Differentiated Service-Aware Group Paging for Massive Machine-Type Communication

Wei Cao, *Student Member, IEEE*, Alex Dytso, *Member, IEEE*, Gang Feng, *Senior Member, IEEE*, H. Vincent Poor, *Fellow, IEEE*, and Zhi Chen, *Senior Member, IEEE*

***Abstract*— Massive machine-type communication (mMTC) has been identified as one of the three generic 5G services, with the aim of providing connectivity to a large number of devices. The concurrent massive access in mMTC may lead to congestion due to limited access resources. Group paging (GP) is emerging as one of the promising solutions to alleviate network congestion by controlling access load. However, the performance of GP deteriorates drastically with the number of devices per paging group. This paper explores GP with a pre-backoff strategy for a general mMTC scenario in which devices are allowed to have diverse access success probability (ASP) requirements, and proposes a differentiated service-aware GP scheme with specific pre-backoff times (GPSP), with the aim of maximizing the total access rate while guaranteeing the ASP requirements for individual devices. An optimal solution to the GPSP problem is derived to provide a performance upper bound. As low-complexity algorithms are of key importance for mMTC applications, an efficient heuristic algorithm is further designed. Numerical results demonstrate that the proposed GPSP scheme can effectively improve the system performance in terms of average ASP, average access delay, and the average number of preamble transmissions.**

***Index Terms*— Massive MTC, random access, group paging, pre-backoff, differentiated services.**

I. Introduction

WITHIN the emerging Internet of Things paradigm, machine-type communication (MTC) enables a broad range of applications in 5G networks, such as mission-critical services, intelligent transportation systems and massive deployment of autonomous devices [1], [2]. It has been predicted that the number of MTC devices will be significantly larger than that of human type communication devices (e.g., smartphones) [1], [2]. Moreover, massive MTC (mMTC), with a focus on providing connectivity to a large number of MTC devices, has been identified as one of the three generic 5G services [3]. For mMTC, however, the concurrent massive access of devices to the radio network may easily lead to congestion and system overload in both radio access network (RAN) and core network (CN) parts [4]. Hence, effective overload control mechanisms are essential to support mMTC service while guaranteeing an adequate quality of service for devices.

Overload control of the uplink random access channel (RACH) in a RAN is one of the fundamental working issues of the 3GPP group [5]. Existing overload control schemes in RANs can be categorized into push- and pull-based approaches. On the one hand, in push-based approaches (e.g., access class barring, separate RACH resources for MTC, slotted access, and MTC-specific backoff schemes, etc.), traffic is pushed from devices to the network without any restrictions until the RAN overloads. The push-based approach can efficiently relieve the overload problem in RANs, but it cannot alleviate the congestion in the CN [7]. On the other hand, the pull based approach could effectively control the access load and prevent overload and congestion in both the RAN and CN, by using paging messages to activate specific devices to access the network. However, the legacy paging mechanism does not work well for mMTC, as it was originally designed for small-scale human-type communications. Each paging message can only page up to sixteen devices, and there are only two paging occasions per radio frame [6]. As one-by-one paging will lead to long paging delay in mMTC applications, group paging (GP), which uses a single group-paging message to activate a group of devices, has become a promising new solution [6].

In GP, devices are organized into groups, and each group is assigned a unique group identity. After joining a group, devices monitor the downlink control channel for the GP message. Upon receiving a GP message with the matched group identity, the group of devices start the random access (RA) procedure immediately [7], [8]. Unfortunately, the access rate of GP dramatically degrades as the number of devices being paged increases [9]. Intuitively, it is beneficial to scatter the devices of the group being paged over an available time access interval instead of letting them start the RA procedure simultaneously [9], [10]. Specifically, the system forces the devices to wait

Manuscript received February 12, 2018; revised June 2, 2018; accepted June 7, 2018. Date of publication June 19, 2018; date of current version November 16, 2018. This work was supported in part by the National Natural Science Foundation of China Major Project under Grant 61631004, the Chinese Fundamental Research Funds for the Central Universities under Grant ZYGX2015Z005, and the U.S. National Science Foundation under Grants CNS-1702808 and ECCS-1647198. The associate editor coordinating the review of this paper and approving it for publication was S. Coleri Ergen. *(Corresponding author: Wei Cao.)*

W. Cao and G. Feng are with the National Key Laboratory of Science and Technology on Communications and the Center for Intelligent Networking and Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: weicao@princeton.edu; fenggang@uestc.edu.cn).

A. Dytso and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: adytso@princeton.edu; poor@princeton.edu).

Z. Chen is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: chenzhi@uestc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2018.2848940

for different pre-backoff times before their first transmission attempts. Hence, the pre-backoff GP optimization problem consists of allocating appropriate pre-backoff times within the access interval to individual devices in the paging group. As the pre-backoff time allocation directly affects the total access rate and individual access success probabilities (ASPs), the allocation mechanism for pre-backoff GP should be carefully designed to maximize the access resource utilization.

The term MTC covers a wide range of use cases and applications with diverse service requirements in terms of ASP, transmission rate, etc. For example, the devices of intelligent transportation systems may require an ultra-high ASP, while the devices of environmental monitoring systems would be much more tolerant. Therefore, in this study, we introduce the ASP requirement as a new dimension to the pre-backoff GP problem. This new formalism guarantees that the ASP requirements of the individual devices being served are satisfied.

In this paper, we propose a scheduling scheme to determine the pre-backoff times for individual devices according to the estimated achievable ASP. We refer to this scheme as GP with specific pre-backoff times (GPSP). Specifically, the optimal GPSP maximizes the total access rate while guaranteeing the ASP requirement of each device being served. Note that the GPSP problem is indeed a generalized GP problem since the conventional GP problem [7]–[10] can be considered to be a special case in which the ASP requirements of all devices are identical. Naturally, the methods used to solve the conventional GP problem are not applicable for solving the GPSP problem. Moreover, the pre-backoff is performed before the devices encounter collisions, which differs from the random access backoff schemes studied in [11]–[13].

Our contributions are as follows:

- A generalized GP problem for mMTC applications with the new dimension of ASP requirements is investigated, which enables the network to provide differentiated services;
- An optimal algorithm is developed by carefully studying the properties of the GPSP problem; and
- According to the structure of the derived optimal solution, an efficient low-complexity algorithm is further developed.

The rest of the paper is organized as follows. Section II presents a brief overview of the related work. Section III introduces basic background about the random access procedure and the ASP estimation model. Based on this model, the GPSP optimization problem is formulated. Section IV derives an optimal solution to the GPSP problem, and proposes a low-complexity algorithm. Numerical results as well as discussion are given in Section V, followed by concluding remarks in Section VI.

II. Related Work

Massive concurrent access to a wireless network is prone to congestion or/and signaling overload in the RAN and CN. To address this problem, most of the existing research focuses on improving the performance of access control in the RAN, from either the network or the user point of view.

Access control on the network side is mainly based on the push- and pull-based mechanisms proposed by 3GPP [5]. On the one hand, the push-based mechanisms include access class barring (ACB) schemes [14], dynamic allocation of RACH resources [5], MTC specific backoff schemes [15], and slotted access [16]. Besides the basic individual ACB schemes studied in [17]–[20], dynamic ACB schemes and cooperative ACB schemes are also investigated in [21]–[23]. Pang *et al.* [24] provided a game-theoretic approach that allocates the preambles for human-type and machine-type communications. A resource allocation scheme was proposed in [25] to prevent the evolved node B (eNB) from allocating the physical uplink shared channel to the collided preambles. To obtain further performance gain, Oh *et al.* [26] provided a joint scheme combining ACB and RACH resource allocation. The performance of uniform backoff and binary exponential backoff algorithms was examined in [12]. Yang *et al.* [13] studied the performance of a random access backoff algorithm under different parameter settings, while taking into account the physical loss.

On the other hand, in the pull-based mechanisms, the network can control the access load via paging. As aforementioned, one-by-one paging cannot adapt to the mMTC service, and GP has emerged as a promising solution. Wei *et al.* [8] presented an analytical model for GP with no pre-backoff, which is useful for further study of GP methods. As the performance of GP dramatically degrades with the number of devices, Arouk *et al.* [9] devised a uniform pre-backoff GP method that uniformly partitions the devices of one paging group into smaller access groups of the same size. In addition, each access group starts the RA process at different RA slots in the access interval. A random pre-backoff GP method was proposed in [10], allowing the devices to uniformly choose a pre-backoff time over the access interval. The pre-backoff GP methods in [9] and [10] effectively improve the GP performance. However, in both methods, devices are treated in the same way, without taking ASP requirement dimension into account.

There are various access control mechanisms from the user perspective as well, such as data aggregation and access point optimization. Guo *et al.* [27] considered a cellular-based mMTC network in which the devices transmit data to some aggregators, and the aggregated data is then relayed to base stations. Moreover, the aggregators not only aggregate data, but also schedule resources to devices, and hence improve the average ASP. Furthermore, the numerical results show that the provision of more resources at the aggregation phase is not always beneficial to the mMTC performance. Lin and Chen [28] studied the problem of constructing maximum-lifetime data aggregation trees in wireless sensor networks, which was shown to be NP-hard. Further, Lin and Chen [28] provided an approximation algorithm that constructs a data aggregation tree whose inverse lifetime is guaranteed to be within a bound from that of the optimal solution. In [29], an optimal device-to-device communication assisted access scheme was developed, in which devices can either access the network directly or via relaying by an idle device. The system proposed in [29] can achieve better load

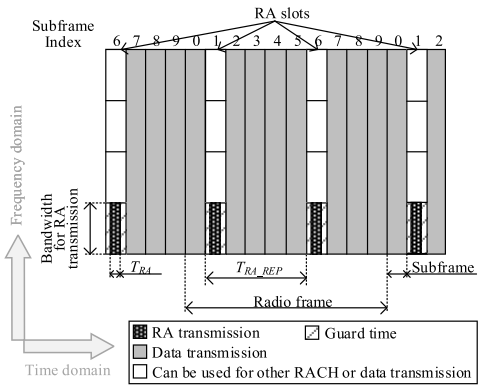


Fig. 1. An example of time-frequency mapping of physical random access transmissions of LTE.

balance and hence higher access rate with the assistance of device-to-device communications.

III. System Model and Problem Formulation

We first briefly provide some preliminary background on RA transmissions. Then, we introduce a probabilistic model to estimate the achievable ASP for individual devices with given pre-backoff times. Next, we formulate the GPSP optimization problem based on this model.

A. Preliminary Background

To access a wireless network, devices should first perform RA transmissions at specific RA occasions using RACH. Fig. 1 illustrates an example of time-frequency mapping of the physical RA transmissions [11]. In the time domain, time is divided into frames and each frame consists of multiple subframes. For an orthogonal separation between the RA and data transmissions, the RA transmissions are restricted to certain subframes, which are called RA slots. LTE defines 64 possible physical RACH (PRACH) configuration indices [12], which further determine the indices of the subframes that are assigned to the RA transmissions. In accordance with the assumption used in [5], [8], and [9], we use RACH configuration index 6 (*i.e.*, the subframe indices for RA transmissions are 1 and 6 [12] where the RA slot period is $T_{\text{RA\_REP}} = 5$ subframes). Note that some portions of RA slots are reserved as guard time to account for the timing uncertainty in the uplink [11]. In the frequency domain, there could be several frequency bands. According to [9], let there be just one frequency band in each RA slot for the RA transmission. Denote by R the number of preambles in an RA slot, and then the number of RA opportunities per RA slot is also R [13]. It is supposed that a collision occurs only if two or more devices make RA transmissions in the same RA time slot and the same frequency band using the same preamble signature [13].

The RA procedure, with the aim to obtain uplink synchronization and transmission resource assignment [11], includes several signaling steps, as shown in Fig. 2. In the procedure, once a device receives the GP message with the matched identity, it starts its pre-backoff timer. After the pre-backoff timer expires, the device starts to transmit a randomly

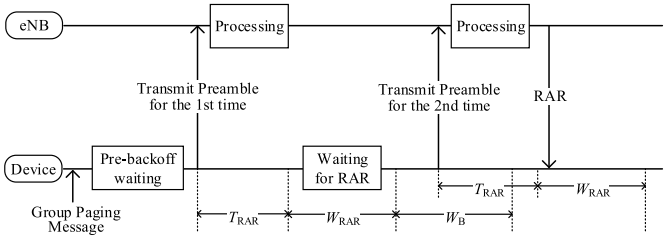


Fig. 2. An example of the RA transmission procedure of one device.

TABLE I
Parameters and Variables

Notation	Parameter
$T_{\text{RA\_REP}}$	The interval between two RA subframes
R	The number of RA preamble signatures
T_{RAR}	The processing time required by the eNB to detect a preamble
W_{RAR}	The RAR window size
N	The maximum number of preamble transmissions
W_B	The backoff window size
I	The number of RA slots in the access interval
Notation	**Variable**
$\mathbf{T}$	$\mathbf{T} = [t_{im}]_{I\times M}$ is the pre-backoff time matrix where $t_{im} = 1_{\{\mathcal{A}^m_{i,1}\}}$
$\mathbf{t}_m$	$\mathbf{t}_m$ is the pre-backoff time I-vector of d_m, such that $\mathbf{T} = [\mathbf{t}_1\ \mathbf{t}_2\ \cdots\ \mathbf{t}_M]$

selected preamble using the reserved PRACH. The eNB needs T_{RAR} processing time to detect the preambles. After the eNB receives the preamble, it transmits a response message, which carries a medium access control header and random access responses (RARs), on the physical downlink shared channel. The header may carry a backoff indicator for the collided or undetected devices [31]. The RAR carries the information for further signaling transmission and the identity of the preamble transmitted by the device. The device expects to receive the RAR message within an RAR window, denoted by W_{RAR}. If the device fails in receiving the RAR carrying the corresponding identity, the device will select a random backoff time according to a uniform distribution on $[0, B]$ where B is the backoff indicator value (thus, the backoff window size is $W_{\text{B}} = B+1$) and will retransmit the preamble after the backoff time [31]. Once the device successfully receives the RAR, it adjusts the uplink synchronization and performs further signaling to prepare for data transmission. For concise system modeling, we summarize the related system parameters and the variables in Table 1.

B. ASP Estimation Model

This study considers a paging group of M devices $\mathcal{D} = \{d_1, d_2, \cdots, d_M\}$ in a paging area. The basic idea of the GPSP scheme is to scatter the devices of a paging group into several access groups with different pre-backoff times within the access interval. Let there be I RA slots in the access interval reserved for group $\mathcal{D}$ (i.e., the access interval starts from the first slot after receiving the paging message and ends in slot I). Since we focus on mMTC scenarios, we have that $I \ll M$. For clarity, we define the following events:

- $\mathcal{A}_{i,n}^m$: device d_m performs its n-th attempt in slot i;
- $\mathcal{X}_{i,n}^m$: device d_m succeeds in its n-th attempt in slot i; and
- $\mathcal{Y}_{i,n}^m$: device d_m fails in its n-th attempt in slot i.

Clearly, $\mathcal{X}_{i,n}^m \cup \mathcal{Y}_{i,n}^m = \mathcal{A}_{i,n}^m$. Denote by $\mathbf{T} = [t_{im}]_{I\times M}$ the pre-backoff time matrix where

$$t_{im} = \mathbf{1}_{\{\mathcal{A}_{i,1}^m\}}, \tag{1}$$

and $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Denote by $\mathbf{t}_m = [t_{im}]_{I\times 1}$ the pre-backoff time vector of d_m. Since each device can choose at most one RA slot for its first attempt, we have that

$$\|\mathbf{t}_m\|_1 \le 1, \quad \forall m, \tag{2}$$

where $\|\mathbf{t}_m\|_1 = \sum_i t_{im}$. Note that $\|\mathbf{t}_m\|_1 = 1$ indicates that d_m transmit in the access interval, and $\mathbf{t}_m = \mathbf{0}$ (or equivalently $\|\mathbf{t}_m\|_1 = 0$) indicates that d_m does not transmit.

We denote a probability measure that describes this model by $\mathbb{P}(\cdot)$. Let N be the maximum number of transmission attempts that each device is allowed. The ASP of device d_m can be expressed as

$$\begin{aligned} P_m &\triangleq \mathbb{P}\left(\bigcup_{i,n} \mathcal{X}_{i,n}^m\right) \\ &\overset{a)}{=} \sum_{i,n} \mathbb{P}(\mathcal{X}_{i,n}^m) \\ &\overset{b)}{=} \sum_{i,n} \mathbb{P}(\mathcal{A}_{i,n}^m \cap \mathcal{X}_{i,n}^m) \\ &= \sum_{i,n} \mathbb{P}(\mathcal{A}_{i,n}^m)\mathbb{P}(\mathcal{X}_{i,n}^m|\mathcal{A}_{i,n}^m), \end{aligned} \tag{3}$$

where the labeled equalities follow from: a) the fact that $\mathcal{X}_{i,n}^m$ are disjoint events; and b) $\mathcal{X}_{i,n}^m \subseteq \mathcal{A}_{i,n}^m$.

Denote by M_i the average number of devices that transmit preambles in slot i, and observe that

$$\begin{aligned} M_i &\triangleq \mathbb{E}\left[\sum_{m=1}^{M}\sum_{n=1}^{N} \mathbf{1}_{\{\mathcal{A}_{i,n}^m\}}\right] \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} \mathbb{E}\left[\mathbf{1}_{\{\mathcal{A}_{i,n}^m\}}\right] \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} \mathbb{P}(\mathcal{A}_{i,n}^m). \end{aligned} \tag{4}$$

According to the derivations in [32], the success probability and collision probability of each transmission attempt depend only on the average number of devices that transmit in that slot, and are respectively given by

$$\mathbb{P}(\mathcal{X}_{i,n}^m|\mathcal{A}_{i,n}^m) = e^{-\frac{M_i}{R}}, \tag{5}$$

$$\mathbb{P}(\mathcal{Y}_{i,n}^m|\mathcal{A}_{i,n}^m) = 1 - e^{-\frac{M_i}{R}}. \tag{6}$$

Note that, we do not take into account the power-ramping effect [5], as the power-ramping effect actually encourages retransmissions.

For all devices the probability of the first attempt depends only on the pre-backoff time, *i.e.*,

$$\mathbb{P}(\mathcal{A}_{i,n=1}^m) = t_{im}. \tag{7}$$

For $n > 1$, a device performs the n-th attempt if and only if its $(n-1)$-th attempt fails. The probability that a device, which fails in its $(n-1)$-th attempt in slot j, performs its n-th attempt in slot i is given in [9] by

$$\begin{aligned} &\mathbb{P}(\mathcal{A}_{i,n}^m|\mathcal{Y}_{j,n-1}^m) \\ &= q_{ji} = \begin{cases} \alpha_1, & \text{if } i = j + \left\lceil \dfrac{T_{\text{RAR}} + W_{\text{RAR}}}{T_{\text{RA\_REP}}} \right\rceil \\ \alpha_2, & \text{if } j + \left\lceil \dfrac{T_{\text{RAR}} + W_{\text{RAR}}}{T_{\text{RA\_REP}}} \right\rceil < i \le j \\ & \quad + \left\lfloor \dfrac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{B}}}{T_{\text{RA\_REP}}} \right\rfloor \\ \alpha_3, & \text{if } i = j + \left\lfloor \dfrac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{B}}}{T_{\text{RA\_REP}}} \right\rfloor + 1 \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{8}$$

where

$$\begin{aligned} \alpha_1 &= \frac{\left\lceil \frac{T_{\text{RAR}}+W_{\text{RAR}}}{T_{\text{RA\_REP}}} \right\rceil T_{\text{RA\_REP}} - (T_{\text{RAR}} + W_{\text{RAR}})}{W_{\text{B}}}, \\ \alpha_2 &= \frac{T_{\text{RA\_REP}}}{W_{\text{B}}}, \\ \alpha_3 &= \frac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{B}}}{W_{\text{B}}} \\ &\quad - \frac{T_{\text{RA\_REP}}}{W_{\text{B}}} \left\lfloor \frac{T_{\text{RAR}} + W_{\text{RAR}} + W_{\text{B}}}{T_{\text{RA\_REP}}} \right\rfloor. \end{aligned}$$

Since $\mathcal{A}_{i,n}^m \cap \mathcal{A}_{j,n}^m = \varnothing, \forall i \ne j$, we have that

$$\begin{aligned} &\mathbb{P}(\mathcal{A}_{i,n>1}^m) \\ &\overset{a)}{=} \sum_{j=1}^{i-1} \mathbb{P}(\mathcal{Y}_{j,n-1}^m)\mathbb{P}(\mathcal{A}_{i,n}^m|\mathcal{Y}_{j,n-1}^m) \\ &\overset{b)}{=} \sum_{j=1}^{i-1} \mathbb{P}(\mathcal{A}_{j,n-1}^m)\mathbb{P}(\mathcal{Y}_{j,n-1}^m|\mathcal{A}_{j,n-1}^m)\mathbb{P}(\mathcal{A}_{i,n}^m|\mathcal{Y}_{j,n-1}^m) \\ &= \sum_{j=1}^{i-1} \mathbb{P}(\mathcal{A}_{j,n-1}^m)(1 - e^{-\frac{M_j}{R}})q_{ji}, \end{aligned} \tag{9}$$

where the labeled equalities follow from: a) the law of total probability; and b) $\mathbb{P}(\mathcal{Y}_{j,n-1}^m) = \mathbb{P}(\mathcal{Y}_{j,n-1}^m \cap \mathcal{A}_{j,n-1}^m)$. Therefore, once the pre-backoff matrix $\mathbf{T}$ is determined, the ASP of each device can be derived from (3)-(9).

C. Problem Formulation of the GPSP Optimization

Constrained by limited resources, the system is unable to admit all devices within the given access interval. To maximize the total access rate, the system needs to carefully choose a fraction of devices to serve and schedule their pre-backoff times. Denote by p_m the minimum ASP required by device d_m (p_m is a given system parameter, and $p_m \in (0,1)$). Clearly, the larger p_m is the more resources d_m needs. Therefore, an incentive measure should be introduced to encourage the network to serve the devices with larger p_m values. In this work, we assign p_m to be the weight of d_m. Recall that $\mathbf{1}_{\{\|\mathbf{t}_m\|_1=1\}}$ indicates that d_m is chosen to transmit. Let $D = \mathbf{1}_{\{\|\mathbf{t}_X\|_1=1\}}$ be a random variable $\mathcal{D} \to \{0,1\}$ with

probability $\mathbb{P}(D = \mathbf{1}_{\{\|\mathbf{t}_x\|_1=1\}}) = \frac{p_x}{\sum_m p_m}$. The total access rate can be defined as follows:

$$E \triangleq \mathbb{E}\left[\mathbf{1}_{\{\|\mathbf{t}_X\|_1=1\}}\right] = \frac{\sum_{m=1}^{M} p_m \|\mathbf{t}_m\|_1}{\sum_{m=1}^{M} p_m}. \tag{10}$$

Note that $\sum_{m=1}^{M} p_m$ is a constant, and thus we let $E = \sum_{m=1}^{M} p_m \|\mathbf{t}_m\|_1$. Then, the GPSP optimization problem can be formulated as follows:

$$(\mathcal{P}1) \quad \max_{\mathbf{T}} \ E = \sum_{m=1}^{M} p_m \|\mathbf{t}_m\|_1, \tag{11a}$$

$$s.t.\ P_m \geq p_m \mathbf{1}_{\{\|\mathbf{t}_m\|_1=1\}} = p_m \|\mathbf{t}_m\|_1, \quad \forall m, \tag{11b}$$

$$\mathbf{t}_m \in \{\mathbf{0}, \mathbf{e}_1, \cdots, \mathbf{e}_I\}, \tag{11c}$$

where (11b) guarantees that the ASP requirements of the devices being chosen to serve are satisfied, and $\mathbf{e}_i$ in (11c) is the binary column I-vector with a single 1 at the i-th entry.

Next, let us show that the GPSP problem can be reduced from the I-Dimensional Knapsack problem [35], which is computationally harder than the well-known NP-Complete Knapsack problem [36] and does not have an efficient polynomial-time approximation scheme (EPTAS) unless P=NP [37]. Therefore, we can conclude that there is no polynomial time algorithm to the optimal solution or EPTAS unless P=NP [35].

Theorem 1: I-Dimensional Knapsack problem $\leq_P$ GPSP problem (where $\leq_P$ denotes the polynomial-time reduction).

Proof: See Appendix A. ∎

IV. Solutions

A. Optimal Algorithm

To avoid using brute force for an optimal solution, let us first examine the properties of the GPSP problem. Intuitively, once the devices are assigned with the same pre-backoff time, the ASPs of the devices are the same. Then, we have Lemma 1.

Lemma 1: Suppose $d_m, d_{m'} \in \mathcal{D}$, and $\mathbf{t}_m = \mathbf{t}_{m'}$. Then for all i, n, $\mathbb{P}(\mathcal{A}_{i,n}^m) = \mathbb{P}(\mathcal{A}_{i,n}^{m'})$, $\mathbb{P}(\mathcal{X}_{i,n}^m) = \mathbb{P}(\mathcal{X}_{i,n}^{m'})$, $\mathbb{P}(\mathcal{Y}_{i,n}^m) = \mathbb{P}(\mathcal{Y}_{i,n}^{m'})$, and $P_m = P_{m'}$.

Proof: See Appendix B. ∎

For simplicity, denote by $\mathcal{G}_i = \{d_m | \mathbf{t}_m = \mathbf{e}_i\}$ the set of devices assigned to slot i for their first transmission attempts. Determining the pre-backoff matrix $\mathbf{T}$ is equivalent to determining $\mathcal{G}_i$'s. Let P_i be the ASP of devices that are assigned to slot i for their first transmission (*i.e.*, $P_m = P_i, \forall d_m \in \mathcal{G}_i$). Denote by $p_i = \max\{p_m : d_m \in \mathcal{G}_i\}$ the requirement of slot i. Then constraints (11b) can be re-written as

$$P_i \geq p_i, \quad \forall i. \tag{12}$$

Next, as intuition suggests, we show that it is preferable to allocate the devices with similar requirements into the same RA slot.

Theorem 2: Suppose $d_a, d_b, d_c \in \mathcal{D}$, $p_a \geq p_b \geq p_c$, and there exists a feasible solution $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_M]$ where $\mathbf{t}_a = \mathbf{t}_c = \mathbf{e}_j$ and $\mathbf{t}_b \neq \mathbf{e}_j$ (or equivalently $d_a, d_c \in \mathcal{G}_j$, and $d_b \notin \mathcal{G}_j$). Then, interchanging the pre-backoff time of d_b and d_c (i.e., $\mathbf{T}' = [\mathbf{t}'_1, \cdots, \mathbf{t}'_M]$ where $\mathbf{t}'_b = \mathbf{t}_c$, $\mathbf{t}'_c = \mathbf{t}_b$, and $\mathbf{t}'_m = \mathbf{t}_m$, $\forall m \neq b$ and $m \neq c$,) does not reduce the total access rate (i.e., $E' \geq E$), and $\mathbf{T}'$ is also a feasible solution.

Proof: Since $\mathbf{T}$ is a feasible solution, from constraint (12) we have that $P_i \geq p_i, \forall i$. Interchanging the pre-backoff time of d_b and that of d_c does not change the number of devices that transmit in each slot, and therefore $P'_i = P_i, \forall i$. Moreover, due to the assumption $\mathbf{t}_a = \mathbf{t}_c = \mathbf{e}_j$, devices d_a and d_c are assigned to slot j. In addition, with the assumption $\mathbf{t}_b \neq \mathbf{e}_j$, we have two cases $\mathbf{t}_b \neq \mathbf{0}$ and $\mathbf{t}_b = \mathbf{0}$. Next we examine these two cases separately.

For the case $\mathbf{t}_b \neq \mathbf{0}$, we can assume that $d_b \in \mathcal{G}_l, l \neq j$. After interchanging the pre-backoff time of d_b and that of d_c, the set of devices in slot i is $\mathcal{G}'_j = \mathcal{G}_j \backslash \{d_c\} \cup \{d_b\}$ and $\mathcal{G}'_l = \mathcal{G}_l \backslash \{d_b\} \cup \{d_c\}$. With the assumption $p_a \geq p_b \geq p_c$ and the definition of p_i, we have that $p'_i \leq p_i, \forall i$. Therefore, $P'_i = P_i \geq p_i \geq p'_i, \forall i$, and thus $\mathbf{T}'$ (which is determined by $\mathcal{G}'_i, i = 1, \cdots, I$) is also a feasible solution. In this case, $E' = E$.

For the case $\mathbf{t}_b = \mathbf{0}$ (*i.e.*, d_b does not transmit), after interchanging the pre-backoff time of d_b and d_c, the set of devices in slot i is $\mathcal{G}'_l = \mathcal{G}_l \backslash \{d_b\} \cup \{d_c\}$. By the assumption $p_a \geq p_b \geq p_c$ and the definition of p_i, we have that $p'_i \leq p_i, \forall i$. Therefore, $P'_i = P_i \geq p_i \geq p'_i, \forall i$, and thus $\mathbf{T}'$ (which is determined by $\mathcal{G}'_i, i = 1, \cdots, I$) is also a feasible solution. In this case, $E' = E + p_b - p_c \geq E$.

This concludes the proof. ∎

Remark 1: We sort the devices according to p_m, so that $p_1 \geq p_2 \geq \cdots \geq p_M$. Assume that there exists a feasible solution $\mathbf{T}$ (or equivalently $\mathcal{G}_i, i = 1, \cdots, I$). Denote by $n_i = |\mathcal{G}_i|$ the group size, and $f_i = \min\{m : d_m \in \mathcal{G}_i\}$ the smallest index of the devices in each group. Then $\mathbf{T}'$ (or equivalently $\mathcal{G}'_i, i = 1, \cdots, I$) is also a feasible solution and $E' \geq E$, if $\mathcal{G}'_i = \{d_{f'_i}, d_{f'_i+1}, \cdots, d_{f'_i+n_i-1}\}, \forall i$, where $f'_i \in \bigcup_{i=1}^{I} \mathcal{G}_i, \forall i$.

Specifically, we construct $\mathbf{T}'$ by the following steps: a) Assume that the device with the smallest index within all groups is $m_1 = \min\{m : d_m \in \bigcup_{i=1}^{I} \mathcal{G}_i\}$, and $d_{m_1} \in \mathcal{G}_j$. Let $f'_j = m$; b) If $d_{f'_j+1} \notin \mathcal{G}_j$, then interchange the pre-backoff time of $d_{f'_j+1}$ and that of d_{l_j}, where $l_j = \max\{m : d_m \in \mathcal{G}_j\}$ is the greatest index. Update $\mathcal{G}'_i$'s; c) Repeat Step b) until $\mathcal{G}'_j = \{d_{f'_j}, d_{f'_j+1}, \cdots, d_{f'_j+n_i-1}\}$; and d) Put $\mathcal{G}'_j$ aside and repeat Step a). Assume that the smallest index of the devices is $m_2 = \min\{m : d_m \in \bigcup_{i \neq j} \mathcal{G}'_i\}$, and $d_{m_2} \in \mathcal{G}'_k$. Note that, the devices in $\mathcal{G}'_k$ come from either $\mathcal{G}_k$ or $\mathcal{G}_j$. Therefore, $f'_k = \in \bigcup_{i=1}^{I} \mathcal{G}_i$. Repeating the interchange operation, we have $\mathcal{G}'_i = \{d_{f'_i}, d_{f'_i+1}, \cdots, d_{f'_i+n_i-1}\}, \forall i$, where $f'_i \in \bigcup_{i=1}^{I} \mathcal{G}_i, \forall i$. Finally, from Theorem 2, we can conclude that all these operations do not reduce the total access rate, i.e., $E' \geq E$.

Therefore, an optimal solution is to select I disjoint subsets, $\mathcal{G}_1, \cdots, \mathcal{G}_I$, from $\mathcal{D}$ (note that some subsets can be empty so as not to omit any potential solutions). For each subset $\mathcal{G}_i$ we only need to select f_i and n_i, and the pre-backoff matrix can be given by $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_M]$, where

$$\mathbf{t}_m = \begin{cases} \mathbf{e}_i, & \text{if } m \in [f_i, f_i + n_i) \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Denote by $\mathbf{f} = [f_1, \cdots, f_I]$ the index vector and let $\mathbf{n} = [n_1, \cdots, n_I]$ be the group size vector. Next, we study the

possible values of $\mathbf{f}$ and $\mathbf{n}$. As the probability of the n-th attempt is small when n is large, we do not take into account the limitation of the maximum number of transmissions (*i.e.*, assuming that $N \geq I$). Then, intuitively the device assigned to an early slot (*i.e.*, slot $i \ll I$) has a higher chance of transmitting and hence has a higher ASP. The following theorem solidifies this intuition.

Theorem 3: For any $0 < j < k \leq I$, $P_j \geq P_k$.

Proof: According to Lemma 1 and the expression in (3),

$$\begin{aligned} P_j &= \sum_{n=1}^{N}\sum_{i=j}^{I}\mathbb{P}(\mathcal{A}_{i,n}^m)\mathbb{P}(\mathcal{X}_{i,n}^m|\mathcal{A}_{i,n}^m) \\ &= \sum_{i=j}^{k-1}\sum_{n=1}^{N}\mathbb{P}(\mathcal{A}_{i,n}^m)\mathbb{P}(\mathcal{X}_{i,n}^m|\mathcal{A}_{i,n}^m) \\ &\quad + \sum_{i=k}^{I}\sum_{n=1}^{N}\mathbb{P}(\mathcal{A}_{i,n}^m)\mathbb{P}(\mathcal{X}_{i,n}^m|\mathcal{A}_{i,n}^m) \\ &= \sum_{i=j}^{k-1}\sum_{n=1}^{N}\mathbb{P}(\mathcal{A}_{i,n}^m)\mathbb{P}(\mathcal{X}_{i,n}^m|\mathcal{A}_{i,n}^m) + P_k \geq P_k. \end{aligned} \quad (14)$$

This concludes the proof. ■

Remark 2: Assume that there exists a feasible solution $\mathcal{G}_i = \{d_{f_i}, d_{f_{i+1}}, \cdots, d_{f_i+n_i-1}\}$, $i = 1, \cdots, I$. *For any* $j < k$, *if* $f_j > f_k + n_k$, *there exists another feasible solution* $\mathcal{G}_i' = \{d_{f_i'}, d_{f_i'+1}, \cdots, d_{f_i+n_i-1}\}$, $i = 1, \cdots, I$ *with* $E' \geq E$, *such that* $1 \leq f_1' + n_1 \leq f_2' + n_2 \leq \cdots \leq f_I' + n_I \leq M$.

Proof: Since $f_j > f_k + n_k$, we have that $p_{f_k} \geq \cdots \geq p_{f_k+n_k} \geq p_{f_j} \geq \cdots \geq p_{f_j+n_j-1}$. From $p_i = \max_{d_m \in \mathcal{G}_i} p_m$, we have that $p_j = p_{f_j}$ and $p_k = p_{f_k}$. From constraint (12), we have that $P_j \geq p_j$ and $P_k \geq p_k$. From Theorem 3, the ASPs of the devices in slots j and k satisfy $P_j \geq P_k$, and hence $P_j \geq p_k$. Next, we examine the following three cases:

For the case $n_j > n_k$, we interchange the pre-backoff times of the devices in $\{d_{f_j+n_j-n_k}, \cdots, d_{f_j+n_j-1}\}$ and $\mathcal{G}_k$ respectively (*i.e.*, $\mathcal{G}_j' = \{d_{f_k}, \cdots, d_{f_k+n_k-1}, d_{f_j}, \cdots, d_{f_j+n_j-n_k-1}\}$ and $\mathcal{G}_k' = \{d_{f_j+n_j-n_k}, \cdots, d_{f_j+n_j-1}\}$). Thus $p_j' = p_{f_k} = p_k$ and $p_k' = p_{f_j+n_j-n_k} \leq p_k$.

For the case $n_j = n_k$, we interchange the pre-backoff times of the devices in $\mathcal{G}_j$ and $\mathcal{G}_k$ respectively (*i.e.*, $\mathcal{G}_j' = \mathcal{G}_k$ and $\mathcal{G}_k' = \mathcal{G}_j$). Thus $p_j' = p_{f_k} = p_k$ and $p_k' = p_{f_j} \leq p_k$.

For the case $n_j < n_k$, we interchange the pre-backoff times of the devices in $\{d_{f_k}, \cdots, d_{f_k+n_j-1}\}$ and $\mathcal{G}_j$ respectively (*i.e.*, $\mathcal{G}_j' = \{d_{f_k}, \cdots, d_{f_k+n_j-1}\}$ and $\mathcal{G}_k' = \{d_{f_k+n_j}, \cdots, d_{f_k+n_k-1}, d_{f_j}, \cdots, d_{f_j+n_j-1}\}$). Thus $p_j' = p_{f_k} = p_k$ and $p_k' = p_{f_j+n_j} \leq p_k$.

Combining the three cases, we have that $p_i' \leq p_i, \forall i$. Moreover, the number of devices that transmit in each slot does not change after interchanging the pre-backoff times of the devices in two sets, and thus $P_i' = P_i, \forall i$. Therefore, $P_i' = P_i \geq p_i \geq p_i', \forall i$, and $\mathbf{T}'$ is also a feasible solution with $E' = E$. Then by repeating the re-sequencing in Remark 1 we can finally obtain a feasible solution $\mathbf{T}'$ with $E' \geq E$, such that $\mathbf{n}' = \mathbf{n}$, and $\mathbf{f}' = [f_1', \cdots, f_I']$, where $1 \leq f_1' + n_1 \leq f_2' + n_2 \leq \cdots \leq f_I' + n_I \leq M$. This concludes the proof. ■

From Remarks 1 and 2, we have that $\mathbf{f}$ and $\mathbf{n}$ should satisfy $1 \leq f_i + n_i \leq M$, and $f_i + n_i \leq f_{i+1}, \forall i$. Therefore, the set of all candidate index vectors and group size vectors can be respectively given by

$$\mathcal{F} = \{\mathbf{f} | 1 \leq f_1 \leq f_2 \leq \cdots \leq f_I \leq M\}, \quad (15)$$

$$\mathcal{N} = \{\mathbf{n} | 0 \leq n_i \leq f_{i+1} - f_i, \forall i < I, n_I \leq M - f_I\}. \quad (16)$$

Based on these conclusions, we can obtain an optimal solution using the following procedure: Search for all the candidate optimal solutions according to (15) and (16), and compute the corresponding total access rate E by (11a). Then the optimal solution is the $\mathbf{T}$ with the maximum E. We can summarize the steps for computing an optimal solution as Algorithm 1.

Algorithm 1 An Optimal Algorithm for $\mathcal{P}_1$

Input: Network parameters, M, I, $p_m (m = 1, , M)$
Output: $\mathbf{T}_{\text{OPT}}$ and E_{OPT}
1: **Sort** devices so that $p_1 \geq p_2 \geq \cdots \geq p_M$.
2: **Initialize** $E_{\text{OPT}} = 0$;
3: **for** $\mathbf{f} \in \mathcal{F}$ **do**
4: Compute $\mathcal{N}$ by (16);
5: **for** $\mathbf{n} \in \mathcal{N}$ **do**
6: Compute $\mathbf{T}$ by (13);
7: **if** constrains (11c)-(12) hold **then**
8: Compute the corresponding E by (11a);
9: **if** $E \geq E_{\text{OPT}}$ **then**
10: Update $\mathbf{T}_{\text{OPT}} = \mathbf{T}$ and $E_{\text{OPT}} = E$;
11: **end if**
12: **end if**
13: **end for**
14: **end for**
15: Return $\mathbf{T}_{\text{OPT}}$ and E_{OPT}.

Let us now examine the complexity of Algorithm 1. First, it takes $O(M \log M)$ operations for the sorting. Second, there are $|\mathcal{N} \times \mathcal{F}|$ loops. Since the size of some groups could be zero, denote by $\mathcal{I}_N = \{i | n_i \neq 0\}$ the set of the indices of non-empty subsets. Clearly, $\mathcal{I}_N \subseteq \{1, \cdots, I\}$. Let $m = |\mathcal{I}_N|$, and hence there are $\sum_{m=1}^{I} C_I^m$ possible ways to choose $\mathcal{I}_N$. For each $\mathcal{I}_N$, there are C_M^m possible ways of choosing m devices from M devices, *i.e.*, $|\mathcal{F}| = C_M^m$. According to (16), we have that $|\mathcal{N}| = \prod_{i=1}^{m} N_i$, where $N_i = f_{i+1} - f_i$, $i = 1, \cdots, m-1$, and $N_m = M - f_m$. In summary, it takes $O_1 = O\left(M \log M + \sum_{m=1}^{I} (C_I^m C_M^m \prod_{i=1}^{m} N_i)\right)$ operations to obtain an optimal solution to $(\mathcal{P}1)$.

Next, we simplify the complexity expression. From the definition of N_i we have that $N_i = M$. By the product property of O-notation, $\prod_{i=1}^{m} N_i = O(M^m)$. Moreover, since $I \ll M$, we have that $C_I^m \ll C_M^m$, and hence $\sum_{m=1}^{I} C_I^m C_M^m = O(C_M^I)$. Furthermore, Stirling's approximation $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$ yields $C_M^I = O\left(\frac{M^M}{I^I (M-I)^{M-I}}\right) = O(M^I)$. Therefore,

$$\begin{aligned} O_1 &= O\left(M \log M + \sum_{m=1}^{I}\left(C_I^m C_M^m \prod_{i=1}^{m} N_i\right)\right) \\ &= O\left(M \log M + \sum_{m=1}^{I}(C_I^m C_M^m M^m)\right) \end{aligned}$$

$$\stackrel{a)}{=} O\left(M\log M + M^I\right)$$
$$\stackrel{b)}{=} O\left(M^I\right), \tag{17}$$

where the labeled equalities come from: a) the sum property of O-notation, and $O\left(\sum_{m=1}^{I} C_I^m C_M^m\right) = O(M^I)$; and b) $M\log M = o\left(M^I\right)$. Since $I \ll M$, O_1 is much lower than that of the brute force method given by $O(I^M)$. Unfortunately, O_1 is still very high even for problems with moderate numbers of devices and RA slots. For example, when $M = 100$ and $I = 5$, $M^I = 10^{10}$. Thus, we may use it for small-scale problems to infer the structure of an optimal solution, which is useful to develop efficient heuristic algorithms.

B. Low-Complexity Algorithm

Since group paging is a pull-based method, the development of a low-complexity algorithm is of key importance. This stimulates us to further develop an efficient algorithm. One of the ideas of our proposed low-complexity scheme is to simplify the backoff rule in (8). The main goal of the GPSP scheme is to estimate the ASPs correctly and schedule the pre-backoff times to maximize the total access rate. We thus set the backoff rule to be the following: the device that fails in its n-th attempt in slot i performs its $(n+1)$-th attempt in slot $(i+1)$. In other words, the transit probability is chosen to be

$$\mathbb{P}(\mathcal{A}_{i,n+1}^m|\mathcal{Y}_{j,n}^m) = q_{ji} = \begin{cases} 1, & \text{if } i = j+1 \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Then (9) can be re-written for all i and n as

$$\begin{aligned}
&\mathbb{P}(\mathcal{A}_{i+1,n+1}^m) \\
&\quad= \mathbb{P}(\mathcal{Y}_{i,n}^m) \\
&\quad= \mathbb{P}(\mathcal{A}_{i-n+1,1}^m) \prod_{j=i-n+1}^{i}\left(1-e^{-\frac{M_j}{R}}\right) \\
&\quad= \begin{cases} \prod_{j=i-n+1}^{i}\left(1-e^{-\frac{M_j}{R}}\right), & \text{if } t_{(i-n+1)m} = 1 \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{19}$$

From (11b), we know that for any m there is at most one i such that $t_{mi} = 1$. Therefore, once the pre-backoff time is determined, the slot in which each device performs its n-th attempt is known. Moreover, in Lemma 1 it has been shown that for $\mathbf{t}_m = \mathbf{t}_{m'}$, $\mathbb{P}(\mathcal{A}_{i,n}^m) = \mathbb{P}(\mathcal{A}_{i,n}^{m'})$, $\forall i, n$. So we re-define the events as follows:

- $\mathcal{B}_{i,n}$: d_m with $\mathbf{t}_m = \mathbf{e}_i$ performs its n-th attempt (in slot $(i+n-1)$);
- $\mathcal{S}_{i,n}$: d_m with $\mathbf{t}_m = \mathbf{e}_i$ succeeds in its n-th attempt; and
- $\mathcal{F}_{i,n}$: d_m with $\mathbf{t}_m = \mathbf{e}_i$ fails in its n-th attempt.

Then (5) and (6) can be re-written as, for all m, i,

$$\mathbb{P}(\mathcal{S}_{i,n}|\mathcal{B}_{i,n}) = e^{-\frac{M_{i+n-1}}{R}}, \tag{20}$$
$$\mathbb{P}(\mathcal{F}_{i,n}|\mathcal{B}_{i,n}) = 1 - e^{-\frac{M_{i+n-1}}{R}}. \tag{21}$$

Each device with $\mathbf{t}_m = \mathbf{e}_i$ will perform its first attempt (in slot i), and thus

$$\mathbb{P}(\mathcal{B}_{i,1}) = 1, \quad \forall i. \tag{22}$$

According to (18), we have that

$$\begin{aligned}
\mathbb{P}(\mathcal{B}_{i,n+1}) &= \mathbb{P}(\mathcal{F}_{i,n}) \\
&= \mathbb{P}(\mathcal{B}_{i,1}) \prod_{k=1}^{n} \mathbb{P}(\mathcal{F}_{i,k}|\mathcal{B}_{i,k}) \\
&= \prod_{k=1}^{n}\left(1-e^{-\frac{M_{i+k-1}}{R}}\right), \quad \forall i, n.
\end{aligned} \tag{23}$$

Moreover, for all i, n,

$$\mathbb{P}(\mathcal{S}_{i,1}) = \mathbb{P}(\mathcal{B}_{i,1})\mathbb{P}(\mathcal{S}_{i,1}|\mathcal{B}_{i,1}) = e^{-\frac{M_i}{R}}, \tag{24}$$

$$\begin{aligned}
\mathbb{P}(\mathcal{S}_{i,n+1}) &= \mathbb{P}(\mathcal{B}_{i,n+1})\mathbb{P}(\mathcal{S}_{i,n+1}|\mathcal{B}_{i,n+1}) \\
&= \mathbb{P}(\mathcal{F}_{i,n})e^{-\frac{M_{i+n}}{R}} \\
&= e^{-\frac{M_{i+n}}{R}} \prod_{k=1}^{n}\left(1-e^{-\frac{M_{i+k-1}}{R}}\right).
\end{aligned} \tag{25}$$

According to (4), the average number of devices that transmit preamble in each slot can be written as

$$\begin{aligned}
M_1 &= \sum_{m=1}^{M} t_{1m}, \\
M_{i+1} &= \sum_{m=1}^{M}\sum_{n=1}^{i+1} \mathbb{P}(\mathcal{B}_{i+1,n}) \\
&= \sum_{m=1}^{M} \mathbb{P}(\mathcal{B}_{i+1,1}) + \sum_{m=1}^{M}\sum_{n=1}^{i}\left[\mathbb{P}(\mathcal{B}_{i,n})\left(1-e^{-\frac{M_i}{R}}\right)\right] \\
&= \sum_{m=1}^{M} t_{(i+1)m} + \left(1-e^{-\frac{M_i}{R}}\right) M_i.
\end{aligned} \tag{26}$$

According to (3), the ASP of the devices that are assigned to slot i for its first attempt can be written as

$$\begin{aligned}
P_i &= \mathbb{P}\left(\bigcup_{n=1}^{I-i+1} \mathcal{S}_{i,n}\right) \\
&= \mathbb{P}(\mathcal{S}_{i,1}) + \sum_{n=1}^{I-i} \mathbb{P}(\mathcal{S}_{i,n+1}) \\
&= e^{-\frac{M_i}{R}} + \sum_{n=1}^{I-i}\left[e^{-\frac{M_{i+n}}{R}} \prod_{k=1}^{n}\left(1-e^{-\frac{M_{i+k-1}}{R}}\right)\right].
\end{aligned} \tag{27}$$

Then we have that

$$P_I = e^{-\frac{M_I}{R}}, \tag{28}$$

$$\begin{aligned}
P_i &= e^{-\frac{M_i}{R}} + \left(1-e^{-\frac{M_i}{R}}\right) \\
&\quad \times \sum_{n=1}^{I-i}\left[e^{-\frac{M_{i+n}}{R}} \prod_{k=2}^{n}\left(1-e^{-\frac{M_{i+k-1}}{R}}\right)\right] \\
&= e^{-\frac{M_i}{R}} + (1-e^{-\frac{M_i}{R}}) \\
&\quad \cdot \left(e^{-\frac{M_{i+1}}{R}} + \sum_{n=1}^{I-i-1}\left[e^{-\frac{M_{i+n+1}}{R}} \prod_{k=1}^{n}\left(1-e^{-\frac{M_{i+k}}{R}}\right)\right]\right) \\
&= e^{-\frac{M_i}{R}} + \left(1-e^{-\frac{M_i}{R}}\right) P_{i+1} \geq e^{-\frac{M_i}{R}}, \quad \forall i < I.
\end{aligned} \tag{29}$$

Therefore, constraint (12) holds if $e^{-\frac{M_i}{R}} \geq p_i$, $\forall i$. Obviously,

$$e^{-\frac{M_i}{R}} \geq p_i \Leftrightarrow M_i \leq -R\log p_i. \tag{30}$$

Recall that the pre-backoff matrix $\mathbf{T}$ can be expressed by the index vector $\mathbf{f}$ and the group size vector $\mathbf{n}$ according to (13). Similar to the method in Section IV-A, we sort the devices such that $p_1 \geq p_2 \geq \cdots \geq p_M$, and then $p_i = \max_{d_m \in G_i}\{p_m\} = p_{f_i}$. In addition, the observations of the optimal solutions obtained by Algorithm 1 suggest that the system tries to satisfy the ASP requirements for the devices at their first transmission to reduce the number of retransmissions. Therefore, we approximate the average number of devices that transmit in slot i by the number of devices that transmit for the first time in slot i (*i.e.*, $M_i = n_i, \forall i$). Then, the simplified GPSP problem can be expressed as follows:

$$(\mathcal{P}2) \quad \max_{\mathbf{f},\mathbf{n}} \; E = \sum_{i=1}^{I} \sum_{n=0}^{n_i-1} p_{f_i+n}, \tag{31a}$$

$$s.t.\; n_i \leq -R \log p_{f_i}, \quad \forall i, \tag{31b}$$

$$f_i \in \{1, 2, \cdots, M\}, \quad \forall i. \tag{31c}$$

To maximize E, the inequality of constrain (31b) should be tight, *i.e.*,

$$n_i = -R \log p_{f_i}, \quad \forall i. \tag{32}$$

In other words, once the first device of a group is chosen, the size of the group is determined. Lemma 2 shows an efficient way of choosing $\mathbf{f}$.

Lemma 2: Assume that there are two feasible solutions $[\mathbf{f}, \mathbf{n}]$ *and* $[\mathbf{f}', \mathbf{n}']$ *where* $f_j > f'_j$, *and* $f_i = f'_i$, $\forall i \neq j$. *If* $p_{f_j} = p_{f'_j}$, *then* $E \geq E'$.

Proof: Since $p_{f_j} = p_{f'_j}$, we have that $n_j = n'_j$. For all $i \neq j$, $f_i = f'_i$, and thus $\mathbf{n} = \mathbf{n}'$. Therefore,

$$\begin{aligned} E - E' &= \left(\sum_{i \neq j} \sum_{n=0}^{n_i-1} p_{f_i+n} + \sum_{n=0}^{n_j-1} p_{f_j+n} \right) \\ &\quad - \left(\sum_{i \neq j} \sum_{n=0}^{n'_i-1} p_{f'_i+n} + \sum_{n=0}^{n'_j-1} p_{f'_j+n} \right) \\ &= \sum_{n=0}^{n_j-1} \left(p_{f_j+n} - p_{f'_j+n} \right) \geq 0. \end{aligned}$$

This concludes the proof. ∎

Therefore, for any i, if f_i is determined, then f_{i+1} should be chosen from $\mathcal{F}_{i+1} = \{f_i + n_i\} \cup \{m | f_i + n_i < m \leq M, p_m < p_{m-1}\}$. Clearly, $\mathcal{F}_1 = \{1\} \cup \{m | 1 < m \leq M, p_m < p_{m-1}\}$. The set of all possible solutions is given by $\mathcal{F} = \mathcal{F}_1 \times \cdots \mathcal{F}_I$. Assume that the ASP requirements p_m are quantized into k ASP scales (*i.e.*, $p_m \in \{\frac{1}{k}, \frac{2}{k}, \cdots, \frac{k-1}{k}, 1\}$). Then, $|\mathcal{F}_i| \leq k, \forall i$ (*i.e.*, each group has at most k choices for its first device), and hence $|\mathcal{F}| \leq k^I$.

Based on these conclusions, we design an efficient low-complexity algorithm. We first sort the devices by p_m, so that $p_1 \geq p_2 \geq \cdots \geq p_M$. Then, we arrange the devices into I groups. As aforementioned, once the first device is determined, the group size can be computed by (32). Therefore, we need only to choose the first device for each group such that $f_{i+1} \in \mathcal{F}_i$. We summarize these steps as Algorithm 2.

Algorithm 2 A Low-Complexity Algorithm

Input: Network parameters, M, I, $p_m (m = 1, , M)$, k
Output: $\mathbf{T}_{\text{HEU}}$ and E_{HEU}
Sort devices so that $p_1 \geq p_2 \geq \cdots \geq p_M$.
Initialize $E_{\text{HEU}} = 0$.
for $\mathbf{f} \in \mathcal{F}$ **do**
  Compute $\mathbf{n}$, $\mathbf{T}$, and P_i's by (32), (18) and (29), respectively;
  for $m = 1 : M$ **do**
    if (12) is not satisfied **then**
      Set $\mathbf{t}_m = \mathbf{0}$;
    end if
  end for
  Compute the P_i's and E by (29) and (11a) respectively;
  if $E > E_{\text{HEU}}$ **then**
    Set $E_{\text{HEU}} = E$ and $\mathbf{T}_{\text{HEU}} = \mathbf{T}$.
  end if
end for
Return $\mathbf{T}_{\text{HEU}}$ and E_{HEU}.

Next we analyze the complexity of Algorithm 2. It takes $O(M \log M)$ operations to sort the sequence of d_m's. Moreover, there are $|\mathcal{F}| \leq k^I$ loops, and in each loop it takes $O(M)$ operations to update $\mathcal{F}_i$. Since $|\mathcal{F}| \leq k^I$, it thus takes $O(Mk^I)$ operations for searching. In summary, the complexity is $O_2 = O(M \log M + Mk^I)$. Since $I \ll M$ and $k \ll M$, we have that $O_2 \ll O_1$.

V. Numerical Results and Discussion

In this section, we first compare our analytical results with simulation results to verify the correctness of the ASP estimation model. We also implement both the optimal algorithm (Algorithm 1) and the heuristic algorithm (Algorithm 2) to obtain the corresponding solutions to the GPSP problem. We compare the performance of the GPSP schemes with that of a number of conventional GP schemes, including the original GP scheme [7], the random pre-backoff GP (PGP) scheme [10], and the uniform PGP scheme [9].

The key performance measures, as described in the 3GPP report [38], include the average access success probability (ASP), average access delay (AD), and average number of preamble transmissions (NPT). Apart from these three performance measures, we also evaluate the total access rate (TAR) defined in (10) as a reference. In addition, we compare the number of iterations[1] of the two GPSP algorithms. We consider two scenarios: a) a small-scale scenario where $I = 2$ and M varies from 5 to 30; and b) a large-scale scenario where $I = 30$ and M varies from 1000 to 5000. The small-scale scenario is used to validate the performance upper bound provided by the optimal GPSP scheme and examine the difference between the optimal performance and that of the heuristic GPSP algorithm. The large-scale scenario, which is of primary interest in this paper, is used to verify the correctness of the ASP estimation model, and evaluate the

[1]The number of iterations (or loops) indicates the number of all candidate solutions that the algorithms search through.

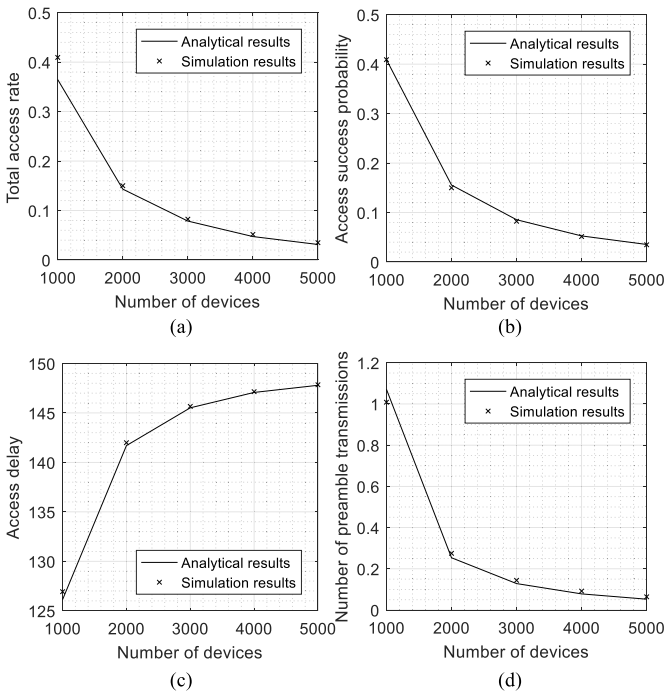


Fig. 3. Comparison of the analytical results and simulation results: (a) The total access rate; (b) The average access success probability; (c) The average access delay; and (d) The average number of preamble transmissions.

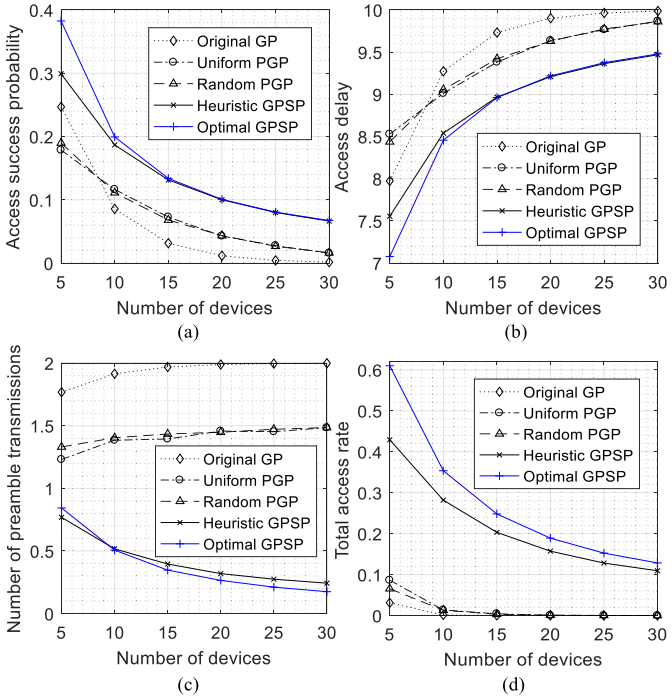


Fig. 4. Comparison of the system performance in the small-scale scenarios with $I = 2$ and M varying from 5 to 30: (a) The average access success probability; (b) The average access delay; (c) The average number of preamble transmissions; and (d) The total access rate.

performance gain of the heuristic GPSP scheme (the optimal GPSP algorithm is not evaluated due to its high complexity in this regime). To obtain statistical results, for each scenario we generate 100 random samples of ASP requirement p_m for each device according to the Gaussian distribution with $\mu = 0.5$ and $\sigma = 0.2$, and then compute the average results. Set $T_{\text{RAR}}+W_{\text{RAR}} = \frac{1}{2}T_{\text{RA\_REP}}$ and $W_{\text{B}} = \frac{1}{2}T_{\text{RA\_REP}}$, which implies that the transit probability is given by (18). As mentioned in Section III-A, we use RACH configuration index 6 where $T_{\text{RA\_REP}} = 5$ subframes. Also, without loss of generality, the AD of the failed devices is assumed to be the length of the access interval.

First, we verify the accuracy of the ASP estimation model by comparing the analytical results with the simulation results. The simulations are based on the Monte Carlo method, and each point represents the average value of 10^7 samples [8]. The processing latency in the simulations is set according to the guideline given by [33]. For each sample, we employ Algorithm 2 to determine the pre-backoff matrix, and then compute the analytical results and simulation results respectively. From the comparison of the analytical and simulation results (Fig. 3), we can see that the analytical model effectively describes the system.

Second, we demonstrate the performance upper bound of the GPSP scheme in the small-scale scenario. The number of preambles is set to $R = 5$ in this scenario. Note that, the optimization objective of the GPSP scheme is only the TAR, so the optimal GPSP scheme is not guaranteed to outperform other schemes in terms of other metrics (*e.g.*, NPT). Interestingly, according to the numerical results, the optimal GPSP scheme usually outperforms the other schemes in terms of all these metrics. The comparison of the system performance in a small-scale scenario (Fig. 4) shows that the ASP of each scheme decreases with the number of devices, and the ASPs of GPSP schemes are always higher than those of the conventional GP schemes. This is because the GPSP schemes provide a differentiated service, which allows the system to allocate the access resources more efficiently. In addition, the GPSP schemes select only a fraction of the devices to transmit, which can alleviate collisions. In contrast, in the conventional schemes every device attempts to access the network, and is treated equally. Therefore, the GPSP schemes can obtain both a diversity gain and a load-alleviation gain. Moreover, the ASP of the heuristic GPSP is slightly lower than that of the optimal GPSP when $M = 6$, but converges to that of the optimal one.

Intuitively, a lower ASP leads to a higher AD. Therefore, as shown in the comparison of the AD (Fig. 4(b)), the AD of each scheme increases with the number of devices, and the ADs of the GPSP schemes are lower than those of the conventional schemes.

In the comparison of the NPT (Fig. 4(c)), the NPTs of conventional GP schemes increase with the number of devices, while the NPTs of both the optimal and heuristic GPSP schemes decrease with the number of devices and are always less than those of conventional schemes. This is because the average ASP deceases with the number of devices (see Fig. 4(a)), and when the ASP is lower than the ASP requirement, the devices do not transmit during the whole access interval. Therefore, the NPTs of the GPSP schemes decrease. Note that the NPT is related to the energy consumption of the system and is desired to be small.

In the comparison of the TAR (Fig. 4(d)), the TAR of each scheme decreases with the number of devices. The TARs of the

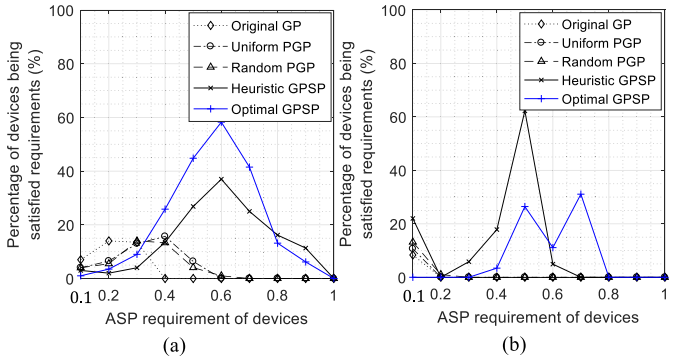


Fig. 5. Comparison of the percentage of the devices, whose ASP requirements are satisfied, in the small-scale scenario with $I = 2$: (a) $M = 5$; and (b) $M = 30$.

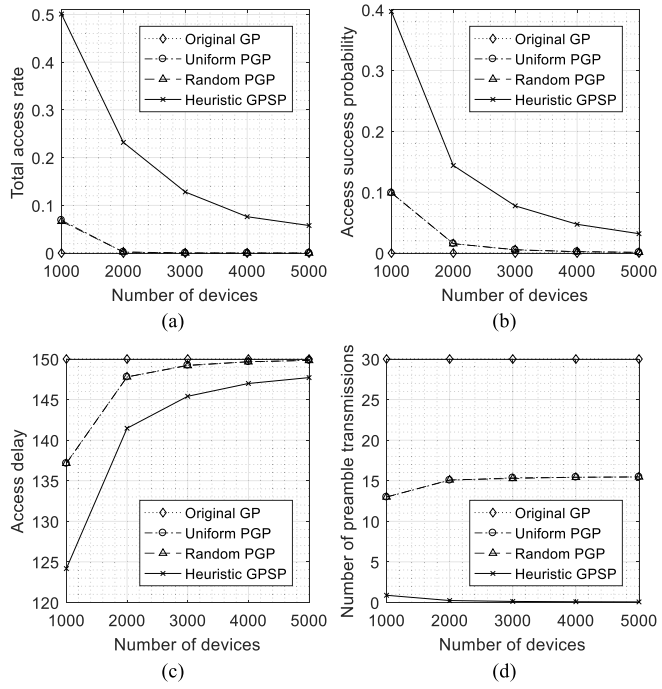


Fig. 6. Comparison of the system performance in the large-scale scenario with $I = 30$ and M varying from 1000 to 5000: (a) The total access rate; (b) The average access success probability; (c) The average access delay; and (d) The average number of preamble transmissions.

GPSP schemes are always higher than those of conventional schemes. The difference between the TAR of the optimal GPSP and that of the heuristic GPSP becomes smaller when the number of devices increases. The TAR of conventional schemes dramatically descends with the number of devices and eventually drops to zero, since the ASP requirements of devices are not satisfied.

The comparison of the percentage of the devices whose ASP requirements are satisfied (Fig. 5) shows that the GPSP schemes can provide better service to the devices with stricter ASP requirement in comparison with the conventional schemes. This is because the GPSP schemes assign higher weights to the devices with stricter requirements.

Next, we examine the performance of the heuristic GPSP scheme and the conventional schemes in the large-scale scenario. In the large-scale scenario we use the typical value of

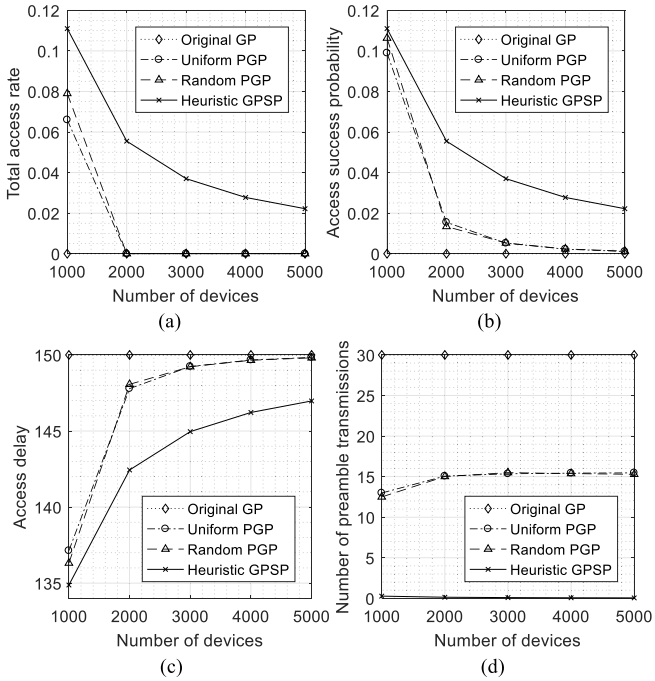


Fig. 7. Comparison of system performance in the identical ASP requirement scenario with $I = 30$ and M varies from 1000 to 5000, and all devices have the same ASP requirement $p_m = 0.5$: (a) The total access rate; (b) The average access success probability; (c) The average access delay; and (d) The average number of preamble transmissions.

the number of preambles $R = 54$. Similar to the small-scale scenario, the comparison of the system performance in the large-scale scenario (Fig. 6) demonstrates that the heuristic GPSP scheme outperforms the conventional schemes in terms of all these metrics and obtains a higher performance gain than the gain in the small-scale scenario. This is because the more devices there are, the higher diversity gain can be obtained. Therefore, we can conclude that the proposed GPSP schemes can efficiently improve the system performance especially in the large-scale scenarios.

As mentioned in Section I, the conventional GP problem can be considerd to be a special case of the GPSP problem with the identical ASP requirements. Next, we examine the performance of the heuristic GPSP and the conventional GP schemes in the case where $I = 30$ and M varies from 1000 to 5000, and all devices have the same ASP requirement $p_m = 0.5$. The comparison of system performance in the identical ASP requirement scenario is shown in Fig. 7. We can see that the GPSP scheme outperforms the conventional GP schemes even in this special case, although the performance gain is not as high as that in general cases where the ASP requirements of devices are different (see Fig. 6). This is because the identical ASP requirements provide no diversity gain, but the load-alleviation gain can still be enjoyed by preventing a proportion of devices from transmissions.

The comparison of the number of iterations in the small-scale scenario (Fig. 8) shows the number of iterations as a function of the number of devices M and the number of RA slots I. The number of iterations of the optimal algorithm

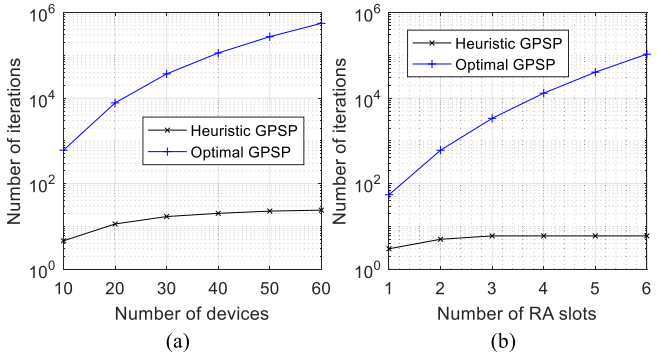


Fig. 8. Comparison of the number of iterations of the optimal GPSP algorithm and the heuristic GPSP algorithm in the small-scale scenario: (a) with $I = 2$ and M varying from 10 to 60; and (b) with $M = 10$ and I varying from 1 to 6.

increases with both M and I exponentially, while that of the heuristic algorithm grows considerably slower. Therefore, the heuristic GPSP algorithm can fully adapt to mMTC applications.

VI. Conclusion

In this work, we have investigated an optimal pre-backoff group paging scheduling scheme using an access success probability estimation model. The proposed system, based on the estimated achievable access success probability, provides differentiated services to individual devices in order to maximize the total access rate. To solve the underlying optimization problem, we have proposed an optimal algorithm and an efficient low-complexity algorithm. Numerical results have shown that the GPSP scheme can effectively improve network access rate.

Providing connectivity to a large number devices is a major challenge. Several interesting directions could be considered in the future. Although many pragmatic access control mechanisms have been proposed in the literature, theoretical analysis is still needed. For example, the performance gain of the heuristic algorithm in this work is not theoretically guaranteed, and it would be interesting to obtain theoretical performance guarantees. Also, as the new ASP requirement dimension attains a notable multiplexing gain, it would be interesting to investigate whether other performance metrics, such as access delay and energy consumption, can bring extra multiplexing gains or not. Finally, it is important to keep in mind that low-complexity algorithms are of vital importance due to the nature of mMTC applications.

Appendix A
A Proof of Theorem 1

The I-dimensional knapsack problem seeks to maximize the sum of the values of the items in the knapsack so that the sum of the weights in each dimension does not exceed the knapsack's capacity. Formally, the I-dimensional knapsack problem consists of a set of items $\mathcal{D} = \{d_1, d_2, \cdots, d_M\}$, each with an I-dimensional weight vector $\mathbf{w}_m = [w_{m1}, \cdots, w_{mI}]$ and a value v_m, along with an I-dimensional maximum weight capacity vector $[W_1, \cdots, W_I]$. Also, let $x_m \in \{0, 1\}$ indicate whether or not the item d_m is included in the knapsack. A general I-dimensional knapsack problem can be expressed as follows:

$$(\mathcal{P}3) \quad \max_{\mathbf{x}} \sum_{m=1}^{M} v_m x_m, \tag{33a}$$

$$s.t. \sum_{m=1}^{M} w_{mi} x_m \le W_i, \quad \forall i = 1, 2, \cdots, I, \tag{33b}$$

$$x_m \in \{0, 1\}. \tag{33c}$$

Now we construct the corresponding GPSP problem. The basic idea is to map the items in the knapsack to the devices in the GPSP problem, and the knapsack to the RA slots. For all m, let $x_m = \mathbf{1}_{\{\|\mathbf{t}_m\|>0\}}$, and then (33c) is equivalent to (11c). The value of the items in the knapsack is mapped to the weight of the device (*i.e.*, let $p_m = \frac{v_m}{\max v_m}$, such that $p_m \in [0, 1]$), and hence (33a) is equivalent to (11a).

Next we show that the constraint in (33b) is a special case of the constraint in (11b). The constraint in (11b) can be written as $P_i \ge p_i, \forall i$ (see Lemma 1), and can be deemed as the constraint on M_i. The capacity of each slot and the "weight" of each device on a particular slot depend on the t_{im}'s. Observe that the knapsack capacity vector and the item weight vector in $(\mathcal{P}3)$ are fixed. Therefore, $(\mathcal{P}3)$ is actually a special case of $(\mathcal{P}1)$ in which we only determine whether or not to serve a device. This concludes the proof.

Appendix B
A Proof of Lemma 1

The proof follows by examining at the following two cases.

Case 1: $\mathbf{t}_m = \mathbf{t}_{m'} = \mathbf{0}$. From (7), we have that $\mathbb{P}(\mathcal{A}_{i,n}^m) = \mathbb{P}(\mathcal{A}_{i,n}^{m'}) = 0$, $\mathbb{P}(\mathcal{X}_{i,n}^m) = \mathbb{P}(\mathcal{X}_{i,n}^{m'}) = 0$, $\mathbb{P}(\mathcal{Y}_{i,n}^m) = \mathbb{P}(\mathcal{Y}_{i,n}^{m'}) = 0$, $\forall i, n$. Then from (3), we have that $P_m = P_{m'} = 0$, $\forall i, n$.

Case 2: $\mathbf{t}_m = \mathbf{t}_{m'} \neq \mathbf{0}$. From (5) and (6), we have that for all i and n,

$$\mathbb{P}(\mathcal{X}_{i,n}^m | \mathcal{A}_{i,n}^m) = \mathbb{P}(\mathcal{X}_{i,n}^{m'} | \mathcal{A}_{i,n}^{m'}) \tag{34}$$

$$\mathbb{P}(\mathcal{Y}_{i,n}^m | \mathcal{A}_{i,n}^m) = \mathbb{P}(\mathcal{Y}_{i,n}^{m'} | \mathcal{A}_{i,n}^{m'}). \tag{35}$$

Moreover, according to (7), we have that for all i,

$$\mathbb{P}(\mathcal{A}_{i,1}^m) = t_{mi} = \mathbb{P}(\mathcal{A}_{i,1}^{m'}). \tag{36}$$

Combining (34), (35) and (36), we have that for all i,

$$\mathbb{P}(\mathcal{X}_{i,1}^m) = \mathbb{P}(\mathcal{X}_{i,1}^{m'}) \quad \text{and } \mathbb{P}(\mathcal{Y}_{i,1}^m) = \mathbb{P}(\mathcal{Y}_{i,1}^{m'}). \tag{37}$$

In addition, (8) yields

$$\mathbb{P}(\mathcal{A}_{i,n}^m | \mathcal{Y}_{i,n}^m) = \mathbb{P}(\mathcal{A}_{i,n}^{m'} | \mathcal{Y}_{i,n}^{m'}), \quad \forall i, n. \tag{38}$$

Then from (9), (34), and (38) we can derive that

$$\mathbb{P}(\mathcal{A}_{i,n}^m) = \mathbb{P}(\mathcal{A}_{i,n}^{m'}), \quad \forall i, n. \tag{39}$$

Combining (34) and (39), we have for all i, n,

$$\mathbb{P}(\mathcal{X}_{i,n}^m) = \mathbb{P}(\mathcal{X}_{i,n}^{m'}), \quad \text{and } \mathbb{P}(\mathcal{Y}_{i,n}^m) = \mathbb{P}(\mathcal{Y}_{i,n}^{m'}). \tag{40}$$

Finally, from (3), we have that $P_m = P_{m'} = 0$.

This concludes the proof.

REFERENCES

[1] H. Shariatmadari *et al.*, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.

[2] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[3] H. Tullberg *et al.*, "The METIS 5G system concept: Meeting the 5G requirements," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 132–139, Dec. 2016.

[4] *Service Requirements for Machine-Type Communications (MTC)—Stage 1*, document 3GPP TS 22.368, V13.2.0, Dec. 2016.

[5] *Study on RAN Improvements for Machine-Type Communications*, document 3GPP TR 37.868, V11.0.0, Sep. 2011.

[6] *Pull Based RAN Overload Control*, document 3. R2-104870, Huawei, Shenzhen, China, RAN2#71, Aug. 2010.

[7] *Group Paging for MTC Devices*, document 3. R2-104004, LG Electronics, RAN2#70, Jul. 2010.

[8] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao, "Performance analysis of group paging for machine-type communications in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3371–3382, Sep. 2013.

[9] O. Arouk, A. Ksentini, and T. Taleb, "Group paging-based energy saving for massive MTC accesses in LTE and beyond networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1086–1102, May 2016.

[10] R. Harwahyu, X. Wang, R. F. Sari, and R.-G. Cheng, "Analysis of group paging with pre-backoff," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 34–42, 2015.

[11] *Backoff Enhancements for RAN Overload Control*, document 3. R2-111916, ZTE, RAN WG2#73, Apr. 2011.

[12] J.-B. Seo and V. C. M. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.

[13] X. Yang, A. Fapojuwo, and E. Egbogah, "Performance analysis and parameter optimization of random access backoff algorithm in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Quebec City, QC, Canada, Sep. 2012, pp. 1–5.

[14] *Access Class Barring and Overload Protection*, document 3GPP TS 23.898, V7.0.0, Mar. 2005.

[15] *Backoff Enhancements for RAN Overload Control*, document 3. R2-112863, ZTE, RAN2#73, May 2011.

[16] *Merits of the Slotted Access Methods for MTC*, document 3. R2-112247, Alcatel-Lucent Shanghai Bell, RAN2#73, Apr. 2011.

[17] Z. Wang and V. W. S. Wong, "Optimal access class barring for stationary machine type communication devices with timing advance information," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5374–5387, Oct. 2015.

[18] H. Jin, W. Toor, B. C. Jung, and J.-B. Seo, "Recursive pseudo-Bayesian access class barring for M2M communications in LTE systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, Sep. 2017.

[19] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[20] H. He, Q. Du, H. Song, W. Li, Y. Wang, and P. Ren, "Traffic-aware ACB scheme for massive access in machine-to-machine networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 617–622.

[21] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, Jun. 2014.

[22] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[23] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 27–32, Jan. 2012.

[24] Y.-C. Pang, S.-L. Chao, G.-Y. Lin, and H.-Y. Wei, "Network access for M2M/H2H hybrid systems: A game theoretic approach," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 845–848, May 2014.

[25] N. Zhang, G. Kang, J. Wang, Y. Guo, and F. Labeau, "Resource allocation in a new random access for M2M communications," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 843–846, May 2015.

[26] C.-Y. Oh, D. Hwang, and T.-J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4182–4192, Aug. 2015.

[27] J. Guo, S. Durrani, X. Zhou, and H. Yanikomeroglu, "Massive machine type communication with data aggregation and resource scheduling," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 4012–4026, Sep. 2017.

[28] H.-C. Lin and W.-Y. Chen, "An approximation algorithm for the maximum-lifetime data aggregation tree problem in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3787–3798, Jun. 2017.

[29] W. Cao, G. Feng, S. Qin, and M. Yan, "Cellular offloading in heterogeneous mobile networks with D2D communication assistance," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4245–4255, May 2017.

[30] *E-UTRA Random Access*, document 3. R1-060584, Ericsson, RAN WG1#44, Feb. 2006.

[31] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document 3GPP TS 36.211, V10.4.0, Dec. 2011.

[32] *LTE Random-Access Capacity and Collision Probability*, document 3. R1-061369, Ericsson, RAN WG1#45, May 2006.

[33] *Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced)*, document 3GPP TS 36.912, V13.0.0, Dec. 2015.

[34] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, document 3GPP TS 36.321, V14.1.0, Dec. 2016.

[35] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*. Berlin, Germany: Springer-Verlag, 2012.

[36] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of Computer Computations*. New York, NY, USA: Springer, 1972, pp. 85–103.

[37] B. Korte and R. Schrader, "On the existence of fast approximation schemes," in *Proc. Nonlinear Program.*, vol. 4, 1981, pp. 415–437.

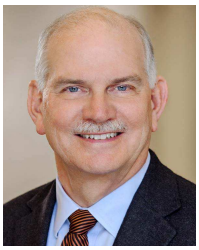
[38] *Further Analysis of Group Paging for MTC*, document 3. R2-113198, ITRI, RAN WG2#74, May 2011.

Wei Cao (S'18) received the B.Eng. degree in electronic engineering from the University of Electronic Science and Technology of China (UESTC) in 2009. She is currently pursuing the Ph.D. degree with the National Key Laboratory of Science and Technology on Communications, UESTC. She is also a Visiting Student Research Collaborator with the Department of Electrical Engineering, Princeton University. Her research interests include theoretical analysis, optimization, and algorithmics in the next generation cellular network for the massive machine-type-communication and Internet of Things.

Alex Dytso (S'08–M'13) received the B.S. degree from the University of Illinois, Chicago, in 2011, where he also received the International Engineering Consortium's William L. Everitt Student Award of Excellence for outstanding seniors, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Illinois, in 2016. He is currently a Post-Doctoral Researcher with the Department of Electrical Engineering, Princeton University.

His current research interests are in the areas of multi-user information theory and estimation theory, and their applications in wireless networks.

Gang Feng (M'01–SM'06) received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), in 1986 and 1989, respectively, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong in 1998. He joined the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2000, as an Assistant Professor and was promoted as an Associate Professor in 2005. He is currently a Professor with the National Laboratory of Communications, UESTC. He has extensive research experience and has published widely in computer networking and wireless networking research. His research interests include AI-enabled wireless access and content distribution, and next generation cellular networks.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty of Princeton University, where he is currently the Michael Henry Strater University Professor of electrical engineering. From 2006 to 2016, he served as the Dean of the School of Engineering and Applied Science, Princeton University. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Information Theoretic Security and Privacy of Information Systems* (Cambridge University Press, 2017).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, Honorary Professorships at Peking University and Tsinghua University, both conferred in 2017, and a D.Sc. *(honoris causa)* from Syracuse University also awarded in 2017.

Zhi Chen received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from the University of Electronic Science and Technology of China (UESTC) in 1997, 2000, and 2006, respectively. In 2006, he joined the National Key Laboratory of Science and Technology on Communications, UESTC, and has been a Professor since 2013. He was a Visiting Scholar with the University of California at Riverside from 2010 to 2011. His current research interests include 5G mobile communications, terahertz communication, and tactile Internet. He has served as a Reviewer for various international journals and conferences, including the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the IEEE TRANSACTIONS ON SIGNAL PROCESSING.