Massive Random Access of Machine-to-Machine Communications in LTE Networks: Modeling and Throughput Optimization

Wen Zhan, *Student Member, IEEE*, and Lin Dai, *Senior Member, IEEE*

***Abstract*— A key challenge for enabling machine-to-machine (M2M) communications in long-term evolution (LTE) networks is the intolerably low access efficiency in the presence of massive access requests. To address this issue, a new analytical framework is proposed in this paper to optimize the random access performance of the M2M communications in LTE networks. Specifically, a novel double-queue model is established, which can both incorporate the queueing behavior of each machine-type device (MTD) and be scalable in the massive access scenarios. To evaluate the access efficiency, the network throughput is further characterized, and optimized by properly choosing the backoff parameters including the access class barring (ACB) factor and the uniform backoff (UB) window size. The analysis reveals that the maximum network throughput is solely determined by the number of preambles, and can be achieved by either tuning the ACB factor or the UB window size based on statistical information such as the traffic input rate of each MTD. Simulation results corroborate that with the optimal tuning of backoff parameters, the network throughput can remain at the highest level regardless of how many MTDs in the network, and is robust against feedback errors of the traffic input rate and burstiness of data arrivals.**

***Index Terms*— Machine-to-machine (M2M) communications, modeling, throughput, optimization, random access.**

I. Introduction

MACHINE-TO-MACHINE (M2M) communications is one of the most important enabling technologies for the emerging Internet-of-Things paradigm that has found wide applications in various domains such as smart grid, intelligent transportation, and e-health. Aimed at providing pervasive connectivity for autonomous devices with minimum or no human intervention, M2M communications has garnered significant attention in recent years, and is expected to play an instrumental role in next-generation data communication networks [1], [2].

To facilitate M2M communications, the most natural and appealing solution is the Long Term Evolution (LTE) cellular system owing to its ubiquitous coverage. With many Machine-Type Devices (MTDs) attempting to initiate connection with the Base Station (BS), however, the deluge of access requests will cause severe congestion with low chances of success [3]. Considering the exponential increase in the number of MTDs [4], there is an urgent need to optimize the access efficiency of LTE networks for accommodating massive access from M2M communications.

For LTE networks, the random access process is designed based on Aloha [5], the simplest and one of the most representative random-access schemes. To understand the new challenges that M2M communications may pose to LTE networks, there have been a plethora of studies on the random access performance with massive access requests from MTDs, most of which follow the classical analysis of Aloha.

A. Modeling: Aggregate-Traffic-Centric or Node-Centric

Due to the uncoordinated nature of transmitters, the aggregate traffic, i.e., the total number of requests in each time slot, in a random-access network varies with time. In Abramson's landmark paper [6], the aggregate traffic of Aloha networks was modeled as a Poisson random variable with parameter G, based on which the network throughput, i.e., the average number of successful requests per time slot, can easily be obtained as $G\mathrm{e}^{-G}$. Such an aggregate-traffic-centric modeling approach captures the essence of contention among nodes, and has been widely adopted in the follow-up studies on various random-access networks [7]–[10]. For M2M communications in LTE networks, in addition to Poisson [11]–[14], Beta distribution was also adopted to model the number of requests (which were also referred to as "active MTDs" or "backlogged MTDs" in the literature) in the bursty scenarios [15]–[17].

Note that the aggregate traffic is indeed determined by the input traffic and backoff mechanism of each node. To characterize its distribution, a more refined model should be established to include the queueing behavior of nodes. Various node-centric models have been proposed for Aloha networks, yet most of them lead to prohibitively high complexity when considering the interactions among nodes' queues [18]–[20]. For M2M communications in LTE networks, queueing models for individual MTDs were established in [21] and [22] to evaluate the network performance under various backoff schemes. In both cases, iterative/recursive algorithms need to be adopted to solve the system equations, with which the computational

Manuscript received February 23, 2017; revised August 17, 2017 and November 27, 2017; accepted January 24, 2018. Date of publication February 9, 2018; date of current version April 8, 2018. This work was supported by the CityU Strategic Research under Grant 7004843. This paper was presented in part at the IEEE International Conference on Communications, Paris, France, May 2017. The associate editor coordinating the review of this paper and approving it for publication was R. Hu. *(Corresponding author: Wen Zhan.)*

The authors are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: wzhan5-c@my.cityu.edu.hk; lindai@cityu.edu.hk).

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/TWC.2018.2803083

complexity increases with the number of MTDs, rendering the models unscalable in massive access scenarios.

As we pointed out in [23] and [24], a random-access network can be regarded as a multi-queue-single-server system, and the key to performance analysis lies in proper characterization of the service time distribution, which is difficult to obtain with either each node's queue completely ignored or interactions among nodes' queues taken into full consideration. To reduce the modeling complexity, we demonstrated in [23] and [24] that a scalable node-centric model can be established by treating each node's queue as an independent queueing system with identically distributed service time, which consists of two parts: 1) state characterization of each individual head-of-line (HOL) packet; and 2) characterization of network steady-state points based on the fixed-point equations of limiting probability of successful transmission of HOL packets. The proposed modeling methodology has been successfully applied to various random-access networks, and shown to be accurate [25]–[28]. In most scenarios, the network steady-state points can be obtained as explicit functions of key system parameters such as the number of nodes and the backoff window size/transmission probability of each node, based on which the optimal network performance can further be characterized.

Note that the above model cannot be directly applied to M2M communications in LTE networks. Specifically, a basic assumption of the aforementioned studies is that each single packet in nodes' queues has to contend for channel access. In LTE networks, however, a connection would first be established between a device and the BS before the device starts to transmit its data packets [29]. That is, each device with data packets to transmit first sends a connection request to the BS, and if a device's request is successfully received, then the BS will allocate resource blocks for the device to clear its data queue. We can see that the connection-based random access adopted in the LTE networks differs from the conventional packet-based random access in that the data packets do not contend for the channel individually, which calls for new scalable node-centric models.

B. Optimal Tuning of Backoff Parameters

It has long been observed that a random-access network may suffer from significant performance degradation if backoff parameters are not properly selected. For Aloha networks, adaptive strategies have been developed to adjust the transmission probability of each node based on the realtime information of the aggregate traffic [30]–[32]. As such information is usually unknown at the transmitter side, decentralized control algorithms have been proposed to estimate the aggregate traffic [33], [34].

To improve the probability of successful access of MTDs in LTE systems, various algorithms were also proposed to adaptively tune backoff parameters including the Access Class Barring (ACB) factor (i.e., the initial transmission probability of each MTD) [35]–[38] and the Uniform Backoff (UB) window size [39], or system resources including the number of preambles and the number of slots for random access [40]–[42], according to the number of access requests in each time slot. Similar to Aloha networks, the effectiveness of adaptive tuning depends on how accurate the estimation of aggregate traffic is. Therefore, most efforts have been focused on developing estimation algorithms to track the time-varying number of access requests based on some measurable network status [36]–[42].

While the aforementioned developments have been substantial, how to optimally tune the backoff parameters of each device to maximize the access efficiency has remained largely unknown. The challenge originates from the lack of proper modeling of the random access process of LTE networks. As we mentioned previously, existing models either exclude the device-level input parameters by ignoring the queueing behavior of each MTD [11]–[17], or become unscalable in the massive access scenarios [21], [22], neither of which can facilitate the optimization of access performance of massive MTDs. Moreover, most of the adaptive tuning algorithms were developed based on the estimation of time-varying network status [36]–[42], which in practice is difficult to track accurately. As we will demonstrate in this paper, if the objective is to optimize the long-term network performance such as the network throughput, such estimation is indeed unnecessary. Instead, the optimal tuning of backoff parameters can solely be based on statistical information such as the traffic input rate of each device and the total number of devices.

C. Our Contributions

In this paper, we propose a new analytical framework for optimizing the access efficiency of M2M communications in LTE networks. Specifically, to capture the key feature of connection-based random access process, a novel double-queue model is established, where each MTD has one request queue and one data queue, and only the request queue is involved in the contention. By characterizing the state transition of each access request, the network steady-state points are obtained as the non-zero roots of the single fixed-point equation of the limiting probability of successful transmission of access requests. The complexity is independent of the number of MTDs even with the queueing behavior of each MTD taken into consideration, which is highly attractive in the massive access scenario.

To evaluate the access efficiency, the network throughput, which is defined as the average number of successful access requests per time slot, is derived as a function of 1) the number of preambles M, 2) the number of MTDs n, 3) the traffic input rate of each MTD λ, 4) the ACB factor q and 5) the UB window size W. Given the system parameters such as the number of preambles M, the network size n and the traffic input rate of each MTD λ, the network throughput can further be maximized by optimally tuning the backoff parameters of each MTD, i.e., the ACB factor q and the UB window size W. Explicit expressions of the maximum network throughput and the corresponding optimal backoff parameters are obtained, which reveal that the maximum network throughput is solely determined by the number of preambles M, and yet to achieve it, either the ACB factor q or the UB window size W should be tuned based on the number of MTDs n, the traffic input rate of each MTD λ and the number of preambles M.

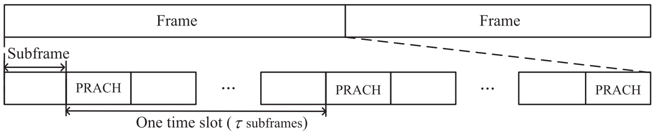


Fig. 1. Frame structure of the LTE system in the Frequency Division Duplex (FDD) mode.

It is further shown that although the tunings of ACB factor and UB window size are equally effective in optimizing the network throughput performance, the latter is more robust against the variation of network size and traffic input rate.

The remainder of this paper is organized as follows. Section II presents the system model. The network steady-state points and throughput with one single preamble are characterized in Section III and Section IV, respectively, and extended to the multi-preamble scenario in Section V. Practical considerations, such as the effects of outdated information of system parameters and bursty input traffic, are discussed in Section VI. Finally, concluding remarks are summarized in Section VII.

II. System Model

Consider a single-cell LTE system with $n \in \{1, 2, \ldots\}$ MTDs attempting to access the BS. Different from [43]–[47] where various access schemes were proposed for M2M communications, in this paper, we focus on the random access process of the current LTE networks without assuming any modification to the standard. Moreover, out of the four steps of the random access process of LTE standard [29], the success of contention is mainly determined by the first step [3], [5]. Therefore, similar to [12], [13], [17], [21], [22], [35]–[38], and [40]–[42], in this paper, we only focus on the access performance of the first step.[1]

Specifically, in the random access process, each MTD randomly selects one out of $M \in \{1, 2, \ldots\}$ orthogonal preambles and transmits via the Physical Random Access CHannel (PRACH) to the BS [29]. The PRACH consists of a series of subframes that appear periodically [48], as Fig. 1 shows. If more than one MTDs transmit the same preamble over the same PRACH subframe, then a collision occurs and all of them fail. The access request is successful if and only if there is one single MTD transmitting for a given preamble at each PRACH subframe. Upon successful request, a connection will be established and the BS will allocate resources to the MTD for its data transmission.

In this paper, we are interested in the network throughput performance of the random access process, which is evaluated by the average number of MTDs that can successfully access the BS per PRACH subframe. Define a time slot as the interval between two consecutive PRACH subframes, as shown in Fig. 1. The network throughput $\hat{\lambda}_{\text{out}}$ can then be defined as the average number of successful access requests per time slot.

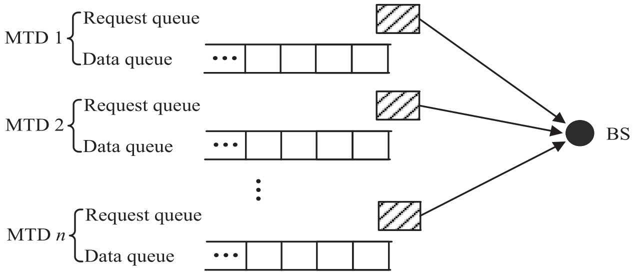


Fig. 2. Double-queue model of each MTD.

Note that due to the orthogonality of preambles, MTDs do not affect each other's chance of successful access if they choose different preambles. Only those who share the same preamble would contend with each other. Therefore, for better illustration, let us start by considering the scenario where all n MTDs share one preamble, i.e., $M = 1$. The analysis will further be extended to the multi-preamble scenario in Section V.

A. Double-Queue Model of Each MTD

For each MTD, a connection would first be established with the BS before it starts to transmit its data packets. Specifically, it generates an access request once it has data packets in the buffer. The access request stays until it is successfully transmitted, upon which the BS will assign sufficient resources for the MTD to clear its data buffer. For such connection-based random access, we propose a double-queue model, that is, each MTD has one data queue and one request queue, as Fig. 2 illustrates. Assume that the data buffer has an infinite size and the arrivals of data packets follow a Bernoulli process[2] with parameter $\lambda \in (0, 1)$. Each newly arrival data packet generates an access request, but only one request can be kept since each MTD can have at most one ongoing access request regardless of how many data packets in its buffer [29]. Each MTD's request queue can then be modeled as a $Geo/G/1/1$ queue. The service time distribution is clearly determined by the state transition of each access request, which will be characterized in the following subsection.

B. State Characterization of Access Request

According to the current LTE standard [29], [49], each MTD needs to perform the ACB check before transmitting its access request. That is, the MTD generates a random number between 0 and 1, and compares it with the ACB factor $q \in (0, 1]$. If the number is less than q, then the MTD proceeds to transmit the access request. Otherwise, it is barred temporarily. Once the MTD passes the ACB check but involves in a collision, it randomly selects a value from $\{0, \ldots, W_s\}$, where W_s is the UB window size in unit of milliseconds,[3] and counts down until it reaches zero.

[1] Note that it was shown in [11], [14]–[16], and [39] that the remaining three steps of the random access process may pose additional limitations on the access performance, e.g., the BS may not be able to acknowledge a successfully transmitted access request due to the downlink control channel resource constraint. How to refine the proposed analytical framework to optimize the access performance with the above constraint taken into consideration is an important future direction.

[2] Note that although the analysis is based on the assumption of Bernoulli arrivals, simulation results with bursty data arrivals will be presented in Section VI-B to verify the effectiveness of the analysis.

[3] Note that the UB window size W_s in the LTE standard has the unit of milliseconds [29]. In a time-slotted system, it needs to be converted into the unit of time slots, which is denoted as $W \in \{1, 2, \ldots\}$, and we have $W = \left\lfloor \frac{W_s}{\tau} \right\rfloor + 1$, where τ is the length of the time slot.

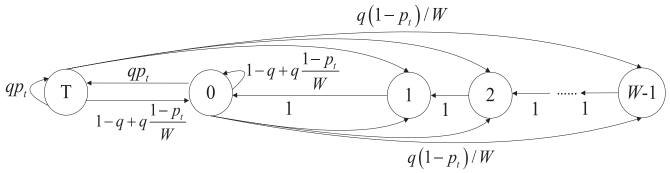


Fig. 3. State transition diagram of each individual access request.

Fig. 3 shows the state transition process of each individual access request. Let p_t denote the probability of successful transmission of access requests at time slot $t = 1, 2, \ldots$. As Fig. 3 illustrates, a fresh access request 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, it then goes to State $i \in \{0, 1, \ldots, W-1\}$ with equal probability $\frac{q(1-p_t)}{W}$. Otherwise, it shifts to State 0. In State $i \in \{1, 2, \ldots, W-1\}$, the access request counts down at each time slot until it reaches State 0. In State 0, the access request remains in State 0 if it fails in the ACB check. If it passes the ACB check and is successfully transmitted, it shifts back to State T. Otherwise, it goes to State $i \in \{0, 1, \ldots, W-1\}$ with equal probability $\frac{q(1-p_t)}{W}$.

The steady-state probability distribution of the Markov chain in Fig. 3 can be obtained as

$$\begin{cases} \pi_T = \left(\dfrac{1}{qp} + \dfrac{(1-p)(W-1)}{2p}\right)^{-1}, \\ \pi_0 = \dfrac{1-qp}{qp}\pi_T, \\ \pi_j = \dfrac{(1-p)(W-j)\pi_T}{pW}, \quad j = 1, 2, \ldots, W-1, \end{cases} \tag{1}$$

where $p = \lim_{t\to\infty} p_t$ is the steady-state probability of successful transmission of access requests. Note that π_T is the service rate of each request queue as the successful output occurs if and only if it is in State T.

C. Network Throughput

In this paper, we are interested in the network throughput,[4] which is denoted by $\hat{\lambda}_{\text{out}}$ and defined as the average number of successful access requests per time slot. According to Fig. 2, the network throughput is determined by the aggregate departure rate of n request queues. Based on the $Geo/G/1/1$ model of each request queue, the network throughput can be obtained as

$$\hat{\lambda}_{\text{out}} = \hat{\lambda}(1-\rho), \tag{2}$$

where $\hat{\lambda} = n\lambda$ denotes the aggregate input rate of MTDs, and ρ denotes the probability that each request queue is nonempty, which is given by [51]

$$\rho = \frac{\lambda}{\lambda + \pi_T}. \tag{3}$$

By combining (1), (2) and (3), we have

$$\hat{\lambda}_{\text{out}} = \frac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{qp} + \frac{(1-p)(W-1)}{2p}\right) + 1}. \tag{4}$$

We can see from (4) that the network throughput $\hat{\lambda}_{\text{out}}$ is closely determined by p, the steady-state probability of successful transmission of access requests. In the following section, we will specifically characterize the network steady-state points based on the fixed-point equation of p.

III. Steady-State Point Analysis

For any given MTD, its access request is successful if and only if all the other $n-1$ devices are either with an empty request queue, or busy with a non-empty request queue but not transmitting. For each MTD, the probability that it has an empty request queue is $1-\rho$, and the probability that it has a non-empty request queue but not transmitting is $\rho\left(\sum_{j=1}^{W-1}\pi_j + (1-q)(\pi_0 + \pi_T)\right)$, according to Fig. 3. The steady-state probability of successful transmission of access requests, p, can then be obtained as

$$\begin{aligned} p &= \left(1 - \rho + \rho\left(\sum_{j=1}^{W-1}\pi_j + (1-q)(\pi_0+\pi_T)\right)\right)^{n-1} \\ &= \left(1 - \frac{\rho\pi_T}{p}\right)^{n-1}. \end{aligned} \tag{5}$$

By combining (1), (3) and applying $n-1 \approx n$, $(1-x)^n \approx \exp\{-nx\}$ for $0 < x < 1$ if n is large, (5) can be approximated by

$$p \overset{\text{with a large } n}{\approx} \exp\left(-\frac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{q} + \frac{W-1}{2}\right) + p\left(1 - \frac{\hat{\lambda}(W-1)}{2n}\right)}\right). \tag{6}$$

Theorem 1 shows that (6) has either one or three non-zero roots.[5]

Theorem 1: The fixed-point equation (6) of p has three non-zero roots $0 < p_A \le p_S \le p_L \le 1$ if $n > 2\left(\frac{2}{q} + W - 1\right)$ and $\hat{\lambda}_1 \le \hat{\lambda} \le \hat{\lambda}_2$, where

$$\hat{\lambda}_1 = \frac{2n}{\dfrac{n - \frac{2}{q} - W + 1 - \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}{\exp\left(\dfrac{-2n}{n - \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}\right)} + W - 1}, \tag{7}$$

[4] It is worth mentioning that the access delay performance can also be characterized based on the proposed double-queue model. Specifically, for each access request, the probability generating function of its service time (i.e., the access delay) can be obtained from the Markov chain given in Fig. 3, based on which the moments of access delay can further be derived as explicit functions of backoff parameters including the ACB factor q and the UB window size W. Due to limited space, in this paper, we limit our focus to the network throughput analysis. How to optimize the delay performance by properly choosing the backoff parameters is another important issue that will be addressed in our future work.

[5] Note that it was also observed in [28] and [50] that for Aloha and CSMA networks with a finite retry limit of HOL packets, the fixed-point equation of steady-state points may have one or three non-zero roots. Yet they should be distinguished from this paper as they assume packet-based random access, rather than connection-based random access.

$$\hat{\lambda}_2 = \frac{2n}{\frac{n-\frac{2}{q}-W+1+\sqrt{n\left(n-\frac{4}{q}-2(W-1)\right)}}{\exp\left(\frac{-2n}{n+\sqrt{n\left(n-\frac{4}{q}-2(W-1)\right)}}\right)}+W-1}. \tag{8}$$

Otherwise, (6) has only one non-zero root $0 < p_L \leq 1$.

Proof: See Appendix A. ■

Note that not all the roots of (6) are steady-state points. We follow the approximate trajectory analysis proposed in [23], and find that:

1) If (6) has only one non-zero root p_L, then p_L is a steady-state point;
2) If (6) has three non-zero roots $p_A \leq p_S \leq p_L$, then only p_L and p_A are steady-state points. Similar to [52], we refer to p_L as the desired steady-state point and p_A as the undesired steady-state point.

A. Bistable Region and Monostable Region

It is clear from Theorem 1 that the number of steady-state points is determined by the number of MTDs n, the aggregate input rate of MTDs $\hat{\lambda}$, the ACB factor q and the UB window size W. Accordingly, we can define the following stable regions:

- **Bistable region** $\mathcal{B} = \left\{(n, \hat{\lambda}, q, W) | n > 2\left(\frac{2}{q} + W - 1\right), \hat{\lambda}_1 \leq \hat{\lambda} \leq \hat{\lambda}_2\right\}$, in which the network has two steady-state points p_L and p_A.
- **Monostable region** $\mathcal{M} = \bar{\mathcal{B}}$, in which the network has only one steady-state point p_L.

A graphical illustration of the bistable region $\mathcal{B}$ and monostable region $\mathcal{M}$ is presented in Fig. 4. It can be observed from Fig. 4 that when $n \leq 2\left(\frac{2}{q} + W - 1\right)$, the network always falls into the monostable region $\mathcal{M}$ and operates at the desired steady-state point p_L, regardless of the aggregate input rate $\hat{\lambda}$. On the other hand, if the aggregate input rate $\hat{\lambda} \geq \frac{4e^{-2}}{1+\frac{W-1}{\frac{2}{q}+W-1}e^{-2}}$, then the network stays in the monostable region $\mathcal{M}$, regardless of the number of MTDs n.

Corollary 1 further summarises the monotonicity property of the steady-state points with regard to the aggregate input rate $\hat{\lambda}$, the number of MTDs n, the ACB factor q and the UB window size W.

Corollary 1: The steady-state points p_L *and* p_A *are monotonic decreasing functions of* $\hat{\lambda}$, n *and* q, *and monotonic increasing functions of* W.

Proof: See Appendix B. ■

Fig. 5 demonstrates how the steady-state points p_L and p_A vary with the aggregate input rate $\hat{\lambda}$ under different values of the ACB factor q and the UB window size W. It can be seen from Fig. 5 that when the aggregate input rate $\hat{\lambda}$ is low, with a smaller ACB factor q or larger UB window size W, the network has a higher chance of falling into the monostable region with a single steady-state point p_L. As q increases or W decreases, the network may shift to the bistable region with two steady-state points p_L and p_A. Similarly, Fig. 6 demonstrates how the steady-state points p_L and p_A vary with the number of MTDs n. It can be seen from both Fig. 5 and Fig. 6 that as the aggregate input rate $\hat{\lambda}$ or the number of MTDs n increases, the network may first move from the monostable region to the bistable region, and eventually stays at the monostable region, as Fig. 4 illustrates.

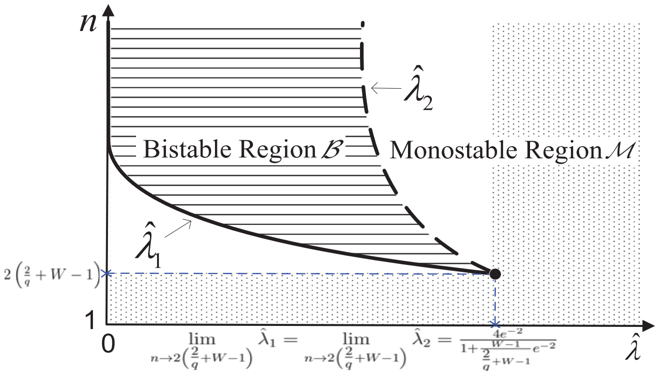

n
$\hat{\lambda}_2$
Bistable Region $\mathcal{B}$
Monostable Region $\mathcal{M}$
$\hat{\lambda}_1$
$2\left(\frac{2}{q}+W-1\right)$
1
0
$\lim_{n\to 2\left(\frac{2}{q}+W-1\right)} \hat{\lambda}_1 = \lim_{n\to 2\left(\frac{2}{q}+W-1\right)} \hat{\lambda}_2 = \frac{4e^{-2}}{1+\frac{W-1}{\frac{2}{q}+W-1}e^{-2}}$
$\hat{\lambda}$

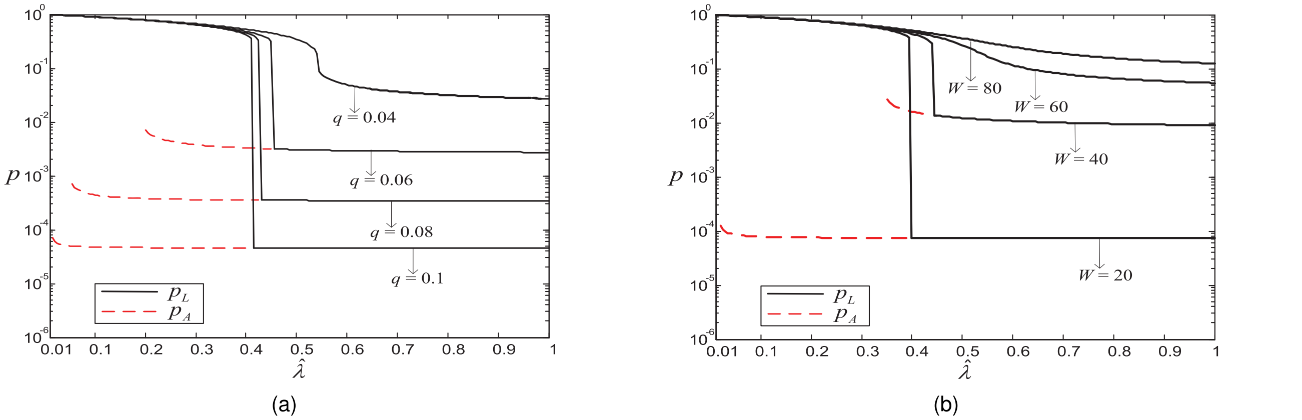
Fig. 4. Bistable region $\mathcal{B}$ and monostable region $\mathcal{M}$.

Fig. 5 and Fig. 6 also corroborate Corollary 1 that both steady-state points p_L and p_A decrease with the ACB factor q, the aggregate input rate $\hat{\lambda}$ and the number of MTDs n, and increase with the UB window size W. A closer look at Fig. 5 shows that with a high aggregate input rate $\hat{\lambda}$, e.g., $\hat{\lambda} > 0.8$, the steady-state points p_L and p_A become insensitive to the increment of $\hat{\lambda}$. In contrast, as shown in Fig. 6, when the number of MTDs n is large, both steady-state points rapidly decrease as n increases.

B. Simulation Results

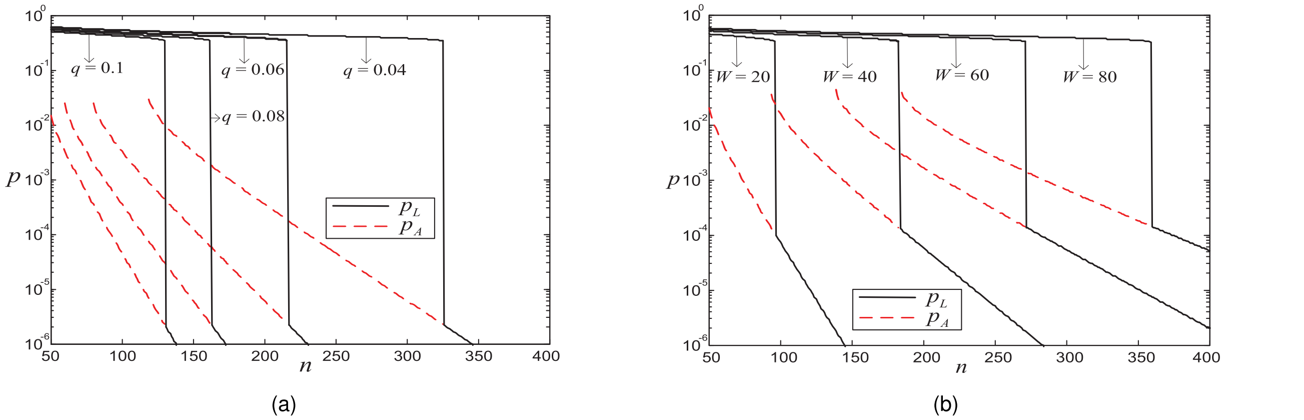
The above analysis is verified by the simulation results presented in Fig. 7 and Fig. 8. The simulation setting is the same as the system model described in Section II, and we omit the details here. Each simulation is carried out for 10^8 time slots. In simulations, we count the total number of transmitted access requests from all MTDs and the total number of successful access requests. The steady-state probability of successful transmission of access requests p is then obtained by calculating the ratio of the number of successful access requests to the total number of transmitted access requests.

Specifically, it has been demonstrated that in the monostable region $\mathcal{M}$, the network has only one steady-state point p_L; in the bistable region $\mathcal{B}$, the network has two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A. Both p_L and p_A are non-zero roots of the fixed-point equation (6) of the steady-state probability of successful transmission of access requests p. As Fig. 7 illustrates, with the aggregate input rate $\hat{\lambda} = 0.3$ and the number of MTDs $n = 100$, the network operates at the monostable region $\mathcal{M}$ when the ACB factor $q \in (0, 0.053)$ for $W = 1$ and $W \in (38, +\infty)$ for $q = 1$. As q increases or W decreases, the network will move to the bistable region $\mathcal{B}$, and may drop from the desired steady-state point p_L to the undesired steady-state point p_A. Moreover, when the aggregate input rate $\hat{\lambda}$ is sufficiently high, e.g., $\hat{\lambda} \geq 4e^{-2} \approx 0.54$, Fig. 4 has shown that the network will stay in the monostable region regardless of the number of MTDs n. As we can see from Fig. 8, with $\hat{\lambda} = 0.6$, the network always operates at the desired



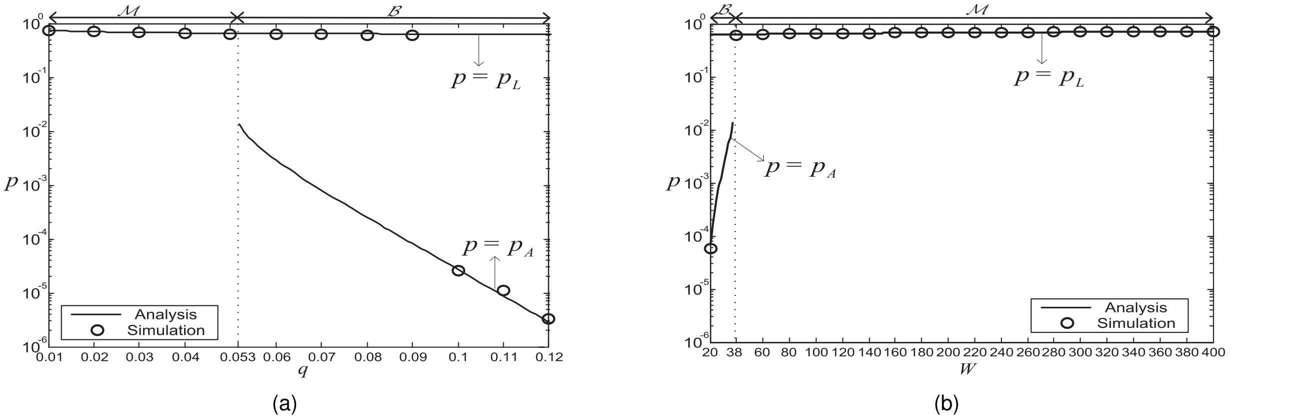
(a) (b)

Fig. 5. Steady-state points p_L and p_A versus the aggregate input rate $\hat{\lambda}$. $M = 1$. $n = 100$. (a) $W = 1, q \in \{0.04, 0.06, 0.08, 0.1\}$. (b) $q = 1$, $W \in \{20, 40, 60, 80\}$.



(a) (b)

Fig. 6. Steady-state points p_L and p_A versus the number of MTDs n. $M = 1$. $\hat{\lambda} = 0.4$. (a) $W = 1, q \in \{0.04, 0.06, 0.08, 0.1\}$. (b) $q = 1$, $W \in \{20, 40, 60, 80\}$.



(a) (b)

Fig. 7. Steady-state probability of successful transmission of access requests p versus the ACB factor q and the UB window size W. $M = 1$. $\hat{\lambda} = 0.3$. $n = 100$. (a) $W = 1$. (b) $q = 1$.

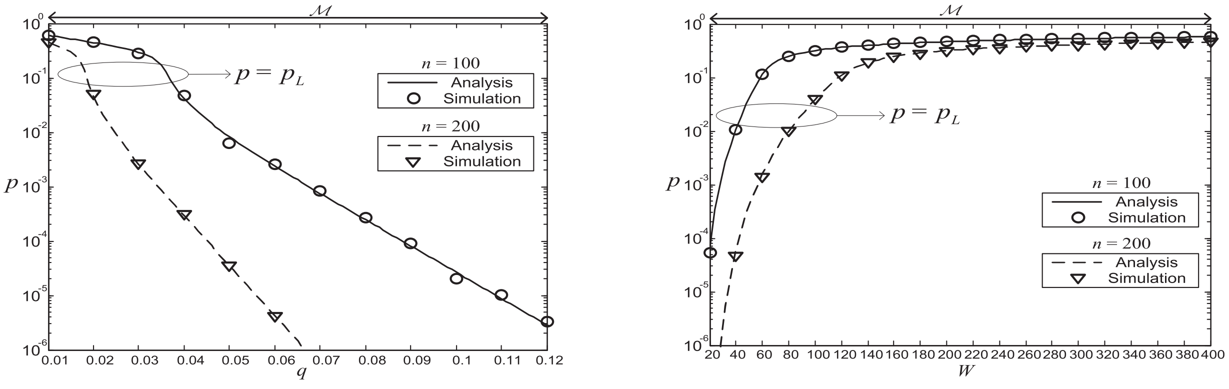
steady-state point p_L. Simulation results presented in Fig. 7 and Fig. 8 well agree with the analysis.

IV. Network Throughput Analysis

It has been shown in Section II-C and Section III that the network throughput $\hat{\lambda}_{\text{out}}$ is crucially determined by the steady-state probability of successful transmission of access requests p, which is a function of the number of MTDs n, the aggregate input rate of MTDs $\hat{\lambda}$, the ACB factor q and the UB window size W. Typically, n and $\hat{\lambda}$ are system input parameters. Therefore, in this section, we focus on optimizing the network throughput by tuning backoff parameters including q and W for given n and $\hat{\lambda}$.

A. Maximum Network Throughput

Define the maximum network throughput as $\hat{\lambda}_{\max} = \max_{(q,W)} \hat{\lambda}_{\text{out}}$. The following theorem presents the maximum network throughput $\hat{\lambda}_{\max}$ and the optimal setting of q^* and W^*.



(a) (b)

Fig. 8. Steady-state probability of successful transmission of access requests p versus the ACB factor q and the UB window size W. $M = 1$. $\hat{\lambda} = 0.6$. $n \in \{100, 200\}$. (a) $W = 1$. (b) $q = 1$.

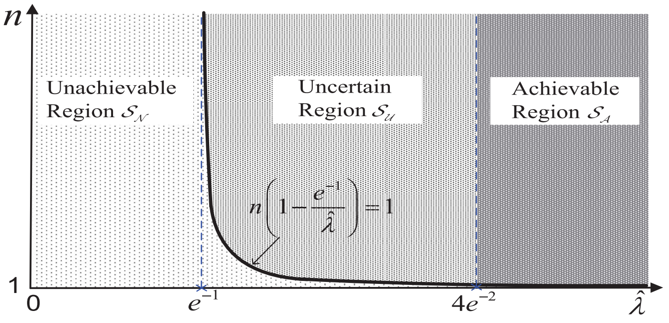

n
Unachievable Region $\mathcal{S_N}$
Uncertain Region $\mathcal{S_U}$
Achievable Region $\mathcal{S_A}$
$n\left(1-\frac{e^{-1}}{\hat{\lambda}}\right)=1$
1
0
e^{-1}
$4e^{-2}$
$\hat{\lambda}$

Fig. 9. Unachievable region $\mathcal{S_N}$, achievable region $\mathcal{S_A}$ and uncertain region $\mathcal{S_U}$.

Theorem 2: The maximum network throughput $\hat{\lambda}_{\max} = e^{-1}$, which is achieved if and only if the network operates at the desired steady-state point p_L, and (q^*, W^*) together satisfy

$$\frac{1}{q^*} + \frac{1-e^{-1}}{2}\left(W^* - 1\right) = n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right). \tag{9}$$

Proof: See Appendix C. ■

We can see from Theorem 2 that when $n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right) < 1$, the maximum network throughput $\hat{\lambda}_{\max}$ cannot be achieved, since (9) does not hold for any $q \in (0, 1]$ and $W \geq 1$. On the other hand, when the aggregate input rate $\hat{\lambda} \geq 4e^{-2} \geq \frac{4e^{-2}}{1+\frac{W-1}{\frac{2}{q}+W-1}e^{-2}}$, Fig. 4 shows that the network is guaranteed to operate at the desired steady-state point p_L. In this case, $\hat{\lambda}_{\max}$ can always be achieved if $n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right) > 1$, and q and W are properly tuned based on (9). Accordingly, we can define the following regions of $(n, \hat{\lambda})$, which are illustrated in Fig. 9:

- **Unachievable region** $\mathcal{S_N} = \left\{(n, \hat{\lambda}) | n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right) < 1\right\}$, in which $\hat{\lambda}_{\max}$ is unachievable regardless of what values of q and W are chosen. Intuitively, to achieve the maximum network throughput $\hat{\lambda}_{\max} = e^{-1}$, the aggregate input rate $\hat{\lambda}$ needs to be sufficiently large, i.e., $\hat{\lambda} > \frac{e^{-1}}{1-\frac{1}{n}} \approx e^{-1}$ for large n. As we can see from Fig. 10a, with $\hat{\lambda} = 0.3$, the network throughput $\hat{\lambda}_{\text{out}}$ is always below $\hat{\lambda}_{\max}$ as $(n, \hat{\lambda}) \in \mathcal{S_N}$.
- **Achievable region** $\mathcal{S_A} = \left\{(n, \hat{\lambda}) | n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right) \geq 1, \hat{\lambda} \geq 4e^{-2}\right\}$, in which $\hat{\lambda}_{\max}$ can be achieved when q and W are tuned according to (9). As we can see from Fig.10b, with $(n, \hat{\lambda}) \in \mathcal{S_A}$, since the network is guaranteed to operate at the desired steady-state point p_L, the network throughput is maximized at $\hat{\lambda}_{\max}$ as long as (9) holds. In this case, the maximum network throughput $\hat{\lambda}_{\max}$ can be achieved by either tuning the ACB factor q or the UB window size W. For instance, if W is fixed to 1, then the optimal ACB factor $q^*_{W=1}$ can be obtained from (9) as

$$q^*_{W=1} = \frac{\hat{\lambda}}{n(\hat{\lambda} - e^{-1})}. \tag{10}$$

  Similarly, if q is fixed to 1, then the optimal UB window size $W^*_{q=1}$ can be obtained from (9) as

$$W^*_{q=1} = \frac{2\left(n - \frac{ne^{-1}}{\hat{\lambda}} - 1\right)}{1 - e^{-1}} + 1. \tag{11}$$

- **Uncertain region** $\mathcal{S_U} = \left\{(n, \hat{\lambda}) | n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right) \geq 1, \hat{\lambda} < 4e^{-2}\right\}$, in which the network may operate at the desired steady-state point p_L or the undesired steady-state point p_A, and $\hat{\lambda}_{\max}$ is achievable only if the network operates at p_L. As we can see from Fig. 10c, the network throughput $\hat{\lambda}_{\text{out}}$ may drastically degrade if the network drops to the undesired steady-state point p_A, in which case the maximum network throughput cannot be achieved.

B. Simulation Results

The above analysis is verified by the simulation results presented in Fig. 11. In simulations, we count the total number of successful access requests in each simulation run, i.e., 10^8 time slots. The network throughput is then obtained by calculating the ratio of the number of successful access requests to the number of time slots 10^8.

Specifically, the expression of network throughput $\hat{\lambda}_{\text{out}}$ has been given in (4), which is determined by the ACB factor q,

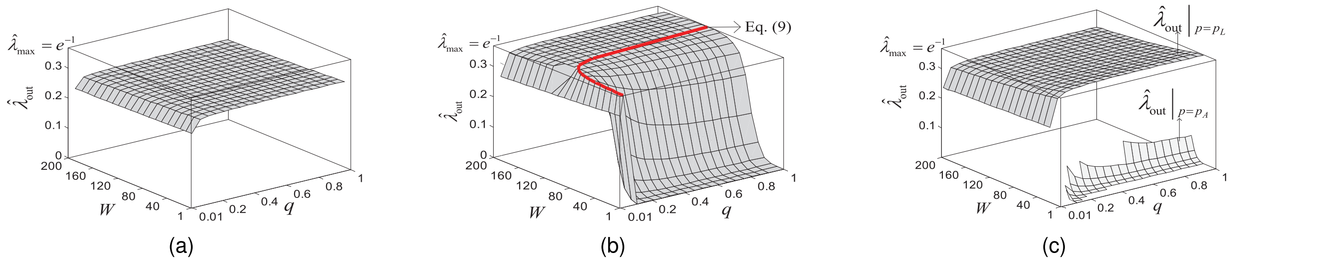


Fig. 10. Network throughput $\hat{\lambda}_{\text{out}}$ versus the ACB factor q and the UB window size W. $M = 1$. $n = 100$. (a) $\hat{\lambda} = 0.3$, $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{N}}$. (b) $\hat{\lambda} = 0.6$, $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{A}}$. (c) $\hat{\lambda} = 0.4$, $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{U}}$.

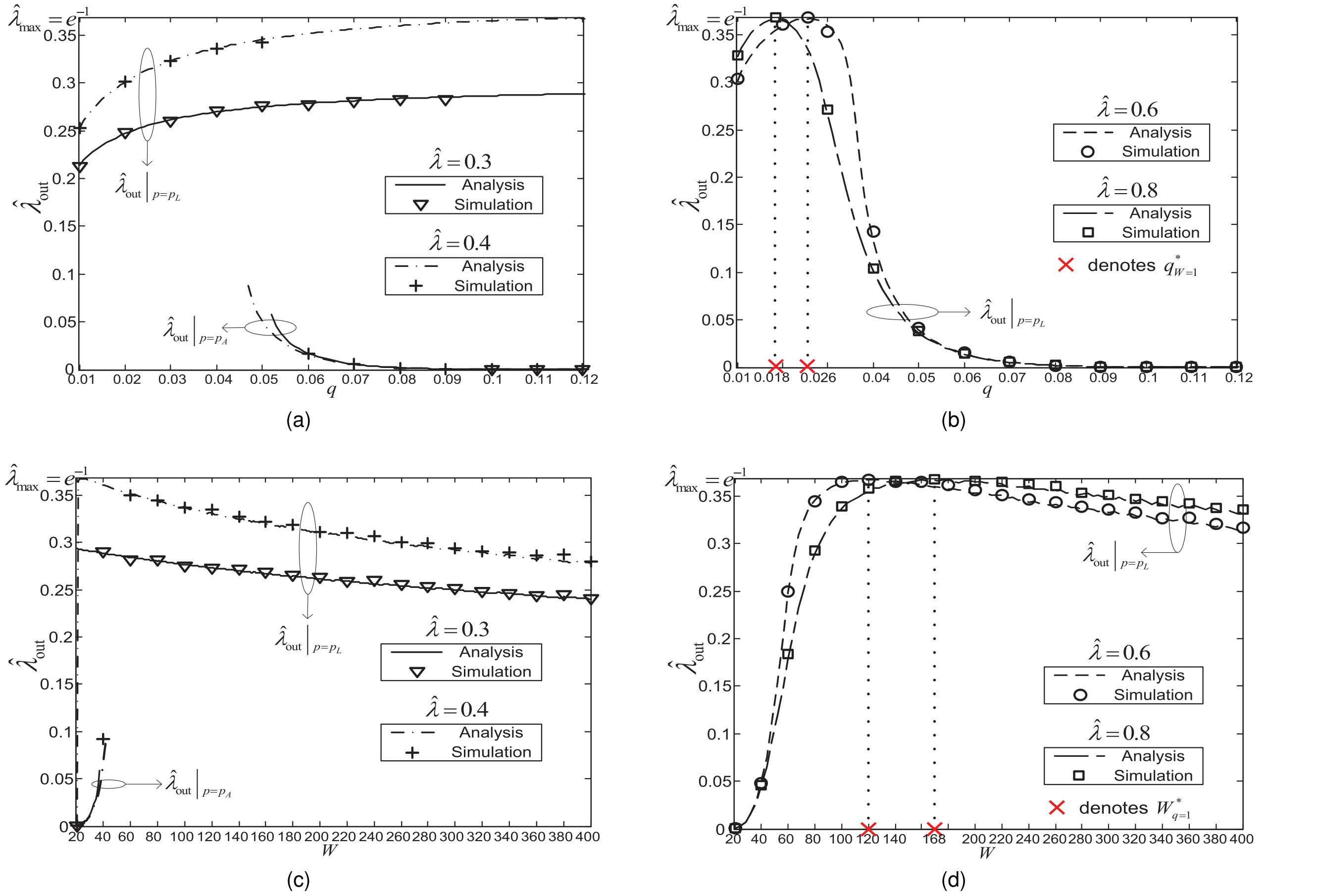


Fig. 11. Network throughput $\hat{\lambda}_{\text{out}}$ versus the ACB factor q and the UB window size W. $M = 1$. $n = 100$. $\hat{\lambda} \in \{0.3, 0.4, 0.6, 0.8\}$. (a)-(b) $W = 1$. (c)-(d) $q = 1$.

the UB window size W, the number of MTDs n and the aggregate input rate $\hat{\lambda}$. Fig. 11a-b illustrate how the network throughput $\hat{\lambda}_{\text{out}}$ varies with the ACB factor q with the UB window size $W = 1$ and the number of MTDs $n = 100$. In Fig. 11a, when the aggregate input rate $\hat{\lambda}$ is 0.3, we have $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{N}}$, in which the maximum network throughput $\hat{\lambda}_{\max}$ cannot be achieved regardless of what value of q is chosen. On the other hand, with $\hat{\lambda} = 0.4$, we have $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{U}}$, in which $\hat{\lambda}_{\max}$ again cannot be achieved, because the network shifts to the undesired steady-state point p_A as the ACB factor q increases. In Fig. 11b, as the aggregate input rate $\hat{\lambda}$ increases to 0.6 or 0.8, we have $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{A}}$, where $\hat{\lambda}_{\max}$ can be achieved when q is tuned according to (10), i.e., $q = q^*_{W=1}$. Similar observations can be obtained from Fig. 11c-d, where the maximum network throughput $\hat{\lambda}_{\max}$ is achieved when $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{A}}$, i.e., $\hat{\lambda} = 0.6$ or 0.8 with $n = 100$, and the UB window size $W = W^*_{q=1}$, which is given in (11).

V. Extension to Multi-Preamble $M > 1$

Note that the analysis in previous sections is based on the assumption that all n MTDs share one preamble, i.e., $M = 1$. In this section, the analysis will be extended to the multi-preamble scenario in which n MTDs choose $M > 1$ preambles.

A. Multi-Group Model

By virtue of orthogonality among preambles, only the MTDs who share the same preamble contend with each other.

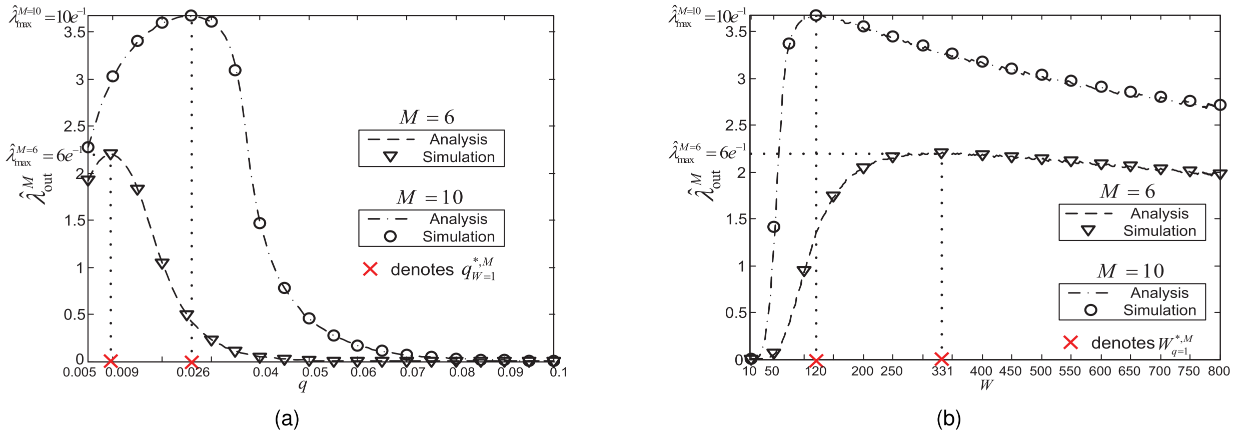


Fig. 12. Network throughput $\hat{\lambda}_{\text{out}}^{M}$ versus the ACB factor q and the UB window size W. $n=1000$. $\lambda=0.006$. $M\in\{6,10\}$. (a) $W=1$. (b) $q=1$.

Accordingly, the MTDs in the network can be divided into M groups according to the preamble each MTD chooses. By doing so, we can extend the previous analytical model to a multi-group one, with the group parameters defined as follows:

- $n^{(i)}$ denotes the number of MTDs in Group i, $i=1,2,\ldots,M$, and $\sum_{i=1}^{M} n^{(i)}=n$.
- $\hat{\lambda}^{(i)}$ denotes the aggregate input rate of MTDs in Group i, and $\hat{\lambda}^{(i)}=n^{(i)}\lambda$, where λ is the input rate of each MTD.
- $q^{(i)}$ denotes the ACB factor of each MTD in Group i.
- $W^{(i)}$ denotes the UB window size of each MTD in Group i.

For each group, denote $p^{(i)}$ as the steady-state probability of successful transmission of access requests for MTDs in Group $i, i=1,2,\ldots,M$. By replacing $n,\hat{\lambda},q,W$ in (6) with $n^{(i)}$, $\hat{\lambda}^{(i)}$, $q^{(i)}$, $W^{(i)}$, the steady-state points of Group i, i.e., $p_L^{(i)}$ and $p_A^{(i)}$, can be obtained. Similarly, denote $\hat{\lambda}_{\text{out}}^{(i)}$ as the aggregate throughput of MTDs in Group $i, i=1,2,\ldots,M$, which can be calculated based on (4) by replacing $n,\hat{\lambda},q,W$ with $n^{(i)}$, $\hat{\lambda}^{(i)}$, $q^{(i)}$, $W^{(i)}$. According to Theorem 2, the maximum group throughput $\hat{\lambda}_{\max}^{(i)}=e^{-1}$, which is achieved when the group steady-state point is at $p_L^{(i)}$, and $q^{(i)}$ and $W^{(i)}$ are chosen according to (9) for given $n^{(i)}$ and $\hat{\lambda}^{(i)}$. The network throughput with M preambles is then given by $\hat{\lambda}_{\text{out}}^{M}=\sum_{i=1}^{M}\hat{\lambda}_{\text{out}}^{(i)}$, which is maximized at

$$\hat{\lambda}_{\max}^{M}=Me^{-1}, \tag{12}$$

when all the groups achieve the maximum group throughput $\hat{\lambda}_{\max}^{(i)}=e^{-1}, i=1,2,\ldots,M$.

Note that according to the standard, each MTD independently and randomly selects a preamble in each access attempt [29]. Therefore, the group size $n^{(i)}, i=1,2,\ldots,M$, may change over time. Nevertheless, when the total number of MTDs n is large, $n^{(i)}$ can be approximated by $n^{(i)}\approx\frac{n}{M}$. In this case, the network throughput can be obtained by replacing n and $\hat{\lambda}$ with $n^{(i)}\approx\frac{n}{M}$ and $\hat{\lambda}^{(i)}\approx\frac{n\lambda}{M}$ in (4), respectively, as

$$\hat{\lambda}_{\text{out}}^{M}=\frac{n}{M}\sum_{i=1}^{M}\frac{1}{\frac{1}{q^{(i)}p^{(i)}}+\frac{(1-p^{(i)})(W^{(i)}-1)}{2p^{(i)}}+\frac{1}{\lambda}}. \tag{13}$$

Moreover, the optimal setting $\left(q^{*,M},W^{*,M}\right)$ for achieving the maximum network throughput $\hat{\lambda}_{\max}^{M}=Me^{-1}$ can be obtained from (9) by replacing n and $\hat{\lambda}$ with $n^{(i)}\approx\frac{n}{M}$ and $\hat{\lambda}^{(i)}\approx\frac{n\lambda}{M}$, respectively, as

$$\frac{1}{q^{*,M}}+\frac{1-e^{-1}}{2}\left(W^{*,M}-1\right)=\frac{n}{M}-\frac{e^{-1}}{\lambda}. \tag{14}$$

Specifically, with $W=1$, the optimal ACB factor $q_{W=1}^{*,M}$ with M preambles can be written as

$$q_{W=1}^{*,M}=\frac{\lambda}{\frac{n\lambda}{M}-e^{-1}}. \tag{15}$$

With $q=1$, the optimal UB window size $W_{q=1}^{*,M}$ with M preambles can be written as

$$W_{q=1}^{*,M}=\frac{2\left(\frac{n}{M}-\frac{e^{-1}}{\lambda}-1\right)}{1-e^{-1}}+1. \tag{16}$$

Note that (15) and (16) reduce to (10) and (11), respectively, when $M=1$. It can be clearly seen from (15) and (16) that $q_{W=1}^{*,M}$ and $W_{q=1}^{*,M}$ are monotonic increasing and decreasing functions of the number of preambles M, respectively, indicating that the ACB factor q and the UB window size W should be adaptively increased or reduced as more preambles are adopted.

B. Simulation Results

In this subsection, we present the simulation results in the multi-preamble scenario to verify the above analysis. In simulations, each MTD independently and randomly selects one out of M orthogonal preambles in each access attempt. We count the total number of successful access requests in each simulation run, i.e., 10^8 time slots, and then obtain the network throughput by calculating the ratio of the number of successful access requests to the number of time slots 10^8.

Specifically, the expression of the network throughput $\hat{\lambda}_{\text{out}}^{M}$ with M preambles has been given in (13). Fig. 12 illustrates how the network throughput $\hat{\lambda}_{\text{out}}^{M}$ varies with the ACB factor q and the UB window size W with $M=6$ and 10. A perfect match between simulation results and the analysis can be

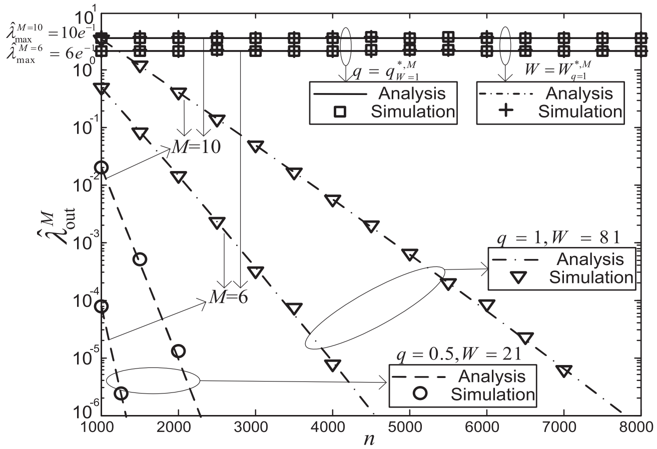


Fig. 13. Network throughput $\hat{\lambda}_{\text{out}}^{M}$ versus the number of MTDs n. $\lambda = 0.006$. $M \in \{6, 10\}$.

observed, which verifies that $n^{(i)} \approx \frac{n}{M}, i = 1, 2, \ldots, M$, can serve as a good approximation when the number of MTDs n is large. Moreover, we can see that the maximum network throughput $\hat{\lambda}_{\max}^{M}$ linearly increases with the number of preambles M, and can be achieved by either tuning the ACB factor q according to $q_{W=1}^{*,M}$ in (15), or the UB window size W according to $W_{q=1}^{*,M}$ in (16).

Note that in the current standard setting, the ACB factor q and the UB window size W are preselected from a certain range [29]. To see the performance loss without adaptive tuning of q and W, Fig. 13 illustrates the network throughput performance with two representative settings of parameters: $\{q = 0.5, W = 21\}$ and $\{q = 1, W = 81\}$ [53].[6] It can be clearly observed that in both cases, the network throughput $\hat{\lambda}_{\text{out}}^{M}$ quickly deteriorates as the number of MTDs n increases, and becomes much lower than the maximum network throughput $\hat{\lambda}_{\max}^{M}$ when n is large. In sharp contrast, if the ACB factor q or the UB window size W is optimally selected, i.e., $q = q_{W=1}^{*,M}$ or $W = W_{q=1}^{*,M}$, the maximum network throughput $\hat{\lambda}_{\max}^{M}$ can always be achieved, which does not vary with the number of MTDs n. It corroborates that adaptive tuning of backoff parameters is indispensable especially for massive access scenarios.

VI. Discussions

In this section, we will further discuss how the optimal network throughput performance is affected by practical network conditions such as the outdated information on the number of MTDs n and the input rate of each MTD λ, and the bursty traffic.

A. Effect of Outdated Information on n and λ

So far we have shown that to optimize the network throughput performance, the backoff parameters of MTDs should be adaptively tuned according to the number of MTDs n and the input rate of each MTD λ. In practice, the BS can count the total number of MTDs,[7] collect the traffic input rate information from the feedback of MTDs, calculate the optimal backoff parameters based on (14), and broadcast the optimal configuration via the system information block. Each MTD then updates its backoff parameters accordingly.

Due to the constant change of network states, sometimes the BS may not be able to update the information on the total number of MTDs n and the input rate of each MTD λ in time. To see how much the network throughput may degrade with outdated information on n and λ, we define $\gamma_n = \frac{\tilde{n}-n}{\tilde{n}}$ and $\gamma_\lambda = \frac{\tilde{\lambda}-\lambda}{\tilde{\lambda}}$ as the relative error on n and λ, respectively, where $\tilde{n}$ and $\tilde{\lambda}$ denote the outdated information on the number of MTDs and the input rate of each MTD at the BS, respectively. Suppose that the optimal backoff parameters $q_{W=1}^{*,M}$ and $W_{q=1}^{*,M}$ are calculated based on $\tilde{n}$ and $\tilde{\lambda}$ according to (15) and (16), respectively. Fig. 14 presents the simulation results of the corresponding network throughput under various values of relative errors γ_n and γ_λ.

It can be clearly seen from Fig. 14a that due to outdated information of the number of MTDs, the network throughput would deflect from the maximum value, and the degradation becomes more significant as the relative error γ_n increases. Recall that it has been shown in Fig. 6 that both steady-state points are quite sensitive to the change of the number of MTDs n. Therefore, the network throughput may quickly drop when n is not updated in time. On the other hand, Fig. 5 shows that when the aggregate input rate $\hat{\lambda}$ is large, the steady-state points become insensitive to $\hat{\lambda}$. As a result, the network throughput can stay at the maximum value even with the relative error γ_λ as large as ±50%, as Fig. 14b illustrates. We can also see from Fig. 14 that although the maximum network throughput can be achieved by either tuning the ACB factor q or the UB window size W, the throughput performance is more sensitive to the relative errors with the optimal tuning of q than W. It suggests that the optimal tuning of W is more robust against the variation of network size and traffic input rate.

B. Effect of Bursty Arrivals of Data Packets

Note that the preceding analysis is based on the assumption that data packet arrivals of each MTD independently follow a Bernoulli process. In practical M2M communication scenarios, packet arrival processes could be bursty in some cases [54]. Hence, in this subsection, we will investigate the effect of bursty arrivals on the optimal network throughput performance.

There are two kinds of burstiness: temporal burstiness and spatial burstiness. As Fig. 15a illustrates, temporal burstiness means that a continuous stream of packets is generated in a short time period for a given MTD. On the other hand, spatial burstiness means that multiple MTDs generate packets

[6] Note that in the standard [29], the values of UB window size W_s are given in unit of milliseconds. If the *PRACH configuration index* is 14, for instance, the time slot length $\tau = 1$ msec [48], and then $W = 21$ and $W = 81$ are corresponding to $W_s = 20$ msec and $W_s = 80$ msec, respectively.

[7] Note that here the total number of MTDs should be distinguished from the number of *active* MTDs, which is usually assumed in the literature [11]–[17], [35]–[42]. Specifically, the BS can easily keep a record of registered MTDs without knowing if they are active or not at each time slot. In contrast, tracking and estimating the time-varying number of active MTDs can be highly challenging and demanding.

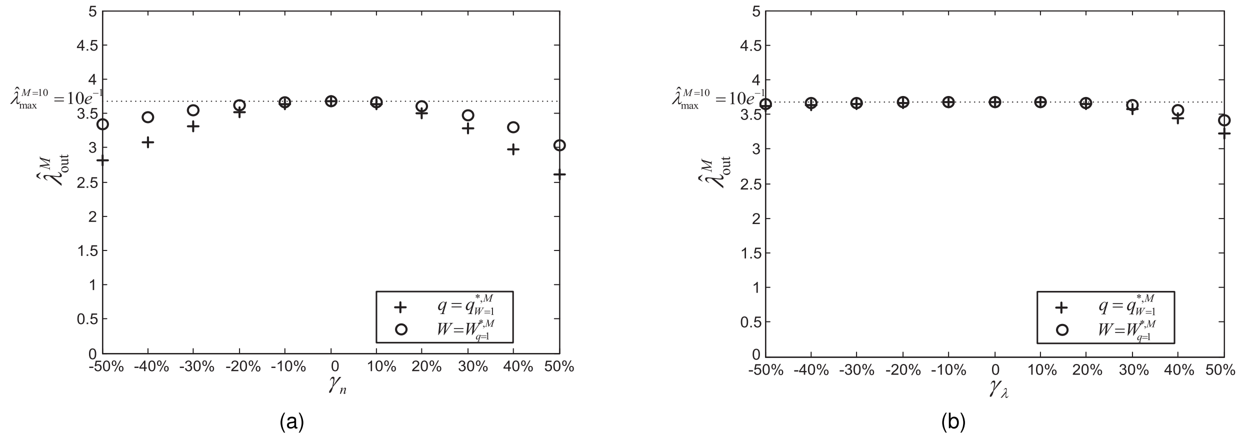


Fig. 14. Simulated network throughput $\hat{\lambda}_{\text{out}}^{M}$ versus the relative errors γ_n and γ_λ. $M = 10$. $\tilde{n} = 1000$. $\tilde{\lambda} = 0.01$. (a) $\gamma_\lambda = 0$. (b) $\gamma_n = 0$.

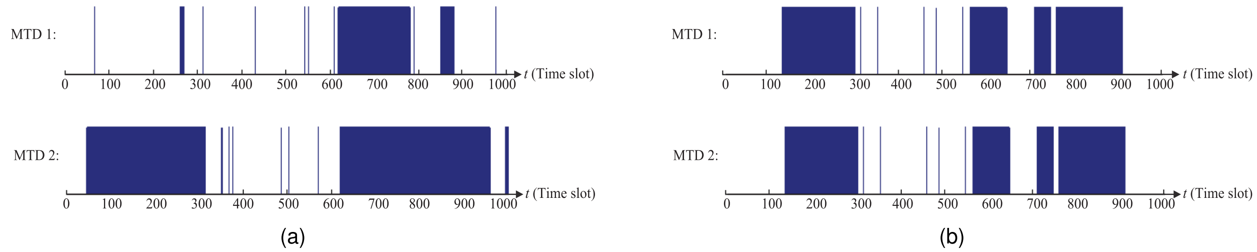


Fig. 15. Illustration of data packet arrival processes at two tagged MTDs. $\sigma_0 = 0.005$. $\lambda_0 = 0.006$. $\sigma_1 = 0.005$. $\lambda_1 = 1$. (a) Temporal burstiness. (b) Spatial burstiness.

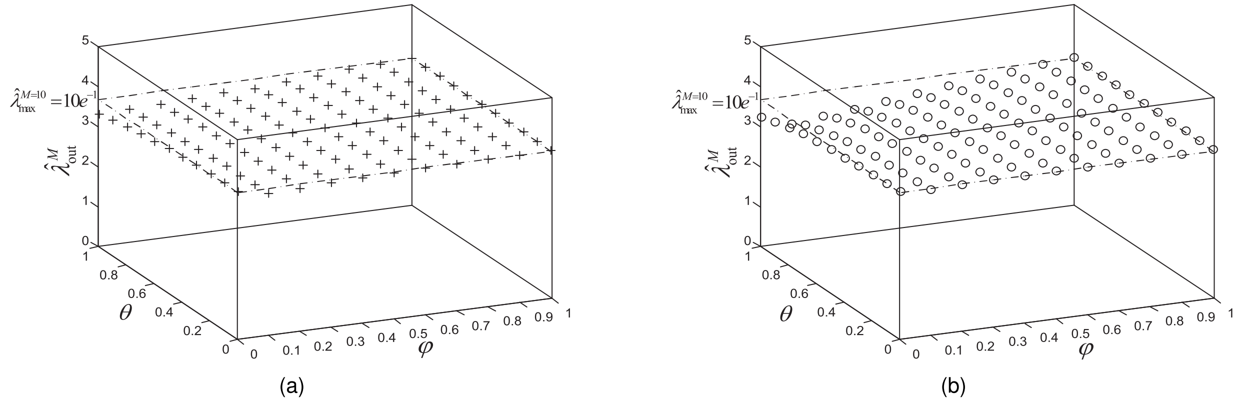


Fig. 16. Simulated network throughput $\hat{\lambda}_{\text{out}}^{M}$ versus the synchronization ratio θ and the frequency of bursty arrivals φ. $M = 10$. $n = 1000$. $\sigma_0 = 0.005$. $\lambda_0 = 0.006$. $\lambda_1 = 1$. (a) $q = q_{W=1}^{*,M}$. (b) $W = W_{q=1}^{*,M}$.

in a synchronous manner, as Fig. 15b shows. To capture the temporal burstiness, the data packet arrival process of a single MTD is modeled as a two-state Markov modulated Bernoulli process (MMBP) $\{(X_t, Y_t), t = 1, 2, \ldots.\}$, where $X_t \in \{0, 1\}$ and $Y_t \in \{0, 1\}$ denote the number of arrivals and the phase of MMBP at time slot t, respectively. We refer to $Y_t = 1$ as the bursty phase and $Y_t = 0$ as the regular phase. Assume that the bursty phase and regular phase have packet arrival rates of $\lambda_1 = 1$ and $\lambda_0 \ll \lambda_1$, respectively, and last for a geometrically distributed amount of time slots with parameter σ_1 and σ_0, respectively. To capture the spatial burstiness, assume that n_s out of n MTDs have identical data packet arrival processes, i.e., they are all driven by a common MMBP $\{(X_t, Y_t), t = 1, 2, \ldots.\}$, and the remaining MTDs have independent data packet arrival processes. Define the frequency of bursty arrivals as $\varphi = \frac{\frac{1}{\sigma_1}}{\frac{1}{\sigma_1}+\frac{1}{\sigma_0}}$, and the synchronization ratio as $\theta = \frac{n_s}{n}$. Apparently, a large φ and θ indicate high temporal burstiness and spatial burstiness, respectively.

Fig. 16 presents the simulation results of network throughput with bursty arrivals under various values of the synchronization ratio θ and the frequency of bursty arrivals φ. We set the backoff parameters q and W as $q_{W=1}^{*,M}$ and $W_{q=1}^{*,M}$,

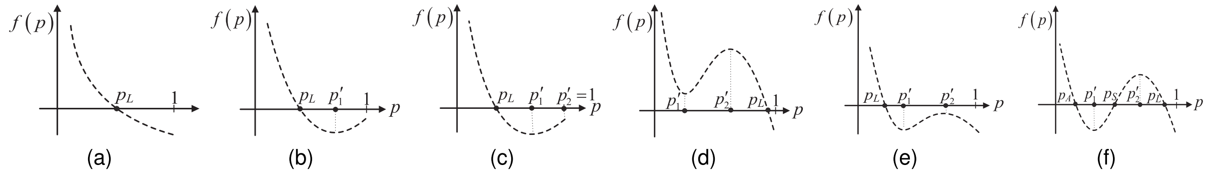


Fig. 17. $f(p)$ has one root in (a), (b), (c), (d), (e) and three non-zero roots in (f).

respectively, which are calculated based on (15) and (16) with $\lambda = \varphi\lambda_1 + (1-\varphi)\lambda_0$. We can observe from Fig. 16 that the network throughput is close to the maximum value within a wide range of θ and φ. The simulation results corroborate that by optimally tuning the ACB factor q or the UB window size W, the bursty input traffic can be sufficiently randomized. Therefore, the maximum network throughput can be achieved even when the input traffic has a high degree of temporal burstiness and spatial burstiness.

VII. Conclusion

In this paper, a new analytical framework is proposed to optimize the random access performance of M2M communications in LTE networks. Starting from the single-preamble case, the analysis shows that the network can either have one or two steady-state points, depending on whether it operates at the monostable or bistable region. The network throughput is derived as an explicit function of the network steady-state points, and optimized by properly adjusting the backoff parameters of MTDs including the ACB factor q and the UB window size W according to the number of MTDs and the traffic input rate of each MTD. The analysis is further extended to the multi-preamble scenario, where explicit expressions of the maximum network throughput and the corresponding optimal backoff parameters are both obtained.

The analysis sheds important light on practical network design for supporting massive access of M2M communications in LTE networks. Specifically, it shows that a preselection of the ACB factor q and the UB window size W always leads to severe degradation of network throughput as more MTDs attempt to access the BS. Only by optimally tuning q or W can the network throughput remain at the highest level regardless of how many MTDs in the network. The proposed optimal tuning is solely based on statistical information including the number of MTDs and the traffic input rate of each MTD, with which the throughput performance is found to be robust against feedback errors of the traffic input rate and burstiness of data arrivals.

Note that although this paper focuses on the network throughput analysis, the access delay performance of each MTD, including the first and second moments of access delay, can also be characterized based on the proposed double-queue model. It is of practical importance to further perform delay optimization under the proposed analytical framework. Moreover, in this paper, we focus on the homogeneous case where each MTD has an identical traffic input rate. The analysis should be extended to heterogeneous scenarios to incorporate distinct traffic characteristics and quality-of-service requirements of MTDs, where fairness would become an important issue. Finally, a key assumption in this paper is that each access request stays in the queue until it is successfully transmitted. In practice, however, an access request could be dropped if it reaches the maximum number of retransmissions [13]. How such a retry limit affects the maximum network throughput and the optimal backoff parameter tuning is another interesting issue that deserves much attention in the future study.

Appendix A
Proof of Theorem 1

Proof: Let $f(p) = -\ln p - \frac{a}{b+p}$, where

$$a = \frac{2n}{\frac{2n}{\hat{\lambda}} - W + 1}, \quad \text{and } b = \frac{\frac{2}{q} + W - 1}{\frac{2n}{\hat{\lambda}} - W + 1}. \tag{17}$$

It can be seen from (6) that $f(p) = 0$ has the same non-zero roots as the fixed-point equation (6). Hence, we will focus on $f(p) = 0$ in the following. The derivative of $f(p)$ can be obtained as $f'(p) = \frac{g(p)}{p(b+p)^2}$, where

$$g(p) = -\left(p + b - \frac{a}{2}\right)^2 + \frac{a^2}{4} - ab. \tag{18}$$

Lemma 1 shows that the number of non-zero roots of $f(p) = 0$ for $p \in (0, 1]$ is crucially related with the number of non-zero roots of $g(p) = 0$ for $p \in (0, 1]$.

Lemma 1: $f(p) = 0$ has three non-zero roots $0 < p_A \leq p_S \leq p_L \leq 1$ if and only if $g(p) = 0$ has two non-zero roots $0 < p'_1 < p'_2 < 1$ with $f(p'_1) \leq 0$ and $f(p'_2) \geq 0$; Otherwise, $f(p) = 0$ has only one non-zero root $0 < p_L \leq 1$.

Proof: Since $\lim_{p\to 0} f(p) = +\infty$ and $f(1) = -\frac{a}{b+1} = -\frac{n}{\frac{1}{q}+\frac{n}{\hat{\lambda}}} < 0$, $f(p) = 0$ has at least one non-zero root for $p \in (0, 1]$. To further determine the number of non-zero roots of $f(p) = 0$ for $p \in (0, 1]$, let us consider the following scenarios.

1) If $g(p) = 0$ has no non-zero roots for $p \in (0, 1]$, then $g(p) < 0$ for $p \in (0, 1]$. As a result, $f'(p) < 0$ for $p \in (0, 1]$, indicating that $f(p)$ monotonically decreases for $p \in (0, 1]$, as shown in Fig. 17a. We can then conclude that in this case, $f(p) = 0$ has only one non-zero root $0 < p_L \leq 1$.
2) If $g(p) = 0$ has one non-zero root $0 < p'_1 \leq 1$, then $g(p) < 0$ for $p \in (0, p'_1)$ and $g(p) > 0$ for $p \in (p'_1, 1]$. As a result, $f'(p) < 0$ for $p \in (0, p'_1)$ and $f'(p) > 0$ for $p \in (p'_1, 1]$, indicating that $f(p)$ monotonically decreases for $p \in (0, p'_1)$, and then increases for $p \in (p'_1, 1]$, as shown in Fig.17b. Since $f(1) < 0$, we can conclude that in this case, $f(p) = 0$ has only one non-zero root $0 < p_L \leq 1$.

3) If $g(p) = 0$ has two non-zero roots $0 < p_1' < p_2' \leq 1$, then we have:

   a) If $p_2' = 1$, then $g(p) < 0$ for $p \in (0, p_1')$ and $g(p) > 0$ for $p \in (p_1', 1]$. As a result, $f'(p) < 0$ for $p \in (0, p_1')$ and $f'(p) > 0$ for $p \in (p_1', 1]$, indicating that $f(p)$ monotonically decreases for $p \in (0, p_1')$, and then increases for $p \in (p_1', 1]$, as shown in Fig.17c. Since $f(1) < 0$, we can conclude that in this case, $f(p) = 0$ has only one non-zero root $0 < p_L \leq 1$.

   b) If $p_2' < 1$, then $g(p) < 0$ for $p \in (0, p_1') \bigcup (p_2', 1)$ and $g(p) > 0$ for $p \in (p_1', p_2')$. As a result, $f'(p) < 0$ for $p \in (0, p_1') \bigcup (p_2', 1)$ and $f'(p) > 0$ for $p \in (p_1', p_2')$, indicating that $f(p)$ decreases for $p \in (0, p_1') \bigcup (p_2', 1)$, and then increases for $p \in (p_1', p_2')$, as shown in Fig.17d-f. We can see from Fig.17d-e that if $f(p_1') > 0$ or $f(p_2') < 0$, $f(p) = 0$ has one non-zero root $0 < p_L \leq 1$; Otherwise, $f(p) = 0$ has three non-zero roots $0 < p_A \leq p_S \leq p_L \leq 1$, as shown in Fig.17f, in which $f(p_1') \leq 0$ and $f(p_2') \geq 0$. ■

Lemma 2 further presents the necessary and sufficient condition for $g(p) = 0$ having two non-zero roots $0 < p_1' < p_2' < 1$ with $f(p_1') \leq 0$ and $f(p_2') \geq 0$.

Lemma 2: $g(p) = 0$ has two non-zero roots $0 < p_1' < p_2' < 1$ with $f(p_1') \leq 0$ and $f(p_2') \geq 0$ if and only if $n > 2(\frac{2}{q} + W - 1)$ and $\hat{\lambda}_1 \leq \hat{\lambda} \leq \hat{\lambda}_2$, where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given in (7) and (8), respectively.

Proof: According to (18), $g(p) = 0$ has two non-zero roots $0 < p_1' < p_2' < 1$ if and only if $0 < b < 1$ and $4b < a < (b+1)^2$, which can be further written as

$$\hat{\lambda} < \min\left\{\frac{2n}{W-1}, \frac{n}{W-1+\frac{1}{q}}\right\} = \frac{n}{W-1+\frac{1}{q}}, \tag{19}$$

$$2\left(\frac{1}{q} + \frac{n}{\hat{\lambda}}\right)^2 > n\left(\frac{2n}{\hat{\lambda}} - W + 1\right), \tag{20}$$

$$n > 2\left(\frac{2}{q} + W - 1\right), \tag{21}$$

according to (17).

The two non-zero roots of p_1' and p_2' of $g(p) = 0$ can be written as

$$\begin{cases} p_1' = \dfrac{n - \frac{2}{q} - W + 1 - \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}{\frac{2n}{\hat{\lambda}} - W + 1} \\ p_2' = \dfrac{n - \frac{2}{q} - W + 1 + \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}{\frac{2n}{\hat{\lambda}} - W + 1} \end{cases} \tag{22}$$

if (21) holds, i.e., $n > 2\left(\frac{2}{q} + W - 1\right)$. It can be obtained from (22) that $f(p_1') \leq 0$ and $f(p_2') \geq 0$ if and only if $\hat{\lambda}_1 \leq \hat{\lambda} \leq \hat{\lambda}_2$, where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given in (7) and (8), respectively.

Now, we prove that if (21) holds and $\hat{\lambda} \leq \hat{\lambda}_2$, then (19) and (20) hold.

(1) Let us first show that when $n > 2\left(\frac{2}{q} + W - 1\right)$, $\hat{\lambda}_2$ monotonically decreases with n. Specifically, according to (8), $\hat{\lambda}_2$ can be written as $\hat{\lambda}_2 = \frac{2}{F(n)}$, where

$$F(n) = \frac{1}{n}\left(\left(n - \frac{2}{q} - W + 1 + \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}\right) \cdot \exp\left(\frac{2n}{n + \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}\right) + W - 1\right). \tag{23}$$

It can be obtained that $\left.\frac{\mathrm{d}F(n)}{\mathrm{d}n}\right|_{n>2\left(\frac{2}{q}+W-1\right)} = \frac{(q(W-1)+2)\exp\left(\frac{2n}{n+\sqrt{n\left(n-2(W-1)-\frac{4}{q}\right)}}\right) - q(W-1)}{qn^2} > \frac{q(W-1)(e^2-1)}{qn^2} > 0$. Therefore,

$$\hat{\lambda}_2 < \lim_{n \to 2\left(\frac{2}{q}+W-1\right)} \hat{\lambda}_2 = \frac{2n}{\left(\frac{2}{q} + W - 1\right)e^2 + W - 1} < \frac{n}{W - 1 + \frac{1}{q}}. \tag{24}$$

It can be seen from (24) that if (21) holds and $\hat{\lambda} \leq \hat{\lambda}_2$, then (19) holds.

(2) We rewrite (20) as $H(\hat{\lambda}) > 0$, where $H(\hat{\lambda})$ is given by $H(\hat{\lambda}) = \frac{2n^2}{\hat{\lambda}^2} + \frac{1}{\hat{\lambda}}\left(\frac{4n}{q} - 2n^2\right) + \frac{2}{q^2} + n(W - 1)$. It can be obtained that $H(\hat{\lambda}) = 0$ has two non-zero roots of $\bar{\lambda}_1 = \frac{n\left(n - \frac{2}{q} - \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}\right)}{n(W-1) + \frac{2}{q^2}}$ and $\bar{\lambda}_2 = \frac{n\left(n - \frac{2}{q} + \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}\right)}{n(W-1) + \frac{2}{q^2}}$ with $\bar{\lambda}_1 < \bar{\lambda}_2$, if (21) holds, i.e., $n > 2\left(\frac{2}{q} + W - 1\right)$. When $\hat{\lambda} < \bar{\lambda}_1$ or $\hat{\lambda} > \bar{\lambda}_2$, $H(\hat{\lambda}) > 0$. According to (8), we can have

$$\frac{\bar{\lambda}_1}{\hat{\lambda}_2} > \left(n - \frac{2}{q} - \sqrt{n\left(n - 2(W-1) - \frac{4}{q}\right)}\right) \cdot \frac{\left(n - \frac{2}{q} + \sqrt{n\left(n - 2(W-1) - \frac{4}{q}\right)}\right)}{2\left((W-1)n + \frac{2}{q^2}\right)} = 1. \tag{25}$$

Therefore, we can conclude that for $\hat{\lambda} < \hat{\lambda}_2$, $H(\hat{\lambda}) > 0$, which indicates that (20) holds. ■

Finally, Theorem 1 can be obtained by combing Lemma 1 and Lemma 2. ■

Appendix B
Proof of Corollary 1

Proof: According to the fixed-point equation (6), we can obtain that

$$\frac{\partial p}{\partial \hat{\lambda}} = \frac{p}{g(p)}\left(\frac{\ln p}{\hat{\lambda}}\right)^2 \left(\frac{\hat{\lambda}}{n}\left(\frac{1}{q} + \frac{W-1}{2}\right) + p\left(1 - \frac{\hat{\lambda}(W-1)}{2n}\right)\right)^2, \tag{26}$$

$$\frac{\partial p}{\partial n} = \frac{p}{g(p)}\left(\frac{1}{n}\cdot\frac{\hat{\lambda}}{1-\frac{(W-1)\hat{\lambda}}{2n}}\right)^2\left(\frac{1}{q}+(1-p)\frac{W-1}{2}\right), \tag{27}$$

$$\frac{\partial p}{\partial q} = \frac{1}{ng(p)}\left(\frac{\ln p}{q}\right)^2\cdot\left(\frac{\hat{\lambda}}{n}\left(\frac{1}{q}+\frac{W-1}{2}\right) + p\left(1-\frac{\hat{\lambda}(W-1)}{2n}\right)\right)^2, \tag{28}$$

$$\frac{\partial p}{\partial W} = -\frac{(\ln p)^2(1-p)}{2ng(p)}\cdot\left(\frac{\hat{\lambda}}{n}\left(\frac{1}{q}+\frac{W-1}{2}\right) + p\left(1-\frac{\hat{\lambda}(W-1)}{2n}\right)\right)^2, \tag{29}$$

where $g(p)$ is given in (18). Let us consider the following scenarios:

1) If $g(p) = 0$ has no non-zero roots for $p \in (0, 1]$, then $g(p) < 0$ for $p \in (0, 1]$. In this case, as Fig.17a shows, (6) has one non-zero root p_L, which is a steady-state point according to the approximate trajectory analysis in [23]. We then have $g(p_L) < 0$.
2) If $g(p) = 0$ has one non-zero root $0 < p'_1 \leq 1$, then $g(p) < 0$ for $p \in (0, p'_1)$ and $g(p) > 0$ for $p \in (p'_1, 1]$. In this case, as Fig.17b shows, (6) has one non-zero root $p_L < p'_1$, which is a steady-state point according to the approximate trajectory analysis in [23]. We then have $g(p_L) < 0$.
3) If $g(p) = 0$ has two non-zero roots $0 < p'_1 < p'_2 \leq 1$, then we have:
   a) If $p'_2 = 1$, then $g(p) < 0$ for $p \in (0, p'_1)$ and $g(p) > 0$ for $p \in (p'_1, 1]$. In this case, as Fig.17c shows, (6) has one non-zero root $p_L < p'_1$, which is a steady-state point according to the approximate trajectory analysis in [23]. We then have $g(p_L) < 0$.
   b) If $p'_2 < 1$, then $g(p) < 0$ for $p \in (0, p'_1) \bigcup (p'_2, 1)$ and $g(p) > 0$ for $p \in (p'_1, p'_2)$. In this case, as Fig.17d-f show, (6) may have one steady-state point $p_L \in (0, p'_1) \bigcup (p'_2, 1)$, or three non-zero roots $p_A < p'_1 < p_S < p'_2 < p_L$, among which p_A and p_L are the steady-state points according to the approximate trajectory analysis in [23]. We then have $g(p_L) < 0$ and $g(p_A) < 0$.

Finally, we can conclude that $g(p_L) < 0$ and $g(p_A) < 0$. It can then be obtained from (26)-(29) that $\frac{\partial p}{\partial \hat{\lambda}}\Big|_{p=p_L,p_A} < 0$, $\frac{\partial p}{\partial q}\Big|_{p=p_L,p_A} < 0$, $\frac{\partial p}{\partial n}\Big|_{p=p_L,p_A} < 0$ and $\frac{\partial p}{\partial W}\Big|_{p=p_L,p_A} > 0$. ■

Appendix C
Proof of Theorem 2

Proof: According to (3), (4) and (6), the network throughput $\hat{\lambda}_{\text{out}}$ can be written as $\hat{\lambda}_{\text{out}} = -p\ln p$. It can be clearly seen that $\hat{\lambda}_{\max} = e^{-1}$, which is achieved when $p^* = e^{-1}$. (9) can then be obtained by substituting $p^* = e^{-1}$ into (6).

At the bistable region, the network may operate at the desired steady-state point p_L or the undesired steady-state point p_A. In the following, we prove that when (9) holds, $p_A < e^{-1}$, implying that $\hat{\lambda}_{\max}$ can only be achieved when the network operates at p_L. Specifically, (9) can be written as

$$\frac{2n}{\hat{\lambda}} - (W^*-1) = \left(2n - \frac{2}{q^*} - (W^*-1)\right)e. \tag{30}$$

From Fig. 17f, we can see $p_A \leq p'_1 < p'_2 \leq p_L$, where p'_1 and p'_2 are roots of $g(p) = 0$, which are given by (22), respectively. If (30) holds, then we have $p'_2 = \frac{n-\frac{2}{q^*}-W^*+1+\sqrt{n(n-\frac{4}{q^*}-2(W^*-1))}}{2n-\frac{2}{q^*}-W^*+1}e^{-1} < e^{-1}$. Therefore, $p_A < p'_2 < e^{-1}$. ■

References

[1] V. B. Mišić and J. Mišić, *Machine-to-Machine Communications: Architectures, Standards and Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[2] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, 2nd Quart., 2015.

[3] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digit. Commun. Netw.*, vol. 1, no. 1, pp. 1–19, 2015.

[4] "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020," Cisco, San Jose, CA, USA, White Paper, Feb. 2016. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf

[5] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.

[6] N. Abramson, "The ALOHA SYSTEM: Another alternative for computer communications," in *Proc. Fall Joint Comput. Conf.*, vol. 44. Nov. 1970, pp. 281–285.

[7] H. Takagi and L. Kleinrock, "Throughput analysis for persistent CSMA systems," *IEEE Trans. Commun.*, vol. 33, no. 7, pp. 627–638, Jul. 1985.

[8] R. MacKenzie and T. O'Farrell, "Throughput and delay analysis for p-persistent CSMA with heterogeneous traffic," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2881–2891, Sep. 2010.

[9] P. K. Wong, D. Yin, and T. T. Lee, "Analysis of non-persistent CSMA protocols with exponential backoff scheduling," *IEEE Trans. Commun.*, vol. 59, no. 8, pp. 2206–2214, Aug. 2011.

[10] J.-B. Seo, H. Jin, and V. C. M. Leung, "Throughput upper-bound of slotted CSMA systems with unsaturated finite population," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2477–2487, Jun. 2013.

[11] P. Osti, P. Lassila, S. Aalto, A. Larmo, and T. Tirronen, "Analysis of PDCCH performance for M2M traffic in LTE," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4357–4371, Nov. 2014.

[12] R. R. Tyagi, F. Aurzada, K.-D. Lee, and M. Reisslein, "Connection establishment in LTE-A networks: Justification of poisson process modeling," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2383–2394, Dec. 2017.

[13] R. R. Tyagi, F. Aurzada, K.-D. Lee, S. G. Kim, and M. Reisslein, "Impact of retransmission limit on preamble contention in LTE-advanced network," *IEEE Syst. J.*, vol. 9, no. 3, pp. 752–765, Sep. 2015.

[14] D. R. Morgan, "Random-Access channel queuing model," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2522–2527, Apr. 2016.

[15] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, Feb. 2016.

[16] C. H. Wei, G. Bianchi, and R. G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.

[17] M. Koseoglu, "Lower bounds on the LTE-A average random access delay under massive M2M arrivals," *IEEE Trans. Wireless Commun.*, vol. 64, no. 5, pp. 2104–2115, May 2016.

[18] V. Anantharam, "The stability region of the finite-user slotted ALOHA protocol," *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 535–540, May 1991.

[19] W. Szpankowski, “Stability conditions for some distributed systems: Buffered random access systems,” *Adv. Appl. Probab.*, vol. 26, no. 2, pp. 498–515, Jun. 1993.

[20] W. Luo and A. Ephremides, “Stability of *N* interacting queues in random-access systems,” *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1579–1587, Jul. 1999.

[21] 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.

[22] D. Niyato, P. Wang, and D. I. Kim, “Performance modeling and analysis of heterogeneous machine type communications,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2836–2849, May 2014.

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

[24] L. Dai, “Toward a coherent theory of CSMA and Aloha,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3428–3444, Jul. 2013.

[25] L. Dai and X. Sun, “A unified analysis of IEEE 802.11 DCF networks: Stability, throughput, and delay,” *IEEE Trans. Mobile Comput.*, vol. 12, no. 8, pp. 1558–1572, Aug. 2013.

[26] Y. Gao, X. Sun, and L. Dai, “IEEE 802.11e EDCA networks: Modeling, differentiation and optimization,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3863–3879, Jul. 2014.

[27] Y. Li and L. Dai, “Maximum sum rate of slotted Aloha with capture,” *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 690–705, Feb. 2016.

[28] X. Sun and L. Dai, “Performance optimization of CSMA networks with a finite retry limit,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5947–5962, Sep. 2016.

[29] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, document TS 36.321 V12.5.0, 3GPP, Apr. 2015.

[30] S. Lam and L. Kleinrock, “Packet switching in a multiaccess broadcast channel: Dynamic control procedures,” *IEEE Trans. Commun.*, vol. 23, no. 9, pp. 891–904, Sep. 1975.

[31] M. Ferguson, “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.

[32] G. Fayolle, E. Gelenbe, and J. Labetoulle, “Stability and optimal control of the packet switching broadcast channel,” *J. ACM*, vol. 24, no. 3, pp. 375–386, Jul. 1977.

[33] B. Hajek and T. van Loon, “Decentralized dynamic control of a multiaccess broadcast channel,” *IEEE Trans. Autom. Control*, vol. 27, no. 3, pp. 559–569, Jun. 1982.

[34] R. Rivest, “Network control by Bayesian broadcast,” *IEEE Trans. Inf. Theory*, vol. 33, no. 3, pp. 323–328, May 1987.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] G.-Y. Lin, S.-R. Chang, and H.-Y. Wei, “Estimation and adaptation for bursty LTE random access,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2560–2577, Apr. 2016.

[40] 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.

[41] J. Choi, “On the adaptive determination of the number of preambles in RACH for MTC,” *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1385–1388, Jul. 2016.

[42] 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.

[43] H. Thomsen, N. K. Pratas, Č. Stefanović, and P. Popovski, “Code-expanded radio access protocol for machine-to-machine communications,” *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 355–365, May 2013.

[44] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, “Fundamentals of throughput maximization with random arrivals for M2M communications,” *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4094–4109, Nov. 2014.

[45] D. T. Wiriaatmadja and K. W. Choi, “Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 33–46, Jan. 2015.

[46] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, “Probabilistic rateless multiple access for machine-to-machine communication,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6815–6826, Dec. 2015.

[47] J. Mišić, V. B. Mišić, and N. Khan, “Sharing it my way: Efficient M2M access in LTE/LTE-A networks,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 696–709, Jan. 2017.

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

[49] *Service Accessibility*, document TS 22.011 V11.3.0, 3GPP, Apr. 2013.

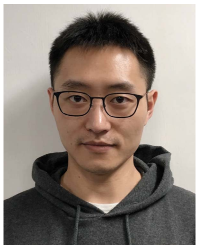
[50] K. Sakakibara, T. Seto, D. Yoshimura, and J. Yamakita, “Effect of exponential backoff scheme and retransmission cutoff on the stability of frequency-hopping slotted ALOHA systems,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 714–722, Jul. 2003.

[51] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. Hoboken, NJ, USA: Wiley, 1998.

[52] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1992.

[53] *MTC Simulation Results With Specific Solutions*, document TSG RAN WG2 #71 R2-104662, 3GPP, Aug. 2010.

[54] *Study on RAN Improvements for Machinetype Communications*, document TR 37.868 V11.0.0, 3GPP, Oct. 2011.



Wen Zhan (S’17) received the B.S. and M.S. degrees from University of Electronic Science and Technology of China, Chengdu, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. His research interests include the Internet of Things, machine-to-machine communications, and wireless random access networks.

Lin Dai (S’00–M’03–SM’13) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1998, and the M.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2003, all in electronic engineering. She was a Post-Doctoral Fellow with The Hong Kong University of Science and Technology and the University of Delaware. Since 2007, she has been with City University of Hong Kong, where she is currently an Associate Professor. She has broad interests in communications and networking theory, with special interests in wireless communications. She was a co-recipient of the Best Paper Award at the IEEE Wireless Communications and Networking Conference in 2007 and the IEEE Marconi Prize Paper Award in 2009.