Received 11 July 2022; revised 10 August 2022; accepted 29 August 2022. Date of publication 1 September 2022; date of current version 16 September 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3203803

Sum Rate and Access Delay Optimization of Short-Packet Aloha

XINGHUA SUN[1] (Member, IEEE), WEN ZHAN[1] (Member, IEEE), WEIHUA LIU[1], YITONG LI[2], AND QI LIU[3,4] (Member, IEEE)

[1]School of Electronics and Communication Engineering, Sun Yat-sen University (Shenzhen Campus), Shenzhen 518107, China

[2]School of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China

[3]School of Future Technology, South China University of Technology, Guangzhou 511442, Guangdong, China

[4]Pazhou Lab, Guangzhou 510330, China

CORRESPONDING AUTHOR: W. ZHAN (e-mail: zhanw6@mail.sysu.edu.cn)

The work of Xinghua Sun was supported in part by the National Key Research and Development Program of China under Grant 2019YFE0196400, and in part by the Guangdong Engineering Technology Research Center for Integrated Space-Terrestrial Wireless Optical Communication. The work of Wen Zhan was supported in part by the National Natural Science Foundation of China under Grant 62001524, and in part by the Shenzhen Science and Technology Program under Grant RCBS20210706092408010 and Grant 2021A04. The work of Yitong Li was supported in part by the National Natural Science Foundation of China under Grant 61801433, and in part by the Science and Technology Planning Project of Henan Province under Grant 222102210278. The work of Qi Liu was supported by the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004. Part of this work was presented at the IEEE/CIC ICCC, August 2020 [1] [DOI: 10.1109/ICCC49849.2020.9238933].

ABSTRACT Shortening the packet length has been a consensus in wireless network design for supporting the ultra-low latency Internet of Things (IoT) applications. Yet, with short-packet transmission, the rate loss would occur, which further depends on the blocklength, making the network optimization notoriously difficult, especially for random access networks. This paper focuses on the representative random access network, i.e., Aloha, with short packet transmission, namely, short-packet Aloha. Specifically, we aim to optimize the sum rate and access delay of short-packet Aloha. By deriving the probability of successful transmissions of packets, both the network sum rate and the probability generating function of access delay are obtained as explicit functions of key system parameters. The maximum sum rate and the minimum mean access delay are further derived by jointly tuning the packet transmission probability and the blocklength of packets. The effect of system parameters on the optimal sum rate and access delay performance is investigated. It is shown that the maximum sum rate is insensitive to the retry limit M, while deteriorates as the information bits per packet k decreases. In contrast, the optimal delay performance can be improved with a small M or k. The reliability performance is also evaluated and shown to be enhanced with a large retry limit M. The analysis sheds important light on the access design of practical short-packet Aloha networks. By taking LTE-M as an example, it is found that to improve access delay performance, the information bits per packet k should not exceed an upperbound, which polynomially decreases as the network size increases.

INDEX TERMS Aloha, finite block length region, low latency, maximum sum rate, short-packet.

I. INTRODUCTION

THE EMERGING mission-critical Internet of Things (IoT) applications, such as autonomous driving, remote surgery and smart grid automation, require Ultra-Reliable Low Latency Communication (URLLC). For instance, autonomous driving usually requires status messages delivered within less than 10 ms to enable cooperative vehicle maneuver, dense platooning, and so on [2]. To meet the ever-demanding latency requirement, various wireless techniques and paradigms have been proposed, among which the grant-free random access scheme along with short-packet transmission gains significant attention.

Short-packet transmission is a straightforward way for latency reduction and the main feature of IoT communications, which adopts finite blocklength codewords for data transmissions. However, packets would experience decoding error due to the operation in the finite blocklength regime, and the error probability is a complicated function of the blocklength [3], with which the network performance optimization becomes challenging. This issue gets even more difficult with grant-free random access scheme, where each transmitter sends its packet to the receiver without seeking the grant from a central controller. With distributed nature of access behavior of nodes, the number of active transmitters varies over time and the service rate of each node's data queue is determined by the aggregate activities of all nodes. Therefore, how to optimize the network performance of the grant-free random access scheme with short-packet transmission is a challenging issue that remains largely unexplored.

A. DELAY ANALYSIS OF ALOHA

Due to the contention in the shared channel, the grant-free random access often suffers from low efficiency and poor delay performance as successful channel access cannot be guaranteed. In this paper, we focus on the representative grant-free random access scheme, i.e., slotted Aloha.

Extensive works have been done on evaluating the delay performance of the slotted Aloha networks. Specifically, by approximating the channel aggregate traffic as a Poisson variable, the mean access delay [4], [5], [6], [7], and the probability mass function of access delay [8], [9] have been characterized. By assuming a saturated network, i.e., each user always has packets to transmit, the mean access delay was characterized in [10], while the probability mass function of access delay was derived in [11]. By further taking the queueing dynamics into consideration, the one-packet buffer assumption was assumed in [12], [13], [14] while a buffer with infinite length was considered in [15], [16]. General cases with a finite buffer length and various queueing models were further investigated in [17], [18], [19], [20] where the probability generating functions of access delay were given. It was found that the delay performance deteriorates as the packet arrival rate [14], [17] or the number of nodes [11], [17] increases. A tradeoff between the throughput and delay performance was observed in [8], [16], [19], [20].

Above studies mainly focus on numerical evaluation of the delay performance of slotted Aloha networks for given system parameters. How to adjust system configurations to optimize the network performance is another interesting problem. In [21], [22], an analytical framework of Aloha networks was proposed, where explicit expressions of moments of access delay were derived, which further enabled delay optimization. The analysis was extended to the scenario with a finite retry limit in [23], where the packet is dropped if the number of retransmissions reaches this limit. It was shown that to optimize the delay performance, the packet transmission probability and the retry limit should be carefully tuned according to the network size.

B. SHORT-PACKET TRANSMISSION

To satisfy the stringent latency requirement of the mission-critical IoT applications, a consensus has gradually come into being: the packet length should be shorten. It has both its pros and cons. The advantages lie in low signalling overhead, energy consumption and delay. However, the short-packet transmission also leads to a large decoding error probability and loss of information encoding rate.

Specifically, the classical Shannon capacity of the point-to-point channel assumed that the codeword length is long enough. In the finite blocklength regime, nevertheless, the rate loss would occur [24]. Tight approximation of the point-to-point channel capacity in the finite blocklength regime has been characterized in [3], [25]. Based on the approximation, a tradeoff among the reliability, throughput and latency performance has been characterized [26]. An energy-efficient packet scheduling scheme was proposed in [27] by optimizing the blocklength and transmission power under delay constraints. Performances of Automatic Repeat reQuest (ARQ) mechanisms were discussed in [28], [29] to find suitable values of the blocklength.

For the multiple access channel, the maximum achievable rate region was characterized from an information-theoretical perspective of view in [30]. For the grant-free random access, [31] considered an Aloha network in which nodes tune their packet transmission probabilities to satisfy distinct delay guarantees. Yet, the blocklength is fixed in these studies. Note that in the finite blocklength regime, the blocklength has a significant impact on the network reliability and delay performances. With a larger blocklength, the chance of successful decoding is enhanced while the packet transmission time is enlarged. As such, the blocklength becomes an important parameter that should be carefully selected. A TDMA scheme was considered in [32], [33], where the network throughput is maximized by jointly tuning the packet length and the packet error rate [32], or by tuning the number of information bits per packet [33].

For Aloha networks in the finite blocklength region, both the blocklength and the packet transmission probability could be jointly tuned to optimize the network performance. Yet no systematic study has been conducted towards that purpose. As a result, how to optimize the rate, delay and reliability performance of the short-packet Aloha networks remains unknown.

C. OUR CONTRIBUTIONS

To address above open issues, in this paper, the analytical framework in [21] is extended to analyze the short-packet n-node Aloha network with packet dropping. Specifically, each node transmits a packet to a single receiver via an Additive White Gaussian Noise (AWGN) channel with a certain probability in each time slot. If the packet transmission fails for M times, then the packet is dropped. Each node encodes k information bits into one packet of blocklength N. Both the network sum rate and the mean access delay are derived as explicit functions of the system parameters including the packet transmission probability, blocklength N and retry limit M. Based on these expressions, the optimal rate and delay performance is characterized by jointly tuning the initial transmission probability of data packets q_0 and the blocklength N.

We start by focusing on the network sum rate performance. The analysis shows that the maximum sum rate can be improved with a larger information bits per packet k while is independent of the retry limit M. To achieve the maximum sum rate, however, the corresponding optimal initial transmission probability of data packets should be properly enlarged as M increases if the binary exponential backoff is adopted. On the other hand, for the access delay performance, it is found that the mean access delay of successfully-transmitted packets ED decreases as the initial transmission probability of data packets q_0 increases or the retry limit M decreases. With $M \rightarrow \infty$, i.e., there is no packet dropping, the network sum rate and the mean access delay can be optimized simultaneously by optimally tuning q_0 and the blocklength N. By further considering the reliability performance, the tradeoff between the access delay and reliability is revealed, where the minimum mean access delay deteriorates while the reliability can be greatly enhanced with a larger M. Since the reliability is insensitive to k and ED can be reduced with a smaller k,[1] a combination of a large retry limit M and a small number of information bits per packet k is suggested to balance the tradeoff between the access delay and reliability.

The analysis is further applied to a single-cell LTE-M system, where the minimum mean access delay is characterized. It is found that in order to satisfy the requirement of the mean access delay, the number of information bits per packet k should not exceed a certain threshold, and the threshold value polynomially decreases with the network size n.

The remainder of this paper is organized as follows. Section II first presents the system model. The maximum sum rate and the minimum mean access delay are derived in Section III and Section IV, respectively. Section V presents how the analysis can be applied by an example of a single-cell LTE-M network. Conclusions are drawn in Section VI.

II. SYSTEM MODEL

Consider a n-node slotted Aloha network with one common receiver. Assume that the network is saturated, i.e., each node always has packets to send.[2] Each node starts a transmission only at the beginning of a time slot, and receives a perfect feedback of the transmission outcome from the receiver by the end of the time slot.

In this paper, we consider an AWGN channel between each transmitter and the receiver, and the mean received SNR is identical for each node,[3] which is denoted as ρ.

A. PACKET ENCODING AND DECODING

Since Aloha was first proposed, a time-slotted and packet-based network has been assumed [4], [5], [6], [7], [8], [9], [11], [12], [13], [14], [15], [17], [18], [19], [20], [31], [34], [35]. Here we also assume that each node independently encodes k information bits into a codeword, i.e., a sequence of symbols, as one packet. Each packet lasts for only one time slot, that is, no coding over successive time slots. Since the channel statistics are identical, it is commonly assumed that each node has the same information encoding rate. Let the number of symbols N denote the blocklength, i.e., the packet length. The information encoding rate R can then be denoted as the ratio of the number of information bits in each packet k to the blocklength N,

$$R = \frac{k}{N}. \tag{1}$$

For a point-to-point AWGN channel, with a large blocklength N, the information encoding rate R can be close to the channel capacity $\log_2(1+\rho)$ by random coding, where ρ is the mean received SNR. In the finite blocklength region, nevertheless, rate loss and decoding error would occur. The information encoding rate is approximately given by [3]

$$R \approx \log_2(1+\rho) - \sqrt{\frac{V}{N}} Q^{-1}(\epsilon) + \frac{1}{2N}\log_2 N, \tag{2}$$

where $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian Q function, the channel dispersion V is given by $V = \rho\frac{2+\rho}{1+\rho^2}(\log_2 e)^2$, and ϵ denotes the packet error probability, which can be obtained as

$$\epsilon = Q\left(\frac{N\log_2(1+\rho) - k + \left(\log_2 N\right)/2}{\sqrt{NV}}\right), \tag{3}$$

by combining (1) and (2).

In this paper, we assume a single-user detector at the receiver, i.e., each node's packet is decoded independently by treating others' as background noise. For one node that has concurrent transmissions from other i nodes, its received

[1] Different from the infinite blocklength case in [21], [34] where the access delay is evaluated in unit of time slots, with finite blocklength, the access delay is evaluated in unit of seconds which depends on the blocklength and thus becomes a function of the number of information bits per packet.

[2] Note that the network sum rate is pushed to the limit under saturated condition. For the delay performance, the saturated condition presents a worst case. As the queueing delay (which includes the waiting time in each node's queue) would become unbounded, we focus on the access delay performance in this paper.

[3] If nodes have different mean received SNRs, then those with larger mean received SNRs would have a larger information encoding rate. Identical received SNR is assumed here to ensure fairness among nodes, that is, all the nodes have the same rate performance.

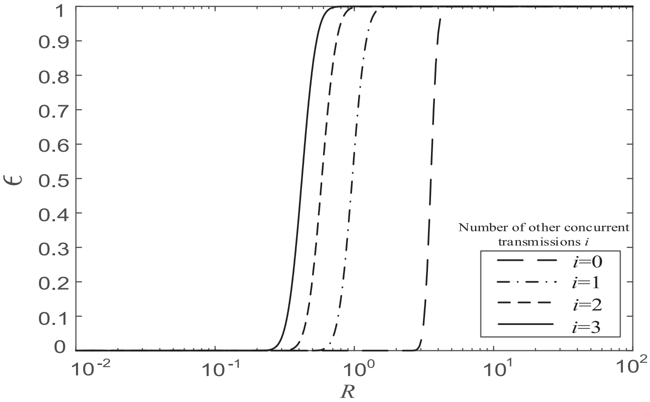


FIGURE 1. Packet error probability ϵ versus the information encoding rate R. $k = 100$, $\rho = 10$ dB.

SINR is given by $\frac{1}{i+1/\rho}$, where $i \in \{0, 1, \dots, n-1\}$. Then the packet error probability becomes

$$\epsilon = Q\left(\frac{N \log_2\left(1 + \frac{1}{i+1/\rho}\right) - k + \left(\log_2 N\right)/2}{\sqrt{NV}}\right). \tag{4}$$

Therefore, whether one packet can be successfully decoded or not critically depends on the information encoding rate R and the number of concurrent transmissions according to (4). As each node decides when to transmit by its own, the subset of active nodes randomly varies over each time slot.

Fig. 1 illustrates how the packet error probability ϵ varies with the information encoding rate R given different numbers of concurrent transmissions. With a fixed number of information bits per packet k, each node can tune[4] the blocklength N so as to achieve a varied value of the information encoding rate R according to (1). The corresponding packet error probability is then derived according to (4). We can observe from Fig. 1 that the packet error probability ϵ increases as R increases, and deteriorates sharply within a small range of R. For instance, with $k = 100$, ϵ increases almost from 0 to 1 when R grows from 2.86 bit/s/Hz to 4.55 bit/s/Hz. Moreover, it can be seen from Fig. 1 that for a given information encoding rate R, the packet error probability ϵ increases significantly as the number of concurrent transmissions increases. Motivated by this observation, we adopt a useful simplification of the receiver, i.e., one node's packet transmission would fail if there exist concurrent transmissions from other nodes.[5] Therefore, the successful transmission of one packet is conditioned on no other concurrent transmissions and no decoding error due to the operation in the finite blocklength region.

Moreover, each node would transmit its head-of-line (HOL) packet with probability q_i in each time slot if the HOL packet has experienced ith transmission failure. To avoid excessive access delay, we consider a finite retry limit M, where the HOL packet is dropped after M-th transmission failure. Without loss of generality, let $q_i = q_0 \mathcal{Q}(i)$, where q_0 is the initial packet transmission probability, $\mathcal{Q}(0) = 1$ and $\mathcal{Q}(i) \le \mathcal{Q}(i-1)$, $i = 1, \dots, M$.

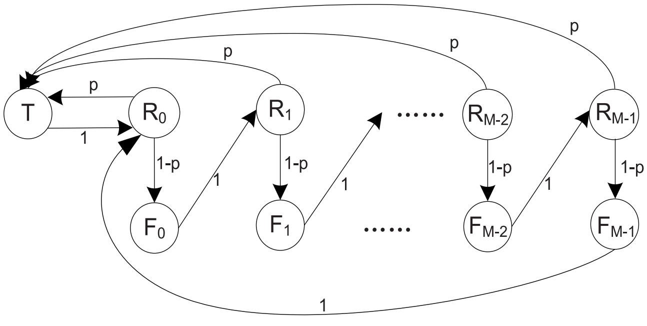


FIGURE 2. Embedded Markov chain of the state transition process of an individual HOL packet with retry limit M.

B. STATE CHARACTERIZATION OF HOL PACKETS

The behavior of the HOL packet in each node's queue in an Aloha network can also be characterized by an embedded Markov chain.[6] As Fig. 2 illustrates, the states of the embedded Markov chain include 1) waiting to request (State R_i, $i = 0, \dots, M-1$) 2) collision (State F_i, $i = 0, \dots, M-1$) and 3) successful transmission (State T). If the transmission succeeds, the HOL packet shifts from State R_i to State T; otherwise, it remains in State F_i until the end of the failure and then moves to State R_{i+1}. A HOL packet would be dropped if it is retransmitted for M times, i.e., after State F_{M-1}, and then a new HOL packet enters State R_0.

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

$$\pi_{R_i} = \frac{(1-p)^i}{1-(1-p)^M} \cdot \pi_T, \tag{5}$$

and

$$\pi_{F_i} = \frac{(1-p)^{i+1}}{1-(1-p)^M} \cdot \pi_T, \tag{6}$$

$i = 0, \dots, M-1$, where p denotes the steady-state probability of successful transmissions of HOL packets.

Since both State T and State F_i indicate that the packet is transmitting, the mean holding time τ_T and the mean holding time τ_{F_i} should be equal to the transmission time of each packet, i.e., one time slot. For State R_i, the mean holding time τ_{R_i} is given by the expected time interval at this state before the HOL packet is transmitted. Recall that q_i denotes the transmission probability of a State-R_i HOL packet in each time slot. The time interval from the instance the HOL packet enters State R_i to the instance it is transmitted is therefore a

[4] To be more specific, as unit bandwidth is assumed in (2), the encoding block varies in the time domain.

[5] This is de facto the classic collision model that is widely adopted in previous studies [4], [7], [8], [9], [10], [11], [12], [13], [14], [15], [17], [18], [20], [36], [37], [38].

[6] Note that a similar embedded Markov chain has been established for CSMA networks in [38]. Different from that in [38], the mean sojourn time of each HOL packet in State R_i does not depend on the channel status, i.e., whether the channel is idle or not, since nodes do not regulate the access behavior according to the channel status in Aloha networks.

geometrically distributed random variable, whose expected value is given by $\frac{1}{q_i}$. As each packet transmission lasts for one time slot, the mean sojourn time of each HOL packet in State R_i, can then be obtained as

$$\tau_{R_i} = \frac{1}{q_i} - 1. \tag{7}$$

Finally, the limiting state probabilities are given by

$$\tilde{\pi}_j = \frac{\pi_j \cdot \tau_j}{\pi_T \cdot \tau_T + \sum_{i=0}^{M-1} \pi_{F_i} \cdot \tau_F + \sum_{i=0}^{M-1} \pi_{R_i} \cdot \tau_{R_i}}, \tag{8}$$

$j \in \{R_0, \ldots, R_{M-1}, T, F_0, \ldots, F_{M-1}\}$. Note that the limiting state probability represents the probability of one HOL packet being in each state. In particular, the probability of one HOL packet being in State T is given by

$$\tilde{\pi}_T = \frac{1}{\frac{1}{1-(1-p)^M} \sum_{i=0}^{M-1} \frac{(1-p)^i}{q_i}}, \tag{9}$$

according to (5)-(8).

C. PERFORMANCE METRICS

To evaluate the network performance, in this paper, we consider three key performance metrics:

- Network sum rate C: the mean successfully-transmitted information bits per channel use, i.e.,

$$C = \hat{\lambda}_{out} \cdot R, \tag{10}$$

  where $\hat{\lambda}_{out}$ is the network throughput, which equals the long-term fraction of time slots that have successful packet transmissions.
- Mean access delay of successfully-transmitted packets ED: the mean interval from the time slot one packet becomes HOL until the time slot it is successfully transmitted.
- Reliability η: the probability that a HOL packet is not dropped, i.e.,

$$\eta = 1 - (1-p)^M, \tag{11}$$

  where M is the retry limit.

In the following sections, we study how to optimize the network sum rate performance and mean access delay performance in Section III and Section IV, respectively, by properly tuning the initial transmission probability and blocklength. The discussion on the reliability performance will also be presented.

III. NETWORK SUM RATE OPTIMIZATION

In this section, we first derive the probability of successful transmissions of HOL packets, and then characterize the network sum rate, and finally optimize the sum rate performance by tuning the initial packet transmission probability q_0 and packet blocklength N.

A. PROBABILITY OF SUCCESSFUL TRANSMISSIONS OF HOL PACKETS

As one HOL packet is successfully received if there are no other concurrent transmissions and it is decoded by the receiver, we have

$$\begin{aligned} p = \Pr\{&\text{all the other HOL packets do not attempt} \\ &\text{to access the channel}\} \cdot (1-\epsilon), \end{aligned} \tag{12}$$

where ϵ is the packet error probability. According to Fig. 2, we have

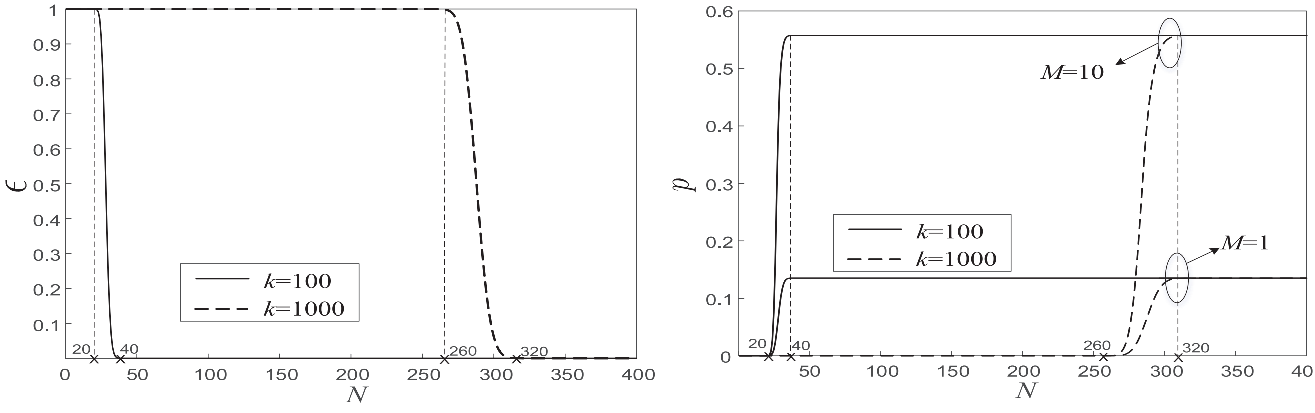
$$p = \left\{1 - \sum_{i=0}^{M-1} \tilde{\pi}_{R_i} \frac{1}{\tau_{R_i}}\right\}^{n-1} \cdot (1-\epsilon). \tag{13}$$

Recall that the sojourn time of each HOL packet in State R_i, τ_{R_i}, is given by $\frac{1}{q_i} - 1$ according to (7). When n is large, by further combining (7)–(9) and (13), we then have[7]

$$\begin{aligned} p &= \exp\left\{-n \sum_{j=0}^{M-1} \tilde{\pi}_{R_j} \frac{q_j}{1-q_j}\right\} \cdot (1-\epsilon) \\ &= \exp\left\{-n \sum_{j=0}^{M-1} \frac{\pi_{R_j} \cdot \tau_{R_j} \cdot \frac{q_j}{1-q_j}}{\pi_T \cdot \tau_T + \sum_{i=0}^{M-1} \pi_{F_i} \cdot \tau_F + \sum_{i=0}^{M-1} \pi_{R_i} \cdot \tau_{R_i}}\right\} \\ &\quad \cdot (1-\epsilon) \\ &= \exp\left\{-n \sum_{j=0}^{M-1} \pi_{R_j} \frac{\tilde{\pi}_T}{\pi_T}\right\} \cdot (1-\epsilon) \\ &= \exp\left\{-n \sum_{j=0}^{M-1} \frac{(1-p)^j}{1-(1-p)^M} \tilde{\pi}_T\right\} \cdot (1-\epsilon) \\ &= \exp\left\{-\frac{n\tilde{\pi}_T}{p}\right\}(1-\epsilon) = \exp\left\{-\frac{n}{\frac{p\sum_{i=0}^{M-1} \frac{(1-p)^i}{q_i}}{1-(1-p)^M}}\right\}(1-\epsilon). \end{aligned} \tag{14}$$

Fig. 3 demonstrates how the packet error probability ϵ and the steady-state probability of successful transmissions of HOL packets p vary with the blocklength N with the information bits per packet $k = 100$ or 1000 and the retry limit $M = 1$ or 10. Note that throughout the paper, the widely-adopted binary exponential backoff [39], i.e., $\mathcal{Q}(i) = 2^{-i}$, is used in the analysis and simulation. We can see from Fig. 3(a) that as the information bits per packet k decreases or the blocklength N increases, the packet error probability ϵ decreases. Particularly, ϵ decreases sharply within a small range of N. For instance, with $k = 100$, the packet error probability ϵ decreases almost from 1 to 0 when N increases from 20 to 40. It can be shown that the point $\frac{k}{\log_2(1+\rho)}$ is included in this small range of N where ϵ changes rapidly. In this case, the point $k/\log_2(1+\rho) \approx 29 \in (20, 40)$.

[7] It is based on the approximation of $n - 1 \approx n$ and $(1-x)^n \approx \exp\{-nx\}$ for $0 \le x \le 1$ with a large n. In practice, a large-n scenario is of particular importance with the rising challenge of massive access of M2M communications.



(a)

(b)

FIGURE 3. (a) Packet error probability ϵ versus the blocklength N. $\rho = 10$ dB. (b) The probability of successful transmissions of HOL packets p versus the blocklength N. $\rho = 10$ dB, $q_0 = 0.1$, $n = 20$ and $\mathcal{Q}(i) = 2^{-i}$.

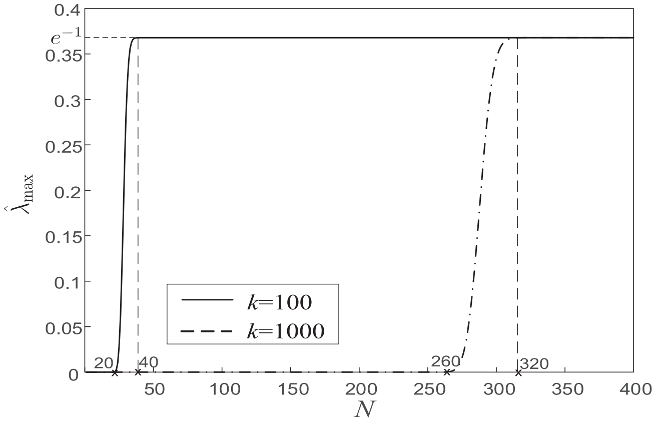


FIGURE 4. Maximum network throughput $\hat{\lambda}_{\max}$ versus the blocklength N. $\rho = 10$ dB.

Moreover, we can see from Fig. 3(b) that the steady-state probability of successful transmissions of HOL packets p increases as the blocklength N increases or the number of information bits k decreases, because of a reduced packet error probability as shown in Fig. 3(a). With a larger retry limit M, on the other hand, nodes can backoff to deeper states, i.e., have a smaller probability of accessing the channel, which relieves the channel contention. Accordingly, the probability of successful transmissions of HOL packets p can be improved.

B. MAXIMUM SUM RATE

According to (1) and (10), the network sum rate is given by $C = \frac{k}{N}\hat{\lambda}_{out}$. Therefore, to derive C, let us first characterize the network throughput $\hat{\lambda}_{out}$. Recall that the saturated condition is considered in this paper, where each node always has packets to transmit. In this case, the node throughput should be equal to the service rate of its data queue, which is the probability of the HOL packet in State T, $\tilde{\pi}_T$, as shown in Fig. 2. The network throughput can thus be derived as

$$\hat{\lambda}_{out} = n\tilde{\pi}_T = -p \ln \frac{p}{1-\epsilon}, \tag{15}$$

by combining (9) and (14). It can be obtained from (15) that $\hat{\lambda}_{out}$ is maximized when the successful transmission probability $p_\lambda = e^{-1}(1-\epsilon)$. Accordingly, by submitting p_λ into (14) and (15), we can obtain the maximum network throughput as

$$\hat{\lambda}_{\max} = \max_{\{q_0\}} \hat{\lambda}_{out} = \frac{1-\epsilon}{e}, \tag{16}$$

and the corresponding optimal initial packet transmission probability

$$q_0 = q_\lambda = \frac{e^{-1}(1-\epsilon)\sum_{i=0}^{M-1}\frac{\left(1-e^{-1}(1-\epsilon)\right)^i}{\mathcal{Q}(i)}}{n\left[1-\left(1-e^{-1}(1-\epsilon)\right)^M\right]}. \tag{17}$$

It is indicated in (16) that different from p, $\hat{\lambda}_{\max}$ does not depend on the retry limit M.

Fig. 4 demonstrates how the maximum network throughput $\hat{\lambda}_{\max}$ varies with the blocklength N with the information bits per packet $k = 100$ or 1000. It can be observed that $\hat{\lambda}_{\max}$ increases as N increases or k decreases. As N becomes large, $\hat{\lambda}_{\max}$ approaches e^{-1} since the packet error probability ϵ decreases to zero in this case, as Fig. 3(a) illustrates.

The maximum sum rate $C_{\max}$ can be written as

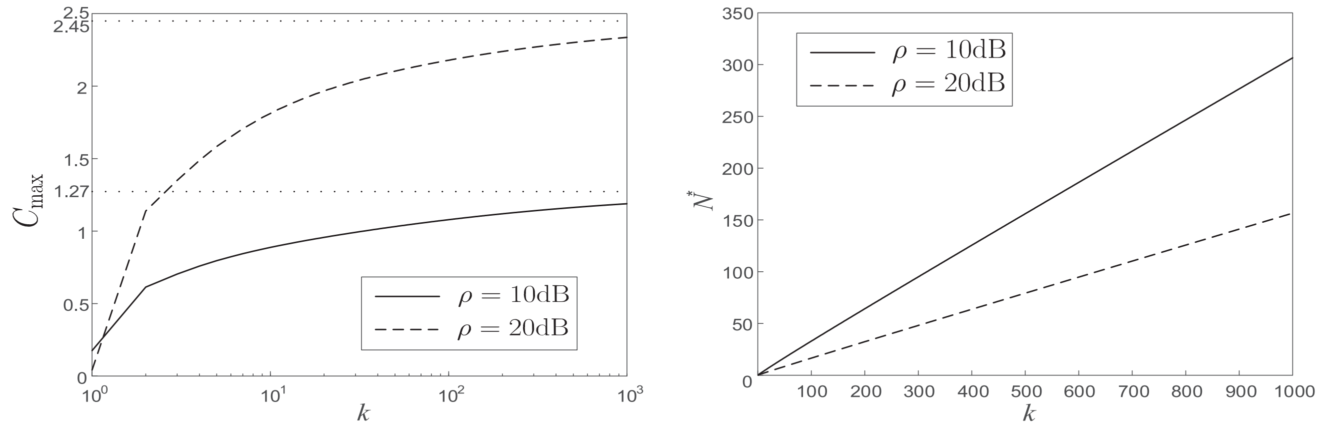
$$C_{\max} = \max_{\{N,q_0\}} \frac{k}{N} \cdot \hat{\lambda}_{out}, \tag{18}$$

according to (10).

Given the number of information bits per packet k and the mean received SNR ρ, the network throughput $\hat{\lambda}_{out}$ can be maximized by tuning the initial packet transmission probability q_0, and the corresponding maximum network throughput $\hat{\lambda}_{\max}$ is solely determined by the blocklength N. Therefore, we have

$$C_{\max} = \max_{N} \frac{k}{N} \cdot \max_{\{q_0\}} \hat{\lambda}_{out} = \max_{N} \frac{k}{N}\hat{\lambda}_{\max}, \tag{19}$$

by combining (16) and (18). Note that the above analysis has revealed that the network throughput performance can be improved with a larger blocklength N because the packet



(a) (b)

FIGURE 5. **(a) Maximum sum rate $C_{\max}$ versus the number of the information bits per packet k. (b) Optimal blocklength N^* versus the number of the information bits per packet k. $M = 1$.**

error probability ϵ is reduced. However, a larger blocklength N would deteriorate the information encoding rate R according to (1). Thus, one should strike a balance between these two in order to optimize the network sum rate.

The following theorem gives the maximum sum rate $C_{\max} = \max_{\{N,q_0\}} C$ and the corresponding optimal parameter settings.

Theorem 1: The maximum sum rate $C_{\max}$ is given by

$$C_{\max} = \frac{k}{N^*} \frac{1 - Q\left(\frac{N^* \log_2(1+\rho) - k + (\log_2 N^*)/2}{\sqrt{N^* V}}\right)}{e}, \tag{20}$$

which is achieved when the blocklength N is set to be $N = N^*$ and the initial packet transmission probability q_0 is set to be $q_0 = q_C$, where $N^* = \arg\max_{\{N \in \mathcal{N}\}} \frac{k}{N} \cdot \frac{1-\epsilon}{e}$ and $\mathcal{N}$ is the set of non-zero roots of

$$\begin{aligned} 1 = {} & Q\left(\frac{N \log_2(1+\rho) - k + (\log_2 N)/2}{\sqrt{NV}}\right) \\ & + \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\left(N \log_2(1+\rho) - k + (\log_2 N)/2\right)^2}{2NV}\right\} \\ & \cdot \frac{N \log_2(1+\rho) + k + \frac{1}{\ln 2} - (\log_2 N)/2}{2\sqrt{NV}}, \end{aligned} \tag{21}$$

and q_C is given by

$$\begin{aligned} q_C = {} & \left(1 - Q\left(\frac{N^* \log_2(1+\rho) - k + (\log_2 N^*)/2}{\sqrt{N^* V}}\right)\right) / \\ & \left(e \cdot n \left(1 - \left(1 - \left(1 - \right.\right.\right.\right. \\ & \left.\left.\left.\left. Q\left(\frac{N^* \log_2(1+\rho) - k + (\log_2 N^*)/2}{\sqrt{N^* V}}\right)\right)/e\right)^M\right)\right) \\ & \cdot \sum_{i=0}^{M-1} \frac{\left(1 - \left(1 - Q\left(\frac{N^* \log_2(1+\rho) - k + (\log_2 N^*)/2}{\sqrt{N^* V}}\right)\right)/e\right)^i}{\mathcal{Q}(i)}. \end{aligned} \tag{22}$$

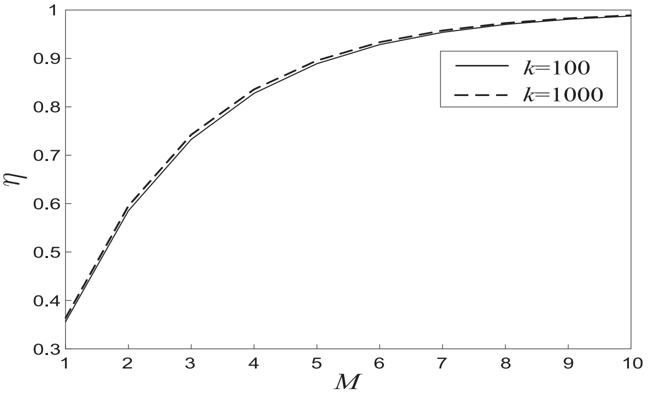
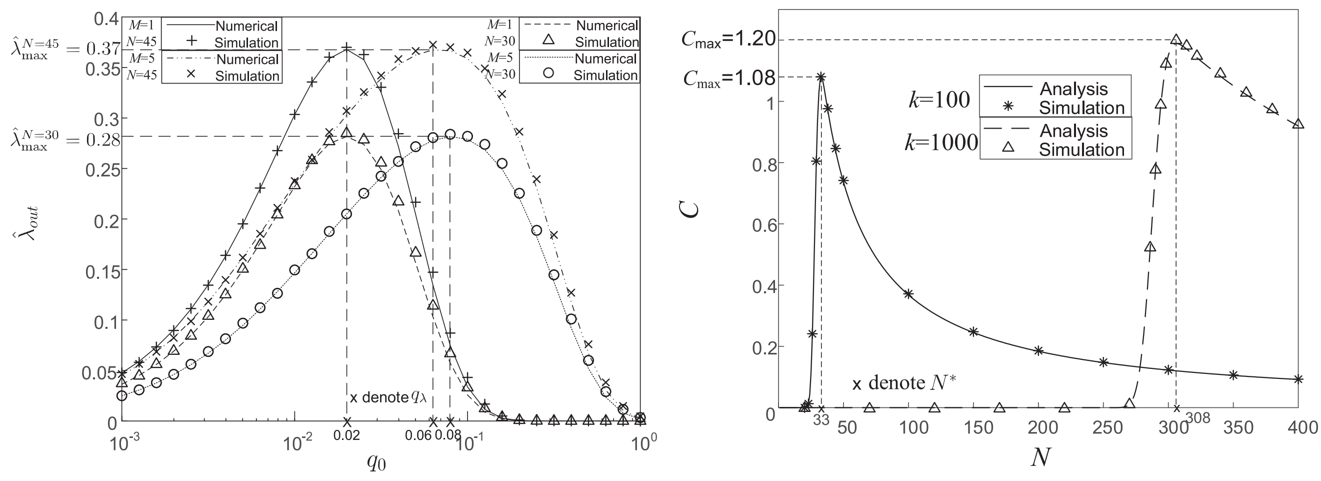


FIGURE 6. **The reliability η versus the retry limit M when the network sum rate C is maximized, $\rho = 10$ dB, $n = 20$, $N = N^*$ and $q_0 = q_C$.**

Proof: See Appendix A. ■

Theorem 1 shows that both maximum sum rate $C_{\max}$ and the corresponding optimal blocklength N^* depend on the number of information bits per packet k and the signal-to-noise ratio ρ. The maximum sum rate $C_{\max}$ is not sensitive to the retry limit M and the number of nodes n, since the initial packet transmission probability is tuned to be q_C, which offsets the effects of M and n. Fig. 5 illustrates how $C_{\max}$ and N^* vary with k with $\rho = 10$ dB or 20 dB. We can see from Fig. 5 that both the maximum sum rate $C_{\max}$ and optimal blocklength N^* increase as k grows. With a small k, a loss in the network sum rate would incur in the short blocklength region. As the number of information bits per packet $k \to \infty$, we have $\lim_{k \to \infty} C_{\max} = e^{-1} \log_2(1+\rho)$ according to (20). In such case, if the optimal blocklength N^* also grows according to (21), then the information encoding rate R approaches the channel capacity $\log_2(1+\rho)$.

To take a closer look at the reliability performance of the network, Fig. 6 illustrates how the reliability η varies with the retry limit M with $k = 100$ or 1000 when the maximum sum rate $C_{\max}$ is achieved. It is in sharp contrast to



(a)

(b)

FIGURE 7. (a) The network throughput $\hat{\lambda}_{out}$ versus the initial probability of accessing the channel q_0, $\rho = 10$ dB, $n = 20$, $k = 100$. (b) The network sum rate C versus the blocklength N. $\rho = 10$ dB, $n = 20$, $M = 1$, $q_0 = q_\lambda$, $\mathcal{Q}(i) = 2^{-i}$.

$C_{\max}$, which is independent of M, the reliability η can be significantly improved with a larger M, as Fig. 6 illustrates. Moreover, the number of information bits per packet k crucially affects $C_{\max}$ while it has limited effect on the reliability η. The above observation suggests that from the perspective of the sum rate performance optimization, a larger retry limit M is preferable in practical system design, because the maximum sum rate $C_{\max}$ is insensitive to M meanwhile the reliability performance can be improved.

C. SIMULATION RESULTS

This section is devoted to presenting event-driven simulation results that verify the proceeding analysis. The simulation setting is consistent with the system model, and therefore details are omitted here for brevity. Each simulation is carried out for 10^8 time slots. The network throughput is obtained by calculating the ratio of the number of successful packets to the number of time slots 10^8. Note that in simulations, one packet is received in error according to the analytical packet error probability calculated by (2)-(3). The reason is two fold: 1) Equation (3) has been verified by [3] and numerous subsequent studies; 2) Equation (2) essentially presents an upper-bound of the information encoding rate. The actual information encoding rate is determined by the encoding scheme, which may not reach the upper-bound. To evaluate the performance limit, such as the maximum sum rate, we then adopt the analytical packet error rate in simulations.

Fig. 7(a) shows how the network throughput $\hat{\lambda}_{out}$ varies with the initial transmission probability q_0 with various values of the blocklength N and the retry limit M. We can see from Fig. 7(a) that $\hat{\lambda}_{out}$ is sensitive to the variation of the packet initial transmission probability q_0, implying that to maximize $\hat{\lambda}_{out}$, q_0 should be carefully tuned. As shown in Fig. 7(a), by optimally tuning q_0 according to (22), the maximum network throughput $\hat{\lambda}_{\max}$ can be achieved. It is interesting to see that with a fixed blocklength N, e.g., $N = 45$, the optimal initial transmission probability q_λ increases from 0.05 to 0.16 as the retry limit M grows from 1 to 5. This is because with the binary exponential backoff, i.e., $\mathcal{Q}(i) = 2^{-i}$, the transmission probability of each packet q_i sharply decreases with the number of retransmissions i. In particular, with a large M, backlogged HOL packets may back off to deeper phases to make the attempt rate arbitrarily small. In this case, to boost the throughput performance, a larger initial transmission probability should be chosen. Moreover, it is also demonstrated in Fig. 7(a) that the maximum network throughput does not vary with M, while grows as the blocklength N increases.

Fig. 7(b) further presents how the network sum rate C varies with the blocklength N with $\hat{\lambda}_{out} = \hat{\lambda}_{\max}$. Theorem 1 has revealed that even with optimal tuning of the transmission probability, the blocklength N also needs to be properly tuned for maximizing the network sum rate C. The simulation results in Fig. 7(b) demonstrated that if N is too small, then the sum rate C is small as well because of a large packet error probability ϵ. On the other hand, if N is large, although the packet error probability ϵ can be improved, each node's information encoding rate becomes small, resulting in a small network sum rate as well. Only by optimally tuning the blocklength N according to (21) can the maximum sum rate be achieved.

The maximum sum rate and the corresponding optimal settings of key system parameters have been characterized so far, based on which we will further study how to optimize the access delay performance in the finite blocklength region in the following section.

IV. MEAN ACCESS DELAY

In this section, we first characterize the mean access delay of successfully-transmitted packets ED, and then study how to minimize it by optimally tuning the initial packet transmission probability q_0 and the blocklength N, and finally discuss the effect of system parameter settings on *ED* and the reliability η.

A. MINIMUM MEAN ACCESS DELAY

Let Y_i denote the sojourn time of each HOL packet in State R_i, and D_i denote the time interval from the beginning of State R_i until the HOL packet leaves the queue, $i = 0, \ldots, M-1$, in unit of time slots. As the holding time in state T and state F is equal to 1, i.e., one time slot, we have

$$D_i = \begin{cases} Y_i + 1 & \text{with prob. } p \\ Y_i + 1 + D_{i+1} & \text{with prob. } 1-p, \end{cases} \tag{23}$$

$i = 0, \ldots, M-2$, and

$$D_{M-1} = Y_{M-1} + 1, \tag{24}$$

according to Figure 2. We then have

$$\begin{cases} G_{D_i}(z) = pzG_{Y_i}(z) + (1-p)zG_{Y_i}(z)G_{D_{i+1}}(z), \\ i = 0, \ldots, M-2 \\ G_{D_{M-1}}(z) = zG_{Y_{M-1}}(z), \end{cases} \tag{25}$$

where $G_{D_i}(z)$ denotes the probability generating function of D_i.

By taking the derivation of (25) with respective to z, we have

$$\begin{cases} G'_{D_i}(z) = pG_{Y_i}(z) + pzG'_{Y_i}(z) + \\ (1-p)\Big[G_{Y_i}(z)G_{D_{i+1}}(z) + \\ zG'_{Y_i}(z)G_{D_{i+1}}(z) + zG_{Y_i}(z)G'_{D_{i+1}}(z)\Big], \\ i = 0, \ldots, M-2 \\ G'_{D_{M-1}}(z) = G_{Y_{M-1}}(z) + zG'_{Y_{M-1}}(z). \end{cases} \tag{26}$$

Subsequently, we have

$$E[D_0] = G'_{D_0}(1) = \frac{1-(1-p)^M}{p} + \sum_{i=0}^{M-1}(1-p)^i G'_{Y_i}(1), \tag{27}$$

where $G'_{Y_i}(1)$ equals the mean holding time of each HOL packet in State R_i, i.e., $G'_{Y_i}(1) = \tau_{R_i} = \frac{1}{q_i} - 1$ according to (7). $E[D_0]$ can then be obtained as

$$E[D_0] = \sum_{i=0}^{M-1} \frac{(1-p)^i}{q_i}. \tag{28}$$

Let D_d and D_s denote the sojourn time of dropped and successfully-transmitted packets, respectively. We have

$$E[D_0] = (1-p)^M E[D_d] + \Big(1-(1-p)^M\Big)E[D_s]. \tag{29}$$

With the retry limit M, D_d can be obtained as

$$D_d = M + \sum_{i=0}^{M-1} Y_i, \tag{30}$$

with the probability generating function given by

$$G_{D_d}(z) = z^M + \prod_{i=0}^{M-1} G_{Y_i}(z). \tag{31}$$

We then have

$$E[D_d] = G'_{D_d}(1) = M + \sum_{i=0}^{M-1} G'_{Y_i}(1) = \sum_{i=0}^{M-1} \frac{1}{q_i}. \tag{32}$$

Finally, by combining (28), (29) and (32), $E[D_s]$ in unit of time slots can be written as

$$E[D_s] = \frac{\sum_{i=0}^{M-1} \frac{(1-p)^i}{q_i} - \sum_{i=0}^{M-1} \frac{(1-p)^M}{q_i}}{1-(1-p)^M}. \tag{33}$$

The mean access delay *ED* then equals the product of $E[D_s]$ and the blocklength N, i.e.,

$$ED = N \cdot E[D_s]. \tag{34}$$

According to (33) and (34), the mean access delay *ED* is determined by the initial packet transmission probability q_0 and the blocklength N. To minimize *ED*, we have the following optimization problem

$$ED_{\min} = \min_{\{N, q_0\}} \frac{N\Big\{\sum_{i=0}^{M-1} \frac{(1-p)^i}{q_i} - \sum_{i=0}^{M-1} \frac{(1-p)^M}{q_i}\Big\}}{1-(1-p)^M}. \tag{35}$$

Due to the implicit nature of the expression in (35), the minimum mean access delay $ED_{\min}$ is hard, if impossible, to be explicitly characterized when $1 < M < \infty$. In Section IV-B, numerical evaluation of $ED_{\min}$ for $1 < M < \infty$ will be presented. Before that, interesting but insightful views can be obtained by considering the cases of $M = 1$ and $M \to \infty$. Specifically, with $M = 1$, one packet is dropped on the first transmission failure. In this case, we have

$$ED|_{M=1} = \frac{N}{q_0}, \tag{36}$$

indicating that the mean access delay decreases as the blocklength N decreases or the initial packet transmission probability q_0 increases.

Moreover, by combining (9), (14), (10), (33) and (34), the mean access delay *ED* can be further written as

$$ED = \frac{nk}{C} - N\frac{(1-p)^M}{1-(1-p)^M}\sum_{i=0}^{M-1}\frac{1}{q_i}. \tag{37}$$

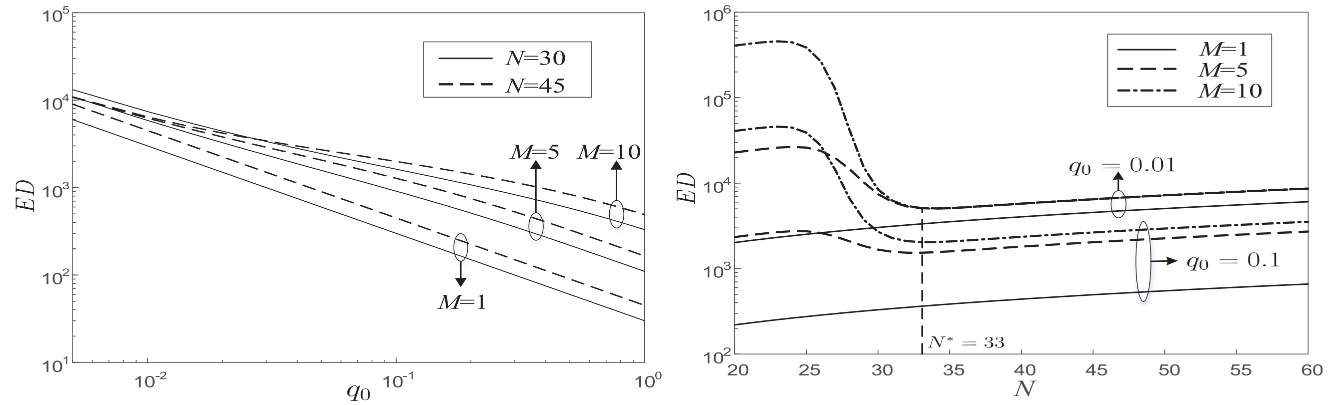
As $M \to \infty$, we have

$$ED|_{M\to\infty} = \frac{nk}{C}, \tag{38}$$

indicating that *ED* is inversely proportional to the network sum rate C. Therefore, *ED* can be minimized when C is maximized, which is achieved when the blocklength N and the initial packet transmission probability q_0 are tuned according to (21) and (22), respectively. Accordingly, the minimum mean access delay in this case is given by

$$ED_{\min}|_{M\to\infty} = \frac{e \cdot nN^*}{1 - Q\Big(\frac{N^* \log_2(1+\rho) - k + (\log_2 N^*)/2}{\sqrt{N^* V}}\Big)}, \tag{39}$$

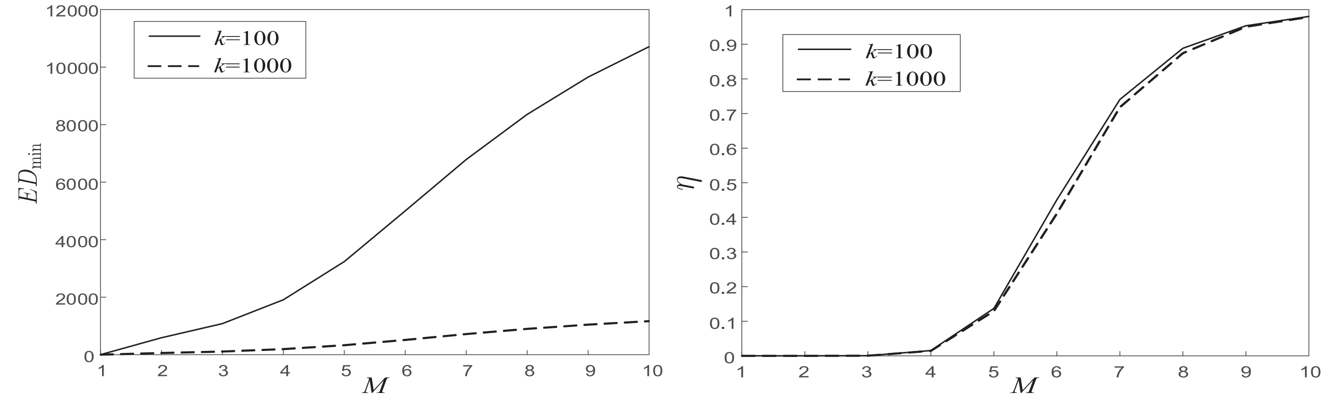
by combining (20) and (38), where N^* is given in (21). According to (39), $ED_{\min}|_{M\to\infty}$ linearly increases with the number of nodes n.



(a)

(b)

FIGURE 8. (a) Mean access delay of successfully-transmitted packets *ED* versus the initial probability of accessing the channel q_0, $k = 100$, $\rho = 10$ dB, $n = 20$ and $\mathcal{Q}(i) = 2^{-i}$. (b) *ED* versus the blocklength *N*, $k = 100$, $\rho = 10$ dB, $n = 20$ and $\mathcal{Q}(i) = 2^{-i}$.



(a)

(b)

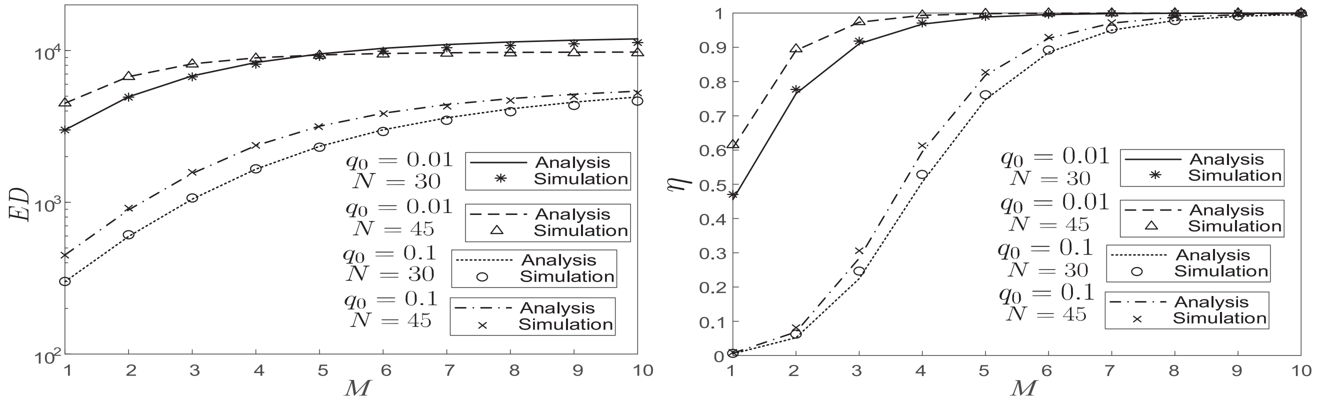
FIGURE 9. (a) Minimum mean access delay of successfully-transmitted packets $ED_{\min}$ versus retry limit *M*. (b) Reliability η versus retry limit *M*, $\rho = 10$ dB, $n = 20$, $\{N,q_0\} = \arg_{\{N,q_0\}} \min ED$ and $\mathcal{Q}(i) = 2^{-i}$.

B. DISCUSSION

The mean access delay of successfully-transmitted packets ED has been obtained in (34) as a function of the blocklength N, the initial packet transmission probability q_0 and the retry limit M. Fig. 8(a) and Fig. 8(b) illustrate how the mean access delay of successfully-transmitted packets ED varies with the initial packet transmission probability q_0 and the blocklength N, respectively. It can be seen from Fig. 8(a) that for a given blocklength N, ED decreases as the retry limit M decreases. Recall that with the binary exponential backoff, i.e., $\mathcal{Q}(i) = 2^{-i}$, the transmission probability of each packet q_i sharply decreases with the number of retransmissions i. Therefore, when the retry limit M grows, a rather small q_i may hold the packet in backoff stage for a long time, which deteriorates the delay performance. Accordingly, a smaller retry limit M or a larger initial transmission probability q_0 can reduce the mean access delay of successfully-transmitted packets ED.

For the effect of the blocklength N on ED, we can see from both Fig. 8(a) and Fig. 8(b) that with the retry limit $M = 1$, the mean access delay of successfully-transmitted packets ED is improved with a smaller blocklength N. This, nevertheless, would deteriorate the network sum rate performance as can be seen from Fig. 7, indicating a severe tradeoff between the network sum rate and the mean access delay performance, especially when the retry limit M is small. On the other hand, as $M \to \infty$, ED becomes inversely proportional to the network sum rate C according to (37). Therefore, ED is minimized when N is tuned to be $N = N^*$, which is given by (21) and verified by the simulation results in Fig. 8(b). Here we can see that there is no tradeoff between the network sum rate and the mean access delay performance when $M \to \infty$. Moreover, in contrast to the optimal blocklength N^* for the the network sum rate maximization, the optimal blocklength to minimize the mean access delay critically depends on the retry limit M.

Fig. 9a illustrates the minimum mean access delay $ED_{\min}$ with $k = 100$ or 1000. Note that $ED_{\min}$ is obtained according to (35) via a brute-force search in a wide range of values of N and q_0. It can be observed from Fig. 9a that $ED_{\min}$ deteriorates as M increases. This is in contrast to the maximum sum rate, which is insensitive to M. To achieve better delay



(a)

(b)

FIGURE 10. (a) Mean access delay of successfully-transmitted packets *ED* versus retry limit *M*, $k = 100$, $\rho = 10$ dB, $n = 20$ and $\mathcal{Q}(i) = 2^{-i}$. (b) Reliability η versus *M*, $k = 100$, $\rho = 10$ dB, $n = 20$ and $\mathcal{Q}(i) = 2^{-i}$.

performance, a small M is usually chosen, with which the reliability performance is however significantly degraded, as shown in Fig. 9b. This reveals a clear tradeoff between the delay and the reliability performance. Here we can see that a small M is not suitable for applications that emphasize on both the reliability and delay performance.

A closer look at Fig. 9 further indicates that by adopting a small number of information bits per packet k, the delay performance can be significantly improved, especially when the retry limit M is large. This suggests that when k is small, one can choose a large M to achieve a much better reliability performance, and in the meanwhile can still have a satisfactory delay performance. The negative effect of having a small k is that the sum rate performance is impaired, as Fig. 5(a) illustrates.

C. SIMULATION RESULTS

Simulation results of the mean access delay ED and the reliability η are presented in Fig. 10 under various values of the blocklength N and the initial packet transmission probability q_0. 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. The reliability is obtained by calculating the ratio of the number of successful packets to the sum of the number of successful packets and the number of dropped packets. It can be seen from Fig. 10a that the mean access delay of successfully-transmitted packets ED increases as M grows, yet the increment becomes marginalized with a large M. Although the access delay performance deteriorates with a large M, the reliability performance is improved, as shown in Fig. 10b. The improvement is significant especially when M is small. When the retry limit M is small, the mean access delay ED increases as the blocklength N increases. For a given initial probability of accessing the channel, ED can be improved when N decreases from $N = 45$ to $N = 30$, as Fig. 10a illustrates. As M increases, ED becomes inversely proportional to the network sum rate C.

V. INSIGHTS FOR M2M COMMUNICATIONS

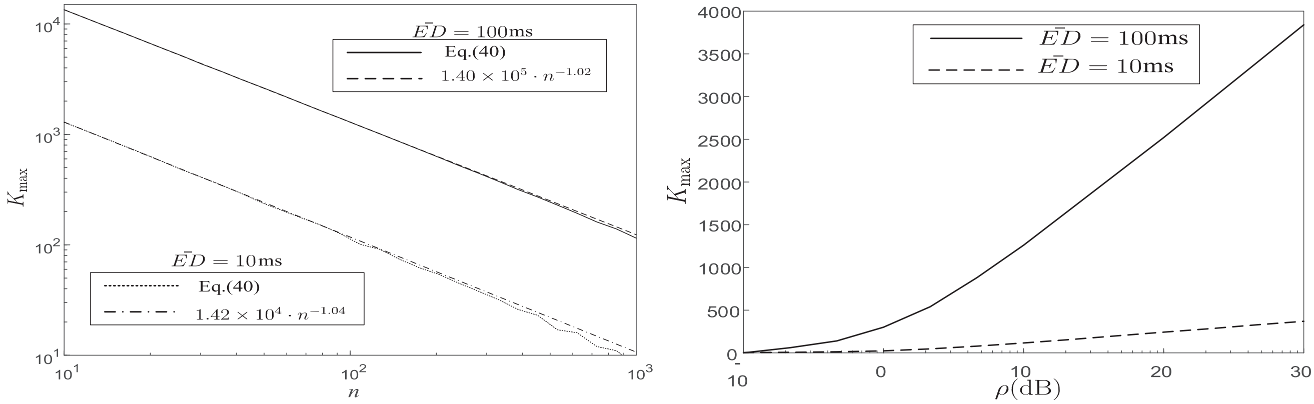
The analysis reveals how to support low-latency service for M2M communications. In the following, we will take the example of LTE-M [40], which was developed by 3GPP to meet the ever-growing need for wide area coverage.

For enabling small packet transmission and reducing the overhead of establishing the connection, 3GPP further introduces the Early Data Transmission (EDT) scheme in LTE-M, where each device send one small packet within the random access procedure [41]. In the following, the analysis will be applied to a single-cell LTE-M system by assuming packet transmissions are performed via the EDT scheme.

The network can adjust the maximum transport block size, i.e., the blocklength, in the random access procedure with EDT and also the packet transmission probability of each device.[8] Recall that to ensure high reliability, the retry limit M should be set to be a large value. Here we let M go to infinity so that the reliability requirement can be guaranteed for simplicity. As $M \to \infty$, it has been shown in (38) that the mean access delay of successfully-transmitted packets ED is inversely proportional to the network sum rate C, to achieve which the blocklength N and the initial probability of accessing the channel q_0 should be tuned according to (21) and (22), respectively. In this case, the minimum mean access delay has been given in (39).

Note that the transmission bandwidth of LTE-M is given by $B = 1.08$ MHz [40]. Therefore, the minimum mean access delay in unit of seconds can be written as $ED_{\min} = \frac{e \cdot nN^*/B}{1-Q(\frac{N^*\log_2(1+\rho)-k+(\log_2 N^*)/2}{\sqrt{N^*V}})}$ according to (39). Note that an excessively large mean access delay could fall short of quality-of-service requirements of latency-critical M2M applications, in which the access delay needs to be bounded under a certain value. For instance, in smart grid, utility operation imposes stringent latency requirements on wireless

[8] In the LTE network, the probability of accessing the channel of each device during the random access procedure is usually referred to as the access class barring factor [42].



(a)

(b)

FIGURE 11. (a) The maximum number of information bits per packet $k_{\max}$ versus the number of nodes n, $\rho = 10$ dB. (b) The maximum number of information bits per packet $k_{\max}$ versus the SNR ρ, $n = 20$.

connectivity because timely actions should be taken to control the grid elements when faults occur. We thus consider a constraint that the minimum mean access delay should not exceed a certain value, i.e., $ED_{\min} \leq \bar{ED}$. Since the minimum mean access delay $ED_{\min}$ monotonically increases as k increases, there exists the maximum allowable number of information bits per packet that the network can support, i.e.,

$$k_{\max} = \max\left\{k | ED_{\min} \leq \bar{ED}\right\}. \tag{40}$$

Given the threshold value $\bar{ED}$, the maximum allowable number of information bits per packet $k_{\max}$ can be easily calculated according to (39), (40) and Theorem 1, which solely depends on the network size n and the received SNR ρ.

Fig.11(a) demonstrates how the maximum allowable number of information bits per packet $k_{\max}$ varies with the number of devices n in the case of the access delay constraint $\bar{ED} = 100$ ms and $\bar{ED} = 10$ ms. It can be seen that $k_{\max}$ polynomially decreases as n increases. In particular, we have $k_{\max} \propto n^{-1.03}$. When n approaches 1000, the maximum allowable number of information bits per packet $k_{\max}$ is close to 100 bits. Note that for improving $k_{\max}$, the network may enlarge the SNR ρ. For illustrating the effect of the SNR ρ, Fig. 11(b) further shows how the maximum allowable number of information bits per packet $k_{\max}$ varies with ρ. It can be observed that $k_{\max}$ steadily increases as ρ increases. When ρ is large, a logarithmic increase can be observed.

VI. CONCLUSION

This paper focus on the sum rate and access delay performance of the short-packet Aloha network with packet dropping. Both the network sum rate and the mean access delay of successfully-transmitted HOL packets are optimized by jointly tuning the transmission probabilities of nodes and the blocklength of packets. The effect of the number of information bits per packet k and the retry limit M on the network optimal performance is characterized. It is found that the retry limit M does not affect the maximum sum rate, and yet strikes a tradeoff between the reliability and delay performances. With a small k, the maximum sum rate deteriorates, while the mean access delay performance can be significantly enhanced, especially when M is large. The analysis is further applied to a single-cell LTE-M system with early data transmission scheme. By taking the constraint of the mean access delay into consideration, the maximum allowable number of information bits per packet is characterized, and shown to be a polynomial decreasing function of the network size, which sheds light on the capability of slotted Aloha to support delay-sensitive service with short-packet transmissions.

Note that we focus on the mean access delay performance in this paper. Based on the proposed analytical framework, higher moments of access delay can be further characterized, which deserves future study. Moreover, we adopt the collision model to simplify the analysis in an AWGN channel. With a fading channel, nevertheless, the collision model can be overly pessimistic, since packets can be decoded with a small error probability even with other concurrent transmissions. In this case, the capture model, i.e., packets would not be successfully received once its received SINR is below certain threshold, can be adopted as a useful simplification.

APPENDIX A
PROOF OF THEOREM 1

According to (16) and (18), we have

$$C_{\max} = \max_{\{N\}} \frac{k}{N} \cdot \frac{1-\epsilon}{e} = \max_{\{N\}} C_{\lambda}, \tag{41}$$

where $C_{\lambda} = \frac{k}{N} \cdot \frac{1-\epsilon}{e}$ denotes the network sum rate when the network throughput is maximized. Recall that the packet error probability is given by

$$\epsilon = Q\left(\frac{N \log_2(1+\rho) - k + \left(\log_2 N\right)/2}{\sqrt{NV}}\right). \tag{42}$$

According to (42), we have

$$\lim_{N\to+\infty} \epsilon = 0,\ \lim_{N\to 1} \epsilon = 1 \text{ and } \tfrac{\partial \epsilon}{\partial N} < 0. \tag{43}$$

To determine the optimal blocklength for maximizing the sum rate, let us first derive

$$\frac{\partial C_\lambda}{\partial N} = \frac{k}{eN^2}\Big(\epsilon - 1 - N\tfrac{\partial \epsilon}{\partial N}\Big), \tag{44}$$

which indicates that the roots of $\frac{\partial C_\lambda}{\partial N} = 0$ are determined by

$$f(N) = \epsilon - 1 - N\tfrac{\partial \epsilon}{\partial N} = 0. \tag{45}$$

Note that for (4), the Gaussian Q function $Q(x) = \frac{1}{2} - \frac{1}{2}\text{erf}(x)$, where $\text{erf}(\cdot)$ is the Gaussian error function and

$$\tfrac{\partial \text{erf}(x)}{\partial x} = \tfrac{2}{\sqrt{\pi}} \exp\Big(x^2\Big). \tag{46}$$

According to (4) and (46), we can have

$$\lim_{N\to+\infty} N\tfrac{\partial \epsilon}{\partial N} = \lim_{N\to+\infty} N\tfrac{V(\log(N)-2(k\log(2)+N\log(\rho+1)+1))}{4\sqrt{2\pi}\log(2)\ \ (NV)^{3/2}} \\ \cdot \exp\left(-\frac{(-k\log(4)+2N\log(\rho+1)+\log(N))^2}{8NV\log^2(2)}\right) = 0. \tag{47}$$

By combining (43), (44) and (47), it can be obtained that

$$\lim_{N\to 1} f(N) > 0 \text{ and } \lim_{N\to+\infty} f(N) < 0. \tag{48}$$

By combining (44)–(45), (48), we can conclude that $\frac{\partial C_\lambda}{\partial N} = 0$ has non-zero roots, i.e., the set $N = \{N|\epsilon - 1 - N\frac{\partial \epsilon}{\partial N} = 0\}$ is non-empty. The optimal blocklength is then given by $N^* = \arg\max_{\{N\in N\}} \frac{k}{N} \cdot \frac{1-\epsilon}{e}$ and the maximum sum rate $C_{\max}$ is thus obtained by combining (41) and $N = N^*$. By combining (17) and (21), (22) can then be derived.

REFERENCES

[1] W. Liu, X. Sun, W. Zhan, and X. Wang, "Maximum sum rate of slotted Aloha for mMTC with short packet," in *Proc. ICCC*, Aug. 2020, pp. 1068–1073.

[2] Z. Jiang, S. Fu, S. Zhou, Z. Niu, S. Zhang, and S. Xu, "AI-Assisted low information latency wireless networking," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 108–115, Feb. 2020.

[3] 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.

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

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

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

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

[8] 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.

[9] 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.

[10] B. J. Kwak, N.-O. Song, and L. E. Miller, "Performance analysis of exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 343–355, Apr. 2005.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted ALOHA with multipacket reception," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2636–2656, Jun. 2005.

[16] M. Sidi and A. Segall, "Two interfering queues in packet-radio networks," *IEEE Trans. Commun.*, vol. 31, no. 1, pp. 123–129, Jan. 1983.

[17] 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.

[18] S. C. Liew, Y. J. 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.

[19] S. B. Rasool and A. U. H. 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.

[20] T. Saadawi and A. Ephremides, "Analysis, stability, and optimization of slotted aloha with a finite number of buffered users," in *Proc. 19th IEEE CDCSAP*, Dec. 1980, pp. 628–633.

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

[22] Y. Li, W. Zhan, and L. Dai, "Rate-constrained delay optimization for slotted ALOHA," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5283–5298, Aug. 2021.

[23] W. Zhan, X. Sun, X. Wang, Y. Fu, and Y. Li, "Performance optimization for massive random access of mMTC in cellular networks with preamble retransmission limit," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 8854–8867, Sep. 2021.

[24] 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.

[25] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the AWGN channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2430–2438, May 2015.

[26] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[27] S. Xu, T.-H. Chang, S.-C. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[28] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475–2485, Nov. 2018.

[29] R. Devassy, G. Durisi, P. Popovski, and E. G. Ström, "Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel," in *Proc. 6th ISCCSP*, May 2014, pp. 173–177.

[30] Y.-W. Huang and P. Moulin, "Finite blocklength coding for multiple access channels," in *Proc. IEEE ISIT*, Jul. 2012, pp. 831–835.

[31] L. Zhao, X. Chi, and Y. Zhu, "Martingales-based energy-efficient D-ALOHA algorithms for MTC networks with delay-insensitive/URLLC terminals co-existence," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1285–1298, Apr. 2018.

[32] J. Chen, L. Zhang, Y.-C. Liang, X. Kang, and R. Zhang, "Resource allocation for wireless-powered IoT networks with short packet communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1447–1461, Feb. 2019.

[33] C. K. Kourtellaris, C. Psomas, and I. Krikidis, "Stability and throughput analysis of multiple access networks with finite blocklength constraints," 2018. *arXiv:1808.01986v1*.

[34] 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.

[35] 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.

[36] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182–185, Apr. 2018.

[37] C. Boyd, R. Kotaba, O. Tirkkonen, and P. Popovski, "Non-orthogonal contention-based access for URLLC devices with frequency diversity," in *Proc. IEEE SPAWC*, Jul. 2019, pp. 1–5.

[38] X. Sun and L. Dai, "Performance optimization of CSMA networks with a finite retry limit," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5947–5962, Sep. 2016.

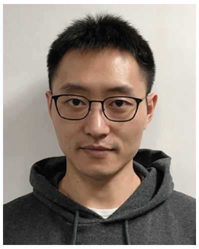
[39] X. Sun and L. Dai, "Backoff design for IEEE 802.11 DCF networks: Fundamental tradeoff and design criterion," *IEEE/ACM Trans. Netw.*, vol. 23, no. 1, pp. 300–316, Feb. 2015.

[40] O. Liberg, M. Sundberg, Y. P. E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things Technologies, Standards and Performance*. Amsterdam, Netherlands: Elsevier, 2019.

[41] A. Höglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, and E. A. Yavuz, "3GPP release 15 early data transmission," *IEEE Commun. Stand. Mag.*, vol. 2, no. 2, pp. 90–96, Jun. 2018.

[42] "Evolved universal terrestrial radio access (E-UTRA); Radio resource control (RRC); protocol specification," 3GPP, Sophia Antipolis, France, 3GPP Rep. TS 36.331 V13.12.0 R13, Jan. 2019.

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



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 and a Postdoctoral Fellow with the City University of Hong Kong. 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 Internet of Things, modeling, and performance optimization of next-generation mobile communication systems.

WEIHUA LIU received the bachelor's degree from Shenzhen University, Shenzhen, China, in 2018, and the master's degree from Sun Yat-sen University, China, in 2021. His research interests include performance analysis and optimization of massive machine-type communications.

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.

QI LIU (Member, IEEE) received the bachelor's and master's degrees from Harbin Engineering University, Harbin, China, in 2013 and 2016, respectively, and the Ph.D. degree in electrical engineering from the City University of Hong Kong, Hong Kong, China, in 2019.

He is currently a Professor with the School of Future Technology with the South China University of Technology. From 2018 to 2019, he was a Visiting Scholar with the University of California at Davis, Davis, CA, USA. From 2019 to 2022, he worked as a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include machine learning, optimization methods, and neuromorphic computing with applications to image/video/speech signal processing. He was a recipient of the Best Paper Award of IEEE International Conference on Signal, Information and Data Processing in 2019. He has been an Associate Editor of the IEEE SYSTEMS JOURNAL since 2022, and *Digital Signal Processing* since 2022. He was also a Guest Editor for the *IET Signal Processing*, *International Journal of Antennas and Propagation*, and *Wireless Communications and Mobile Computing*.