Received March 27, 2021, accepted April 15, 2021, date of publication April 28, 2021, date of current version May 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076153

A Reconfigurable Access Scheme for Massive-MIMO MTC Networks

ATOOSA DALILI SHOAEI, DUC TUONG NGUYEN, AND THO LE-NGOC, (Life Fellow, IEEE)
Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada

Corresponding author: Duc Tuong Nguyen (tuong.nguyen2@mail.mcgill.ca)

This work was supported in part by the Huawei Technologies Canada, and in part by the Natural Sciences and Engineering Research Council of Canada.

ABSTRACT This paper presents an efficient reconfigurable access scheme for massive machine-type communication networks, where to provide massive connectivity, the base station is equipped with a large-scale antenna array. In particular, to maximize the expected throughput of the network, the scheme uses a frame divided into two segments: grant-based and grant-free that are more efficient for devices with high and low expected throughput, respectively. At the beginning of each frame, the base station decides how to partition the two segments, the resource allocation in the grant-based segment and the access probabilities in the grant-free segment to maximize the expected throughput based on the device traffic profiles. The corresponding optimization problem is formulated, and a sub-optimal solution algorithm with low computational complexity is proposed. The performance of the proposed access scheme is evaluated under different conditions to demonstrate its advantages in terms of achieved throughput and average packet delay.

INDEX TERMS Wireless communication, access control, massive MIMO, optimization, access protocols.

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

Massive machine-type communication (mMTC) has been identified as one of the three key applications of 5G networks. The goal of mMTC services is to provide connection to large numbers of devices that transmit data between themselves or to a base station (BS) with little human intervention. mMTC is the enabling technology of internet of things (IoT) applications, such as smart homes, smart cities and industrial internet of things (IIoT).

Compared to human-type communication, mMTC has several unique features posing new challenges to the wireless networks. Devices in mMTC may transmit their packets in a frequent or sporadic manner [1], and their data packets are usually short. This necessitates an access scheme that can support heterogeneous traffic at low signaling overhead. In addition, while wireless resources are limited, mMTC networks have to support large numbers of devices, which makes the massive access a problem in mMTC [2]. A promising technology to support massive connectivity and alleviate the wireless resource insufficiency is massive multiple-input multiple-output (MIMO). Thanks to the large number of base station (BS) antennas in massive MIMO, the channel vectors between the BS and devices become asymptotically orthogonal and linear processing can be utilized to separate devices using the same time-frequency resources effectively. Thus, massive MIMO has the potential to overcome the resource scarcity when a large number of machine-type devices access the network.

In massive MIMO, it is vital for the BS to have an accurate estimation of device channels, which is typically accomplished by uplink training with preambles sent by devices. If the number of devices associated to a BS is small, each preamble can be assigned to a device and no intra-cell pilot collision happens [3]. However, mMTC networks usually contain a huge number of devices associated to a BS whereas the preamble length and the number of unique preambles are limited due to the channel coherence time [4]. Thus, how devices accessing these preambles becomes a problem. In grant-based random access schemes such as in [3], [5]–[10], devices have to compete for preambles and collisions are resolved by different methods. The authors of [5], [6], [11] propose the strongest-user collision resolution (SUCRe) protocol, which allows devices to randomly access the preambles and when collision happens, only the one with the strongest signal can transmit. This protocol can resolve collisions, however, it can cause problems if devices with weak signals are delay-sensitive. The authors of [3] propose an improved version of SUCRe,

The associate editor coordinating the review of this manuscript and approving it for publication was Zesong Fei.

which allows devices with weaker signals to contend for idle preambles. However, it might require broadcasting idle preambles many times and its effectiveness might not be high when there are many devices since there will be a few idle preambles. Other improved versions of SUCRe such as in [9], [10] also have the problem of transmitting the preambles multiple times. In [7], the BS allows weak-signal devices to contend for idle preambles and also performs successive-interference-cancellation (SIC) to estimate the channel response of each device. This approach can reduce the times of broadcasting pilot signals but might cause high error in channel estimation. In general, due to the time spent on the handshake procedure, grant-based random access might cause low channel utility when packets are short.

Grant-free access schemes are also given attention thanks to its potential to reduce the signaling overhead in mMTC by allowing devices to transmit immediately without the complicated handshake procedure [12]. One approach of grant-free access is assigning each device a preamble, which is not necessarily orthogonal to other preambles and using compressive sensing techniques to detect devices activities. In each time slot, the BS tries to detect which devices have transmitted and estimates their channels based on the received metadata, which is the transmitted preambles, and then decodes the received data based on the channel estimation. In [13], [14], approximate-message-passing (AMP) algorithm is utilized to estimate the channel matrix of all devices under the assumption that this matrix is sparse because only a small portion of devices are active every time slot. Despite the use of non-orthogonal preambles, the AMP-based approach can detect devices activities without error when the number of antennas in the BS approaches infinity [15], the channel estimation error is higher due to the use of non-orthogonal preambles and the preamble length should be longer to compensate for non-orthogonality [16]. In [17] and [18], the device activity and data detection problem is formulated as a maximum likelihood estimation problem based on the covariance matrix of the received signal, which has lower probability of error than AMP-based methods. Other approaches based on compressive sensing techniques are proposed in [19], [20]. However, false alarm and missed detection might still happen and potentially cause unwanted errors. A grant-free random access scheme using orthogonal preambles is proposed in [21]. In this access scheme, each active device randomly chooses a preamble from available ones and transmits it along with its data. This approach enables high-accuracy estimation of device channels, however, when the number of contending devices is high, collisions will become a problem. In [22], [23], each device divides its data into several codewords and uses pilot-hopping to transmit each codeword and in [24], each device transmits several preambles before transmitting the data. These approaches are promising to reduce preamble collisions, however, the signaling overhead is increased by multiple times of preamble transmission.

There has been a research interest for hybrid access schemes [25], [26], combining grant-based and random access schemes to alleviate the number of collisions in the random access schemes and consequently improve the network throughput. In [27], [28], each device wishing to transmit has to contend for channel access by sending a request message, and only successful ones are allocated time slots while the BS utilizes a distributed queuing algorithms to resolve collisions. This approach enables data transmission without collisions, however, it requires a portion of each time frame for the request signal transmission rather than data transmission. In [29], each time frame is divided into a deterministic access phase, in which time slots are allocated to several devices by the BS, and a random access phase, in which the rest of devices contend for transmission by using p-persistent carrier-sense multiple access (CSMA). An optimization problem to maximize the expected throughput is solved by the BS to decide which devices should be in the deterministic access phase in the next time frame. This access scheme does not require any time slots spent on sending request message and can allocate devices with high traffic to deterministic access to reduce collisions. However, only one device can transmit in one time slot, which might be unsuitable in future networks, where massive MIMO is adopted at the BS. Therefore, a hybrid access scheme in the scenario of massive MIMO BS to leverage the capacity of massive MIMO is necessary, and this paper aims to develop such an access scheme.

B. CONTRIBUTIONS

The contributions of this paper are two-fold. First, aiming to provide massive connectivity, we propose a reconfigurable access scheme for mMTC networks with a massive MIMO BS, allowing a single radio resource to be used by multiple devices simultaneously. As these networks have large numbers of devices, assigning orthogonal preamble sequences to all devices would be impractical. However, because of the sporadic feature of mMTC traffic, at each frame, only a portion of devices have a packet for transmission [30]. Hence, we propose a hybrid access scheme in which each time frame is divided into two segments: grant-based segment and grant-free segment. In the grant-based segment, preambles and radio resources are assigned to devices with high probabilities of having non-empty queues while the rest of devices with low chance of having non-empty queues can contend with each other in the grant-free segments. The proposed access scheme does not use the handshake procedure of the grant-based schemes to reduce signaling overhead. In addition, it reduces the problem of severe collisions of grant-free schemes and improve throughput by proactively allocating resources to devices having high traffic demand to help them transmit without collisions and to reduce the traffic load on the grant-free segment.

Second, we formulate an optimization problem to schedule devices in the grant-based segment and calculate the access probabilities for devices in the grant-free segment to maximize the expected throughput based on the device traffic profiles. This problem is high-dimensional and mixed-integer and has decision variables in the sum bounds. Therefore, traditional methods cannot be directly applied. To solve this problem, we propose an iterative algorithm that alternatively solves the grant-based scheduling sub-problem and grant-free access probability sub-problem.

We carry out simulations to compare the proposed access scheme with a pure grant-free access scheme with optimized access probabilities in terms of average throughput and packet delay. Illustrative results show that our proposed access scheme achieves significantly higher throughput and lower packet delay as compared to the pure grant-free random access scheme.

C. STRUCTURE

The remaining of this paper is divided into five sections. Section II introduces the system model considered in this paper, the detail of our reconfigurable access scheme, the traffic model and the channel model. An overview of zero-forcing and conjugate beamforming is also provided in this section. Section III presents the problem formulation and the proposed algorithm to solve the problem along with its complexity is described in Section III. Section V presents the simulation results and Section VI concludes the paper.

Table 1 summarizes the key notations used in this paper.

II. SYSTEM MODEL

We consider a multipoint-to-point MTC network consisting of one BS and D devices as illustrated in Figure 1. Devices can be distributed arbitrarily within a coverage area. As illustrated in Figure 1, we consider only uplink data transmission from devices to the BS, as most applications involve reporting information to the BS in MTC networks. The BS is equipped with M antennas while each device has only one antenna since devices in MTC networks are usually size- and power-limited.

In order to estimate the channel state information (CSI) at the BS, each device wishing to transmit its data has to transmit a preamble beforehand. There are $N_{\rm p}$ orthogonal preambles available in the network and the BS periodically announces them to devices. When receiving data from multiple devices transmitting in the same time slot, the BS uses zero-forcing (ZF) or conjugate beamforming (CB) processing to distinguish the data of each device from the others. In addition, all devices utilize power control so that the expected received power at the BS is the same for all devices.

A. RECONFIGURABLE ACCESS SCHEME

In this work, we present the access scheme in the time domain for the ease of presentation. The proposed access scheme can also be applied to the frequency domain. The network operates on a frame-by-frame basis. Each time frame is started with a beacon, followed by two segments as illustrated in Figure 2. The time frame consists of $N_{\rm ts}$ time slots.

TABLE 1. Table of key notations.

Notation	Definition
$\theta_d(t)$	Probability of device d having packets at time frame t
a_d	Packet arrival probability of device d
$v_d(t)$	The last time that the BS received a packet from device d
ℓ_d	Large-scale fading coefficient of the channel between device d and the BS
$\mathcal{D}_{\rm s}$	Set of devices transmitting in time slot s
$\mathbf{h}_d$	Small-scale fading vector of the channel between device d and the BS
$P_{\rm R}$	Received power of each device at the BS
$\mathbf{b}_1^{\rm T}$	Received conjugate beamformer for device 1
$\gamma_{\rm CB}^1$	Signal-to-interference-and-noise ratio (SINR) of device 1 after conjugate beamforming (CB)
M	Number of antennas at the BS
$\rho_{\rm R}$	Signal-to-noise ratio (SNR) of each device at the BS
Q	Number of preambles chosen by all devices other than device 1
$\gamma_{\rm ZF}^1$	SINR of device 1 after zero-forcing (ZF) beamforming
$\mathcal{W}_q$	The set of devices choosing preamble $q \in \mathcal{Q}$
$\mathbf{X}$	Time slot allocation matrix
$x_{d,s}$	Element at the $d^{\rm th}$ row and $s^{\rm th}$ column of matrix $\mathbf{X}$
$\mathbf{P}$	Vector containing the device access probabilities
p_d	The $d^{\rm th}$ element of vector $\mathbf{P}$
$S_{\rm gb}$	Expected throughput of the grant-based segment
$S_{\rm gf}$	Expected throughput of the grant-free segment
γ^d	SINR of device d after beamforming
$\gamma^{\rm th}$	SINR threshold
$\gamma_{\rm zf}^d$	SINR of device d after ZF beamforming
$\gamma_{\rm cb}^d$	SINR of device d after CB beamforming
K	Expected number of device other than device d in the grant-free segment
$N_{\rm ts}$	Number of time slots in a time frame
$N_{\rm gf}$	Number of time slots of the grant-free segment
$N_{\rm gb}$	Number of time slots of the grant-based segment
$N_{\rm gb,max}$	Maximum number of time slots of the grant-based segment
$\mathcal{D}_{\rm gf}$	Set of devices in the grant-free segment
$U_{\rm gf}$	Expected number of devices in the grant-free segment
$N_{\rm p}$	Number of preambles
$\mathcal{P}_{\rm cb}^d$	Success probability of device d by CB beamforming
$\mathcal{P}_{\rm zf}^d$	Success probability of device d by ZF beamforming

1) BEACON

At the start of each time frame, using the non-empty queue probabilities of devices, the BS decides the length of the grant-based segment, the set of devices allocated to this segment along with their assigned preambles and time slots, and the access probabilities of other devices. Then it broadcasts this information to devices via a beacon.

2) GRANT-BASED SEGMENT

In this segment, granted devices transmit their packets in the corresponding allocated time slots. Up to $N_{\rm p}$ devices can be granted the same time slot and each device in this segment

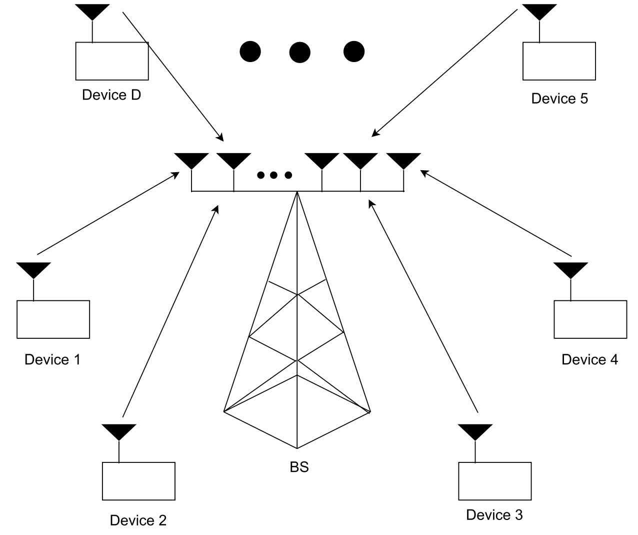


FIGURE 1. System model.

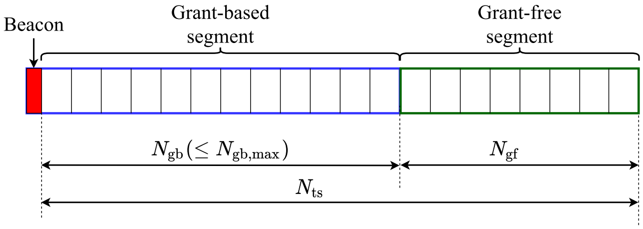


FIGURE 2. Time frame structure.

will be allocated a preamble. Devices transmit their assigned preambles first, followed by their data payload. The BS uses the estimated device channel information for beamforming to cancel interference for the data of each device. In this segment, transmission is considered successful if the SINR after beamforming is larger than or equal to a threshold value. The length of this segment is N_{gb} time slots, which is limited to at most $N_{\text{gb,max}}$ time slots.

3) GRANT-FREE SEGMENT

In this segment, devices that have packets for transmission but are not granted a time slot, contend with each other, by using p-persistent ALOHA as follows [21]. In each time slot, with probability p, a device having packets in its queue randomly chooses a preamble from N_{p} available ones and transmits this preamble followed by its data. The BS also uses the received preamble signal to estimate the device channel information and then uses the device channel estimation for beamforming to cancel interference for the data of each device. If two or more devices select the same preamble, their activity cannot be detected, and their channels cannot be estimated. Therefore, their transmission is unsuccessful. The transmission of each device will be considered successful if its chosen preamble is not selected concurrently by any other devices and its signal-to-interference-and-noise ratio (SINR) after beamforming is larger than or equal to a threshold value, γ_{th}. The access probability, p, helps to control the traffic load in this segment to reduce collisions and thus, improve throughput. In each time frame, after one successful transmission, the device is not allowed to transmit any other packets. The length of the grant-free segment is $N_{\text{gf}} = N_{\text{ts}} - N_{\text{gb}}$.

In the context of mMTC, a pure grant-based access scheme can cause low channel utility since resources might be allocated to devices with low traffic demand. On the other hand, a pure grant-free access scheme can cause severe collisions when traffic demand is high. The combination of the two transmission segments enables devices having high traffic demand to transmit without collisions. In addition, when devices having high traffic being allocated to the grant-based segment, there will be less devices in the grant-free segment and thus, there will be less collisions.

B. TRAFFIC MODEL

We assume that a packet is generated at device d with probability of a_d in each time frame and is added to the device queue. It is assumed that the BS is aware of the packet arrival probabilities of the devices and it keeps a vector $\boldsymbol{V}(t) = [v_d(t)]_{\forall d}$, where v_d indicates the last time that the BS has received a packet from the device d. Moreover, 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 time frame t, the BS updates $\theta_d(t)$ which is the probability of the device d having a non-empty queue at t as

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

C. CHANNEL MODEL

The channel model under consideration is uncorrelated Rayleigh fading channel and the channel response is assumed to be unchanged within each time slot. The channel response vector between device d and the BS is $\mathbf{g}_d = \sqrt{\ell_d}\mathbf{h}_d \in \mathbb{C}^M$. ℓ_d is the large-scale coefficient of the channel between device d and the BS. $\mathbf{h}_d \sim \mathcal{CN}(0, \mathbf{I_M})$ is the small-scale fading vector of the channel between device d and the BS.

D. OVERVIEW OF ZF BEAMFORMING AND CONJUGATE BEAMFORMING

This section briefly reviews the formulas to calculate the SINR after ZF beamforming and CB of each device as provided in [21]. Let $\mathcal{D}_s$ be the set of devices transmitting in time slot s. In the following subsections, we will present the formulas of SINR of the first device in $\mathcal{D}_s$ as an example.

1) CB

Let $\mathbf{h}_1$ denote the channel vector of the first device and $\mathbf{h}_d$ denote the channel vector of device $d \in \mathcal{D}_s$, the SINR of the first device after CB is

$$\gamma_{\mathrm{CB}}^{1} = \frac{P_{\mathrm{R}}|\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_1|^2}{|\mathbf{h}_1^{\mathrm{H}}\mathbf{n}|^2 + P_{\mathrm{R}} \sum\limits_{\substack{d\in\mathcal{D}_{\mathrm{gf}}\\ d\neq 1}} |\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_d|^2}, \tag{2}$$

where $\mathbf{n} \sim \mathcal{CN}(0, \sigma_{\mathrm{n}}^2\mathbf{I}_{\mathrm{M}})$ is a vector of the additive white Gaussian noise and P_{R} is the received power at the BS of each device. By the strong law of large number, $[\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_1/M] \xrightarrow{M\to\infty} 1$, thus we can simplify (2) as

$$\gamma_{\mathrm{CB}}^{1} = \frac{\rho_{\mathrm{R}}M}{1 + (\rho_{\mathrm{R}}/M) \sum\limits_{\substack{d\in\mathcal{D}_{\mathrm{gf}}\\ d\neq 1}} |\mathbf{h}_1^{\mathrm{H}}\mathbf{h}_d|^2}, \tag{3}$$

where $\rho_{\mathrm{R}} = P_{\mathrm{R}}/\sigma_{\mathrm{n}}^2$ is the SNR at the BS of each device.

2) ZF BEAMFORMING

In order to present the formula of the SINR of the first device after ZF beamforming, we define $\mathcal{Q} = \{1, 2, \ldots, Q\}$ as the set of preambles chosen by all the transmitting devices *other than* device 1, $\mathcal{W}_q$ as the set of devices selecting preamble $q \in \mathcal{Q}$ and w_q as an arbitrary element in $\mathcal{W}_q$. In addition, let $\mathbf{a}_q = \sum_{d\in\mathcal{W}_q} \mathbf{h}_d$, $\mathbf{A} = [\mathbf{h}_1, \mathbf{a}_2, \ldots, \mathbf{a}_Q]$ and $\mathbf{B} = (\mathbf{A}^{\mathrm{H}}\mathbf{A})^{(-1)}\mathbf{A}^{\mathrm{H}}$, the SINR of device 1 is

$$\gamma_{\mathrm{ZF}}^{1} = \frac{\rho_{\mathrm{R}}}{\rho_{\mathrm{R}}\left|\sum\limits_{d\in\mathcal{W}_q\setminus w_q : s\in\mathcal{Q}} \sqrt{2}\mathbf{b}_1^{\mathrm{T}}\mathbf{h}_d\right|^2 + ||\mathbf{b}_1^{\mathrm{T}}||^2}, \tag{4}$$

where $\mathbf{b}_1^{\mathrm{T}}$ is the first row of the matrix $\mathbf{B}$ and the notation $\mathcal{W}_q \setminus w_q$ means the set $\mathcal{W}_q$ excluding the element w_q.

III. PROBLEM FORMULATION

The reconfigurable access scheme aims to maximize the expected throughput of each time frame. The time frame is divided into the grant-based and grant-free segments. The grant-based access scheme is more efficient for devices with a high probability of having a packet for transmission, while the grant-free access scheme can achieve better throughput for devices with low probabilities of having a packet for transmission. In the following subsections, we will first derive the expected throughput of each segment, and then formulate the problem to optimize the sum of the expected throughputs achieved in these two segments.

A. EXPECTED THROUGHPUT OF GRANT-BASED ACCESS SEGMENT

In this subsection, the expected throughput of the grant-based access segment is derived. Let $\mathbf{X} : D \times N_{\mathrm{ts}}$ be the time slot allocation matrix, where $x_{d,s} = 1$ indicates device d is assigned to time slot s. The expected throughput of the grant-based access segment of that time frame can be written as

$$S_{\mathrm{gb}} = \sum_{s}^{N_{\mathrm{ts}}}\sum_{d}^{D} x_{d,s}\theta_d\mathcal{P}(\gamma^d \geq \gamma_{\mathrm{th}}), \tag{5}$$

where γ^d is the SINR of device d at the BS *after* ZF or CB processing, and γ_{th} is an SINR threshold value. $\mathcal{P}(\gamma^d \geq \gamma_{\mathrm{th}})$, called success probability, is the probability that the SINR of device d after ZF or CB processing is at least equal to the threshold, In the following, the success probabilities derived in [21] for ZF and CB techniques are presented.

Note that in equations (6) and (7), the subscripts zf and cb denote the type of beamforming used.

1) SUCCESS PROBABILITY IN GRANT-BASED SEGMENT USING ZF BEAMFORMING

In the grant-based segment, devices that are granted the same time slot, use different preambles, which are assigned to them by the BS. Therefore, in this segment, no preamble collision happens. Consequently, the probability that the device transmits its packet successfully only depends on its SINR at the BS. In [21], the probability that the SINR of device d is not smaller than a threshold in case of ZF beamforming is

$$\mathcal{P}(\gamma_{\mathrm{zf}}^d \geq \gamma_{\mathrm{th}}) = e^{-\frac{\gamma_{\mathrm{th}}}{\rho_{\mathrm{R}}}} \sum_{p=0}^{M-K_s-1} \frac{1}{p!}\left(\frac{\gamma_{\mathrm{th}}}{\rho_{\mathrm{R}}}\right)^p, \tag{6}$$

where $K_s = \sum_{d=1}^{D} x_{d,s}\theta_d - 1$ is the average number of devices *other than* device d in time slot s, and γ_{zf}^d is the SINR of device d after ZF beamforming processing.

2) SUCCESS PROBABILITY IN GRANT-BASED SEGMENT USING CB

Using the same notation of K_s, the probability of the SINR of device d is larger than a threshold when CB is used is given as

$$\begin{aligned}
&\mathcal{P}(\gamma_{\mathrm{cb}}^d \geq \gamma_{\mathrm{th}})\\
&\quad = 1 - \eta^{-K_s+1}e^{-\beta\Lambda}\\
&\qquad + (1-\eta)\sum_{n=0}^{K_s-2}\frac{1}{n!}\eta^{n-K_s+1}\Gamma(n+1, \frac{\sqrt{M}}{\sqrt{M}-1}\Lambda),
\end{aligned} \tag{7}$$

where $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t}dt$ is the upper-incomplete Gamma function, $\eta = K_s/(\sqrt{M} + K_s - 1)$, $\beta = \sqrt{M}/(\sqrt{M} + K_s - 1)$ and $\Lambda = [M/\gamma_{\mathrm{th}}] - [1/\rho_{\mathrm{R}}]$.

B. EXPECTED THROUGHPUT OF GRANT-FREE ACCESS SEGMENT

In this subsection, the expected throughput of the grant-free access segment for both ZF and CB processing is derived. Denote $\mathbf{P} = [p_d]_{\forall d}$ as a vector containing the access probability of each device and define a vector $\mathbf{Z} = [z_d]_{\forall d}$, where $z_d = \sum_{s=1}^{N_{\mathrm{ts}}} x_{d,s}$ indicates whether any time slot is allocated to device d or not. We have

$$S_{\mathrm{gf}} = \sum_{d\in\mathcal{D}_{\mathrm{gf}}} p_d\theta_d\mathcal{P}^d N_{\mathrm{gf}}, \tag{8}$$

where N_{gf} is the duration of the grant-free access segment of a time frame in time slots, and $\mathcal{D}_{\mathrm{gf}} = \{d \mid z_d = 0\}$ is the set of devices allocated in grant-free segment of the according time frame. Consequently, N_{gf} can be calculated as

$$N_{\text{gf}} = N_{\text{ts}} - \sum_{s=1}^{N_{\text{ts}}} H(\sum_{d=1}^{D} x_{d,s}), \tag{9}$$

where $H(x)$ is defined as

$$H(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Furthermore, $\mathcal{P}^d$ is the success probability of device d in the grant-free segment. The following subsections present the derivation of $\mathcal{P}^d$ for ZF beamforming and CB.

1) SUCCESS PROBABILITY IN GRANT-FREE SEGMENT UZING ZF BEAMFORMING

Let $\mathcal{P}_{\text{zf}}^d$ denote the success probability of device d when ZF beamforming is used. In [21], this probability is given as

$$\mathcal{P}_{\text{zf}}^d = \sum_{Q=1}^{\min\{N_{\text{P}}-1,K\}} \tilde{\mathcal{P}}_{\text{zf}}(Q|K)\mathcal{P}(\gamma_{\text{zf}}^d \geq \gamma_{\text{th}}|K,Q). \tag{11}$$

In equation (11), K is the average number of devices in grant-free access segment *other than* device d and is calculated as follows

$$K = \left\lfloor \sum_{d\in\mathcal{D}_{\text{gf}}} p_d\theta_d - 1 \right\rfloor, \tag{12}$$

where the function $\lfloor x \rfloor$ is the floor function mapping x to the largest integer that is smaller than or equal to x. $\tilde{\mathcal{P}}_{\text{zf}}(Q|K)$ is the probability that Q preambles are selected by K devices and no collision occurs between device d and the other K devices. $\tilde{\mathcal{P}}_{\text{zf}}(Q|K)$ is given in [21] as

$$\tilde{\mathcal{P}}_{\text{zf}}(Q|K) = \frac{\binom{N_{\text{P}}-1}{Q} Q! \left\{ {K \atop Q} \right\}}{N_{\text{P}}^K}, \tag{13}$$

where $\left\{ {K \atop Q} \right\} = (1/Q!)\sum_{j=0}^{Q}(-1)^{Q-j}\binom{Q}{j}j^K$ and $\binom{n}{r} = [n!/(r!(n-r)!)]$.

When $K > Q$, $\mathcal{P}(\gamma_{\text{zf}}^d \geq \gamma_{\text{th}}|K,Q)$ can be approximated as [21],

$$\mathcal{P}(\gamma_{\text{zf}}^d \geq \gamma_{\text{th}}|K,Q) \approx e^{-\frac{\gamma_{\text{th}}}{\rho_{\text{R}}}} \sum_{p=0}^{M-Q-1}\sum_{q=0}^{p}\binom{p}{q}\frac{\gamma_{\text{th}}^p}{p!}\frac{\rho_R^{q-p}}{(K-Q-1)!}\frac{\Gamma(K-Q+q,0)}{(1+\gamma_{\text{th}})^{K-Q+q}}, \tag{14}$$

where $\Gamma(s,x) = \int_x^\infty t^{s-1}e^{-t}dt$ is the upper incomplete Gamma function.

Furthermore, for $K = Q$, we have

$$\mathcal{P}(\gamma_{\text{zf}}^d \geq \gamma_{\text{th}}|K,Q) \approx e^{-\frac{\gamma_{\text{th}}}{\rho_{\text{R}}}} \sum_{p=0}^{M-K-1}\frac{1}{p!}\left(\frac{\gamma_{\text{th}}}{\rho_{\text{R}}}\right)^p. \tag{15}$$

2) SUCCESS PROBABILITY IN GRANT-FREE SEGMENT USING CB

Adopting the notation of K, the success probability of CB is given in [21] as

$$\mathcal{P}_{\text{cb}}^d = \tilde{\mathcal{P}}_{\text{cb}}(K)\tilde{\mathcal{P}}(\gamma_{\text{cb}}^d \geq \gamma_{\text{th}}|K), \tag{16}$$

where $\tilde{\mathcal{P}}_{\text{cb}}(K)$ is the probability that no preamble collisions occur between K other devices and device d and is provided as

$$\tilde{\mathcal{P}}_{\text{cb}}(K) = \left(1 - \frac{1}{N_{\text{P}}}\right)^K. \tag{17}$$

$\tilde{\mathcal{P}}(\gamma_{\text{cb}}^d \geq \gamma_{\text{th}}|K)$ is the probability that the SINR of device d after CB processing is larger than the threshold γ_{th}, which is the same as equation (7).

C. OPTIMIZATION PROBLEM

In the proposed access scheme, before a time frame starts, the BS decides which devices should be allocated to the grant-based segment as well as the access probabilities for the rest of devices that compete with each other in the grant-free segment so that the total expected throughput of that time frame is maximized. This problem can be formulated as follows

$$\begin{aligned} \underset{\mathbf{X},\mathbf{P}}{\text{maximize}}\ & S_{\text{gb}} + S_{\text{gf}}, \\ \text{subject to: } & \text{C18.1: } x_{d,s} \in \{0,1\}, \quad \forall d,s, \\ & \text{C18.2: } z_d p_d = 0, \quad \forall d, \\ & \text{C18.3: } z_d \in \{0,1\}, \quad \forall d, \\ & \text{C18.4: } 0 \leq p_d \leq 1, \quad \forall d, \\ & \text{C18.5: } N_{\text{gb}} \leq N_{\text{gb,max}}, \end{aligned} \tag{18}$$

where S_{gb} and S_{gf} are given in (5) and (8), respectively. Condition C18.2 is to ensure that each device is only assigned to either grant-based or grant-free segment. Condition C18.3 guarantees that each device is allocated at most one time slot and condition C18.4 ensures the access probability is between 0 and 1. Finally, condition C18.5 is to limit the length of the grant-based segment to $N_{\text{gb,max}}$ time slots.

IV. PROPOSED ALGORITHM

It is obvious that (18) is a mixed-integer optimization problem. In addition, in (6), (7) and (11), since the decision variables $\mathbf{X}$ and $\mathbf{P}$ appear in the upperbound of the sums, taking derivatives becomes difficult. Furthermore, branch-and-bound method is also inapplicable since $\mathbf{X}$ has a huge number of elements and, consequently, the search space is inherently large. To address this issue, we propose a suboptimal algorithm with affordable computational complexity, namely Algorithm 1.

In this iterative Algorithm 1, the length of the grant-based segment is initially set to 0 and then increased by one time slot at each iteration. Furthermore, in each iteration, the problem is divided into two sub-problems performed in two steps:

Algorithm 1 Hill-Climbing Algorithm for Solving Problem (18)

```
Initialize $N_{\mathrm{gb}} = 0, S = 0, S' = 0, \mathbf{X}' = 0, \mathbf{P}' = 0$
repeat
    Step 1: $S \leftarrow S', \mathbf{X} \leftarrow \mathbf{X}', \mathbf{P} \leftarrow \mathbf{P}', N_{\mathrm{gb}} \leftarrow N_{\mathrm{gb}} + 1$
    Step 2: Solve $\mathbf{X}'$ by Algorithm 2
    Step 3: Solve $\mathbf{P}'$ by the method in section IV-B
        1) Solve (23) to obtain the initial value of $U_{\mathrm{gf}}$
        2) Use the obtained initial $U_{\mathrm{gf}}$ to solve (22) by hill-climbing integer search to obtain the optimal value of $U_{\mathrm{gf}}$
        3) Solve (24) using the optimal value of $U_{\mathrm{gf}}$
    Step 4: $S' \leftarrow S_{\mathrm{gb}} + S_{\mathrm{gf}}, \Delta S \leftarrow S' - S$
until $\Delta S < 0$ or $N_{\mathrm{gb}} = N_{\mathrm{gb,max}}$
if $\Delta S < 0$ then
    return $\mathbf{X}$ and $\mathbf{P}$
else
    return $\mathbf{X}'$ and $\mathbf{P}'$
end if
```

1) Choosing devices for the grant-based segment and assigning them to their corresponding time slots, 2) Deriving ***P*** for the remaining devices to maximize the throughput of this segment. The iteration continues until no further throughput enhancement can be achieved by increasing the length of the grant-based segment or the length of grant-based segment exceeds its set limit $N_{\mathrm{gb,max}}$. In the following, we present how to solve each of these two sub-problems.

A. GRANT-BASED SCHEDULING SUB-PROBLEM

In this subsection, we present an algorithm to maximize the expected throughput of this segment. In particular, given the fixed length of the grant-based segment, the expected throughput of the segment depends on which devices are chosen for this segment and how these devices are grouped such that devices belong to the same group are allocated the same time slot to transmit over it.

In order to maximize the expected throughput of this segment, devices with the highest non-empty queue probabilities are chosen. The reason is that in the grant-based segment, time slots are allocated to devices, and to maximize the expected throughput, time slots should be allocated to devices that have a packet for transmission. However, when the number of devices transmitting in a same time slot increases, interference also increases. Consequently, the probability that the SINR of each device is larger than or equal to the threshold decreases and the expected throughput of that time slot will decrease if the interference is too high. Therefore, in the proposed Algorithm 2, first, we sort devices in the set $\mathcal{D}$ according to their probabilities of having non-empty queues in a descending order. Then, an iterative procedure starts and iterates over the devices in $\mathcal{D}$. In each iteration, the algorithm considers a device in $\mathcal{D}$ and iterates over the time slots in the grant-based segment to find a time slot that the device can be allocated to without causing throughput decrease. Figure 3 shows an example of this algorithm with 11 devices, 3 time slots and 3 preambles.

Algorithm 2 Grant-Based Scheduling Algorithm

```
Sort devices in $\mathcal{D}$ with respect to their non-empty queue probabilities in descending order.
Set $\mathbf{X} = \mathbf{0}, \mathcal{D}_1 = \emptyset, \mathcal{D}_2 = \emptyset, \ldots, \mathcal{D}_{N_{\mathrm{gb}}} = \emptyset, s = 1, i = 0.$
repeat
    Set $c = 0$.
    $i \leftarrow i + 1$.
    Set $d \leftarrow$ device $i_{\mathrm{th}}$ in $\mathcal{D}$.
    while $c < N_{\mathrm{gb}}$ and $x_{d,s'} = 0 \; \forall \, s' \in [1, N_{\mathrm{gb}}]$ do
        Step 1: Consider allocating device $d$ to time slot $s$:
            Add $d$ to $\mathcal{D}_s$ and calculate the expected throughput of $\mathcal{D}_s$ before and after adding $d$.
            if $|\mathcal{D}_s| < N_{\mathrm{p}}$ and adding $d$ does not decrease the expected throughput of $\mathcal{D}_s$ then
                $x_{d,s} \leftarrow 1$.
                Add $d$ to $\mathcal{D}_s$.
            end if
        Step 2: Update the number of time slots that have been considered:
            $c \leftarrow c + 1$.
        Step 3: Move to the next time slot:
            $s \leftarrow s + 1$.
            if $s > N_{\mathrm{gb}}$ then
                $s \leftarrow 1$.
            end if
    end while
until Each time slot in the grant-based segment has $N_{\mathrm{P}}$ devices or $i > D$
return $\mathbf{X}$
```

*B. **P**-DERIVATION SUB-PROBLEM*

Given the fixed value of $\mathbf{X}$, the sub-problem of solving $\mathbf{P}$ is

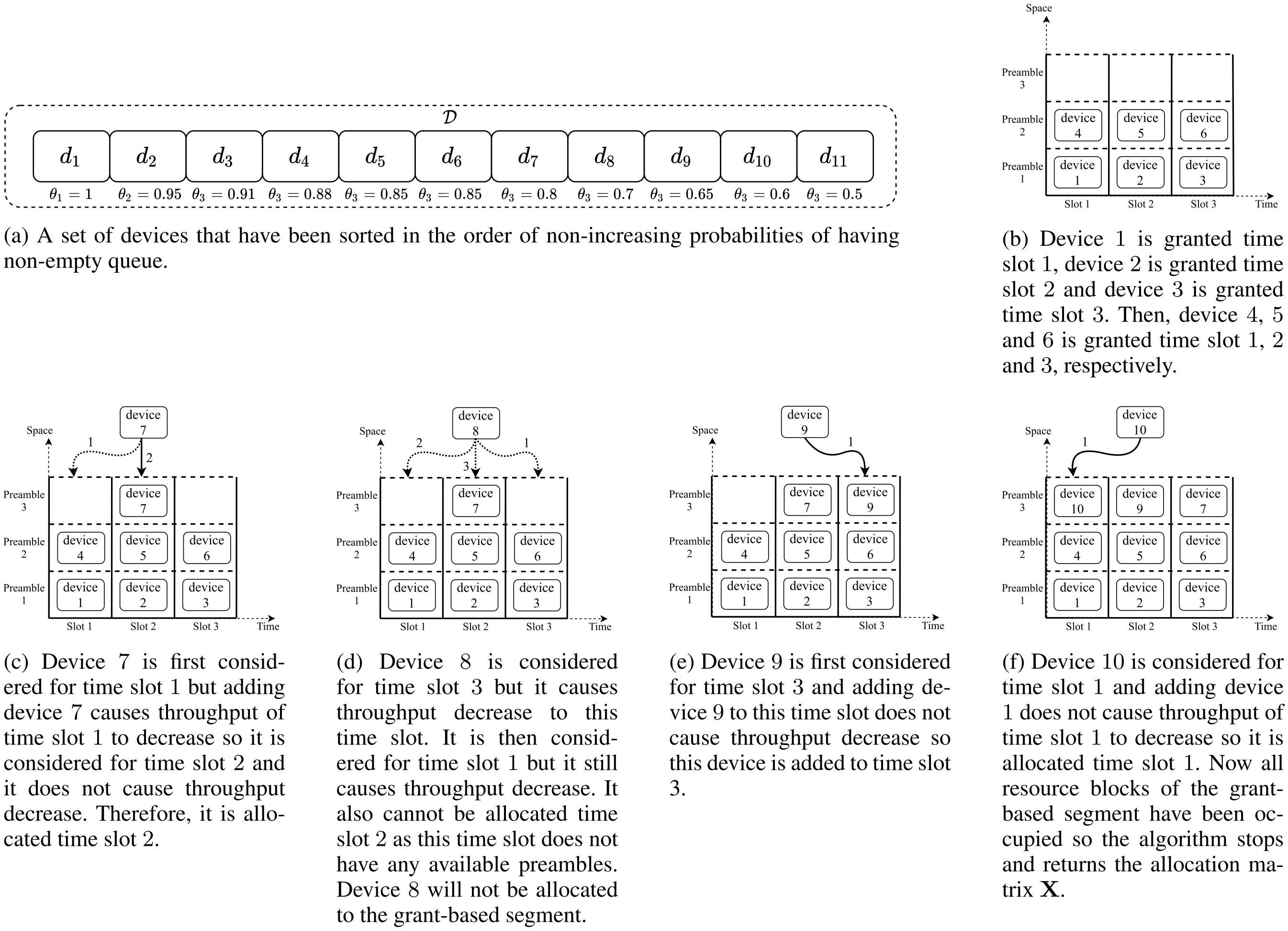
$$\begin{aligned} &\underset{\mathbf{P}}{\text{maximize}} \; S_{\mathrm{gf}}, \\ &\text{subject to: C19.1: } 0 \leq p_d \leq 1, \quad \forall d. \end{aligned} \tag{19}$$

Based on equation (8), this problem can be written as

$$\begin{aligned} &\underset{\mathbf{P}}{\text{maximize}} \; \sum_{d \in \mathcal{D}_{\mathrm{gf}}} p_d \theta_d \mathcal{P}^d U_{\mathrm{gf}}, \\ &\text{subject to: C19.1: } 0 \leq p_d \leq 1, \quad \forall d. \end{aligned} \tag{20}$$

In problem (20), the objective function is discrete and the decision vector $\mathbf{P}$ appears in the sum bound of the success probability $\mathcal{P}^d$, therefore traditional methods such as convex optimization are inapplicable. To tackle this issue, a new variable $U_{\mathrm{gf}} = \sum_{d:z_d=0} p_d \theta_d$ is introduced, which is the average number of devices in the grant-free segment. Obviously, $K = \lfloor U_{\mathrm{gf}} - 1 \rfloor$ and problem (20) is equivalent to

$$\begin{aligned} &\underset{\mathbf{P}, N_{\mathrm{gf}}}{\text{maximize}} \; S_{\mathrm{gf}}, \\ &\text{subject to: C21.1: } 0 \leq p_d \leq 1, \quad \forall d, \\ &\qquad\qquad \text{C21.2: } \sum_{d:z_d=0} p_d \theta_d = U_{\mathrm{gf}}. \end{aligned} \tag{21}$$


$\mathcal{D}$
d_1 d_2 d_3 d_4 d_5 d_6 d_7 d_8 d_9 d_{10} d_{11}
$\theta_1 = 1$ $\theta_2 = 0.95$ $\theta_3 = 0.91$ $\theta_3 = 0.88$ $\theta_3 = 0.85$ $\theta_3 = 0.85$ $\theta_3 = 0.8$ $\theta_3 = 0.7$ $\theta_3 = 0.65$ $\theta_3 = 0.6$ $\theta_3 = 0.5$
Space
Preamble 3
Preamble 2
Preamble 1
Slot 1
Slot 2
Slot 3
Time
device 1
device 2
device 3
device 4
device 5
device 6
device 7
device 8
device 9
device 10
1
2
3

(a) A set of devices that have been sorted in the order of non-increasing probabilities of having non-empty queue.

(b) Device 1 is granted time slot 1, device 2 is granted time slot 2 and device 3 is granted time slot 3. Then, device 4, 5 and 6 is granted time slot 1, 2 and 3, respectively.

(c) Device 7 is first considered for time slot 1 but adding device 7 causes throughput of time slot 1 to decrease so it is considered for time slot 2 and it does not cause throughput decrease. Therefore, it is allocated time slot 2.

(d) Device 8 is considered for time slot 3 but it causes throughput decrease to this time slot. It is then considered for time slot 1 but it still causes throughput decrease. It also cannot be allocated time slot 2 as this time slot does not have any available preambles. Device 8 will not be allocated to the grant-based segment.

(e) Device 9 is first considered for time slot 3 and adding device 9 to this time slot does not cause throughput decrease so this device is added to time slot 3.

(f) Device 10 is considered for time slot 1 and adding device 1 does not cause throughput of time slot 1 to decrease so it is allocated time slot 1. Now all resource blocks of the grant-based segment have been occupied so the algorithm stops and returns the allocation matrix **X**.

FIGURE 3. An example of Algorithm 2 for $N_{\rm gb} = 3$ and $N_{\rm P} = 3$.

The benefit of introducing $U_{\rm gf}$ is that when it is fixed, the success probabilities are fixed and optimization problem (21) becomes a linear programming problem, which can be solved efficiently. In other words, the introduction of $U_{\rm gf}$ and constraint C21.2 aids in dealing with the problem of decision variables in the sum bounds. Having introduced $U_{\rm gf}$, problem (21) can be solved in two steps: first, obtaining $U_{\rm gf}$ and then solving **P**. The following subsections will discuss these two steps in detail.

1) DERIVATION OF $U_{\rm gf}$

With the introduction of $U_{\rm gf}$, the sub-problem to obtain this variable can be written as

$$\begin{aligned} &\underset{U_{\rm gf}}{\text{maximize}}\ \mathcal{P}^d N_{\rm gf} U_{\rm gf}, \\ &\text{subject to: C22.1: } 0 \leq U_{\rm gf} \leq \sum_{d:z_d=0} \theta_d. \end{aligned} \tag{22}$$

The remaining issue is how to efficiently search for the optimal value of $U_{\rm gf}$. It can be seen that $S_{\rm gf}$ is a piecewise function. Moreover, when $U_{\rm gf}$ takes values between any two consecutive integers x and $x+1$, $S_{\rm gf}$ is an increasing function since K will be a fixed value and the success probability remains the same. The jumps (either up or down) of $S_{\rm gf}$ occur at points making $U_{\rm gf}$ integer values. Thus, we only need to search within the set of integer values of $U_{\rm gf}$ for the optimal value. For integer value search, hill-climbing can be utilized. However, for hill-climbing search, we need a good initial value, otherwise, the number of iterations will be large. The proposed approach to initialize a value of $U_{\rm gf}$ is as follows. First, consider the case that M is very large, according to [21], the success probability approaches $(1-(1/N_{\rm P}))^{U_{\rm gf}-1}$. In this case, problem (22) reduces to

$$\begin{aligned} &\underset{U_{\rm gf}}{\text{maximize}}\ (1-\frac{1}{N_{\rm p}})^{U_{\rm gf}-1} U_{\rm gf} N_{\rm gf} \\ &\text{subject to: C23.1: } 0 \leq U_{\rm gf} \leq \sum_{d:z_d=0} \theta_d, \end{aligned} \tag{23}$$

which is a strictly quasi-concave problem with differentiable objective function and constraints. Thus, its local maximum is also a global maximum, and we can utilize a non-linear programming technique to obtain its optimal value. After obtaining the solution of problem (23), we use this value of

$U_{\rm gf}$ to start hill-climbing search for solving problem (21). Note that problem (23) is *not equivalent* to problem (22), it is only a problem that we solve to obtain a fairly good initial value of $U_{\rm gf}$.

2) DERIVATION OF P

After obtaining the optimal value of $N_{\rm gf}$, the success probability $\mathcal{P}^d$ can be calculated and problem (21) becomes

$$\begin{aligned} &\underset{\mathbf{P}}{\text{maximize}}\; S_{\rm gf} \\ &\text{subject to: C24.1: } 0 \le p_d \le 1, \quad \forall d, \\ &\qquad\qquad \text{C24.2: } \sum_{d:z_d=0} p_d\theta_d = U_{\rm gf}. \end{aligned} \tag{24}$$

In problem (24), the problem of decision variables in the sum bound has been solved. Therefore, it becomes a linear programming problem, which can be solved efficiently by traditional linear programming methods.

C. COMPLEXITY OF THE PROPOSED ALGORITHM

In this subsection, we present the complexity of Algorithm 1. For the grant-based scheduling sub-problem, in the worst case, we have to consider $N_{\rm gb}$ time slots for each device, therefore we have to perform $DN_{\rm gb}$ operations for each value of $N_{\rm gb}$. If the optimal value of $N_{\rm gb}$ is $N_{\rm gb,max}$, we have to vary $N_{\rm gb}$ from 0 to $N_{\rm gb,max}$. Therefore, the maximum number of operations for the grant-based scheduling sub-problem is $D(1 + 2 + \ldots + N_{\rm gb,max}) = \frac{DN_{\rm gb,max}(N_{\rm gb,max}+1)}{2}$. Thus, the complexity of this subproblem is $\mathcal{O}(DN^2_{\rm gb,max})$, which is polynomial.

For the grant-free sub-problem, for each value of $N_{\rm gb}$, we have to solve the single-variable problem (23), and to calculate the objective function of (22) at most $N_{\rm P}$ times. Finally, we have to solve the linear programming problem (24), which has polynomial complexity [31].

On the other hand, if we need to solve problem (18) optimally, we can use the branch-and-bound algorithm for the grant-based scheduling sub-problem with an exponential complexity, which becomes infeasible when D is large.

V. NUMERICAL RESULTS

In order to evaluate the performance of the proposed access scheme, simulations on Matlab environment are conducted and its optimization toolbox is used to solve problem (23). To simulate the heterogeneous traffic in mMTC network, the arrival rates of devices are generated as follows. Devices are divided into four groups, each of which has 25% of devices. For the first group, the arrival rates are generated from the uniform distribution on the interval $[0.01, 0.25]$. For the second group, the arrival rates of devices are randomly and uniformly distributed in the interval $[0.25, 0.5]$. Similarly, the arrival rates of the third and the fourth groups are randomly distributed in the intervals $[0.5, 0.75]$ and $[0.75, 1]$, respectively. Other simulation parameters are listed in Table 2.

TABLE 2. Simulation parameters.

D	700
$N_{\rm p}$	20
M	50
$\rho_{\rm R}$	1 dB
$\gamma_{\rm th}$	8 dB
$N_{\rm ts}$	20
$N_{\rm gb,max}$	12

In order to show the effectiveness of our proposed scheme in terms of achieved throughput and average packet delay, it is compared to the pure grant-free random access scheme with optimized access probability, referred to as Optimized RA from this point onward. In this scheme, in each time slot, with an access probability of p_d, each device d randomly chooses a preamble and transmits this preamble along with its data packet. The transmission of a device is considered successful if it does not have preamble collision and its SINR after beamforming is larger than a threshold. The access probabilities are optimized by solving problem (19). For simulations, the throughput is measured in terms of the average number of packets successfully received at the BS per time frame (pckt/tf) and the delay is measured in terms of the number of time frames. The results are shown in Figures 4-9.

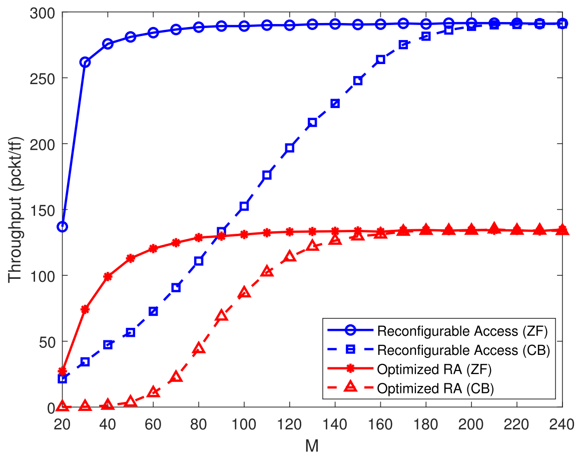


FIGURE 4. Throughput vs *M*.

It can be seen from Figure 4 that when M increases from 20 to 30, there is a significant improvement in throughput of the proposed access scheme with ZF beamforming since when M increases the received SINR also increases as pointed out in [21]. However, from $M = 30$ to $M = 240$ the improvement is not significant, which means that $M = 30$ for ZF almost converges to the case of large M while for Optimized RA with ZF beamforming, the proportion of improvement is still significant in this interval. In addition, in the case of ZF beamforming, the throughput achieved by the proposed approach is usually three times more than that of Optimized RA. This is because the grant-based seg-

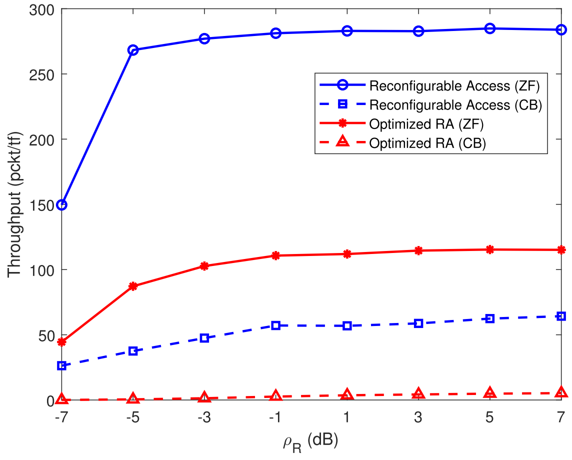


FIGURE 5. Throughput vs ρ_R.

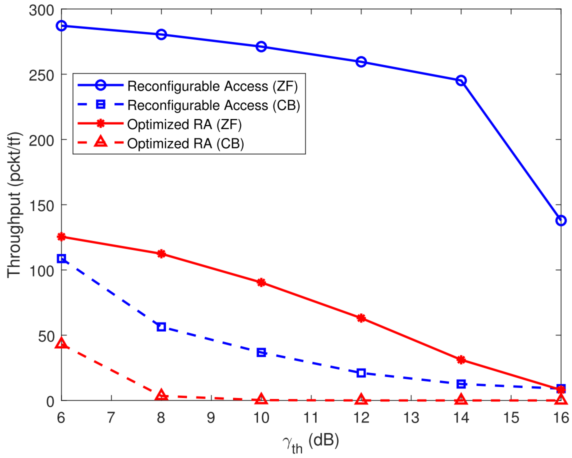


FIGURE 6. Throughput vs γ_{th}.

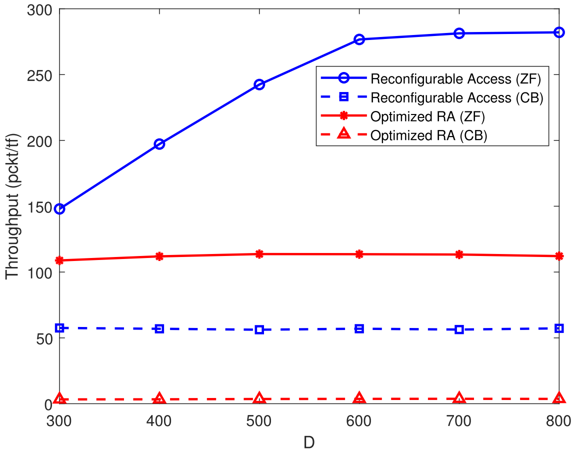


FIGURE 7. Throughput vs *D*.

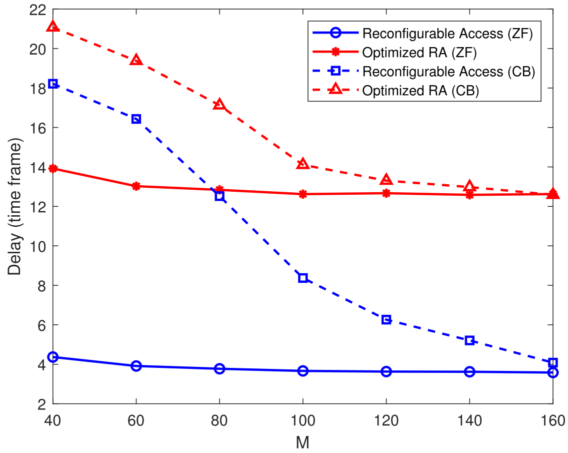


FIGURE 8. Delay vs *M*.

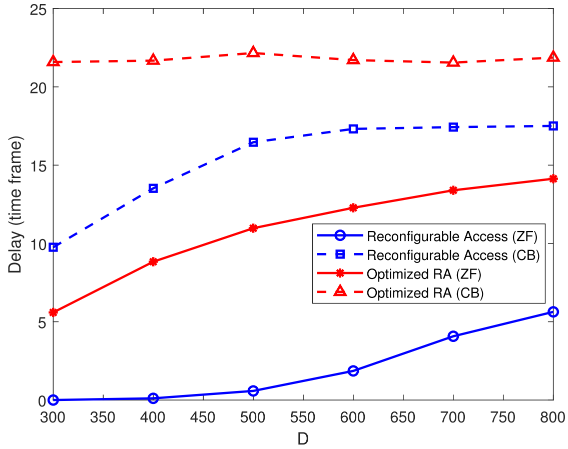


FIGURE 9. Delay vs *D*.

ment allows devices with high expected throughput to transmit without preamble collisions, the achieved throughput of each time slot in the grant-based segment will be higher than that of the grant-free segment. For CB, when M is small, the achieved throughput is low for Optimized RA and the reconfigurable access scheme. However, with higher M, the throughput also increases and eventually converges to the throughput achieved by ZF beamforming.

In Figure 5, the throughput achieved by different values of ρ_R is plotted. When ρ_R varies from -7dB to -5dB, the throughput achieved by the proposed scheme with ZF beamforming is improved significantly and from -5dB to higher values, the improvement becomes insignificant. Meanwhile, for Optimized RA with ZF, the improvement is not as much as the proposed scheme with ZF beamforming and then it almost saturates from $\rho_R = -5$dB. Similar to Figure 5, the throughput achieved by the proposed scheme with ZF beamforming is approximately three times more than that of Optimized RA with ZF beamforming. For CB, the proposed scheme also yields much higher throughput than that of Optimized RA with the same beamforming technique. Overall, when ρ_R is low, increasing it (by raising the transmission power) can improve throughput significantly since this helps in overcoming the influence of noise. However, further increase in transmission power (and subsequently the SNR) does not cause significant throughput improvement since interference also increases. In this situation, having more antennas at the BS is desirable. For example, for the proposed scheme using CB, increasing ρ_R to higher than -1dB does not improve its throughput, however, as shown in Figure 4, increasing M leads to further throughput gain since the interference cancellation becomes better at high M.

In Figure 6, throughput for different values of γ_{th} is shown. In general, when γ_{th} increases, the throughput of all access schemes deteriorates. However, with ZF beamforming, the throughput of the reconfigurable access scheme only decreases significantly when γ_{th} is 14dB while Optimized RA with ZF beamforming decreases sharply since 8dB. For CB, the throughput of both the reconfigurable access scheme and Optimized RA exhibits a similar trend, i.e., it decreases quickly when γ_{th} increases from 6dB to 8dB. In general, for the same beamforming technique, the proposed scheme always achieves significantly higher throughput than Optimized RA as expected.

In Figure 7, the proposed scheme achieves higher throughput when D increases from 300 to 600 and then saturates at approximately 280 while the Optimized RA with ZF beamforming offers the same throughput of about 115 over a wide range of D. Similar to other figures, in Figure 7, the proposed access scheme also achieves significantly higher throughput than Optimized RA with the same beamforming technique, which shows the advantage of the proposed scheme in terms of achieved throughput. Overall, the benefit of the proposed scheme is more pronounced when D is high because in this case, the traffic demand is high and a pure random access scheme will suffer from severe collisions. In contrast, the proposed scheme allocates devices having high non-empty queue probabilities to the grant-based segment, which allows them to transmit in a collision-free manner and reduces the traffic load on the grant-free segment.

Figure 8 illustrates the average packet delay versus M of the proposed access scheme and Optimized RA. For the same beamforming and M, the proposed access scheme offers significantly lower delay than Optimized RA. For the same access scheme (Optimized RA or proposed scheme), ZF-beamforming yields a delay almost the same for a wide range of M and lower than CB-beamforming. The delay provided by CB-beamforming is significantly reduced by increasing M and approaches that provided by ZF beamforming at large M.

In Figure 9, the average packet delay versus D is plotted. For the same access scheme, ZF-beamforming provides shorter delay and its delay increases faster with increasing D, as compared to CB. For the same beamforming type, the proposed scheme offers much shorter delay than Optimized RA, e.g., about $16/22$ for CB and $4/12$ for ZF beamforming when $D = 700$.

For further performance assessment of the proposed algorithm, we develop the optimum algorithm for solving (18), called Algorithm 3, in the appendix, and compare the achieved throughput versus M of the proposed algorithm and Algorithm 3 (Optimal) using ZF beamforming (Figure 10) or CB (Figure 11). In Figure 10, we use the simulation parameters in Table 2, however, in Figure 11, we reduce D to 245 devices, N_{ts} to 10 time slots, $N_{gb,max}$ to 6 time slots and N_P to 14 preambles. The reason for these reductions is that in the case of CB, the optimal grant-based scheduling (25) takes too long to converge when D is large. In both figures, when M is small, the proposed algorithm offers a slightly worse throughput than the optimal Algorithm 3. However, when M becomes larger, this performance gap vanishes. The reason is that when M is small, $U_{gb,max} \ll N_P$ and thus, devices must be carefully chosen for the grant-based segment so that the expected number of devices in each time slot does not exceed $U_{gb,max}$. When M is large enough, N_P devices can be allocated to a time slot without causing throughput decrease so both algorithms will try to allocate devices having highest non-empty queue probabilities to the grant-based segment. In other words, as M increases, the proposed algorithm can offer the same throughput as the optimal Algorithm 3.

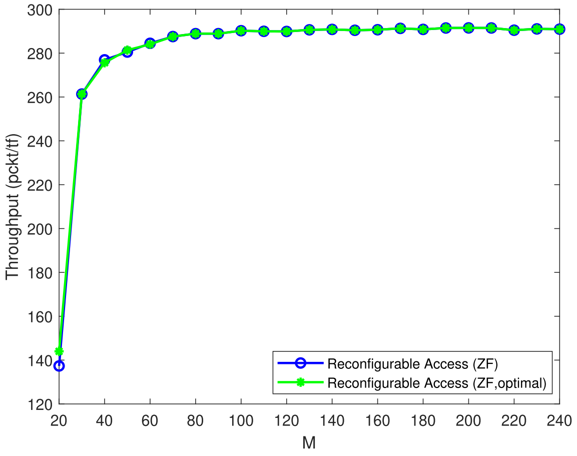


FIGURE 10. **Throughput vs *M* using ZF beamforming.**

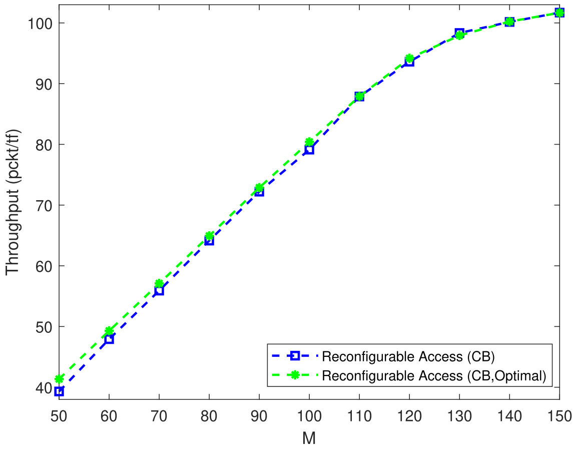


FIGURE 11. **Throughput vs *M* when CB is used.**

VI. CONCLUSION

This paper considers a network of massive machine-type devices, where to provide massive connectivity, a massive MIMO BS is deployed. Aiming to maximize the expected throughput of the network consisting of devices with diverse packet arrival probabilities, a reconfigurable and traffic-aware access scheme is proposed. At each time frame, the proposed scheme dynamically switches from the grant-based scheme to the grant-free scheme, as the grant-based scheme is more efficient for devices with high chance of packet transmission and the grant-free scheme can gain better throughput for devices with low probability of having packet for transmission. To maximize the expected throughput of each time frame, an optimization problem to decide which devices should be assigned to the grant-based or grant-free segment along with the optimal access probabilities for devices in the grant-free segment is formulated. To solve this optimization problem, an iterative algorithm with affordable computational complexity is proposed. Performance comparison of the proposed access scheme and the grant-free random access scheme with optimized access probability under different numbers of antennas at the BS, different SNR and SINR threshold is conducted. We also compare the proposed algorithm with the optimal algorithm. Simulation results reveal that the proposed access scheme offers significant improvement in both throughput and average packet delay as compared to optimized random access using the same beamforming technique. Moreover, the proposed low-complexity algorithm can approach the performance of the optimal algorithm when the number of antennas is sufficiently large.

APPENDIX A
OPTIMAL ALGORITHM

From (18), it is clear that for a given N_{gb}, maximum total throughput is obtained when both S_{gb} and S_{gf} are maximized. S_{gf} is globally maximized by solving problem (19). Therefore, optimum solution of (18) can be achieved by optimally solving for $\mathbf{X}$ instead of using Algorithm 2.

From (6) and (7), it is clear that $\mathcal{P}(\gamma_{\text{zf}}^{d} \geq \gamma_{\text{th}})$ is non-continuous and has the decision variable $\mathbf{X}$ in its sum bounds. Therefore, S_{gb} is non-convex even if the integer condition of $\mathbf{X}$ is relaxed. To overcome this problem, we use the constraint that the expected number of devices in each time slot should not exceed a certain threshold, $N_{\text{gb,max}}$, beyond which the interference will be too high that the expected throughput of the time slot decreases. Using this constraint, an optimization problem to select and schedule devices in the grant-based segment for a given N_{gb} is formulated as follows

$$\begin{aligned} \underset{\mathbf{Y}}{\text{maximize}} \ & \sum_{j=1}^{N_{\text{gb}}} \sum_{d=1}^{D} \theta_d y_{d,j} \\ \text{subject to: } & \text{C25.1: } \sum_{j=1}^{N_{\text{gb}}} y_{d,j} \leq 1, \quad \forall d, \\ & \text{C25.2: } \sum_{d=1}^{D} y_{d,j} \leq N_{\text{P}}, \quad \forall j, \\ & \text{C25.3: } \sum_{d=1}^{D} \theta_d y_{d,j} \leq U_{\text{gb,max}}, \quad \forall j, \\ & \text{C25.4: } y_{d,j} \in \{0,1\}, \quad \forall d, j, \end{aligned} \tag{25}$$

where $\mathbf{Y} : D \times N_{\text{gb}}$ corresponds to the first N_{gb} columns of $\mathbf{X}$ and $y_{d,j}$ denotes the element at row d and column j of $\mathbf{Y}$. Condition 25.1 ensures that each device is allocated at most one time slot. Condition 25.2 ensures that at most N_{P} devices are allocated to a time slot. Condition 25.3 limits the total expected number of devices in each time slot to $U_{\text{gb,max}}$ so that the throughput of that time slot does not decrease due to high interference. $U_{\text{gb,max}}$ is obtained by solving the following optimization problem

$$\begin{aligned} \underset{U_{\text{gb}}}{\text{maximize}} \ & \mathcal{P}(\gamma_d \geq \gamma_{\text{th}}) U_{\text{gb}} \\ \text{subject to: } & \text{C26.1: } 0 \leq U_{\text{gb}} \leq N_{\text{P}}, \\ & \text{C26.2: } U_{\text{gb}} \text{ is integer.} \end{aligned} \tag{26}$$

In (26), depending on the type of beamforming technique used, $\mathcal{P}(\gamma_d \geq \gamma_{\text{th}})$ is calculated by (6) and (7) with $K_s = U_{\text{gb}} - 1$. This problem is one-dimensional so it can be solved by hill-climbing method or exhaustive search. After obtaining $U_{\text{gb,max}}$, (25) becomes a mixed integer linear programming, which can be solved by many solvers. In this work, we use MOSEK [32] to solve it.

The optimal algorithm to solve (18) is summarized in Algorithm 3.

Algorithm 3 Optimal Algorithm for Solving (18)

```
Initialize $N_{gb} = 0$, $S = 0$, $S' = 0$, $\mathbf{X}' = 0$, $\mathbf{P}' = 0$
repeat
    Step 1: $S \leftarrow S'$, $\mathbf{X} \leftarrow \mathbf{X}'$, $\mathbf{P} \leftarrow \mathbf{P}'$, $N_{gb} \leftarrow N_{gb} + 1$
    Step 2: Obtaining $U_{gb,max}$ by solving (26) then solve for
    $\mathbf{Y}$ from problem (25). Set the first $N_{gb}$ columns of $X'$ to
    $\mathbf{Y}$ and other columns to zeros.
    Step 3: Solve $\mathbf{P}'$ by the method in section IV-B
        1) Solve (23) to obtain the initial value of $U_{gf}$
        2) Use the obtained initial $U_{gf}$ to solve (22) by
    hill-climbing integer search to obtain the optimal value
    of $U_{gf}$
        3) Solve (24) using the optimal value of $U_{gf}$
    Step 4: $S' \leftarrow S_{gb} + S_{gf}$, $\Delta S \leftarrow S' - S$
until $\Delta S < 0$ or $N_{gb} = N_{gb,max}$
if $\Delta S < 0$ then
    return $\mathbf{X}$ and $\mathbf{P}$
else
    return $\mathbf{X}'$ and $\mathbf{P}'$
end if
```

REFERENCES

[1] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia, "Machine learning for wireless communications in the Internet of Things: A comprehensive survey," *Ad Hoc Netw.*, vol. 93, Oct. 2019, Art. no. 101913.

[2] Q. Zhang, S. Jin, and H. Zhu, "A hybrid-grant random access scheme in massive MIMO systems for IoT," *IEEE Access*, vol. 8, pp. 88487–88497, 2020.

[3] H. Han, X. Guo, and Y. Li, "A high throughput pilot allocation for M2M communication in crowded massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9572–9576, Oct. 2017.

[4] S. Kusaladharma, G. Amarasuriya, W.-P. Zhu, and W. Ajib, "Rate analysis for NOMA in massive MIMO based stochastic cellular networks with pilot contamination," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[5] E. Bjornson, E. de Carvalho, J. H. Sorensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.

[6] E. Bjornson, E. de Carvalho, E. G. Larsson, and P. Popovski, "Random access protocol for massive MIMO: Strongest-user collision resolution (SUCR)," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[7] H. Han, Y. Li, and X. Guo, "A graph-based random access protocol for crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7348–7361, Nov. 2017.

[8] E. De Carvalho, E. Bjornson, J. H. Sorensen, P. Popovski, and E. G. Larsson, "Random access protocols for massive MIMO," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 216–222, May 2017.

[9] H. Han, Y. Li, and X. Guo, "User identity-aided pilot access scheme for massive MIMO-IDMA system," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6197–6201, Jun. 2019.

[10] J. C. Marinello, T. Abrão, R. D. Souza, E. de Carvalho, and P. Popovski, "Achieving fair random access performance in massive MIMO crowded machine-type networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 503–507, 2020.

[11] J. C. Marinello and T. Abrao, "Collision resolution protocol via soft decision retransmission criterion," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4094–4097, Apr. 2019.

[12] N. H. Mahmood, H. Alves, O. A. Lãpez, M. Shehab, D. P. M. Osorio, and M. Latva-Aho, "Six key features of machine type communication in 6G," in *Proc. 2nd 6G Wireless Summit*, 2020, pp. 1–5.

[13] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.

[14] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.

[15] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

[16] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, Jun. 2018.

[17] Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu, "Covariance based joint activity and data detection for massive random access with massive MIMO," in *Proc. ICC - IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[18] Z. Chen and W. Yu, "Phase transition analysis for covariance based massive random access with massive MIMO," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 36–40.

[19] L. Li, W. Meng, and C. Li, "Compressed sensing based semidefinite relaxation detection algorithm for overloaded uplink multiuser massive MIMO system," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[20] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020.

[21] J. Ding, D. Qu, H. Jiang, and T. Jiang, "Success probability of grant-free random access with massive MIMO," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 506–516, Feb. 2019.

[22] E. de Carvalho, E. Bjornson, J. H. Sorensen, E. G. Larsson, and P. Popovski, "Random pilot and data access in massive MIMO for machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7703–7717, Dec. 2017.

[23] J. Gao, Y. Wu, and F. Wei, "Random pilot and data access for massive MIMO spatially correlated Rayleigh fading channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[24] H. Jiang, D. Qu, J. Ding, and T. Jiang, "Multiple preambles for high success rate of grant-free random access with massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4779–4789, Oct. 2019.

[25] A. Rajandekar and B. Sikdar, "A survey of MAC layer issues and protocols for Machine-to-Machine communications," *IEEE Internet Things J.*, vol. 2, no. 2, pp. 175–186, Apr. 2015.

[26] O. A. Amodu and M. Othman, "A survey of hybrid MAC protocols for machine-to-machine communications," *Telecommun. Syst.*, vol. 69, no. 1, pp. 141–165, Sep. 2018, doi: 10.1007/s11235-018-0434-4.

[27] A. Laya, C. Kalalas, F. Vazquez-Gallego, L. Alonso, and J. Alonso-Zarate, "Goodbye, ALOHA!," *IEEE Access*, vol. 4, pp. 2029–2044, 2016.

[28] J. Yuan, H. Shan, A. Huang, T. Q. S. Quek, and Y. Yao, "Massive machine-to-machine communications in cellular network: Distributed queueing random access meets MIMO," *IEEE Access*, vol. 5, pp. 2981–2993, 2017.

[29] A. D. Shoaei, M. Derakhshani, and T. Le-Ngoc, "Reconfigurable and traffic-aware MAC design for virtualized wireless networks via reinforcement learning," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5490–5505, Aug. 2019.

[30] M. Ke, Z. Gao, Y. Wu, X. Gao, and K.-K. Wong, "Massive access in cell-free massive MIMO-based Internet of Things: Cloud computing and edge computing paradigms," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 756–772, Mar. 2021.

[31] L. G. Khachiyan, "Polynomial algorithms in linear programming," *USSR Comput. Math. Math. Phys.*, vol. 20, no. 1, pp. 53–72, Jan. 1980.

[32] *The MOSEK Optimization Toolbox for MATLAB Manual*, MOSEK Aps, Copenhagen, Denmark, 2019. [Online]. Available: http://docs.mosek.com/9.0/toolbox/index.html

ATOOSA DALILI SHOAEI received the B.Sc. degree in information technology engineering from the Isfahan University of Technology, Isfahan, Iran, in 2009, the M.Sc. degree in information technology engineering from the Amirkabir University of Technology, Tehran, Iran, in 2012, and the Ph.D. degree in electrical engineering from McGill University, in 2019. She is currently a Postdoctoral Fellow with McGill University. Her current research interests include medium access control techniques, the Internet of Things (IoT), and wireless virtualization.

DUC TUONG NGUYEN received the B.Eng. degree in electrical and electronics engineering from the University of Danang–University of Science and Technology, Danang, Vietnam, in 2017. He is currently pursuing the M.Eng. degree with the Department of Electrical and Computer Engineering, McGill University, Canada. His main research interests include multiple access in machine-type communications and the applications of machine learning into multiple access.

THO LE-NGOC (Life Fellow, IEEE) received the B.Eng. degree (Hons.) in electrical engineering and the M.Eng. degree from McGill University, Montréal, QC, Canada, in 1976 and 1978, respectively, and the Ph.D. degree in digital communications from the University of Ottawa, Canada, in 1983. From 1977 to 1982, he was a Research and Development Senior Engineer with Spar Aerospace Ltd., Sainte-Anne-de-Bellevue, Canada, where he was involved in the development and design of satellite communications systems. From 1982 to 1985, he was an Engineering Manager with the Radio Group, Department of Development Engineering, SR Telecom 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 includes 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 was a recipient of the 2004 Canadian Award in Telecommunications Research, and the IEEE Canada Fessenden Award 2005. He is a Distinguished James McGill Professor.

• • •