Rate-Constrained Delay Optimization for Slotted Aloha

Yitong Li, Wen Zhan, *Member, IEEE*, and Lin Dai, *Senior Member, IEEE*

***Abstract*— Slotted Aloha provides a simple way for accommodating the massive access of Machine-to-Machine (M2M) communications. Yet, the delay performance of slotted Aloha has long been observed to significantly deteriorate as the network size grows. It is therefore important to study how to optimize the delay performance of slotted Aloha in a large-scale network. This paper focuses on the optimization of access delay of a buffered slotted Aloha network, where n nodes transmit to a common receiver in fading channels. Specifically, by deriving the closed-form expressions of the network steady-state points in both unsaturated and saturated conditions, the first and second moments of access delay of each packet are obtained as explicit functions of system parameters, and minimized by optimizing the transmission probability of each node. The analysis shows that to achieve the minimum mean access delay, the transmission probability of each node should be reduced as the network size increases, leading to a diminishing node data rate unless the information encoding rate is jointly optimized. The minimum mean access delay for a given data rate requirement is further characterized, and effects of key parameters such as the minimum required data rate for each node, the mean received signal-to-noise ratio of each packet and the number of nodes on the rate-constrained minimum mean access delay are discussed. The practical insights of the analysis are also demonstrated by taking the example of an LTE-M system with smart grid applications.**

***Index Terms*— Slotted Aloha, access delay, machine-to-machine (M2M) communications.**

I. Introduction

RANDOM access is a fundamental way for multiple users to share a common channel under distributed control, where each user independently determines when and how to access. Depending on whether sensing the channel or not before each user's transmission, random access schemes can be broadly divided into two categories: Aloha-based [1] and Carrier Sense Multiple Access (CSMA)-based [2], both of which have found wide applications in various wireless networks, such as cellular networks [3] and WiFi networks [4]. With the simplest slotted Aloha [1], for instance, each user transmits with a certain probability whenever it has packets in the buffer. It provides a simple solution for facilitating the massive access of Machine-to-Machine (M2M) communications that has attracted great attention in recent years, and is expected to play an instrumental role in next-generation data communication networks [5], [6].

Despite the simplicity, it has long been observed that the performance of slotted Aloha significantly deteriorates when the number of nodes is large. Specifically, with each node transmitting at a fixed probability, the number of access requests quickly increases with the network size, leading to a diminishing chance of success. Such performance degradation is especially significant for the massive access of M2M communications, where packets may be backlogged for an excessively long time before being successfully transmitted [7]. How to optimize the delay performance of slotted Aloha is therefore becoming an increasingly pressing issue, which is of crucial importance for its successful support for delay-critical services of M2M communications [8].

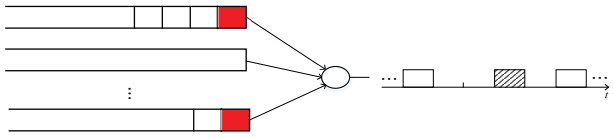


Fig. 1. An n-node buffered slotted Aloha network is essentially a multi-queue single-server system.

A. Delay Characterization of Buffered Slotted Aloha

As Fig. 1 illustrates, an n-node buffered slotted Aloha network is essentially a multi-queue-single-server system where each node has a buffer for accommodating incoming packets and competes for access. The queueing delay of each packet consists of two parts: 1) the waiting time, i.e., the time from arriving till being the head-of-line (HOL) packet, and 2) the service time, i.e., the time from being the HOL packet till being successfully transmitted. The service time, also referred to as the *access delay*, is crucially determined by nodes' aggregate activities, channel conditions, and receiver models. The queueing delay is further dependent on the arrival process.

Manuscript received October 17, 2020; revised March 1, 2021; accepted May 3, 2021. Date of publication May 12, 2021; date of current version August 16, 2021. The work of W. Zhan was supported in part by the National Natural Science Foundation of China under Grant 62001524, and in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant JCYJ20200828214622001. The work of Y. Li was supported by the National Natural Science Foundation of China (NSFC) under Grant 61801433. The work of L. Dai 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 N. Pappas. *(Corresponding author: Wen Zhan.)*

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

Wen Zhan is with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: zhanw6@mail.sysu.edu.cn).

Lin Dai is with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong (e-mail: lindai@cityu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3079513.

Digital Object Identifier 10.1109/TCOMM.2021.3079513

Early work focused on the access delay by ignoring the queueing behavior of each node. A representative model proposed in [9] is that each node has a one-packet-buffer that can be either in idle or backlogged states. When the number of nodes is large, the aggregate traffic, i.e., the number of newly arrived and retransmitted packets, can be approximated as a Poisson random variable, based on which the delay analysis can be greatly simplified [9]–[14]. Another simplification is to only consider the saturated condition [15], [16], where each user always has a packet to transmit, i.e., with zero idle probability. The mean access delay [9]–[12], [15], [17], the probability mass function of access delay [13], [14], [16], [18] and the probability generating function of access delay [19] have been characterized in various scenarios. Though well capturing the essence of contention from HOL packets, the effect of the arrival rate of packets on the access delay performance cannot be properly characterized based on the one-packet-buffer model. Moreover, by excluding the nodes' queues in the model, the analysis cannot be extended to further evaluate the queueing delay performance.

For a two-node buffered slotted Aloha network, the mean queueing delay of packets with Bernoulli arrivals was first characterized in [20] by deriving the generating function of the stationary joint queue length distribution. Closed-form expressions of the optimal transmission probability of nodes for minimizing the mean queueing delay were further obtained in [21], and [22] for a variant of Aloha with queue-aware transmissions. With more than two nodes, nevertheless, it is difficult to characterize the stationary joint queue length distribution. The focus of the analysis was then shifted to the queue length distribution of each *individual* node in the symmetric case, i.e., all the nodes have identical arrival processes and backoff parameters, and thus the same probability of successful/failed transmissions [23]–[27]. The numerical methods usually involve jointly solving multiple nonlinear equations. Due to the lack of closed-form expressions and high computational complexity, little light has been shed to the optimal tuning of system parameters for delay optimization.

In general, for performance analysis of a multi-queue-single-server system, the main challenge lies in the characterization of service process: With multiple nodes sharing the same server, the service time distribution is determined by their aggregate activities. As demonstrated in [28], the key to characterization of service process includes 1) proper modeling of HOL packets' behavior, as only the HOL packets of nodes' queues are involved in the service process, and 2) derivation of steady-state probability of successful transmission of HOL packets p, which is determined by the states of all the nodes. In [28], by establishing the fixed-point equations of p under unsaturated and saturated conditions, two network steady-state points were obtained as explicit functions of system parameters, based on which both the maximum network throughput and minimum mean access delay were further derived.

B. Our Contributions

The analysis in [28] was based on a few ideal assumptions that need to be further relaxed. First of all, the effect of noise was ignored in [28], and thus the received signal-to-noise ratio (SNR) of each packet was excluded from the analytical framework. Secondly, the channel gain was regarded as constant in [28], which is a good approximation for wired networks but fails to capture the random fluctuations of channel gains in wireless networks. Due to channel fading, a packet cannot be correctly decoded even without concurrent transmissions if the channel condition is too poor to support its information encoding rate [29].

In this paper, the delay analysis is extended from constant noiseless channels to fading channels, where the information encoding rate and mean received SNR of each packet are key parameters that have crucial impact on the delay performance. Note that channel fading was also considered in our recent work [30], [31], where the maximum sum rate, i.e., the maximum total number of successfully decoded information bits in unit time and unit bandwidth, of slotted Aloha was characterized under various receiver models including the capture model [30] and the successive interference cancellation (SIC) [31]. A key assumption in [30], [31] is that the network is saturated, i.e., each node always has packets in the buffer. With the objective of maximizing the sum rate, the saturated condition is of more interest, with which the network throughput is pushed to the limit. When the delay performance becomes the primary concern, however, it is crucial to include the unsaturated case, where the delay performance is significantly better than that in the saturated condition.

Specifically, we focus on the characterization and optimization of access delay of an n-node buffered slotted Aloha network where all the nodes transmit to a common receiver in fading channels. We consider the symmetric case, that is, all the nodes have identical input rates of packets, backoff parameters, and channel gain distribution. The network steady-state points in both unsaturated and saturated conditions are derived as explicit functions of key system parameters such as the information encoding rate and transmission probability of each node, based on which the minimum mean access delay and the corresponding optimal transmission probability of each node are further characterized.

The analysis shows that to achieve the minimum mean access delay, the transmission probability of each node should be reduced as the number of nodes increases. The corresponding node throughput also declines, indicating that the data rate of each node decreases if the information encoding rate is fixed. In practice, however, the data rate is an important performance metric in many M2M applications [32], where a minimum data rate for each node should be guaranteed. It is therefore of great practical importance to study how to optimize the mean access delay while satisfying a certain data rate requirement. For a given minimum required data rate of each node, the rate-constrained minimum mean access delay D_R^* is further obtained by jointly optimizing the transmission probability and the information encoding rate. It is shown that D_R^* sharply increases with the minimum required data rate R_0 when R_0 is large. The growth is particularly significant when the traffic input rate is low, in which case the information encoding rate of each node has to be sufficiently large to satisfy the rate requirement.

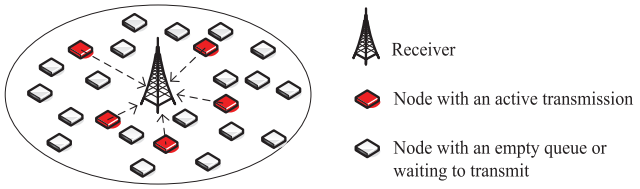


Fig. 2. With Aloha, each node transmits with a certain probability in each time slot when it has packets in its buffer.

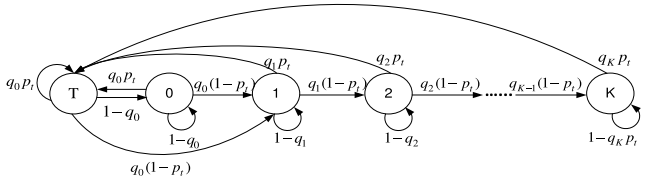


Fig. 3. State transition diagram of an individual HOL packet in slotted Aloha networks [30].

The analysis in this paper sheds important light on the practical system design for facilitating the massive access of M2M communications. For illustration, we take the LTE-M system with smart grid applications as an example and evaluate the rate-constrained minimum mean access delay of each packet under different traffic scenarios with distinct quality-of-service requirements. It is found that for delay-insensitive light traffic scenarios, LTE-M can support a large number of smart grid devices, e.g., more than 10^4 devices in a cell. However, if the quality-of-service requirement becomes strict, then the number of devices that LTE-M can support drastically decreases even when the mean received SNR is large.

The remainder of this paper is organized as follows. Section II presents the system model. The network steady-state points in the unsaturated condition and the saturated condition are characterized in Section III. In Section IV, the mean access delay at both steady-state points is derived and minimized by optimally tuning the transmission probability of each node. In Section V, the rate-constrained minimum mean access delay is characterized and the analysis is applied to an LTE-M network with smart grid applications. Conclusions are summarized in Section VI.

II. System Model

Consider a buffered slotted Aloha network where n nodes transmit to a single receiver over fading channels. Assume that all the nodes are synchronized and can start a transmission only at the beginning of a time slot. With Aloha, each node transmits with a certain probability in each time slot when it has packets in its buffer. Assume that each packet transmission lasts for one time slot. As Fig. 2 shows, for a given time, multiple nodes may have concurrent transmissions and interfere with each other.

Assume that each node is equipped with a buffer of infinite size to accommodate the arrival packets. For each node, assume that the input rate, i.e., the long-term average number of packets arrived in each time slot, is λ. As we mentioned in Section I, the key to performance analysis of a buffered slotted Aloha network is the characterization of its service process, which is crucially determined by aggregate activities of HOL packets, channel conditions and receiver design. In the following, we will present details on the HOL-packet model, channel model and receiver model.

A. HOL-Packet Model

For random-access networks, the performance closely depends on the transmission probabilities of nodes, which may change with time. Among numerous designs, a prevalent one is to adjust the transmission probability according to the number of transmission failures that a HOL packet has experienced, i.e., $q_i = q_0 \cdot \mathcal{Q}(i)$, where q_i is the transmission probability after the i-th failure, and the backoff function $\mathcal{Q}(i)$ is an arbitrary monotonic non-increasing function of the number of transmission failures i. To prevent q_i from being excessively small, a cutoff phase K is usually imposed, exceeding which the transmission probability does not vary with the number of transmission failures, i.e., $\mathcal{Q}(i)$ is constant when $i \geq K$. A backoff scheme can then be characterized by the sequence of transmission probabilities $\{q_i\}_{i=0,\ldots,K}$.

In [30], the behavior of each HOL packet in a slotted Aloha network was modeled as a discrete-time Markov process, as shown in Fig. 3. Specifically, a fresh HOL packet is initially in State T, and stays in State T if it is successfully transmitted. Otherwise, it moves to State 0 if it is not transmitted, or State 1 if its transmission fails. For a State-i HOL packet, it moves to State T if it is successfully transmitted. Otherwise, it stays in State i if it is not transmitted, or State $\min(K, i+1)$ if its transmission fails. q_i denotes the transmission probability of a State-i HOL packet, and p_t denotes the probability of successful transmission of HOL packets at time slot t.

Note that the steady-state probability distribution of the Markov chain in Fig. 3 exists only when $p = \lim_{t\to\infty} p_t$ exists. It has been obtained in [30] that the steady-state probability distribution is given by

$$\pi_T = \frac{1}{\sum_{i=0}^{K-1} \frac{(1-p)^i}{q_i} + \frac{(1-p)^K}{pq_K}}, \tag{1}$$

and

$$\pi_0 = \frac{1-q_0}{q_0}\pi_T, \quad \pi_K = \frac{(1-p)^K}{pq_K}\pi_T, \ \pi_i = \frac{(1-p)^i}{q_i}\pi_T, \tag{2}$$

for $i = 1, \ldots, K-1$, if $K \geq 1$. When the cutoff phase K is 0, States 0 and K in Fig. 3 merge into one state, i.e., State 0, and we have $\pi_0 = \frac{1-pq_0}{pq_0}\pi_T$. In this case, the transmission probability of each node is q_0 regardless of how many transmission failures it has experienced. For each node, π_T is the service rate of its queue as the queue has a successful output if and only if the HOL packet is in State T.

B. Channel Model

Let g_k denote the channel gain from node k to the receiver, which can be further written as $g_k = \gamma_k \cdot h_k$, where h_k is the small-scale fading coefficient of node k that varies from time slot to time slot and is modeled as a complex Gaussian random variable with zero mean and unit variance. The large-scale

fading coefficient γ_k characterizes the long-term channel effect such as path loss and shadowing.

Due to the differences in large-scale fading effects, the mean received signal-to-noise ratio (SNR) would vary from node to node if they adopt the same transmission power. For random-access networks, the near-far effect would cause severe unfairness among nodes [12], [30], [33]–[35]. That is, nodes with larger mean received power would have higher throughput. To ensure fairness, in this paper, we follow the assumption that uplink power control is performed to overcome the effect of large-scale fading,[1] i.e., the transmission power of each node is properly adjusted according to the large-scale fading coefficient γ_k such that each node has the same mean received SNR, denoted by ρ. With Rayleigh fading, the received SNR of each packet is exponentially distributed with mean ρ.

C. Receiver Model

Throughout the paper, we assume that the receiver always has perfect channel state information but the transmitters are unaware of the instantaneous realizations of the small-scale fading coefficients. As a result, each node independently encodes its information at a constant rate R_{in} bit/s/Hz. Assume that each codeword lasts for one time slot. That is, no joint decoding is performed among nodes' packets or with previously received packets.

As nodes do not coordinate their transmissions, more than one packets could be received within one time slot. Whether the packets can be successfully decoded or not crucially depends on the receiver model. In this paper, we consider the classical collision model, with which a packet transmission is unsuccessful as long as there are concurrent transmissions. When further taking the channel fading into consideration, a packet transmission could be unsuccessful even without concurrent transmissions if its received SNR is too low to support the encoding rate R_{in}. We assume that the codeword length is sufficiently large such that a single packet can be successfully decoded if its received SNR η satisfies $\log_2(1+\eta) \geq R_{in}$.[2] Let

$$\mu = 2^{R_{in}} - 1 \tag{3}$$

denote the SNR threshold. For each packet, its transmission is successful if and only if there are no concurrent transmissions and its received SNR $\eta \geq \mu$.

D. Access Delay, Throughput and Data Rate of Nodes

In this paper, we focus on the access delay performance. For each packet, its access delay is defined as the time interval from the instant that it becomes a HOL packet to its successful transmission. In Section IV, we will derive the probability generating function of the access delay based on the discrete-time Markov process of each HOL packet. As it is crucially determined by the steady-state probability of successful transmission of HOL packets p, in the following section, we will first establish the fixed-point equations of p to characterize the network steady-state points in unsaturated and saturated conditions. The mean access delay at both steady-state points will be derived and minimized by optimally tuning the transmission probability of nodes in Section IV.

Similar to access delay, the number of successfully decoded packets in each time slot is also a time-varying variable due to the lack of coordination among nodes. For random-access networks, the node throughput λ_{out}, which is defined as the long-term average number of successfully decoded packets per time slot per node, is another important performance metric. Let N_t denote the total number of successfully decoded packets in time slot t. The node throughput can then be written as $\lambda_{out} = \frac{1}{n}\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t} N_i$. As each packet has an information encoding rate of R_{in} bit/s/Hz, the long-term average received information rate per node can be written as

$$R_{out} = \frac{1}{n}\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t} N_i \cdot R_{in} = R_{in}\cdot\lambda_{out}. \tag{4}$$

We refer to R_{out} as the effective data rate.

Both the information encoding rate R_{in} and the node throughput λ_{out} depend on the SNR threshold μ. Intuitively, for larger μ, on average fewer packets can be successfully decoded. Yet each packet carries more information owing to a larger information encoding rate according to (3). Therefore, the effective data rate R_{out} is crucially determined by the SNR threshold μ. In practice, many applications have requirements on the minimum effective data rate, i.e., a certain amount of information has to be sent out within a time window. In Section V, we will further study how to minimize the mean access delay with a certain constraint on the effective data rate by optimally tuning both the transmission probability of nodes and the SNR threshold μ.

III. Network Steady-State Points

It has been shown in (1) that the service rate of each node's queue is determined by the steady-state probability of successful transmission of HOL packets p. Depending on whether the network is saturated, two fixed-point equations of p will be established, and the corresponding solutions are referred to as the steady-state points of the network.

Specifically, for each HOL packet, its transmission is successful if and only if its received SNR η is above the threshold μ, and all the other $n-1$ nodes are either idle with empty queues or busy but not requesting transmissions. For each node, let p_{emp} denote the probability of being idle with an empty queue, and p_{not} denote the probability of being busy with a HOL packet but not requesting transmission. The probability of successful transmission of HOL packets p can then be written as

$$p = \Pr\{\eta \geq \mu\}\cdot\left(p_{emp} + (1-p_{emp})\cdot p_{not}\right)^{n-1}. \tag{5}$$

[1] In practice, due to the slow-varying nature, the large-scale fading coefficients are usually available at the transmitter side through channel measurement and feedback. Therefore, uplink power control has been widely adopted in practical networks such as cellular systems.

[2] Note that with $\log_2(1+\eta) \geq R_{in}$, by random coding the error probability of a single packet (codeword) is exponentially reduced to zero as the codeword length goes to infinity. In practice, powerful AWGN-capacity-achieving codes can be adopted to efficiently suppress the error probability to a desirable level with moderate codeword length [29]. When the codeword length is not sufficiently large, nevertheless, (3) may be updated by further considering the error probability requirement determined by the specific coding/decoding schemes.

Note that the second part of the right-hand side of (5) is based on the assumption that events of node i being idle or busy without requesting are independent, $i = 1, \ldots, n$. With random arrivals and random transmissions of nodes, the independence assumption works well for Aloha networks when the number of nodes n is large.

As the received SNR of each HOL packet is exponentially distributed with mean ρ, we have

$$\Pr\{\eta \geq \mu\} = \exp\left(-\frac{\mu}{\rho}\right). \tag{6}$$

According to the Markov chain of HOL packets shown in Fig. 3, the probability that a HOL packet is not requesting transmission can be written as

$$p_{not} = \pi_T(1 - q_0) + \sum_{i=0}^{K} \pi_i(1 - q_i). \tag{7}$$

The probability of being idle with an empty queue, p_{emp}, depends on whether the network is saturated or not. In the saturated case, each node's queue is busy with probability 1. Otherwise, the probability that a queue is busy is given by $\lambda/\pi_T < 1$, where λ and π_T are the input rate and service rate of each node's queue, respectively. We then have

$$p_{emp} = \begin{cases} 1 - \lambda/\pi_T & \lambda < \pi_T \\ 0 & \lambda \geq \pi_T. \end{cases} \tag{8}$$

A. Steady-State Point in the Unsaturated Condition p_L

By combining (5)-(8), the steady-state probability of successful transmission of HOL packets p in the unsaturated condition can be written as

$$\begin{aligned} p &= \left(1 - \frac{\lambda}{\pi_T} + \frac{\lambda}{\pi_T}\left(\pi_T(1-q_0) + \sum_{i=0}^{K} \pi_i(1-q_i)\right)\right)^{n-1} \exp\left(-\frac{\mu}{\rho}\right) \\ &\overset{\text{for large } n}{\approx} \exp\left(-\frac{\hat{\lambda}}{p} - \frac{\mu}{\rho}\right), \end{aligned} \tag{9}$$

which has two non-zero roots:

$$\begin{aligned} p_L &= \exp\left\{\mathbb{W}_0\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right) - \frac{\mu}{\rho}\right\}, \\ p_S &= \exp\left\{\mathbb{W}_{-1}\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right) - \frac{\mu}{\rho}\right\}, \end{aligned} \tag{10}$$

if and only if the aggregate input rate $\hat{\lambda} = n\lambda$ is no larger than $\hat{\lambda}_{\max}^{p=p_L} = \exp\left\{-1 - \frac{\mu}{\rho}\right\}$. $\mathbb{W}_0(\cdot)$ and $\mathbb{W}_{-1}(\cdot)$ are two branches of the Lambert W function [36]. We have $p_L \geq p_S$ and the equality holds when $\hat{\lambda} = \hat{\lambda}_{\max}^{p=p_L}$, at which $p_L = p_S = \exp\left\{-1 - \frac{\mu}{\rho}\right\}$. By following the approximate trajectory analysis proposed in [28], it can be found that only the larger root p_L is the steady-state point, which we refer to as the desired steady-state point. According to (10), we can see that p_L is determined by the aggregate input rate $\hat{\lambda}$, the SNR threshold μ and the mean received SNR ρ.

Note that for an unsaturated node, its throughput $\lambda_{out}^{p=p_L}$ is determined by its input rate λ. As a result, the corresponding network throughput $\hat{\lambda}_{out}^{p=p_L}$ is equal to the aggregate input rate $\hat{\lambda}$, and we have

$$\hat{\lambda}_{out}^{p=p_L} = \hat{\lambda} \leq \hat{\lambda}_{\max}^{p=p_L} = \exp\left\{-1 - \frac{\mu}{\rho}\right\}. \tag{11}$$

B. Steady-State Point in the Saturated Condition p_A

As the input rate λ grows and exceeds the service rate, each node's queue is busy with probability 1, and the network becomes saturated. By combining (5)-(8), the steady-state probability of successful transmission of HOL packets p in the saturated condition can be obtained as

$$\begin{aligned} p &= \exp\left(-\frac{\mu}{\rho}\right) \cdot \left(1 - \frac{\pi_T}{p}\right)^{n-1} \\ &\overset{\text{for large } n}{\approx} \exp\left\{-\frac{n\pi_T}{p} - \frac{\mu}{\rho}\right\} \\ &= \exp\left\{-\frac{\mu}{\rho} - \frac{n}{\sum_{i=0}^{K-1} \frac{p(1-p)^i}{q_i} + \frac{(1-p)^K}{q_K}}\right\}. \end{aligned} \tag{12}$$

It has been shown in [30] that when $\{q_i\}_{i=0,\ldots,K}$ is a monotonic non-increasing sequence, the fixed-point equation (12) has a single non-zero root p_A, which is referred to as the undesired steady-state point. We can see that p_A is closely dependent on transmission probabilities of nodes $\{q_i\}_{i=0,\ldots,K}$. For instance, with $K = 0$, the undesired steady-state point p_A can be explicitly written as

$$p_A = \exp\left\{-nq_0 - \frac{\mu}{\rho}\right\}. \tag{13}$$

Note that for a saturated node, its throughput $\lambda_{out}^{p=p_A}$ is equal to the service rate $\pi_T(p_A)$, which is below the input rate λ. The corresponding network throughput is then given by

$$\hat{\lambda}_{out}^{p=p_A} = \frac{n}{\sum_{i=0}^{K-1} \frac{(1-p_A)^i}{q_i} + \frac{(1-p_A)^K}{p_A q_K}} \leq \hat{\lambda}, \tag{14}$$

according to (1). By combining (12) and (14), the network throughput $\hat{\lambda}_{out}^{p=p_A}$ can further be written as

$$\begin{aligned} \hat{\lambda}_{out}^{p=p_A} &= -p_A \ln p_A - \frac{p_A \mu}{\rho} \\ &\leq \hat{\lambda}_{\max}^{p=p_A} = \max_{p_A} \hat{\lambda}_{out}^{p=p_A} = \exp\left\{-1 - \frac{\mu}{\rho}\right\}. \end{aligned} \tag{15}$$

We can conclude from (11) and (15) that the maximum network throughput $\hat{\lambda}_{\max} = \exp\left\{-1 - \frac{\mu}{\rho}\right\}$.[3]

C. When to Operate at the Desired Steady-State Point p_L?

So far we have shown that a slotted Aloha network has two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A, at which the throughput performance varies. Specifically, the network throughput is determined by the aggregate input rate $\hat{\lambda}$ while insensitive to the transmission probabilities $\{q_i\}_{i=0,\ldots,K}$ when it operates at the desired steady-state point p_L. On the other hand, the network throughput at the undesired steady-state point p_A is crucially dependent on the transmission probabilities $\{q_i\}_{i=0,\ldots,K}$.

In practice, it is important to know at which steady-state point the network operates for given transmission probabilities $\{q_i\}_{i=0,\ldots,K}$. Given the backoff function $\mathcal{Q}(i)$, it is desirable to characterize a region of initial transmission probability q_0, within which the network operates at the desired steady-state

[3]Note that with $\rho \to \infty$, the above results reduce to those presented in [28] with the ideal collision channel, where a packet transmission is successful as long as there are no concurrent transmissions.

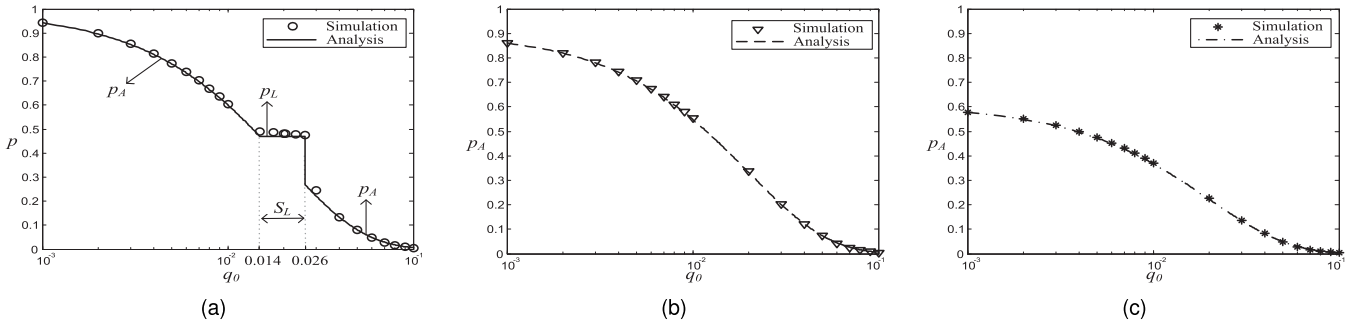


Fig. 4. Steady-state probability of successful transmission of HOL packets p versus the initial transmission probability q_0. $n = 50$. $\hat{\lambda} = 0.35$. (a) $\mu = 0.1$, $\rho = 10$ dB. (b) $\mu = 0.1$, $\rho = 0$ dB. (c) $\mu = 5$, $\rho = 10$ dB.

point p_L. Otherwise, the network operates at the undesired steady-state point p_A. Such a region is in general difficult to characterize due to the lack of the dynamic trajectory of the probability p_t of successful transmission of HOL packets at time slot t. Therefore, similar to [28], we derive the absolute-stable region $S_L = \{q_0 | \lambda \leq \pi_T(p_L), \min_t p_t \geq p_S\}$. With the initial transmission probability $q_0 \in S_L$, the network is guaranteed to operate at the desired steady-state point p_L. If $q_0 \notin S_L$, the network may shift to the undesired steady-state point p_A. With $K = 0$, an explicit expression of the absolute-stable region S_L of q_0 can be obtained as

$$S_L = \left[-\tfrac{1}{n}\mathbb{W}_0\left(-\hat{\lambda}\exp\left(\tfrac{\mu}{\rho}\right)\right), -\tfrac{1}{n}\mathbb{W}_{-1}\left(-\hat{\lambda}\exp\left(\tfrac{\mu}{\rho}\right)\right)\right]. \quad (16)$$

Appendix A presents the detailed derivation of (16). (16) indicates that the absolute-stable region S_L shrinks as the aggregate input rate $\hat{\lambda}$ or the SNR threshold μ increases, or the mean received SNR ρ decreases. S_L is a non-empty set if and only if $\hat{\lambda} \leq e^{-1}$ and $\mu \leq \mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$. The tightness of the absolute-stable region S_L derived in (16) will be validated by simulation results presented below.

D. Simulation Results

In this section, simulation results are presented to verify the preceding analysis. In this paper, event-driven simulations are conducted and each simulation is carried out for 10^8 time slots. The simulation setting is the same as the system model characterized in Section II and thus we omit the details here due to limited space. In simulations, we count the total number of transmitted packets from all nodes and the total number of successful packets. The steady-state probability of successful transmission of HOL packets p is then obtained by calculating the ratio of the number of successful packets to the total number of transmitted packets.

Specifically, the analysis has revealed that if the initial transmission probability q_0 is selected from the absolute-stable region S_L, then the network operates at the desired steady-state point p_L. Otherwise, the network may shift to the undesired steady-state point p_A. With $K = 0$, expressions of p_L, p_A and S_L have been given in (10), (13) and (16), respectively. As Fig. 4 illustrates, the absolute-stable region S_L is crucially determined by the SNR threshold μ and the mean received SNR ρ. For instance, with $\mu = 0.1$ and $\rho = 10$ dB, as shown in Fig. 4a, the absolute-stable region $S_L = [0.014, 0.026]$ for the aggregate input rate $\hat{\lambda} = 0.35$ and the number of nodes $n = 50$. In this case, if the initial transmission probability $q_0 \in S_L$, then the network would operate at the desired steady-state point p_L, which is independent of q_0. Otherwise, it shifts to the undesired steady-state point p_A, which is a decreasing function of q_0. As the mean received SNR ρ decreases, e.g., $\rho = 0$ dB, or the SNR threshold μ increases, e.g., $\mu = 5$, the absolute-stable region S_L vanishes. In this case, the network always operates at the undesired steady-state point p_A, as Fig. 4b–c show. Simulation results presented in Fig. 4 well agree with the analysis.

IV. Mean Access Delay at Two Steady-State Points

In this section, by deriving the probability generating function of access delay, we will characterize moments of access delay and study how to properly select the initial transmission probability q_0 to minimize the mean access delay.

A. Moments of Access Delay

Denote Y_i as the sojourn time of a HOL packet in State i for $i = 0, 1, \ldots, K$ and D_i as the time from the beginning of State i to the service completion for $i = T, 0, 1, \ldots, K$. According to the Markov chain in Fig. 3, we have

$$D_T = \begin{cases} 1 & \text{with probability } q_0\, p \\ 1 + D_1 & \text{with probability } q_0(1-p) \\ 1 + D_0 & \text{with probability } 1 - q_0, \end{cases} \quad (17)$$

and

$$D_i = \begin{cases} Y_i & \text{with probability } p \\ Y_i + D_{i+1} & \text{with probability } 1 - p, \end{cases} \quad (18)$$

for $i = 0, \ldots, K-1$, and $D_K = Y_K$. Note that D_T is the service time of HOL packets, which is also the access delay. Let $G_{D_T}(z)$ denote its probability generating function. According to (17) and (18), we have

$$\begin{cases} G_{D_T}(z) = q_0 pz + (1-q_0) z G_{D_0}(z) + q_0(1-p) z G_{D_1}(z), \\ G_{D_i}(z) = pG_{Y_i}(z) + (1-p) G_{Y_i}(z) G_{D_{i+1}}(z),\ i = 0, \ldots, K-1, \\ G_{D_K}(z) = G_{Y_K}(z), \end{cases} \quad (19)$$

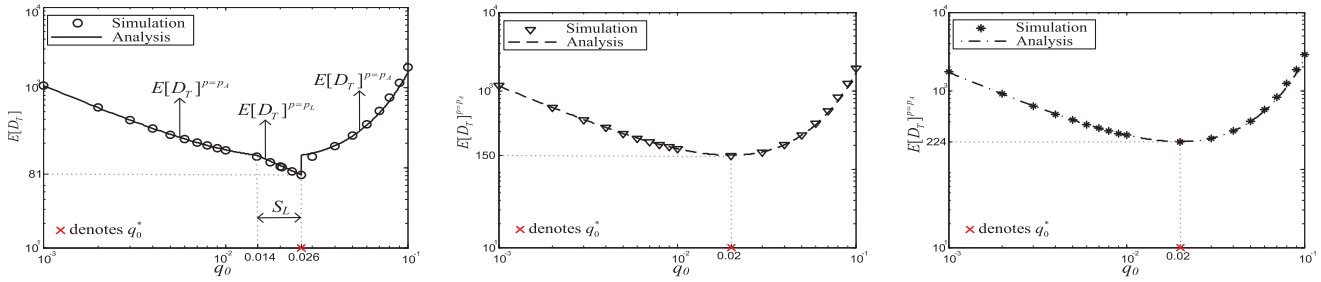


Fig. 5. Mean access delay $E[D_T]$ (in unit of time slots) versus the initial transmission probability q_0. $n = 50$. $\hat{\lambda} = 0.35$. (a) $\mu = 0.1$, $\rho = 10$ dB. (b) $\mu = 0.1$, $\rho = 0$ dB. (c) $\mu = 5$, $\rho = 10$ dB.

where the sojourn time of a HOL packet in State i, Y_i, follows the geometric distribution with the probability generating function given by

$$G_{Y_i}(z) = \begin{cases} \frac{q_i z}{1-z(1-q_i)} & \text{for } i = 0, \ldots, K-1 \\ \frac{pq_K z}{1-z(1-pq_K)} & \text{for } i = K. \end{cases} \tag{20}$$

It can be obtained from (19) that

$$\begin{aligned} &G_{D_T}(z) \\ &= q_0 pz + q_0 z\Big(p\sum_{j=1}^{K-1}(1-p)^j \prod_{i=1}^{j} G_{Y_i}(z) \\ &\quad + (1-p)^K \prod_{i=1}^{K} G_{Y_i}(z)\Big) + (1-q_0)z \\ &\quad \times \Big(p\sum_{j=0}^{K-1}(1-p)^j \prod_{i=0}^{j} G_{Y_i}(z) \\ &\quad + (1-p)^K \prod_{i=0}^{K} G_{Y_i}(z)\Big). \end{aligned} \tag{21}$$

By combining (20) and (21), the first moment of the access delay $E[D_T]$, i.e., the mean access delay (in unit of time slots), and the second moment of access delay $E[D_T^2]$ can be obtained as

$$E[D_T] = G'_{D_T}(1) = \sum_{i=0}^{K-1} \frac{(1-p)^i}{q_i} + \frac{(1-p)^K}{pq_K}, \tag{22}$$

$$\begin{aligned} E[D_T^2] &= G'_{D_T}(1) + G''_{D_T}(1) \\ &= \sum_{i=0}^{K-1} \frac{(1-p)^i}{q_i} + \frac{(1-p)^K}{pq_K} + \sum_{i=0}^{K-1} \frac{2(1-p)^i}{q_i}\Big(\frac{1-q_i}{q_i} \\ &\quad + \sum_{j=1}^{K-i-1} \frac{(1-p)^j}{q_{i+j}} + \frac{(1-p)^{K-i}}{pq_K}\Big) + \frac{2(1-p)^K(1-pq_K)}{p^2 q_K^2}. \end{aligned} \tag{23}$$

We can observe from (22)-(23) that both $E[D_T]$ and $E[D_T^2]$ are crucially determined by the transmission probabilities $\{q_i\}_{i=0,\ldots,K}$. In the following, we consider the case of $K = 0$. The first moment of access delay $E[D_T]$ and the second moment of access delay $E[D_T^2]$ in (22) and (23) can then be simplified as

$$E[D_T] = \frac{1}{pq_0}, \quad \text{and } E[D_T^2] = \frac{2-q_0 p}{q_0^2 p^2}. \tag{24}$$

Note that with the probability generating function of access delay, we can also obtain the mean queueing delay for a given arrival process. With Bernoulli arrivals, for instance, the mean queueing delay is determined by the first and second moments of the access delay. Moreover, we can see from (24) that $E[D_T^2] = 2(E[D_T])^2 - E[D_T]$, indicating that the minimization of $E[D_T^2]$ is equivalent to that of the mean access delay $E[D_T]$. Therefore, in the following, we only focus on the optimization of mean access delay.

B. Minimizing Mean Access Delay

The analysis in Section III has revealed that the network has two steady-state points, i.e., the desired steady-state point p_L and the undesired steady-state point p_A. By combining (10), (13) and (24), we can obtain the mean access delay $E[D_T]^{p=p_L}$ at the desired steady-state point p_L and $E[D_T]^{p=p_A}$ at the undesired steady-state point p_A as

$$\begin{aligned} E[D_T]^{p=p_L} &= \frac{1}{q_0}\exp\left\{-\mathbb{W}_0\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)+\frac{\mu}{\rho}\right\}, \\ E[D_T]^{p=p_A} &= \frac{1}{q_0}\exp\left\{nq_0+\frac{\mu}{\rho}\right\}, \end{aligned} \tag{25}$$

respectively. We can see from (25) that both $E[D_T]^{p=p_L}$ and $E[D_T]^{p=p_A}$ are determined by the number of nodes n, the initial transmission probability q_0, the SNR threshold μ and the mean received SNR ρ. In this subsection, we are interested in minimizing the mean access delay $E[D_T]$ by optimally tuning the initial transmission probability q_0, i.e., $\min_{q_0} E[D_T]$.

Note that it has been shown in Section III-C that when the initial transmission probability $q_0 \in S_L$, the network operates at the desired steady-state point p_L, with which we can see from (25) that $E[D_T]^{p=p_L}$ monotonically decreases as q_0 increases. As a result, to minimize the mean access delay $E[D_T]^{p=p_L}$, q_0 should be set to the upper-bound of the absolute-stable region S_L, i.e., $-\frac{1}{n}\mathbb{W}_{-1}\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right)$, with which we can obtain from (25) that $\min_{q_0\in S_L} E[D_T]^{p=p_L} = \frac{\mathbb{W}_0\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}{\lambda\mathbb{W}_{-1}\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}$. On the other hand, if the network operates at the undesired stable point p_A, then it can be easily derived from (25) that $\min_{q_0\notin S_L} E[D_T]^{p=p_A} = n\exp\left\{1+\frac{\mu}{\rho}\right\}$, for achieving which q_0 should be set to $\frac{1}{n}$. Since the absolute-stable region $S_L \neq \emptyset$ if and only if the aggregate

input rate $\hat{\lambda} \leq e^{-1}$ and the SNR threshold $\mu \leq \mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$, as shown in Section III-C, the minimum mean access delay $\min_{q_0} E[D_T]$ can be written as

$$\min_{q_0} E[D_T] = \begin{cases} \frac{\mathbb{W}_0\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}{\lambda\mathbb{W}_{-1}\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)} & \text{if } \mu \leq \mu_0 \text{ and } \hat{\lambda} \leq e^{-1}, \\ n\exp\left\{1+\frac{\mu}{\rho}\right\} & \text{otherwise,} \end{cases} \tag{26}$$

which is achieved when q_0 is set to

$$q_0^* = \begin{cases} -\frac{1}{n}\mathbb{W}_{-1}\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right) & \text{if } \mu \leq \mu_0 \text{ and } \hat{\lambda} \leq e^{-1}, \\ \frac{1}{n} & \text{otherwise.} \end{cases} \tag{27}$$

C. Simulation Results

In this subsection, simulation results are presented to verify the above analysis. The simulation setting is the same as the system model 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 packets to the total number of successfully transmitted packets.

Specifically, the expressions of mean access delay at the desired steady-state point p_L and the undesired steady-state point p_A have been given in (25) and illustrated in Fig. 5. As Fig. 5a shows, when the initial transmission probability q_0 is chosen from the absolute stable region S_L, the mean access delay $E[D_T]^{p=p_L}$ decreases as q_0 increases. To minimize $E[D_T]^{p=p_L}$, q_0 should be set to the upper-bound of S_L. On the other hand, if the absolute-stable region S_L does not exist, then the network always operates at the undesired stable point p_A. In this case, as shown in Fig. 5b-c, to minimize the mean access delay $E[D_T]^{p=p_A}$, the optimal initial transmission probability q_0^* should be set as $\frac{1}{n}$.

To see the performance gain of optimal tuning of q_0, Fig. 6 illustrates how the mean access delay varies with the number of nodes n with the initial transmission probability $q_0 = 0.01$, 0.05 or q_0^*. It has been shown in (25) that with fixed q_0, the mean access delay $E[D_T]^{p=p_A}$ exponentially increases with n. In sharp contrast, the minimum mean access delay linearly increases with the number of nodes n when n is large according to (26). Fig. 6 corroborates that substantial gains in the mean access delay can be achieved by optimally tuning the initial transmission probability q_0 especially in the massive access scenario.

Note that for large n, even with q_0 optimally tuned, the throughput of each node $\lambda_{out} = \frac{\hat{\lambda}_{\max}}{n} = \frac{1}{n}\exp\left\{-1-\frac{\mu}{\rho}\right\}$ still decreases as the number of nodes n grows, indicating that the effective data rate of each node $R_{out} = R_{in}\cdot\lambda_{out}$ would also decline if the information encoding rate of each node R_{in} is fixed. In practice, however, many applications may have requirements on the minimum data rate of each node. As we will demonstrate in the next section, to minimize the mean access delay while taking the data rate requirement into consideration, both the initial transmission probability q_0 and the information encoding rate of each node R_{in} need to be optimally chosen.

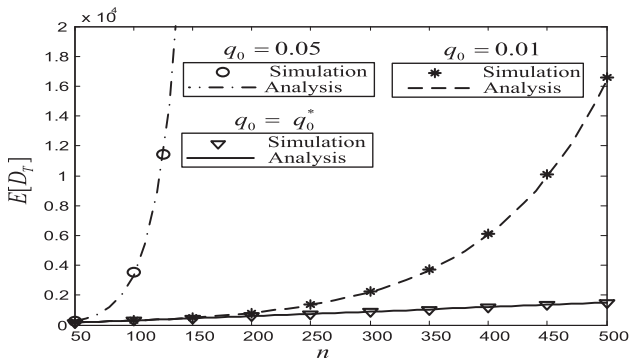


Fig. 6. Mean access delay $E[D_T]$ (in unit of time slots) versus the number of nodes n. $\lambda = 0.01$. $\mu = 0.1$. $\rho = 0$ dB.

V. Rate-Constrained Minimum Mean Access Delay

In this section, we will study how to minimize the mean access delay with a certain constraint of the effective data rate of each node. Specifically, it has been shown in previous sections that both the effective data rate of each node $R_{out} = R_{in}\cdot\lambda_{out}$ and the mean access delay $E[D_T]$ are determined by the SNR threshold μ and the initial transmission probability q_0. Let R_0 denote the minimum required data rate for each node. We are interested in characterizing the rate-constrained minimum mean access delay:

$$D_R^* = \min_{\mu>0, 0<q_0\leq 1} E[D_T] \quad \text{s.t.} \quad R_{in}\cdot\lambda_{out} \geq R_0. \tag{28}$$

Note that if the minimum required data rate R_0 is too large, then the constraint $R_{in}\cdot\lambda_{out} \geq R_0$ may not hold for any $q_0 \in (0,1]$ and $\mu \in (0,+\infty)$, in which case the optimization problem in (28) has no solution. Let us define the maximum achievable rate of a slotted Aloha network as $\bar{C} = \max_{\mu>0,\ 0<q_0\leq 1} R_{in}\cdot\hat{\lambda}_{out}$. The data rate constraint cannot be satisfied when $R_0 > \frac{\bar{C}}{n}$. The following lemma presents the maximum achievable rate $\bar{C}$.

Lemma 1: The maximum achievable rate of a slotted Aloha network is given by

$$\bar{C} = \begin{cases} C_u & 0 < \hat{\lambda} \leq \hat{\lambda}_\rho, \\ C_s & \hat{\lambda} > \hat{\lambda}_\rho, \end{cases} \tag{29}$$

where C_u is the maximum achievable rate in the unsaturated case, which is given by

$$C_u = \hat{\lambda}\log_2(1-\rho-\rho\ln\hat{\lambda}), \tag{30}$$

and C_s is the maximum achievable rate in the saturated case, which is given by

$$C_s = \exp\left(-1-\frac{e^{\mathbb{W}_0(\rho)}-1}{\rho}\right)\cdot\log_2(e^{\mathbb{W}_0(\rho)}), \tag{31}$$

and $\hat{\lambda}_\rho = \exp\left(-1-\frac{e^{\mathbb{W}_0(\rho)}-1}{\rho}\right)$.

Proof: See Appendix B. ∎

Theorem 1: If $0 \le R_0 \le \frac{\bar{C}}{n}$*, then the rate-constrained minimum mean access delay* D_R^* *is given by*

$$D_R^* = \begin{cases} \dfrac{\mathbb{W}_0\left(-n\lambda \exp\left(\frac{2^{\frac{R_0}{\lambda}}-1}{\rho}\right)\right)}{\lambda \mathbb{W}_{-1}\left(-n\lambda \exp\left(\frac{2^{\frac{R_0}{\lambda}}-1}{\rho}\right)\right)} & \text{if } 0 < \lambda \le \frac{e^{-1}}{n} \text{ and } R_0 \le \frac{C_u}{n}, \\ n \exp\left(1+\frac{\mu_1}{\rho}\right) & \text{if } \frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \le \frac{C_s}{n}, \text{ or } \lambda \ge \frac{e^{-1}}{n} \text{ and } R_0 \le \frac{C_s}{n}, \end{cases} \tag{32}$$

which is achieved when the SNR threshold μ *is set to*

$$\mu_R^* = \begin{cases} 2^{\frac{R_0}{\lambda}} - 1 & \text{if } 0 < \lambda \le \frac{e^{-1}}{n} \text{ and } R_0 \le \frac{C_u}{n}, \\ \mu_1 & \text{if } \frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \le \frac{C_s}{n}, \text{ or } \lambda \ge \frac{e^{-1}}{n} \text{ and } R_0 \le \frac{C_s}{n}, \end{cases} \tag{33}$$

and the initial transmission probability q_0 *is set to (34)*

$$q_{0,R}^* = \begin{cases} -\frac{1}{n}\mathbb{W}_{-1}\left(-n\lambda \exp\left(\frac{2^{\frac{R_0}{\lambda}}-1}{\rho}\right)\right) & \text{if } 0 < \lambda \le \frac{e^{-1}}{n} \text{ and } R_0 \le \frac{C_u}{n}, \\ \frac{1}{n} & \text{if } \frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \le \frac{C_s}{n}, \text{ or } \lambda \ge \frac{e^{-1}}{n} \text{ and } R_0 \le \frac{C_s}{n}, \end{cases} \tag{34}$$

where μ_1 *is the smaller root of the following equation*

$$\frac{1}{n}\exp\left(-1-\frac{\mu}{\rho}\right)\log_2(1+\mu) = R_0. \tag{35}$$

Otherwise, the optimization problem (28) *has no feasible solution.*

Proof: See Appendix C. ∎

Theorem 1 shown at the top of this page presents the rate-constrained minimum mean access delay D_R^* and the corresponding optimal settings of the initial transmission probability $q_{0,R}^*$ and the SNR threshold μ_R^* for the minimum required data rate $R_0 \le \frac{\bar{C}}{n}$. Note that the optimal information encoding rate R_{in}^* can be obtained from (33) as $R_{in}^* = \log_2(1+\mu_R^*)$ according to (3).

A. Unsaturated Region $\mathcal{S}_\mathcal{U}$*, Saturated Region* $\mathcal{S}_\mathcal{S}$ *and Infeasible Region* $\mathcal{S}_\mathcal{I}$

It is shown in the proof of Theorem 1 that the network operates in the unsaturated condition when $0 < \lambda \le \frac{e^{-1}}{n}$ and $0 < R_0 \le \frac{C_u}{n}$, in which the rate-constrained minimum mean access delay D_R^* is determined by the input rate of each node λ, the number of nodes n, the minimum required data rate for each node R_0 and the mean received SNR ρ. On the other hand, when $\frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n}$ and $\frac{C_u}{n} < R_0 \le \frac{C_s}{n}$, or $\lambda \ge \frac{e^{-1}}{n}$ and $0 < R_0 \le \frac{C_s}{n}$, the network operates in the saturated condition and the corresponding D_R^* is only determined by n, R_0 and ρ. We can define the following regions in terms of (n, λ, R_0, ρ):

- **Unsaturated region** $\mathcal{S}_\mathcal{U} = \left\{(n,\lambda,R_0,\rho)|0 < \lambda \le \frac{e^{-1}}{n} \text{ and } 0 < R_0 \le \frac{C_u}{n}\right\}$, in which D_R^* is achieved when the network is unsaturated.
- **Saturated region** $\mathcal{S}_\mathcal{S} = \left\{(n,\lambda,R_0,\rho)|\frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \le \frac{C_s}{n}, \text{ or } \lambda \ge \frac{e^{-1}}{n} \text{ and } 0 < R_0 \le \frac{C_s}{n}\right\}$, in which D_R^* is achieved when the network is saturated.
- **Infeasible region** $\mathcal{S}_\mathcal{I} = \overline{\mathcal{S}_\mathcal{U} \bigcup \mathcal{S}_\mathcal{S}}$, in which the optimization problem (28) has no solution.

A graphic illustration of the unsaturated region $\mathcal{S}_\mathcal{U}$, saturated region $\mathcal{S}_\mathcal{S}$ and infeasible region $\mathcal{S}_\mathcal{I}$ is presented in Fig. 7. As Fig. 7a shows, the network operates at the unsaturated region $\mathcal{S}_\mathcal{U}$ when both the input rate of each node λ and the minimum required data rate for each node R_0 are small. As λ increases, the network would shift to the saturated region $\mathcal{S}_\mathcal{S}$ and eventually falls into the infeasible region $\mathcal{S}_\mathcal{I}$ when R_0 is large, i.e., $R_0 > \frac{C_s}{n}$.

Note that those three regions can also be interpreted in terms of (ρ, λ) and (n, λ), as shown in Fig. 7b and Fig. 7c, respectively. Specifically, ρ_u and ρ_s in Fig. 7b are the roots of $R_0 = \frac{C_u}{n}$ and $R_0 = \frac{C_s}{n}$ for ρ, respectively, which can be obtained as

$$\begin{aligned} \rho_u &= \tfrac{1-2^{R_0/\lambda}}{1+\ln\hat{\lambda}}, \\ \rho_s &\approx \left(\mathbb{W}_0\left(\tfrac{1}{neR_0\ln 2}\right)\right)^{-1} \exp\left\{\left(\mathbb{W}_0\left(\tfrac{1}{neR_0\ln 2}\right)\right)^{-1}\right\}. \end{aligned} \tag{36}$$

Similarly, n_u and n_s in Fig. 7c are roots of $R_0 = \frac{C_u}{n}$ and $R_0 = \frac{C_s}{n}$ for n, respectively, which can be obtained as

$$\begin{aligned} n_u &= \tfrac{1}{\lambda} \cdot \exp\left(\tfrac{1}{\rho} - 1 - \tfrac{2^{R_0/\lambda}}{\rho}\right), \\ n_s &= \tfrac{1}{R_0} \cdot \exp\left(-1 - \tfrac{e^{\mathbb{W}_0(\rho)}-1}{\rho}\right) \cdot \log_2(e^{\mathbb{W}_0(\rho)}). \end{aligned} \tag{37}$$

We can see from Fig. 7b and Fig. 7c that for a given minimum required data rate for each node R_0, the network falls into the infeasible region $\mathcal{S}_\mathcal{I}$ when either the mean received SNR ρ is too small or the number of nodes n is too

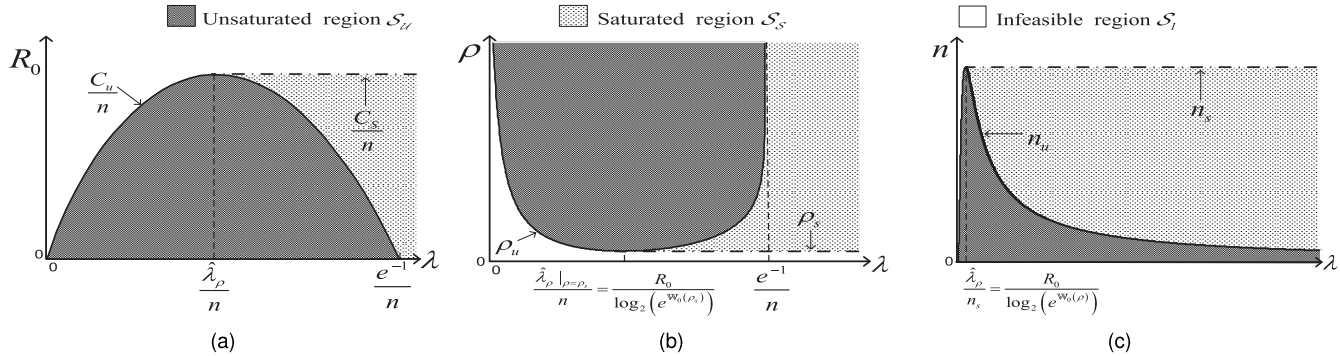


Fig. 7. Unsaturated region $\mathcal{S}_{\mathcal{U}}$, saturated region $\mathcal{S}_{\mathcal{S}}$ and infeasible region $\mathcal{S}_{\mathcal{I}}$ for given (a) ρ and n, (b) R_0 and n, and (c) ρ and R_0.

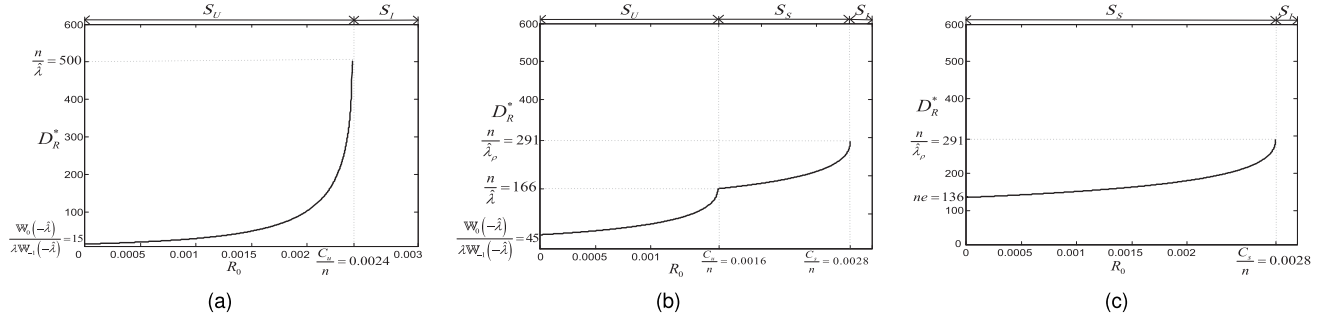


Fig. 8. Rate-constrained minimum mean access delay D_R^* (in unit of time slots) versus the minimum required data rate for each node R_0 (in unit of bit/s/Hz). $n = 50$. $\rho = 0$ dB. $\hat{\lambda}_\rho = 0.1715$. (a) $\hat{\lambda} = 0.1$. (b) $\hat{\lambda} = 0.3$. (c) $\hat{\lambda} = 0.5$.

large. Similar to Lemma 1 where the maximum achievable rate $\bar{C}$ is characterized as an upperbound of R_0, we can also characterize the minimum required mean received SNR $\bar{\rho}$ as

$$\bar{\rho} = \begin{cases} \rho_u & 0 < \lambda \leq \frac{R_0}{\log_2(e^{\mathbb{W}0(\rho_s)})}, \\ \rho_s & \lambda > \frac{R_0}{\log_2(e^{\mathbb{W}0(\rho_s)})}, \end{cases} \tag{38}$$

and the maximum allowable number of nodes $\bar{n}$ as

$$\bar{n} = \begin{cases} n_u & 0 < \lambda \leq \frac{R_0}{\log_2(e^{\mathbb{W}0(\rho)})}, \\ n_s & \lambda > \frac{R_0}{\log_2(e^{\mathbb{W}0(\rho)})}. \end{cases} \tag{39}$$

When the mean received SNR $\rho < \bar{\rho}$ or the number of nodes $n > \bar{n}$, the data rate constraint cannot be satisfied.

B. Discussions

To take a closer look at Theorem 1, Fig. 8 and Fig. 9 illustrate how the rate-constrained minimum mean access delay D_R^* varies with the minimum required data rate for each node R_0 and the mean received SNR ρ, respectively, under various values of the aggregate input rate $\hat{\lambda}$. For the asymptotic cases, it can be easily obtained from Theorem 1 that

$$\lim_{R_0 \to 0} D_R^* =$$

$$\lim_{\rho \to +\infty} D_R^* = \begin{cases} \frac{\mathbb{W}_0(-\hat{\lambda})}{\lambda \mathbb{W}_{-1}(-\hat{\lambda})} & 0 < \hat{\lambda} < e^{-1}, \\ ne & \hat{\lambda} \geq e^{-1}. \end{cases} \tag{40}$$

With a positive rate requirement $R_0 > 0$ and finite mean received SNR $\rho < \infty$, D_R^* would increase as R_0 grows or ρ declines. As shown in Fig. 8a and Fig. 9a, when the aggregate input rate $\hat{\lambda} = 0.1$, we have $\hat{\lambda} < \hat{\lambda}_\rho = 0.1715$ for $\rho = 0$ dB, and $\hat{\lambda} < \hat{\lambda}_\rho|_{\rho=\rho_s} = 0.174$ for $R_0 = 0.003$ bit/s/Hz. Therefore, the network operates at the unsaturated region $\mathcal{S}_{\mathcal{U}}$ with $R_0 < \frac{C_u}{n}$ or $\rho > \rho_u$. If the aggregate input rate $\hat{\lambda}$ increases to 0.3, then as shown in Fig. 8b and Fig. 9b, the network would first operate at the unsaturated region $\mathcal{S}_{\mathcal{U}}$, and then shifts to the saturated region $\mathcal{S}_{\mathcal{S}}$ when R_0 exceeds $\frac{C_u}{n}$ or ρ drops below ρ_u. If the aggregate input rate $\hat{\lambda}$ further increases to 0.5 such that $\hat{\lambda} > e^{-1}$, then as shown in Fig. 8c and Fig. 9c, the network always operates at the saturated region as long as $R_0 < \frac{C_s}{n}$ or $\rho > \rho_s$.

It is interesting to note from Fig. 8 and Fig. 9 that for high rate requirement R_0 or small mean received SNR ρ, a lower traffic input rate may even lead to larger rate-constrained minimum mean access delay D_R^*. Specifically, it can be seen from Fig. 8a and Fig. 9a that with the aggregate input rate $\hat{\lambda} = 0.1$, D_R^* sharply increases when R_0 (or ρ) is close to the limit $\frac{C_u}{n}$ (or ρ_u). Intuitively, when the traffic input rate is small, to satisfy the rate requirement, the information encoding rate of each packet has to be sufficiently high, which leads to low chances of successful transmission of HOL packets and thus poor delay performance. It outperforms the heavy traffic input rate case only when the data rate requirement is loose (i.e., small R_0) or the mean received SNR ρ is large.

Fig. 10a illustrates how the rate-constrained minimum mean access delay D_R^* varies with the number of nodes n under various values of the mean received SNR ρ. We can see that the rate-constrained minimum mean access delay D_R^* superlinearly increases with n for large n. It is in sharp contrast to the unconstrained case, as shown in (26) and Fig. 6, where the minimum mean access delay $\min_{q_0} E[D_T]$

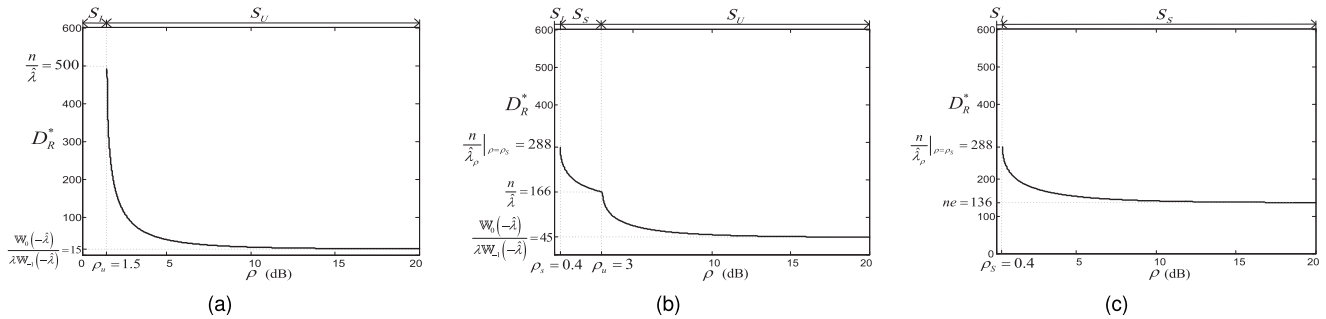


Fig. 9. Rate-constrained minimum mean access delay D_R^* (in unit of time slots) versus the mean received SNR ρ. $n = 50$. $R_0 = 0.003$ bit/s/Hz. (a) $\hat{\lambda} = 0.1$. (b) $\hat{\lambda} = 0.3$. (c) $\hat{\lambda} = 0.5$.

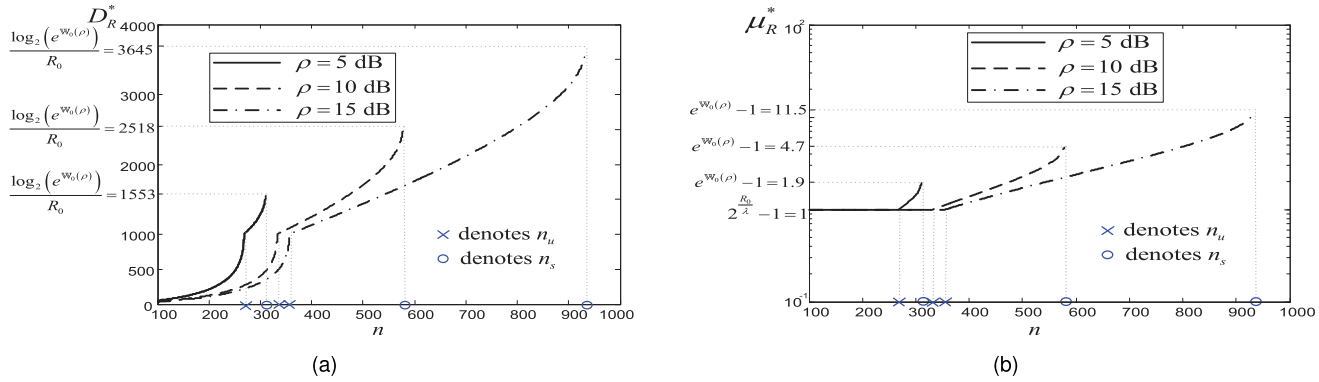


Fig. 10. (a) Rate-constrained minimum mean access delay D_R^* (in unit of time slots) versus the number of nodes n. (b) Optimal SNR threshold μ_R^* versus the number of nodes n. $R_0 = 0.001$ bit/s/Hz. $\lambda = 0.001$. $\rho = 5$ dB, 10 dB or 15 dB.

linearly increases with n for fixed SNR threshold μ (or equivalently, information encoding rate R_{in}). Intuitively, as the number of nodes n increases, the throughput of each node λ_{out} declines. Therefore, to satisfy the required data rate R_0, each node should enlarge the information encoding rate R_{in}. As Fig. 10b shows, the optimal SNR threshold μ_R^* increases with the number of nodes n, leading to an increasing information encoding rate and thus superlinearly increasing rate-constrained minimum mean access delay D_R^*.

C. Insights for Massive Access of M2M Communications

The above analysis sheds important light on how to facilitate massive access of M2M communications. For illustration, let us take the example of LTE-M [38], which was developed by the Third-Generation Partnership Project (3GPP) for addressing the fast-expanding market for low power wide area connectivity. Similar to the legacy LTE networks, LTE-M also adopts an Aloha-based random access procedure. Yet, different from the legacy LTE networks, where the data packets are transmitted after a connection is established in the random access procedure, the early data transmission scheme is introduced in LTE-M, where each device sends its packet within the random access procedure [39].

In the following, we will apply our analysis to a single-cell LTE-M system with smart grid applications. We consider three representative traffic models: delay-insensitive light traffic model, delay-insensitive heavy traffic model and delay-sensitive traffic model, with traffic characteristics summarized in Table I [32], [37]. LTE-M has the transmission bandwidth of $B = 1.08$ MHz with the length of the random access procedure typically 15 milliseconds [40]. For each traffic model, we can calculate the input rate of each device $\lambda = \frac{\sigma}{\text{Reporting Period}}$ packet/slot, where $\sigma = 15$ milliseconds is the length of a time slot. The minimum required data rate normalized by the system bandwidth B is $R_0 = \frac{\text{Payload Size}}{\text{Reporting Period} \times B}$ bit/s/Hz.

Let us first focus on delay-insensitive traffic models, i.e., Traffic model 1 and Traffic model 2. Fig. 11a demonstrates how the rate-constrained minimum mean access delay D_R^* (in unit of seconds) varies with the number of devices n with the mean received SNR $\rho = 0$ dB. We can see from Fig. 11a that the maximum number of devices per cell that LTE-M can support is quite large, i.e., 34090 for Traffic model 1 and 11360 for Traffic model 2, respectively, though the corresponding minimum mean access delay D_R^* is also very high, i.e., $D_R^* = 2934$ seconds (48.9 minutes) for Traffic model 1 and $D_R^* = 969$ seconds (16.15 minutes) for Traffic model 2. Even with a delay constraint of 900 seconds (15 minutes), LTE-M can still support 18320 devices with Traffic model 1 and 11290 devices with Traffic model 2. It corroborates that LTE-M is well suited for massive access of machine-type devices with loose quality-of-service requirements.

As the delay constraint becomes stringent, however, the number of devices that can be supported would drastically decrease. We can see from Fig. 11b that for Traffic model 3, only 435 devices can be supported for a delay constraint of 1 second with $\rho = 0$ dB. Although the number of devices

TABLE I
CHARACTERISTICS OF THREE TRAFFIC MODELS IN SMART GRID [32], [37]

	Payload Size	Reporting Period	Delay Requirement	Use-case
Traffic model 1 (Delay-insensitive light traffic)	500 bytes	Every 15 minutes	15 minutes	Periodical power grid state reporting
Traffic model 2 (Delay-insensitive heavy traffic)	500 bytes	Every 5 minutes	15 minutes	
Traffic model 3 (Delay-sensitive traffic)	500 bytes	Every 60 minutes	1 second	Control message exchange

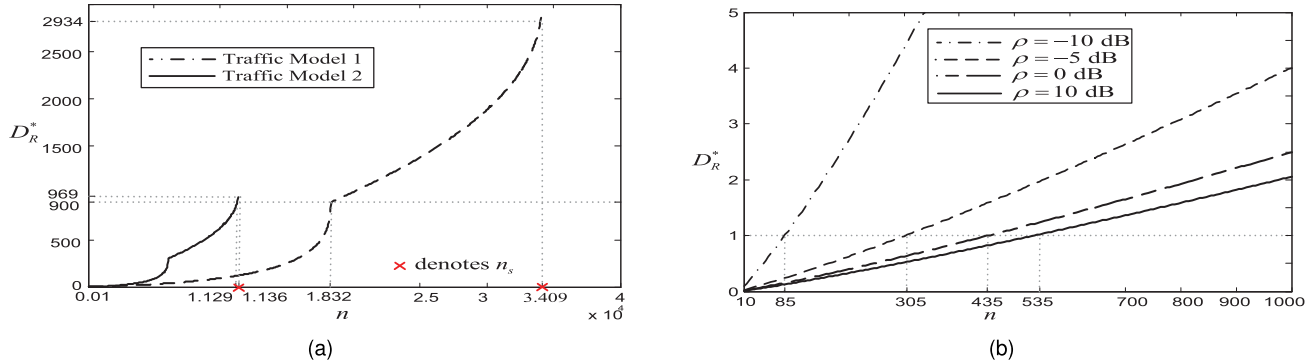


Fig. 11. Rate-constrained minimum mean access delay D_R^* (in unit of seconds) versus the number of devices n. (a) Traffic model 1 and Traffic model 2. $\rho = 0$ dB. (b) Traffic model 3. $\rho = -10$ dB, -5 dB, 0 dB or 10 dB.

can be enlarged by increasing the mean received SNR ρ, the gain is quite marginal, i.e., 535 devices with $\rho = 10$ dB. In this case, with sporadic transmissions from devices, i.e., the traffic input rate $\lambda = 4.1 \times 10^{-6}$ packet/time slot, the network operates at the unsaturated region with rate-constrained minimum mean access delay D_R^* linearly increasing with the number of devices n when n is small.

Note that despite the useful insights, there are caveats when applying the analysis to practical scenarios. First of all, the analysis focus on the *optimal* access delay performance, to achieve which the initial transmission probability and the information encoding rate of each device should be optimally tuned according to the total number of devices n, the mean received SNR ρ, the traffic input rate of each device λ and the minimum required data rate for each device R_0, as Theorem 1 shows. In practice, such an optimal tuning can be implemented at the receiver side, since all the devices communicate with a common receiver, e.g., Base Station (BS). The receiver can keep a record of all registered devices and collect the traffic characteristics. It then calculates the optimal initial transmission probability and the information encoding rate and broadcasts the configuration periodically for devices to update their parameter setting accordingly.

Secondly, in this paper, we only consider the access delay of each packet. For massive access of M2M communications, due to sporadic transmissions (i.e., low input rate) of devices, the waiting time in the buffer, i.e., the time for waiting-to-be-HOL-packet, is usually quite small. It is therefore important to optimize the access delay performance, which is the main contributor, key indicator and the potential bottleneck of the whole end-to-end delay performance. Nevertheless, for heavy traffic load scenarios where the waiting time of each packet cannot be neglected any more, the mean queueing delay would become the primary concern, which can further be characterized based on the probability generating function of access delay derived in Section IV-A.

There are also a few key assumptions that may be relaxed when extending the analysis to a variety of M2M communications systems:

1) Power Control: In this paper, power control is assumed to be adopted to ensure that each node has the same long-term performance, including throughput, mean access delay, and effective data rate. In some low-cost M2M communications systems where power control may not be supported, the mean received power of packets would vary from device to device, causing serious performance disparity. In that case, fairness constraints need to be further considered when optimizing the transmission probability and the information encoding rate of each device.

2) Collision Model: With the collision model, at most one packet can be successfully decoded in each time slot regardless of the difference in received power of nodes. In practice, however, that difference can be well utilized to enhance the throughput performance. It has been shown in our previous studies [31] that with the capture model, the maximum network throughput of slotted Aloha in fading channels can be significantly improved at the low SNR region, though the gain in the maximum sum rate is marginal. For M2M communications systems with low-power devices, it is

important to further extend the delay analysis to incorporate more advanced receivers.

3) SNR Threshold: In fading channels, even without concurrent transmissions, a packet (codeword) cannot be correctly decoded if its received SNR is too low to support its information encoding rate. In this paper, we set the SNR threshold based on the channel capacity by assuming that the codeword length is sufficiently large. For M2M communications systems featured with short packets, however, the maximum achievable rate could be significantly lower than the channel capacity due to the small codeword length [8]. In that case, the relationship between the SNR threshold and the information encoding rate needs to be updated by further taking the codeword length into account [41].

4) Infinite Retry Limit: In this paper, we assume that every packet stays in the buffer until it is successfully transmitted. In some M2M applications, however, packets would be dropped after a few failed transmission attempts as the information they carry becomes outdated. The maximum number of allowable retransmissions is usually referred to as the retry limit, which has significant effects on the access performance of packets. It has been shown in [42] that for a CSMA network, the minimum mean access delay can be substantially improved by reducing the retry limit, but at the cost of throughput performance. Such a tradeoff offers important insights to practical system design, which should further be characterized for slotted Aloha networks.

5) Symmetric Setting: In this paper, we consider the symmetric case where all the nodes have identical arrival processes and backoff parameters. For M2M communications systems where devices may have distinct traffic characteristics and parameter settings, the proposed analytical framework can further be extended by grouping devices with the same characteristics into one group, with parameters differing from group to group. For CSMA-based WiFi networks, a multi-group model has been proposed to optimize the throughput performance in various heterogeneous scenarios including diverse traffic input rates [43], service differentiation requirements [44], and multiple standards [45]. It would be interesting to further generalize the delay analysis of slotted Aloha to heterogeneous scenarios.

VI. Conclusion

This paper presents the access delay analysis of slotted Aloha in fading channels. By characterizing closed-form expressions of the network steady-state points in both unsaturated and saturated conditions, the minimum mean access delay and the corresponding optimal transmission probability of each node are obtained as explicit functions of key system parameters including the number of nodes, the aggregate traffic input rate, the mean received SNR and the information encoding rate. The analysis shows that even with the transmission probability of each node optimally tuned to minimize the mean access delay, the effective data rate of each node still diminishes as the number of nodes increases if the information encoding rate of each node is fixed. Therefore, to further take the data rate requirement into consideration, the rate-constrained minimum mean access delay is characterized by jointly optimizing the transmission probability and information encoding rate of each node. Bounds on the minimum required data rate, the mean received SNR of each node and the total number of nodes are also derived, only within which the data rate constraint can be satisfied. To illustrate the practical insights of the analysis for massive access of M2M communications, a single-cell LTE-M network is further considered with smart grid applications in various scenarios. The rate-constrained minimum mean access delay is evaluated for three representative traffic models, which indicates that LTE-M is well suited for massive access of machine-type devices with loose quality-of-service requirements.

Appendix A
Derivation of (16)

It has been shown in [28] that two conditions should be satisfied for the network to operate at p_L:

Condition 1: The network should be unsaturated;

Condition 2: The probability of successful transmission of HOL packets at time slot t, p_t, should be no smaller than p_S.

For Condition 1, the input rate of each queue should be smaller than the service rate, i.e., $\lambda < \pi_T(p_L)$, where the service rate of each node's queue, $\pi_T(p_L)$, can be written as $\pi_T(p_L) = p_L q_0$ with $K = 0$. With $n\lambda = \left(-\ln p_L - \frac{\mu}{\rho}\right) p_L$ according to (10), we can obtain the lower-bound of q_0 as $-\frac{1}{n}\mathbb{W}_0\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right)$.

Note that when $q_0 = -\frac{1}{n}\mathbb{W}_0\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right)$, we have $\lambda = \pi_T(p_L) = \pi_T(p_A)$. In this case, the network operates at the undesired-stable point p_A, but we have $p_A = \exp\left\{-\mathbb{W}_0\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right) - \frac{\mu}{\rho}\right\} = p_L$. Therefore, the lower-bound of q_0 is included in the absolute-stable region S_L.

For Condition 2, the probability of successful transmission of HOL packets at time slot t, p_t, is determined by the number of nodes requesting transmissions at time slot t, which varies with time. Specifically, assume that for a given HOL packet, among the $n-1$ interfering nodes, there are n_i nodes with HOL packets at State i, $i = 0, \ldots, K$. We have

$$\begin{aligned} p_t &= \exp\left\{-\tfrac{\mu}{\rho}\right\} \cdot \prod_{i=0}^{K} (1-q_i)^{n_i} \\ &\geq \exp\left\{-\tfrac{\mu}{\rho}\right\} \cdot (1-q_0)^{\sum_{i=0}^{K} n_i} \\ &\geq \exp\left\{-\tfrac{\mu}{\rho}\right\} \cdot (1-q_0)^{n-1} \overset{n \gg 1}{\approx} \exp\left\{-nq_0 - \tfrac{\mu}{\rho}\right\}, \end{aligned} \quad (41)$$

where the first inequality is due to $q_i \leq q_0$, for $i = 0, \ldots, K$, and the second inequality is due to $\sum_{i=0}^{K} n_i \leq n-1$. We can then obtain the upper-bound of q_0 from $\min_t p_t = \exp\left\{-nq_0 - \frac{\mu}{\rho}\right\} \geq p_S$ as $-\frac{1}{n}\mathbb{W}_{-1}\left(-\hat{\lambda}\exp\left(\frac{\mu}{\rho}\right)\right)$.

Appendix B
Proof of Lemma 1

Proof: The maximum achievable rate of the network $\bar{C} = \max_{\mu>0,\ 0<q_0\leq 1} R_{in} \cdot \hat{\lambda}_{out}$ can be written as

$$\bar{C} = \max_{\mu>0}\left(\log_2(1+\mu) \cdot \max_{0<q_0\leq 1} \hat{\lambda}_{out}\right), \quad (42)$$

according to (3). Section III has shown that the network throughput $\hat{\lambda}_{out}$ closely depends on whether the network is saturated or not. In the following, let us consider two cases:

1) Aggregate input rate $\hat{\lambda} > e^{-1}$: In this case, the network is always saturated and operates at the undesired steady-state point p_A with the corresponding maximum network throughput given in (15). Let C_s denote the maximum achievable rate in the saturated case. By combining (15) and (42), we have

$$C_{\hat{\lambda}>e^{-1}} = C_s = \exp\left(-1-\frac{e^{\mathbb{W}_0(\rho)}-1}{\rho}\right)\cdot\log_2(e^{\mathbb{W}_0(\rho)}), \quad (43)$$

which is achieved when the SNR threshold $\mu = \mu^* = e^{\mathbb{W}_0(\rho)} - 1$.

2) Aggregate input rate $\hat{\lambda} \le e^{-1}$: When the aggregate input rate $\hat{\lambda} \le e^{-1}$, whether the network is saturated or unsaturated depends on the SNR threshold μ. If $\mu \le \mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$, then the network is unsaturated when the initial transmission probability $q_0 \in S_L$. Otherwise, the network is saturated. Let $C_1 = \max\limits_{0<\mu\le\mu_0,\hat{\lambda}\le e^{-1}}\left(\log_2(1+\mu)\cdot\max\limits_{0<q_0\le 1}\hat{\lambda}_{out}\right)$ and $C_2 = \max\limits_{\mu\ge\mu_0,\hat{\lambda}\le e^{-1}}\left(\log_2(1+\mu)\cdot\max\limits_{0<q_0\le 1}\hat{\lambda}_{out}\right)$.

For C_1, it has been shown in (11) that the network throughput $\hat{\lambda}_{out} = \hat{\lambda}$ when the network is unsaturated. Let C_u denote the maximum achievable rate in the unsaturated case. We have

$$C_1 = C_u = \hat{\lambda}\cdot\max_{0<\mu\le\mu_0}\log_2(1+\mu) = \hat{\lambda}\log_2(1-\rho-\rho\ln\hat{\lambda}), \quad (44)$$

achieved when the SNR threshold $\mu = \mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$.

For C_2, as the network is saturated when the SNR threshold $\mu \ge \mu_0$, we have $C_2 = \max\limits_{\mu\ge\mu_0}\exp\left(-1-\frac{\mu}{\rho}\right)\cdot\log_2(1+\mu)$ according to (15). Let $g(\mu) = \exp\left(-1-\frac{\mu}{\rho}\right)\cdot\log_2(1+\mu)$ denote the objective function of C_2. It can be easily obtained that $g(\mu)$ has one global maximum at $\mu^* = e^{\mathbb{W}_0(\rho)} - 1$, with $g(\mu^*) = C_s$ and $g(\mu_0) = C_u$. Moreover, for $\mu \in [\mu_0, +\infty)$, we have i)

1) if $\mu_0 \ge \mu^*$, then $g(\mu)$ is a monotonically decreasing function of μ for $\mu \in [\mu_0, +\infty)$. In this case, C_2 is maximized at C_u when $\mu = \mu_0$, and
2) if $\mu_0 < \mu^*$, then $g(\mu)$ is a monotonically increasing function of μ for $\mu \in [\mu_0, \mu^*]$ and a monotonically decreasing function of μ for $\mu \in (\mu^*, +\infty)$. In this case, C_2 is maximized at C_s when $\mu = \mu^*$.

For $\mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$ and $\mu^* = e^{\mathbb{W}_0(\rho)} - 1$, it can be shown that $\mu_0 \ge \mu^*$ is equivalent to $\hat{\lambda} \le \hat{\lambda}_\rho$, where $\hat{\lambda}_\rho = \exp\left(-1-\frac{e^{\mathbb{W}_0(\rho)}-1}{\rho}\right)$. We can then obtain that $C_2 = C_u$ if $\hat{\lambda} \le \hat{\lambda}_\rho$, achieved when $\mu = \mu_0$, and $C_2 = C_s$ if $\hat{\lambda}_\rho < \hat{\lambda} \le e^{-1}$, achieved when $\mu = \mu^*$. Further note that $C_1 = C_u$ according to (44) and $C_s > C_u$ when $\hat{\lambda}_\rho < \hat{\lambda} \le e^{-1}$, we can obtain that

$$C_{\hat{\lambda}\le e^{-1}} = \max(C_1, C_2) = \begin{cases} C_u & 0 < \hat{\lambda} \le \hat{\lambda}_\rho, \\ C_s & \hat{\lambda}_\rho < \hat{\lambda} \le e^{-1}. \end{cases} \quad (45)$$

Finally, (29) can be obtained by combining (43) and (45). ∎

APPENDIX C
PROOF OF THEOREM 1

Proof: It has been shown in Sections III and IV that both the throughput of each node λ_{out} and the mean access delay $E[D_T]$ depend on whether the network is unsaturated or saturated. When the initial transmission probability q_0 is selected from the absolute-stable region S_L, which exists when $\hat{\lambda} \le e^{-1}$ and $\mu \le \mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$, the network is unsaturated and operates at the desired steady-state point p_L. Otherwise, the network is saturated and operates at the undesired steady-state point p_A. According to

$$\lambda_{out}^{p=p_A} = \pi_T^{p=p_A} < \lambda = \lambda_{out}^{p=p_L} < \pi_T^{p=p_L}, \quad (46)$$

we have

$$E[D_T]^{p=p_L} < \frac{1}{\lambda} < E[D_T]^{p=p_A} = \frac{1}{\lambda_{out}^{p=p_A}}, \quad (47)$$

because $E[D_T] = \frac{1}{\pi_T}$ by combining (1) and (22).

Let $D_R^{*,p=p_L}$ and $D_R^{*,p=p_A}$ denote the rate-constrained minimum mean access delay in the unsaturated and saturated cases, respectively. We have $D_R^* = \min(D_R^{*,p=p_L}, D_R^{*,p=p_A})$. By combining (3), (28), (46) and (47), they can be written as

$$\begin{aligned} D_R^{*,p=p_L} = \min_{0<\mu\le\mu_0}\ \min_{q_0\in S_L} & E[D_T]^{p=p_L} \\ \text{s.t.}\ \ & 0 < \hat{\lambda} \le e^{-1}, \\ & \lambda\log_2(1+\mu) \ge R_0, \end{aligned} \quad (48)$$

and

$$\begin{aligned} D_R^{*,p=p_A} = \min_{\mu>0}\ \min_{q_0\notin S_L, 0<q_0\le 1} & E[D_T]^{p=p_A} \\ \text{s.t.}\ \ & \frac{1}{E[D_T]^{p=p_A}}\cdot\log_2(1+\mu) \ge R_0. \end{aligned} \quad (49)$$

In the following, we will consider $D_R^{*,p=p_L}$ and $D_R^{*,p=p_A}$ separately.

For $D_R^{*,p=p_L}$, it has been shown in Section IV-B that $\min\limits_{q_0\in S_L} E[D_T]^{p=p_L} = \frac{\mathbb{W}_0\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}{\lambda\mathbb{W}_{-1}\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}$, achieved when the initial transmission probability $q_0 = -\frac{1}{n}\mathbb{W}_{-1}\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)$. Accordingly, (48) can be rewritten as $D_R^{*,p=p_L} = \min\limits_{2^{R_0/\lambda}-1\le\mu\le\mu_0}\frac{\mathbb{W}_0\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}{\lambda\mathbb{W}_{-1}\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}$ for $\hat{\lambda} \le e^{-1}$. Note that for the existence of $D_R^{*,p=p_L}$, we need $\mu_0 \ge 2^{R_0/\lambda} - 1$, or equivalently, $R_0 \le \frac{C_u}{n}$ according to $\mu \le \mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right)$ and (30). As $\frac{\mathbb{W}_0\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}{\lambda\mathbb{W}_{-1}\left(-n\lambda\exp\left(\frac{\mu}{\rho}\right)\right)}$ is a monotonic increasing function of the SNR threshold μ, we have

$$D_R^{*,p=p_L} = \frac{\mathbb{W}_0\left(-n\lambda\exp\left(\left(2^{\frac{R_0}{\lambda}}-1\right)/\rho\right)\right)}{\lambda\mathbb{W}_{-1}\left(-n\lambda\exp\left(\left(2^{\frac{R_0}{\lambda}}-1\right)/\rho\right)\right)}, \quad (50)$$

for $0 < \hat{\lambda} \le e^{-1}$ and $0 < R_0 \le \frac{C_u}{n}$, achieved when the SNR threshold $\mu = 2^{R_0/\lambda} - 1$.

For $D_R^{*,p=p_A}$, note that if the network is saturated, then the initial transmission probability q_0 should not be selected from the absolute-stable region S_L. As S_L exists only when

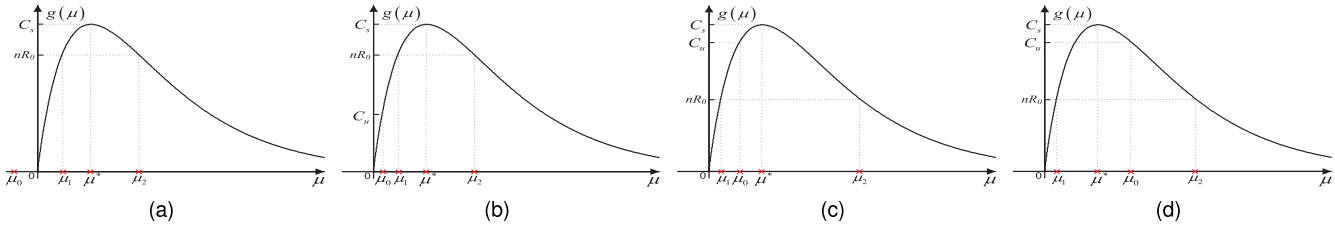


Fig. 12. $g(\mu)$ versus μ. (a) $\mu_0 \le 0 < \mu_1 \le \mu^*$. (b) $0 < \mu_0 < \mu_1 \le \mu^*$. (c) $0 < \mu_1 \le \mu_0 \le \mu^*$. (d) $0 < \mu_1 \le \mu^* \le \mu_0$.

$\hat{\lambda} \le e^{-1}$ and $\mu \le \mu_0$, in the following, we will consider the delay performance in two cases: 1) $q_0 \notin S_L$, $0 < \hat{\lambda} \le e^{-1}$ and $0 < \mu \le \mu_0$, and 2) $\mu \ge \mu_0$.

1) $q_0 \notin S_L$, $0 < \hat{\lambda} \le e^{-1}$ and $0 < \mu \le \mu_0$: According to (46) and (47), we have $\min_{q_0 \in S_L} E[D_T]^{p=p_L} < \min_{q_0 \notin S_L, 0<q_0\le 1} E[D_T]^{p=p_A}$ and $\lambda > \frac{1}{E[D_T]^{p=p_A}}$. We can then conclude from (48) and (49) that in this case, $D_R^{*,p=p_L} < D_R^{*,p=p_A}$, and D_R^* is given in (50).

2) $\mu \ge \mu_0$: It has been shown in Section IV that $\min_{q_0 \notin S_L, 0<q_0\le 1} E[D_T]^{p=p_A} = n \exp\left(1 + \frac{\mu}{\rho}\right)$, achieved when the initial transmission probability $q_0 = \frac{1}{n}$. Accordingly, (49) can be rewritten as

$$D_R^{*,p=p_A} = \min_{\mu \ge \mu_0} \; n \exp\left(1 + \frac{\mu}{\rho}\right) \quad \text{s.t.} \quad \exp\left(-1 - \frac{\mu}{\rho}\right) \cdot \log_2(1+\mu) \ge nR_0. \tag{51}$$

Let $g(\mu) = \exp\left(-1 - \frac{\mu}{\rho}\right) \cdot \log_2(1+\mu)$. It has been shown in Appendix B that $g(\mu)$ has one global maximum at $\mu^* = e^{\mathbb{W}_0(\rho)} - 1$ with $g(\mu^*) = C_s$ and $g(\mu_0) = C_u$. It can be seen from (51) that the objective function is a monotonically increasing function of μ. Therefore, we aim to find the minimum μ for satisfying the constraints of $\mu \ge \mu_0$ and $g(\mu) \ge nR_0$. As Fig. 12 illustrates, $g(\mu) = nR_0$ has two roots when $nR_0 < C_s$. Denote the smaller root as μ_1 and the larger root as μ_2. Apparently, the constraint $g(\mu) \ge nR_0$ cannot be satisfied when $\mu_0 > \mu_2$. For $\mu_0 \le \mu_2$, let us consider the following cases:

i) $\mu_0 \le 0 < \mu_1$ and $0 < \mu_0 < \mu_1$: As Fig. 12a and Fig. 12b illustrate, in both cases, the minimum μ for satisfying $g(\mu) \ge nR_0$ and $\mu \ge \mu_0$ is μ_1. We then have $D_R^{*,p=p_A} = n \exp\left(1 + \frac{\mu_1}{\rho}\right)$. Further note that $\mu_0 = \rho\left(\ln\frac{1}{\hat{\lambda}} - 1\right) \le 0$ is equivalent to $\hat{\lambda} \ge e^{-1}$, and $\mu_0 < \mu^* = e^{\mathbb{W}_0(\rho)} - 1$ is equivalent to $\hat{\lambda} > \hat{\lambda}_\rho = \exp\left(-1 - \frac{e^{\mathbb{W}_0(\rho)} - 1}{\rho}\right)$. We can then conclude that $D_R^{*,p=p_A} = n \exp\left(1 + \frac{\mu_1}{\rho}\right)$, for $\hat{\lambda} \ge e^{-1}$ and $0 < R_0 \le \frac{C_s}{n}$, or $\hat{\lambda}_\rho < \hat{\lambda} < e^{-1}$ and $\frac{C_u}{n} < R_0 \le \frac{C_s}{n}$. As $D_R^{*,p=p_L}$ does not exist in the above cases, we have

$$D_R^* = D_R^{*,p=p_A} = n \exp\left(1 + \frac{\mu_1}{\rho}\right), \tag{52}$$

for $\hat{\lambda} \ge e^{-1}$ and $0 < R_0 \le \frac{C_s}{n}$, or $\hat{\lambda}_\rho < \hat{\lambda} < e^{-1}$ and $\frac{C_u}{n} < R_0 \le \frac{C_s}{n}$, achieved when $\mu = \mu_1$.

ii) $0 < \mu_1 \le \mu_0$: As Fig. 12c and Fig. 12d illustrate, in this case, the minimum μ for satisfying $g(\mu) \ge nR_0$ and $\mu \ge \mu_0$ is μ_0. We then have $D_R^{*,p=p_A} = n \exp\left(1 + \frac{\mu_0}{\rho}\right) = \frac{1}{\lambda}$. Further note that $\mu_0 > 0$ is equivalent to $\hat{\lambda} < e^{-1}$. We can then conclude that $D_R^{*,p=p_A} = \frac{1}{\lambda}$, for $0 < \hat{\lambda} < e^{-1}$ and $0 < R_0 \le \frac{C_u}{n}$. By comparing with (50) and noting that $\frac{\mathbb{W}_0(x)}{\mathbb{W}_{-1}(x)} < 1$ for $x \in (-e^{-1}, 0)$, we have $D_R^{*,p=p_L} < \frac{1}{\lambda} = D_R^{*,p=p_A}$ in this case, and D_R^* is given in (50).

Finally, (32) and (33) can be obtained by combining (50) and (52). (34) can be obtained by combining (27) and (33). ■

REFERENCES

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

[2] L. Kleinrock and F. Tobagi, "Packet switching in radio channels: Part I-carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE Trans. Commun.*, vol. COM-23, no. 12, pp. 1400–1416, Dec. 1975.

[3] 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.

[4] J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, 3rd ed. Reading, MA, USA: Addison-Wesley, 2004.

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

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

[7] 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.

[8] 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.

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

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

[11] M. Ferguson, "On the control, stability, and waiting time in a slotted ALOHA random-access system," *IEEE Trans. Commun.*, vol. COM-23, no. 11, pp. 1306–1311, Nov. 1975.

[12] D. J. Goodman and A. A. M. Saleh, "The near/far effect in local ALOHA radio communications," *IEEE Trans. Veh. Technol.*, vol. VT-36, no. 1, pp. 19–27, Feb. 1987.

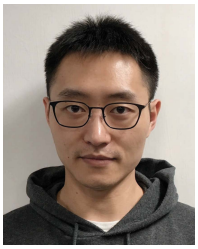
[13] Y. Yang and T.-S. P. Yum, "Delay distributions of slotted ALOHA and CSMA," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1846–1857, Nov. 2003.

[14] Y.-J. Choi, S. Park, and S. Bahk, "Multichannel random access in OFDMA wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 603–613, Mar. 2006.

[15] B. J. Kwak, N. O. Song, and L. E. Miller, "Performance analysis of exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 343–353, Apr. 2005.

[16] L. Barletta, F. Borgonovo, and I. Filippini, "The throughput and access delay of slotted-ALOHA with exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 451–464, Feb. 2018.
[17] W. Yue, "The effect of capture on performance of multichannel slotted ALOHA systems," *IEEE Trans. Commun.*, vol. 39, no. 6, pp. 818–822, Jun. 1991.
[18] M. E. Rivero-Angeles, D. Lara-Rodriguez, and F. A. Cruz-Perez, "Gaussian approximations for the probability mass function of the access delay for different backoff policies in S-ALOHA," *IEEE Commun. Lett.*, vol. 10, no. 10, pp. 731–733, Oct. 2006.
[19] F. A. Tobagi, "Distributions of packet delay and interdeparture time in slotted ALOHA and carrier sense multiple access," *J. ACM*, vol. 29, no. 4, pp. 907–927, Oct. 1982.
[20] M. Sidi and A. Segall, "Two interfering queues in packet-radio networks," *IEEE Trans. Commun.*, vol. COM-31, no. 1, pp. 123–129, Jan. 1983.
[21] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted ALOHA with mutipacket reception," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2636–2656, Jul. 2005.
[22] I. Dimitriou and N. Pappas, "Stable throughput and delay analysis of a random access network with queue-aware transmission," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3170–3184, May 2018.
[23] T. Saadawi and A. Ephremides, "Analysis, stability, and optimization of slotted ALOHA with a finite number of buffered users," *IEEE Trans. Autom. Control*, vol. AC-26, no. 3, pp. 680–689, Jun. 1981.
[24] A. Ephremides and R.-Z. Zhu, "Delay analysis of interacting queues with an approximate model," *IEEE Trans. Commun.*, vol. COM-35, no. 2, pp. 194–201, Feb. 1987.
[25] S. Rasool and A. Sheikh, "An approximate analysis of buffered S-ALOHA in fading channels using tagged user analysis," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1320–1326, Apr. 2007.
[26] S. C. Liew, Y. Zhang, and D. R. Chen, "Bounded-mean-delay throughput and nonstarvation conditions in ALOHA network," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1606–1618, Oct. 2009.
[27] A. Mutairi, S. Roy, and G. Hwang, "Delay analysis of OFDMA-ALOHA," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 89–99, Jan. 2013.
[28] L. Dai, "Stability and delay analysis of buffered ALOHA networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2707–2719, Aug. 2012.
[29] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[30] Y. Li and L. Dai, "Maximum sum rate of slotted ALOHA with capture," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 690–705, Feb. 2016.
[31] Y. Li and L. Dai, "Maximum sum rate of slotted ALOHA with successive interference cancellation," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5385–5400, Nov. 2018.
[32] Sunseed. (Mar. 2015). *Traffic Modelling, Communication Requirements and Candidate Network Solutions for Real-Time Smart Grid Control, Version 2.0*. [Online]. Available: http://sunseed-fp7.eu/wp-content/uploads/2015/04/SUNSEED-WP3-D31-V20-03072015.pdf
[33] C. van der Plas and J.-P.-M. G. Linnartz, "Stability of mobile slotted ALOHA network with Rayleigh fading, shadowing, and near-far effect," *IEEE Trans. Veh. Technol.*, vol. 39, no. 4, pp. 359–366, Nov. 1990.
[34] J.-P.-M. G. Linnartz, R. Hekmat, and R.-J. Venema, "Near-far effects in land mobile random access networks with narrow-band Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 41, no. 1, pp. 77–90, Feb. 1992.
[35] X. Sun and L. Dai, "Fairness-constrained maximum sum rate of multi-rate CSMA networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1741–1754, Mar. 2017.
[36] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.
[37] *Smart Grid Traffic Behaviour Discussion*, document TSG RAN WG2 #69b R2-102340, 3GPP, Apr. 2010.
[38] O. Liberg, M. Sundberg, Y. P. E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things Technologies, Standards and Performance*. Amsterdam, The Netherlands: Elsevier, 2019.
[39] A. Höglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.
[40] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, Dec. 2017.
[41] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[42] 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.
[43] Y. Gao, X. Sun, and L. Dai, "Throughput optimization of heterogeneous IEEE 802.11 DCF networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 398–411, Jan. 2013.
[44] Y. Gao, X. Sun, and L. Dai, "IEEE 802.11e EDCA networks: Modeling, differentiation and optimization," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3863–3879, Jul. 2014.
[45] Y. Gao, X. Sun, and L. Dai, "Sum rate optimization of multi-standard IEEE 802.11 WLANs," *IEEE Trans. Commun.*, vol. 67, no. 4, pp. 3055–3068, Apr. 2019.

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



Wen Zhan (Member, IEEE) received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, 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 with the City University of Hong Kong, where he held a post-doctoral position. Since 2020, he has been with the School of Electronics and Communication Engineering, Sun Yat-sen University, China, where he is currently an Assistant Professor. His research interests include the Internet of Things and modeling and performance optimization of next-generation mobile communication systems.

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