Throughput Optimization for Massive Random Access of M2M Communications in LTE Networks

Wen Zhan and Lin Dai
Department of Electronic Engineering, City University of Hong Kong, Hong Kong
wzhan5-c@my.cityu.edu.hk, lindai@cityu.edu.hk

***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 M2M communications in LTE networks. Both the maximum network throughput and the corresponding optimal backoff parameters including the Access Class Barring (ACB) factor and the backoff window size are obtained as explicit functions of key system parameters such as the number of preambles, the number of Machine Type Devices (MTDs) and the aggregate input rate. The analysis is verified by simulations and sheds important light on practical network design for supporting massive access of M2M communications in LTE networks.**

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

I. Introduction

Machine-to-Machine (M2M) communications has gained worldwide attention and facilitated numerous innovative applications such as smart grid, intelligent transportation and e-health [1]. To enable M2M communications, the existing cellular system, i.e, Long Term Evolution (LTE), has become the most widely considered option owing to its ubiquitous coverage [2].

A significant challenge for supporting M2M communications in LTE networks is that when a large number of Machine Type Devices (MTDs) attempt to initiate connection with the Base Station (BS), massive access requests will be generated and transmitted via a shared channel, causing severe congestion and intolerably low access efficiency [3], [4]. With a drastic increase in the number of MTDs [5], it is of paramount importance to study how to improve the access efficiency of M2M communications, especially in massive access scenarios.

According to the contention-based random access procedure in the current LTE standard [3], each MTD determines when to access the BS in a distributed manner. A key performance metric to evaluate the random access efficiency is network throughput, which is defined as the average number of MTDs that can successfully access the BS per unit time. Although it has been long observed that the network performance drastically degrades when the number of MTDs is large [4], [6], how to properly tune the backoff parameters to optimize the network throughput remains largely unknown.

Specifically, to evaluate the access performance, various analytical models have been proposed. [7]–[9] focused on modeling and estimating the aggregate traffic, i.e., the total number of access requests in each time unit. Due to ignoring the queueing behavior of each individual MTD, little light can be shed on the effects of device-level input parameters such as the traffic input rate of each MTD on the network performance. In [10]–[12], queueing models for each MTD were further incorporated into the analysis. All of them, nevertheless, require to develop iterative algorithms to solve system equations, where the computational complexity increases with the number of MTDs, and becomes prohibitively high in the massive access scenario.

To improve the access efficiency, various algorithms were also proposed to adaptively tune backoff parameters including the Access Class Barring (ACB) factor [13]–[16] and the backoff window size [17] based on the realtime information of the aggregate traffic. The effectiveness of those algorithms is largely dependent on the accuracy of estimation of the aggregate traffic, which in practice is difficult to capture [18]. As we will demonstrate in this paper, such information is indeed unnecessary if the objective is to optimize long-term network performance such as the network throughput, in which case the optimal tuning can be solely based on statistical input parameters such as the traffic input rate of each MTD.

In this paper, a new analytical framework is proposed to optimize the random access performance of M2M communications in LTE networks. By introducing a novel double-queue model for each MTD and characterizing the state transition of each access request, the network throughput is derived as an explicit function of key system parameters such as the total number of MTDs and the aggregate input rate, based on which explicit expressions of the maximum network throughput and the corresponding optimal backoff parameters including the ACB factor and the backoff window size are further obtained. The analysis shows that although the maximum network throughput is solely determined by the number of preambles, to achieve it, either the ACB factor or the backoff window size should be tuned based on the total number of MTDs, the number of preambles and the traffic input rate of each MTD.

The remainder of the paper is organized as follows. Section II illustrates the system model. The throughput analysis with one single preamble is presented in Section III, and extended to the multi-preamble scenario in Section IV. The analysis is verified by the simulation results provided in Section V. Finally, concluding remarks are summarized in Section VI.

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

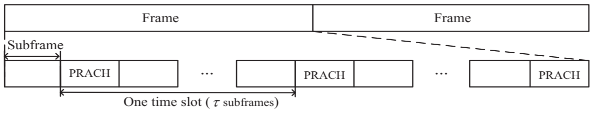


Fig. 1. Frame structure of the LTE system in the Frequency Division Duplex (FDD) mode.

II. System Model

Consider a single-cell LTE system with n MTDs attempting to access the BS. In the random access procedure, each MTD randomly selects one out of M orthogonal preambles and transmits via the Physical Random Access CHannel (PRACH) to the BS [3]. The PRACH consists of a series of subframes that appear periodically [19], 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[1].

We are interested in the network throughput performance of the random access procedure, 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.

A. Queuing Model of Each MTD

For each MTD, one access request is generated once it has data packets in the buffer. Assume that each access request will not be dropped until its successful transmission, upon which the BS will assign sufficient resources for the MTD to clear its data buffer. As Fig. 2 illustrates, each MTD has one data queue and one request queue. Assume that the data buffer has an infinite size and the arrival of data packets follows a Bernoulli process with parameter λ. Each newly arrival data packet generates an access request, but only one request can be kept since each MTD can have at most one access request regardless of how many data packets in its buffer [3]. 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 [3], [20], each MTD needs to perform the Access Class Barring (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 Uniform Backoff (UB) window size in unit of milliseconds[2], and counts down until it reaches zero. Fig. 3 shows the state transition process of each individual access request. The states of access request can be divided into two categories: 1) successful transmission (State T), and 2) waiting to transmit (State i, $i \in \{0, 1, \ldots, W-1\}$).

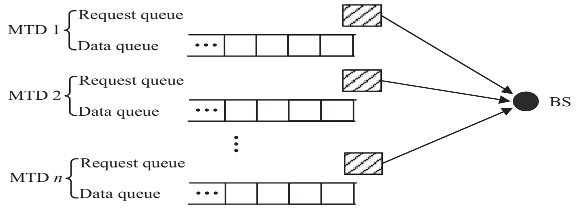


Fig. 2. In the proposed double-queue model, each MTD has one data queue and one request queue.

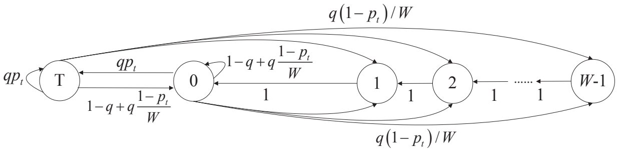


Fig. 3. State transition diagram of each individual access request.

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.

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

$$\pi_T = \left(\frac{1}{qp} + \frac{(1-p)(W-1)}{2p}\right)^{-1}, \tag{1}$$

and

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

where $p = \lim_{t\to\infty} p_t$ is the steady-state probability of successful transmission of access requests.

[1] Note that the random access procedure of LTE standard has four steps [3]. Similar to [10]–[15], this paper assumes that the success of connection is mainly determined by the first step, and thus only focuses on the access performance of the first step.

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

III. Throughput Analysis with $M = 1$

In this paper, we focus on the network throughput $\hat{\lambda}_{\text{out}}$, which is defined as the average number of successful access requests per time slot. Let us first consider the scenario where all n MTDs share one preamble, i.e., $M = 1$. Extension to the multi-preamble scenario will be presented in Section IV.

Based on the $Geo/G/1/1$ model of each request queue, the network throughput can be obtained as

$$\hat{\lambda}_{\text{out}} = \hat{\lambda}(1 - \rho), \tag{3}$$

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 [21]

$$\rho = \frac{\lambda}{\lambda + \pi_T}, \tag{4}$$

where π_T is the steady-state probability that the access request is in State T, which is also the service rate of each request queue as the successful output occurs if and only if the access request is in State T. By combining (1), (3) and (4), we have

$$\hat{\lambda}_{\text{out}} = \frac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{qp} + \frac{(1-p)(W-1)}{2p}\right) + 1}. \tag{5}$$

It can be observed from (5) that the throughput performance is closely determined by p, the steady-state probability of successful transmission of access requests. In the following, we first characterize the network steady-state points based on the fixed-point equation of p, and then derive the maximum network throughput and the corresponding optimal setting of the ACB factor q and the UB window size W.

A. Steady-State Points

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. The probability that an MTD 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

$$\begin{aligned} p &= \left(1 - \rho + \rho\left(\sum_{j=1}^{W-1} \pi_j + (1-q)(\pi_0 + \pi_T)\right)\right)^{n-1} \\ &= \left(1 - \frac{\rho\pi_T}{p}\right)^{n-1}. \end{aligned} \tag{6}$$

By combining (1), (4) and applying $n - 1 \approx n$, $(1-x)^n \approx \exp\{-nx\}$ for $0 < x < 1$ if n is large, (6) can be approximated by

$$p \overset{\text{with a large } n}{\approx} \exp\left(-\frac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{q} + \frac{W-1}{2}\right) + p\left(1 - \frac{\hat{\lambda}(W-1)}{2n}\right)}\right). \tag{7}$$

Theorem 1 shows that (7) has either one or three non-zero roots. The proof is omitted due to limited space.

Theorem 1. *The fixed-point equation (7) 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*

$$\hat{\lambda}_1 = \frac{2n}{\frac{n - \frac{2}{q} - W + 1 - \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}{\exp\left(-\frac{2n}{n - \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}\right)} + W - 1}, \tag{8}$$

$$\hat{\lambda}_2 = \frac{2n}{\frac{n - \frac{2}{q} - W + 1 + \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}{\exp\left(-\frac{2n}{n + \sqrt{n\left(n - \frac{4}{q} - 2(W-1)\right)}}\right)} + W - 1}. \tag{9}$$

Otherwise, (7) has only one non-zero root $0 < p_L \le 1$.

Note that not all the roots of (7) are steady-state points. We follow the approximate trajectory analysis proposed in [22], and find that:

(1) If (7) has only one non-zero root p_L, then p_L is a steady-state point;

(2) If (7) has three non-zero roots $p_A \le p_S \le p_L$, then only p_L and p_A are steady-state points. Similar to [23], we refer to p_L as the desired steady-state point and p_A the undesired steady-state point.

B. Throughput Optimization

It can be seen from (5) and (7) that the network throughput $\hat{\lambda}_{\text{out}}$ depends on the number of MTDs n, the aggregate input rate $\hat{\lambda}$, the ACB factor q and the UB window size W. Typically, n and $\hat{\lambda}$ are system input parameters. Therefore, we focus on optimizing the network throughput by tuning q and W for given n and $\hat{\lambda}$. Specifically, 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^*, W^*). The proof is omitted due to limited space.

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*

$$\frac{1}{q^*} + \frac{1-e^{-1}}{2}(W^* - 1) = n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right). \tag{10}$$

It can be seen 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 (10) does not hold for any $q \in (0, 1]$ and $W \ge 1$. On the other hand, when the aggregate input rate

$$\hat{\lambda} \ge 4e^{-2} \ge \lim_{n \to 2\left(\frac{2}{q} + W - 1\right)} \hat{\lambda}_2 = \frac{4e^{-2}}{1 + \frac{W-1}{\frac{2}{q} + W - 1} e^{-2}}, \tag{11}$$

the network is guaranteed to operate at p_L according to

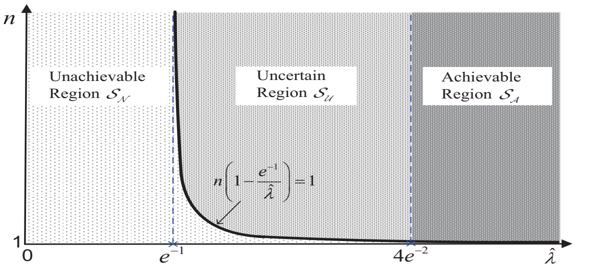

n
Unachievable Region $\mathcal{S}_{\mathcal{N}}$
Uncertain Region $\mathcal{S}_{\mathcal{U}}$
Achievable Region $\mathcal{S}_{\mathcal{A}}$
$n\left(1-\frac{e^{-1}}{\hat{\lambda}}\right)=1$
1
0
e^{-1}
$4e^{-2}$
$\hat{\lambda}$

Fig. 4. Unachievable region $\mathcal{S}_{\mathcal{N}}$, achievable region $\mathcal{S}_{\mathcal{A}}$ and uncertain region $\mathcal{S}_{\mathcal{U}}$.

Theorem 1. 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 (10). Accordingly, we can define the following regions of $(n, \hat{\lambda})$:

- Unachievable region $\mathcal{S}_{\mathcal{N}} = \left\{(n, \hat{\lambda}) | n\left(1-\frac{e^{-1}}{\hat{\lambda}}\right) < 1\right\}$, in which $\hat{\lambda}_{\max}$ is unachievable regardless of what values of q and W are chosen.
- Achievable region $\mathcal{S}_{\mathcal{A}} = \left\{(n, \hat{\lambda}) | n\left(1-\frac{e^{-1}}{\hat{\lambda}}\right) \geq 1, \hat{\lambda} \geq 4e^{-2}\right\}$, in which $\hat{\lambda}_{\max}$ can be achieved when q and W are tuned according to (10).
- Uncertain region $\mathcal{S}_{\mathcal{U}} = \left\{(n, \hat{\lambda}) | n\left(1-\frac{e^{-1}}{\hat{\lambda}}\right) \geq 1, e^{-1} < \hat{\lambda} < 4e^{-2}\right\}$, in which the network may operate at the desired steady-state point p_L or the undesired steady-state point p_A. $\hat{\lambda}_{\max}$ is achievable only if the network operates at p_L.

Fig. 4 illustrates the above regions. Note that the maximum network throughput $\hat{\lambda}_{\max}$ can be achieved by tuning the ACB factor q or the UB window size W. For instance, according to Theorem 2, by fixing $W = 1$, the optimal ACB factor $q^*_{W=1}$ is given by

$$q^*_{W=1} = \frac{\hat{\lambda}}{n(\hat{\lambda} - e^{-1})}. \tag{12}$$

On the other hand, if $q = 1$, then the optimal UB window size $W^*_{q=1}$ is given by

$$W^*_{q=1} = \frac{2n\left(1-\frac{e^{-1}}{\hat{\lambda}}\right) - 2}{1 - e^{-1}} + 1. \tag{13}$$

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

Note that the throughput analysis in Section III 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.

By virtue of the orthogonality among preambles, only the MTDs who share the same preamble contend with each other. 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 (7) 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 (5) 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 (10) for given $n^{(i)}$ and $\hat{\lambda}^{(i)}$. The network throughput is then given by

$$\hat{\lambda}_{\text{out}} = \sum_{i=1}^{M} \hat{\lambda}_{\text{out}}^{(i)}, \tag{14}$$

which is maximized at Me^{-1} 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 [3]. 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 $n^{(i)} \approx \frac{n}{M}$. As we will demonstrate in the next section, such an approximation is accurate in the massive access scenario, i.e., $n \gg M$.

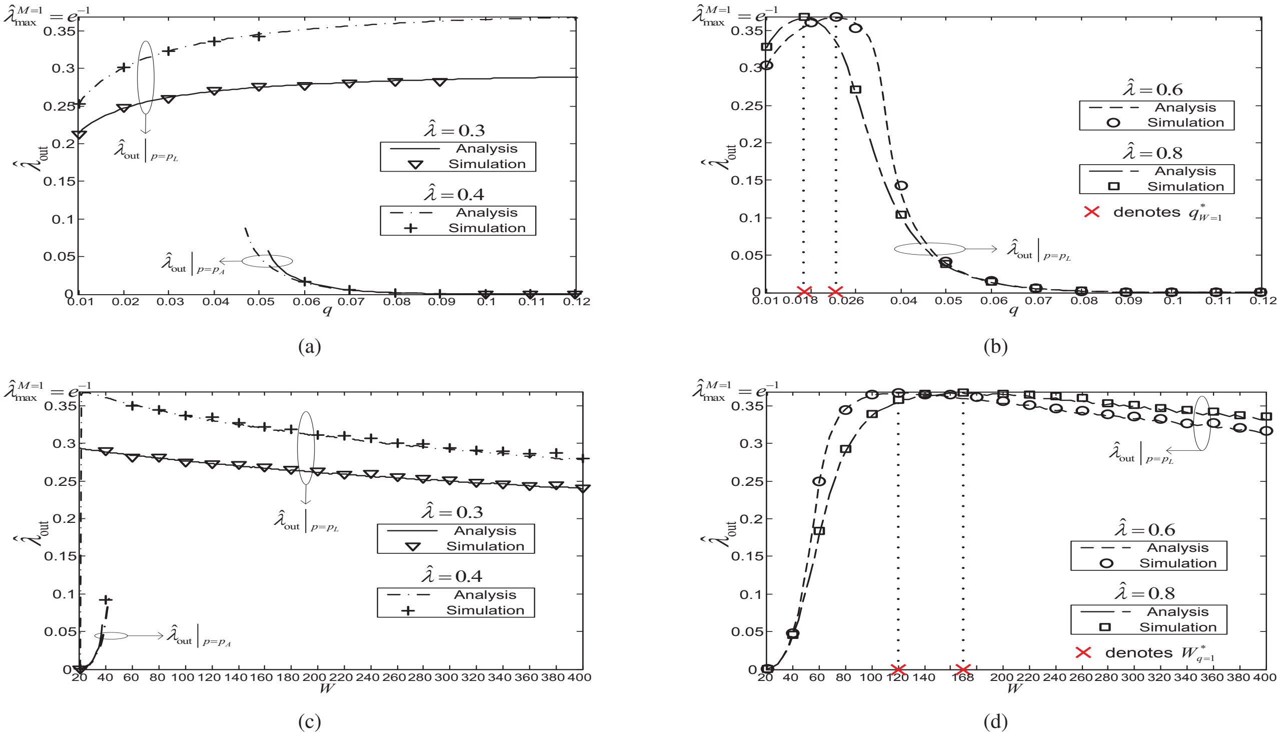
V. Simulation Results

In this section, we present simulation results to validate the preceding analysis. The simulation setting is the same as the system model described in Section II, and we omit details here. Each simulation is carried out for 10^8 time slots. In the following, let us first consider the scenario in which all n MTDs share the same preamble, i.e., $M = 1$.

A. Network Throughput with $M = 1$

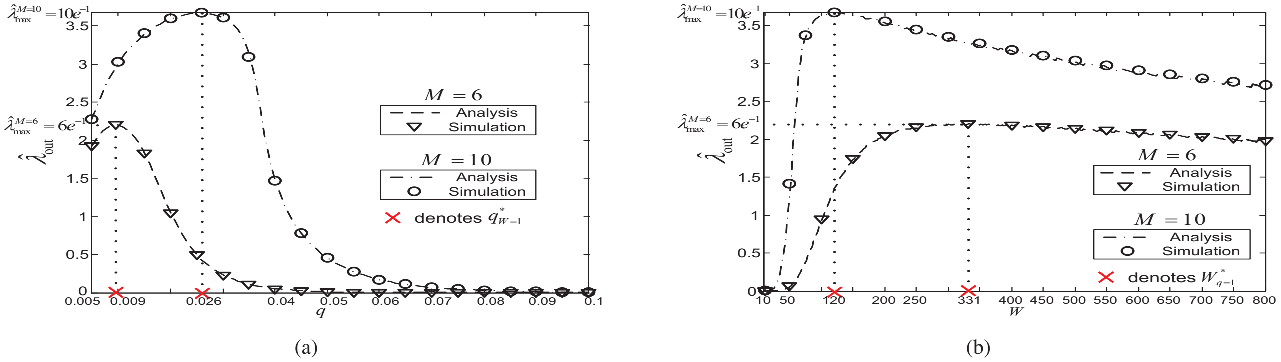
The network throughput $\hat{\lambda}_{\text{out}}$ with $M = 1$ is given by (5) in Section III, and is shown to be crucially determined by the number of MTDs n, the aggregate input rate $\hat{\lambda}$, the ACB factor q, the UB window size W, and which steady-state point the network operates at, i.e., the desired steady-state point p_L or the undesired steady-state point p_A. In simulations, the network throughput is obtained by calculating the ratio of the number of the successful access requests to the number of time slots, i.e., 10^8 time slots. The analysis is verified by the simulation results presented in Fig. 5.

Specifically, Fig. 5a-b illustrate how the network throughput $\hat{\lambda}_{\text{out}}$ varies with the ACB factor q with the UB window size



(a) (b) (c) (d)

Fig. 5. 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) $q = 1$.



(a) (b)

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

$W = 1$ and the number of MTDs $n = 100$. In Fig. 5a, 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. 5b, 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 (12), i.e., $q = q^*_{W=1}$. Similar observations can be obtained from Fig. 5c-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 (13).

B. Network Throughput with $M > 1$

In this section, we provide simulation results to verify the throughput analysis in the multi-preamble scenario presented in Section IV. In simulations, each MTD independently and randomly selects one out of M orthogonal preambles in each access attempt. Fig. 6 illustrates how the network throughput $\hat{\lambda}_{\text{out}}$ 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 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 $\hat{\lambda}_{\max}$ linearly increases with the number of preambles M, i.e., $\hat{\lambda}_{\max} = Me^{-1}$. To achieve $\hat{\lambda}_{\max}$, the ACB factor q or the UB window size W

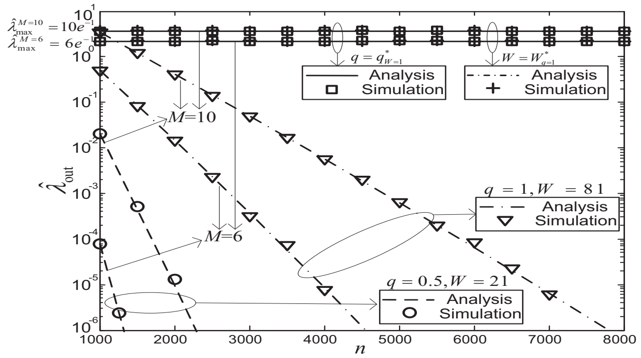


Fig. 7. Network throughput $\hat{\lambda}_{\text{out}}$ versus the number of MTDs n. $\lambda = 0.006$.

should be tuned adaptively based on the aggregate input rate $\hat{\lambda}$, the number of MTDs n and the number of preambles M.

Note that in the current standard setting, the ACB factor q and the UB window size W are preselected from a certain range [3]. To see the performance loss without adaptive tuning of q and W, Fig. 7 illustrates the network throughput performance with two representative settings of parameters: $\{q = 0.5, W = 21\}$, $\{q = 1, W = 81\}$ [6].[3] It can be clearly observed that in both cases, the network throughput $\hat{\lambda}_{\text{out}}$ quickly deteriorates as the number of MTDs n increases and becomes much lower than the maximum network throughput $\hat{\lambda}_{\max}$ when n is large. In sharp contrast, if the ACB factor q or the UB window size W is optimally selected, i.e., $q = q^*_{W=1}$ or $W = W^*_{q=1}$, the maximum network throughput $\hat{\lambda}_{\max}$ 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. Conclusion

In this paper, a new analytical framework is proposed to optimize the random access performance of M2M communications in LTE networks. The modeling complexity is shown to be 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 scenarios. Explicit expressions of the network throughput and the optimal setting are obtained, which reveal that the maximum network throughput is solely determined by the number of preambles. Yet to achieve the maximum network throughput, either the ACB factor q or the UB window size W should be tuned based on the aggregate input rate $\hat{\lambda}$, the number of MTDs n and the number of preambles M.

The analysis sheds important light on practical network design for supporting massive access of M2M communications in LTE networks. In the current standard, a preselection of the ACB factor q and the UB window size W would lead to drastic throughput degradation as more MTDs attempt to access the BS. By optimally tuning q or W, substantial gains can be achieved, where the network throughput remains at the highest level regardless of the number of MTDs.

[3] Note that in the standard [3], the values of UB backoff 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 [19], and then $W = 21$ and $W = 81$ are corresponding to $W_s = 20$ msec and $W_s = 80$ msec, respectively.

References

[1] V. B. Misic and J. Misic, *Machine-to-machine Communications: Architectures, Standards and Applications*. CRC Press, 2014.

[2] 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, Feb. 2015.

[3] 3GPP TS 36.321 V12.5.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," Apr. 2015.

[4] 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, First Quarter 2014.

[5] Cisco whitepaper, "Cisco visual networking index: global mobile data traffic forecast update, 2015-2020," Feb. 2016.

[6] 3GPP TSG RAN WG2 #71 R2-104662, "MTC simulation results with specific solutions," Aug. 2010.

[7] 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.

[8] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372-375, Feb. 2015.

[9] R. G. Cheng, J. Chen, D. W. Chen, and C. H. Wei, "Modeling and analysis of an extended access barring scheme for machine-type communications in LTE networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956-2968, Feb. 2015.

[10] 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, Apr. 2014.

[11] J. B. Seo and V. 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.

[12] X. Yang, A. O. Fapojuwo, and E. E. Egbogah, "Performance analysis and parameter optimization of random access backoff algorithm in LTE," in *Proc. VTC*, Sept. 2012.

[13] 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.

[14] H. Wu, C. Zhu, R. 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.

[15] 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, Mar. 2015.

[16] Z. Wang and V. 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, Jun. 2015.

[17] 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.

[18] 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.

[19] 3GPP TS 36.211 V10.4.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," Dec. 2011.

[20] 3GPP TS 22.011 V11.3.0, "Service accessibility," Apr. 2013.

[21] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. Wiley, 1998.

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

[23] D. Bertsekas and R. Gallager, *Data Networks*. Prentic Hall, 1992.