of MEC users via optimizing the user clustering, computing and communication resource allocation, and transmit powers. In particular, similar to frequency Resource Blocks (RBs), we divide the computing capacity available at the cloudlet to computing RBs. Accordingly, we explore the joint allocation of the frequency and computing RBs to the users that are assigned to different order indices within the NOMA clusters. We also design an efficient heuristic algorithm for user clustering and RBs allocation, and formulate a convex optimization problem for the power control to be solved independently per NOMA cluster. The performance of the proposed NOMA scheme is evaluated via simulations.

Keywords

Mobile edge computing, NOMA, power control

I. INTRODUCTION

With the fast development of mobile Internet and Internet of Things (IoT), mobile data traffic is anticipated to witness explosive growth in the years to come. To support this unprecedented growth, both academic and industrial communities have conducted extensive research to design the fifth generation (5G) wireless networks. The 5G networks are to offer significant improvements of wireless network capacity and the user experience [1], and demand spectral-efficient multiple access techniques. To this end, Non-Orthogonal Multiple Access (NOMA) techniques [2] have been recognized as promising solutions for 5G and have attracted extensive research recently. In contrast with Orthogonal Multiple Access (OMA) techniques, where the radio resources are allocated orthogonally to multiple users, NOMA allows multiple users to share the same resources. By serving multiple users simultaneously over the same radio resources, more users can be supported, thus leading to a significant increase in the network capacity. This improvement is nevertheless available at the expense of intra-cell interference as well as additional complexity at the receiver side. To deal with the intra-cell interferences and the complexity, NOMA splits the users in the power domain based on their respective channel conditions and employs efficient Multi-User Detection (MUD) technique such as Successive Interference Cancellation at the receiver side [3].

In parallel with the explosive growth of mobile data traffic, our daily life witnesses a significant increase in demands for running sophisticated applications in the mobile devices for social networking, business, etc. [4]. Moreover, in future user-centric 5G networks, the IoT users participate in sensing and computing tasks, and computation-intensive tasks need to be offloaded to either the cloud or the computing resources at the edge. To this end, Mobile Edge Computing (MEC), which is being standardized by an Industry Specification Group (ISG) launched by the European Telecommunications Standards Institute (ETSI) [5], is recognized as one of the key emerging technologies for 5G networks. The idea of MEC is to provide computing capabilities in proximity of users and within the Radio Access Network (RAN), thereby reducing the latency and improving the Quality of Service (QoS) [5].

In this paper, we focus on two aforementioned emerging technologies of 5G, i.e., NOMA and MEC, and propose a novel MEC aware NOMA technique for 5G networks. Our proposed scheme is motivated by the fact that the joint allocation of communication and computing resources greatly improves the performance of the system. In other words, it may happen that one type of resources is wasted due to congestion of other type of resources. While several works such as [6], [7] have investigated the joint allocation of computing and communication resources, non of the existing works consider a joint optimization technique in the context of NOMA with consideration of intra-cell interferences. To this end, the current study aims to address the aforementioned issue by proposing a joint optimization technique to allocate the computing and communication resources based on the requirements of both MEC and NOMA.

Contributions: We have made three major contributions. 1) We propose a novel NOMA augmented edge computing model that captures the gains of uplink NOMA in MEC users’ energy consumption. Specifically, we design a NOMA based optimization framework that minimizes the energy consumption of MEC users via optimizing the user clustering, computing and communication resource allocation, and transmit powers. To this end, similar to frequency Resource Blocks (RBs), we define the notion of computing RBs and investigate the joint allocation of the frequency and computing RBs. More importantly, we consider a time constraint for each edge computing task, and accordingly the minimum data rate requirement of each user is established based on its deadline. 2) We design an efficient heuristic algorithm for user clustering and RBs allocation. Moreover, we formulate a convex optimization problem for the transmission power control to be solved independently per NOMA cluster. 3) We evaluate the performance of our proposed NOMA scheme and the heuristic algorithm via extensive simulations in which we show the benefits of NOMA in reducing the MEC users’ uplink energy consumption. We also evaluate the effects of computing capacity and its division strategy of computing RBs on the total energy consumption.
Related works: The related works to this paper include MEC, NOMA, and sub-carrier scheduling. In the past few years, a large and cohesive body of work investigated the major challenges of MEC and the researchers came up with a variety of policies and algorithms. Recently, Chiang et al. [8] summarized the opportunities and challenges of edge computing in the networking context of IoT and Gonzalez et al. [9] explored the applications of edge computing in IoT. Yu et al. [7] proposed a joint subcarrier and CPU time allocation algorithm for MEC. A hierarchical MEC model designed based on the principle of LTE-Advanced backhaul network is introduced in [10] in which the so called field, shallow, and deep cloudlets are located hierarchically in three different tiers of the network. A task scheduling scheme for code partitioning over time and the hierarchical cloudlets is also proposed in [11]. Moreover, a novel approach to mobile edge computing for the IoT architecture is presented in [12]. A hybrid architecture that harnesses the synergies between edge caching and C-RAN is proposed in [13].

Lee et al. [14] explored the fundamental problem of LTE Single-carrier FDMA uplink scheduling by adopting the conventional time domain proportional fair algorithm. As discussed earlier, this paper proposes NOMA based model for MEC. Recently, several research studies have identified the potential benefits of NOMA in both the downlink and uplink. For instance, Al-Imari [15] proposed a NOMA scheme for uplink that allows more than one user to share the same subcarrier while a joint processing is implemented at the receiver to detect the users’ signals. Zhang et al. [16] designed an uplink power control scheme where eNB distinguishes the multiplexing users in the power domain, and theoretically analyzed the outage performance and the achievable sum-rate of the proposed scheme. Liang et al. [17] proposed the so called non-orthogonal random access (NORA) scheme based on SIC to tackle the access congestion problem. In NORA, the difference of time of arrival is used to identify multiple users. Shipon et al. [18] also designed a sum-throughput maximization problem under transmission power constraints for both uplink and downlink NOMA. Moreover, Tabassum et al. [19] characterized the rate coverage probability of a user in NOMA cluster with a given rank as well as the mean rate coverage probability of all users in the cluster for perfect SIC, imperfect SIC, and imperfect worst case SIC.

While there are numerous research activities that investigate NOMA technique and its benefits in 5G networks, there is no prior work that study the advantages of NOMA in the context of edge computing. To this end, the current studied aims at proposing a novel edge computing aware NOMA model to reduce the uplink energy consumption of MEC users via utilizing the gains of uplink NOMA. Moreover, we take into consideration of the deadline requirements of MEC users in the user clustering.

The rest of the paper is organized as follows. Section II describes the system model and problem formulation. We propose our optimization framework and the corresponding heuristic algorithms in Section III. Finally, Sections IV and V present numerical results and conclude the paper, respectively.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a single-cell scenario, where one eNB equipped with a cloudlet serves the uniformly distributed edge computing users. Denote \( U = \{1, \ldots, U\} \) as the set of users each with a task to be offloaded to the cloudlet via eNB. In the following, the users and tasks are used interchangeably. Each task \( u \) is characterized by the workload \( \lambda_u \), i.e., the number of CPU cycles required to complete the execution of the task, and the input \( L_u \), i.e., the number of bits that must be transferred from the user to the eNB.

A. Communication Resources

We assume that the available bandwidth is divided into a set of frequency resource blocks \( R_f = \{1, \ldots, M_f\} \) and the bandwidth of each resource block is \( B \). According to the NOMA schemes, it is assumed that the users transmit over the resource blocks in an non-orthogonal manner, i.e., more than one user can share the same resource blocks. Therefore, the users are assumed to be divided into different groups called the NOMA clusters, where a set of frequency RBs are allocated to each NOMA cluster. Denote \( I = \{1, \ldots, N\} \) as the set of NOMA clusters and \( \beta_{r,i} \) as the binary variable to indicate the allocation of RB \( r \in R_f \) to NOMA cluster \( i \in I \). Here, \( \beta_{r,i} = 1 \) if RB \( r \) is allocated to NOMA cluster \( i \), and \( \beta_{r,i} = 0 \) otherwise. Given the principles of NOMA in which at least two users must share the same frequency resource blocks, we set \( N \leq \lceil \frac{U}{M_f} \rceil \).

We assume that an efficient MUD technique such as SIC [3] is applied at the eNB to decode the message signals in which the users are required to be ordered in each NOMA cluster. We define \( J = \{1, 2, \ldots, u_{max}\} \) as the set of the orders in a cluster. Here, \( u_{max} \) is defined as the maximum number of users allowed to share a RB, hereby, reducing the complexity at the receiver side. It is also assumed that \( Nu_{max} \geq U \). According to the principles of MUD techniques [15], the message of the \( j \)-th user in the cluster is decoded before all the users with higher indices. Consequently, the \( j \)-th user of a cluster experiences interference from all the users \( \{j + 1, j + 2, \ldots\} \) in that cluster. In other words, the first user to be decoded (\( j = 1 \)) will see interference from all the other users \( j = 2, \ldots, u_{max} \), and the second user to be decoded will see interference from the users \( j = 3, \ldots, u_{max} \), and so on.

Denote \( p_r^u \) as the transmission power of user \( u \) over RB \( r \) and \( \alpha_{u,j}^i \) as the binary variable to indicate the assignment of user \( u \) to the \( j \)-th order in cluster \( i \). Here, \( \alpha_{u,j}^i = 1 \) if the assignment occurs, and \( \alpha_{u,j}^i = 0 \) otherwise. The achievable data rate of user \( u \) is given by

\[
R_u = \sum_{i \in I} \sum_{j \in J} \alpha_{u,j}^i \sum_{r \in R_f} \beta_{r,i} B \log(1 + \frac{h_u^r p_r^u}{\sigma^2 + \sum_{k \in U \setminus u} \sum_{l=j+1}^{u_{max}} \alpha_{u,l}^i h_k^r p_k^r})
\]
where $h_{ru}$ denotes the channel gain between user $u$ and the eNB on RB $r$, and $\sigma^2$ is the noise power. Note that by $h_{ru}$, we assume that the channel conditions vary across RBs as well as users. Therefore, the uplink transmission time of task $u$ is

$$T_u = \frac{L_u}{R_u}$$

(2)

and the energy consumption of user $u$ is given by

$$E_u = T_u \sum_{r \in R_f} p^r_u$$

(3)

B. Computing Resources

Analogous to the communication resources, we assume the computing capacity of the cloudlet is divided into different computing RBs. For example, one computing RB can be one virtual machine or one core. Denote $R_c = \{1, ..., M_c\}$ as the set of all computing RBs. We also assume the capacity of one computing RB is equal to $C$ CPU cycles per second. It is assumed that a number of computing RBs is allocated to each order index of each NOMA cluster. Therefore, the computing time of task $u$ is

$$Q_u = \sum_{i \in I} \sum_{j \in J} \alpha_{i,j} u \lambda_u x_{i,j} C$$

(4)

where $x_{i,j}$ denotes the number of computing RBs allocated to order index $j$ of NOMA cluster $i$.

III. Optimization Problem

In this section, we formulate an optimization problem to minimize the summation of the energy consumption of all the users with constraint on the total transmission and computing time. In particular, we enforce a deadline $D_u$ as an upper limit on the total time of task $u$ as follows

$$T_u + Q_u \leq D_u$$

(5)

Constraint (5) is equivalent to the following data rate requirement

$$R_u \geq \frac{L_u}{D_u - Q_u}$$

(6)

where $Q_u \leq D_u$. Therefore, we propose to solve the following optimization problem,

P1 : minimize $\sum_{u \in U} E_u$

s.t. C1 : $R_u \geq \frac{L_u}{D_u - Q_u}$ $\forall u \in U$

C2 : $Q_u \leq D_u$ $\forall u \in U$

C3 : $\sum_{i \in I} \sum_{j \in J} \alpha_{i,j} u \lambda_u = 1$ $\forall u \in U$

C4 : $\sum_{u \in U} \sum_{j \in J} \alpha_{i,j} u \geq 2$ $\forall i \in I$

C5 : $\sum_{i \in I} \beta_{r,i} \leq 1$ $\forall r \in R_f$

C6 : $\sum_{i \in I} \sum_{j \in J} x_{i,j} \leq M_c$

C7 : $\alpha_{i,j} u \leq \alpha_{i,j-1} u \forall u \in U$, $i \in I$, $2 \leq j \leq u_{\text{max}}$

C8 : $\sum_{r \in R_f} p^r_u \leq P_u^{\text{max}}$ $\forall u \in U$

C9 : $p^r_u \geq 0$ $\forall u \in U$, $r \in R_f$
C10: $\alpha_u^{i,j} \in \{0, 1\}$ $\forall u \in U$, $i \in I$, $j \in J$

C11: $\beta^{r,i} \in \{0, 1\}$ $\forall r \in R_f$, $i \in I$

C12: $x^{i,j} \in \mathbb{N}$ $\forall i \in I$, $j \in J$

Inequality constraints C1 is the computing aware minimum data rate requirement per user. $Q_u$ is upper bounded by $D_u$ in constraint C2. The equality constraint C3 is to ensure that each user is assigned to only one NOMA cluster and also only one order index within that NOMA cluster. Constraint C4 is to ensure that at least two users are assigned to each cluster. In addition, we use the inequality constraints C5 to ensure that each frequency RB is allocated to only one cluster and C6 to bound the total allocated computing RBs by the available computing RBs. Constraint C7 is designed to give assignment priority to a lower value of order in one NOMA cluster over all the higher values of order in the same cluster. Moreover, by constraints C8, the total transmission power of user $u$ is limited to power budget $P_u^{\text{max}}$. Finally, constraint C9 restricts variable $p_u^r$ to positive values, constraints C10 and C11 restrict the variables $\alpha_u^{i,j}$ and $\beta^{r,i}$ to binary choices, and constraint C12 is to restrict variables $x^{i,j}$ to the integer values.

Note that P1 defines a flexible Mixed Integer Non-Linear Programming (MINLP) problem which involves binary, integer and real variables. However, finding an optimal solution to this problem is intractable and presents computing complexity where the complexity grows fast with the number of variables. Furthermore, the objective function in P1 is not generally a convex function. Therefore, in order to reduce the complexity and obtain high quality solutions in reasonable time, we follow a two-phase approach. First, we propose an efficient heuristic algorithm for user clustering and RBs allocation. In fact, the heuristic algorithm is designed to decide about the binary and the integer variables. Second, having the binary and integer variables removed, we formulate a convex optimization problem for the transmission power control to be solved independently per NOMA cluster.

### A. User Clustering and RBs allocation

The pseudo code for the user clustering and RBs allocation is summarized in Algorithms [1]. As shown in this algorithm, we carry out the user clustering, computing RBs allocation and frequency RBs allocation in three separated phases. Denote $\bar{h}_u = \frac{\sum_{r \in R} h_u^r}{M_f}$ as the average channel condition of user $u$. In the clustering phase (lines 2-9), we follow a clustering method based on the average channel conditions. In other words, the users with the higher average channel gains are assigned to the lower order indices of the clusters. By doing so, a user with a higher channel gain does not interfere to the users with the lower channel gains since its interference is canceled out by the SIC receiver. Thus, the users with higher channel gains can transmit with the maximum transmission power, thereby, improving the sum-rate of the cluster. Let $u^{i,j}$ be the user assigned to the $j$th index of cluster $i$ and $U^i$ as the set of all the users assigned to cluster $i$. Note that $u^{i,j}$ and $U^i$ are known after the clustering phase. In the next phase, i.e., the computing RBs allocation phase (lines 11-31), we first allocate the computing RBs to satisfy condition $Q_u < D_u$ for all the users. Here, we assume the available number of computing RBs are sufficient to satisfy such a condition. Then, for each of the remaining computing RBs, we search the set of all the clusters to find a favorite cluster, i.e., $i^*$. Denote $Q_u^{x^{i,j}+1}$ as the computing time of user $u^{i,j}$ if we increase $x^{i,j}$ to $x^{i,j} + 1$. The favorite user is identified by comparing the terms $\frac{L_{u^{i,j}}^{r,i} - Q_u^{x^{i,j}}}{D_{u^{i,j}} - Q_u^{x^{i,j}+1}} - \frac{L_{u^{i,j}}^{r,i} - Q_u^{x^{i,j}+1}}{D_{u^{i,j}} - Q_u^{x^{i,j}} + 1}$. By doing so, we not only take into consideration of the input size of the users but also their deadlines in the computing resource allocation.

The procedure for the frequency RBs allocation phase is presented in lines 33-54. As shown in lines 33-43, we initially allocate the channels to satisfy the minimum data rate requirement for all the users. Denote $R^i$ as the set of channels allocated to cluster $i$ and $I^i$ as the set of clusters consisting of some users with unsatisfied minimum data rate requirement. Let $R_{f, i}$ be the set of already allocated channels. The favorite cluster for each channel, i.e., $i$, is the cluster that achieves the best sum-rate over that channel as compared to those clusters which have users with unsatisfied minimum data rate. The sets $R_{f, i}$, $R^i$, and powers $P^u_r$ are accordingly updated after each channel allocation. Note that we assume the maximum transmit power of each user is equally divided among all the allocated channels to its corresponding cluster. Then, for each of the remaining frequency RBs to be allocated, we search over all the clusters and identify a favorite cluster by comparing $\sum_{u \in U_i} (\frac{L_u^r}{R_u^r} - \frac{L_u^r}{R_u^r})^{P_{u}^{\text{max}}}$ (lines 44-54). Here, $R_{u}^{R_u^{\text{max}}}$ and $R_{u}^{R_u^{\text{max}}}$ denote the data rate of user $u$ based on the current allocation and that of based on the current allocation as well as the allocation of RB $r$ to cluster $i$, respectively. Note that these data rates are calculated based on the current values of $p_u^r$. In fact, the favorite cluster for each channel is the one that achieves the maximum increase in the objective function of P1. After identification of favorite cluster $i^*$, we accordingly update set $R^i$ as well as powers $P^u_r$ for all the users in $U^{i^*}$.

In problem P1, we do not take into consideration of contiguous RB constraint. Under RB contiguity constraint, all the RBs allocated to a single user must be contiguous in frequency [14]. Therefore, one can extend problem P1 by adding the following
Algorithm 1

1: User Clustering
2: sort the users in $U$ such that $h_1 \geq h_2 \geq \ldots \geq h_U$
3: for all $j \in J$ do
4:  if $jN \leq U$ then
5:    assign users $\{(j-1)N+1, \ldots, jN\}$ as the $j$-th users of clusters $\{1, \ldots, N\}$, respectively
6:  else
7:    assign users $\{(j-1)N+1, \ldots, U\}$ as the $j$-th users of clusters $\{1, \ldots, U-(j-1)N\}$, respectively
8:  end if
9: end for
10: Computing RBs Allocation
11: for all $i \in I$ do
12:  for all $j \in J$ do
13:   repeat
14:    $x^i,j = x^i,j + 1$
15:    $M_c = M_c - 1$
16:  until condition $Q_{u,i,j} < D_{u,i,j}$ is satisfied
17: end for
18: end for
19: repeat
20:  $\hat{q} = 0$, $\hat{i} \leftarrow \emptyset$ and $\hat{j} \leftarrow \emptyset$
21:  for all $i \in I$ do
22:    for all $j \in J$ do
23:      if $\frac{L_{u,i,j}}{D_{u,i,j}} - \frac{L_{u,i,j}}{D_{u,i,j} - Q_{u,i,j}} \geq \hat{q}$ then
24:        $\hat{i} \leftarrow i$ and $\hat{j} \leftarrow j$
25:        $\hat{q} \leftarrow \frac{L_{u,i,j}}{D_{u,i,j} - Q_{u,i,j}} - \frac{L_{u,i,j}}{D_{u,i,j}}$
26:      end if
27:    end for
28:  end for
29:  $x^\hat{i},\hat{j} = x^\hat{i},\hat{j} + 1$
30:  $M_c = M_c - 1$
31: until $M_c > 0$
32: Frequency RBs Allocation
33: $p'_{u} = P_{u}^{\max} \forall u \in U, r \in R_f$
34: $R_{f_0} \leftarrow \emptyset$, $I' \leftarrow I$ and $R^\hat{i} \leftarrow \emptyset$
35: for all $r \in R_f$ do
36:  $i = \arg\max_{i \in I'} \left( \sum_{u \in U_i} R_u \right)$
37:  $\beta^{r,i} = 1$, $R_{f_0} \leftarrow R_{f_0} \cup r$ and $R^{\hat{i}} = R^{\hat{i}} \cup r$
38:  $P^r_{u} = \frac{P^r_{u}}{|R^r_{u}|+1} \forall u \in U^i, r \in R_f$
39:  if $R_u \geq \frac{L_u}{D_u}$ then
40:    $I' \leftarrow I' \setminus \hat{i}$
41: end if
42: end for
43: $R_f = R_f \setminus R_{f_0}$
44: for all $r \in R_f$ do
45:  $\hat{e} = 0$ and $\hat{i} \leftarrow \emptyset$
46:  for all $i \in I$ do
47:   if $\sum_{u \in U_i} \left( \frac{L_u}{R_u} - \frac{L_u}{R_u^{\max}} \right) P_{u}^{\max} \geq \hat{e}$ then
48:     $\hat{i} \leftarrow i$
49:     $\hat{e} \leftarrow \sum_{u \in U_i} \left( \frac{L_u}{R_u} - \frac{L_u}{R_u^{\max}} \right) P_{u}^{\max}$
50: end if
51: end for
52: $\beta^{r,i} = 1$ and $R^{\hat{i}} \leftarrow R^{\hat{i}} \cup r$
53: $P^r_{u} = \frac{P^r_{u}}{|R^r_{u}|+1} \forall u \in U^i, r \in R_f$
54: end for
contiguity constraint,
\[ \sum_{r=k_1}^{k_2} \beta^{r,i} = k_2 - k_1 + 1 \quad \forall i \in \mathcal{I}, \beta^{k_1,i} = \beta^{k_2,i} = 1 \]  

(7)

While the user clustering and computing RBs allocation phases in Algorithm 1 are still valid to solve the extended problem under constraint (7), the frequency RBs allocation must be changed to comply with this constraint. Notice that the frequency RBs allocation under the contiguity constraint is known to be NP-hard [14]. To address this hardness, we can adopt the RB grouping algorithm proposed in [14]. In a nutshell, \( M_f \) available frequency RBs are divided into \( n \) groups, and the frequency RBs allocation phase in Algorithm 1 is accomplished with the granularity of RB groups. As a result, by extending the unit of consideration from a single RB to the RB group, there is a wider view to obtain optimal frequency RB allocation.

**B. Power Control**

Given the user clustering and frequency RBs allocation, the binary variables \( \alpha_{i,j}^u \) and \( \beta^{r,i} \) in problem P1 are fixed to 0 or 1. Given the computing RBs allocation, the integer constraints vanish and the power control per NOMA cluster follows. In fact, we can now eliminate all the terms that do not depend on the transmit powers. However, the objective function of problem P1 is non-convex in transmit powers. To this end, for the power control, we propose to minimize the total power consumption of the users instead of their energy consumption since the communication time is taken care of in constraint C1. In other words, power consumption minimization problem can provide an accurate solution for the energy consumption minimization as the communication time is upper bounded in constraint C1. Let consider \( i \) as an NOMA cluster of interest. Denote \( U_i = \{1,...,u_{\text{max}}\} \) as the set of the users assigned to order indices \( \{1,...,u_{\text{max}}\} \) of cluster \( i \), respectively. Recall that \( \mathcal{R}_i \) is the set of frequency RBs allocated to cluster \( i \). Therefore, the power control optimization problem for cluster \( i \) can be written as,

\[
P_2: \minimize p_u^r \sum_{u \in U_i} \sum_{r \in \mathcal{R}_i} p^r_u
\]

s.t. C1 : \( R_u \geq \frac{L_u}{D_u - Q_u} \) \( \forall u \in U_i \)

C2 : \( \sum_{r \in \mathcal{R}_i} p^r_u \leq P_{\text{max}}^u \) \( \forall u \in U_i \)

C3 : \( p^r_u \geq 0 \) \( \forall u \in U_i, r \in \mathcal{R}_i \)

Nevertheless, we can prove the following theorems:

**Theorem 1** Optimization problem P2 is equivalent to the optimization problem P3.

\[
P_3: \minimize S_u^r, Z_u^r, R_u \sum_{u \in U_i} \sum_{r \in \mathcal{R}_i} e^{S_u^r}
\]

C1 : \( R_u \geq \frac{L_u}{D_u - Q_u} \) \( \forall u \in \mathcal{R}_i \)

C2 : \( \sum_{r \in \mathcal{R}_i} e^{S_u^r} \leq P_{\text{max}}^u \) \( \forall u \in U_i \)

C3 : \( Z_u^r \leq B \log(1 + \frac{h_r^u e^{S_u^r}}{\sigma^2 + \sum_{k=1+u+1}^{u_{\text{max}}} h_k^r e^{S_k^r}}) \) \( \forall u \in U_i \)

C4 : \( R_u = \sum_{r \in \mathcal{R}_i} Z_u^r \) \( \forall u \in U_i \)

**Proof:** Problem P2 is not a convex problem in the current form. However, let consider the following change of variables

\[
p_u^r = e^{S_u^r} \forall u \in U_i
\]

where obviously (8) is an one to one mapping, and it is always possible to determine the variable \( S_u^r \) from \( p_u^r \) and viceversa. We also add the new variables \( Z_u^r \) and \( R_u \) to the problem formulation, where \( Z_u^r \) is given by

\[
Z_u^r = B \log(1 + \frac{h_r^u e^{S_u^r}}{\sigma^2 + \sum_{k=1+u+1}^{u_{\text{max}}} h_k^r e^{S_k^r}}) \]

(9)
Figure 1: Energy consumption comparison between heuristic and optimal approaches.

and

\[ R_u = \sum_{r \in \mathcal{R}} Z_u^r \]  

Consequently, by substituting the new variables \( S_u^r, Z_u^r \) and \( R_u \) in P2, and after simple algebraic manipulation, we can obtain the formulation of problem P3. Note that in the formulation of P3, we have changed the equality relations \( Z_u^r = B \log(1 + \frac{h_u^r e^{S_u^r}}{\sigma^2 + \sum_{k=u+1}^{\text{max}} h_k^r e^{S_k^r}}) \) to the inequality constraints C3. This change is needed for the convexity of problem P3, and does not affect the optimal solution since the data rate of user \( u \) on RB \( r \) at the optimality cannot be less than \( B \log(1 + \frac{h_u^r e^{S_u^r}}{\sigma^2 + \sum_{k=u+1}^{\text{max}} h_k^r e^{S_k^r}}) \). 

**Theorem 2** Optimization problem P3 is convex in the high SINR regime.

**Proof:** The objective function of P3 is the sum of exponentials and is thus convex. While it is straightforward to prove the convexity of constraints C1, C2, and C4, the inequality constraint C3 is not convex since the throughput function is a non-convex function of the powers. A commonly used solution to deal with the non-convexity of the throughput function is the approximation \( \log(1 + x) \approx \log(x) \) which is a valid approximation in the high SINR regime \([20]-[22]\). As a result of this approximation, constraint C3 becomes

\[ Z_u^r + \log(\sigma^2 h_u^{r-1} e^{-S_u^r} + \sum_{k=u+1}^{\text{max}} h_k^{r-1} h_u^r e^{S_k^r-S_u^r}) \leq 0 \quad \forall u \in \mathcal{U}^i \]

where \( \log(\sigma^2 h_u^{r-1} e^{-S_u^r} + \sum_{k=u+1}^{\text{max}} h_k^{r-1} h_u^r e^{S_k^r-S_u^r}) \) is the log of a sum of exponentials, and thus is convex. Therefore, problem P3 is convex in the high SINR regime and the proof is complete.

Nonetheless, problem P3 is convex and can be solved by efficient optimization techniques such as interior point methods. Note that the complexity of this problem may increase as the numbers of users per NOMA cluster increases. However, each NOMA cluster is assumed to be limited to a few number of users. Moreover, we assume that the channel coefficients \( h_u^r \) are known only to the BS and this problem is solved centrally at the BS for each NOMA cluster.

**IV. Simulation Results**

In this section, we evaluate the performance of the proposed edge computing aware NOMA scheme. We consider a single-cell with 1 km radius in which the users are uniformly distributed within the cell. We set the maximum transmit power of each
Figure 2: Energy consumption of 10 users versus number of frequency RBs for $u_{\text{max}}=1$, 2, 3, and 4.

Figure 3: Spectral efficiency versus average input.
Figure 4: Computing time versus number of users for different numbers of computing RBs and computing capacities.

Figure 5: Computing time fairness index versus number of users for different numbers of computing RBs and computing capacities.
user to 1 W, the maximum available frequency RBs to 30, and the bandwidth of each resource block to 180 kHz. The ITU pedestrian B fast fading model, the COST231 Hata propagation model for micro cell environment [15], [23], and the Lognormal shadowing with 8 dB standard deviation are implemented. The noise power spectral density is set to be 173 dBm/Hz. Unless it is stated otherwise, we assume a total computing capacity of 300 Giga cycles per second is available as 30 computing RBs each with a capacity of 10 Giga cycles per second. The workload of each user ($\lambda_u$) is randomly generated according to a uniform distribution between 0.5 Giga cycles and 1 Giga cycles. The input of each user ($I_u$) and the deadline ($D_u$) are also randomly generated according to uniform distributions between 5000 bits and 7000 bits, and 400 ms to 500 ms, respectively.

Fig. 1 compares the performance of the heuristic (Algorithm 1) and optimal (problem P1) approaches by providing the total energy consumption for different number of users where $u_{\text{max}} = 3$. As we can see in this figure, the heuristic algorithm incurs a total energy consumption quite close to that of the optimal approach. Meanwhile, the heuristic algorithm provides the suboptimal solution within a few seconds while the computation time of the optimal approach grows fast with the number of users.

Fig. 2 illustrates the effect of $u_{\text{max}}$, i.e., maximum number of users allowed to share a frequency RB, on the total energy consumption of 10 users for different number of available frequency RBs. As shown in this figure, the energy consumption decreases by increasing the available number of frequency RBs; second, the energy consumption improves by increasing $u_{\text{max}}$ from 1 to 2, and 3 due to the fact that the spectral efficiency improves by increasing $u_{\text{max}}$. The energy consumption gap is more significant between $u_{\text{max}} = 1$ and $u_{\text{max}} = 2$ or 3 as compared to that of between $u_{\text{max}} = 2$ and $u_{\text{max}} = 3$. This result is attributed to the fact that the intra-cell interference becomes more considerable by increasing $u_{\text{max}}$. The optimal approach provides the performance of the proposed scheme with $u_{\text{max}} = 3$ in terms of the spectral efficiency. As the results show, the spectral efficiency is increased by increasing both the number of users and the average input of each user.

Fig. 3 shows the impact of the total computing capacity as well as its division into the computing RBs on the total energy consumption. In particular, we assume two different cases in terms of the total computing capacity, each with two different division scenarios. In the first case, we consider a total computing capacity of 120 Giga cycles per second, which is divided into 30 computing RBs, each with a capacity of 3 Giga cycles per second and 120 computing RBs each with a capacity of 1 Giga cycles per second as the first and second scenarios, respectively. Similarly, for the second case, the total computing capacity of 90 Giga cycles per second is assumed to be divided to 30 RBs as the first scenario and 90 RBs as the second scenario.

As Fig. 4 shows, the energy consumption can be improved specifically for higher number of users if the computing capacity is divided into smaller RBs, i.e., the scenarios with 120 and 90 RBs. This observation is due to the fact that the users have different amounts of workloads, and thus the computing capacity can be allocated more fairly to the users when it is divided to smaller blocks. Moreover, this improvement is more considerable for the case with the capacity of 120 Giga cycles per second since the difference between the two division scenarios is more pronounced. To understand the reason of the observation in Fig. 4, we should analyze the performance of the computing RBs allocation scheme in terms of the fairness. To this end, we have adopted the Jain’s fairness index [24] for the computing time,

$$\text{Fairness index} = \frac{\left(\sum_{u=1}^{U} Q_u\right)^2}{U \sum_{u=1}^{U} (Q_u)^2}$$  \hspace{1cm} (11)

Fig. 5 shows the Jain’s fairness index which is bounded between 0 and 1 for the aforementioned scenarios. As we can see in this figure, the scenarios with computing RBs with granularity of 1 Giga cycles per second are fairer as compared to those with granularity of 3 Giga cycles per second.

V. CONCLUSION

In this study, we have proposed an edge commuting aware NOMA technique, which can leverage the gains of uplink NOMA in reducing MEC users’ energy consumption. Specifically, we have formulated a NOMA based optimization framework that minimizes the energy consumption of MEC users via optimizing the user clustering, computing and communication resource allocation, and transmit powers. In particular, we have investigated the joint allocation of the frequency and computing RBs to the users that are assigned to different order indices within the NOMA clusters. We have also designed an efficient heuristic algorithm for user clustering and RBs allocation, and formulated a convex optimization problem for the power control to be solved independently per NOMA cluster. Moreover, we have evaluated and demonstrated the effectiveness of the proposed NOMA scheme in lowering the energy consumption via simulations.

REFERENCES

[1] J. Thompson, X. Ge, H. Wu, R. Irmer, H. J., G. Fettweis, and S. Alamouti, “5g wireless communication systems: prospects and challenges [guest editorial],” IEEE Communications Magazine, vol. 52, no. 2, pp. 62–64, 2014.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, “Non-orthogonal multiple access (noma) for cellular future radio access,” pp. 1–5, 2013.

[3] B. Kim, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, “Non-orthogonal multiple access in a downlink multiuser beamforming system,” pp. 1278–1283, 2013.
[4] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, “Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks,” IEEE Signal Processing Magazine, vol. 31, no. 6, pp. 45–55, 2014.

[5] Y. C. Hu et al., “Mobile edge computing: a key technology towards 5G,” ETSI White Paper, vol. 11, 2015.

[6] P. Di Lorenzo, S. Barbarossa, and S. Sardellitti, “Joint optimization of radio resources and code partitioning in mobile edge computing,” arXiv preprint arXiv:1307.3835v3, 2016.

[7] Y. Yu, J. Zhang, and K. B. Letaief, “Joint subcarrier and cpu time allocation for mobile edge computing,” pp. 1–6, 2016.

[8] M. Chiang and T. Zhang, “Fog and IoT: An overview of research opportunities,” IEEE Internet of Things Journal, vol. 3, no. 6, pp. 854–864, 2016.

[9] N. M. Gonzalez et al., “Fog computing: Data analytics and cloud distributed processing on the network edges,” pp. 1–9, 2016.

[10] A. Kiani and N. Ansari, “Towards hierarchical mobile edge computing: An auction-based profit maximization approach,” IEEE Internet of Things Journal, vol. 4, no. 6, pp. 2082–2091, 2017.

[11] ——, “Optimal code partitioning over time and hierarchical cloudlets,” IEEE Communications Letters, to be published, DOI: 10.1109/LCOMM.2017.2764904, 2017.

[12] X. Sun and N. Ansari, “EdgeIoT: Mobile edge computing for internet of things,” IEEE Comm. Mag., vol. 54, no. 12, pp. 22–29, Dec. 2016.

[13] R. Tandon and O. Simeone, “Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks,” IEEE Communications Magazine, vol. 54, no. 8, pp. 44–50, 2016.

[14] S.-B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, “Proportional fair frequency-domain packet scheduling for 3gpp lte uplink,” pp. 2611–2615, 2009.

[15] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, “Uplink non-orthogonal multiple access for 5G wireless networks,” pp. 781–785, 2014.

[16] N. Zhang, J. Wang, G. Kang, and Y. Liu, “Uplink nonorthogonal multiple access in 5G systems,” IEEE Communications Letters, vol. 20, no. 3, pp. 458–461, 2016.

[17] Y. Liang, X. Li, J. Zhang, and Z. Ding, “Non-orthogonal random access (nora) for 5G networks,” IEEE Transactions on Wireless Communications, vol. 16, no. 7, pp. 4817 – 4831, 2017.

[18] M. S. Ali, H. Tabassum, and E. Hossain, “Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems,” IEEE Access, vol. 4, pp. 6325–6343, 2016.

[19] H. Tabassum, E. Hossain, and M. J. Hossain, “Modeling and analysis of uplink non-orthogonal multiple access (noma) in large-scale cellular networks using poisson cluster processes,” IEEE Transactions on Communications, vol. 65, no. 8, pp. 3555 – 3570, 2017.

[20] X. Qiu and K. Chawla, “On the performance of adaptive modulation in cellular systems,” IEEE transactions on Communications, vol. 47, no. 6, pp. 884–895, 1999.

[21] D. Julian, M. Chiari, D. O’Neill, and S. Boyd, “Qos and fairness constrained convex optimization of resource allocation for wireless cellular and ad hoc networks,” vol. 2, pp. 477–486, 2002.

[22] S. S. Ram, V. V. Veeravalli, and A. Nedic, “Distributed non-autonomous power control through distributed convex optimization,” pp. 3001–3005, 2009.

[23] G. T. 25.996, “Spatial channel model for multiple input multiple output (mimo) simulations,” Release 11, Sept. 2012.

[24] R. Jain, D.-M. Chiu, and W. R. Hawe, A quantitative measure of fairness and discrimination for resource allocation in shared computer system. Eastern Research Laboratory, Digital Equipment Corporation Hudson, MA, 1984, vol. 38.