Signaling Overhead-Constrained Throughput Optimization for 5G Packet-Based Random Access with mMTC

Yiwen Liang*, Wen Zhan*, Xinghua Sun*, Kai Xie* and Xiang Chen†

*School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University
†School of Electronics and Information Technology, Sun Yat-sen University
liangyw39@mail2.sysu.edu.cn, {zhanw6, sunxinghua, xiek8, chenxiang}@mail.sysu.edu.cn

GLOBECOM 2021 - 2021 IEEE Global Communications Conference | 978-1-7281-8104-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GLOBECOM46510.2021.9685888

***Abstract*—To reduce the signaling overhead for sporadic small packet transmission in massive Machine Type Communications (mMTC), Packet-Based Random Access (PBRA) scheme is introduced in 5G system, where devices can transmit data packets in the random access procedure without connection establishment. Yet, even with PBRA, the signaling overhead may surge if the system parameters are configured improperly. This paper aims to address this issue by studying how to tune the Access Class Barring (ACB) factor to maximize the throughput while maintaining the signaling-to-throughput ratio below a certain level. Explicit expressions of maximum throughput and the corresponding optimal ACB factor in saturated and unsaturated cases are derived. It reveals that with a demanding requirement on signaling-to-throughput ratio, the throughput performance has to be sacrificed even with optimal tuning of ACB factor. To boost the throughput performance, the system should either loose the signaling constraint or enlarge the packet length. The analysis is verified by simulations and sheds important light on practical 5G network design for supporting mMTC with PBRA.**

***Index Terms*—Machine type communications, Packet-based Random access, Signaling overhead, Throughput, Optimization.**

I. Introduction

Massive Machine Type Communications (mMTC) is a new kind of communication paradigm in which Machine Type Devices (MTDs), such as sensors and actuators, process and exchange information packets without human intervention. It is forecasted that the number of MTDs will surpass 14 billion in 2023 [1]. Due to its huge market potential, mMTC has been regarded as one of three generic service categories for 5G system and the traffic generated by MTDs will have a major influence on the design of 5G system.

The work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101120003, in part by the State's Key Project of Research and Development Plan under Grants 2019YFE0196400, in part by National Natural Science Foundation of China under Grant 62001524, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515011906, in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant 2021A04 and in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, 2021qntd11. *(Corresponding author: Wen Zhan.)*

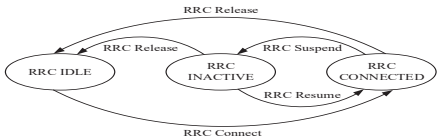


Fig. 1. RRC State Machine in 5G.

In 4G/5G system, the Connection-Based Random Access (CBRA) scheme has been adopted for a long time, in which each device has to perform the random access procedure for establishing a connection with the gNB prior to its data transmission. The CBRA scheme was originally designed to support traditional Human-Type Communications (HTC), such as video streaming, where the number of devices is small but each device transmits a significant amount of data. Therefore, the signaling overhead for the connection establishment is negligible.

However, mMTC traffic is radically different: it contains a large number of MTDs while each MTD transmits small packets sporadically. As such, establishing a connection prior to data transmission is inefficient, because the comparatively heavy signaling overhead results in additional latency and drains the limited battery of MTD fast. To address the signaling issue, the Packet-Based Random Access (PBRA) scheme is introduced in 5G standards, where MTDs can transmit one small data packet in the random access procedure without connection establishment [2]. By doing so, the simulation results in [3], [4] show that the energy consumption, latency and signaling overhead can be significantly reduced, compared to that in the CBRA scheme.

Although the PBRA scheme is a promising way for supporting mMTC in 5G networks, its performance crucially depends on the network configuration [5], such as the transmission probability of each device. The fundamental reason lies in the Aloha-type access paradigm that PBRA takes root in, i.e., each MTD independently determine when/how to send the request. This distributed behavior implies that in the massive access case, severe collision issue and intolerably low chance

of successful access may emerge, leading to frequent packet retransmissions along with enormous signaling overhead. Thus, it is of great importance to study how to tune the system parameters to optimize the access efficiency and further reduce the signaling overhead of 5G networks with PBRA scheme.

Extensive studies for Aloha-type access networks have focused on throughput performance, i.e., the long-term average number of successfully-transmitted packets/bits per time slot. For instance, various strategies have been developed to adjust the transmission probability of devices for throughput maximization based on the periodical estimation of the channel load [7]–[9]. In [10], [11], analytical frameworks were developed to characterize the behavior of each device or the Head-of-Line (HOL) packet, with which optimal backoff parameter settings for throughput maximization were obtained. While the aforementioned developments have been substantial, none of them take the signaling overhead into consideration. Questions naturally arise: Is the throughput maximization equivalent to signaling overhead minimization? If a certain requirement on signaling overhead is given, then how to properly tune the system parameters to optimize the throughput performance in the PBRA scheme?

In this paper, we provide a closed-form solution to above open questions by focusing on mMTC in 5G networks. Specifically, we evaluate the signaling overhead via the signaling-to-throughput ratio, which represents the average amount of signaling overhead for successfully delivering one information bit. By extending the analytical framework in [11], both the throughput and signaling-to-throughput ratio are characterized. The key contribution of this paper is that for a given upper-bound on the signaling-to-throughput ratio, explicit expressions of the maximum throughput and corresponding optimal transmission probability (refer to as the Access Classing Barring (ACB) factor in the 5G standard [6]) are obtained in both unsaturated and saturated cases. The analysis shows that to meet a stringent upper-bound on the signaling-to-throughput ratio, the throughput performance has to be sacrificed even with optimal tuning of the transmission probability. Otherwise, to maintain the optimum throughput performance, the data packet length should be enlarged. The analysis is verified by simulation results, which provides direct guidance on practical network design for mMTC in 5G system.

The rest of the paper is organized as follows. Section II presents technical backgrounds of RRC handling and PBRA in 5G before the system model in Section III. Signaling-constrained throughput maximization problem is addressed in Section IV and the results are validated via extensive simulation results in Section V. Finally, concluding remarks are summarized in Section VI.

II. Background of PBRA and RRC Handling in 5G

To ease the understanding, we would like to first present technical backgrounds of PBRA and Radio Resource Control (RRC) handling in 5G in this section before the system model in Section III.

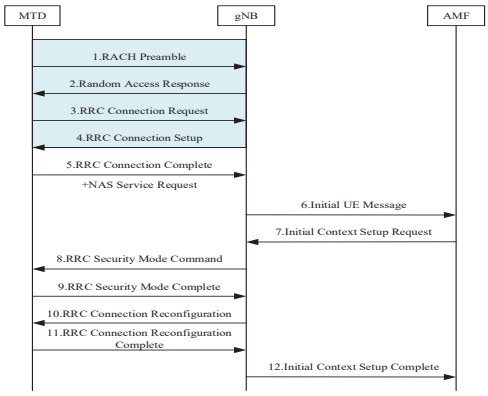


Fig. 2. Signaling diagram for the state transition from RRC IDLE to RRC CONNECTED.

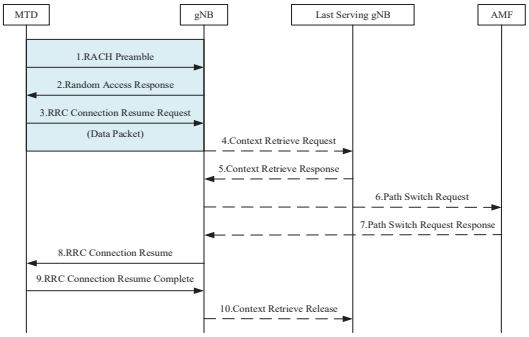


Fig. 3. Signaling diagram for the state transition from RRC INACTIVE to RRC CONNECTED.

Specifically, from the 5G RRC layer point of view, three RRC states are defined: RRC CONNECTED, RRC IDLE and RRC INACTIVE, as shown in Fig. 1. According to 5G standard [6], the MTD is in RRC CONNECTED state when an RRC connection has been established between itself and gNB. In this case, it can obtain data transmission resource for clearing its data queue. If an MTD is in RRC IDLE state, then its connection with gNB and 5GCN (5G Core Network) is released. Accordingly, it has to perform random access procedure to reestablish the connection with gNB and 5GCN.

Fig. 2 demonstrates the signaling diagram for the state transition from RRC IDLE to RRC CONNECTED. Start with the four-way handshake random access procedure, we can clearly see that in the best case, where there is no transmission failure, 12 messages have to be exchanged between the MTD and the gNB, and 9 of them are transmitted over the air interface. The entailed signaling overhead, additional delay and power consumption are negligible for traditional human-type communications, but comparatively significant for mMTC. To reduce the signaling overhead and additional latency, RRC INACTIVE state is introduced in 5G RRC state machine. It is a new RRC state compared to its counterpart in 4G system, where only RRC CONNECTED state and RRC IDLE state are included. The signaling diagram for the state transition from RRC INACTIVE to RRC CONNECTED is presented in Fig. 3.

Introducing RRC INACTIVE state brings twofold of benefits:

(1) If an MTD is in RRC INACTIVE state, then its RAN/CN connection and UE AS context are kept at the 5GCN and gNB while the on-the-air connection with gNB is released. Therefore, as shown in Fig. 3, although random access procedure is needed for connection re-establishment, fewer signalings are required, compared to the state transition from RRC IDLE state to RRC CONNECTED state, shown in Fig. 2.

(2) The MTD can transmit one small packet in the third step of random access procedure, while remains in the RRC INACTIVE state, which avoids signaling overhead and additional latency due to state transition and is beneficial for mMTC. Such kind of random access scheme is referred to as the Packet-Based Random Access scheme (PBRA), because each single packet in the device's queues has to contend for channel access. In the following, we will take a closer look at the signaling overhead of 5G networks with PBRA and study how to configure the system parameters for optimizing the network throughput performance.

III. System Model

This paper considers a single cell 5G system in which one gNB serves n MTDs. The arrival of data packets at each MTD follows a Bernoulli process with parameter λ, and the buffer of each MTD is infinite. With a busy buffer, the MTD would access gNB for data transmission via random access procedure. In the random access procedure, each MTD randomly selects one out of $M \geq 1$ orthogonal preambles and transmits via the physical random access channel to the gNB. If more than one MTDs transmit the same preamble simultaneously, then a collision occurs and all of them fail. The access request is successful as long as there is no collision. The time is slotted and assume the four-way handshake is completed in one time slot.

In this paper, we only consider the PBRA scheme. Accordingly, all MTDs are assumed to be in RRC INACTIVE state. In each access attempt, the MTD transmits one data packet[1] in the third step of the random access procedure, and do not shift into the RRC CONNECTED state. Moreover, since preambles are orthogonal to each other, MTDs do not affect each other's chance of successful access if they choose different preambles. For simplicity, we only consider the single-preamble case $M = 1$ in this paper. Note that the extension to the multi-preamble case can be implemented based on the multi-group model in [12].

A. Preliminary Analysis

To characterize behavior of each packet, we use the analytical model in [11], where a Markov chain $\mathbf{X} = \{X_j\}$ is established as shown in Fig. 4. There are two states of HOL packet: successful transmission (State T) and waiting to transmit (State 0).

[1]Since each access request transmission corresponds to one data packet transmission, the terminologies "access request" and "data packet" are used interchangeably.

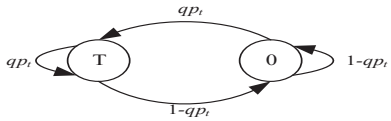


Fig. 4. State transition diagram of each individual access request.

Note that according to 5G standard [6], to control the intensive contention in massive access scenario, the Access Class Barring (ACB) scheme can be used, that is, each MTD transmits access request in each time slot with probability $q \in (0, 1]$. Let p_t denote the probability that the MTD can successfully transmit the packet in time slot $t = 1, 2, \ldots$. A fresh packet is initially in State T, and remains in State T if it passes the ACB check and is successfully transmitted with probability qp_t. If it passes the ACB check but encounters a collision, then it goes to State 0. In State 0, if it passes the ACB check and is successfully transmitted with probability qp_t, then it shifts to State T.

The steady-state probability distribution of the Markov chain in Fig. 4 can be derived as $\pi_T = pq,$ and $\pi_0 = 1 - pq$, where $p = \lim_{t\to\infty} p_t$ is the steady-state probability of successful transmission of packets. Note that π_T is the service rate of each node's queue. The offered load ρ of each node's queue can then be written as

$$\rho = \frac{\lambda}{\pi_T} = \frac{\lambda}{pq}. \tag{1}$$

B. Performance Metrics

This paper focuses on the throughput performance and signaling overhead of PBRA for mMTC in 5G networks. The throughput is defined as the long-term average number of successfully transmitted information bits per time slot and can be written as $\hat{\boldsymbol{\lambda}}_{out} = \hat{\lambda}_{out} L$, where $\hat{\lambda}_{out}$ is the long-term average number of successfully transmitted packets per time slot, and L is the number of information bits per packet.

Let TS denote the time-average amount of signaling overhead per time slot. To characterize the signaling overhead in PBRA, we let s (in a unit of bits) denote the average size of control messages exchanged between MTD and gNB[2]. Depending on whether the transmission is successful or not, we can consider the following two cases:

- If the MTD fails in the four-way handshake random access process, then the signaling overhead is $4s$. The long-term signaling overhead due to failed transmissions can then be written as $4sF$, where F denotes the average number of failed packet transmissions per slot.
- If the MTD succeeds in the random access process, then the signaling overhead is $3s$, because the third step is for data transmission. The long-term signaling overhead for successful transmissions can then be written as $3s\hat{\lambda}_{out}$.

We can finally conclude that

$$TS = 4sF + 3s\hat{\lambda}_{out}. \tag{2}$$

[2]It is clear that the sizes of control messages may be different from each other in the random access procedure. Nevertheless, the analysis can easily be extended to incorporate such practical consideration.

TABLE I
MAXIMUM THROUGHPUT AND THE OPTIMAL ACB FACTOR.

Region	$\{n,\lambda\}$	$\{\beta, L\}$	q^*	$\hat{\boldsymbol{\lambda}}_{\max}$
$S1$	$n\lambda < e^{-1}$	$\frac{4s\exp(-\mathbb{W}_0(-\hat{\lambda}))-s}{L} \le \beta$	$[\frac{\hat{\lambda}}{np_L}, -\frac{1}{n}\ln p_S]$	$L\hat{\lambda}$
$S2$		$\frac{4s\exp(-\mathbb{W}_0(-\hat{\lambda}))-s}{L} > \beta$	no solution	no solution
$S3$	$n\lambda \ge e^{-1}$	$\frac{(4e-1)s}{L} \le \beta$	$\frac{1}{n}$	Le^{-1}
$S4$		$\frac{3s}{L} < \beta < \frac{(4e-1)s}{L}$	$\frac{1}{n}\ln\frac{\beta L+s}{4s}$	$\frac{4Ls}{\beta L+s}\ln\frac{\beta L+s}{4s}$
$S5$		$0 < \beta \le \frac{3s}{L}$	no solution	no solution

It is clear that for mMTC services, a higher throughput with a lower signaling-to-throughput ratio $\frac{TS}{L\hat{\lambda}_{out}}$ are desired, where $\frac{TS}{L\hat{\lambda}_{out}}$ represents the average amount of signaling overhead per successfully-transmitted information bit. An excessively large $\frac{TS}{L\hat{\lambda}_{out}}$ usually indicates significant energy consumption for packet deliveries that is too costly for battery-limited MTDs. Therefore, in this paper, we let β denote the upper-bound of the signaling-to-throughput ratio, and study how to optimize the throughput performance while satisfying the signaling-to-throughput ratio constraint by properly tuning the ACB factor q, i.e.,

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\max} = \max_{0<q\le 1} \quad & \hat{\boldsymbol{\lambda}}_{out} \\ \text{s.t.} \quad & \tfrac{TS}{L\hat{\lambda}_{out}} \le \beta.\end{aligned} \tag{3}$$

In the following section, we will derive the expressions of $\hat{\boldsymbol{\lambda}}_{out}, TS$ and address the above optimization problem.

IV. PERFORMANCE OPTIMIZATION

For each HOL packet, it can be successfully transmitted if and only if all the other $n-1$ nodes are idle with an empty queue with probability $1-\rho$, or busy with a non-empty queue but not requesting transmission with probability $\rho(\pi_0+\pi_T)(1-q)$. Depending on $\rho = 1$ or not, we consider the saturated case and unsaturated case, separately, in the following.

1) Saturated Case: When the aggregate input rate $\hat{\lambda}(=n\lambda)$ is large, the network is likely to be in saturated case, where each nodes' queue is always busy and $\rho = 1$. The steady-state probability of successful transmission of HOL packets p in the saturated condition can be obtained as $p_A = (1-q)^{n-1} \overset{\text{with a large } n}{\approx} \exp(-nq)$, based on which the average number of failed packet transmissions per slot can be further written as

$$F = nq(1-p_A). \tag{4}$$

In saturated case, the long-term average number of successfully transmitted packets per time slot is determined by the aggregate service rate, i.e.,

$$\hat{\lambda}_{out} = n\pi_T = n\exp(-nq)\,q. \tag{5}$$

By combining (2), (4)–(5), the optimization problem in (3) can be rewritten as

$$\begin{aligned}\hat{\boldsymbol{\lambda}}^{sa}_{\max} = \max_{0<q\le 1} \quad & Ln\exp(-nq)\,q \\ \text{s.t.} \quad & \tfrac{4s\exp(nq)-s}{L} \le \beta.\end{aligned} \tag{6}$$

Solving (6) yields the maximum throughput in saturated case as

$$\hat{\boldsymbol{\lambda}}^{sa}_{\max} = \begin{cases} Le^{-1} & \text{if } \frac{(4e-1)s}{L} \le \beta, \\ \frac{4Ls}{\beta L+s}\ln\frac{\beta L+s}{4s} & \text{if } \frac{3s}{L} < \beta < \frac{(4e-1)s}{L}, \\ \text{no solution} & \text{if } 0 < \beta \le \frac{3s}{L}, \end{cases} \tag{7}$$

which is achieved when the ACB factor is set to

$$q^*_{sa} = \begin{cases} \frac{1}{n} & \text{if } \frac{(4e-1)s}{L} \le \beta, \\ \frac{1}{n}\ln\frac{\beta L+s}{4s} & \text{if } \frac{3s}{L} < \beta < \frac{(4e-1)s}{L}, \\ \text{no solution} & \text{if } 0 < \beta \le \frac{3s}{L}. \end{cases} \tag{8}$$

Note that in the ideal case, where all packets can be successfully transmitted in one-shot, the signaling-to-throughput ratio is $\frac{3s}{L}$. Accordingly, there is no solution for (6) if the upper-bound of this ratio $\beta \le \frac{3s}{L}$.

2) Unsaturated Case: As the number of devices or the traffic input rate decreases, the network may become unsaturated and we have $\rho < 1$. The probability of successful transmission of packets is given by $p = (1-\rho+\rho(1-q))^{n-1} \overset{\text{with a large } n}{\approx} \exp\left(-\frac{\hat{\lambda}}{p}\right)$, which has two non-zero roots:

$$p_L = \exp(\mathbb{W}_0(-\hat{\lambda})), \quad \text{and} \quad p_S = \exp(\mathbb{W}_{-1}(-\hat{\lambda})). \tag{9}$$

By following the approximate trajectory analysis in [11], we can find that only p_L is the steady-state point. Based on (1) and (9), the average number of failed packet transmissions per slot can be further written as

$$F = n\rho q(1-p_L) = \hat{\lambda}\left(\exp(-\mathbb{W}_0(-\hat{\lambda})) - 1\right). \tag{10}$$

Since the network is stable in the unsaturated case, it is straightforward to have [11]

$$\hat{\lambda}_{out} = \hat{\lambda}. \tag{11}$$

By combining (2), (10)–(11), we can rewrite the optimization problem in (3) as

$$\begin{aligned}\hat{\boldsymbol{\lambda}}^{unsa}_{\max} = \max_{q\in S_L} \quad & L\hat{\lambda} \\ \text{s.t.} \quad & \tfrac{4s\exp(-\mathbb{W}_0(-\hat{\lambda}))-s}{L} \le \beta.\end{aligned} \tag{12}$$

Solving (12) yields the maximum throughput in unsaturated case as

$$\hat{\boldsymbol{\lambda}}^{unsa}_{\max} = \begin{cases} L\hat{\lambda} & \text{if } \frac{4s\exp(-\mathbb{W}_0(-\hat{\lambda}))-s}{L} \le \beta, \\ \text{no solution} & \text{if } \frac{4s\exp(-\mathbb{W}_0(-\hat{\lambda}))-s}{L} > \beta, \end{cases} \tag{13}$$

which is achieved when the ACB factor is set to

$$q^*_{unsa} \in S_L = [q_l, q_u] = \begin{cases} [\frac{\hat{\lambda}}{np_L}, -\frac{1}{n}\ln p_S] & \text{if } n\lambda < e^{-1}, \\ \emptyset & \text{else}, \end{cases} \tag{14}$$

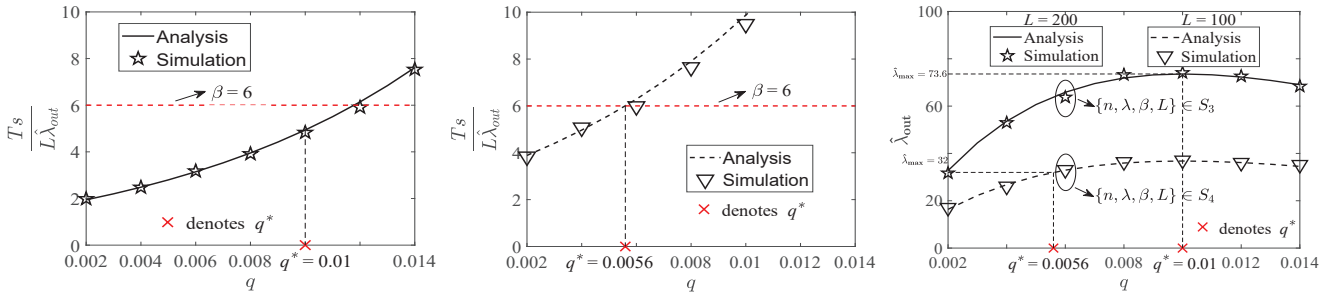


Fig. 5. Signaling-to-throughput ratio $\frac{TS}{L\hat{\lambda}_{out}}$ and throughput $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ versus the ACB factor q. $\lambda = 0.01$. $n = 100$. $s = 100$. $L = 100$ bits or $L = 200$ bits. $\beta = 6$. (a) $\frac{TS}{L\hat{\lambda}_{out}}$ versus q, $L = 200$ bits. $\{n, \lambda, \beta, L\} \in S3$. (b) $\frac{TS}{L\hat{\lambda}_{out}}$ versus q, $L = 100$ bits. $\{n, \lambda, \beta, L\} \in S4$. (c) $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ versus q.

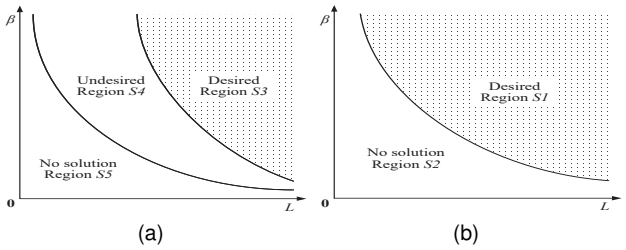


Fig. 6. (a) A graphic illustration of Regions $S3, S4$ and $S5$ in saturated case in terms of L. (b) A graphic illustration of Regions $S1$ and $S2$ in unsaturated case in terms of L.

where S_L is referred to as the absolute-stable region, within which the network is unsaturated and operates at the desired steady-state point p_L. Otherwise, the network operates at the undesired steady-state point p_A and becomes saturated, where $p_A \leq p_L$ [11]. It is very interesting to see from (12) that both the network throughput and signaling-to-throughput ratio are independent of the ACB factor q, which implies that in this case, q can be a random value chosen from the absolute-stable region S_L.

3) Summary: So far, we have obtained the optimal ACB factor for maximizing the network throughput with the signaling-to-throughput ratio constraint in saturated and unsaturated cases, respectively. To summarize, we present the results in Table I, where we define region Si in terms of $\{n, \lambda, \beta, L\}$ for $i \in \{1, 2, 3, 4, 5\}$. Regions $S1$ and $S2$ belong to the unsaturated case while Regions $S3, S4$ and $S5$ belong to the saturated case.

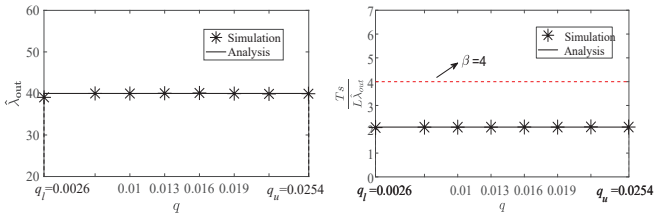
- In regions $S1$ and $S3$, the constraint on the signaling-to-throughput ratio can be guaranteed and the maximum throughput (i.e., $L\hat{\lambda}$ in unsaturated case and Le^{-1} in saturated case) is achieved as well. Accordingly, regions $S1$ and $S3$ are referred to as **the desired regions**.
- In region $S4$, the constraint on the signaling-to-throughput ratio can be guaranteed while the maximum throughput is sacrificed because $\frac{4Ls}{\beta L+s} \ln \frac{\beta L+s}{4s} < Le^{-1}$. Accordingly, region $S4$ is referred to as **the undesired region**.
- In regions $S2$ and $S5$, there is no solution for the optimization problem (3), implying that in this case the constraint on the signaling-to-throughput ratio cannot be satisfied regardless of what value of q. Therefore, they are referred to as **the no solution region**.

To take a further look at how the regions Si vary with the system parameters, Fig. 6a and Fig. 6b give a graphic illustration of Regions $S3, S4$ and $S5$ in saturated case and that of Regions $S1$ and $S2$ in unsaturated case in terms of the packet length L and the signaling-to-throughput ratio threshold β. We can clearly see from Fig. 6a that if the packet length L grows, then the desired region $S3$ rapidly expands while the undesired region $S4$ and the no solution region $S5$ shrinks. Similar observations can also be obtained in Fig. 6b, which indicates that with PBRA scheme, the network can always be benefited from a larger packet length. However, in practical 5G system, due to limited uplink time-frequency resources, the packet length that the third step of the random access procedure can carry is limited as well. On the other hand, to enlarge the desired region $S3$ or $S1$, Fig. 6 also shows that the network can also enlarge the upper-bound of the signaling-to-throughput ratio β if the MTC service is insensitive to the signaling overhead, delay or power consumption.

V. Simulation and Discussion

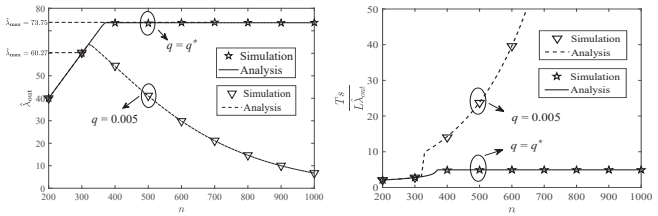
In this section, simulation results are presented to verify the preceding theoretical analysis. The simulation setting is the same as the system model described in Section II, and we omit the details here. Each simulation lasts for 10^7 time slots.

Fig. 5 demonstrates how the signaling-to-throughput ratio $\frac{TS}{L\hat{\lambda}_{out}}$ and the throughput $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ vary with the ACB factor q with the upper-bound of the signaling-to-throughput ratio $\beta = 6$, the number of nodes $n = 100$, input rate $\lambda = 0.01$ and the network is in saturated case. Recall that it has been shown in the above section, if the network is in the desired region $S3$, then by optimally tuning the ACB factor $q = q^* = \frac{1}{n}$, the signaling-to-throughput ratio constraint can be guaranteed and the maximum throughput Le^{-1} can be achieved as well. Therefore, we can see from Fig. 5a that with the packet length $L = 200$ bits, the network belongs to the desired region $S3$, i.e., $\{n, \lambda, \beta, L\} \in S_3$. When $q = \frac{1}{n}$, the signaling-to-throughput



(a) (b)

Fig. 7. Throughput $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ and signaling-to-throughput $\frac{TS}{L\hat{\lambda}_{out}}$ versus the ACB factor q in unsaturated case. $\lambda = 0.002$. $n = 100$. $s = 100$. $L = 200$ bits. $\beta = 4$.



(a) (b)

Fig. 8. Throughput $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ and signaling-to-throughput $\frac{TS}{L\hat{\lambda}_{out}}$ versus the number of MTDs n. $\lambda = 0.001$. $s = 100$. $L = 200$ bits. $\beta = 6$. $q = 0.005$ or $q = q^*$.

ratio $\frac{TS}{L\hat{\lambda}_{out}}$ is below the upper-bound β, indicating that the constraint can be satisfied with $q = \frac{1}{n}$. As shown in Fig. 5c, the throughput is maximized at $\hat{\boldsymbol{\lambda}}_{\mathbf{max}} = Le^{-1} = 73.6$ bit/slot. On the other hand, if $L = 100$ bits, then the network falls into the undesired region $S4$, i.e., $\{n, \lambda, \beta, L\} \in S_4$, and $\frac{TS}{L\hat{\lambda}_{out}} > \beta$ with $q = \frac{1}{n}$, as shown in Fig. 5b. In this case, only by choosing the ACB factor $q = \frac{1}{n} \ln \frac{\beta L + s}{4s}$ can signaling overhead constraint be satisfied. Yet, as shown in Fig. 5c, the throughput is maximized at $\hat{\boldsymbol{\lambda}}_{\mathbf{max}} = \frac{4Ls}{\beta L + s} \ln \frac{\beta L + s}{4s} = 32 < Le^{-1} \approx 36.8$ bit/slot, implying the throughput performance is sacrificed.

Fig. 7 demonstrates how the throughput $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ and signaling-to-throughput ratio $\frac{TS}{L\hat{\lambda}_{out}}$ vary with $q \in S_L = [q_l, q_u]$ with $\beta = 4$, $n = 100$, $\lambda = 0.002$ and the network is in unsaturated case. Table I has shown that both $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ and $\frac{TS}{L\hat{\lambda}_{out}}$ are insensitive to the variation of q if $q \in S_L = [q_l, q_u]$. The maximum throughput $\hat{\boldsymbol{\lambda}}_{\mathbf{max}} = n\lambda L = 40$ bit/slot is achieved as long as the threshold $\frac{4s \exp(-\mathbb{W}_0(-\hat{\lambda})) - s}{L} \leq \beta$. Those observations are clearly confirmed by the simulation results in Fig. 7.

Finally, Fig. 8 shows how the throughput $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ and the signaling-to-throughput ratio $\frac{TS}{L\hat{\lambda}_{out}}$ vary with the network size $n \in [200, 1000]$ with the ACB factor q optimally tuned according to Table I, i.e., $q = q^*$, or fixed at $q = 0.005$. We can see that when the number of nodes n is small, e.g., $n = 200$, both $\hat{\boldsymbol{\lambda}}_{\boldsymbol{out}}$ and $\frac{TS}{L\hat{\lambda}_{out}}$ increase with n while remain same for $q = q^*$ or $q = 0.005$, because they are insensitive to the variation of q if $q \in S_L = [q_l, q_u]$. In sharp contrast, when n is large, if q is fixed, then the throughput rapidly drops and the signaling-to-throughput ratio grows, indicating an intolerably deteriorated network performance. However, if q is adaptively configured according to n, we can see that the maximum throughput can always be achieved and the signaling-to-throughput ratio $\frac{TS}{L\hat{\lambda}_{out}}$ remains low, which corroborates that adaptive tuning of the ACB factor q is indispensable especially for massive access scenarios.

VI. Conclusion

This paper presents the throughput and signaling overhead analysis of mMTC in 5G networks with PBRA. Based on the analytical framework in [11], closed-form expressions of the throughput and signaling-to-throughput ratio in both unsaturated and saturated cases are obtained. By further taking the signaling-to-throughput ratio constraint into consideration, the maximum throughput is characterized by optimizing the ACB factor. The analysis sheds important light on practical 5G network design with PBRA for efficient support of mMTC. It reveals that to achieve the optimum throughput performance while keeping the signaling-to-throughput ratio below a certain level, a larger data packet length is preferred and the ACB factor has to be adaptively tuned according to the network size.

Note that in this paper, we do not consider the CBRA scheme. In practice, the PBRA scheme and CBRA scheme coexist in 5G system. How to dynamically select a random access mechanism according to the traffic characteristics of mMTC is an interesting topic that deserves much attention in future study.

References

[1] Cisco whitepaper, "Cisco visual networking index: global mobile data traffic forecast update, 2018–2023," Mar. 2020.

[2] A. Höglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.

[3] S. Ryoo, J. Jung and R. Ahn,"Energy efficiency enhancement with RRC connection control for 5G new RAT," in *Proc. IEEE WCNC*, Barcelona, Spain, Apr. 2018.

[4] S. Hailu, M. Säily, and O. Tirkkonen, "RRC state handling for 5G," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 106–113, Jan. 2019.

[5] A. Khlass, D. Laselva and R. Jarvela, "On the flexible and performance-enhanced radio resource control for 5G NR networks," in *Proc. IEEE VTC-Fall*, Honolulu, USA, Sept. 2019.

[6] 3GPP TS 38.331 V15.6.0, "5G;NR;Radio Resource Control (RRC);Protocol specification," Jul. 2019.

[7] M. Feguson, "On the control, stability, and waiting time in a slotted ALOHA random-access system," *IEEE Trans. Commun.*, vol. 23, no. 11, pp. 1306–1311, Nov. 1975.

[8] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[9] C. Di, B. Zhang, Q. Liang, S. Li and Y. Guo, "Learning automata based access class barring scheme for massive random access in machine-to-machine communications," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6007–6017, Aug. 2019.

[10] T. Wan and A. U. Sheikh, "Performance and stability analysis of buffered slotted Aloha protocols using tagged user approach," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 582-593, Mar. 2000.

[11] L. Dai, "Stability and delay analysis of buffered Aloha networks," *IEEE. Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2707–2719, Aug. 2012.

[12] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Throughput optimization with a finite data transmission rate," *IEEE. Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5749–5763, Dec. 2019.