A NOMA-enhanced reconfigurable access scheme with device pairing for M2M networks

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- © 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Metadata Record: https://dspace.lboro.ac.uk/2134/37190

Version: Accepted for publication

Publisher: © IEEE

Please cite the published version.
A NOMA-Enhanced Reconfigurable Access Scheme with Device Pairing for M2M Networks

Atoosa Dalili Shoaei∗, Mahsa Derakhshani†, Tho Le-Ngoc∗
∗Department of Electrical & Computer Engineering, McGill University, Montreal, QC, Canada
† Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK
Email: atoosa.dalilishoaei@mail.mcgill.ca; m.derakhshani@lboro.ac.uk; tho.le-ngoc@mcgill.ca

Abstract—This paper aims to address the distinct requirements of machine-to-machine networks, particularly heterogeneity and massive transmissions. To this end, a reconfigurable medium access control (MAC) with the ability to choose a proper access scheme with the optimal configuration for devices based on the network status is proposed. In this scheme, in each frame, a separate time duration is allocated for each of the non-orthogonal multiple access (NOMA)-based, orthogonal multiple access (OMA)-based, and random access-based segments, where the length of each segment can be optimized. To solve this optimization problem, an iterative algorithm consisting of two sub-problems is proposed. The first sub-problem deals with selecting devices for the NOMA/OMA-based transmissions, while the second one optimizes the parameter of the random access scheme. To show the efficacy of the proposed scheme, the results are compared with the reconfigurable scheme which does not support NOMA. The results demonstrate that by using a proper device pairing scheme for the NOMA-based transmissions, the proposed reconfigurable scheme achieves better performance when NOMA is adopted.

I. INTRODUCTION

In the fifth generation (5G) wireless networks, three different types of services are to be supported, including enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) [1]. The key focus in eMBB is providing high data rates to a small number of human-type communications in the downlink, while mMTC and URLLC mainly target Internet of Things (IoT) applications, where variety of machine-type devices generate small data packets mostly transmitted in the uplink direction, towards a central controller, such as a base station (BS). The machine-type traffics are different in terms of delay and reliability requirements. The URLLC technology enables ultra reliable data transmissions between devices within low latency requirements for mission critical applications. On the other hand, mMTC facilitates the connectivity of a massive number of devices with relaxed latency requirements coexisting in one cell.

In this paper, we focus on uplink-centric mMTC services with short packet sizes, which pose major challenges on the network optimization and multiple access. To date, some potential candidates have been proposed to support massive connectivity, such as massive multiple-input-multiple-output (MIMO), millimeter wave communications, ultra dense networks, and non-orthogonal multiple access (NOMA). In this paper, we adopt NOMA, which is highly expected to increase the network throughput and accommodate massive connectivity.

NOMA allows multiple devices to transmit over the same resource simultaneously using power-domain or code-domain techniques, while in conventional orthogonal multiple access (OMA) schemes, radio resources are orthogonally assigned to devices to avoid or alleviate interference [1]. Accordingly, it is expected that by using NOMA, the network throughput significantly increases compared to using OMA schemes, since if each resource is simultaneously used by multiple devices, the total network efficiency becomes multiple fold too. However, in machine-to-machine (M2M) networks, also referred to as machine-type communication networks, since devices might have sporadic transmissions, even though a resource is shared among them, it may still left unused, leading to low spectrum utilization. In other words, the NOMA scheme is more useful for devices having periodic or frequent packets for transmissions compared to sporadic transmissions for which grant-free access schemes are more pertinent.

In this paper, to achieve the spectral efficiency optimality, we propose a NOMA-enhanced reconfigurable access scheme (NERA) which is able to switch between the grant-based NOMA-OMA and grant-free schemes, taking into account the device traffic statistics. In the grant-based segment of the proposed scheme, the fast uplink grant method proposed by 3GPP is used [2]. In this approach, devices do not send a scheduling request for uplink transmissions and the access point (AP) allocates resources to them in a proactive manner. Thus, this approach eliminates the signaling overhead required in the request-based scheduling schemes [3]. Furthermore, as device traffic statistics are considered in the scheduling procedure, the probability that the allocated resources left idle is minimized.

To obtain the optimal length of each segment, we formulate a network throughput maximization problem such that the device pairing in the NOMA segment can be optimized under minimum rate requirements of the devices. Due to the combinatorial nature of the formulated mixed integer non-linear programming (MINLP) problem, we propose a decomposition-based scheme which solves the problem into two steps. In the first step, to choose a set of paired devices for the NOMA...
segment, we formulate the problem by a weighted matching problem and propose a sub-optimal algorithm to solve it. Then, in the next step, for a given set of the NOMA pairs, we obtain the length of each segments such that the total network throughput is maximized. To that end, the second step is formulated by an optimization problem which is solved in an iterative manner. The performance of the proposed scheme is compared with a reconfigurable scheme which does not support NOMA. Considering different scenarios, the results show when using the NOMA-enhanced scheme is beneficial.

The rest of this paper is organized as follows. In Section II, we review the existing works on uplink NOMA transmissions. Section III presents the system model and the proposed reconfigurable access scheme. In Section IV, an optimization problem to maximize the network throughput using the proposed scheme is provided. Sections V and VI discuss the algorithms to solve the proposed problem. Finally, Section VII presents the simulation results and Section VIII concludes the work.

II. RELATED WORKS

In the research community, NOMA has received significant attentions, due to its potential capability to improve the networks capacity. In addition to its spectral efficiency gain, research studies have also shown that NOMA can accommodate a large number of devices, which is important for massive MTC networks [4].

Most of these research studies employ NOMA for downlink scenarios [5]–[9]. However, as in MTC networks, the traffic generated at uplink is heavier, the focus of this section is on the existing works that deploy NOMA for uplink scenarios.

In [10], to improve the network throughput, first, users are clustered using a sub-optimal algorithm and then for the given clusters, the optimal power allocation is derived. The work in [11] also employs the decomposition-based approach in which users are clustered using a graph-based algorithm, and then power and bandwidth are optimized. In [12], a power control algorithm for uplink NOMA is proposed, where the outage probability and the achievable sum rate of the proposed scheme are theoretically analyzed. The work in [13] proposes optimum and sub-optimum user pairing schemes for both single and multi-antenna BSs as well as for multi-antenna users. Moreover, an interference cancellation technique for asynchronous uplink NOMA systems is introduced in [14]. Furthermore, there are some studies targeted to maximize the energy efficiency of the MTC networks with NOMA [15], [16].

In [17], a distributed power control algorithm is designed for the uplink of a NOMA system consisting of two cells. It is assumed that each cell has one BS and two users, where the aim is to minimize the total power of the two users, while their rate requirements are satisfied. This problem is formulated by a two-player non-cooperative game, where the BSs correspond to the players. The properties of Nash equilibrium is investigated and a distributed algorithm is proposed to converge to it. The work in [18], considers a multi-cell network where to enhance the performance of cell edge users, fractional frequency reuse (FFR) is used. In the FFR, the entire frequency band is divided into cell-center bands set and cell-edge bands set. In order to avoid interference between neighboring cells, the cell-edge bands are separated among the three neighboring cells. Furthermore, the cell users are divided into cell-interior and cell-edge users based on their reference signal receiving power (RSRP) from the serving base station. On the cell-center bands, only cell-interior users are allowed to transmit using NOMA, while for the cell-edge bands, cell-interior users can be multiplexed with cell-edge users, where the large difference in channel gain between cell-interior users and cell-edge users can benefit NOMA. In order to schedule users for NOMA uplink, a proportional fairness-based scheme is proposed.

Different form the aforementioned studies which use grant-based NOMA, the works in [19]–[22] exploit NOMA for random access in multi-channel networks. More specifically, in [19] a set of power levels is defined, where each device with a packet for transmission randomly chooses a power level and a channel. The performance analysis of this scheme for a single channel network is investigated in [23]. The drawback of this scheme is that the set of power levels is predefined, thus it may lead to low performance for the network if the power levels are not suited for it. This issue is addressed in [21], where an adaptive set of power levels is used. In this work, to obtain the proper power level set, the number of active devices is estimated by the gateway and then based on that the set is determined. In a similar approach, in [20], the estimated number of active devices is broadcasted to the devices, where using this information, each device adjusts its transmission power properly. In [22], the cell area is divided into different layers, where devices in different layers have different target received power. More specifically, in this approach, each device having a packet for transmission, chooses its transmit power equal to its target received power divided by its channel gain.

There are also several works which deploy code-domain NOMA [24]–[27], where various compressive sensing (CS) techniques are done for multi-user detection (MUD). The main drawback of these works is that they need prior information about user activity, which results in high computational complexity on the receiver side. Moreover, these approaches are only suitable for the scenarios that user activity is time-related and sparse. In [28], each device having a packet to transmit, randomly selects a channel. It also chooses a random seed to construct a Raptor code to encode its packet and transmits the encoded packet over the selected channel. The BS is able to decode the packet, if the seed is not chosen by other devices transmitting over the same channel.

Although, the above works enhance the performance of considered networks, none of them address the unique requirements of MTC networks which consists of large number of devices with heterogeneous traffic rates. In these net-
works, using only grant-based NOMA schemes lead to poor performance for devices with sporadic transmissions while NOMA-based random access may not perform well due to large number of devices. In fact, the chance of choosing the same power and channel increases with increasing number of devices leading to large collision rate. Thus, in this work, we propose a reconfigurable scheme which to enhance the network throughput adaptively switches between a grant-based NOMA-OMA and contention-based grant-free scheme based on CSMA.

III. SYSTEM MODEL

We consider an MTC network consisting of one AP and $D$ devices. We assume that all devices transmit their packets with the same transmission power $\beta$. The channel power gain coefficient between device $d \in D = \{1, \cdots, D\}$ and the AP is $\gamma_d$ and the received power from device $d$ at the AP is $\beta \gamma_d + n_d$, where $n_d$ denotes the Gaussian noise power.

A. Overview of Uplink NOMA

In the uplink NOMA, multiple devices non-orthogonally transmit to the AP on the same radio resource. At the AP, these signals are received in a superimposed form, where they cause interference to each other. In order to decode these signals, the successive interference cancellation (SIC) technique can be used, where to ensure successful SIC, the received signal power of the devices should be distinctive [10]. As devices are different, each received signal at the AP experiences a distinct channel gain, which makes it possible to use NOMA, even if all devices transmit at the same power level.

In the SIC technique, the AP first decodes the signal of the device with the highest channel gain, since it likely is the strongest at the AP. To do that, the AP treats the signals from other devices as additive noise. After that, the decoded signal of the first device is subtracted from the received signal and the AP continues the decoding for the second device and treats the remainder as additive noise. This process is continued until all the devices are successfully decoded [29]. As a result, the highest channel gain device experiences interference from all devices and the lowest channel gain device effectively enjoys interference-free transmission if SIC is done without any errors.

In the case of the imperfect SIC, when the device’s signal is decoded, there is a difference between the actual and estimated signal. In other words, in the imperfect SIC, some portion of the received power of devices remains as interference which is called as residual interfering signal power and it is denoted by $I_d$. The magnitude of the SIC error is dependent on the type of SIC employed, the number of signals being canceled, and channel and device mobility conditions [30]. Thus, unlike the perfect SIC, where the signal of the lowest channel gain device is decoded with no interference from other devices, here it contains the residual interfering signal powers from devices with larger channel gains. To consider all sources of error, we define the expected level of cancellation achieved by SIC as $\sigma^2 = \mathbb{E}[|s_d - \hat{s}_d|]$, where $s_d$ and $\hat{s}_d$ are the received and estimated signals of device $d$ at the AP, respectively. Thus, we have

$$I_d = \sigma^2 \beta \gamma_d$$

(1)

B. NOMA-Enhanced Reconfigurable Access Scheme

The network operates on a frame-by-frame basis. Each frame is started with a beacon followed by three traffic segments: grant-based NOMA (GBN), grant-based OMA (GBO), and contention-based grant-free (CGF) as shown in Fig. 1. The beacon is transmitted to the devices by the AP for notifying them on the scheduling of the frame. The three traffic segments have a total of $N_f$ time-slots and each time-slot has a duration of $T_s$. Thus, the length of each frame denoted by $T_f$ is $N_f \times T_s$. The GBN and GBO have a total length (time) of less than $T_{\text{max}}$. Operations in the three traffic segments are described in the following paragraphs.

Beacon: At the start of each time frame, the AP broadcasts a beacon frame including the information about the grant-based NOMA/OMA and contention-based grant-free schemes.

Grant-based NOMA (GBN): In this segment, the granted devices transmit their packets in the corresponding allocated time-slots using NOMA. Here, we consider the case that only up to two devices can simultaneously transmit their packets in an allocated time-slot. In fact, as the received signal at the AP is superposition of at most two signals, the decoding complexity is acceptable.

Grant-based OMA (GBO): In this traffic segment, similar to the GBN, granted devices transmit their packets in the corresponding allocated time-slots. The only difference is that, each time-slot is only assigned to one device.

Contention-based Grant-Free (CGF): In this traffic segment, devices that have packets for transmission but are not granted a time-slot, contend with each other, based on the $p$-persistent CSMA protocol. In the $p$-persistent CSMA, the device $d$ listens to the channel and if the channel is idle, it transmits its packet with the contending probability denoted by $p_d$. Here, the optimized values of $p_d$ is announced by the AP in the beacon.

For convenience, the time duration of the time-slot $T_s$ is represented in backoff-time units.

2The frame structure can be applied to both single-carrier and multi-carrier systems.
C. Traffic Model

We assume that a packet is generated at device $d$ with probability of $a_d$ in each frame, and is added to the device’s queue if the queue length is less than $Q_{\text{max}}$. Otherwise, it is dropped. Moreover, it is assumed that the AP is aware of the packet arrival probabilities of the devices and it keeps vector $\mathbf{V}(t) = [v_d(t)]_{d \in \mathcal{D}}$, where $v_d$ denotes the last time that the AP has received a packet from the device $d$. We also consider that at each time the device sends a packet, it piggybacks an extra bit, $q_d(t)$, reporting whether its queue is empty ($q_d(t) = 0$) or non-empty ($q_d(t) = 1$, i.e., it has packets backlogged in the queue to transmit). Thus, at each frame $t$, the AP updates $\theta_d(t)$ which is the probability of the device $d$ having a non-empty queue at $t$

$$
\theta_d(t) = \begin{cases} 
1 - (1 - a_d)^{t-v_d(t)}, & \text{if } q_d(v_d(t)) = 0 \\
1 & \text{if } q_d(v_d(t)) = 1.
\end{cases}
$$

IV. Problem Formulation

NOMA can provide higher throughput for MTC networks as it can serve multiple devices using the same radio resource. However, in MTC networks, devices may transmit in a sporadic manner, where grant-free schemes may have better performance. This is because the allocated resources might be left idle due to their low rate of transmissions. Thus, deploying a combination of NOMA and grant-free scheme in each frame can be beneficial if devices with a higher probability of transmission are scheduled to transmit in the grant-based segments and the rest for the grant-free segment. In fact, using this scheme, the network throughput depends on which devices are selected for the NOMA segment and how these devices are paired. In the following, we derive the throughput of CGF segment and then present the problem formulation for the scheduling of the proposed scheme.

A. CGF Analytical Throughput

We consider the achievable throughput of $D$ devices using p-persistent CSMA, where device $d$ attempts for transmission with probability $p_d$ once it has a packet in its queue. Each device transmits at most one packet during the CGF segment. As an approximation, we assume the number of active devices is constant over a frame. Let $P_{\text{idle}}$ be the probability that the channel is idle in a backoff unit. We can obtain this probability as

$$
P_{\text{idle}} = \prod_{d \in \mathcal{D}} (1 - \theta_d p_d),
$$

where $\theta_d p_d$ is the transmission probability of device $d$.

As device $d$ transmits a packet successfully if there is no other transmissions on the channel, i.e., the probability of successful transmission of device $d$ is

$$
P_{\text{succ},d} = \theta_d p_d \prod_{d' \in \mathcal{D}, d' \neq d} (1 - \theta_d p_d).
$$

According to [31], we can derive the normalized throughput of device $d$ denoted by $\rho_d$, which is defined as the fraction of time that the channel is used for its successful transmission

$$
\rho_d = \frac{P_{\text{succ},d} T_s}{P_{\text{idle}} + (1 - P_{\text{idle}})T_s},
$$

where $\phi$ is the duration of a backoff unit and $T_s$ is the duration of a successful transmission, which includes the data transmission for a fixed time, inter-frame spaces, and signaling overheads. Since signaling and inter-frame spaces are relatively small (in the order of $\mu$s) compared with the data transmission length (in the order of ms), we approximately assume that both collided and successful transmissions are of the same size (i.e., $T_s$). Consequently, the denominator in (5) represents the expected length of a general time-slot.

Finally, to simplify (5), we define a new variable $y_d$ as [32]

$$
y_d = \frac{\theta_d p_d}{1 - \theta_d p_d}.
$$

Having defined $y_d$, we can rewrite $P_{\text{idle}}$ and $P_{\text{succ},d}$ in terms of $y_d$ as

$$
P_{\text{idle}} = \prod_{d \in \mathcal{D}} \frac{1}{(1 + y_d)},
$$

$$
P_{\text{succ},d} = \prod_{d \in \mathcal{D}} \frac{y_d}{(1 + y_d)} = y_d P_{\text{idle}}.
$$

Therefore, the throughput of device $d$ in the CGF segment becomes

$$
\rho_d = \prod_{d \in \mathcal{D}} \frac{y_d}{(1 + y_d)} - y_d',
$$

where $y' = \frac{T_s - \phi}{T_s}$.

B. Optimization Problem

To schedule devices based on the proposed reconfigurable scheme, we formulate an optimization problem aiming to maximize the network throughput. Let define $x_{d,d'}$ as a binary variable indicating whether devices $d$ and $d'$ are paired together to be assigned a time-slot in the NOMA segment. In NOMA, devices are said to be in a pair, if they simultaneously use the same time-slot. In fact, the efficiency of the NOMA scheme depends on how devices are paired. With a constraint that the rate of device $d$ is sufficiently large to transmit one packet during one time-slot, the achievable throughput for the NOMA segment can be calculated as $\sum_{d \in \mathcal{D}} \sum_{d' \in \mathcal{D}} \theta_d x_{d,d'}$. For the CGF segment, based on (9), the total throughput is equal to $\sum_{d \in \mathcal{D}} \rho_d T_{\text{CGF}}$ where $T_{\text{CGF}}$ is $T_f - 0.5T_s \sum_{d \in \mathcal{D}} \sum_{d' \in \mathcal{D}, d' \neq d} x_{d,d'} +$
Therefore, the optimization problem becomes
\[
\max_{X,Y} \sum_{d \in D} \left( \sum_{d' \in D} \theta_{d} x_{d,d'} + \rho_{d} T_{CGF} \right)
\]
subject to:
\[
\begin{align*}
\text{C10.1: } & \log_{2} \left( 1 + \frac{\beta \gamma_{d}}{\gamma_{d} n_{d}} \right) \geq R x_{d,d'}, \forall d \in D, \forall d' < d \\
\text{C10.2: } & \log_{2} \left( 1 + \frac{\beta \gamma_{d}}{\sigma^{2} \beta \gamma_{d} + n_{d}} \right) \geq R x_{d,d'}, \forall d \in D, \forall d' > d \\
\text{C10.3: } & \log_{2} (1 + \frac{\beta \gamma_{d}}{n_{d}}) \geq R x_{d,d}, \forall d \in D, \\
\text{C10.4: } & y_{d} \sum_{d' \in D} x_{d,d'} = 0, \quad \forall d \in D, \\
\text{C10.5: } & 0.5 T_{s} \sum_{d \in D} \sum_{d' \in D} x_{d,d'} + T_{s} \sum_{d \in D} x_{d,d} \leq T_{\max}, \\
\text{C10.6: } & \frac{y_{d}}{\theta_{d}(1 + y_{d})} \leq 1, \quad \forall d \in D, \\
\text{C10.7: } & x_{d,d'} \in \{0, 1\}, \quad \forall d \in D, \forall d' \in D.
\end{align*}
\]

In C10.1, \( R \) is the required rate for transmitting a packet in one time-slot. In the SIC, the AP first decodes the signal from the device with the higher channel gain. Consequently, the higher channel gain device experiences interference from its paired device. Assuming that devices are sorted in ascending order with respect to their channel gains, constraint C10.1 ensures that the rate of device \( d \) with higher channel gain is sufficiently large to transmit one packet during one time-slot.

Constraint C10.2 guarantees the minimum rate requirement of the paired device with the weaker signal, which in the case of imperfect SIC suffers from residual interfering signal power. Constraint C10.3 ensures that the required rate of the GBO device is met. Constraint C10.4 indicates that device \( d \) is either selected for GBN, GBO or CGF segment. Constraint C10.5 guarantees that the duration of grant-based NOMA/OMA access is less than \( T_{\max} \). Constraint C10.6 indicates that the \( \rho_{d} \) of each device for the CSMA scheme should be less than 1. Finally, constraint C10.7 defines \( x_{d,d'} \) as a binary variable.

The problem in (10) is a mixed-integer optimization problem due to the binary constraint of C10.7. Furthermore, it is non-convex, as the objective function and constraint C10.7 are non-convex. Consequently, it is a non-convex mixed-integer optimization problem. Generally, there is no computational efficient approach to solve this class of optimization problems. However, the problem can be decomposed into two sub-problems; 1) device pairing sub-problem, 2) NERA optimization sub-problem. The device pairing sub-problem deals with how to partition devices into the groups while the solution of NERA optimization sub-problem determines the optimal number of NOMA groups and \( \rho_{d} \) for devices in the CGF segment. In the following, we describe how to obtain an efficient and tractable solution for each of these sub-problems.

V. DEVICE PAIRING FOR NOMA

In this section, we discuss the proposed device pairing scheme. In NOMA, to be able to perform SIC, devices should be paired such that the difference between their received powers at the AP is sufficiently large. Otherwise, if the difference in the received powers is small, the signals will not be successfully decoded at the AP. The device pairing sub-problem is inherently combinatorial, which requires an exhaustive search to obtain the optimal solution. Here, we propose a simple yet high performance algorithm to solve this sub-problem. More specifically, we model it as a weighted graph matching problem, which is shortly explained in the following subsection.

A. Introduction to Weighted Matching Problem

Let consider a bipartite graph \( G = (V, E) \), where \( V \) and \( E \) denote the set of vertices and edges respectively and \( w \) is a function which assigns a weight to the edges of \( G \), i.e., \( w : E \to \mathbb{R} \). A matching \( M \) is a sub-graph of \( G \) with edges of \( F \subseteq E \), where no two edges in \( F \) share a common vertex. The weight of this matching is the sum of the weights in \( M \), i.e., \( w(M) = \sum_{e \in M} w(e) \). The maximum weighted matching is the matching of a graph \( G \) with the largest value of \( w(M) \).

In [33], Edmonds proposed an exact algorithm with running time of \( O(|V|^4) \), which is not affordable computationally. The fastest algorithm for solving this problem has a running time of \( O(|V||E| + |V|^2 \log |V|) \) [34], which is still costly for large graphs. Therefore, approximation algorithms have been proposed which are faster than the exact algorithms. Furthermore, the other drawback of Edmonds’ algorithm is that it is non-intuitive and consequently difficult to understand and implement [35]. In the following, we discuss an approach which is simple but yet efficient.

The device pairing sub-problem can be modeled as a weighted matching problem in which the vertices denote devices and an edge between two vertices \( d \) and \( d' \) indicates that the simultaneous transmission of those devices over the same time-slot does not violate the SIC constraint, i.e., the signals of devices \( d \) and \( d' \) can be successfully decoded at the AP. Finally, the weight of each edge is the sum of the expected throughput of its endpoints, i.e., for edge \( e \) connecting devices \( d \) and \( d' \), the weight is
\[
w(e_{d,d'}) = \theta_{d} + \theta_{d'}. \]

B. Proposed Device Pairing Algorithm

Here, we propose a heuristic approach to solve the device pairing sub-problem, which nominates maximum \( T_{\max} \) edges of the graph iteratively for the NOMA transmissions. The algorithm works as follows. The algorithm iterates \( T_{\max} \) times, where at each iteration, the edge with the largest weight is chosen. The devices corresponding to the vertices of that edge are added as a pair to the NOMA-pairing set, while they are eliminated from the graph. The algorithm terminates once either \( T_{\max} \) pairs are nominated or no more edges with non-zero weights is remained. The result of the algorithm is
Algorithm 1 Device Pairing for NOMA

Input: $G(V, E), T_{\text{max}}$
$K \leftarrow [0]_D \times D$
for $c = 1 : T_{\text{max}}$ do
  $e_{i,j} \leftarrow$ Choose the edge of $E$ with the largest weight
  $K_{i,j} \leftarrow e_{i,j}$
end for
Output: $K$

presented by the matrix $K$ with dimension of $D \times D$, where $k_{d,d'}$ is set to the weight of $e_{d,d'}$ if the pair is in the NOMA-pairing set, otherwise it is set to 0. Note that the reason to choose maximum $T_{\text{max}}$ edges is that the length of the grant-based NOMA segment is restricted to $T_{\text{max}}$.

VI. SCHEDULING ALGORITHM FOR NERA SCHEME

In this section, given the device pairing graph obtained from Algorithm 1, we obtain the optimal selection of NOMA pairs along with $p_d$ values for CGF devices.

To solve the optimization problem, first we simplify the CGF throughput by using approximations. Assuming that the network consists of a large number of devices, we have

$$\rho_d = \frac{y_d}{\prod_{d' \in D} (1 + y_{d'}) - t'} \approx \frac{y_d}{\prod_{d' \in D} (1 + y_{d'})} \approx y_d (1 - \sum_{d' \in D} y_{d'}).$$

(11)

Substituting the above approximation into (10), we reach to the following optimization problem.

$$\max_{X, Y} \sum_{d \in D} \left[ \sum_{d' \in D} \theta_d f_{d,d'} x_{d,d'} + y_d (1 - \sum_{d' \in D} y_{d'}) (T_f - T_s \sum_{d' \in D} \sum_{d'' \in D} f_{d,d'} x_{d''}) \right],$$

subject to: C10.4 – C10.7

The optimization problem in (12) is still non-convex. However, it can be solved by using an iterative algorithm, in which at each iteration the problem is decomposed into two sub-problems: 1) GBN/GBO scheduling, 2) $p$-probability derivation. More specifically, the whole algorithm consists of variable initialization followed by the iterative phase. In the iterative phase, first, for fixed values of $Y$, optimal $X$ are derived and then in the next step, for the given $X$, optimal $Y$ are obtained. The iterations continue until the results converge, which mathematically can be stated as

$$\|X^*(t) - X^*(t - 1)\| \leq \varepsilon_1,$$

and

$$\|Y^*(t) - Y^*(t - 1)\| \leq \varepsilon_2.$$

The procedure of finding optimal $X$ and $Y$ by using the proposed iterative algorithm can be expressed as

$$X(0) \rightarrow Y(0) \rightarrow \ldots X^*(t) \rightarrow Y^*(t) \rightarrow X^* \rightarrow Y^*,$$

Initialization Iteration $t$ Optimal solution

In the following, we discuss how to solve each of these sub-problems.

A. GBN/GBO Scheduling Sub-problem

This sub-problem takes the device pairing set obtained from Algorithm 1 and determines the selected pairs for the grant-based NOMA or single devices chosen for the grant-based OMA scheme. In the following, we discuss how the results of Algorithm 1 are represented in the GBN/GBO scheduling formulation.

Consider the matrix $F_{D \times D}$ with each element $f_{d,d'}$ defined as

$$f_{d,d'} = \begin{cases} 0.5 & \text{if } k_{d,d'} > 0 \& d \neq d', \\ 1 & \text{if } k_{d,d'} > 0 \& d = d', \\ 0 & \text{otherwise} \end{cases}$$

(13)

The GBN/GBO scheduling sub-problem is expressed as

$$\max_X \sum_{d \in D} \left[ \sum_{d' \in D} \theta_d f_{d,d'} x_{d,d'} + y_d (1 - \sum_{d' \in D} y_{d'}) (T_f - T_s \sum_{d'' \in D} \sum_{d' \in D} f_{d,d''} x_{d''}) \right],$$

subject to:

C14.1: $T_s \sum_{d'' \in D} \sum_{d'' \in D} f_{d,d''} x_{d,d''} \leq T_{\text{max}},$

C14.2: $x_{d,d'} = x_{d',d}, \{\forall (d, d') | f_{d,d'} > 0\}$

C14.3: $\sum_{d'' \in D} x_{d,d''} \leq 1, \forall d \in D$

C14.4: $x_{d,d'} \geq 0, \forall d \in D, \forall d' \in D.$

The above optimization problem is linear and can be solved by existing techniques.

B. $p$-Probability Derivation Sub-problem

For the fixed $X$, the probability derivation sub-problem is defined as

$$\max_Y \sum_{d \in D} y_d (1 - \sum_{d' \in D} y_{d'}),$$

subject to:

C15.1: $(1 + \theta_d) y_d \leq \theta_d, \forall d \in D,$

C15.2: $y_d \sum_{d'' \in D} x_{d,d''} = 0, \forall d \in D.$

Notably, in the above optimization problem, the terms containing $X$ are omitted in the objective function, while the resulting optimization is equivalent to the original problem.

The optimization problem in (15) is convex and hence the optimal solution can be derived.
C. Reconfigurable Access Scheme

In order to show how NERA can enhance the network performance leveraging NOMA, we compare its performance with the reconfigurable access (RA) scheme, which is similar to the proposed NERA scheme, however it restricts the number of simultaneous transmissions in each slot to one. More specifically, in this scheme, at each frame, the following optimization problem should be solved.

\[
\max_{X,Y} \sum_{d \in D} \left[ \theta_d x_d + \rho_d (T_f - T_s \sum_{d \in D} x_d) \right]
\]

subject to:

C16.1: \( x_d y_d = 0, \quad \forall d \in D \),

C16.2: \( T_s \sum_{d \in D} x_d \leq T_{\max} \),

C16.3: \( \frac{\theta_d (1 + y_d)}{\rho_d} \leq 1, \quad \forall d \in D \),

C16.4: \( x_d \in \{0, 1\}, \quad \forall d \in D \).

In order to solve this problem, in a similar approach to the one for solving problem (10), we first apply the approximation (11) for \( \rho_d \), and then we divide the problem into two sub-problems. Specifically, in this scheme, at each frame, the following optimization problem is written as

\[
\max_X \sum_{d \in D} \theta_d x_d + y_d (1 - \sum_{d \in D} y_d) (T_f - \sum_{d \in D} x_d),
\]

subject to:

C17.1: \( \log_2 \left( 1 + \frac{\theta_d y_d}{n_d} \right) \geq R x_d \), \( \forall d \in D \),

C17.2: \( T_s \sum_{d \in D} x_d \leq T_{\max} \),

C17.3: \( x_d \leq 1, \forall d \in D \).

In the second sub-problem, we obtain \( Y \), for fixed \( X \).

\[
\max_Y \sum_{d \in D} y_d (1 - \sum_{d \in D} y_d),
\]

subject to:

C18.1: \( (1 + \theta_d) y_d \leq \theta_d, \quad \forall d \in D \),

C18.2: \( x_d y_d = 0, \forall d \in D \).

These sub-problems are solved sequentially over multiple rounds until the results of two last rounds converge.

D. Computational Complexity

In this sub-section, we analyze the computational complexity of the proposed algorithm. As the proposed algorithm is iterative, its computational complexity is the product of the number of iterations and complexity of each iteration.

Let first discuss the complexity of each iteration. In each iteration, two sub-problems should be solved. Thus, the computational complexity of each iteration that of these two sub-problems. The GBN/GBO scheduling sub-problem is a linear problem with \( D^2 + D \) variables and \( D^2 + D + T_{\max} + 1 \) constraints, which can be efficiently solved.

The second sub-problem for \( p \)-probability derivation is a convex optimization problem with \( D \) variables and \( 2D \) constraints, which can be solved in an efficient manner due to the low number of constraints.

Note that the number of iterations cannot be analytically studied. However, simulation results can be performed to show its order of computational complexity.

VII. Performance Evaluation

The results are obtained in Matlab environment and CVX is used for solving the optimization problems. We assume the network consists of 300 devices, each with a packet arrival probability randomly chosen from the uniform distribution over \((0, 1)\). Devices are randomly located (with uniform distribution) in the coverage area of the AP. Furthermore, we set \( N_f = 100 \), \( T_{\max} = 80 \) time-slots, and \( T_s = 12 \) backoff units. We also consider \( \sigma^2 = 0 \) and \( Q_{\max} = 4 \). The results are obtained for 5 runs, each run consists of 100 frames.

To show the gain of the proposed scheme, we compare the results of the RA scheme, with the same parameter values used in the NERA scheme. As a performance metric, we use normalized throughput. Here, the normalized throughput is defined as the average number of successfully transmitted packets per frame divided by the maximum throughput that can be achieved in an OMA scheme. Note that in OMA, the normalized throughput is not greater than one; however, as NOMA allows multiple devices to simultaneously transmit over the same time-slot, the normalized throughput can exceed one for NERA.

Fig. 2 shows the normalized throughput versus different values of \( R \) for \( \beta = 15 \) and 23 dBm. As seen in this figure, by increasing \( R \) the performance of the NERA scheme decreases. The reason is that less number of devices can be paired for NOMA transmissions when \( R \) is larger. Furthermore, as seen for the lower values of \( R \), the same normalized throughput is achieved for both \( \beta = 15 \) and 23 dBm. However,
after some point, the performance degradation for $\beta = 15$ dBm is higher. The reason is that, satisfying a larger rate requirement needs higher transmission powers. Therefore, with $\beta = 23$ dBm, higher throughput can be achieved. Moreover, the proposed NERA scheme outperforms the RA scheme, and the performance gap is larger for lower rates.

Furthermore, in Figs. 3 and 4, we plot the average number of devices transmitting in the GBN, GBO and CGF. By increasing $R$, the number of devices transmitting in NOMA decreases, while the number of devices transmitting in GBO and CGF increases. In fact, these plots explain why by increasing $R$, the performance gap between NERA and RA scheme decreases.

Fig. 5 shows the normalized throughput versus different values of $\beta$ for two values of $R = 3$ and 5 bits/s/Hz. Evidently, NERA outperforms the RA scheme when $R = 3$, since a larger number of devices can be paired. However, for $R = 5$, the performance of the NERA is the same or close to the RA scheme for lower powers, but it improves by increasing the power. The reason is that when the required rate is large, the pairing is not feasible.

Fig. 6 illustrates the results for $\beta = 15$ dBm for $R = 3$ and 5. As observed, by increasing $D$ the performance gap between NERA and RA increases for both values of $R$. The reason is that more traffic is generated in the network, therefore devices with larger expected throughput are paired together. Moreover, the RA throughput remains unchanged for $D \geq 200$, as GBO throughput reaches to its capacity and CGF throughput remains the same by controlling $p$ values of $p$-persistent CSMA.

Fig. 7 demonstrates the effect of SIC error on the NERA performance. The results are obtained for $\sigma^2 \in \{0, 0.05, 0.1\}$ and $R = \{1, 2, 3\}$. As observed, for $R = 1$, the presence
of SIC error for all values of $\sigma^2$ does not affect the NERA performance. For $R = 2$, and low $\beta$, increasing $\sigma^2$ causes degradation on NERA throughput. For $R = 3$, in the presence of SIC error for $\sigma^2 = 0.05$ or 0.1, pairing is impossible, resulting in the similar performance as RA. Thus, for this setting, in the case of imperfect SIC, using NOMA is only beneficial if $R$ is low. Otherwise, no additional gain can be obtained.

Finally, Fig. 8 illustrates the average number of iterations that takes for NERA and RA schemes to converge for the same setting of Fig. 6. As seen, both algorithms converge after a few iterations indicating a low computational complexity. Furthermore, for NERA, the number of iterations is larger when $R = 5$ compared to the case of $R = 3$.

VIII. CONCLUSIONS

In this paper, we propose a NOMA-enhanced reconfigurable access scheme to increase network throughput, and support heterogeneity and massive connectivity in machine-to-machine networks. The proposed scheme, at each frame, adopts three different access schemes including grant-based NOMA, grant-based OMA and contention-based grant-free according to the network condition. An optimization problem is formulated to allocate devices to the proper regime. Furthermore, simulation results are obtained to compare the performance of the proposed scheme with a reconfigurable scheme which does not support NOMA.

REFERENCES


Mahsa Derakhshani (S10–M13) received the Ph.D. degree in electrical engineering from McGill University, Montreal, Canada, in 2013. She was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada, a Research Assistant with the Department of Electrical and Computer Engineering, McGill University, from 2013 to 2015, and an Honorary NSERC Post-Doctoral Fellow with the Department of Electrical and Electronic Engineering, Imperial College London, from 2015 to 2016. She is currently a Lecturer (Assistant Professor) in digital communications with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K. Her research interests include non-orthogonal multiple access (NOMA), ultra-reliable low-latency communications (URLLC), software-defined wireless networking, and the applications of optimization and machine learning for radio resource management. She has received several awards and fellowships, including the John Bonsall Porter Prize, the McGill Engineering Doctoral Award, the Fonds de Recherche du Quebec Nature et Technologies (FRQNT), and the Natural Sciences and Engineering Research Council of Canada (NSERC) Post-Doctoral Fellowships. She currently serves as an Associate Editor for IEEE Signal Processing Journal.

Tho Le-Ngoc (F97) obtained his B.Eng. (with Distinction) in Electrical Engineering in 1976, his M.Eng. in Microprocessor Applications in 1978 from McGill University, Montreal, and his Ph.D. in Digital Communications 1983 from the University of Ottawa, Canada. From 1977 to 1982, he was with Spar Aerospace Ltd., Sainte-Anne-de-Bellevue, QC, Canada, involved in the development and design of satellite communications systems. From 1982 to 1985, he was with SRITelecom Inc., Saint-Laurent, QC, Canada, where he developed the new point-to-multipoint DA-TDMA/TDM Subscriber Radio System SR500. From 1985 to 2000, he was a Professor with the Department of Electrical and Computer Engineering, Concordia University, Montréal. Since 2000, he has been with the Department of Electrical and Computer Engineering, McGill University. His research interest is in the area of broadband digital communications. He is a fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada. He is a recipient of the 2004 Canadian Award in Telecommunications Research, and the IEEE Canada Fessenden Award 2005. He holds a Canada Research Chair (Tier I) on Broadband Access Communications.

Atoosa Dalili Shoaei received her B.Sc. degree in information technology engineering from Isfahan University of Technology, Isfahan, Iran, in 2009 and her M.Sc. degree in information technology engineering from Amirkabir University of Technology, Tehran, Iran, in 2012. Since 2014, she has been with the Department of Electrical and Computer Engineering, McGill University, Canada, where she is currently working toward her Ph.D. degree. Her main areas of research interest are media access control techniques, Internet of Things (IoT) and wireless virtualization.