Access Delay Optimization of M2M Communications in LTE Networks

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

***Abstract*—This letter focuses on the optimization of access delay performance of Machine-to-Machine (M2M) communications in Long Term Evolution (LTE) networks. Specifically, by deriving the probability generating function of access delay, both the first and second moments of access delay are obtained as functions of system parameters. The mean access delay is further minimized by optimally tuning the Access Class Barring (ACB) factor, and significant gains are demonstrated over the standard setting where the ACB factor is fixed.**

***Index Terms*—M2M communications, random access, LTE, access delay, optimization.**

I. Introduction

M2M COMMUNICATIONS has experienced a remarkable growth over the past few years. Many emerging M2M use-cases, such as industrial automation, rely on the real-time control, which imposes stringent latency requirements on wireless connectivity [1]. However, recent studies have demonstrated that as a key enabler for M2M communications, the LTE networks would suffer from severe congestion at the random access channel when a massive number of Machine-Type Devices (MTDs) simultaneously send access requests for connection establishment [2]. In that case, excessively long access delay occurs due to low chances of successful access. How to optimize the access delay performance has thus become a significant challenge for supporting M2M communications over LTE networks.

In LTE networks, the access delay of each MTD is defined as the time spent from the generation of an access request until its successful transmission. It closely depends on the backoff parameters including the Uniform Backoff (UB) window size and the ACB factor (i.e., the initial transmission probability of each MTD). The access delay performance with UB or ACB scheme has been analyzed in [3]–[10], where the mean access delay [3]–[8] or the probability mass function of access delay [9], [10] was calculated based on the number of access requests[1] in each time slot. In practice, however, such information is usually unavailable due to the uncoordinated nature of MTDs. Therefore, various algorithms were developed to estimate the number of access requests according to the number of idle slots [3], [4], or collision slots [5], [6]. In [7]–[10], iterative methods were further proposed to calculate the current number of access requests based on previous ones.

Despite the above extensive studies, effects of key system parameters on the access delay performance have not been well understood. For instance, the data arrival rate of each MTD, which determines the traffic load of the network, has been ignored in previous studies, where MTDs were usually assumed to be bufferless. How the mean access delay varies with the total number of MTDs or the number of preambles under different traffic conditions thus remains largely unknown. More importantly, due to the lack of accurate information of the number of access requests, the mean access delay has to be numerically calculated based on the estimated or iteratively updated number of access requests. The implicit nature renders it extremely difficult to further study how to optimize the access delay performance of each MTD by properly tuning backoff parameters.

In our recent work [11], a new analytical framework was proposed for M2M communications in LTE networks to maximize the network throughput. In this letter, the proposed analytical framework will further be extended to optimize the access delay performance of MTDs. Specifically, based on the discrete-time Markov process of each access request, the probability generating function of the access delay is derived, from which the first and second moments of access delay can further be obtained. The analysis shows that when the UB window size $W = 1$, the minimization of the second moment of access delay is equivalent to the minimization of the first moment of access delay, i.e., the mean access delay. Explicit expressions of the minimum mean access delay and the corresponding optimal ACB factor are obtained, which show that the minimum mean access delay linearly increases with the number of MTDs and is inversely proportional to the number of preambles when the total data arrival rate is large.

The remainder of this letter is organized as follows. Section II presents the system model and preliminary analysis. The minimum mean access delay and the corresponding optimal ACB factor are characterized in Section III, and verified by simulation results presented in Section IV. Finally, concluding remarks are summarized in Section V.

II. System Model and Preliminary Analysis

Consider a single-cell LTE system with n MTDs attempting to access the Base Station (BS). Assume that for each MTD, the arrivals of data packets follow a Bernoulli process with parameter $\lambda \in (0, 1)$, and one access request is generated once the MTD has data packets in its buffer. In the random access procedure, each MTD randomly selects one out of M orthogonal preambles and transmits its access request to the BS through the Physical Random Access CHannel (PRACH), which appears periodically [12]. We define the time slot as the interval between two consecutive PRACHs. In each time slot,

Manuscript received June 5, 2019; accepted August 11, 2019. Date of publication August 16, 2019; date of current version December 6, 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 H. Zhou. *(Corresponding author: Wen Zhan.)*

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

Digital Object Identifier 10.1109/LWC.2019.2935722

[1] In [3]–[10], the data buffer of each MTD was ignored. Thus, the terms “MTD” and “access request” were often used interchangeably in those studies.

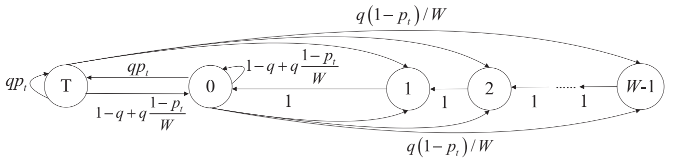

Fig. 1. State transition diagram of each individual access request.

an access request is successful if and only if there are no concurrent access request transmissions using the same preamble. Otherwise, a collision occurs and all of them fail.

In our recent work [11], a new analytical framework has been proposed for optimizing the access efficiency of M2M communications in LTE networks, in which a discrete-time Markov process $\{X_j, j = 0, 1, \ldots\}$ has been established to characterize the behavior of each access request, with the state transmission diagram of each access request shown in Fig. 1. A detailed description of the state transition process of each individual access request can be found in [11, Sec. II-B].

Let D_i denote the time spent from the beginning of State i until the service completion, where $i \in \{T, 0, 1, \ldots, W-1\}$. It can be obtained from Fig. 1 that

$$D_j = \begin{cases} 1 & \text{with probability } qp, \\ 1{+}D_0 & \text{with probability } 1-q+\frac{q(1-p)}{W}, \\ 1{+}D_i & \text{with probability } \frac{q(1-p)}{W}, i=1,\ldots,W-1, \end{cases} \quad (1)$$

$j \in \{T, 0\}$ and

$$D_i = 1 + D_{i-1}, \quad i = 1, 2, \ldots, W-1, \quad (2)$$

where $p = \lim_{t\to\infty} p_t$ is the steady-state probability of successful transmission of access requests, $q \in (0, 1]$ denotes the ACB factor and $W \in \{1, 2, \ldots\}$ represents the UB window size.

Note that a fresh access request is initially in State T and its service completes when it shifts back into State T. Therefore, D_T is the access delay of each MTD. Let $G_{D_T}(z)$ denote its probability generating function. According to (1) and (2), we have

$$G_{D_T}(z) = \frac{W(1-z)zqp}{W(1-z)(1-(1-q)z)-zq(1-z^W)(1-p)}. \quad (3)$$

The first moment of the access delay, i.e., the mean access delay (in unit of time slots), $E[D_T]$, and the second moment of the access delay $E[D_T^2]$ can then be obtained from (3) as

$$E[D_T] = G'_{D_T}(1) = \frac{1}{qp} + \frac{(1-p)(W-1)}{2p}, \quad (4)$$

and

$$\begin{aligned} E[D_T^2] &= G'_{D_T}(1) + G''_{D_T}(1) \quad (5)\\ &= E[D_T](2E[D_T]-1) + \tfrac{1}{3}(\tfrac{1}{p}-1)(W^2-1), \end{aligned}$$

respectively.

We can see from (4) and (5) that both $E[D_T]$ and $E[D_T^2]$ are crucially determined by the ACB factor q and the UB window size W. It is interesting to observe from (5) that when the UB window size $W = 1$, the minimization of the second moment of access delay $E[D_T^2]$ is equivalent to the minimization of the mean access delay $E[D_T]$. Therefore, in this letter, we assume $W = 1$ and focus on the optimization of mean access delay. We are interested in minimizing the mean access delay $E[D_T]$ by optimally tuning the ACB factor q, i.e., $\min_q E[D_T]$.

III. Minimum Mean Access Delay

Let us start by establishing the relationship between the mean access delay and network throughput. In [11], the network throughput, which is defined as the average number of successful access requests per time slot, was derived as

$$\hat{\lambda}_{out} = \frac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{qp}+\frac{(1-p)(W-1)}{2p}\right)+1}, \quad (6)$$

where $\hat{\lambda} = n\lambda$ is the aggregate input rate. By combining (4) and (6), we can see that

$$E[D_T] = \frac{n}{\hat{\lambda}_{out}} - \frac{1}{\lambda}, \quad (7)$$

indicating that the mean access delay $E[D_T]$ is minimized when the network throughput $\hat{\lambda}_{out}$ is maximized.

It has been shown in [11] that the network has either two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A with $p_L > p_A$, or one steady-state point p_L, depending on whether it operates at the bistable region or monostable region. Both p_L and p_A are nonzero roots of the fixed-point equation of p, which was derived in [11] as

$$p = \exp\left(-\frac{\hat{\lambda}/M}{p+\hat{\lambda}/(nq)}\right), \quad (8)$$

when the UB window size $W = 1$. The maximum network throughput $\hat{\lambda}_{\max} = Me^{-1}$ is achieved only when the network operates at the desired steady-state point p_L and the ACB factor q is set to

$$q^* = \frac{\hat{\lambda}/n}{\hat{\lambda}/M - e^{-1}}. \quad (9)$$

It is clear from (7) that the minimum mean access delay is achieved at $\frac{n}{\hat{\lambda}_{\max}} - \frac{1}{\lambda} = \frac{ne}{M} - \frac{1}{\lambda}$ (time slots), when the ACB factor $q = q^*$ and the network operates at the monostable region. If the network operates at the bistable region for $q = q^*$, however, then the network may shift to the undesired steady-state point p_A, at which the minimum mean access delay $\frac{ne}{M} - \frac{1}{\lambda}$ is unachievable. To see how to optimally tune the ACB factor q to minimize the mean access delay in that case, in the following, we will 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 III-B.

A. Minimum Mean Access Delay With $M = 1$

Let us first determine the conditions for the network to operate at the desired steady-state point p_L with $q = q^*$. In [11], the bistable region and monostable region were defined in terms of the quadruple $(n, \hat{\lambda}, q, W)$. With the UB window size $W = 1$, for given the number of MTDs n and the aggregate input rate $\hat{\lambda}$, the bistable region and monostable region in terms of the ACB factor q can be written as:

- *Bistable region* $\mathcal{B}_q = \{q | q_1 \le q \le q_2\}$, in which the network has two steady-state points p_L and p_A.
- *Monostable region* $\mathcal{M}_q = \bar{\mathcal{B}}_q$, in which the network has only one steady-state point p_L. q_1 and q_2 are given by

$$q_1 = -\frac{4\mathbb{W}_{-1}^2\left(-\sqrt{\hat{\lambda}}/2\right)}{n+2n\mathbb{W}_{-1}\left(-\sqrt{\hat{\lambda}}/2\right)}, \quad q_2 = -\frac{4\mathbb{W}_0^2\left(-\sqrt{\hat{\lambda}}/2\right)}{n+2n\mathbb{W}_0\left(-\sqrt{\hat{\lambda}}/2\right)}, \quad (10)$$

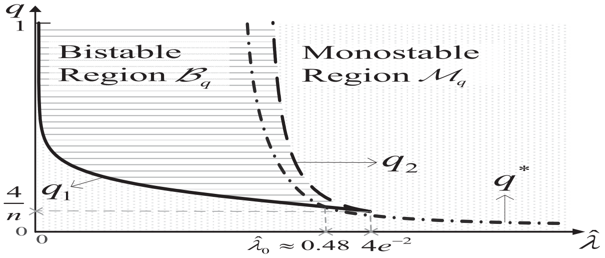

Fig. 2. Bistable region $\mathcal{B}_q$ and monostable region $\mathcal{M}_q$.

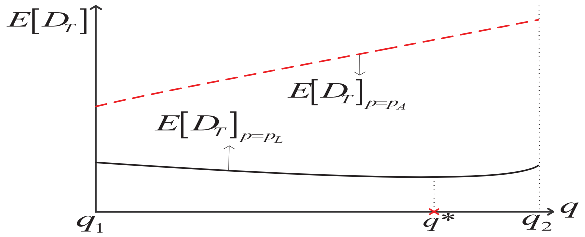

Fig. 3. Mean access delay $E[D_T]$ versus the ACB factor q for $q \in \mathcal{B}_q$ when $\hat{\lambda} \leq \hat{\lambda}_0$.

respectively,[2] where $\mathbb{W}_0(\cdot)$ and $\mathbb{W}_{-1}(\cdot)$ are two real-valued branches of the Lambert W function with $\mathbb{W}_0(\cdot)$ as the principal branch.

Fig. 2 illustrates the bistable region $\mathcal{B}_q$ and monostable region $\mathcal{M}_q$. It can be clearly seen from Fig. 2 that q^* is included in the monostable region $\mathcal{M}_q$ only when the aggregate input rate $\hat{\lambda}$ is sufficiently high, i.e., $\hat{\lambda} > \hat{\lambda}_0$, where $\hat{\lambda}_0 \approx 0.48$ is the single non-zero root of the equation $\hat{\lambda} - \hat{\lambda}(1+1/\mathbb{W}_{-1}(-\sqrt{\hat{\lambda}}/2))^2 = 4(\hat{\lambda} - e^{-1})$, which is derived by combining $q^* = q_1$, (9) and (10). With $\hat{\lambda} > \hat{\lambda}_0$, the network is guaranteed to operate at the desired steady-state point p_L when the ACB factor q is set to q^*, with which the network throughput is maximized and the mean access delay is minimized. We then have

$$\arg\min_q E[D_T]|_{\hat{\lambda} > \hat{\lambda}_0} = q^*. \tag{11}$$

On the other hand, if $\hat{\lambda} \leq \hat{\lambda}_0$, then $q^* \in \mathcal{B}_q$, indicating that the network has two steady-state points with $q = q^*$. In this case, q^* may not be the optimal ACB factor for minimizing the mean access delay as the network may shift to the undesired steady-state point p_A that is much lower than the desired steady-state point p_L. As Fig. 3 illustrates, for $q \leq q^* \in \mathcal{B}_q$, the mean access delay $E[D_T]_{p=p_L}$ monotonically decreases as the ACB factor q increases if the network operates at the desired steady-state point p_L. If it operates at the undesired steady-state point p_A, on the other hand, $E[D_T]_{p=p_A}$ increases with q, and $E[D_T]_{p=p_A} > E[D_T]_{p=p_L}$. It is unknown when the network would shift from the desired steady-state point p_L to the undesired steady-state point p_A. Yet, the larger q, the higher risk of the network shifting from p_L to p_A. Therefore, to minimize the chance of operating at the undesired steady-state point p_A, q should be set at the lowerbound q_1, i.e.,

$$\arg\min_q E[D_T]|_{\hat{\lambda} \leq \hat{\lambda}_0} = q_1. \tag{12}$$

Finally, by denoting the optimal ACB factor for minimizing the mean access delay as q_D^* and combining (9)–(12), we have

$$q_D^{*,M=1} = \begin{cases} \dfrac{\hat{\lambda}}{n(\hat{\lambda} - e^{-1})} & \text{if } \hat{\lambda} > \hat{\lambda}_0, \\ \dfrac{4\mathbb{W}_{-1}^2\left(-\sqrt{\hat{\lambda}}/2\right)}{n\left(-2\mathbb{W}_{-1}\left(-\sqrt{\hat{\lambda}}/2\right)-1\right)} & \text{otherwise.} \end{cases} \tag{13}$$

It can be seen from (13) that the optimal ACB factor $q_D^{*,M=1}$ decreases as the number of MTDs n or the aggregate input rate $\hat{\lambda}$ increases.

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

The above analysis is based on the assumption that all n MTDs share one preamble, i.e., $M = 1$. In this section, by applying the multi-group model proposed in [11], the analysis will be extended to the multi-preamble scenario in which n MTDs choose $M > 1$ preambles. Specifically, as the preambles are orthogonal to each other, we can divide n MTDs into M groups according to the preamble each MTD chooses. Since each MTD independently and randomly selects a preamble in each access attempt, the group size $n^{(i)}$ can then be approximated as $n^{(i)} \approx \frac{n}{M}$ if the number of MTDs n is large. By replacing n and $\hat{\lambda}$ with $n^{(i)} \approx \frac{n}{M}$ and $\hat{\lambda}^{(i)} \approx \frac{n\lambda}{M}$ in (13), respectively, the optimal ACB factor can be obtained as

$$q_D^* = \begin{cases} \dfrac{\lambda}{\frac{n\lambda}{M} - e^{-1}} & \text{if } \lambda > \frac{M}{n}\hat{\lambda}_0, \\ \dfrac{4M\mathbb{W}_{-1}^2\left(-\sqrt{n\lambda/M}/2\right)}{n\left(-2\mathbb{W}_{-1}\left(-\sqrt{n\lambda/M}/2\right)-1\right)} & \text{otherwise.} \end{cases} \tag{14}$$

By substituting (14) into (4) and (8), the corresponding minimum mean access delay can be derived as

$$\min_q E[D_T] = \begin{cases} \frac{ne}{M} - \frac{1}{\lambda} & \text{if } \lambda > \frac{M}{n}\hat{\lambda}_0, \\ \frac{n\left(-2\mathbb{W}_{-1}\left(-\sqrt{n\lambda/M}/2\right)-1\right)}{4M\mathbb{W}_{-1}^2\left(-\sqrt{n\lambda/M}/2\right)p_L^{q=q_1}} & \text{otherwise,} \end{cases} \tag{15}$$

where $p_L^{q=q_1}$ is the desired steady-state point of the network with $q = q_1$, which is the larger non-zero root of the fixed-point equation $n\lambda(1 + 2\mathbb{W}_{-1}(-\sqrt{n\lambda/M}/2)) - 4Mp\mathbb{W}_{-1}^2(-\sqrt{n\lambda/M}/2) = 4n\lambda\mathbb{W}_{-1}^2(-\frac{\sqrt{n\lambda/M}}{2})/\ln p$.

We can see from (15) that the minimum mean access delay $\min_q E[D_T]$ increases as the number of preambles M decreases or the network size n grows. For large n or λ, i.e., $n\lambda > M\hat{\lambda}_0$, a linear increase of the minimum mean access delay $\min_q E[D_T]$ with regard to n and $\frac{1}{M}$ can be observed.

IV. Simulation Results

In this section, simulation results are presented to verify the above analysis. The simulation setting is the same as the system model described in Section II, and each simulation is carried out for 10^8 time slots. In simulations, the mean access delay is obtained by calculating the ratio of the sum of access delay of all successfully transmitted access requests to the total number of successful access requests.

[2] q_1 and q_2 are the roots of $\hat{\lambda}_1 = \hat{\lambda}$ and $\hat{\lambda}_2 = \hat{\lambda}$, respectively, where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given in [11, eqs. (7)-(8)].

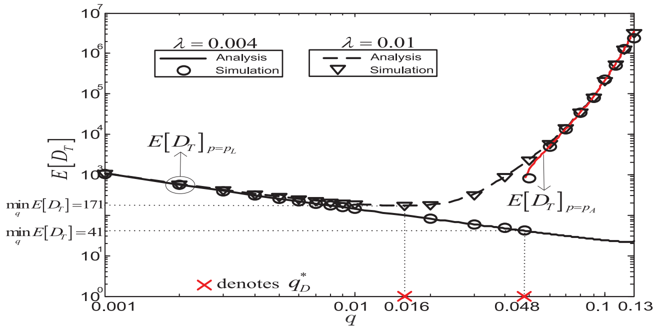

Fig. 4. Mean access delay $E[D_T]$ (in unit of time slots) versus the ACB factor q. $n = 1000$. $M = 10$. $\lambda \in \{0.004, 0.01\}$.

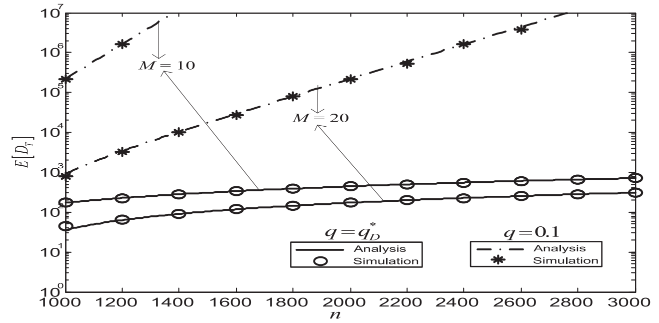

Fig. 5. Mean access delay $E[D_T]$ (in unit of time slots) versus the number of MTDs n. $M \in \{10, 20\}$. $\lambda = 0.01$. $q = q_D^*$ or 0.1.

Specifically, the expression of mean access delay $E[D_T]$ has been given in (4) and illustrated in Fig. 4, where the number of preambles $M = 10$ and the number of MTDs $n = 1000$. For large data arrival rate λ, i.e., $\lambda = 0.01 > \frac{M\hat{\lambda}_0}{n} \approx 0.0048$, the analysis has shown that the network is guaranteed to operate at the desired steady-state point p_L for $q = q^* = 0.016$, with which the mean access delay is minimized. It can be seen from Fig. 4 that the minimum mean access delay $\min_q E[D_T] = \frac{ne}{M} - \frac{1}{\lambda} = 171$ time slots is indeed achieved with $q_D^* = 0.016$.

On the other hand, for $\lambda = 0.004 < \frac{M\hat{\lambda}_0}{n}$, the network operates at the bistable region with two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A, when the ACB factor $q = q^* = 0.125$. As Fig. 4 shows, for $q \in \mathcal{B}_q = \{q | 0.048 \le q \le 0.13\}$, the network quickly shifts to p_A as q increases, at which the mean access delay $E[D_T]$ is much higher than that at p_L. Therefore, to minimize the mean access delay, the optimal ACB factor q^* should be set to the lowerbound $q_D^* = q_1 = 0.048$, with which the minimum mean access delay $\min_q E[D_T] = 41$ time slots is achieved according to (15). Simulation results presented in Fig. 4 well agree with the analysis.

Note that the analysis further shows that with the ACB factor optimally tuned as $q = q_D^*$, the minimum mean access delay $\min_q E[D_T]$ linearly increases with n and $\frac{1}{M}$ for large number of MTDs n, which is verified by simulation results presented in Fig. 5. In sharp contrast, if the ACB factor q is fixed, as the current standard does [13], then the mean access delay $E[D_T]$ exponentially increases with n and $\frac{1}{M}$, as Fig. 5 illustrates. The gap between the minimum mean access delay and the mean access delay with the ACB factor q fixed at 0.1 grows as the network size n increases or the number of preambles M decreases, indicating that adaptive tuning of the ACB factor is crucial for massive access scenarios with a limited number of preambles.

V. Conclusion

In this letter, the analytical framework in [11] is extended to study how to properly tune backoff parameters to optimize the access delay performance of M2M communications in LTE networks. By establishing the relationship between the mean access delay of each MTD and the network throughput, it is shown that the mean access delay is minimized when the network throughput is maximized, to achieve which the network should operate at the desired steady-state point with backoff parameters properly chosen. Explicit expressions of the minimum mean access delay and the corresponding optimal ACB factor are obtained, which reveal that in contrast to the maximum network throughput that is solely determined by the number of preambles, the minimum mean access delay further depends on the network size and the traffic input rate of each MTD. Simulation results corroborate that compared to the standard setting where the ACB factor is fixed, the mean access delay of each MTD can be substantially reduced by optimally tuning the ACB factor according to the number of MTDs and the traffic input rate of each MTD. The improvement is especially significant in the massive access scenarios with a limited number of preambles.

References

[1] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[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] X. Jian, Y. Liu, Y. Wei, X. Zeng, and X. Tan, "Random access delay distribution of multichannel slotted ALOHA with its applications for machine type communications," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 21–28, Feb. 2017.

[11] 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.

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

[13] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) Protocol Specification, V13.3.0*, document TS 36.331, 3GPP, Sophia Antipolis, France, Jan. 2017.