Performance Optimization for Massive Random Access of mMTC in Cellular Networks With Preamble Retransmission Limit

Wen Zhan, *Member, IEEE*, Xinghua Sun, *Member, IEEE*, Xijun Wang, *Member, IEEE*, Yaru Fu, *Member, IEEE*, and Yitong Li

***Abstract*—As one of the three main application scenarios of 5 G cellular system, massive Machine-Type Communications (mMTC) has been regarded as the key solution to facilitate the IoT paradigm. One major bottleneck for accommodating mMTC is the severe congestion at the cellular random access channel when plenty of Machine-Type Devices (MTDs) send access requests concurrently while the preamble resources are limited. To remedy this issue, limiting the number of retransmissions and dropping access requests after the limit is reached can be an effective approach. Yet, the effect of the preamble retransmission limit K on the optimal access performance of mMTC in cellular networks remains largely unexploited, which motivates the study in this paper. Specifically, in this paper, we start by characterizing the network steady-state points based on the limiting probability of successful transmission of access requests. We then obtain explicit expressions of the access throughput and the mean access delay of successfully-transmitted access requests as functions of K and the number of preambles M. The maximum access throughput and the corresponding optimal backoff window size are further derived. It is shown that the maximum access throughput is independent of K, while the mean access delay can be significantly reduced with a small K, yet, at the expense of increased request dropping ratio. In addition, to improve both the throughput and delay performance, the analysis shows that more preambles should be allocated but the performance gain becomes marginal when M is large. Therewith, an algorithm is proposed for determining the least number of preambles M^* that maximizes the access throughput and the preamble resource utilization ratio. Numerical results show that a smaller preamble retransmission limit K can further reduce M^*.**

***Index Terms*—Machine-type communications, optimization, random access, retransmission limit.**

I. Introduction

MACHINE-TYPE Communications (MTC) is a new type of communication paradigm in which devices can automatically operate, communicate, and process information without or with few human intervention. It is the foundation of many emerging Internet of Things (IoT) applications such as smart city, smart agriculture and Internet of vehicles [1]. Recent reports have revealed that the number of Machine-Type Devices (MTDs) will surpass 10 billion in the next few years and the data traffic from MTDs will constitute a substantial portion of the traffic volume in the next-generation cellular networks [2]. Therefore, massive MTC (mMTC) has been identified as one of the three generic services to be supported by 5 G cellular system [3].

To accommodate mMTC in cellular networks, many challenges have to be addressed. One of the cruces that capture much attention is the severe congestion in the cellular random access channel [4]. Specifically, the random access procedure of the cellular system is designed based on the principle of multi-channel slotted Aloha. That is, each MTD randomly chooses one orthogonal preamble from the preamble resource pool and transmits via a shared random access channel. If more than one MTDs transmit the same preamble simultaneously, the collision occurs, resulting in the failures of all involved access requests. Due to limited preamble resources, when a large number of MTDs transmit access requests, frequent preamble collisions occur and the access efficiency becomes intolerably low.

To relieve the congestion, preamble (re)transmission limit, denoted by $K \geq 1$, is adopted in the cellular system at the MAC layer of user equipment side [5]. That is, each MTD with collided preamble would retransmit the access request only if it has not exhausted the retransmission limit. When the limit is reached, the

Manuscript received July 24, 2020; revised December 10, 2020 and April 17, 2021; accepted July 4, 2021. Date of publication July 13, 2021; date of current version September 17, 2021. The work of Wen Zhan was supported in party by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101120003, in part by the National Natural Science Foundation of China under Grant 62001524, in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant 2021A04, and in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, 2021qntd11. The work of Xinghua Sun was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101120003 and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515011906. The work of Xijun Wang was supported in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, 2021qntd09 and in part by GuangDong Basic and Applied Basic Research Foundation under Grant 2021A1515012631. The work of Yaru Fu was supported by Hong Kong President's Advisory Committee on Research and Development (PACRD) under Project 2020/1.6. The work of Yitong Li was supported by the National Natural Science Foundation of China under Grant 61801433. The review of this article was coordinated by Prof. Shibo He. (*Corresponding authors: Xinghua Sun and Xijun Wang*.)

Wen Zhan and Xinghua Sun are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 510275, China (e-mail: zhanw6@mail.sysu.edu.cn; sunxinghua@mail.sysu.edu.cn).

Xijun Wang is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangdong 510275, China (e-mail: wangxijun@mail.sysu.edu.cn).

Yaru Fu is with the School of Science and Technology, The Open University of Hong Kong, Hong Kong SAR, China (e-mail: yfu@ouhk.edu.hk).

Yitong Li is with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: ieytli@zzu.edu.cn).

Digital Object Identifier 10.1109/TVT.2021.3096259

MAC layer indicates a random access problem to upper layers and the MTD drops the access request. It is clear that with a small K, the number of concurrent access requests can be reduced. The probability of successful access of each MTD might be boosted, however, at the cost of a high packet dropping ratio. The preamble retransmission limit K would significantly affect the access performance. It is therefore of great importance to characterize its impact on the access performance and understand how to optimize the access performance of mMTC in cellular systems with finite preamble retransmission limit.

A. Related Works

Extensive works have been done for mMTC in cellular networks in investigating its random access performance. Therein, the classical slotted Aloha analysis is widely adopted and most of the existing literature put the major focus on the network stability. For instance, with an infinite retransmission limit, it was demonstrated in [6], [7] that the slotted Aloha network exhibits bistable behavior, i.e., it may possess multiple statistically steady-state equilibrium points, which can be characterized based on the limiting probability of successful transmission of packets. With finite retransmission limit and node population, [8], [9] revealed that if the retransmission limit is small, then the bistable behavior can be mitigated. Otherwise, the network may still have multiple steady-state points. [10] developed a multi-dimensional Markov process to characterize the stability region of a heterogeneous Aloha network, where nodes have different input rates. With an infinite node population, [11] numerically discussed the region of retransmission limit that guarantees the network stability.

For analyzing the effect of preamble retransmission limit on the random access performance in cellular systems, numerous analytical models were proposed. In [12], by modeling the arrival process of access requests with a Poisson distribution, the request load region, outside of which the system is stable, was obtained as the function of retransmission limit. In [13]–[16], by recursively calculating the average number of the contending devices in each time slot, analytical models were developed for characterizing the transient behavior of aggregate channel traffic and deriving performance measures, such as collision probability and access delay, for given preamble retransmission limit. In [17], [18], queueing models for individual MTD were established, which revealed that a smaller preamble retransmission limit may improve the throughput and delay performance.

In above studies, the main focus is on numerically evaluating the access performance of mMTC in cellular networks for given system parameters, e.g., preamble retransmission limit. How to properly select system parameters to optimize the access efficiency is another interesting problem, which is of great importance in practice. Specifically, it was explicitly demonstrated by [19]–[25] that the system parameters (such as the total amount of preambles, the transmission probability of access requests, and the backoff window size) can crucially affect the access efficiency. Accordingly, numerous schemes were proposed to adaptively tune those system parameters by periodically estimating the aggregate traffic [19]–[30]. The effectiveness of those schemes, nevertheless, relies on how accurate the traffic estimation is.

In recent, a new analytical framework was developed in [31] to optimize the long-term throughput [31] and delay [32] performance of mMTC in cellular networks. By formulating a Markov model to characterize the state transmission process of each access request, the network steady-state points were derived, based on which the maximum access throughput, the minimum mean access delay and the optimal backoff parameters were all obtained. As only the statistical traffic information, such as the number of devices in the network, is needed for the optimal tuning of backoff parameters, realtime traffic estimation is thus no longer required. However, in [31] and [32], an infinite retransmission limit $K = +\infty$ was assumed, with which little light was shed on how to properly tune system parameters to optimize the access efficiency with a finite retransmission limit $K < +\infty$.

B. Our Contributions

In this paper, by extending the analytical framework in [31] to incorporate a finite preamble retransmission limit $K < +\infty$, we will address the above issue and demonstrate the effect of $K < +\infty$ on the optimal access efficiency. Specifically, the network steady-state points, the access throughput and the mean access delay are all obtained as the functions of K, based on which the optimal access efficiency and the corresponding optimal backoff window size and number of preambles are further derived. Our main contributions are summarized as follows:

1) Steady-state Points Analysis: We apply a discrete-time Markov renewal process to characterize the behavior schema regarding each access request, based on which we obtain the fixed-point equation with respect to the limiting probability of the successful transmissions. It is shown that the network has the property of having one or two steady-state points, which is denoted by p_L and p_A, where $p_L > p_A$. Both steady-state points are monotonic increasing functions of backoff window size W. Yet, the monotonicity with regards to the retransmission limit K depends on W. If W is below a certain threshold, then both steady-state points decline as K grows; Otherwise, they increase as K grows. Such a critical threshold is explicitly characterized and is shown to be solely determined by the traffic input rate of each device λ.

2) Performance Optimization: For the throughput maximization, we derive the closed-form expressions in regard to the maximum access throughput as well as the corresponding optimal backoff window size. It is shown that the maximum access throughput has no concern with the preamble retransmission limit K while the optimal backoff window size should be properly enlarged when K increases. The analysis also reveals that the network is of necessity to operate on the desired steady-state point p_L such that the maximum access throughput is achievable. To mitigate the bistability, we explicitly characterize the lower-bound of the aggregate input rate, above which the network has only one steady-state point regardless of K. For

the delay performance, the analysis shows that a small preamble retransmission limit K can effectively reduce the mean access delay of access requests that are successfully transmitted. Yet, the performance improvement is achieved at the cost of packet dropping. When K is large, the minimization of the mean access delay is equivalent to the maximization of the access throughput.

3) Optimal Number of Preambles: To boost the throughput and delay performance, intuitively, the system should allocates more preambles. However, the analysis shows that an over-provisioning of preambles makes little contrition to performance improvement while degrades preamble resource utilization ratio, i.e., access throughput per preamble. As preambles are scarce resources in cellular systems, in this paper, an algorithm is proposed for determining the least number of preambles M^* that maximizes both the access throughput and the preamble resource utilization ratio. It is shown that a smaller preamble retransmission limit K can further reduce M^* because of packet dropping.

The rest parts of this paper are organized as follows. In Section II, we elaborate the details on the considered system model. In Section III, we analyze the network steady-state points. In Section IV and Section V, we evaluate and optimize the access throughput and access delay performance in the single-preamble scenario, respectively. In Section VI, we further focus on the multi-preamble scenario. The implementation to practical 5 G system design is demonstrated in Section VII. Finally, Section VIII gives the conclusion.

II. System Model

This paper considers a general single-cell cellular system, which consists of one Base Station (BS) and n MTDs. Note that although 3GPP has recently developed multiple new radio access technologies [33] (such as NB-IoT or LTE-M) for machine-type services, the random access procedure is, in essence, similar. It contains four steps: 1) preamble transmission, 2) random access response, 3) radio resource control connection request transmission and 4) radio resource control connection setup message transmission. Since existing works have demonstrated that the success of access request is mainly determined by the preamble transmission step [4], [12], [17]–[20], [22], [25], we therefore stipulate that the random access procedure can be successfully completed as long as the preamble transmission is successful.

In the random access procedure, each MTD transmits a randomly selected preamble. The total number of preambles is $M \in \mathbb{N}^+$. According to the standards [34], preambles are transmitted via the shared Physical Random Access CHannel (PRACH), which appears periodically in time axis. Let us define the time slot as the interval between two neighboring PRACH. Therefore, an MTD can transmit one preamble in one time slot. If multiple MTDs transmit the same preamble at the same time, then all of them fail due to preamble collision. The access request can be successfully transmitted only when there is no concurrent transmission of the same preamble at the same time slot.

Assume that the buffer size of each MTD is infinite. In each time slot, the arrivals of data packets follow a Bernoulli process with parameter $\lambda \in (0, 1)$. The MTD with nonempty buffer has an access request. If the access request is successful, then the MTD proceeds to the data transmission process, in which it clears the data queue within one time slot. If the transmission attempts of the access request fail for K times, then the MTD drops it along with data packets in the data queue. It is clear that the scenario in [31], [32] is a special case of $K = +\infty$. When the number of MTDs n is large, the network can be regarded as a multi-queue-single-server system, wherein each individual MTD's request queue is characterized by a $Geo/G/1/1$ queue model [31].

Since preambles are orthogonal to each other, MTDs would not contend with each other if they use different preambles. For better illustration, we focus on the single-preamble scenario $M = 1$ in Sections III–V, and then extend the analysis to the scenario with multiple preambles in Section VI.

III. Steady-State Point Analysis

In this section, we first characterize the behavior of each access request and then derive the limiting probability of successful transmission of access requests, i.e., the network steady-state point.

A. Behavior of Each Access Request

To model the behavior of each access request, a discrete-time Markov renewal process $(\mathbf{X}, \mathbf{T}) = \{(X_j, T_j), j = 0, 1, \ldots\}$ is established, where T_j are the state transition times and X_j are the related states in the embedded Markov chain $\mathbf{X} = \{X_j\}$, as shown in Fig. 1.

Let $\mathcal{S}$ denote the state space of $\mathbf{X}$, and we have $\mathcal{S} = \{T, \mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_{K-1}, D\}$. Those states can be segmented into three categories: 1) successful request transmission (State T), 2) backoff (State $\mathrm{R}_i, i = 1, 2, \ldots, K-1$.) and 3) packet drop (State D), where i represents the total number of the experienced transmission failures by the access request. Let p_t denote the probability of successful transmission of access requests at the t-th time slot. As illustrated by Fig. 1, we see that an access request is successful if and only if an access request remains in State T, swaps from States R_i to T or changes from States D to T.

Otherwise, the access request shifts from States T to R_1 or from States R_i to R_{i+1}. Given that the transmission fails for K times, the access request is dropped and enters State D. Therefore, a new access request is either initially in State T if the previous request is successfully transmitted or in State D if the previous request is dropped.

We can obtain the steady-state probability distribution with respect to the Markov chain in Fig. 1 as

$$\begin{cases} \pi_{\mathrm{R}_i} = \frac{(1-p)^i}{1-(1-p)^K}\pi_T, \text{ for } i = 1, 2, \ldots, K-1, \\ \pi_D = \frac{(1-p)^K}{1-(1-p)^K}\pi_T, \end{cases} \tag{1}$$

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

The holding time in State X_j is the interval between two successive transitions, i.e., $T_{j+1} - T_j$, which depends on State $X_j, j = 1, 2, \ldots$ only. Let τ_i be the average holding time in State

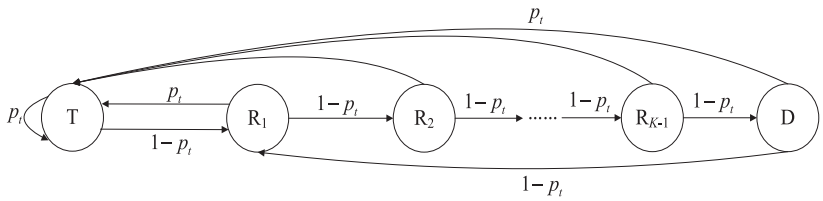


Fig. 1. State transition process of each access request.

i, where $i \in \mathcal{S}$. According to the standard [5], a fresh access request will be transmitted instantly. Therefore, the holding time in State T and that in State D are equal to one unit time slot, namely,

$$\tau_i = 1, i \in \{\mathrm{T}, \mathrm{D}\}. \tag{2}$$

The holding time in State $\mathrm{R}_i, i = 1, 2, \ldots, K-1$, is closely related to the backoff protocol. In this paper, we consider the Uniform Backoff (UB) scheme adopted in the 3GPP MAC protocol [5]. That is, once the MTD involves a collision, a backoff counter is randomly chosen from the set $\{0, \ldots, W\}$, where W depicts the UB window size with the unit of time slots. The backoff counter decreases by one in each time slot. A request will be retransmitted by the MTD given that the backoff counter reaches zero. The mean holding time in State R_i should then be

$$\tau_{\mathrm{R}_i} = \tfrac{W+1}{2}, i = 1, 2, \ldots, K-1. \tag{3}$$

Finally, we can obtain the limiting state probabilities of the Markov renewal process $(\mathbf{X}, \mathbf{T})$ as follows

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

Based on (1)–(4), we have

$$\tilde{\pi}_T = \frac{1-p_d}{1+\frac{(W+1)(1-p-p_d)}{2p}}, \tag{5}$$

where p_d denotes the probability that an access request is dropped, and is given by

$$p_d = (1-p)^K. \tag{6}$$

It is clear that each access request has the probability of $1-p_d$ to be successfully transmitted, served with rate $\tilde{\pi}_T$, or has the probability of p_d to be dropped, served with rate $\tilde{\pi}_D$. With the $Geo/G/1/1$ queueing model, the probability that each MTD's buffer is nonempty can then be written as

$$\rho = \tfrac{\lambda(1-p_d)}{\lambda(1-p_d)+\tilde{\pi}_T} = \tfrac{\lambda p_d}{\lambda p_d + \tilde{\pi}_D}. \tag{7}$$

B. Steady-State Points

In accordance with the above discrete-time Markov renewal process, let us now derive the steady-state points of the network based on the fixed-point equation of p. In specific, each MTD in the network must be either in State S_1 or in State S_2, where

S_1: The buffer is empty;

S_2: The access request is in State $i, i \in \mathcal{S}$.

For each MTD, its access request transmission is successful only when all the other $n-1$ MTDs are either in State S_1 or in State S_2 but not transmitting request. Thus, the probability of successful transmission of access requests p can be expressed as

$$\begin{aligned} p = (&\Pr\{\text{MTD in } S_1\} \\ &+ \Pr\{\text{MTD in } S_2 \text{ with no transmission}\})^{n-1}, \end{aligned} \tag{8}$$

The probability that an MTD is in State S_1 is

$$\Pr\{\text{MTD in } S_1\} = 1-\rho, \tag{9}$$

and the probability that it does not request transmission given that it is in State S_2 is

$$\begin{aligned} &\Pr\{\text{MTD in } S_2 \text{ with no transmission}\} \\ =&\rho\left(1-\tilde{\pi}_T-\tilde{\pi}_D-\sum_{j=1}^{K-1}\tilde{\pi}_{\mathrm{R}_j} r_j\right), \end{aligned} \tag{10}$$

according to Fig. 1. r_j denotes the conditional probability that the MTD performs the access request transmission when the access request is in State R_j, which is given by [35]

$$r_j = \tfrac{2}{W+1}, j \in \{1, 2, \ldots, K-1\}. \tag{11}$$

When n is large, we have $n-1 \approx n$ and $(1-x)^n \approx \exp\{-nx\}$. Together with (7)–(11), we obtain the fixed-point equation as

$$p \overset{\text{with a large } n}{\approx} \exp\left(-\frac{\hat{\lambda}(1-p_d)}{p+\frac{\hat{\lambda}p}{n}+\frac{\hat{\lambda}(W+1)(1-p)(1-p_d)}{2n}}\right), \tag{12}$$

where $\hat{\lambda} = n\lambda$ denotes the aggregate input rate. It can be seen from (12) that if the preamble retransmission limit $K = +\infty$, (12) reduces to (6) in [31].

Note that when $K = 1$, (12) has only one solution, which is explicitly given as

$$p_L^{K=1} = \exp\left(-\frac{\hat{\lambda}}{1+\frac{\hat{\lambda}}{n}}\right). \tag{13}$$

As can be inferred from (13), $p_L^{K=1}$ is solely affected by the number of MTDs and the aggregate input rate, but unrelated to backoff window size, because the access request is transmitted for only once no matter whether it is successful or not.

On the other hand, for $1 < K < +\infty$, it can be easily shown that the fixed-point equation (12) has at least one root based on the Intermediate Value Theorem [38]. However, due to the nonlinearity of (6) and (12), it is difficult to derive the explicit expressions of the non-zero roots of (12).

For illustration, we let $f(p) = \exp(-\frac{\hat{\lambda}(1-(1-p)^K)}{p+\frac{\hat{\lambda}p}{n}+\frac{\hat{\lambda}(W+1)(1-p)(1-(1-p)^{K-1})}{2n}}) - p$, and $f(p) = 0$ has

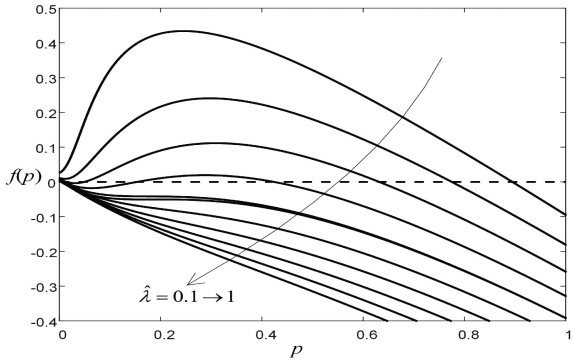


Fig. 2. Roots of (12). Each intersection point of the solid line and the dashed line represents a root. $n = 100$. $W = 40$. $K = 100$.

the same roots as those of (12). Numerical results are presented in Fig. 2. We can see that (12) has either three roots $0 < p_A \leq p_S \leq p_L \leq 1$ or only one root $0 < p_L \leq 1$, which is consistent with the observations in [8], [12]. As shown by Fig. 2, if $\hat{\lambda}$ is either small, for example, $\hat{\lambda} = 0.1$, or large, for instance, $\hat{\lambda} = 1$, then (12) will have only one root, in which case the system is said to be stable, because it is guaranteed to operate at p_L. Otherwise, the system is said to be bistable, in which by following the approximate trajectory analysis in [36], we can find that only p_L and p_A are steady-state points. Since $p_A < p_L$, p_L is referred to as the desired steady-state point and p_A is referred to as the undesired steady-state point.

Corollary 1 shows the monotonicity of the steady-state points in terms of the UB window size W and the preamble retransmission limit K.

Corollary 1: 1) For $K \in (1, +\infty)$, the steady-state points p_L and p_A are monotonic increasing functions of W.

2) For $W \in [1, \tilde{W})$, the steady-state points p_L and p_A are monotonic decreasing functions of K. For $W \in [\tilde{W}, +\infty)$, p_L and p_A are monotonic increasing functions of K, where $\tilde{W} = \frac{2}{\hat{\lambda}}$.

Proof: See Appendix A. ■

We can see from Corollary 1 that enlarging the UB window size can always benefit the limiting probability of successful transmission of access requests p. In sharp contrast, how the preamble transmission limit K affects p closely depends on whether the UB backoff window size $W > \tilde{W}$ or not. If W is sufficiently large, i.e., $W > \tilde{W}$, then the contention in the channel is light already. In this case, the MTD should not drop the request but try more. Thus, a larger K is preferred for improving p. However, as W decreases, i.e., $W < \tilde{W}$, the contention becomes severe as retransmissions of each MTD become more frequent. In this case, a smaller retransmission limit K could relieve the congestion via packet dropping and benefit the network performance. The threshold $\tilde{W}$, on the other hand, only depends on the input rate of each MTD λ, and increases as λ decreases.

C. Simulation Results

In this subsection, simulation results are presented to validate the above analytical results. Note that the simulation setting is in line with the system configuration as elaborated in Section II. The total number of time slots for each simulation run is 10^8. The curves are analytical results obtained from the fixed-point equation in (12), while the symbols demonstrate the simulation results, which is obtained by calculating the ratio of the number of successful access requests to the total number of transmitted access requests.

Fig. 3 illustrates how the preamble retransmission limit K and the UB window size W affect the limiting probability of successful transmission of access requests, p, with the number of MTDs $n = 100$ and the aggregate input rate $\hat{\lambda} = 0.4$. From Fig. 3, it is observed that when $W = 1000$, i.e., $W > \tilde{W} = 500$, the single steady-state point p_L increases with K; When $W = 30$ or 100, i.e., $W < \tilde{W}$, the steady-state points decrease with K. Intuitively, as the preamble retransmission limit K increases, each request has more chances to be retransmitted, which intensifies the channel competition. Thus, we can see from Fig. 3 that when K is large and W is small, two steady-state points, namely, p_L and p_A, exist in the network. It may move from p_L to p_A, on which the access performance is poor. Moreover, a closer look at Fig. 3 a also suggests that both p_L and p_A will be insensitive to the increment of preamble retransmission limit K when K is large. On the other hand, if $K = 1$, as shown in Fig. 3 b, the network has a single steady-state point p_L, which is independent of the UB window size W, according to (13). Finally, we can conclude that simulation results are in tight agreement with the analytical results.

IV. Access Throughput

In this section, the optimal access throughput performance with finite preamble retransmission limit is characterized. Let $\hat{\lambda}_{out}$ denote the access throughput, which measures the time-average of the number of successfully received access requests. With the $Geo/G/1/1$ queueing model, the expression of the access throughput can be obtained as follows [37]

$$\hat{\lambda}_{out} = \hat{\lambda}(1-\rho)(1-p_d) = \frac{\hat{\lambda}(1-p_d)}{1+\frac{\hat{\lambda}}{n}+\frac{\hat{\lambda}(W+1)(1-p-p_d)}{2np}}, \quad (14)$$

by combining (5)–(7). In particular, with the preamble retransmission limit $K = 1$, we have $\hat{\lambda}_{out} = -\mathbb{W}_0(\frac{n\lambda}{1+\lambda})$, which is only determined by the number of MTDs and the traffic input rate of each MTD.

A. Maximum Access Throughput

When the preamble retransmission limit $K > 1$, it is shown in (6) and (14) that the access throughput $\hat{\lambda}_{out}$ further relies on K and the UB window size W. In the following, we only consider the case that $K > 1$ and address how to maximize the access throughput $\hat{\lambda}_{out}$ by properly choosing W. Define the maximum access throughput as $\hat{\lambda}_{\max} = \max_W \hat{\lambda}_{\text{out}}$. The following theorem presents $\hat{\lambda}_{\max}$ and the corresponding optimal UB backoff window size W^*.

Theorem 1: The maximum access throughput $\hat{\lambda}_{\max} = e^{-1}$ is achieved if and only if the following two criteria are satisfied, i.e., the network operates at p_L, and the backoff window size W

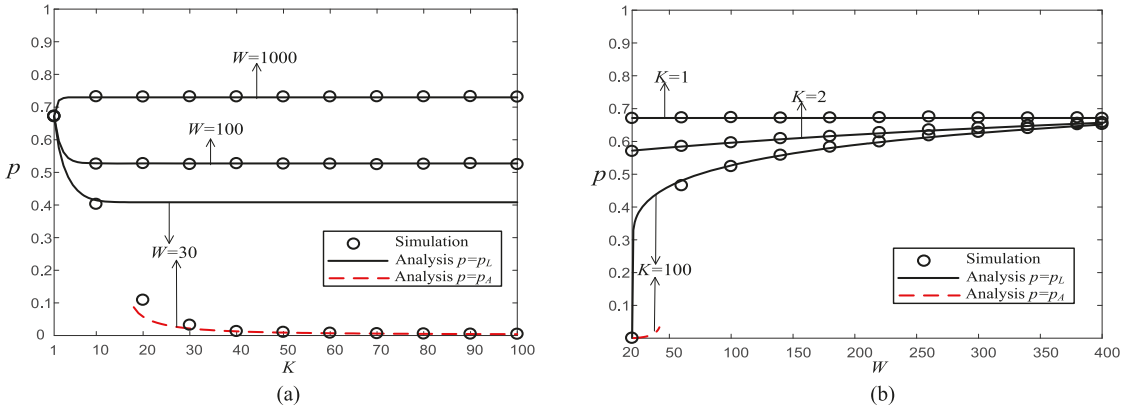


Fig. 3. Limiting probability of successful transmission of access requests, p, versus preamble retransmission limit K and the UB window size W. $n = 100$. $\hat{\lambda} = 0.4$. $M = 1$. (a) p versus K with $W \in \{30100, 1000\}$. (b) p versus W with $K \in \{1, 2100\}$.

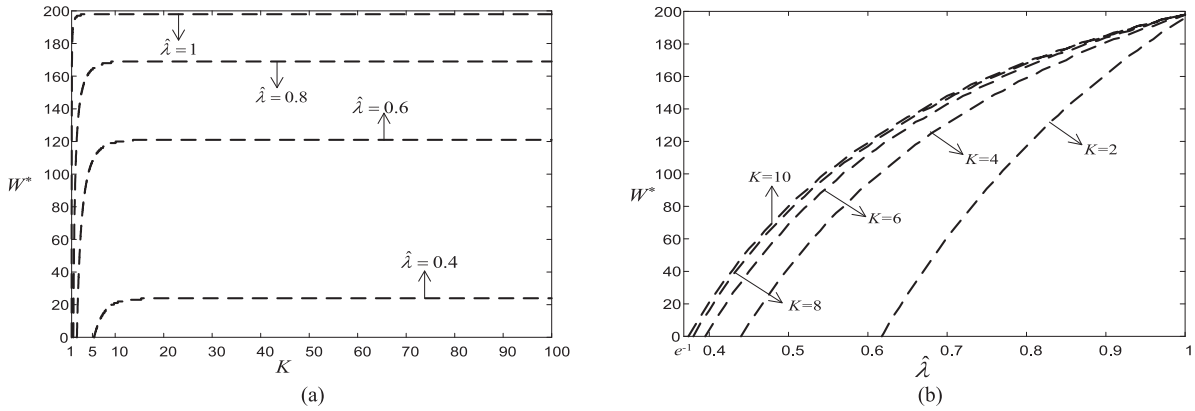


Fig. 4. Optimal backoff window size W^* versus preamble retransmission limit K and aggregate input rate $\hat{\lambda}$. $n = 100$. $M = 1$. (a) W^* versus K. (b) W^* versus $\hat{\lambda}$.

is set as

$$W^* = \frac{\hat{\lambda}\left(e^{-1}+1-(1-e^{-1})^K\right)+2n\left(e^{-1}+\hat{\lambda}(1-e^{-1})^K-\hat{\lambda}\right)}{\hat{\lambda}(e^{-1}-1+(1-e^{-1})^K)}, \text{ for } K > 1. \tag{15}$$

Proof: See Appendix B. ■

From Theorem 1, we can see that the maximum access throughput $\hat{\lambda}_{\max}$ is independent of the preamble retransmission limit K, but the optimal UB window size W^* should adaptively change with K. Fig. 4 illustrates how W^* varies with K and the aggregate input rate $\hat{\lambda}$. We can see from Fig. 4 that W^* monotonically increases with both K and $\hat{\lambda}$. Similar to the network steady-state point p shown in Fig. 3, the optimal UB window size W^* will become insensitive to the preamble retransmission limit K when K is large.

Fig. 4 b also shows that for given preamble retransmission limit K, if $\hat{\lambda}$ is too small, then the analytical result of W^* drops to zero and probably even less, which conflicts with the fact that $W \geq 1$. It indicates that in this case, the network cannot achieve the maximum access throughput $\hat{\lambda}_{\max}$. Moreover, Theorem 1 shows that $\hat{\lambda}_{\max}$ can be achieved only when the network operates at p_L. To ensure that the network has only one steady-state point p_L, Fig. 2 reveals that a sufficiently large aggregate input rate $\hat{\lambda}$ is necessary. To guarantee that the maximum access throughput $\hat{\lambda}_{\max}$ is achievable, the following corollary shows that $\hat{\lambda}$ should be no smaller that 1.

Corollary 2: Given that the aggregate input rate $\hat{\lambda} \geq 1$ for $K > 1$, it is warranted that the maximum access throughput $\hat{\lambda}_{\max}$ is attainable with $W = W^*$.

Proof: See Appendix C. ■

Note that it has been revealed in [31] that with $K = +\infty$, the network can achieve the maximum access throughput with the optimal tuning of backoff parameters when the aggregate input rate is larger than $4e^{-2} \approx 0.541$, which is smaller than 1 in the case of $K < +\infty$. Intuitively, with finite preamble retransmission limit, many access requests are likely to be dropped. To push the access throughput performance to the limit, the aggregate input rate should therefore be sufficiently large to offset the packet loss due to dropping.

B. Simulation Results

Numerical simulations are conducted in this subsection to validate above analytical results. Fig. 5 demonstrates how $\hat{\lambda}_{out}$ varies with W, wherein we set $\hat{\lambda} = 0.4$ or 1 and $K = 2$ or 20. We can see from Fig. 5 a that with $\hat{\lambda} = 0.4$ and $K = 20$, the optimal backoff window size $W^* = 23$. However, even with $W = W^*$, the maximum access throughput $\hat{\lambda}_{\max}$ is not attainable due to the

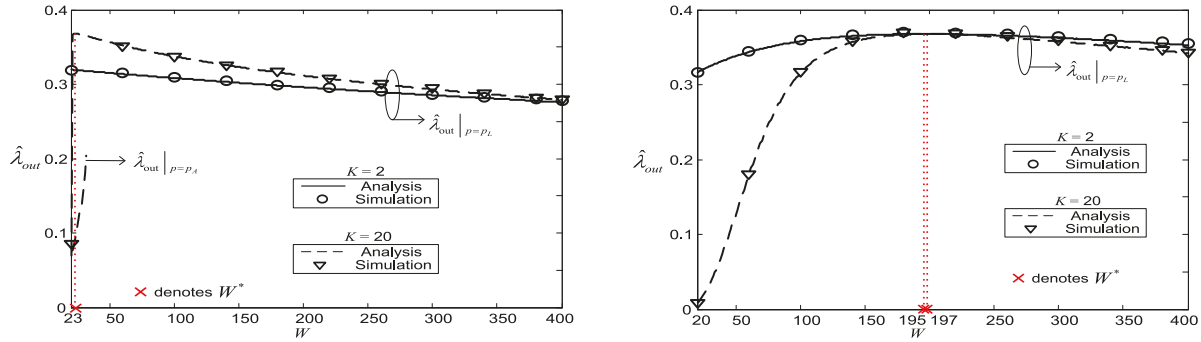


Fig. 5. Access throughput $\hat{\lambda}_{out}$ (in unit of requests/slot) versus backoff window size W. $n = 100$. $M = 1$. $K = 2$ or 20. (a) $\hat{\lambda} = 0.4$. (b) $\hat{\lambda} = 1$.

fact that the network degenerates into the undesired steady-state point p_A. We can also see that with $K = 2$, only one steady-state point exists in the network and the access throughput decreases as W increases. Thus, with $W = 1$, the access throughput can be maximized, which, however, is sill lower than e^{-1}. When the aggregate input rate increases to 1, as Corollary 2 indicates, the maximum access throughput $\hat{\lambda}_{\max}$ is always achievable for $K > 1$. Therefore, it can be seen from Fig. 5 b that with $W = W^*$, the maximum access throughput is obtained even when the preamble retransmission limit $K = 2$.

V. Access Delay

In this section, we will further evaluate how the preamble retransmission limit K affects the access delay performance by deriving the moments of access delay of successfully-transmitted access requests. The access delay is defined as the time length between the generation and the successful transmission of an access request.

For an access request, let us denote Y_i as its holding time length in State R_i; D_T as the time length spent from the beginning of State T until it leaves the queue (i.e., it is either dropped or successfully transmitted), D_i for $i = 1, 2, \ldots, K-1$ as the time length from the beginning of State R_i until it leaves the queue. According to Fig. 1, we have

$$D_T = \begin{cases} 1 & \text{with probability } p, \\ D_1 & \text{with probability } 1-p, \end{cases} \tag{16}$$

and

$$D_i = \begin{cases} Y_i + 1 & \text{with probability } p, \\ Y_i + D_{i+1} & \text{with probability } 1-p, \end{cases} \tag{17}$$

$i = 1, 2, \ldots, K-1$, and

$$D_K = 1. \tag{18}$$

Note that D_T is the sojourn time of an access request in the system. Let $G_{D_T}(z)$ denote its Probability Generating Function (PGF) and $G_{Y_i}(z)$ the PGF of Y_i. It can be derived from (16)–(18) that

$$G_{D_T}(z) = pz + pz\sum_{i=1}^{K-1}(1-p)^i \prod_{j=1}^{i} G_{Y_j}(z) + (1-p)^K z \prod_{j=1}^{K-1} G_{Y_j}(z). \tag{19}$$

Since

$$G_{Y_j}(z) = \frac{z - z^{W+1}}{W(1-z)}, \tag{20}$$

and $G'_{Y_i}(1) = \frac{W+1}{2}$ for $i = 1, 2, \ldots, K-1$, we can further obtain the mean sojourn time $E[D_T]$ as

$$E[D_T] = G'_{D_T}(1) = 1 + \frac{W+1}{2}\left(\frac{1-p-p_d}{p}\right). \tag{21}$$

from (19). Note that the access request leaves the system because of two reasons: 1) it is dropped, or 2) it is successfully transmitted. Let D_d and D_s denote the sojourn time of a dropped access request in the system and that of a successfully-transmitted access request, respectively. With the access request dropping probability p_d, we have

$$E[D_T] = p_d E[D_d] + (1-p_d)E[D_s]. \tag{22}$$

An access request is dropped after Kth transmission failure. Therefore, the corresponding sojourn time is then given by

$$D_d = 1 + \sum_{i=1}^{K-1} Y_i. \tag{23}$$

with the probability generating function

$$G_{D_d}(z) = z\prod_{j=1}^{K-1} G_{Y_j}(z). \tag{24}$$

The mean access delay of successfully-transmitted access requests $E[D_s]$ can then be obtained from (19)–(24) as

$$E[D_s] = \frac{1}{1-p_d}\left(1 + \frac{W+1}{2}\left(\frac{1-p-p_d}{p}\right) - p_d\left(1 + \frac{(K-1)(W+1)}{2}\right)\right), \tag{25}$$

and the probability generating function of D_s

$$G_{D_s}(z) = \frac{1}{1-p_d}\left(pz + pz\sum_{i=1}^{K-1}\left((1-p)\left(\frac{z-z^{W+1}}{W(1-z)}\right)\right)^i\right). \tag{26}$$

Fig. 6 a presents the simulation results on how the mean access delay of successfully-transmitted access requests $E[D_s]$ varies with the UB backoff window size W when the aggregate input rate $\hat{\lambda} = 1$ and the preamble retransmission limit $K = 2$, 20 or

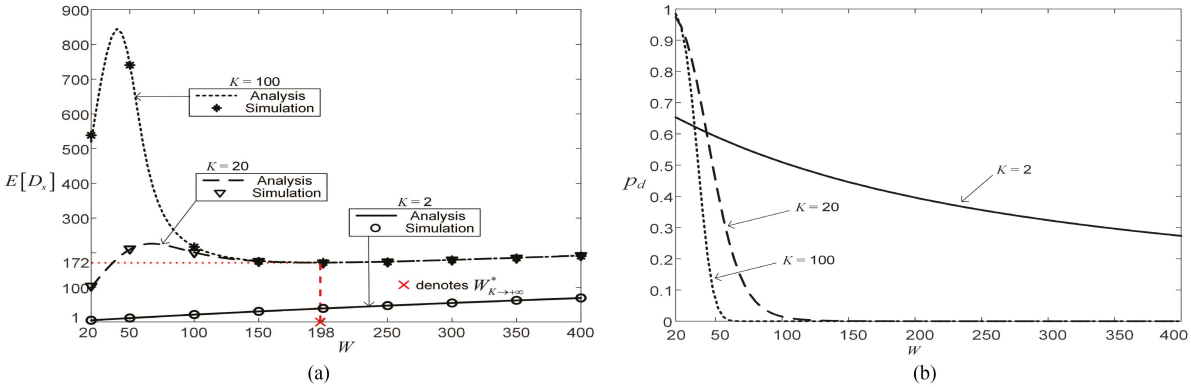


Fig. 6. (a) Mean access delay of successfully-transmitted access requests $E[D_s]$ (in unit of slots) versus backoff window size W. (b) Packet dropping probability p_d versus backoff window size W. $n = 100$. $M = 1$. K =2, 20 or 100. $\hat{\lambda} = 1$.

100. We can clearly see that there are different trends among cases where $K = 2$, 20 or 100. Let us first focus on the case where $K = 2$. Specifically, as shown in Fig. 3 b, with $K = 2$, the steady-state point p is insensitive to the variation of W. According to (6) and (25), the mean access delay $E[D_s]$ can then be approximately regarded as a linearly increasing function of W. As a result, we can see from Fig. 6 a that $E[D_s]$ linearly increases with W when $K = 2$.

In sharp contrast, as shown in Fig. 6 a, if the preamble retransmission limit K increases to 20 or 100, then the variation of $E[D_s]$ with regards to the window size W becomes complicated. Specifically, when W is small, i.e., $W = 20$, the packet dropping probability p_d approaches 1, as shown in Fig. 6 b. The channel contention can be greatly relieved, leading to small mean access delay $E[D_s]$, as shown in Fig. 6 a. If the UB window size W increases, then the packet dropping probability p_d drops quickly, which implies that a large number of devices would contend for channel access. The mounting channel contention boosts $E[D_s]$. Yet, with further growth of W, e.g., $W \approx 70$, $E[D_s]$ drops. Finally, if W is excessively large, i.e., $W \approx 300$, then the access request has to wait a long time before the next transmission once it encounters a collision. Therefore, we can see from Fig. 6 a that the mean access delay $E[D_s]$ grows with the window size W again.

The results in Fig. 6 also reveal that the preamble retransmission limit K determines a crucial tradeoff between the reliability performance and the access delay performance. As shown in Fig. 6 a, $E[D_s]$ with $K = 2$ is much lower than $E[D_s]$ with $K = 100$, and can further be reduced as W declines. Particularly, with $K = 1$, we have $E[D_s]_{K=1} = 1$ according to (6), (13) and (25). It suggests that a small retransmission limit K can improve the delay performance. Yet, the performance improvement is achieved at the cost of reliability, i.e., the probability that the access request is successfully delivered rather than dropped. We can see from Fig. 6 that with $K = 2$, the packet dropping probability p_d is still high even when W is large, indicating that a large proportion of access requests are dropped.

Note that if the preamble retransmission limit K is large, e.g., $K \to +\infty$, then it can be obtained from (14) and (25) that

$$E[D_s]_{K\to+\infty} = 1 + \frac{(W+1)(1-p)}{2p} = \frac{n}{\hat{\lambda}_{out,K\to+\infty}} - \frac{1}{\hat{\lambda}}, \quad (27)$$

which shows that minimizing the mean access delay $E[D_s]_{K\to+\infty}$ is equivalent to maximizing the access throughput $\hat{\lambda}_{out,K\to+\infty}$. In such case, the optimal backoff window size $W^*_{K\to+\infty}$ for optimizing the throughput and delay performance can be derived as

$$W^*_{K\to+\infty} = \frac{2\left(n - \frac{ne^{-1}}{\hat{\lambda}} - 1\right)}{1-e^{-1}} + 1, \quad (28)$$

according to Theorem 1. We can observe from Fig. 6 a that with $K = 100$, the minimum mean access delay of successfully-transmitted access requests $E[D_s]_{\min} = 172$ is achieved with $W^*_{K\to+\infty} = 198$.

VI. Performance Optimization in Multi-Preamble Scenario

So far, we have analyzed the effect of preamble retransmission limit K on the access throughput and access delay performance when the number of preamble $M = 1$. In this section, we will further focus on the multi-preamble scenario $M > 1$ and study how to properly select the number of preambles for optimizing the access efficiency and the preamble resource utilization ratio.

A. Performance Analysis in Multi-Preamble Scenario

Specifically, a multi-group model was proposed in [31] based on the orthogonality among preambles, that is, MTDs who use different preambles would not interfere with others. In the multi-group model, MTDs are classified into M groups based on their selected preambles. Denote $n^{(i)}$ as the group size of the i-th group, where $i = 1, 2, \ldots, M$ and $\sum_{i=1}^{M} n^{(i)} = n$. According to the standards [5], MTDs randomly select preambles in each access request transmission attempt. Therefore, in the long run, $n^{(i)}$ can be approximated as $\frac{n}{M}$, i.e., $n^{(i)} \approx \frac{n}{M}$. Likewise, we have $\hat{\lambda}^{(i)} \approx \frac{n\lambda}{M}$. Replacing n and $\hat{\lambda}$ by $n^{(i)} \approx \frac{n}{M}$ and $\hat{\lambda}^{(i)} \approx \frac{n\lambda}{M}$, respectively, the fixed-point equation in the multi-preamble case can be written as

$$p^{(M)} = \exp\left(-\frac{\frac{\lambda n}{M}(1-p_d)}{p^{(M)} + \lambda p^{(M)} + \frac{\lambda(W+1)(1-p^{(M)})(1-p_d)}{2}}\right), \quad (29)$$

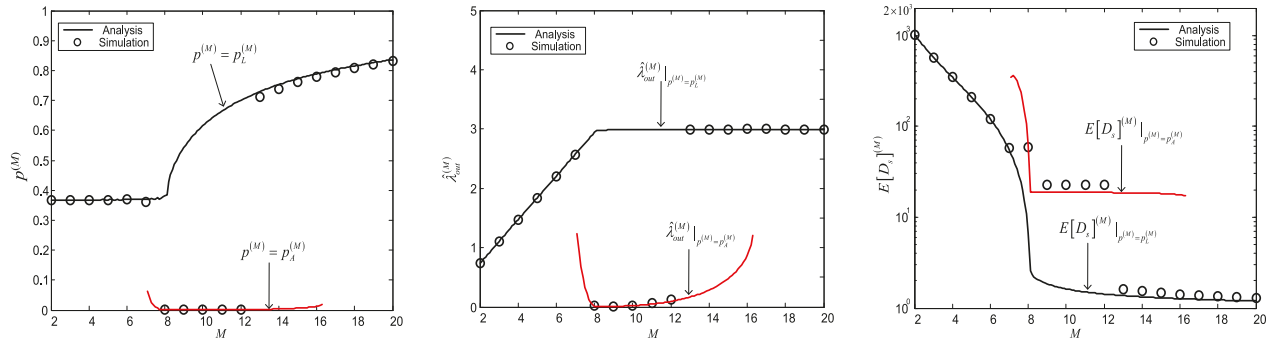


Fig. 7. Probability of successful transmission of access requests $p^{(M)}$, access throughput $\hat{\lambda}_{out}^{(M)}$ (in unit of requests/slot) and mean access delay of successfully-transmitted access requests $E[D_s]^{(M)}$ (in unit of slots) versus the number of preambles M. $n = 1000$. $\lambda = 0.003$. $W = \max\{1, W^{*,(M)}\}$. $K = 30$.

and the access throughput as

$$\hat{\lambda}_{out}^{(M)} = \hat{\lambda}M(1-\rho)(1-p_d) = \frac{\hat{\lambda}(1-p_d)}{1+\lambda+\frac{\lambda(W+1)(1-p^{(M)}-p_d)}{2p^{(M)}}}, \tag{30}$$

which is maximized at

$$\hat{\lambda}_{\max}^{(M)} = Me^{-1}, \tag{31}$$

only if the backoff window size is set to be

$$W^{*,(M)} = \frac{\frac{n\lambda}{M}\left(e^{-1}+1-(1-e^{-1})^K\right)+2n\left(e^{-1}+\frac{n\lambda}{M}(1-e^{-1})^K-\frac{n\lambda}{M}\right)}{\frac{n\lambda}{M}(e^{-1}-1+(1-e^{-1})^K)}. \tag{32}$$

Meanwhile, the mean access delay of successfully-transmitted access requests in multi-preamble cases should be given by

$$E[D_s]^{(M)} = \frac{1}{1-p_d}\left(1+\frac{W+1}{2}\left(\frac{1-p^{(M)}-p_d}{p^{(M)}}\right)-p_d\left(1+\frac{(K-1)(W+1)}{2}\right)\right). \tag{33}$$

Fig. 7 shows how $p^{(M)}$, $\hat{\lambda}_{out}^{(M)}$ and $E[D_s]^{(M)}$ vary with the number of preambles M with the backoff window size $W = \max\{1, W^{*,(M)}\}$.[1] We see that when M is small, with the optimal tuning of W, the steady-state point $p^{(M)}$ operates at e^{-1}, the access throughput linearly increases with M because $\hat{\lambda}_{\max}^{(M)} = Me^{-1}$, and the access delay drops quickly. However, when M is moderately large, e.g., $M = 8$, the bistable behavior can be observed from Fig. 7 a, where the network has two steady-state points, i.e., $p_L^{(M)}$ and $p_A^{(M)}$, and the network shifts to the undesired steady-state point $p_A^{(M)}$. Under such circumstance, as can be observed from Fig. 7 b-c, the access throughput $\hat{\lambda}_{out}^{(M)}$ would be close to 0 and the mean access delay $E[D_s]^{(M)}$ rises to be large. If M further grows, the network again has a single steady-state point $p_L^{(M)}$. Besides, $\hat{\lambda}_{out}^{(M)}$ is getting close to the total input rate, i.e., $\hat{\lambda} = 3$.

In particular, we can see from Fig. 7 that when the number of preambles M is large, allocating more preambles makes marginal contribution to the throughput and delay performance. Without the limitation on the number of preambles M, it is intuitively clear that the network should allocate preambles as many as possible for machine-type services. However, according to the standards [4], the total number of preambles is limited, e.g., only 64. Note that to ensure the quality-of-service of the traditional Human-to-Human (H2H) services, a certain amount of preambles has to be reserved, indicating that the number of preambles for machine-type services would be very limited. The scarcity of the preamble resource necessitates the study on how to optimize the access efficiency of mMTC in cellular networks with least number of preambles.

[1]When M is large, $\frac{n\lambda}{M}$ would be small and $W^{*,(M)} < 0$ according to (32). Thus, we let $W = \max\{1, W^{*,(M)}\}$ to ensure that $W > 0$ for all possibilities of M.

B. Optimal Number of Preambles

This subsection is devoted to addressing the above issue, where the major focus is on the throughput performance. Specifically, we are interested in maximizing the access throughput and the preamble resource utilization ratio, i.e., access throughput per preamble. The maximum access throughput in the single-preamble scenario has been explicitly shown in Theorem 1. Equivalently, the maximum preamble resource utilization ratio can be obtained as $\frac{\hat{\lambda}_{out}^{(M)}}{M} = e^{-1}$. Accordingly, the optimization problem that we aim to address can be formulated as follows

$$M^* = \underset{\{1\le M\le M_{\max}\}}{\arg\max} \ \hat{\lambda}_{out}^{(M)}. \quad \text{s.t.} \ \frac{\hat{\lambda}_{out}^{(M)}}{M} = e^{-1}, \tag{34}$$

where $M_{\max} \in \mathbb{N}^+$ denotes the maximum number of preambles that the network can allocate for machine-type services, and M^* is referred to as the optimal number of preambles.

Note that it has been shown in Theorem 1 and Fig. 7 that the network has to operate at the desired steady-state point to attain the optimal throughput performance. However, the roots of the fixed-point equation as given in (29) can only be determined via numerical calculation. It is therefore difficult, if not impossible, to characterize the explicit expression of M^*. In this regard, we propose an exhaustive search algorithm in Algorithm 1 for determining M^*. The basic idea is to calculate

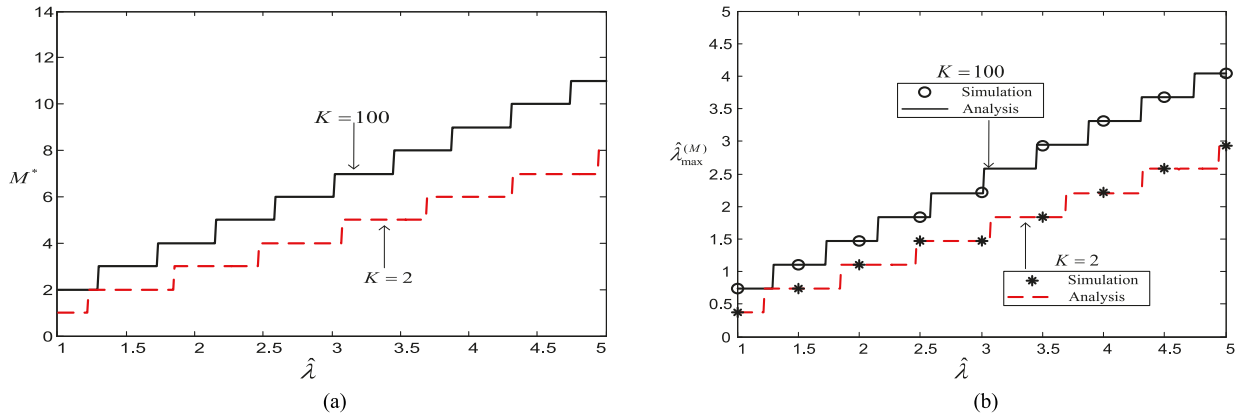


Fig. 8. Optimal number of preambles M^* and the corresponding maximum access throughput $\hat{\lambda}_{\max}^{(M)}$ (in unit of requests/slot) versus the aggregate input rate $\hat{\lambda}$. $n = 1000$. $K = 2$ or 100. $W = W^{*,(M)}$. (a) M^* versus $\hat{\lambda}$. (b) $\hat{\lambda}_{\max}^{(M)}$ versus $\hat{\lambda}$.

Algorithm 1: Calculation of M^*.

Input $n, \lambda, K, M_{\max}$ and initialize $M = 1$ and $\hat{\lambda}_{out}^{(0)} = 0$.
repeat
  Calculate $p^{(M)}$ in (29) with $W = W^{*,(M)}$.
  If $p_A^{(M)}$ does not exist **then**
    Calculate $\hat{\lambda}_{out}^{(M)}$ in (30) with $W = W^{*,(M)}$.
    If $\hat{\lambda}_{out}^{(M)} > \hat{\lambda}_{out}^{(M-1)}$ and $\frac{\hat{\lambda}_{out}^{(M)}}{M} = e^{-1}$ **then**
      $M^* = M$.
    endif
  endif
  $M = M + 1$.
until $M > M_{\max}$
Output M^*.

the access throughput $\hat{\lambda}_{out}^{(M)}$ with $W = W^{*,(M)}$ for each value of M in $\{1, 2, \ldots, M_{\max}\}$, and finally select the value of M which guarantees that the network has a single steady-state point $p_L^{(M)}$ with $W = W^{*,(M)}$ and $\frac{\hat{\lambda}_{out}^{(M)}}{M} = e^{-1}$. Take the case in Fig. 7 b as an example, we can see that the optimal number of preambles $M^* = 7$ if $M_{\max} \geq 7$.

Fig. 8 demonstrates how the optimal number of preambles M^* and the corresponding maximum access throughput $\hat{\lambda}_{\max}^{(M)}$ vary with $\hat{\lambda}$ with the preamble retransmission limit $K = 2$ or 100. Note that with $K = 2$, as shown in Fig. 6 b, lots of requests would be dropped, in which case the system should allocate relatively small amount of preamble resources. Therefore, as shown in Fig. 8, given the aggregate input rate $\hat{\lambda}$, the optimal number of preambles M^* with $K = 2$ is smaller than that with $K = 100$. Moreover, as $\hat{\lambda}$ grows, we can observe from Fig. 8 that M^* increases in a stair-like manner, which indicates that only if the aggregate input rate grows sufficiently large will the system allocate one more preamble. The corresponding maximum access throughput $\hat{\lambda}_{\max}^{(M)}$, which is a linear function of M^* according to (31), also increases with step size of e^{-1}, implying that every single preamble that the system allocates makes its best contribution to the growth of the access throughput.

VII. Implementation to Practical 5 G System Design

The analysis presented in this paper sheds important light on the practical 5 G network design for supporting massive random access of mMTC. In the current 5 G standard [39], [40], various optional values of the UB window size W, the number of preambles M and the preamble retransmission limit K are listed, yet, without specifying how to properly configure them. As Figs. 5-7 demonstrate, the access performance severely degrades if the network configuration is improper. To optimize the access performance, the backoff parameter and the number of preambles need to be adaptively tuned according to the number of MTDs, the aggregate traffic input rate and the preamble retransmission limit. The optimal settings are presented in this paper, and verified by simulation results.

Fig. 9 shows how to implement the optimal configuration in practical 5 G networks. Specifically, in a single-cell scenario, all MTDs communicate with one gNB. The gNB easily keeps a record of registered MTDs by analyzing their MAC addresses.[2] Each MTD can report its input rate to the gNB through RRC message, such as the RRC request in the random access procedure. Given the number of MTDs and the input rate of each MTD, the gNB can obtain the optimal settings of backoff parameter and the number of preambles based on the analysis above. Finally, the gNB can broadcast the number of preambles via the system information block and the backoff indicator, i.e., the UB backoff window size, via the random access response. When either the number of MTDs or the input rate of each MTD changes, the gNB can recalculate the optimal setting, and broadcast the configuration information again.

We can see that no additional signaling overhead is introduced for implementing the optimal configuration. The main overhead may come from Algorithm 1. The computational complexity of Algorithm 1 is given by $\mathcal{O}(M_{\max})$. Since the maximum number of preambles that the network can allocate for machine-type

[2]It can be seen that the number of MTDs n defined in this paper is different from the number of backlogged MTDs considered in existing works [19]–[30].

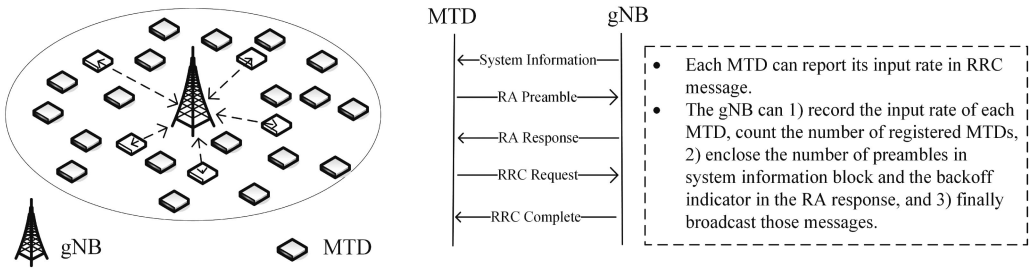


Fig. 9. Implementation of the optimal configuration in practical 5 G networks.

services M_{max} is usually 64 [4], Algorithm 1 wouldn't be computationally expensive and time-consuming. Therefore, the analytical results in this paper are promising to be used in practical networks.

VIII. Conclusion

In this paper, the effect of preamble retransmission limit K on the access efficiency of mMTC in cellular networks is evaluated. Towards this end, the state transition process for each access request is characterized, with which the network's steady-state points are obtained and their monotonicity with regards to the backoff window size and preamble retransmission limit K is identified. The analysis further shows how the maximum access throughput can be achieved, together with the corresponding optimal backoff window size. It reveals that only if the aggregate input rate is sufficiently large can the maximum access throughput is guaranteed to be achieved with optimal backoff window size, regardless of the value of the preamble retransmission limit K. Moreover, the explicit expression of the mean access delay is derived, which shows that the preamble retransmission limit K determines a crucially tradeoff between the reliability performance and the access delay performance. Besides the access efficiency optimization, we further take the preamble resource utilization ratio into consideration, and propose an algorithm for determining the optimal number of preambles. Finally, we demonstrate how to implement the optimal setting in practical systems.

Note that in this paper, we stipulate that the data queue of each individual MTD can be purged within one time slot. In other words, an infinite data transmission rate is implicitly assumed. However, in practice, when the data transmission resource is insufficient, a long time length is inevitable for the MTD to clear all packets in the data queue. In this connection, a finite data transmission rate is assumed in [41]. Yet, an infinite preamble retransmission limit $K = +\infty$ is also assumed therein. Accordingly, one interesting direction for the future study is to reveal the effect of the preamble retransmission limit K on access and data transmission performance of mMTC in cellular networks with a finite data transmission rate. Moreover, this paper only considers the mean access delay although the PGF of access delay has been given in (26). Based on (26), the Cumulative Distribution Function (CDF) of access delay can be obtained. Thus, another interesting direction for the future study is to evaluate the CDF of access delay, which would provide a more insightful view on the effect of retransmission limit K on the delay performance.

Appendix A
Proof of Corollary 1

Proof: This Appendix investigates the monotonicity of the steady-state points in terms of the UB backoff window size W and the preamble retransmission limit K. Let us start by defining

$$f(p) = -\ln p - \frac{\hat{\lambda}\left(1-(1-p)^K\right)}{p+\frac{\hat{\lambda}p}{n}+\frac{\hat{\lambda}(W+1)(1-p)(1-(1-p)^{K-1})}{2n}}. \tag{35}$$

We can see that $f(p) = 0$ and the fixed-point equation in (12) have the same non-zero roots. Let $f'(p)$ denote the derivative of $f(p)$ with regards to p. According to (12) and (35), we can have

$$\frac{\partial p}{\partial W} = -f'(p)\frac{\frac{\hat{\lambda}}{2n}\lambda(1-(1-p)^K)(1-p-(1-p)^K)}{\left(\frac{\hat{\lambda}}{2n}(W+1)(1-p-(1-p)^K)+\left(1+\frac{\hat{\lambda}}{2n}\right)p\right)^2}, \tag{36}$$

$$\frac{\partial p}{\partial K} = f'(p)\frac{(1-p)^K p\left(\frac{\hat{\lambda}}{2n}(W+1)-\left(1+\frac{\hat{\lambda}}{2n}\right)\right)\lambda\ln(1-p)}{\left(\frac{\hat{\lambda}}{2n}(W+1)(1-p-(1-p)^K)+\left(1+\frac{\hat{\lambda}}{2n}\right)p\right)^2}. \tag{37}$$

The following lemma shows that no matter the network operates at p_L or p_A, $f'(p) \leq 0$ always holds.

Lemma 1: For $p = p_L$ or p_A, $f'(p) \leq 0$.

Proof: Depending on the number of roots of $f(p) = 0$, we consider the following two scenarios :

- Suppose that $f(p) = 0$ has one single root p_L. In this case, we should have $\frac{\partial f(p)}{\partial p}|_{p=p_L} \leq 0$. Otherwise, if $\frac{\partial f(p)}{\partial p}|_{p=p_L} > 0$, then with the fact that $f(p)$ is a continuously differentiable function of $p \in (0,1)$, there should exist $\delta > 0$ such that for $p \in (p_L - \delta, p_L + \delta)$, $\frac{f(p)-f(p_L)}{p-p_L} > 0$. Equivalently, we have $f(p) > f(p_L) = 0$ for $p \in (p_L, p_L + \delta)$, and $f(p) < f(p_L) = 0$ for $p \in (p_L - \delta, p_L)$. Since $f(0) > 0$ and $f(1) < 0$, according to the Intermediate Value Theorem, $f(p) = 0$ should have another two roots, i.e., one in $(0, p_L - \delta]$ and another in $[p_L + \delta, 1)$, which results in a contradiction.
- Suppose that $f(p) = 0$ has three roots $p_A \leq p_S \leq p_L$. In this case, we should have $\frac{\partial f(p)}{\partial p}|_{p=p_A} \leq 0$ and $\frac{\partial f(p)}{\partial p}|_{p=p_L} \leq 0$.
  Otherwise, if $\frac{\partial f(p)}{\partial p}|_{p=p_A} > 0$, then there should exist $\delta > 0$ such that for $p \in (p_A - \delta, p_A + \delta)$, $\frac{f(p)-f(p_A)}{p-p_A} > 0$. Equivalently, we have $f(p) > f(p_A) = 0$ for $p \in (p_A, p_A + \delta)$, and $f(p) < f(p_A) = 0$ for $p \in (p_A - \delta, p_A)$. Since $f(0) > 0$, according to the Intermediate Value Theorem, $f(p) = 0$ should have at least one more

root in $(0, p_A - \delta]$, besides p_A, p_S and p_L, which results in a contradiction.

If $\frac{\partial f(p)}{\partial p}|_{p=p_L} > 0$, then there should exist $\delta > 0$ such that for $p \in (p_L - \delta, p_L + \delta)$, $\frac{f(p)-f(p_L)}{p-p_L} > 0$. Equivalently, we have $f(p) > f(p_L) = 0$ for $p \in (p_L, p_L + \delta)$, and $f(p) < f(p_L) = 0$ for $p \in (p_L - \delta, p_L)$. Since $f(1) < 0$, according to the Intermediate Value Theorem, $f(p) = 0$ should have at least one more root in $[p_L + \delta, 1)$, besides p_A, p_S and p_L, which results in a contradiction.

Therefore, we can conclude that $f'(p_A) \leq 0$ and $f'(p_L) \leq 0$.

By combining (36) and Lemma 1, we can obtain that $\frac{\partial p}{\partial W} > 0$; By combining (37) and Lemma 1, we can obtain that if $W > \tilde{W}$, $\frac{\partial p}{\partial K} > 0$, else, $\frac{\partial p}{\partial K} < 0$, where $\tilde{W}$ is given in Corollary 1. ■

APPENDIX B
PROOF OF THEOREM 1

Proof: According to (7), (12) and (14), the access throughput $\hat{\lambda}_{out}$ can be rewritten as $\hat{\lambda}_{out} = -p \ln p$, based on which we can see that $\hat{\lambda}_{\max} = e^{-1}$, achieved when $p^* = e^{-1}$. By substituting $p^* = e^{-1}$ into (12), the optimal backoff window size W^* in (15) can then be derived.

In the following, we prove that only when the network operates at the desired steady-state point p_L can $\hat{\lambda}_{\max}$ be guaranteed to be achieved with $W = W^*$. Specifically, according to (14), we can see that if both ρ and p_d decreases with p, then $\hat{\lambda}_{out}$ should grow with p. Note that based on (7), we can have the derivative of ρ over p

$$\rho' = \frac{2\hat{\lambda}(W+1)h(K)}{(1-p)(\lambda(W+1)((1-p)^K-1)+p(\lambda(W-1)-2))^2}, \tag{38}$$

where

$$h(K) = (1-p)^K(1-p+Kp) - 1 + p. \tag{39}$$

Moreover, the derivative of $h(K)$ over K can be obtained as

$$\begin{aligned} h'(K) &= (1-p)^K(p + (1-p+Kp)\ln(1-p)) \\ &\leq (1-p)^K(p + \ln(1-p)) < 0, \end{aligned} \tag{40}$$

for $p \in (0,1)$. Since $h(0) = 0$, we can have $h(K) \leq 0$ for $p \in (0,1)$ and further $\rho' < 0$ for $K \geq 1$ and $p \in (0,1)$. Accordingly, by combining (38) and (40), we can have $\rho' < 0$, implying that ρ decreases as p increases. On the other hand, based on (6), it is straightforward to have p_d decreases as p increases. Therefore, we can conclude that both ρ and p_d decreases with p, and then the access throughput $\hat{\lambda}_{out}$ grows with p according to (14). Thus, when the network has two steady-state points, p_L and p_A, if the maximum access throughput $\hat{\lambda}_{\max}$ is achieved at $p_A = e^{-1}$, then as $p_L > p_A$, the access throughput $\hat{\lambda}_{out}$ on p_L should be larger than $\hat{\lambda}_{\max}$, which results in a contradiction. Accordingly, $\hat{\lambda}_{\max}$ is guaranteed to be achieved with $W = W^*$ only if the network operates at p_L. ■

APPENDIX C
PROOF OF COROLLARY 2

Proof: Let us focus on $f(p)$ in (35), since $f(p) = 0$ and the fixed-point equation in (12) have the same non-zero roots. We can rewrite (35) as $f(p) = -\ln p - Q(p)$, where

$$Q(p) = \frac{\hat{\lambda}\left(1-(1-p)^K\right)}{p + \frac{\hat{\lambda}p}{n} + \frac{\hat{\lambda}(W+1)(1-p)(1-(1-p)^{K-1})}{2n}}. \tag{41}$$

In the following, it will be demonstrated that when $\hat{\lambda} > 1$ and $W = W^*$, then $Q(p)$ increases with p for $p \in (0,1)$. As $-\ln p$ decreases as p increases for $p \in (0,1)$, accordingly, we can have that $f(p) = 0$, i.e., the fixed-point equation of (12), has a single root p_L for $p \in (0,1)$, implying the network always operates on p_L and the maximum access throughput $\hat{\lambda}_{\max}$ can be certainly achieved, based on Theorem 1.

Specifically, the derivative of $Q(p)$ in regards to p can be derived as

$$Q'(p) = \frac{2n\hat{\lambda}\tilde{Q}(p)h(K)}{(1-p)\left((p(W-1)+((1-p)^K)-1)(W+1)\hat{\lambda}-2np\right)^2}, \tag{42}$$

where $h(K)$ is given in (39), and $h(K) \leq 0$ for $p \in (0,1)$, as shown in Appendix B, and $\tilde{Q}(p) = 2n + \hat{\lambda} - W\hat{\lambda}$. Therefore, we can conclude that if $\tilde{Q}(p) < 0$ for $p \in (0,1)$, then $Q'(p) > 0$ for $p \in (0,1)$, implying that $Q(p)$ is a monotonic increasing function of p for $p \in (0,1)$.

By substituting W^* in (15) into $\tilde{Q}(p)$, it can be obtained

$$\tilde{Q}(p)|_{W=W^*} = \frac{2e(e^K-(e-1)^K)(n(\hat{\lambda}-1)-\hat{\lambda})}{(e-1)^K e + e^K(1-e)}, \tag{43}$$

where $\tilde{Q}(p)|_{W=W^*} < 0$ if and only if $\hat{\lambda} > \frac{n}{n-1} \overset{\text{with a large } n}{\approx} 1$.

REFERENCES

[1] Y. Ni, L. Cai, J. He, A. Vinel, Y. Li, H. M. Jahromi, and J. Pan "Toward reliable and scalable internet of vehicles: Performance analysis and resource management," *Proc. IEEE*, vol. 108, no. 2, pp. 324–340, Feb. 2020.

[2] Cisco whitepaper, "Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022," Feb. 2019.

[3] 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.

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

[5] 3GPP TS 36.321 V15.8.0, "LTE; Evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification," Jan. 2020.

[6] L. Kleinrock and S. Lam, "Packet switching in a multiaccess broadcast channel: Performance evaluation," *IEEE Trans. Commun.*, vol. 23, no. 4, pp. 410–423, Apr. 1975.

[7] A. Carleial and M. Hellman, "Bistable behavior of Aloha-type systems," *IEEE Trans. Commun.*, vol. 23, no. 4, pp. 401–410, Apr. 1975.

[8] K. Sakakibara, H. Muta, and Y. Yuba. "The effect of limiting the number of retransmission trials on the stability of slotted ALOHA systems," *IEEE Trans. Veh. Technol.*, vol. 4, no. 49, pp. 1449–1453, Jul. 2000.

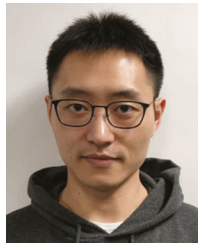
[9] K. Sakakibara, T. Seto, D. Yoshimura, and J. Yamakita, "Effect of exponential backoff scheme and retransmission cutoff on the stability of frequency-hopping slotted ALOHA systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 714–722, Jul. 2003.

[10] J. B. Seo, J. Hu, and V. C. M. Leung, "Impacts of retransmission limit on stability and throughput regions of S-ALOHA systems," *IEEE Trans. Veh. Technol.*, vol. 11, no. 66, pp. 10296–10306, Nov. 2017.

[11] S. W. Kim, "Frequency-hopped spread-spectrum random access with retransmission cutoff and code rate adjustment," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 10, pp. 344–349, Feb. 1992.

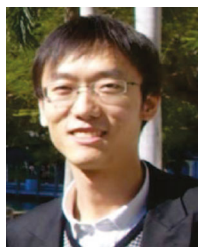
[12] R. R. Tyagi, F. Aurzada, K. D. Lee, and M. Reisslein, "Impact of retransmission limit on preamble contention in LTE-Advanced network," *IEEE Syst. J.*, vol. 9, no. 3, pp. 752–765, Sep. 2015.

[13] 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.
[14] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–376, Feb. 2016.
[15] 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.
[16] T. N. Weerasinghe, I. A. M. Balapuwaduge and F. Y. Li, "Preamble reservation based access for grouped mMTC devices with URLLC requirements," in. *Proc. IEEE ICC*, Shanghai, China, 2019, pp. 1–6.
[17] J. B. Seo and V. C. M. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.
[18] D. Niyato, P. Wang, and D. I. Kim, "Performance modeling and analysis of heterogeneous machine type communications," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2836–2849, Apr. 2014.
[19] H. Wu, C. Zhu, R. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.
[20] J. Choi, "On the adaptive determination of the number of preambles in RACH for MTC," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1385–1388, Jul. 2016.
[21] 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.
[22] 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.
[23] T. M. Lin, C. H. Lee, J. P. Cheng, and W. T. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 5, no. 63, pp. 2467–2472, Jan. 2014.
[24] Y. C. Pang, S. L. Chao, G. Y. Lin, and H. Y. Wei, "Network access for M2M/H2H hybrid systems: A game theoretic approach," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 845–848, May 2014.
[25] M. Koseoglu, "Pricing-based load control of M2M traffic for the LTE-A random access channel," *IEEE. Trans. Commun.*, vol. 65, no. 3, pp. 1353–1365, Mar. 2017.
[26] S. K. Sharma and X. Wang, "Collaborative distributed Q-learning for RACH congestion minimization for LTE-A networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, Apr. 2019.
[27] Z. Jiang, S. Zhou and Z. Niu, "Distributed policy learning based random access for diversified QoS requirements," in. *Proc. IEEE ICC*, Shanghai, China, 2019, pp. 1–6.
[28] M. Vilgelm, S. R. Liñares, and W. Kellerer, "Dynamic binary countdown for massive IoT random access in dense 5G networks," *IEEE Trans. Veh. Technol.*, vol 6, no. 4, pp. 6896–6908, Aug. 2019.
[29] C. Di, B. Zhang, Q. Liang, S. Li, and Y. Guo, "Learning automata based access class barring scheme for massive random access in machine-to-machine communications," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6007–6017, Aug. 2019.
[30] Y. Sim and D. H. Cho,"Performance analysis of priority-based access class barring scheme for massive MTC random access," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5245–5252, Dec. 2020.
[31] 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.
[32] W. Zhan and L. Dai, "Access delay optimization of M2M communications in LTE networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1675–1678, Dec. 2019.
[33] O. Liber, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things: Technologies, Standards, and Performance*, New York, NY, USA: Academic Press, 2017.
[34] 3GPP TS 36.211 V16.2.0.0 "LTE; Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," Sep. 2020.
[35] L. Dai and X. Sun, "A unified analysis of IEEE 802.11 DCF networks: Stability, throughput and delay," *IEEE Trans. Mobile Comput.*, vol. 12, no. 8, pp. 1558–1572, Aug. 2013.
[36] L. Dai, "Stability and delay analysis of buffered aloha networks," *IEEE. Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2707–2719, Aug. 2012.
[37] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. Hoboken, NJ, USA: Wiley, 1998.
[38] X. Sun and L. Dai, "Performance optimization of CSMA networks with a finite retry limit," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5947–5962, Sep. 2016.
[39] 3GPP TS 38.321 V15.11.0, "5 G; NR; medium access control (MAC) protocol specification," Jan. 2021.
[40] 3GPP TS 38.331 V16.4.1, "5 G; NR; radio resource control (RRC); protocol specification," Apr. 2021.
[41] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Throughput optimization with a finite data transmission rate," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5749–5763, Dec. 2019.



Wen Zhan (Member, IEEE) 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, and the Ph.D. degree from the City University of Hong Kong, China, in 2019. He was a Research Assistant and Postdoc with the City University of Hong Kong. Since 2020, he has been with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, China, where he is currently an Assistant Professor. His research interests include Internet of things, modeling and performance optimization of next-generation mobile communication systems.

Xinghua Sun (Member, IEEE) received the Ph.D. degree from the City University of Hong Kong (CityU), Hong Kong, in 2013. In 2010, he was a Visiting Student with INRIA, France. In 2013, he was a Postdoctoral Fellow with CityU. From 2015 to 2016, he was a Postdoctoral Fellow with the University of British Columbia, Vancouver, BC, Canada. From July to August 2019, he was a Visiting Scholar with the Singapore University of Technology and Design, Singapore. From 2014 to 2018, he was an Associate Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. Since 2018, he has been an Associate Professor with Sun Yat-sen University, Guangzhou, China. His research interests include stochastic modeling of wireless networks and machine learning over networks. He was the Technical Program Committee Member for numerous IEEE conferences.



Xijun Wang (Member, IEEE) received the B.S. degree (with high honors) in communications engineering from Xidian University, Xi'an, China, in 2005 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in January 2012. He was an Assistant Professor from 2012 to 2015 and an Associate Professor from 2015 to 2018 with the School of Telecommunications Engineering, Xidian University. From 2015 to 2016, he was a Research Fellow with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore. He is currently an Associate Professor with Sun Yat-sen University, Guangzhou, China. He has authored or coauthored in several IEEE journals and conference proceedings in the areas of wireless networks and patents related to heterogeneous networks. His current research interests include UAV communications, edge computing, and age of information. He was the Technical Program Committee Member for numerous IEEE conferences. He was also served as the Publicity Chair of IEEE ICCC 2013 and the Technical Program Co-Chair of the Wireless Communications Systems Symposium, IEEE ICCC 2016. He was the recipient of the Best Paper Award from the IEEE ICCC 2013. He is currently a Reviewer for several IEEE journals. He was recognized as an Exemplary Reviewer of the IEEE Wireless Communications Letters in 2014. He is currently an Associate Editor for the IEEE ACCESS.

Yaru Fu (Member, IEEE) received the B.Eng. (Hons.) degree in telecommunication engineering from Northeast Electric Power University, Jilin City, China, in 2011, the M.Sc. (Hons.) degree in communication and information system from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014, and the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, in 2018. Then, she joined the Institute of Network Coding (INC), The Chinese University of Hong Kong, Hong Kong, as a Postdoctoral Research Assistant. From September 2018 to June 2020, she was a Research Fellow (Class 2) with the Singapore University of Technology and Design, Singapore. From February 2016 to May 2016, she was a Visiting Researcher with Telecom Paris Tech and Laboratory of Information, Networking and Communication Sciences. She was also an Intern with Nokia Bell Labs, Paris, France. She is currently a Research Assistant Professor with the School of Science and Technology, The Open University of Hong Kong, Hong Kong, China. Her research interests include intelligent wireless communications and networks, caching and recommendation system, Internet-of-Things, and URLLC.

Yitong Li received the B.Eng. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 2011 and 2016, respectively. He is currently an Assistant Professor with the School of Information Engineering, Zhengzhou University, Zhengzhou, China. His research interests include the performance evaluation and optimization of wireless random access networks.