Massive Random Access of Machine-to-Machine Communications in LTE Networks: Throughput Optimization With a Finite Data Transmission Rate

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

***Abstract*— This is a sequel of our previous work [20] on access throughput optimization of Machine-to-Machine (M2M) communications in Long Term Evolution (LTE) networks. By incorporating a finite data transmission rate, this paper aims to characterize the effect of data transmission on the optimal access performance of Machine-Type Devices (MTDs). Specifically, both the maximum access throughput and the corresponding optimal Access Class Barring (ACB) factor are obtained as explicit functions of the data transmission rate, which show that even with the ACB factor optimally tuned, the access throughput may deteriorate as the number of MTDs increases, and even drop to zero if the data transmission rate is too small. To boost the data transmission rate, more resources should be allocated to data transmission, which, however, leads to fewer chances for access. In light of the tradeoff between the data transmission rate and the access frequency, the time slot length is further optimized for maximizing the normalized maximum access throughput. Simulation results corroborate that by properly choosing the time slot length, substantial gains can be achieved over the default setting in various scenarios.**

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

I. Introduction

MACHINE-to-Machine (M2M) communications is a new service type identified by the Third-Generation Partnership Project (3GPP) for the Long Term Evolution (LTE) networks. It usually involves many Machine-Type Devices (MTDs) that can actuate, exchange and process data without human intervention, which has found wide applications in various domains such as smart city, Industry 4.0 and e-health [1]. With the explosive growth of the number of MTDs [2], however, the deluge of access requests generated by MTDs may easily lead to severe congestion with intolerably low chances of success. Therefore, how to efficiently facilitate the access of a massive number of MTDs has become a significant challenge for supporting M2M communications over LTE networks [3].

The random access process of LTE networks is based on Aloha [4]. Yet, different from the classical Aloha where each packet has to contend for channel access, in LTE networks, a connection-based random access process is adopted. That is, each device with packets to transmit first sends an access request to the Base Station (BS) to establish a connection, and then the BS would assign resource blocks for the device to clear its data queue [5]. Another distinguished feature of the LTE random access is that the access requests can only be sent on the Physical Random Access CHannel (PRACH) subframes that appear periodically [6]. The period of PRACH subframes is a key system parameter that determines how the resources are allocated between access and data transmission [7].

Extensive studies have focused on modeling and evaluating the access performance of MTDs in LTE networks, and demonstrated that the access efficiency crucially depends on two access parameters, that is, the Access Class Barring (ACB) factor and the Uniform Backoff (UB) window size [8]–[13]. Accordingly, various algorithms have been developed to adaptively tune the access parameters based on the estimation of the time-varying number of access requests [14]–[19]. To maximize the access efficiency of M2M communications in LTE networks, a new analytical framework was recently proposed in our work [20]. Specifically, to capture the essence of the connection-based random access, a novel double-queue model was established, which can both incorporate the queueing behavior of each MTD and be scalable in the massive access scenario. The access efficiency is evaluated by the access throughput, i.e., the average number of successful access requests per PRACH subframe, which is optimized by properly tuning the access parameters including the ACB factor and the UB window size. Explicit expressions of the maximum access throughput and the corresponding optimal access parameters were obtained, which show that the maximum access throughput 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.

Note that a key assumption in the above studies is that for each MTD, once its access request is successful, it can always clear its data queue within one time slot, i.e., one period of PRACH subframes. Though a good approximation for light-traffic scenarios that are common for M2M communications,

Manuscript received November 23, 2018; revised May 21, 2019; accepted August 21, 2019. Date of publication September 10, 2019; date of current version December 10, 2019. This work was supported by the Research Grants Council (RGC) of Hong Kong under GRF Grant CityU 11212018. The associate editor coordinating the review of this article and approving it for publication was K. R. Chowdhury. *(Corresponding author: Wen Zhan.)*

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

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

Digital Object Identifier 10.1109/TWC.2019.2938955

this assumption may not hold true when the traffic load becomes heavy or the resource for data transmission is insufficient. In that case, it may take more than one time slot for MTDs to clear their data queues. Intuitively, with prolonged data transmission time, the access efficiency would decrease because newly generated access requests may not be accommodated until the ongoing data transmission completes. It may even drop to zero when the data transmission rate is too small to clear the data queues. To improve the access efficiency, more resources may be allocated to data transmissions to boost the data transmission rate, with which, however, the time slot length, i.e., the period of PRACH subframes, would be enlarged, indicating that the MTDs can access the channel less frequently. Apparently, the time slot length determines a crucial tradeoff between the data transmission rate and the access frequency, which should be properly set to optimize the access performance.

In this paper, the analytical framework proposed in [20] is extended to analyze the effect of the data transmission rate β, which is defined as the total number of data packets that can be transmitted per time slot, on the optimal access performance of MTDs in LTE networks. Specifically, based on the double-queue model, a discrete-time Markov renewal process is established to characterize the behavior of each access request, where a data transmission state is introduced to describe the case of a data transmission lasting for more than one time slot. To evaluate the access efficiency, the access throughput is characterized and maximized by optimally tuning the ACB factor. Both the maximum access throughput and the optimal ACB factor are obtained as explicit functions of the data transmission rate β, the number of preambles M, the number of MTDs n and the traffic input rate of each MTD λ. The analysis shows that the maximum access throughput is a monotonic increasing function of the data transmission rate β, which becomes zero when β is smaller than the aggregate traffic input rate. In that case, the data throughput, which is defined as the average number of transmitted data packets per time slot, reaches the maximum value, but the network becomes unstable as the data queues can never be cleared.

For improving the data transmission rate, a larger time slot length should be chosen, which, nevertheless, reduces the access frequency of MTDs. The analysis further demonstrates that to optimize the access performance, the time slot length should be carefully selected based on the number of MTDs, the traffic input rate and data transmission rate per subframe. The optimal time slot length for maximizing the normalized access throughput, i.e., the average number of successful access requests per millisecond, is characterized, and shown to lead to significant gains over the default setting in various scenarios.

It is worth mentioning that for data transmission performance of M2M communications, a lot of efforts have been made to maximize the sum rate [21]–[23] or the energy efficiency [24]–[26] by optimally allocating resource blocks to MTDs based on constraints such as the maximum transmit power of each MTD [22]–[24], or queueing delay bound of data packets [24]. By assuming that connections between the BS and MTDs have already been established, the access process is usually ignored in those studies. Moreover, this paper should also be distinguished from [27]–[32], where the focus is on proposing new access and data transmission schemes for M2M communications, rather than optimizing the performance of the current LTE random access process.

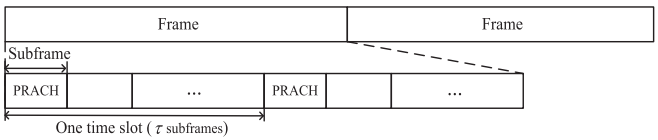


Fig. 1. Frame structure of the LTE system in the Frequency Division Duplex (FDD) mode.

The remainder of this paper is organized as follows. Section II presents the system model. The network steady-state points and throughput performance with one single preamble are characterized in Section III and Section IV, respectively, and extended to the multi-preamble scenario in Section V. The characterization of optimal time slot length is presented in Section VI. Finally, concluding remarks are summarized in Section VII.

II. System Model

Consider a single-cell LTE system with n MTDs attempting to access the BS for uplink data transmission. In the random access procedure, each MTD randomly selects one out of M orthogonal preambles and transmits to the BS via the PRACH [5]. The PRACH consists of a series of subframes that appear periodically, as shown in Fig. 1. We define a time slot as the interval between two consecutive PRACH subframes.[1] Accordingly, each MTD can transmit one access request in each time slot. If more than one MTD transmits the same preamble for a given time slot, then a collision occurs and all of them fail. The access request transmission is successful if and only if there is one single MTD transmitting for a given preamble at each time slot.

A. Double-Queue Model of Each MTD

As we mentioned in [20], different from the conventional packet-based random access where each single data packet has to contend for channel access, the random access process of LTE systems is connection-based, that is, once an MTD's request is successfully received, the BS will allocate resource blocks for the device to clear its data queue. To characterize each MTD's behavior in the connection-based random access, a double-queue model was proposed in [20]. As Fig. 2 illustrates, each MTD has one data queue and one request queue. Only the request queue is involved in the contention, and each access request stays in the queue until it is successfully transmitted. It can be seen from Fig. 2 that for MTDs, the probability of successful access is crucially determined by the state transition process of each access request, which will be characterized in Section III-A.

[1]The time slot defined in this paper should be distinguished from the slot defined in the standard [6]. The length of a slot in the LTE standard has a fixed value of 0.5 millisecond and two slots constitute one subframe.

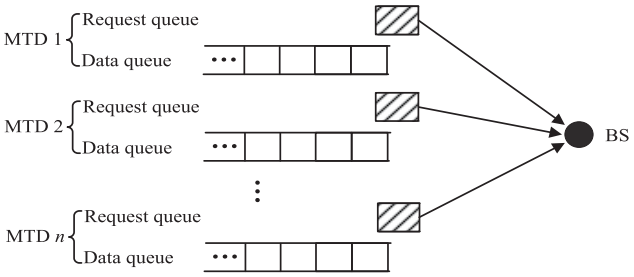


Fig. 2. Double-queue model of each MTD.

Assume that the data buffer has an infinite size and the arrivals of data packets follow a Bernoulli process 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 are in its buffer [5]. Each MTD's request queue can then be modeled as a $Geo/G/1/1$ queue.

B. Data Transmission Rate β

It was assumed in [20] that the BS can always allocate sufficient resources for the MTDs with successful access requests to clear their data queues within one time slot. In practice, however, the number of data packets that can be transmitted per time slot is determined by the time-frequency resources allocated, which may not always be sufficiently large. To further take the resource constraint into consideration,[2] in this paper, a finite data transmission rate $\beta \geq 1$ is assumed, that is, in each time slot, a total of β data packets can be transmitted. When β is small, it may take multiple time slots for those MTDs with successful access requests to clear their data queues. Apparently, the scenario in [20] can be regarded as a special case of $\beta = +\infty$, with which each data queue can always be cleared within one time slot.

C. Access Throughput and Data Throughput

Note that with an unlimited data transmission rate, the average number of transmitted data packets per time slot, which is referred to as the data throughput, is always equal to the aggregate traffic input rate. Therefore, the focus of [20] was on the optimization of access throughput, which evaluates the access efficiency and is defined as the average number of successful access requests per time slot. As we will demonstrate in this paper, when the data transmission rate is finite, the maximization of data throughput is achieved at the cost of sacrificing the access throughput. Both the data throughput and the access throughput will be analyzed and optimized in this paper.

As MTDs contend with each other only when they choose the same preamble, in the following, we start from the single-preamble scenario, where all MTDs share one preamble, i.e., $M = 1$. The analysis will be extended to the multi-preamble scenario in Section V.

[2]It is worth mentioning that the LTE standard does not specify how the BS should allocate resources to MTDs [33]. In practice, the BS may adopt different kinds of resource scheduling algorithms. Despite the differences in the resource scheduling algorithms, their effect on the access performance of MTDs can be well captured by the total number of transmitted data packets in each time slot, i.e., data transmission rate β.

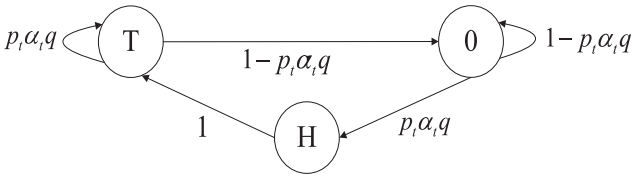


Fig. 3. Embedded Markov chain $\{X_j\}$ of the state transition process of each individual access request.

III. Steady-State Point Analysis

Based on the double-queue model, each MTD has one request queue and one data queue, and only the request queue is involved in the contention. In this section, we will first focus on the state transition process of each access request, i.e., the behavior of the head-of-line packet of the request queue, and then characterize the network steady-state points based on it.

A. State Characterization of Access Request

According to the current LTE standard [5], [34], 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 $\{0, \ldots, W-1\}$, where W is the UB window size in unit of time slots, and counts down until it reaches zero. As it has been demonstrated in [20] that the maximum access throughput can be achieved by either tuning the ACB factor q or the UB window size W, in this paper, we let $W = 1$ and only consider the tuning of ACB factor q.[3]

To model the behavior of each individual access request, a discrete-time Markov renewal process $(\mathbf{X}, \mathbf{V}) = \{(X_j, V_j), j = 0, 1, \ldots\}$ is established, where X_j denotes the state of a tagged access request at the j-th transition and V_j denotes the epoch at which the j-th transition occurs. Fig. 3 shows the embedded Markov chain $\mathbf{X} = \{X_j\}$, which has three states: 1) successful transmission (State T), 2) waiting to request (State 0) and 3) data transmission (State H). Different from Fig. 3 in [20], here a new state, i.e., State H, is introduced to denote the data transmission state, because each data transmission may last for more than one time slot for small data transmission rate β.

Note that an MTD can successfully transmit its access request only when no other MTD is in the data transmission state. Let α_t denote the probability that no access request is in State H at time slot t, and p_t denote the probability

[3]More specifically, it was shown in [20] that for achieving the maximum access throughput, the optimal backoff parameters including the ACB factor q and the UB window size W should satisfy an equality constraint. By fixing one and optimally tuning the other backoff parameter, the maximum access throughput can always be achieved.

of successful transmission of access requests given that no access request is in State H at time slot $t, t = 1, 2, \ldots$. As shown in Fig. 3, a fresh access request is initially in State T. It remains in State T if it passes the ACB check and is successfully transmitted[4] given that no access request is in State H. Otherwise, it moves to State 0. It leaves State 0 for State H if it passes the ACB check and is successfully transmitted given that no other access request is in State H. It eventually shifts from State H to State T when the data transmission finishes.

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

$$\begin{cases} \pi_T = \left(1 - p\alpha q + \dfrac{1}{p\alpha q}\right)^{-1}, \\ \pi_H = (1 - p\alpha q)\pi_T, \\ \pi_0 = \dfrac{1 - p\alpha q}{p\alpha q}\pi_T, \end{cases} \tag{1}$$

where $\alpha = \lim_{t\to\infty} \alpha_t$ is the steady-state probability that no access request is in State H, and $p = \lim_{t\to\infty} p_t$ is the steady-state probability of successful transmission of access requests given that no access request is in State H.

The interval between successive transitions, i.e., $V_{j+1} - V_j$, is called the holding time in State X_j, which solely depends on State $X_j, j = 1, 2, \ldots$. Let τ_i denote the mean holding time in State i, where $i \in \{0, \mathrm{T}, \mathrm{H}\}$. The holding time in State T and that in State 0 are both one time slot, i.e.,

$$\tau_0 = \tau_T = 1. \tag{2}$$

The mean holding time of State H τ_H is determined by the data transmission rate β and the input rate λ of each MTD's data queue. Appendix A shows that τ_H can be obtained as

$$\tau_H \approx \frac{\lambda}{\beta p\alpha q}, \tag{3}$$

where the approximation is introduced mainly by dropping the rounding down operation for analytical tractability.

Finally, the limiting state probability of the Markov renewal process $(\mathbf{X}, \mathbf{V})$ is given by

$$\tilde{\pi}_i = \frac{\pi_i \tau_i}{\sum_{j\in\mathbb{S}} \pi_j \tau_j}, \tag{4}$$

$i \in \mathbb{S}$, where $\mathbb{S}$ is the state space of $\mathbf{X}$. Specifically, the probability of the access request being in State T can be obtained by combining (1)–(4) as

$$\tilde{\pi}_T = \frac{p\alpha q}{1 + \frac{\lambda}{\beta}(1 - p\alpha q)}. \tag{5}$$

Note that $\tilde{\pi}_T$ is also the service rate of each MTD's request queue as each request queue has a successful output if and only if the access request is in State T. Based on the $Geo/G/1/1$ model of each request queue, the probability that each request queue is nonempty is given by

$$\rho = \frac{\lambda}{\lambda + \tilde{\pi}_T}. \tag{6}$$

[4]Note that a fresh access request would not enter State H if its transmission is successful because the data queue has only one data packet (i.e., the data queue was cleared when the previous access request was successfully transmitted), which can be cleared within one time slot.

B. Steady-State Points

The analysis in Section III-A indicates that the steady-state performance of the network is crucially determined by p, the limiting probability of successful transmission of access requests given that no access request is in State H. In this subsection, the network steady-state points will be characterized based on the fixed-point equation of p.

Specifically, for a given access request, its transmission is successful if and only if all the other $n-1$ devices have empty request queues or have non-empty request queues but without requesting any transmission, given that no access request is in State H. Accordingly, we have

$$\begin{aligned} p &= \big(\Pr\{\text{request queue is empty}|\text{no access request is in State H}\} + \Pr\{\text{request queue is non-empty but not transmitting}|\text{no access request is in State H}\}\big)^{n-1} \\ &= \left(\frac{1-\rho}{1-\rho\tilde{\pi}_H} + \frac{\rho(\tilde{\pi}_T + \tilde{\pi}_0)(1-q)}{1-\rho\tilde{\pi}_H}\right)^{n-1}. \end{aligned} \tag{7}$$

By combining (1)–(4), (6) and applying $n - 1 \approx n$, $(1-x)^n \approx \exp\{-nx\}$ for $0 < x < 1$ if n is large, (7) can be approximated by

$$p \overset{\text{with a large } n}{\approx} \exp\left(-\frac{nq}{1 + \frac{p\alpha q}{\lambda}}\right). \tag{8}$$

Appendix B shows that the steady-state probability that no access request is in State H, α, is given by

$$\alpha = 1 + \frac{\lambda \ln p}{\beta q}. \tag{9}$$

By substituting (9) into (8), the fixed-point equation of p can be obtained as

$$p = \exp\left(-\frac{\hat{\lambda} q}{\frac{\hat{\lambda}}{n} + pq\left(1 + \frac{\hat{\lambda}\ln p}{nq\beta}\right)}\right), \tag{10}$$

where $\hat{\lambda} = n\lambda$ denotes the aggregate input rate. It can be seen that with $\beta = +\infty$, (10) reduces to $p = \exp\left(-\frac{\hat{\lambda} q}{\frac{\hat{\lambda}}{n} + pq}\right)$, which is consistent with Eq. (6) in [20] with the UB window size $W = 1$. Theorem 1 shows that (10) has either one or three non-zero roots.

Theorem 1: The fixed-point equation (10) of $p \in (0, 1]$ has either three non-zero roots $0 < p_A \le p_S \le p_L \le 1$ or one non-zero root $0 < p_L \le 1$.

Proof: See Appendix C. ∎

Note that not all the roots of (10) are steady-state points. We follow the approximate trajectory analysis proposed in [35], and find that:

1) If (10) has only one non-zero root p_L, then p_L is a steady-state point;
2) If (10) has three non-zero roots $p_A \le p_S \le p_L$, then only p_L and p_A are steady-state points. We refer to p_L as the desired steady-state point and p_A as the undesired steady-state point.

It could be shown from (10) that the steady-state points p_L and p_A are both monotonic decreasing functions of traffic input rate of each MTD λ, the number of MTDs n, and the

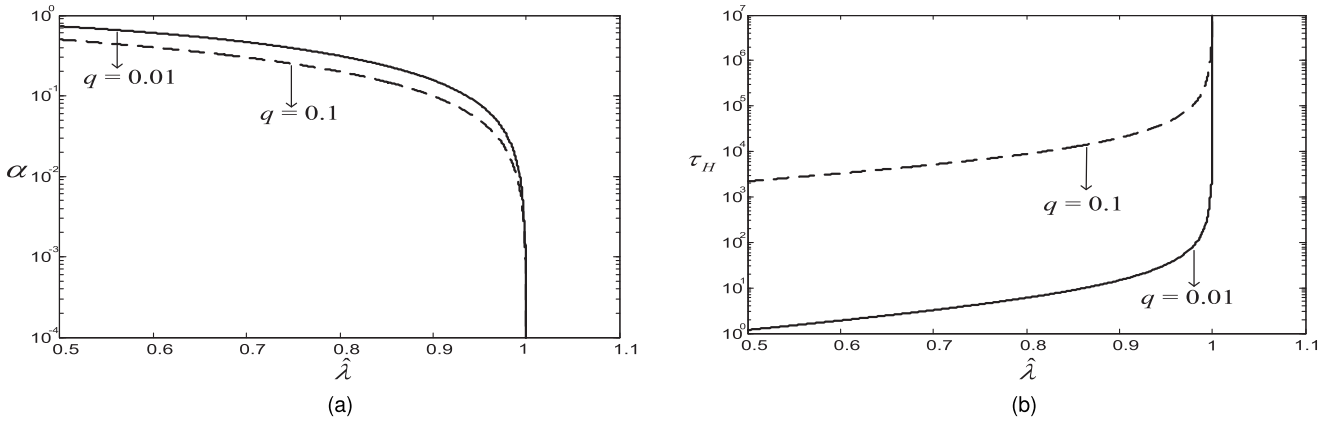


Fig. 4. Limiting probability that no access request is in State H, α, and mean holding time in state H, τ_H, versus the aggregate input rate $\hat{\lambda}$. $n = 100$. $M = 1$. $q \in \{0.01, 0.1\}$. $\beta = 1$. (a) α versus $\hat{\lambda}$. (b) τ_H versus $\hat{\lambda}$.

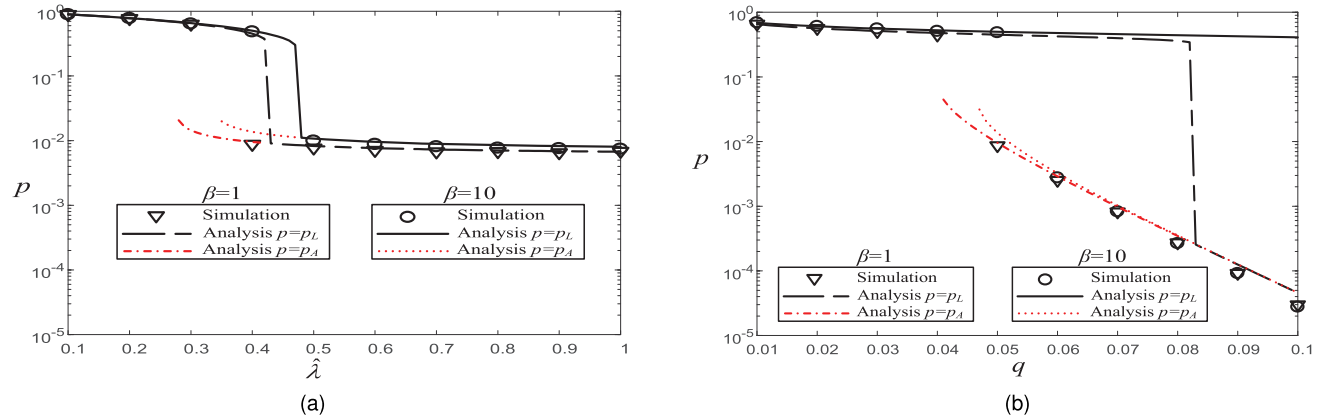


Fig. 5. Limiting probability of successful transmission of access requests given that no access request is in State H, p, versus the aggregate input rate $\hat{\lambda}$ and the ACB factor q. $\beta \in \{1, 10\}$. $n = 100$. $M = 1$. (a) $q = 0.05$. (b) $\hat{\lambda} = 0.4$.

ACB factor q, and monotonic increasing functions of the data transmission rate β.

C. Stability

By combining (9) and (10), we can see that the limiting probability that no access request is in State H, α, is also a monotonic decreasing function of the input rate of each MTD λ. As λ increases, α would eventually drop to zero, with which the mean holding time in State H, τ_H, becomes infinite according to (3), and the Markov chain in Fig. 3 becomes non-recurrent. In this case, one MTD would occupy the channel for data transmission for an unlimited amount of time. Meanwhile, as other MTDs' access requests cannot succeed, the queue length of their data queues would become infinite, indicating that the network has become unstable.

In this paper, we define that the network is stable if and only if the Markov chain in Fig. 3 is recurrent. It can be seen from (9) and (10) that $\alpha = 0$ when the aggregate input rate $\hat{\lambda}$ is equal to the data transmission rate β. As Fig. 4 illustrates, when the aggregate input rate $\hat{\lambda}$ grows, the limiting probability that no access request is in State H, α, decreases and the mean holding time in State H τ_H increases. α drops to zero when $\hat{\lambda} \geq \beta$, with which $\tau_H = +\infty$, and the network becomes unstable.[5] Intuitively, when the data transmission rate β is smaller than the aggregate input rate $\hat{\lambda}$, the data queues can never be cleared.

D. Simulation Results

The above analysis is verified by the simulation results presented in Fig. 5. In this paper, event-driven simulations are conducted and each simulation is carried out for 10^7 time slots. The simulation setting is the same as the system model described in Section II, and we omit the details here due to limited space. In simulations, we count the total number of transmitted access requests from all MTDs and the total number of successful access requests when no MTD is in the data transmission state. The limiting probability of successful transmission of access requests given that no access request is in State H, p, is then obtained by calculating the ratio of the number of successful access requests to the total number of transmitted access requests.

Specifically, the analysis has shown that the network can have either one steady-state point p_L or two steady-state points, i.e., the desired steady-state point p_L and the undesired

[5]Note that in [20], the network is always stable because the data transmission rate $\beta = +\infty$.

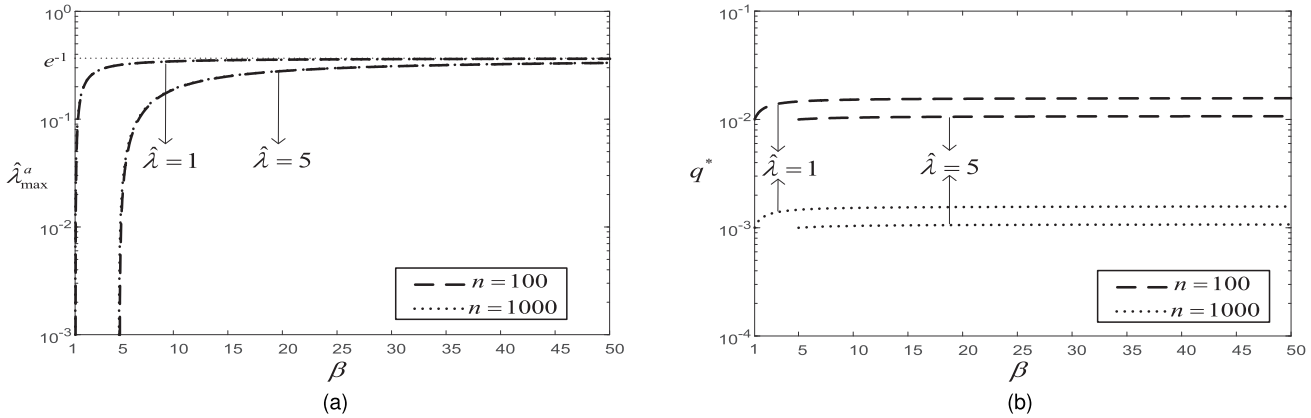


Fig. 6. Maximum access throughput $\hat{\lambda}^a_{\max}$ and optimal ACB factor q^* versus the transmission rate β. $\hat{\lambda} \in \{1, 5\}$. $n \in \{100, 1000\}$. $M = 1$. (a) $\hat{\lambda}^a_{\max}$ versus β. (b) q^* versus β.

steady-state point p_A, which are non-zero roots of the fixed-point equation (10) of p. Fig. 5 presents how the steady-state points p_L and p_A vary with the aggregate input rate $\hat{\lambda}$ and the ACB factor q when the data transmission rate β is 1 or 10. It can be seen that when $\hat{\lambda}$ or q is small, the network has only one steady-state point p_L and operates at p_L. As $\hat{\lambda}$ or q increases, the network will have two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A. It may first operate at the desired steady-state point p_L and then drop to the undesired steady-state point p_A. Both steady-state points decrease as the ACB factor q increases, and are slightly improved when the data transmission rate β increases. Simulation results presented in Fig. 5 well agree with the analysis.

IV. Access Throughput and Data Throughput

Based on the steady-state point analysis, in this section, we will focus on the throughput performance. In particular, we will derive the access throughput $\hat{\lambda}^a_{out}$ and the data throughput $\hat{\lambda}^d_{out}$, and study how to optimally choose the system parameters to maximize $\hat{\lambda}^a_{out}$ and $\hat{\lambda}^d_{out}$, respectively.

A. Access Throughput

In this paper, the access throughput $\hat{\lambda}^a_{out}$ is defined as the average number of successful access requests per time slot. Based on the $Geo/G/1/1$ model of each access request queue, the access throughput $\hat{\lambda}^a_{out}$ can be obtained as

$$\hat{\lambda}^a_{out} = \hat{\lambda}(1-\rho), \tag{11}$$

where ρ denotes the probability that each request queue is nonempty. By combining (1)–(4), (6), (9) and (11), we have

$$\begin{aligned}\hat{\lambda}^a_{out} &= \frac{np\alpha q}{1+\frac{\hat{\lambda}(1-p\alpha q)}{\beta}+\frac{p\alpha q}{\hat{\lambda}}} \\ &= \frac{np\left(q+\frac{\hat{\lambda}\ln p}{n\beta}\right)}{1+\frac{\hat{\lambda}}{n\beta}\left(1-pq-\frac{\hat{\lambda}p\ln p}{n\beta}\right)+\frac{npq}{\hat{\lambda}}+\frac{p\ln p}{\beta}}.\end{aligned} \tag{12}$$

Note that when the data transmission rate $\beta = +\infty$, (12) reduces to $\lim_{\beta\to+\infty} \hat{\lambda}^a_{out} = \frac{n\hat{\lambda}qp}{\hat{\lambda}+npq}$, which is consistent with Eq. (4) in [20] with the UB window size $W = 1$. Moreover, when the aggregate input rate $\hat{\lambda} \geq \beta$, the network will be unstable, in which case $\hat{\lambda}^a_{out} = 0$ because the limiting probability that no access request is in State H, α, is 0. Typically, the number of MTDs n, the aggregate input rate $\hat{\lambda}$ and the data transmission rate β are system input parameters. Therefore, we are interested in how to optimally tune the ACB factor q to maximize $\hat{\lambda}^a_{out}$ for given n, $\hat{\lambda}$ and β.

Define the maximum access throughput as $\hat{\lambda}^a_{\max} = \max_q \hat{\lambda}^a_{out}$. The following theorem presents the maximum access throughput $\hat{\lambda}^a_{\max}$ and the optimal ACB factor q^*.

Theorem 2: The maximum access throughput is given by

$$\hat{\lambda}^a_{\max} = \frac{n\beta(\beta-\hat{\lambda})}{\beta(en\beta-n-1)+\frac{\hat{\lambda}\beta(ne-1)}{n}+\frac{\hat{\lambda}^2}{n}}, \tag{13}$$

which is achieved if and only if the network operates at the desired steady-state point p_L, and

$$q^* = \frac{\hat{\lambda}(\beta-e^{-1})}{n\beta(\hat{\lambda}-e^{-1})}. \tag{14}$$

Proof: See Appendix D. ∎

Fig. 6 illustrates how the maximum access throughput $\hat{\lambda}^a_{\max}$ and the optimal ACB factor q^* vary with the data transmission rate β, the number of MTDs n and the aggregate input rate $\hat{\lambda}$. Specifically, Fig. 6a shows that $\hat{\lambda}^a_{\max}$ is a monotonic increasing function of the data transmission rate β. With $\beta = +\infty$, (13) reduces to $\lim_{\beta\to+\infty} \hat{\lambda}^a_{\max} = e^{-1}$, which is consistent with Theorem 2 in [20]. Moreover, $\hat{\lambda}^a_{\max}$ decreases as the aggregate arrival rate $\hat{\lambda}$ increases, and becomes zero when $\hat{\lambda} = \beta$, in which case the network becomes unstable. As for the optimal ACB factor q^*, it can be seen from Fig. 6b that as the data transmission rate β grows, q^* increases and quickly converges to $\lim_{\beta\to+\infty} q^* = \frac{\hat{\lambda}}{n(\hat{\lambda}-e^{-1})}$, which is consistent with Eq. (10) in [20]. q^* decreases as the aggregate input rate $\hat{\lambda}$ or the number of MTDs n increases.

We can also see from Theorem 2 that when $\frac{n\beta(\hat{\lambda}-e^{-1})}{\hat{\lambda}(\beta-e^{-1})} < 1$, the maximum access throughput $\hat{\lambda}^a_{\max}$ cannot be achieved because (14) does not hold for any ACB

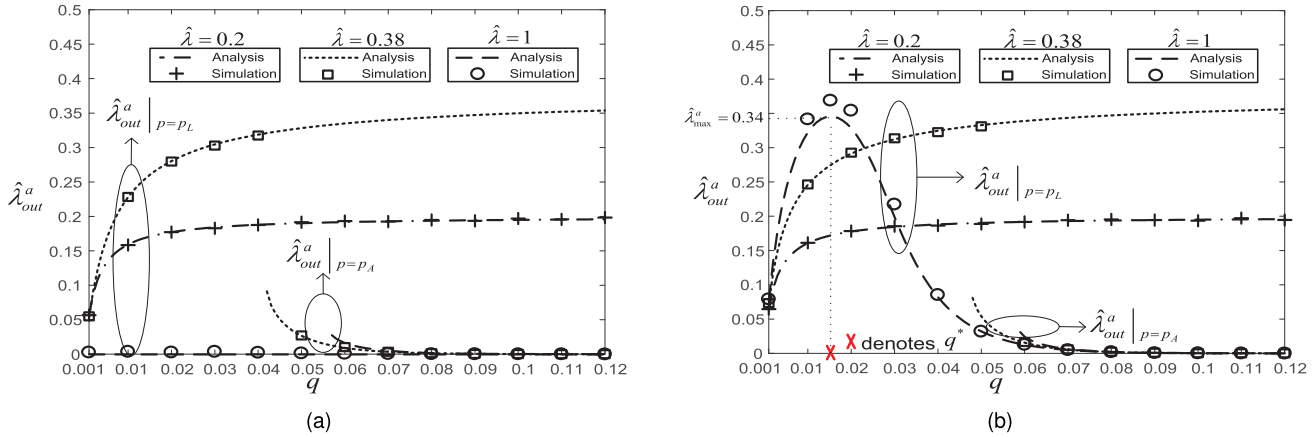


Fig. 7. Access throughput $\hat{\lambda}^a_{out}$ versus the ACB factor q. $\hat{\lambda} \in \{0.2, 0.38, 1\}$. $n = 100$. $M = 1$. (a) $\beta = 1$. (b) $\beta = 10$.

factor $q \in (0, 1]$. Similar to [20], we define the following regions of $(n, \hat{\lambda})$:

- *Unachievable Region:* $\mathcal{S}_{\mathcal{N}} = \left\{(n,\hat{\lambda})|\frac{n\beta(\hat{\lambda}-e^{-1})}{\hat{\lambda}(\beta-e^{-1})} < 1\right\}$, in which $\hat{\lambda}^a_{\max}$ cannot be achieved, because (14) does not hold for any $q \in (0, 1]$.
- *Uncertain Region:* $\mathcal{S}_{\mathcal{U}} = \left\{(n,\hat{\lambda})|\frac{n\beta(\hat{\lambda}-e^{-1})}{\hat{\lambda}(\beta-e^{-1})} \geq 1, \hat{\lambda} < 4e^{-2}\right\}$, in which the network may operate at the desired steady-state point p_L or the undesired steady-state point p_A. $\hat{\lambda}^a_{\max}$ is achieved only if the network operates at p_L.
- *Achievable Region:*

$$\mathcal{S}_{\mathcal{A}} = \left\{(n,\hat{\lambda})|\frac{n\beta(\hat{\lambda}-e^{-1})}{\hat{\lambda}(\beta-e^{-1})} \geq 1, 4e^{-2} \leq \hat{\lambda} < \beta\right\} \overset{\text{with a large } n}{\approx} \left\{(n,\hat{\lambda})|4e^{-2} \leq \hat{\lambda} < \beta\right\},$$

  in which the network is guaranteed to operate at the desired steady-state point p_L, and $\hat{\lambda}^a_{\max}$ can be achieved when the ACB factor q is tuned according to (14).
- *Unstable Region:* $\mathcal{S}_{\mathcal{S}} = \left\{(n,\hat{\lambda})|\hat{\lambda} \geq \beta\right\}$, in which the network is unstable and the access throughput $\hat{\lambda}^a_{out} = 0$.

Note that when the data transmission rate $\beta = +\infty$, the unstable region $\mathcal{S}_{\mathcal{S}}$ vanishes, and the remaining three regions reduce to the counterparts defined in [20].

B. Data Throughput

Let us now consider the data throughput $\hat{\lambda}^d_{out}$, which is defined as the average number of transmitted data packets per time slot. When the network is stable, i.e., $\hat{\lambda} < \beta$, the data throughput is equal to the aggregate input rate $\hat{\lambda}$, i.e.,

$$\hat{\lambda}^d_{out}|_{\hat{\lambda}<\beta} = \hat{\lambda}. \tag{15}$$

On the other hand, when the network is unstable, i.e., $\hat{\lambda} \geq \beta$, it has been shown in Section III-C that in this case, one MTD would capture the channel and transmit its data packets with rate β, while other MTDs cannot access. As a result, the data throughput is given by

$$\hat{\lambda}^d_{out}|_{\hat{\lambda}\geq\beta} = \beta. \tag{16}$$

We can see from (15) and (16) that when $\hat{\lambda} < \beta$, the data throughput $\hat{\lambda}^d_{out}|_{\hat{\lambda}<\beta}$ grows as the aggregate input rate $\hat{\lambda}$ increases, and approaches β as $\hat{\lambda} \to \beta$; When $\hat{\lambda} \geq \beta$, the network achieves its maximum data throughput

$$\hat{\lambda}^d_{\max} = \beta. \tag{17}$$

It can be seen from (12) and (17) that when the maximum data throughput $\hat{\lambda}^d_{\max}$ is achieved, the network is unstable and the access throughput $\hat{\lambda}^a_{out} = 0$. On the other hand, to achieve the maximum access throughput $\hat{\lambda}^a_{\max}$, the network should operate at the achievable region $\mathcal{S}_{\mathcal{A}}$, where the data throughput $\hat{\lambda}^d_{out} = \hat{\lambda} < \hat{\lambda}^d_{\max} = \beta$. We can conclude that the maximum access throughput $\hat{\lambda}^a_{\max}$ and the maximum data throughput $\hat{\lambda}^d_{\max}$ cannot be achieved simultaneously.

C. Simulation Results

The above analysis is verified by the simulation results presented in Figs. 7–8. In simulations, we count the total number of successful access requests and the number of transmitted data packets in each simulation run, i.e., 10^7 time slots. The access throughput and data throughput are then obtained by calculating the ratios of the number of successful access requests and the number of transmitted data packets to the number of time slots 10^7, respectively.

Specifically, the expression of access throughput $\hat{\lambda}^a_{out}$ has been given in (12), which is determined by the number of MTDs n, the aggregate input rate $\hat{\lambda}$, the data transmission rate β and the ACB factor q. Fig. 7 illustrates how the access throughput $\hat{\lambda}^a_{out}$ varies with the ACB factor q with the data transmission rate $\beta = 1$ or 10 and the number of MTDs $n = 100$. As Fig. 7 shows, when the aggregate input rate $\hat{\lambda} = 0.2 \in (0, 0.37)$, we have $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{N}}$, in which the maximum access throughput $\hat{\lambda}^a_{\max}$ cannot be achieved regardless of what value of q is chosen. On the other hand, with $\hat{\lambda} = 0.38 \in [0.37, 0.54)$, we have $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{U}}$, in which $\hat{\lambda}^a_{\max}$ again cannot be achieved, because the network shifts to the undesired steady-state point p_A as the ACB factor q increases. With $\hat{\lambda} = 1$, if the data transmission rate $\beta = 1$, as shown in Fig. 7a, then we have $(n, \hat{\lambda}) \in \mathcal{S}_{\mathcal{S}}$, in which the network is unstable and the access throughput $\hat{\lambda}^a_{out} = 0$.

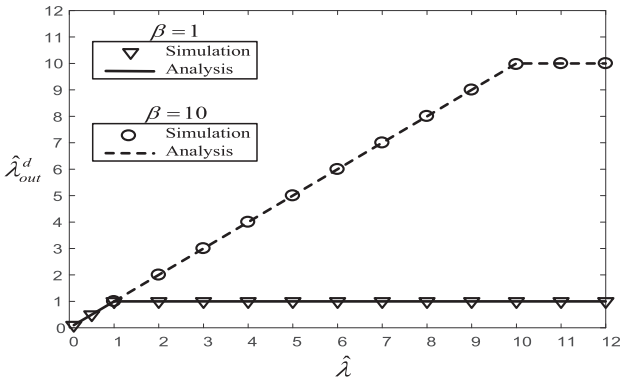


Fig. 8. Data throughput $\hat{\lambda}^d_{out}$ versus the aggregate input rate $\hat{\lambda}$. $n = 100$. $M = 1$. $q = 0.001$. $\beta \in \{1, 10\}$.

If $\beta = 10$, then $(n, \hat{\lambda}) \in \mathcal{S}_A$, where the maximum access throughput $\hat{\lambda}^a_{\max}$ can be achieved when the ACB factor q is tuned according to (14), i.e., $q = q^* = 0.015$, as shown in Fig. 7b. Note that a slight deviation between the analysis and simulation results is observed in this case because of the approximation introduced in (3).

As for the data throughput $\hat{\lambda}^d_{out}$, the analysis has shown that $\hat{\lambda}^d_{out} = \hat{\lambda}$ for $\hat{\lambda} < \beta$ and $\hat{\lambda}^d_{out} = \beta$ for $\hat{\lambda} \geq \beta$. Simulation results presented in Fig. 8 verify that $\hat{\lambda}^d_{out}$ linearly increases with the aggregate input rate $\hat{\lambda}$ when $\hat{\lambda}$ is below the data transmission rate β, and reaches the maximum data throughput $\hat{\lambda}^d_{\max}$ when $\hat{\lambda} \geq \beta$. In this case, however, $(n, \hat{\lambda}) \in \mathcal{S}_S$ and the access throughput $\hat{\lambda}^a_{out} = 0$, as shown in Fig. 7, indicating that the maximization of data throughput is achieved at the cost of sacrificing the access throughput.

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

Sections III and IV are based on the assumption that the number of preambles $M = 1$. In this section, we extend the above analysis to the multi-preamble scenario $M > 1$ by applying the multi-group model proposed in [20]. Specifically, by virtue of orthogonality among preambles, MTDs that use different preambles do not affect each other's chance of successful access. Therefore, we can divide them into M groups according to the preamble that each MTD chooses. The group parameters can be 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 $\hat{\lambda}^{(i)} = n^{(i)}\lambda$.
- $\beta^{(i)}$ denotes the transmission rate in Group i.

By replacing $n, \hat{\lambda}, \beta$ in (10), (12) and (15)–(16) with $n^{(i)}, \hat{\lambda}^{(i)}, \beta^{(i)}$, the steady-state points of Group i, i.e., $p_L^{(i)}$ and $p_A^{(i)}$, the group access throughput $\hat{\lambda}^{(i),a}_{out}$, and the group data throughput $\hat{\lambda}^{(i),d}_{out}$ can be obtained, respectively.

As each MTD independently and randomly selects a preamble in each access attempt [5], when the total number of MTDs n is large, $n^{(i)}$ can be approximated by $n^{(i)} \approx \frac{n}{M}$. Moreover, for fairness, we assume that all the groups have the same data transmission rate, that is, $\beta^{(i)} = \frac{\beta}{M}$. Similar to (15)–(16), the data throughput is still given by

$$\hat{\lambda}^{M,d}_{out} = \begin{cases} \hat{\lambda} & \text{if } \hat{\lambda} < \beta, \\ \beta & \text{otherwise,} \end{cases} \tag{18}$$

with the maximum data throughput $\hat{\lambda}^{M,d}_{\max} = \beta$ when $\hat{\lambda} \geq \beta$.

On the other hand, the access throughput can be obtained by replacing n, $\hat{\lambda}$ and β with $n^{(i)} \approx \frac{n}{M}, \hat{\lambda}^{(i)} \approx \frac{n\lambda}{M}$ and $\beta^{(i)} = \frac{\beta}{M}$ in (12), respectively, as

$$\hat{\lambda}^{M,a}_{out} = \frac{n}{M} \sum_{i=1}^{M} \frac{p^{(i)}\left(q + \frac{\lambda M \ln p^{(i)}}{\beta}\right)}{1 + \frac{\lambda M}{\beta}\left(1 - p^{(i)}q - \frac{\lambda M p^{(i)} \ln p^{(i)}}{\beta}\right) + \frac{p^{(i)}q}{\lambda} + \frac{M p^{(i)} \ln p^{(i)}}{\beta}}, \tag{19}$$

which is maximized at

$$\hat{\lambda}^{M,a}_{\max} = \frac{n\beta(\beta - n\lambda)}{\beta\left(\frac{en\beta}{M} - n - M\right) + \lambda\beta ne - M\lambda(\beta - n\lambda)}, \tag{20}$$

with the optimal ACB factor

$$q^{*,M} = \frac{\lambda M(\beta - Me^{-1})}{\beta(n\lambda - Me^{-1})}. \tag{21}$$

When the data transmission rate $\beta = +\infty$, we can obtain from (20)–(21) that $\lim_{\beta \to +\infty} \hat{\lambda}^{M,a}_{\max} = Me^{-1}$ and $\lim_{\beta \to +\infty} q^{*,M} = \frac{\lambda}{\frac{n\lambda}{M} - e^{-1}}$, which are consistent with Eq. (12) and Eq. (15) of [20], respectively. We can also see from (20) that for large number of MTDs n, $\hat{\lambda}^{M,a}_{\max} \approx \frac{M(\beta - n\lambda)}{e\beta - 1}$, indicating the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ decreases as n increases, and drops to zero when $\beta = n\lambda$, in which case the network becomes unstable. If the data transmission rate β is large, i.e., $\beta \gg n\lambda$, then $\hat{\lambda}^{M,a}_{\max}$ will approach Me^{-1} and become insensitive to the number of MTDs n.

As we can see from Fig. 9, both the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ and the corresponding optimal ACB factor $q^{*,M}$ are monotonic increasing functions of the data transmission rate β and the number of preambles M. Intuitively, with more orthogonal preambles, more MTDs can successfully access the network, and thus the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ is increased. The increasing rate, however, depends on the data transmission rate β. When β is small, a superlinear increase of $\hat{\lambda}^{M,a}_{\max}$ with M can be observed. As β increases, the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ tends to linearly increase with M.

Simulation results are presented in Fig. 10 to validate the above analysis. In simulations, each MTD independently and randomly selects one out of M preambles in each access attempt. If the selected preamble is used by another MTD in the data transmission state, then the MTDs will perform the random preamble selection again in the next time slot.[6] An MTD with a successful access request clears its data queue with the data transmission rate $\frac{\beta}{M}$, and at most M MTDs can be accommodated for concurrent data transmissions.

[6]In practice, the BS knows which MTDs are in the data transmission state, and what preambles are used. If an MTD chooses a preamble that is used by another MTD in the data transmission state, then the BS would consider the request failed and not acknowledge it. The MTD then performs random preamble selection again.

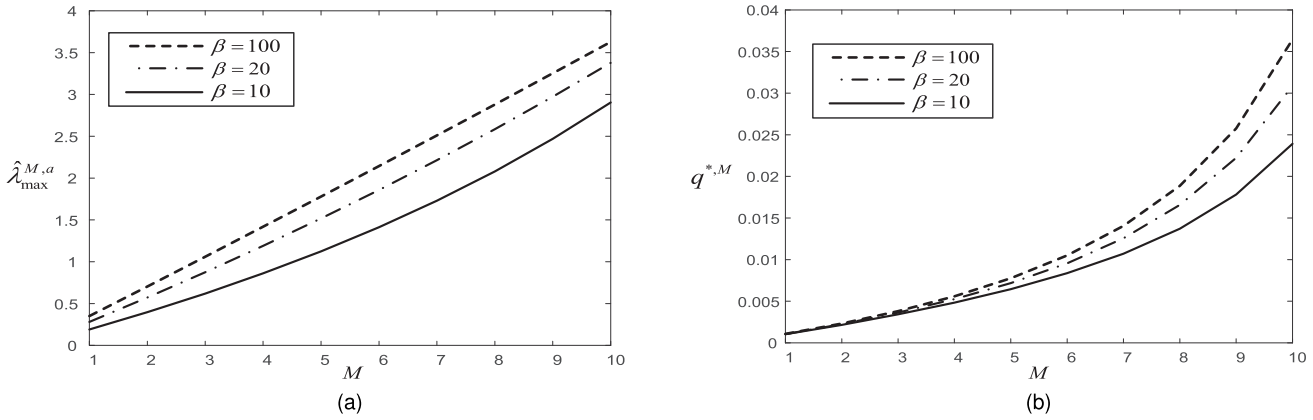


Fig. 9. Maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ and optimal ACB factor $q^{*,M}$ versus the number of preambles M. $n = 1000$. $\lambda = 0.005$. $\beta \in \{10, 20, 100\}$. (a) $\hat{\lambda}^{M,a}_{\max}$ versus M. (b) $q^{*,M}$ versus M.

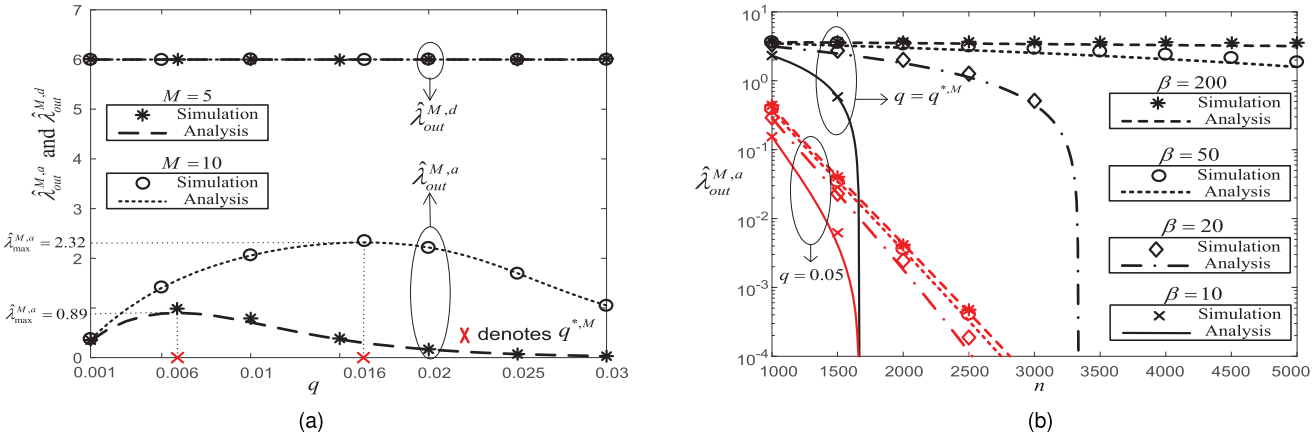


Fig. 10. (a) Access throughput $\hat{\lambda}^{M,a}_{out}$ and data throughput $\hat{\lambda}^{M,d}_{out}$ versus the ACB factor q. $n = 1000$. $\lambda = 0.006$. $\beta = 10$. $M \in \{5, 10\}$. (b) Access throughput $\hat{\lambda}^{M,a}_{out}$ versus the number of MTDs n. $q = 0.05$ or $q^{*,M}$. $\lambda = 0.006$. $\beta \in \{10, 20, 50, 200\}$. $M = 10$.

Specifically, Fig. 10a illustrates how the access throughput $\hat{\lambda}^{M,a}_{out}$ and the data throughput $\hat{\lambda}^{M,d}_{out}$ vary with the ACB factor q when multiple preambles are adopted, i.e., $M = 5$ or 10. It can be clearly seen that the access throughput is quite sensitive to the value of ACB factor q. To achieve the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$, q should be carefully tuned based on the number of preambles M, the number of MTDs n and the input rate of each MTD λ according to (21), i.e., $q = q^{*,M}$. Moreover, in contrast to the case of infinite data transmission rate $\beta = +\infty$ where $\hat{\lambda}^{M,a}_{\max}$ increases linearly with the number of preambles M, with $\beta = 10$, $\hat{\lambda}^{M,a}_{\max}$ is nearly tripled when M is doubled, indicating that the improvement in the maximum access throughput brought by increasing the number of preambles M is more significant when the data transmission rate is small. The data throughput $\hat{\lambda}^{M,d}_{out}$, on the other hand, is independent of the ACB factor q, and equal to the aggregate input rate $\hat{\lambda}$ as long as $\hat{\lambda} < \beta$.

Fig. 10b further illustrates how the access throughput $\hat{\lambda}^{M,a}_{out}$ varies with the number of MTDs n under various values of the data transmission rate β. It can be seen that even with the ACB factor optimally tuned, i.e., $q = q^{*,M}$, the access throughput performance may still deteriorate when the network size n increases if the data transmission rate β is small. The maximum access throughput becomes insensitive to the number of MTDs n only when β is sufficiently large, i.e., $\beta \gg n\lambda$. On the other hand, if the ACB factor q is fixed, e.g., $q = 0.05$, then the access throughput $\hat{\lambda}^{M,a}_{out}$ quickly drops as the number of MTDs n increases no matter how large the data transmission rate β is. Compared to the maximum access throughput, the throughput loss is significant especially when the data transmission rate β and the number of MTDs n are both large.

Note from Fig. 10b that for given data transmission rate β, the access throughput drops to zero when the aggregate input rate exceeds β, in which case the network becomes unstable. Intuitively, to avoid being unstable, the time slot length should be enlarged to allocate more resources for data transmission, which, however, leads to fewer chances for access. In the next section, we will further study how to properly choose the time slot length to optimize the access efficiency.

VI. Optimal Time Slot Length

Recall that the access throughput $\hat{\lambda}^{M,a}_{out}$ is defined as the average number of successful access requests per *time slot*. In practice, however, the access throughput normalized by the time slot length, that is, the average number of successful

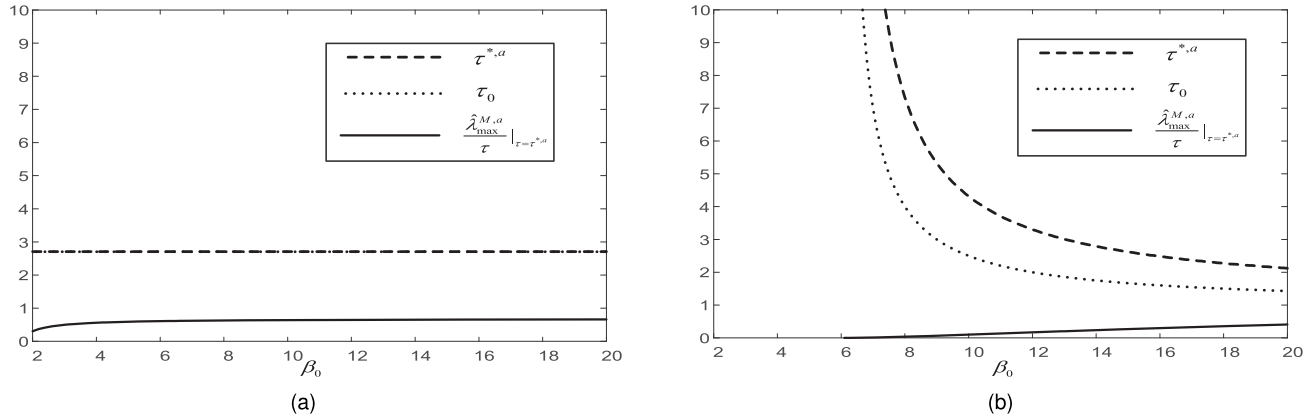


Fig. 11. Optimal time slot length $\tau^{*,a}$ and the corresponding normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}|_{\tau=\tau^{*,a}}$ versus β_0. $n = 1000$. $M = 5$. (a) $\lambda_0 = 0.001$. (b) $\lambda_0 = 0.006$.

access requests per *millisecond*, could be of more interest. In the LTE standard, the length of one time slot, τ (in unit of subframes[7]), is indeed a system parameter that can be adaptively tuned. As Fig. 1 shows, the time slot length τ determines how the system resources are allocated between access and data transmission. With a smaller τ, MTDs can access the channel more frequently, but the data transmission rate would be lower as there are fewer subframes for data transmission per time slot. In this section, we will focus on the optimal tuning of the time slot length τ for maximizing the normalized access throughput.[8]

A. Normalized Maximum Access Throughput

Specifically, as the BS dynamically allocates time-frequency resources every 1 millisecond, i.e., every subframe, in LTE networks [36], the total number of data packets that can be transmitted in each subframe, which is determined by the amount of resources and scheduling algorithm, is a given system parameter and denoted by β_0. With $\tau-1$ subframes for data transmission within each time slot, the data transmission rate β can be written as

$$\beta = \beta_0(\tau - 1). \tag{22}$$

Similarly, denote the probability that an MTD generates a new data packet in one subframe as λ_0. The input rate of each MTD λ can then be written as

$$\lambda = \lambda_0 \tau. \tag{23}$$

By combining (20) and (22)–(23), the normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}$ can be obtained as

$$\frac{\hat{\lambda}^{M,a}_{\max}}{\tau} = \frac{\beta_0 - n\lambda_0 - \frac{\beta_0}{\tau}}{\frac{e\beta_0(\tau-1)}{M} - 1 - \frac{M}{n} + \lambda_0\tau\left(e - \frac{M}{n}\right) + \frac{M\lambda_0^2\tau^2}{\beta_0(\tau-1)}}. \tag{24}$$

Here we are interested in how to optimize the normalized maximum access throughput by properly choosing the time slot length τ : $\max_{\tau>1} \frac{\hat{\lambda}^{M,a}_{\max}}{\tau}$. Note that it has been shown in Section IV-A that to guarantee the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ is achievable, the number of MTDs n and the aggregate input rate $\hat{\lambda}$ should fall into the achievable region $\mathcal{S}_{\mathcal{A}}$. By replacing $\hat{\lambda}$ and β in $\mathcal{S}_{\mathcal{A}}$ with $\frac{n\lambda}{M}$ and $\frac{\beta}{M}$, and further combining (22)–(23), we can obtain the following constraints on the time slot length τ:

$$\tau \geq \frac{4e^{-2}M}{n\lambda_0} \text{ and } \tau > \frac{\beta_0}{\beta_0 - n\lambda_0}. \tag{25}$$

Let $\tau^{*,a}$ denote the optimal time slot length to maximize the normalized maximum access throughput, which can then be written as

$$\tau^{*,a} = \arg\max_{\tau>\tau_0} \frac{\hat{\lambda}^{M,a}_{\max}}{\tau}, \tag{26}$$

where $\tau_0 = \max\{1, \frac{4e^{-2}M}{n\lambda_0}, \frac{\beta_0}{\beta_0 - n\lambda_0}\}$ denotes the minimum requirement on the time slot length τ according to (25). Fig. 11 presents the optimal time slot length $\tau^{*,a}$ and the corresponding normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}|_{\tau=\tau^{*,a}}$ with $\lambda_0 = 0.001$ or 0.006.

Specifically, we can clearly see from Fig. 11a that when the traffic is light, i.e., $n\lambda_0 \ll \beta_0$, the optimal time slot length $\tau^{*,a} \to \tau_0$, indicating that in this case, τ should be tuned as small as possible, approaching the minimum requirement τ_0. Intuitively, with light traffic, the data transmission requires few resources. Therefore, the time slot length τ should be reduced such that MTDs can access more frequently. Moreover, neither the optimal time slot length $\tau^{*,a}$ nor the corresponding normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}|_{\tau=\tau^{*,a}}$ varies with the number of data packets that can be transmitted per subframe β_0 because data queues can always be cleared within two subframes.

In sharp contrast, with heavy traffic, e.g., $\lambda_0 = 0.006$ as Fig. 11b shows, the optimal time slot length $\tau^{*,a}$ becomes much larger than the minimum requirement τ_0, because more resources need to be allocated to data transmission. As β_0 increases, data queues can be cleared within shorter time. The time slot length can then be reduced to allow for more frequent access for MTDs, and thus the corresponding normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}|_{\tau=\tau^{*,a}}$ is significantly improved.

[7]According to the LTE standard [6], the length of one subframe is fixed to be one millisecond.

[8]Note that it has been shown in Sections IV and V that the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$ and the maximum data throughput $\hat{\lambda}^{M,d}_{\max}$ cannot be achieved at the same time. In this section, we only focus on the maximum access throughput $\hat{\lambda}^{M,a}_{\max}$.

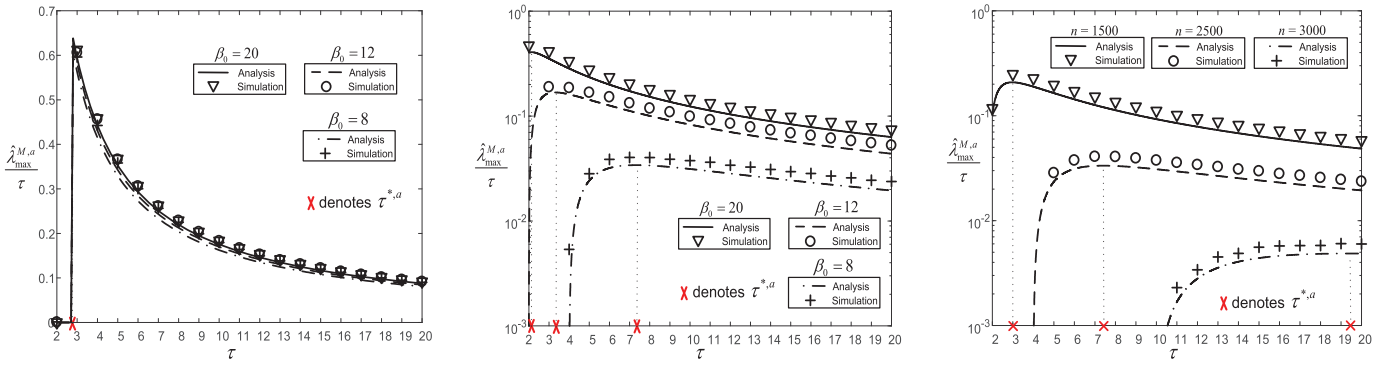


Fig. 12. Normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}$ versus time slot length τ. $M = 5$. $q = q^{*,M}$. (a) $n = 1000$. $\lambda_0 = 0.001$. $\beta_0 \in \{8, 12, 20\}$. (b) $n = 1000$. $\lambda_0 = 0.006$. $\beta_0 \in \{8, 12, 20\}$. (c) $n \in \{1500, 2500, 3000\}$. $\lambda_0 = 0.006$. $\beta_0 = 20$.

B. Simulation Results

Simulation results are presented in Fig. 12 to validate the preceding analysis. The simulation setting is the same as that in Section V, except that each MTD generates a packet with probability λ_0 in each subframe, rather than with probability λ in each time slot.

Specifically, Fig. 12 illustrates how the normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}$ varies with the time slot length τ under various values of the data transmission rate per subframe β_0 and the number of MTDs n. We can see from Fig. 12 that the normalized maximum access throughput $\frac{\hat{\lambda}^{M,a}_{\max}}{\tau}$ crucially depends on the setting of τ, indicating that to optimize the throughput performance, the time slot length τ should be carefully selected. The values of the optimal time slot length $\tau^{*,a}$ have been shown in Fig. 11, and are verified by simulation results presented in Fig. 12. It is worth noting that in LTE networks, the time slot length τ (in unit of subframes) is an integer. Therefore, when choosing the optimal time slot length, $\tau^{*,a}$ should be rounded to the nearest integer properly.

In the current standard, one representative value of the time slot length is $\tau = 5$ subframes[9] [37]. It can be seen from Fig. 12 that with $\tau = 5$, the network may suffer from severe throughput degradation. Specifically, when the traffic is light, i.e., $n\lambda_0 \ll \beta_0$, we can observe from Fig. 12a that the normalized maximum access throughput with $\tau = 5$ is only half of that with the optimal time slot length $\tau^{*,a} \approx 3$. As the traffic input rate of each MTD per subframe λ_0 or the network size n increases, the time slot length should be properly enlarged to allocate more resources to data transmission. If the time slot length τ is still fixed at 5, then the corresponding normalized maximum access throughput would be far below that with the optimal time slot length $\tau^{*,a}$. For instance, with $n = 3000$, $\lambda_0 = 0.006$ and $\beta_0 = 20$ as shown in Fig. 12c, the network becomes unstable with the maximum access throughput dropping to zero when $\tau = 5$. We can conclude that by optimally choosing the time slot length based on the traffic conditions and the data transmission rate, significant gains in the normalized access throughput can be achieved over the default setting.

[9]More specifically, in [37], PRACH configuration index=6 is used as one representative value, which, according to the PRACH configuration index table [6], specifies that within 10 subframes for each frame, the subframes with index 1 and 6 are PRACH subframes. In this case, PRACH subframes appear every 5 subframes, indicating that the time slot length $\tau = 5$ subframes.

VII. Conclusion

In this paper, the analytical framework proposed in [20] is extended to incorporate a finite data transmission rate β for the random access of M2M communications in LTE networks. By introducing a data transmission state into the state transition process of each individual access request, explicit expressions of the maximum access throughput and the optimal ACB factor are obtained. The analysis shows that when the data transmission rate β is small, the maximum access throughput decreases as the number of MTDs grows, and superlinearly increases with the number of preambles, indicating that significant improvements can be brought by allocating more preambles. It is in sharp contrast to the case of infinite data transmission rate where the maximum access throughput is independent of the number of MTDs and linearly increases with the number of preambles.

The analysis also reveals that the maximum access throughput drops to zero when the aggregate input rate exceeds the data transmission rate, in which case the data throughput reaches the maximum, but the network becomes unstable. To efficiently accommodate the massive access of MTDs, the data transmission rate should be sufficiently large, which indicates that the time slot length should be properly increased for allocating more resources to data transmission as the number of MTDs grows. The effect of time slot length on the normalized maximum access throughput is further analyzed, and shown to be crucial. By optimally choosing the time slot length according to the aggregate traffic rate and data transmission rate per subframe, substantial gains are observed over the default setting in various scenarios.

Note that in this paper, we do not consider the service differentiation issue. For M2M communications, some applications may contain critical information and require higher priority. It is therefore of great importance to further extend the analysis to incorporate distinct quality-of-service requirements of MTDs. Moreover, in this paper, we assume that the total number of data packets that can be transmitted in each time

slot, β, does not change with time for the tractability of analysis. In practice, the data transmission rate β may not be constant due to the dynamics of the resource scheduling process. How the time fluctuation of the data transmission rate β affects the optimal access performance of MTDs is an interesting topic that deserves much attention in the future study.

Appendix A
Derivation of (3)

To derive the mean holding time in State H, τ_H, let us first denote the holding time in State i as Y_i, where $i \in \{0, \text{T}, \text{H}\}$, the number of packets in the data queue when the access request enters State H as N_H, and the number of new arrival packets when the access request is in State H as $\hat{N}_H$. We can then obtain that

$$Y_H + Y_T = \lceil \tfrac{\hat{N}_H + N_H}{\beta} \rceil, \tag{27}$$

where β is the data transmission rate. Note that $Y_T = 1$ and $\tau_H = E[Y_H]$. According to (27), the mean holding time in State H, τ_H, can therefore be written as

$$\begin{aligned} \tau_H &= E[Y_H] = E\lceil \tfrac{N_H + \hat{N}_H}{\beta} \rceil - 1 \\ &= E\lfloor 1 + \tfrac{N_H + \hat{N}_H - 1}{\beta} \rfloor - 1 \approx \tfrac{E[N_H]-1}{\beta} + \tfrac{E[\hat{N}_H]}{\beta}, \end{aligned} \tag{28}$$

by applying $\lceil \frac{x}{y} \rceil = \lfloor 1 + \frac{x-1}{y} \rfloor$ for $x, y \in \{1, 2, \ldots\}$, and dropping the rounding down operation for analytical tractability. We can see from (28) that τ_H is determined by the average number of new arrival packets when the access request is in State H, $E[\hat{N}_H]$, and the average number of packets in the data queue when the access request enters State H, $E[N_H]$. In the following, we focus on $E[\hat{N}_H]$ first and then $E[N_H]$.

As the traffic input rate of each MTD is λ and the mean holding time in State H is τ_H, we can easily obtain the average number of new arrival packets when the access request is in State H as

$$E[\hat{N}_H] = \lambda \tau_H. \tag{29}$$

On the other hand, to derive the average number of packets in the data queue when the access request enters State H, $E[N_H]$, let us first define D_i as the time spent from the beginning of State i until the service completion. According to Fig. 3, we can see that before the access request enters State H, it should first be in State T, in which case the data queue has one data packet, and then in State 0, in which case the average number of new arrival packets when the access request is in State 0 is $\lambda E[D_0 - D_H]$. Therefore, we have

$$E[N_H] = 1 + \lambda E[D_0 - D_H]. \tag{30}$$

According to the Markov chain in Fig. 3, it can be obtained that

$$D_0 = \begin{cases} Y_0 + D_H & \text{with probability } p\alpha q, \\ Y_0 + D_0 & \text{with probability } 1 - p\alpha q. \end{cases} \tag{31}$$

Let $G_{D_i}(z)$ denote the probability generating functions of D_i, where $i \in \{0, \text{T}, \text{H}\}$. Based on (31) and $Y_0 = 1$, we can have

$$G_{D_0}(z) = \tfrac{p\alpha q z G_{D_H}(z)}{1-(1-p\alpha q)z}, \tag{32}$$

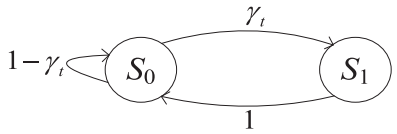


Fig. 13. Embedded Markov chain $\{X_j^H\}$.

and furthermore,

$$E[D_0 - D_H] = \left(\tfrac{G_{D_0}(z)}{G_{D_H}(z)} \right)' |_{z=1} = \tfrac{1}{p\alpha q}. \tag{33}$$

Finally, by further combining (28)–(33), we can obtain the mean holding time in State H, τ_H, as

$$\tau_H = \frac{\lambda}{\beta p \alpha q \left(1 - \frac{\lambda}{\beta}\right)} \approx \tfrac{\lambda}{\beta p \alpha q}, \tag{34}$$

by ignoring $\frac{\lambda}{\beta}$ in the denominator because $\frac{\lambda}{\beta} < \frac{1}{n} \ll 1$ for large n.

Appendix B
Derivation of (9)

To derive the limiting probability that no access request is in State H, α, let us first define a discrete-time Markov renewal process $(\mathbf{X}^H, \mathbf{V}^H) = \{(X_j^H, V_j^H), j = 0, 1, \ldots\}$, where X_j^H denotes the number of access requests in State H at the j-th transition and V_j^H denotes the epoch at which the j-th transition occurs. As there is at most one access request in State H in any time slot, X_j^H only has two states, i.e., State S_0 with $X_j^H = 0$ or State S_1 with $X_j^H = 1$, and the embedded Markov chain $\mathbf{X}^H = \{X_j^H\}$ is shown in Fig. 13, where γ_t denotes the state transition probability from State S_0 to State S_1 at time slot t.

The steady-state probability distribution of the embedded Markov chain can be derived as

$$\pi_{S_0} = \tfrac{1}{\gamma+1}, \quad \pi_{S_1} = \tfrac{\gamma}{1+\gamma}, \tag{35}$$

where $\gamma = \lim_{t \to \infty} \gamma_t$. Since the mean holding time in State S_0 is 1 and that in State S_1 is τ_H, we can then obtain the steady-state probability distribution of the discrete-time Markov renewal process $(\mathbf{X}^H, \mathbf{V}^H)$ as

$$\tilde{\pi}_{S_0} = \tfrac{1}{1+\tau_H \gamma}, \quad \tilde{\pi}_{S_1} = \tfrac{\tau_H \gamma}{1+\tau_H \gamma}. \tag{36}$$

Note that the limiting probability that no access request is in State H, α, is given by $\tilde{\pi}_{S_0}$, which is determined by γ as (36) shows.

To derive γ, let us first focus on γ_t, which is the probability that given no access request in State H at time slot $t-1$, one access request is in State H at time slot t. We can see from Fig. 3 that when one access request shifts to State H at time slot t, it must stay in State 0 at time slot $t-1$, and is successfully transmitted at time slot t. Let w_t denote the probability that one access request is transmitted at time slot t given no access request in State H at time slot $t-1$. γ_t can then be written as

$$\gamma_t = n w_t (1 - w_t)^{n-1} \cdot \tfrac{\tilde{\pi}_0}{\tilde{\pi}_0 + \tilde{\pi}_T}, \tag{37}$$

where $nw_t(1-w_t)^{n-1}$ represents the probability that an access request is successfully transmitted at time slot t, and $\frac{\tilde{\pi}_0}{\tilde{\pi}_0+\tilde{\pi}_T}$ is the probability that the access request stays in State 0 at time slot $t-1$. Note that the probability of successful transmission of access requests given that no access request is in State H at time slot t, p_t, can be written as

$$p_t = (1-w_t)^{n-1} \approx \exp(-nw_t), \tag{38}$$

by applying $n-1 \approx n, (1-x)^n \approx \exp\{-nx\}$ for $0 < x < 1$ if n is large. By combining (1)–(4), (37) and (38), we can obtain that

$$\gamma = \lim_{t\to+\infty} \gamma_t = -p(1-p\alpha q)\ln p \approx -p\ln p, \tag{39}$$

where $p\alpha q$ is ignored because it is typically very small.

Finally, the limiting probability that no access request is in State H, α, can be obtained by combining (3), (36) and (39).

Appendix C
Proof of Theorem 1

Proof: To determine the number of non-zero roots of the fixed-point equation (10), let us first define

$$f(p) = \frac{h(p)}{1+\frac{pq}{\lambda}\left(1+\frac{\lambda\ln p}{\beta q}\right)}, \tag{40}$$

where

$$h(p) = -\frac{p(\ln p)^2}{\beta} - \frac{(pq+\lambda)\ln p}{\lambda} - nq. \tag{41}$$

As $1+\frac{pq}{\lambda}\left(1+\frac{\lambda\ln p}{\beta q}\right) = 1+\frac{pq\alpha}{\lambda} > 0$, it can be seen from (10), (40) and (41) that $h(p)=0$ has the same non-zero roots as the fixed-point equation (10). Therefore, we focus on $h(p)=0$ in the following.

Since $\lim_{p\to 0} h(p) = +\infty$ and $h(1) = -nq < 0$, $h(p)=0$ has at least one non-zero root for $p \in (0,1]$. To further determine the number of non-zero roots of $h(p)=0$, according to (41), we can obtain that

$$h'(p) = -\frac{(\ln p)^2}{\beta} - \left(\frac{2}{\beta}+\frac{q}{\lambda}\right)\ln p - \frac{pq+\lambda}{p\lambda}, \tag{42}$$

with $\lim_{p\to 0} h'(p) = -\infty$, $h'(1) = -\frac{q+\lambda}{\lambda}$, and

$$h''(p) = -\frac{2\ln p}{p\beta} - \frac{2}{p\beta} - \frac{2q}{p\lambda} + \frac{pq+\lambda}{\lambda p^2}, \tag{43}$$

with $\lim_{p\to 0} h''(p) = +\infty$, $h''(1) = 1 - \frac{2\lambda+q\beta}{\lambda\beta}$. The single non-zero root of $h''(p)=0$ is given by

$$p_E = \frac{\beta}{2\mathbb{W}_0\left(\frac{\beta}{2}\exp\left(1+\frac{q\beta}{2\lambda}\right)\right)}, \tag{44}$$

where $\mathbb{W}_0(\cdot)$ is the principal branch of the Lambert W function. It can be obtained from (43)–(44) that if $2\lambda+q\beta \le \lambda\beta$, then $p_E \ge 1$ and $h''(1) \ge 0$; Otherwise, $0 < p_E < 1$ and $h''(1) < 0$.

In the following, we discuss the number of non-zero roots of $h(p)=0$ based on the monotonicity of $h'(p)$ for $p \in (0,1]$.

- If $2\lambda+q\beta \le \lambda\beta$, then $h''(p) \ge 0$ for $p \in (0,1]$. As a result, $h'(p)$ is a non-decreasing function for $p \in (0,1]$. Since $\lim_{p\to 0} h'(p) < 0$ and $h'(1) < 0$, we have $h'(p) < 0$ for $p \in (0,1]$, indicating that $h(p)$ monotonically decreases for $p \in (0,1]$. We can then conclude that in this case, $h(p)=0$ has only one non-zero root $0 < p_L \le 1$.
- If $2\lambda+q\beta > \lambda\beta$, then $h''(p) > 0$ for $p \in (0, p_E)$ and $h''(p) < 0$ for $p \in (p_E, 1]$. As a result, $h'(p)$ monotonically increases for $p \in (0, p_E)$, and then decreases for $p \in (p_E, 1]$.
  - If $h'(p_E) \le 0$, then we have $h'(p) \le 0$ for $p \in (0,1]$, indicating that $h(p)$ is a non-increasing function for $p \in (0,1]$. We can then conclude that in this case, $h(p)=0$ has only one non-zero root $0 < p_L \le 1$.
  - If $h'(p_E) > 0$, then $h'(p) = 0$ should have two non-zero roots p_1 and p_2, where $0 < p_1 < p_E < p_2 < 1$, and $h'(p) < 0$ for $p \in (0,p_1)\bigcup(p_2,1]$ and $h'(p) > 0$ for $p \in (p_1,p_2)$, indicating that $h(p)$ monotonically decreases for $p \in (0,p_1)\bigcup(p_2,1]$, and increases for $p \in (p_1,p_2)$. If $h(p_1) > 0$ or $h(p_2) < 0$, then $h(p)=0$ has one non-zero root $0 < p_L \le 1$; Otherwise, $h(p)=0$ has three non-zero roots $0 < p_A \le p_S \le p_L \le 1$, in which $h(p_1) \le 0$ and $h(p_2) \ge 0$.

■

Appendix D
Proof of Theorem 2

Proof: To derive the maximum access throughput $\hat{\lambda}^a_{\max}$ and the optimal ACB factor q^*, let us first rewrite the expression of the access throughput $\hat{\lambda}^a_{out}$ as

$$\hat{\lambda}^a_{out} = \frac{-n\lambda\beta}{\lambda(1+\lambda)+ng(p)(\beta-\lambda^2)} + \frac{n^2\lambda\beta}{\frac{\lambda}{g(p)}(1+\lambda)+n(\beta-\lambda^2)}, \tag{45}$$

by combining (8) and (12), where $g(p) = -\frac{q}{\ln p}$. It can be clearly seen from (45) that $\hat{\lambda}^a_{out}$ is maximized when $g(p)$ is maximized. Therefore, we focus on $g(p)$ in the following.

The derivative of $g(p)$ with regard to p can be written as

$$g'(p) = \frac{q}{p(\ln p)^2} - \frac{q'(p)}{\ln p}, \tag{46}$$

where $q'(p)$ is the derivative of q with regard to p. According to (10), we can obtain that

$$q = \frac{-\lambda p(\ln p)^2 - \beta\lambda\ln p}{\beta(n\lambda+p\ln p)}, \tag{47}$$

and

$$q'(p) = -\frac{\lambda\left(\beta n\lambda+p(\ln p)^2(n\lambda+p-\beta)+2n\lambda p\ln p\right)}{\beta p(n\lambda+p\ln p)^2}. \tag{48}$$

By combining (46)–(48), we can then rewrite $g'(p)$ as

$$g'(p) = -\frac{\lambda(\beta-n\lambda)(1+\ln p)}{\beta(n\lambda+p\ln p)^2}. \tag{49}$$

It can be seen from (49) that $g'(p) > 0$ for $p \in (0, e^{-1})$ and $g'(p) < 0$ for $p \in (e^{-1}, 1]$, indicating that $g(p)$ monotonically increases for $p \in (0, e^{-1})$ and decreases for $p \in (e^{-1}, 1]$. Thus, it can be concluded that $g(p)$ is maximized when $p^* = e^{-1}$. Accordingly, the optimal ACB factor q^* in (14) can be obtained by substituting $p^* = e^{-1}$ into (10), and the maximum access throughput $\hat{\lambda}^a_{\max}$ can be derived by substituting $p^* = e^{-1}$ and (14) into (12).

Note that it has been shown in Section III that the network may have two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A. In the following, we prove that $p_A < e^{-1}$, implying that the maximum access throughput $\hat{\lambda}^a_{\max}$ can only be achieved when the network operates at p_L. Specifically, when the network has two steady-state points p_L and p_A, we can see from Appendix C that $0 < p_A < p_E < p_L \leq 1$ and $h'(p_E) > 0$, where $h'(p)$ and p_E are given in (42) and (44), respectively. By combining (42) and (44), we have

$$h'(p_E) = \frac{\left(\frac{q\beta}{2\lambda} - \mathbb{W}_0\left(\frac{\beta}{2}\exp(1+\frac{q\beta}{2\lambda})\right)\right)\left(\frac{q\beta}{2\lambda} + \mathbb{W}_0\left(\frac{\beta}{2}\exp(1+\frac{q\beta}{2\lambda})\right)\right)}{\beta} + \frac{1-2\mathbb{W}_0\left(\frac{\beta}{2}\exp(1+\frac{q\beta}{2\lambda})\right)}{\beta}. \quad (50)$$

Note that for $\beta \geq 1$, $1 - 2\mathbb{W}_0\left(\frac{\beta}{2}\exp(1+\frac{q\beta}{2\lambda})\right) < 1 - 2\mathbb{W}_0\left(\frac{\beta e}{2}\right) < 0$. Therefore, it can be seen from (50) that if $h'(p_E) > 0$, then we must have $\frac{q\beta}{2\lambda} > \mathbb{W}_0\left(\frac{\beta}{2}\exp(1+\frac{q\beta}{2\lambda})\right)$, with which $\ln p_E = -1 - \frac{q\beta}{2\lambda} + \mathbb{W}_0\left(\frac{\beta}{2}\exp(1+\frac{q\beta}{2\lambda})\right) < -1$, indicating that $p_A < p_E < e^{-1}$. ■

REFERENCES

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

[2] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*, Cisco Syst., San Jose, CA, USA, Mar. 2017.

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

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] J. J. Nielsen, D. M. Kim, G. C. Madueño, N. K. Pratas, and P. Popovski, "A tractable model of the LTE access reservation procedure for machine-type communications," in *Proc. IEEE Globecom*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[13] 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.

[14] 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.

[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, Aug. 2015.

[16] 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.

[17] 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.

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

[19] Z. Alavikia and A. Ghasemi, "Collision-aware resource access scheme for LTE-based machine-to-machine communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4683–4688, May 2018.

[20] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Modeling and throughput optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771–2785, Apr. 2018.

[21] Z. Feng, Z. Feng, and T. A. Gulliver, "Biologically inspired two-stage resource management for machine-type communications in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5897–5910, Sep. 2017.

[22] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 184–192, Jul. 2012.

[23] F. Ghavimi, Y.-W. Lu, and H.-H. Chen, "Uplink scheduling and power allocation for M2M communications in SC-FDMA-based LTE-A networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6160–6170, Jul. 2017.

[24] A. Aijaz, M. Tshangini, M. R. Nakhai, X. Chu, and A.-H. Aghvami, "Energy-efficient uplink resource allocation in LTE networks with M2M/H2H co-existence under statistical QoS guarantees," *IEEE Trans. Commun.*, vol. 62, no. 7, pp. 2353–2365, Jul. 2014.

[25] G. Zhang, A. Li, K. Yang, L. Zhao, Y. Du, and D. Cheng, "Energy-efficient power and time-slot allocation for cellular-enabled machine type communications," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 368–371, Feb. 2016.

[26] A. Azari and G. Miao, "Network lifetime maximization for cellular-based M2M networks," *IEEE Access*, vol. 5, pp. 18927–18940, 2017.

[27] 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.

[28] 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.

[29] H. S. Jang, S. M. Kim, H.-S. Park, and D. K. Sung, "Message-embedded random access for cellular M2M communications," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 902–905, May 2016.

[30] Y. D. Beyene, R. Jäntti, and K. Ruttik, "Random access scheme for sporadic users in 5G," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1823–1833, Mar. 2017.

[31] M. Centenaro, L. Vangelista, S. Saur, A. Weber, and V. Braun, "Comparison of collision-free and contention-based radio access protocols for the Internet of Things," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3832–3846, Sep. 2017.

[32] X. Zhang, Y. C. Liang, and J. Fang, "Novel Bayesian inference algorithms for multiuser detection in M2M communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7833–7848, Sep. 2017.

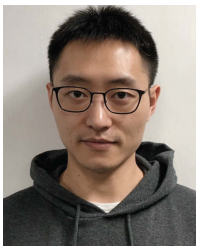
[33] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. Hoboken, NJ, USA: Wiley, 2012.

[34] *Digital Cellular Telecommunications System (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Service Accessibility*, document TS 22.011 V11.3.0, 3GPP, Apr. 2013.

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

[36] E. Dahlman, S. Parlvall, and J. Sköld, *4G: LTE/LTE-Advanced for Mobile Broadband*. Amsterdam, The Netherlands: Elsevier, 2014.

[37] *RAN Improvements for Machine-Type Communications*, document TR 37.868 V11.0.0, 3GPP, Oct. 2011.



Wen Zhan (S'17) received the B.S. and M.S. degrees from the 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 Electrical Engineering, City University of 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 the 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 the City University of Hong Kong, where she is currently a Full 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 (WCNC) 2007 and the IEEE Marconi Prize Paper Award in 2009.