IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 25, 2019, accepted February 15, 2020, date of publication February 21, 2020, date of current version March 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2975545

Martingales-Based ALOHA-Type Grant-Free Access Algorithms for Multi-Channel Networks With mMTC/URLLC Terminals Co-Existence

RUIZHE QI, XUEFEN CHI, LINLIN ZHAO, AND WANTING YANG
College of Communication Engineering, Jilin University, Changchun 130012, China

Corresponding author: Linlin Zhao (zhaoll13@mails.jlu.edu.cn)

This work was supported in part by the Natural Science Foundation of the Department of Science and Technology of Jilin Province under Grant 20180101040JC, and in part by the National Natural Science Foundation of China under Grant 61801191.

ABSTRACT As a simple single-phase transmission strategy, grant-free access is believed to be an effective way to guarantee the stringent quality of service (QoS) requirements for ultra-reliable low-latency communications (URLLCs). However, unless a theory-based fine evaluation on dynamic delay, we cannot hope to overcome the natural defects of random access and so effectively utilize the time-frequency resources. In this paper, we propose a novel multi-channel ALOHA-type (M-ALOHA) grant-free access algorithm for heterogeneous machine type communication (MTC) networks with URLLC-type terminals and delay-tolerant massive MTC (mMTC)-type terminals co-existence. Firstly, we construct a statistical service model characterizing the transmission rate of each terminal with joint consideration of the features of M-ALOHA access scheme, short packet transmissions and frequency-selective fading channel. Then, taking the great advantages of service-martingales theory in random queuing analysis, we present an ingenious delay analysis and obtain the martingales-based formulation of delay-bound violation probability, where the sporadic feature of MTC traffic arrival is carefully addressed. Finally, the M-ALOHA algorithm is formulated as a system throughput maximization problem subject to martingales-based statistical delay-QoS and the total bandwidth of system. The problem is solved by the proposed bi-objective multi-variable-grey wolf optimizer (BOMV-GWO) algorithm. As a result, we obtain the access probability for each terminal and the optimal parameters for the system design, including the number of sub-channels, the bandwidth for each sub-channel and the packets transmission rate. Simulation results demonstrate that the performance of our M-ALOHA algorithm is favorable.

INDEX TERMS Grant-free access, multi-channel, service-martingales theory, ultra-reliable and low-latency communications (URLLCs).

I. INTRODUCTION

More than 50 billion devices are expected to be connected to the Internet of Thing (IoT) by 2020 [1]. With the development of wireless technologies, machine type communication (MTC) network can support devices with different quality of service (QoS) requirements, such as delay-tolerant massive MTC (mMTC)-type devices and ultra-reliable low-latency communication (URLLC) devices [2]. URLLC requires to ensure the reliability of at least 99.999% within the ambitious latency bound of 1ms for short packets [3], which challenges radio access designs of MTC networks [4].

Scheduled uplink transmission contains multiple phases such as handshaking processes, scheduling request, scheduling response, resource allocation and exclusive access transmissions [5]. The complicated multi-phase transmission results in unnecessary latency and may break the constraint of delay-QoS of URLLC applications. Further, the grant-based access models will lead to poor spectrum efficiency and scalability in case of short packets and sporadic arrivals of URLLC traffic [6]. Therefore, grant-based access may not be the favorite solution for uplink MTC with massive

The associate editor coordinating the review of this manuscript and approving it for publication was Saad Qaisar.

devices. In 2017, uplink grant-free random access transmission schemes were proposed by 3rd Generation Partnership Project (3GPP). This simple single-phase transmission strategy allows skipping the lengthy scheduling request and resource allocation process [5], and so reducing the latency significantly.

A handful of works were devoted to studying grant-free access for URLLC and MTC recently. ALOHA is one of the classical contention access schemes. However, it suffers from packet collisions as the number of active user equipments (UEs) increases [4]. Multi-channel technology is an effective method to mitigate collisions. As a result, the reliability of communication is improved and the transmission latency is reduced [7]. In [8], Olga Galinina *et al.* proposed a multi-channel random access control algorithm to achieve low access delay by sending multiple message replicas over multiple channels. They analyzed the number of backlogged MTC devices and the channel access delay. In [9], Jun-Bae Seo *et al.* investigated uplink random access system based on power-domain non-orthogonal multiple access (NOMA) with multiple channels. By exploiting multi-channel selection diversity, they improved the energy efficiency of NOMA random access systems. In [10], Jinho Choi studied the effective capacity and QoS exponent for low-latency communications through multiple fading channels. He obtained an upper-bound on the error probability and a lower-bound on the QoS exponent for the case that the transmitter had statistical channel state information (CSI). However, none of the above literature concentrates on the impact of the number of multi-channel on the reliability and delay QoS guarantee. Thus, it is questioned if URLLCs were achieved in a spectrum efficient way. Besides, the above works focused on QoS guarantee of homogeneous terminals, they ignored diverse QoS guarantees provision for the MTC networks where delay-tolerant and delay-sensitive terminals coexist. In [11], Zhao *et al.* proposed an effective-capacity-based ALOHA-like distributed random access algorithm for multi-packet reception (MPR)-aided uplink visible light communication (VLC) systems having heterogeneous QoS guarantees. However, due to the limitation of effective-capability theory [12], the delay-bound violation probability obtained by effective-capacity theory might be loose. Thus the bandwidth requirement for each terminal might be overestimated [4]. As a result, it is not suitable for the MTC network with URLLCs.

In [13], Poloczek and Ciucu proposed the concept of service-martingales. It constructs arrival-martingale and service-martingale from the perspective of queuing theory, and then performs the delay analysis which were proved to be effective and accurate. Inspired by [13], Liu *et al.* [14] derived the delay bound based on the service-martingales theory for a heterogeneous vehicular network and constructed an optimal task allocation problem with the purpose to minimize the overall delay violation probability. In [15], Hu *et al.* investigated the multi-hop super-martingale end-to-end backlog and delay bound under the first in first out (FIFO) scheduling policy by utilizing the martingale theory and the framework of stochastic network calculus (SNC). Based on the service-martingales theory, Zhao *et al.* [4] proposed an energy-efficient differentiated ALOHA (D-ALOHA) random access algorithms in MTC networks, where delay-insensitive terminals and URLLC terminals compete for one sharing channel. However, in their works, only the system with saturated arrival are considered, where the arrival and service processes are assumed to be statistically independent. Thus, their methods are not applicable when the arrival of traffic is sporadic.

Uplink URLLC traffic with sporadic feature are considered in this paper. In addition, a large bandwidth may be required for meeting its extremely stringent QoS requirements. These features of URLLCs may lead to poor spectrum efficiency. Thus, it is worth exploring that whether the URLLC-type terminals are suitable to coexist with other delay-tolerant terminals, and then improve the bandwidth resources utilization. In addition, we hope to quantitatively obtain the access probabilities and the number of terminals for these co-existence terminals by an elaborate theoretical analysis. In this paper, we focus on the uplink of a multi-channel MTC network that accommodates both delay-sensitive URLLC-type terminals and delay-tolerant mMTC-type terminals. A martingales-based multi-channel ALOHA-type (M-ALOHA) grant-free access mechanism is proposed. Multiple sub-channels are designed to mitigate the collisions of ALOHA. Considering the sporadic arrivals of MTC traffic, we present an ingenious delay analysis relying on the service-martingales theory. Our M-ALOHA algorithm is formulated as a throughput maximization problem, which is high-dimensional, non-convex and non-concave. To handle this intractable mathematical problem, a bi-objective multi-variable-grey wolf optimizer (BOMV-GWO) algorithm is studied. Besides outputting the access probability for each terminal, the BOMV-GWO algorithm also helps to obtain the key parameters for system design, including the optimal number of sub-channels, the bandwidth for each sub-channel and the packets transmission rate. Our main contributions are summarized as follows.

1) We propose a multi-channel ALOHA-type grant-free access scheme. We construct the service process model for a terminal with jointly considering the M-ALOHA access scheme, the characteristics of short packet transmissions and the features of frequency-selective fading channels. In our distributed M-ALOHA access mechanism, the distinct delay and reliability QoS requirements for URLLC-type and mMTC-type terminals are guaranteed by controlling their access probabilities.
2) Considering the sporadic arrivals of MTC traffic, we divide the queue of a terminal's buffer into an empty queue and a non-empty queue logically. Then, a precise and ingenious delay analysis for each terminal is elaborated. For the non-empty queue, relaying on the service-martingales theory, we analyze the queuing process in martingale domain, and calculate the

martingale parameters of our formulated service process. For the empty queue, we derive the probability of successful transmission with respect to the specified delay threshold. Finally, the delay-bound violation probability of each terminal is derived through the full probability formula.

3) We formulate our M-ALOHA algorithm as a system throughput maximization problem subject to martingales-based statistical delay-QoS and the system total bandwidth. The constrained optimization problem (COP) is handled by the BOMV-GWO algorithm. As a result, we obtain the access probability for each terminal and the optimum value of the key parameters for our system design, such as the number of sub-channels and the packets transmission rate. Simulation results show that by dividing the total system bandwidth into multiple sub-channels, system performance is improved greatly. Specially, for a given total bandwidth, only multi-channel ALOHA can meet the extremely stringent QoS requirements of URLLCs. And compared to the single-channel system, the timeslot length for multi-channel system is enlarged, which contributes to both hardware and software.
4) Besides supporting heterogeneous grant-free access, the proposed M-ALOHA algorithm provides a theoretical framework for service configuration of our heterogeneous MTC network. It guides the design of a number of important parameters including the number of terminals and their access probabilities for both URLLC-type and mMTC-type terminals. It helps to realize the co-existence of URLLC/mMTC-type terminals reasonably, and so improves the system spectrum resources utilization effectively.

The rest of this paper is organized as follows. In Section II, the system model is described. In Section III, the M-ALOHA access scheme is proposed and the martingales-based delay analysis for each terminal is presented. In Section IV, our M-ALOHA algorithm is formulated as a COP and the BOMV-GWO algorithm is employed to solve this problem. Our simulation results are provided in Section V. Finally, the conclusions are offered in Section VI.

II. SYSTEM MODEL

In this paper, we consider the uplink communication scenario of MTC network consisting of an access point (AP) and N terminals. The terminals are divided into two types: the mMTC-type terminals and the URLLC-type terminals. The ISM 2.4 GHz band is used. The system bandwidth is divided into M sub-channels, and the allocated bandwidth for each sub-channel is B Hz. Besides, the time is divided into slots. The system model is shown in Fig. 1.

The terminals will access to AP by an ALOHA-type grant-free access mechanism for multi-channel system, which is called M-ALOHA access scheme. In this M-ALOHA access scheme, both the URLLC-type terminal and the mMTC-type terminal will choose a sub-channel

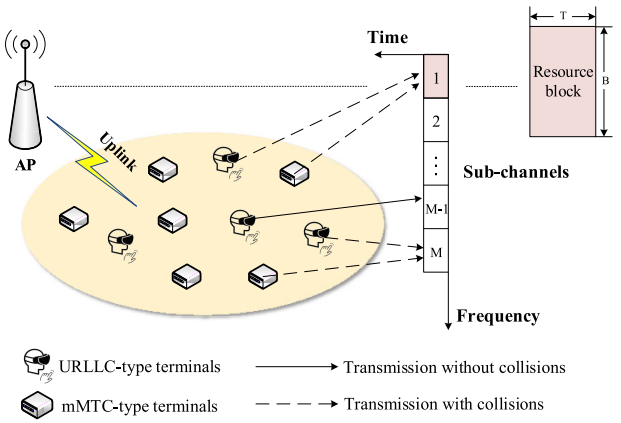

FIGURE 1. The system model.

which possesses the best channel state. Then, the terminal will transmit its data packets over the selected sub-channel with an access probability. The access probability of terminal i is defined as p_i for $i = 1, \cdots, N$. When multiple terminals choose a same sub-channel and transmit simultaneously, their transmissions all fail. Then the information that fails to be transmitted will be retransmitted.

In the slot t, the channel gain of the terminal i can be expressed as

$$h_i(t) = {d_i}^{-l} g_i(t, \tau_n(t)), \tag{1}$$

where ${d_i}^{-l}$ denotes the large-scale channel gain, d_i denotes the distance between the terminal i and the AP. l denotes the path-loss exponent. $g_i(t, \tau_n(t))$ denotes the small-scale channel gain which is mainly effected by multipath fading. $\tau_n(t)$ denotes the delay of path n in the slot t. Considering the influence of frequency selectivity fading and multipath effect, the frequency response experienced by the small-scale fading can be expressed as [16]

$$H_{g_i(t,\tau_n(t))}(f) = \sum_{n=0}^{C(t)} u_n(t) \cdot e^{-j2\pi f \tau_n(t)} \cdot \delta(t - \tau_n(t)), \tag{2}$$

where f denotes the frequency of signal transmission, $C(t)$ denotes the path number of signal propagation in the slot t. $u_n(t)$ denotes the weighting factor of the path n in the slot t, which is a product of transmission and reflection factor. $|u_n(t)| \leq 1$.

The transmission power of the terminal i in the slot t is defined as $\beta_i(t)$. Each terminal is assumed to be aware of its distance to the AP. The transmission power is controlled by resisting large-scale fading. Thus we have

$$\beta_i(t) {d_i}^{-l} = \Theta, \tag{3}$$

and Θ is a constant value.

For the AP, the signal-to-noise (SNR) of the sub-channel selected by terminal i (defined as γ_i) is given by

$$\gamma_i = \frac{\Theta \cdot g_i(t, \tau_n(t))}{\upsilon^2}, \tag{4}$$

where υ^2 denotes the variance of the Gaussian noise.

In MTC networks, short data packets with finite block-length are common form of traffic. $R_i^{\mathrm{b}}(t, m, \rho_i)$ denotes the achievable rate of terminal i for the finite block-length m and the finite block error rate (BLER) ρ_i in the slot t. Different from the Shannon capacity formula which cannot characterize the achieved rate with given block error rate, the achievable rate with finite block-length is expressed as [17]

$$R_i^{\mathrm{b}}(t, m, \rho_i) = B\left\{\log_2(1+\gamma_i) - \sqrt{\frac{\gamma_i(2+\gamma_i)}{(1+\gamma_i)^2 m}(\log_2 e)Q^{-1}(\rho_i)}\right\} \text{[bits/s]}, \tag{5}$$

where $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian Q function.

From the MAC layer, the delay, arrival and service rate are usually described from the perspective of packets. Hence, the unit of achievable rate is transformed from "bits/s" to "packets/slot". Then, the packet transmission rate in the slot t is given by

$$R_i(t, m, \rho_i) = \left\lfloor \frac{R_i^{\mathrm{b}}(t, m, \rho_i) \cdot T}{L} \right\rfloor \text{[packets/slot]}, \tag{6}$$

where T denotes the duration of a slot, and L denotes the size of a data packet. The operation of $\lfloor\cdot\rfloor$ is rounded down to ensure the packet transmission rate is an integer.

For the terminal i, the number of its arrived packets in the slot t is defined as $a_i(t)$, and the number of served packets in the slot t is defined as $s_i(t)$. $A_i(0, t)$ and $S_i(0, t)$ denote the number of accumulated arrived and served packets from the slot 0 to slot t, respectively. Then, we have $A_i(0, t) = \sum_{k=0}^{t} a_i(k)$ and $S_i(0, t) = \sum_{k=0}^{t} s_i(k)$. The corresponding departure process $D_i(t)$ could be expressed as $D_i(0, t) := \min_{0\le l\le t}\{A_i(0, t) + S_i(l, t)\}$ [13]. The delay of the data packets of terminal i in the slot t is represented as $W_i(t)$. Intuitively, it is the horizontal distance between the curves $A_i(t)$ and $D_i(t)$, expressed as [16]

$$W_i(t) := \min\{k \ge 0 | A_i(0, t-k) \le D_i(0, t)\}. \tag{7}$$

In the MTC network, we aim to ensure the QoS guarantee of different type of terminals with different QoS requirements, especially for the URLLC-type terminals with stringent delay and reliability QoS requirements. Consequently, the delay and reliability QoS of terminal i is characterized by [13]

$$\Pr\left\{W_i(t) \ge W_i^{\max}\right\} < \varepsilon_i^{\max}, \tag{8}$$

where $W_i^{\max}$ denotes the delay threshold (i.e., the maximum tolerable delay) for terminal i. $\varepsilon_i^{\max}$ denotes the maximum delay violation probability for terminal i, which gives an expression to the unreliability.

III. MARTINGALES-BASED DELAY ANALYSIS FOR mMTC/URLLC-TYPE TERMINALS

In this section, we construct a statistical model of service process for each terminal for the heterogeneous MTC networks with M-ALOHA access scheme. Then, a precise martingales-based statistical analysis of delay for terminals is carried out.

A. THE SERVICE MODEL OF EACH TERMINAL

In this grant-free network, each terminal may access to AP with an access probability at the beginning of each time slot. The access probability of terminal i is defined as p_i for $i = 1, \cdots, N$. The vector $\boldsymbol{\lambda}^j \in \{0, 1\}^N$ denotes the transmission state j for $j = 1, \cdots, 2^N$. The i-th element in $\boldsymbol{\lambda}^j$ represents the transmission state of the terminal i. Specifically, $\lambda_i^j = 1$ represents that the terminal i transmits data packets to AP; otherwise, $\lambda_i^j = 0$. According to the ALOHA-type grant-free access scheme, we have

$$\Pr\left\{\lambda_i^j = 1\right\} = p_i, \tag{9}$$

$$\Pr\left\{\lambda_i^j = 0\right\} = 1 - p_i. \tag{10}$$

As each terminal's transmission state is statistically independent in the M-ALOHA access mechanism, the probability of the occurrence of state $\boldsymbol{\lambda}^j$, defined as $\Phi(\boldsymbol{\lambda}^j)$, could be expressed as

$$\Phi(\boldsymbol{\lambda}^j) = \prod_{i=1}^{N}\left[(1-\lambda_i^j)(1-p_i) + \lambda_i^j p_i\right]. \tag{11}$$

For analyzing the delay, we first characterize the statistical feature of service process of each terminal. Because of the ALOHA scheme that we employ, the number of data packets transmitted successfully by the terminal i ($s_i(t)$, $t \ge 0$) is independent identical distributed (IID) across slots [4]. Then we have

$$s_i(t) = \begin{cases} R_i(t, m, \rho_i), & p_i^s, \\ 0, & 1 - p_i^s, \end{cases} \tag{12}$$

where p_i^s denotes the probability of successful transmission in a slot for the terminal i. Before accessing to the AP, terminals will choose one sub-channel to transmit data packets. Since the channel is divided into M sub-channels in frequency domain, multiple terminals may transmit data in a single slot simultaneously. We assume that terminals are aware of the CSI of each sub-channel. Therefore, it is certain that which sub-channel should be chosen as the best channel for each terminal. However, because of the randomness and unpredictability of channel states, when multiple terminals transmit information together in state j, collisions happen and their transmissions all fail. To summarize, terminal i transmits to AP successfully only if none of other terminals choose the same sub-channel as terminal i, and none block error occurs during its transmission. The successful transmission probability of terminal i could be expressed as

$$p_i^s = \left[\sum_{\boldsymbol{\lambda}^j\in\{0,1\}^N, g(\boldsymbol{\lambda}^j)\ge 1} \lambda_i^j \Phi(\boldsymbol{\lambda}^j)(1-p_m)^{g(\boldsymbol{\lambda}^j)-1}\right]\cdot(1-\bar{\rho}_i), \tag{13}$$

where $g(\boldsymbol{\lambda}^j)$ denotes the number of terminals that transmit packets simultaneously in state j. p_m denotes the probability of a sub-channel to be chosen as the best channel by each terminal. Since the average SNRs of the sub-channels for a terminal are assumed to be the same, the probability for one sub-channel to be chosen as the best one is the same. So, we have $p_m = 1/M$. Besides, $\bar{\rho}_i$ denotes the average BLER of terminal i in the system that possesses M sub-channels. From (5), ρ_i is

$$\rho_i = Q\left(\frac{\log_2(1+\gamma_i) - R_i^{\mathrm{b}}(t,m,\rho_i)/B}{\sqrt{(1-\frac{1}{(1+\gamma_i)^2})(\log_2 e)/m}}\right). \tag{14}$$

According to [18], Q function can be approximated to be a linear function. Then, we have

$$Q\left(\frac{\log_2(1+\gamma_i) - R_i^{\mathrm{b}}(t,m,\rho_i)/B}{\sqrt{(1-\frac{1}{(1+\gamma_i)^2})(\log_2 e)/m}}\right) \approx \begin{cases} 1, & \gamma_i \le \alpha_i, \\ \frac{1}{2} - w_i\sqrt{m}(\gamma_i - \theta_i), & \alpha_i < \gamma_i < \beta_i, \\ 0, & \gamma_i \ge \beta_i, \end{cases} \tag{15}$$

where the parameter θ_i, w_i, α_i and β_i of terminal i should satisfied: $\theta_i = 2^{R_i^{\mathrm{b}}(t,m,\rho_i)/B} - 1$, $w_i = 1/(2\pi\sqrt{\theta_i})$, $\alpha_i = \theta_i - 1/(2w_i\sqrt{m})$, $\beta_i = \theta_i + 1/(2w_i\sqrt{m})$, and γ_i denotes the instantaneous SNR of the sub-channel selected by terminal i. $\bar{\gamma}_i$ denotes the average SNR of all of the sub-channels for terminal i, then the cumulative distribution function (CDF) of γ_i will be exponential [16]:

$$F(\gamma_i) = \begin{cases} (1 - e^{-\frac{\gamma_i}{\bar{\gamma}_i}})^M, & \gamma_i \ge 0, \\ 0, & otherwise. \end{cases} \tag{16}$$

To derive the average value of BLER for terminal i, we should compute the integral of (15)

$$\begin{aligned} \bar{\rho}_i &= \int_0^{\alpha_i} f(\gamma_i)d\gamma_i + \int_{\alpha_i}^{\beta_i} [\frac{1}{2} - w_i\sqrt{m}(\gamma_i - \theta_i)]f(\gamma_i)d\gamma_i \\ &= F(\alpha_i) - F(0) + \frac{1}{2}[F(\beta_i) - F(\alpha_i)] \\ &\quad - \int_{\alpha_i}^{\beta_i} w_i\sqrt{m}(\gamma_i - \theta_i)f(\gamma_i)d\gamma_i \\ &= w_i\sqrt{m}\int_{\alpha_i}^{\beta_i} (1 - e^{-\frac{\gamma_i}{\bar{\gamma}_i}})^M d\gamma_i \\ &= w_i\sqrt{m}\cdot M\cdot\sum_{j=1}^{M}\frac{C_{M-1}^{M-j}}{j} \\ &\quad \cdot(-1)^j\cdot\left[\beta_i - \alpha_i + \frac{\bar{\gamma}_i}{j}(e^{-\frac{j}{\bar{\gamma}_i}\beta_i} - e^{-\frac{j}{\bar{\gamma}_i}\alpha_i})\right]. \end{aligned} \tag{17}$$

B. MARTINGALES-BASED DELAY ANALYSIS FOR EACH TERMINAL

In this part, for MTC networks with hybrid traffic, we present a delay analysis, which integrates the martingales-based analysis with the canonical probability analysis. Two main parameters are used to describe the reliability and delay QoS requirement, including the delay threshold $W_i^{\max}$ and the delay violation probability $\varepsilon_i^{\max}$, as the formula (8). On one hand, the MTC traffic has sporadic feature. The buffers of terminals may be empty sometimes. On the other hand, due to channel noise or contention among multiple terminals, a data packet may not be transmitted immediately as it arrives. As a result, some data packets may stay in the buffer of terminal and the waiting delay exists. In order to analyze the delay performance in complex situations, we logically divide the queue of a terminal's buffer into two categories: 1) the buffer is empty and 2) the buffer is non-empty when the data packet arrives at the terminal. Similar to [19], for the terminal i, we approximate the probability of the buffer being non-empty as

$$\xi_i = \frac{E[a_i(t)]}{E[s_i(t)]}. \tag{18}$$

$E[\cdot]$ denotes the expectation operator. $E[s_i(t)] = R_i^s \cdot p_i^s$, where

$$R_i^s = \left\lfloor \frac{TB}{L}\left\{\log_2(1+\gamma_i) - \sqrt{\frac{\gamma_i(2+\gamma_i)}{(1+\gamma_i)^2 m}(\log_2 e)}Q^{-1}(\rho_i)\right\}\right\rfloor. \tag{19}$$

The probability of the buffer being empty is given by $1 - \xi_i$.

Next, we start to analysis the delay violation probability of these two queues.

1) THE BUFFER IS EMPTY

If the buffer is empty when a data packet arrives at the terminal, the access delay of this packet is the time that it waits to be transmitted successfully. The maximum number of slots within which the packet must be transmitted is

$$k_i = \left\lfloor \frac{W_i^{\max}}{T} \right\rfloor, \tag{20}$$

and the probability of the packet being transmitted successfully within k_i slots is

$$\begin{aligned} p_{k_i} &= p_i^s + (1-p_i^s)p_i^s + \cdots + (1-p_i^s)^{k_i-1}p_i^s \\ &= \sum_{n=0}^{k_i-1}(1-p_i^s)^n p_i^s. \end{aligned} \tag{21}$$

Then, the delay violation probability in this case is

$$\begin{aligned} \varepsilon_i^1 &= 1 - p_{k_i} \\ &= 1 - \sum_{n=0}^{k_i-1}(1-p_i^s)^n p_i^s. \end{aligned} \tag{22}$$

2) THE BUFFER IS NON-EMPTY

In this situation, the arrival-martingales for arrival flow and the service-martingales for our presented service model of M-ALOHA are all constructed based on the service-martingales theory. Then, the martingales-based delay analysis is studied. Before the analysis, we introduce

some definitions and theorems in the service-martingales theory [13].

Definition 1 (*Super−Martingales*) : If a discrete stochastic sequence $\{X(t), t > 0\}$ satisfy:

$$1)E[X(t)] < \infty;$$
$$2)E[X(t+1)|X(1), X(2), \ldots X(t)] \leq X(t) \quad (23)$$

for all t, then $X(t)$ is said to be a super-martingale.

Based on the definition of super-martingales, two central definitions concerning arrival and service modelling of terminal i have been constructed.

Definition 2 (*Arrival − Martingales*) : The data flow A_i admits arrival-martingales if every $\theta_i > 0$ there is a $K_i^a(\theta_i) \geq 0$ and a function $h_a : rng(a_i(t)) \to R^+$ such that the process

$$M_i^a(t) = h_a(a_i(t))e^{\theta(A_i(t)-tK_i^a(\theta_i))}, \quad t \geq 0, \quad (24)$$

is a super-martingale. In the definition, $rng(\cdot)$ stands for the range operator.

Definition 3 (*Service − Martingales*) : The service process S_i admits service-martingales if every $\theta_i > 0$ there is a $K_i^s(\theta_i) \geq 0$ and a function $h_s : rng(s_i(t)) \to R^+$ such that the process

$$M_i^s(t) = h_s(s_i(t))e^{\theta(S_i(t)-tK_i^s(\theta_i))}, \quad t \geq 0, \quad (25)$$

is a super-martingale.

Definition 4 (*Threshold*) : For $h_a(a_i(t))$ and $h_s(s_i(t))$ as in Definition 2 and 3 define the threshold

$$H_i \coloneqq \min\{h_a(a_i(t))h_s(s_i(t)) : a_i(t) - s_i(t) > 0, t \geq 0\}. \quad (26)$$

Intuitively, H_i is the smallest value of $h_a(a_i(t))h_s(s_i(t))$ such that the instantaneous arrival (i.e., $a_i(t)$) is larger than any value of the stochastic process driving the service process (i.e., $s_i(t)$).

Since the parameters $h_a(a_i(t))$, $K_i^a(\theta_i)$, $h_s(s_i(t))$ and $K_i^s(\theta_i)$ are the functions of θ_i, we could find a special value θ_i^* in the range of θ_i to relate arrival- and service-martingale. The expression of θ_i^* is given by Theorem 1.

Theorem 1 *([4]):* Assume that its arrival process A_i and service process S_i are statistically independent. Further, as the stability condition, assume that

$$\theta_i^* := \sup\{\theta > 0 : K_i^a(\theta_i) \leq K_i^s(\theta_i)\}. \quad (27)$$

Let H_i as in Definition 4, the arrival (service) process admits the arrival- and service-martingale respectively for the terminal i, its delay-bound violation probability is given by

$$\Pr(W_i(t) \geq W_i^{\max}) \leq \frac{E[h_a(a_i(0))]E[h_s(s_i(0))]}{H_i} \cdot e^{-\theta_i^* K_i^s(\theta_i^*)W_i^{\max}}. \quad (28)$$

According to our analysis of the service model, the number of serviced packets transmitted successfully in the slot t, i.e., $s_i(t)$ is IID across slots. Thus $E[e^{-\theta_i s_i(t+1)}] = E[e^{-\theta_i s_i(t)}]$. Let $h_s(s_i(t)) = 1$ for all $s_i(t) \geq 0$, then we have

$$E[h_s(s_i(t+1))e^{\theta_i((t+1)K_i^s(\theta_i)-S_i(t+1))}|s_i(1), \ldots s_i(t)] = e^{\theta_i((t)K_i^s(\theta_i)-S_i(t))}E[e^{-\theta_i s_i(t+1)}]e^{\theta_i K_i^s(\theta_i)}. \quad (29)$$

For making $h_s(s_i(t))e^{\theta_i((t)K_i^s(\theta_i)-S_i(t))}$ as a super-martingale, we have $E[e^{-\theta_i s_i(t+1)}]e^{\theta_i K_i^s(\theta_i)} = 1$. Thus:

$$K_i^s(\theta_i) = -\frac{1}{\theta_i}\ln E[e^{-\theta_i s_i(t)}] = -\frac{1}{\theta_i}\ln\left(1 - p_i^s + p_i^s \cdot e^{-\theta_i R_i^s}\right). \quad (30)$$

For the Bernoulli process and the Markov-modulated as arrival modeling, the parameters of arrival-martingales (i.e., $h_a(a_i(t))$ for $t \geq 0$ and $K_i^a(\theta_i)$ for $\theta_i \geq 0$) have been investigated in [4]. The upper bound of delay violation probability of the case 2) could be expressed as

$$\varepsilon_i^2 = \frac{E[h_a(a_i(0))]}{H_i}e^{-\theta_i^* K_i^s(\theta_i^*)W_i^{\max}}. \quad (31)$$

Based on the analysis above, we could obtain the delay violation probability of terminal i in general condition:

$$\begin{aligned}\varepsilon_i &\leq (1-\xi_i)\cdot\varepsilon_i^1 + \xi_i\cdot\varepsilon_i^2 \\ &\leq (1-\frac{E[a_i(t)]}{E[s_i(t)]})\cdot\left(1-\sum_{n=0}^{k_i-1}(1-p_i^s)^n p_i^s\right) \\ &\quad + \frac{E[a_i(t)]}{E[s_i(t)]}\cdot\frac{E[h_a(a_i(0))]}{H_i}e^{-\theta_i^* K_i^s(\theta_i^*)W_i^{\max}}. \end{aligned} \quad (32)$$

For the MTC traffic with sporadic arrivals, it is often the case that the buffer of terminal is empty, thus the $1 - \xi_i$ accounts for a large proportion. Therefore, a tight delay-bound violation probability is obtained by our martingales-based delay analysis.

IV. PROBLEM FORMULATION AND THE BOMV-GWO ALGORITHM

In this section, we formulate our M-ALOHA algorithm as a system throughput maximization problem with multiple constrains on martingales-based delay-QoS and the total bandwidth of system. Then the BOMV-GWO algorithm is presented to solve the optimization problem. As a result, the parameters for the design of system, including the bandwidth for each sub-channel, the number of sub-channels and the packets transmission rate are obtained for the mMTC/URLLC co-existence network.

A. PROBLEM FORMULATION FOR THE OPTIMAL M-ALOHA ALGORITHM

The system bandwidth is defined as B_{total}. So the number of sub-channels is $M = \lfloor B_{total}/B \rfloor$. The throughput of the network is given by

$$\begin{aligned}\eta(\boldsymbol{p}, B, R_i^s) &= \sum_{i=1}^{N} p_i^s \cdot \frac{R_i^s L}{T} \\ &= \sum_{i=1}^{N}\sum_{\boldsymbol{\lambda}^j\in\{0,1\}^N, g(\boldsymbol{\lambda}^j)\geq 1}[\lambda_i^j \Phi(\boldsymbol{\lambda}^j)(1-p_m)^{g(\boldsymbol{\lambda}^j)-1}] \\ &\quad \cdot(1-\bar{\rho}_i)\cdot\frac{R_i^s L}{T}. \end{aligned} \quad (33)$$

In this paper, we aim to find the access probability vector $\boldsymbol{p} = [p_1, \cdots, p_N]$, the sub-channel bandwidth B and the packet transmission rate R_i^s. Accordingly, the system throughput maximization problem is formulated as

$$\begin{aligned} &\max_{\boldsymbol{p},B,R_i^s} \eta(\boldsymbol{p}, B, R_i^s) \\ &s.t. \Pr(W_i(t) \geq W_i^{\max}) \leq \varepsilon_i^{\max}, \quad 1 \leq i \leq N \\ &\quad 0 < p_i \leq 1, 1 \leq i \leq N \\ &\quad 0 < B \leq B_{total} \\ &\quad R_i^s \geq 1 \end{aligned} \tag{34}$$

Substitute (32) into (34), we have

$$\begin{aligned} &\max_{\boldsymbol{p},B,R_i^s} \eta(\boldsymbol{p}, B, R_i^s) \\ &s.t. (1 - \frac{E[a_i(t)]}{E[s_i(t)]}) \cdot \left(1 - \sum_{n=0}^{k_i-1} (1-p_i^s)^n p_i^s\right) + \frac{E[a_i(t)]}{E[s_i(t)]} \\ &\quad \cdot \frac{E[h_a(a_i(0))]}{H_i} e^{-\theta_i^* K_i^s(\theta_i^*) W_i^{\max}} \leq \varepsilon_i^{\max}, \quad 1 \leq i \leq N \\ &\quad 0 < p_i \leq 1, \quad 1 \leq i \leq N \\ &\quad 0 < B \leq B_{total} \\ &\quad R_i^s \geq 1 \end{aligned} \tag{35}$$

Then, the COP is transformed into an unconstrained bi-objective optimization problem(OP), which is given as

$$\min_{\boldsymbol{p},B,R_i^s} \Pi(\boldsymbol{p}, B, R_i^s) = \left(\frac{1}{\eta(\boldsymbol{p}, B, R_i^s)}, \Omega(\boldsymbol{p}, B, R_i^s)\right), \tag{36}$$

where $\Omega(\boldsymbol{p}, B, R_i^s)$ is the sum of all constraint violations in the OP (35). More specifically, $\Omega(\boldsymbol{p}, B, R_i^s)$ is defined as follows

$$\Omega(\boldsymbol{p}, B, R_i^s) = \sum_i^N \left[\Omega_i^{QoS} + \Omega_i^{p0} + \Omega_i^{p1}\right] + \Omega^{B0} + \Omega^{B1} + \Omega^{R0}, \tag{37}$$

$$\Omega_i^{QoS} = \max\left(0, \frac{\varepsilon_i}{\varepsilon_i^{\max}} - 1\right), \tag{38}$$

$$\Omega_i^{p0} = \max(0, -p_i), \tag{39}$$

$$\Omega_i^{p1} = \max(0, p_i - 1), \tag{40}$$

$$\Omega^{B0} = \max(0, B), \tag{41}$$

$$\Omega^{B1} = \max(0, B - B_{total}), \tag{42}$$

$$\Omega^{R0} = \max(0, 1 - R_i^s). \tag{43}$$

Obviously, we have $\Omega(\boldsymbol{p}, B, R_i^s) \geq 0$, and all the constraints in (35) are satisfied if and only if $\Omega(\boldsymbol{p}, B, R_i^s) = 0$.

B. THE BOMV-GWO ALGORITHM

The Grey Wolf Optimizer (GWO) algorithm [20] was proposed by Mirjalili in 2014, which is a swarm intelligence algorithm with simulation of grey wolves leadership hierarchy and hunting mechanism in nature. Below we start with a brief introduction of the GWO algorithm.

We firstly create initial population of the Q grey wolves that are randomly dispersed over the (N+2)-dimensional search space. The locations of the wolves are indicated as optimum vector candidates, one of which is defined as $\boldsymbol{X} = [\boldsymbol{p}, B, R_i^s]$. The optimization objection $\Pi(\boldsymbol{p}, B, R_i^s)$ indicates the fitness function of the prey. In order to mathematically model the social hierarchy of wolves, the fittest optimum vector solution is defined as the leader wolf σ. And then, the second and the third best solutions are named τ and κ, respectively. They are the wolves with sub-optimum fitness and the candidate of σ simultaneously. The rest of the candidate solutions are assumed to be ϖ. The hunting (optimization) is guided by σ, τ and κ. The responsibility of ϖ is to assist the top three wolves in attacking the prey (i.e., the optimization objection). $Z_{\max}$ is defined as the max number of iterations. After the initialization, each search agent (i.e., ϖ) have to update its distance from the prey to optimize the candidate solutions in the iterating process [21]. The key steps [20] of the GWO algorithm are presented as follows.

Step 1) *Encircling Prey* : After locating the prey, the grey wolves will encircle the prey firstly. At the k-th iteration, the distance between the wolves and the prey could be expressed as

$$\boldsymbol{Y}(k) = \left|\boldsymbol{C}(k) \cdot \boldsymbol{X}_p(k) - \boldsymbol{X}(k)\right|, \tag{44}$$

$$\boldsymbol{X}(k+1) = \boldsymbol{X}_p(k) - \boldsymbol{A}(k) \cdot \boldsymbol{Y}(k), \tag{45}$$

where $\boldsymbol{X}_p(k)$ is the position vector of the prey, $\boldsymbol{X}(k)$ is the position vector of a grey wolf at the k-th iteration. $\boldsymbol{A}(k)$ and $\boldsymbol{C}(k)$ are coefficient vectors, which could be calculated as

$$\boldsymbol{A}(k) = \boldsymbol{q}(k) \cdot (2 \cdot \boldsymbol{r}_1 - 1), \tag{46}$$

$$\boldsymbol{C}(k) = 2 \cdot \boldsymbol{r}_2. \tag{47}$$

where components of $\boldsymbol{q}(k)$ are linearly decreased for 2 to 0 over the course of iterations. The elements in $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ are uniformly distributed random numbers within [0,1].

Step 2) *Hunting* : Hunting is a process of the wolves approach their prey, in other words, the optimum vector approaches the optimal solution gradually. We suppose that the leader wolves σ, τ and κ know the potential location of prey. Therefore, the other wolves ϖ will update their position according to the position of the leader wolves. The approximate distances between the current solution and σ, τ and κ solutions are computed using the formula

$$\boldsymbol{Y}_l(k) = \left|\boldsymbol{C}_l(k) \cdot \boldsymbol{X}_l(k) - \boldsymbol{X}(k)\right|, \quad l \in \{\sigma, \tau, \kappa\}, \tag{48}$$

and using the notations $\boldsymbol{X}^l(k)$ for the updated σ, τ and κ solutions, respectively.

$$\boldsymbol{X}^l(k) = \boldsymbol{X}_l(k) - \boldsymbol{A}_l(k) \cdot \boldsymbol{Y}_l(k), \quad l \in \{\sigma, \tau, \kappa\}, \tag{49}$$

where $\boldsymbol{Y}_\sigma$, $\boldsymbol{Y}_\tau$ and $\boldsymbol{Y}_\kappa$ defined as the distance between the leader wolves σ, τ and κ with the other wolves ϖ, respectively. Then, the formula of updating a wolf's position could be expressed as

$$\boldsymbol{X}(k+1) = \frac{\boldsymbol{X}^\sigma(k) + \boldsymbol{X}^\tau(k) + \boldsymbol{X}^\kappa(k)}{3}. \tag{50}$$

Step 3) *Attacking Prey* : Attacking is the last stage when the prey stops moving. The grey wolves attack and capture the

prey, which represent we have got the optimum solution. This process is mainly realized by the decrease of $\boldsymbol{q}(k)$. When $\boldsymbol{q}(k)$ is decreased from 2 to 0 over the course of iterations, the fluctuation range of $\boldsymbol{A}(k)$ is in the interval $[-2\boldsymbol{q}(k), 2\boldsymbol{q}(k)]$. If random values of $\boldsymbol{A}(k)$ are satisfied $|\boldsymbol{A}(k)| \leq 1$, it suggested that the wolvesnext position will be closer to the prey.

To sum up, the search process starts with creating a random population of grey wolves (candidate solutions). Over the course of iterations, the leader wolves σ, τ and κ estimate the probable position of the prey, and each candidate solution ϖ updated its distance from the prey. After this process, the wolves will approach their prey gradually. Finally, the GWO algorithm is terminated by the satisfaction of an end criterion. We apply the GWO algorithm to our bi-objective multi-variable optimization problem (36), and the whole framework of our BOMV-GWO algorithm is presented in Algorithm 1.

Algorithm 1 The BOMV-GWO Algorithm

1: Initialize the number of iterations: $Z = 0$;
2: Create initial population of Q wolfs $X = \{\boldsymbol{X}_j, j = 1, \ldots, Q\}$ that are randomly dispersed over the (N+2)-dimensional search space;
3: Initialize the position vector of each wolf, $\boldsymbol{X}_j = [\boldsymbol{p}, B, R_i^s]$.
4: **while** Z<$Z_{\max}$ **do**
5: Z=Z+1;
6: **for** each wolf **do**
7: Calculate p_i^s and ε_i of each wolf in X.
8: Calculate the optimized objective function $1/\eta(\boldsymbol{p}, B, R_i^s)$ and constraint function $\Omega(\boldsymbol{p}, B, R_i^s)$ of each wolf in X.
9: **end for**
10: Q^{fit} is expressed as the number of $\Omega(\boldsymbol{X}_j) = 0$.
11: Calculate the weighting factor $\varsigma = Q^{fit}/Q$.
12: **for** each wolf **do**
13: The optimization function is $\Pi(\boldsymbol{X}_j) = \sqrt{\varsigma \cdot (1/\eta(\boldsymbol{X}_j))^2 + (1-\varsigma) \cdot (\Omega(\boldsymbol{X}_j))^2}$.
14: Sort the $\Pi(\boldsymbol{X}_j)$ in ascending order according to their fitness values.
15: id_1 = the best wolf
16: id_2 = the second best wolf
17: id_3 = the third best wolf
18: Calculate $\boldsymbol{q}(k)$, $\boldsymbol{A}(k)$ and $\boldsymbol{C}(k)$;
19: Update the position of the current wolf by equation (50).
20: **end for**
21: **end while**
22: **return** the best fitness and $X(id_1)$ of $Z_{\max}$.

V. SIMULATION RESULTS AND DISCUSSIONS

In this section, we verify the accuracy of our M-ALOHA service model, evaluate our martingales-based delay analysis for the M-ALOHA access scheme, and present the achievable performance of our M-ALOHA algorithm. In the simulations, N_1 URLLC-type terminals and N_2 mMTC-type terminals exist in the MTC network. The total system bandwidth B_{total} is divided into M sub-channels, each of which has a bandwidth of B. Referring to [22], the packet length is set to be 32 bytes. The Bernoulli process is utilized to model the IID traffic arrival, and the Markov-modulated on-off (MMOO) process is utilized to model the arrival with bursty for each terminal. For Bernoulli source, p^a and q^a denote the probabilities of state 1 and state 0, respectively. For MMOO source, p^a and q^a denote the transition probabilities from state 0 to state 1, and from state 1 to state 0. The transition matrix of MMOO process is given by $\begin{bmatrix} 1-p^a & p^a \\ q^a & 1-q^a \end{bmatrix}$. The packets transmission rate R_i^s is set to be 1 packets/slot from Fig. 2 to Fig. 5.

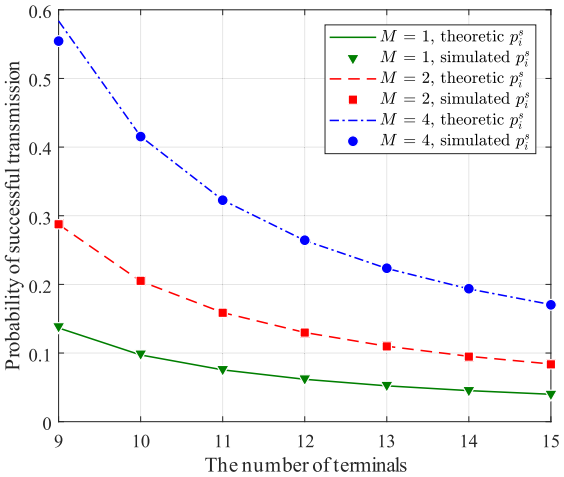

FIGURE 2. Comparison between theoretical and simulated results of successful transmission probability in one slot for each terminal in the context of different number of sub-channels (i.e., M). The number of URLLC-type terminals N_1 is range from 3 to 9. $N_2 = 6$, $B_{total} = 20$ MHz and $B = 20, 10, 5$ MHz, $T = 0.032, 0.064, 0.128$ ms when $M = 1, 2, 4$.

A. ACCURACY ANALYSIS OF THE M-ALOHA SERVICE MODEL

Fig. 2 shows the comparison results between the theoretical and the simulated values in terms of the successful transmission probability (i.e., p_i^s) for each terminal. The access probabilities $p_1 = \min\{1, M/N_1\}$ are for URLLC-type terminals, and $p_2 = 0.01$ are for mMTC-type terminals. The total bandwidth of system B_{total} is set to be 20 MHz. Obviously, the value of B is varies with M. And to ensure R_i^s is constant, the duration of a time slot T is varies with B. The total slots is set to be 10^6. The number of mMTC-type terminals is $N_2 = 6$. The values of N_1 range from 3 to 9. The average SNR is set to be 15 dB. The bandwidth of a sub-channel B is 20 MHz, 10 MHz, 5 MHz, and the duration of a time slot T is set to be 0.032 ms, 0.064 ms, 0.128 ms in the case of M is 1, 2, 4. Obviously, Fig. 2 demonstrates that the results of successful transmission probability obtained by our theoretical analysis match the simulation results well. Further, Fig. 2 verifies

that the probability of successful transmission increases as M increases, and decreases as the number of terminals (i.e., N) increases. It is reasonable because the adoption of multiple sub-channels will distinctly alleviate the collisions caused by simultaneous transmissions of multiple terminals. Notably, the values of p_i^s showed in the Fig. 2 signify the probability of successful transmission in only one time slot, rather than within the specified time slots bounded by delay threshold. Thus, it does not represent the reliability of our M-ALOHA access mechanism for each terminal.

B. DELAY AND RELIABILITY ANALYSIS OF THE M-ALOHA ACCESS SCHEME

In this section, we show the delay-QoS and the reliability results obtained by the martingales-based delay analysis for each terminal. The arrival parameters p^a is set to be 0.9, and q^a is set to be 0.001. The average arrival rate of traffic in a terminal is set to be 0.0825 packets/slot. The access probabilities $p_1 = \min\{1, M/N_1\}$ are for URLLC-type terminals, and $p_2 = 0.01$ are for mMTC-type terminals. The delay values of 10^6 slots for each terminal are recorded for evaluating the simulation results of delay-bound violation probability (i.e., ε_i). The difference between 1 and the value of ε_i is the reliability of our M-ALOHA access scheme.

Fig. 3 shows the performance of delay and reliability for the URLLC-type terminals from different perspectives of system design. In Fig. 3, we assume that there are 3 URLLC-type terminals and 6 mMTC-type terminals in our system. In the simulation scenario of Fig. 3(a), B_{total} is set to be 40 MHz, and it is divided into multiple sub-channels. The value of B is 40 MHz, 20 MHz, 10 MHz, and the duration of a time slot T is 0.0165 ms, 0.033 ms, 0.066 ms in the case of M is 1, 2, 4. From Fig. 3(a), it can be seen that the delay-bound violation probability counted in the simulation matches accurately with the delay-bound violation probability obtained by our martingales-based delay analysis. And we could find that the the reliability (i.e., the value of $1 - \varepsilon_i$) of our M-ALOHA access scheme hits 0.99999 in case that the delay threshold is about 0.95 ms in the situation of $M = 4$, and the delay threshold is about 1.75 ms in the situation of $M = 1$. Obviously, it demonstrates that even if a single channel has the full bandwidth and a small duration of a time slot, it hardly meet the QoS requirements for URLLCs. Therefore, dividing the system bandwidth into multiple sub-channels is an extremely effective method. Besides, the reliability of our grant-free access scheme increases significantly as the delay threshold (i.e., $W_i^{\max}$) increases. It is reasonable because the larger $W_i^{\max}$ indicates more opportunities of retransmissions, and therefore a higher reliability is achieved. In the simulation scenario of Fig. 3(b), the sub-channel bandwidth B is set to be 10 MHz, and T is set to be 0.066 ms. Obviously, when the system only has one channel (as shown in Fig. 3(b), $M = 1$), it is particularly unreliable. Fig. 3(b) demonstrates that more sub-channels with narrow slots are preferred to carry multiple URLLC-type and mMTC-type terminals, if the resource of system bandwidth is abundant. Otherwise, more

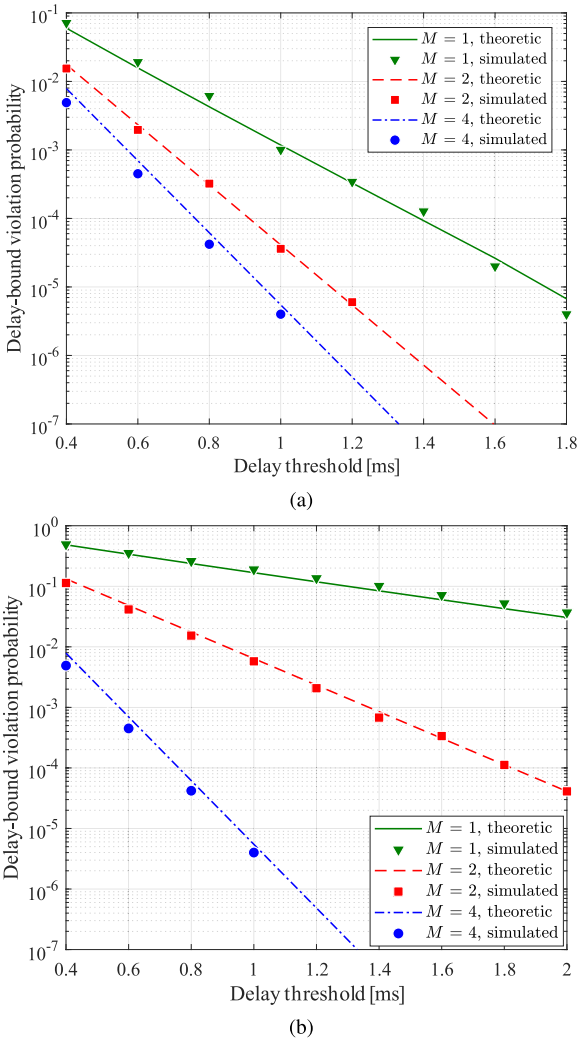

FIGURE 3. The performance of delay and reliability for URLLC-type terminals from two perspectives of system design. $N_1 = 3$, $N_2 = 6$. (a) Constant total bandwidth of system and packets transmission rate, varying bandwidth of each sub-channel, the number of sub-channels and the duration of a time slot. $B_{total} = 40$ MHz, $R_i^s = 1$ packets/slot and $B = 40, 20, 10$ MHz, $T = 0.0165, 0.033, 0.066$ ms when $M = 1, 2, 4$. (b) Constant bandwidth of each sub-channel, packets transmission rate and the duration of a time slot, varying the number of sub-channels and the system total bandwidth. $R_i^s = 1$ packets/slot, $B = 10$ MHz and $T = 0.066$ ms.

elaborate design of system parameters and access probability is crucial to achieve the reliable communications for URLLC type terminals.

Fig. 4 shows the delay-bound violation probability for the URLLC-type terminals in the context of different number of URLLC-type terminals (i.e., N_1) in our system. In the simulation scenario of Fig. 4, the total bandwidth of system B_{total} is 40 MHz, the duration of a time slot is 0.066 ms. The number of sub-channels M is 4, and the number of mMTC-type terminals is 6. It can be seen from Fig. 4 that the performance of reliability deteriorates with the increase of the number of URLLC-type terminals.

Fig. 5 shows the impact of the number of mMTC-type terminals (i.e., N_2) on the delay-bound violation probability of

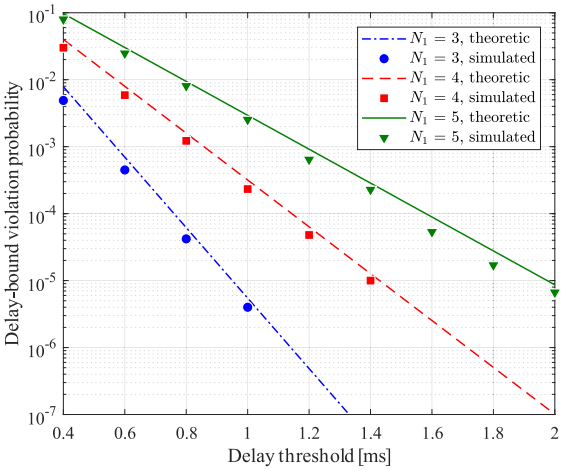

FIGURE 4. The delay-bound violation probability for the URLLC-type terminals in the context of different number of URLLC-type terminals (i.e., N_1). $N_2 = 6$. $R_i^S = 1$ packets/slot, $B_{total} = 40$ MHz, $M = 4$, $B = 10$ MHz and $T = 0.066$ ms.

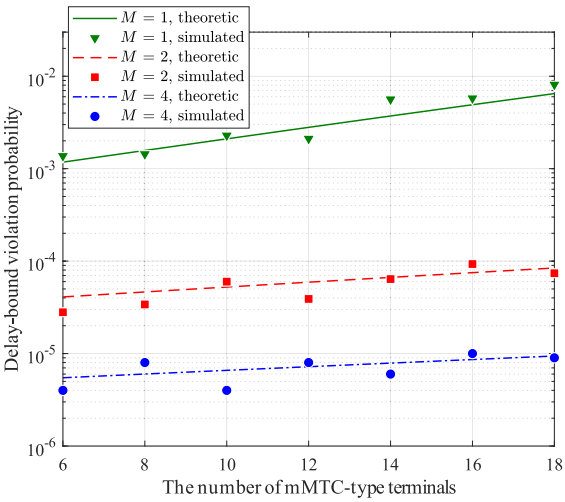

FIGURE 5. The impact of the number of mMTC-type terminals (i.e., N_2) on the delay-bound violation probability of URLLC-type terminals in the context of different number of sub-channels (i.e., M). $B_{total} = 40$ MHz, $R_i^S = 1$ packets/slot and $B = 40, 20, 10$ MHz, $T = 0.0165, 0.033, 0.066$ ms when $M = 1, 2, 4$. The number of URLLC-type terminals is 3, and their delay threshold is 1 ms.

URLLC-type terminals. In the simulation scenario of Fig. 5, the total bandwidth of system B_{total} is 40 MHz, the duration of a time slot is 0.066 ms. The number of URLLC-type terminals (i.e., N_1) is 3, and their delay threshold is set to be 1 ms. The number of mMTC-type terminals is range from 6 to 18. From Fig. 5, it can be seen that as N_2 increases, the reliability QoS of URLLCs decreases (i.e., the delay-bound violation probability increases), but the decline is slow. It demonstrates that the number of mMTC-type terminals has little impact on the QoS performance of URLLCs. In addition, this result also gives us an inspiration on the service configuration, that is, URLLC-type terminals and mMTC-type terminals are suitable for co-existence, which can effectively improve the utilization of bandwidth resources without affecting the QoS performance of critical services. Besides, it is obviously that in the case of $M = 1$, the reliability decreases distinctly as N_2 increases, compared to multiple sub-channels. It indicates that the multi-channel system is better able to accommodate more terminals.

C. PERFORMANCE ANALYSIS OF M-ALOHA ALGORITHM AND COMMUNICATION SYSTEM DESIGN

In this section, we firstly show the convergence performance of the BOMV-GWO algorithm, the access probability for each terminal and the optimum parameters for system design obtained by this BOMV-GWO algorithm. Then, we present the achievable performance of the M-ALOHA algorithm. The average arrival rate is set to be 1 packets/slot. And the average SNR is set to be 15 dB.

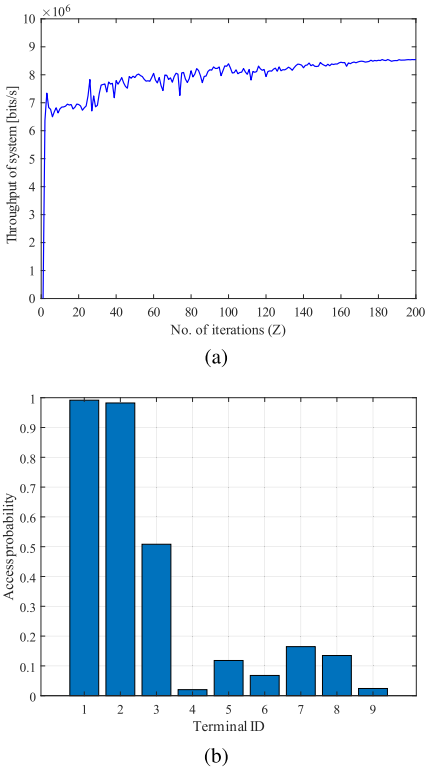

FIGURE 6. (a) Performance of the BOMV-GWO algorithm and (b) Access probability for each terminal obtained by BOMV-GWO algorithm. Terminals 1-3 are URLLC-type terminal and terminals 4-9 are mMTC-type terminal.

Fig. 6(a) depicts the convergence curve of system throughput through 200 iterations of the BOMV-GWO algorithm. The number of grey wolves (i.e., Q) is set to be 50. In this simulation scenario, there are 3 URLLC-type terminals and 6 mMTC-type terminals. The unreliability requirements of the URLLC-type terminals and the mMTC-type terminals are 10^{-7} and 10^{-5} respectively. And their requirements of delay bounds are 1 ms and 100 ms respectively. The arrival parameters are set as follows, $p_1^a = 0.1, p_2^a = 0.001, q_1^a = 0.5$, $q_2^a = 0.75$. It can be seen that during the first 100 iterations, the optimal throughput value fluctuates considerably. But it

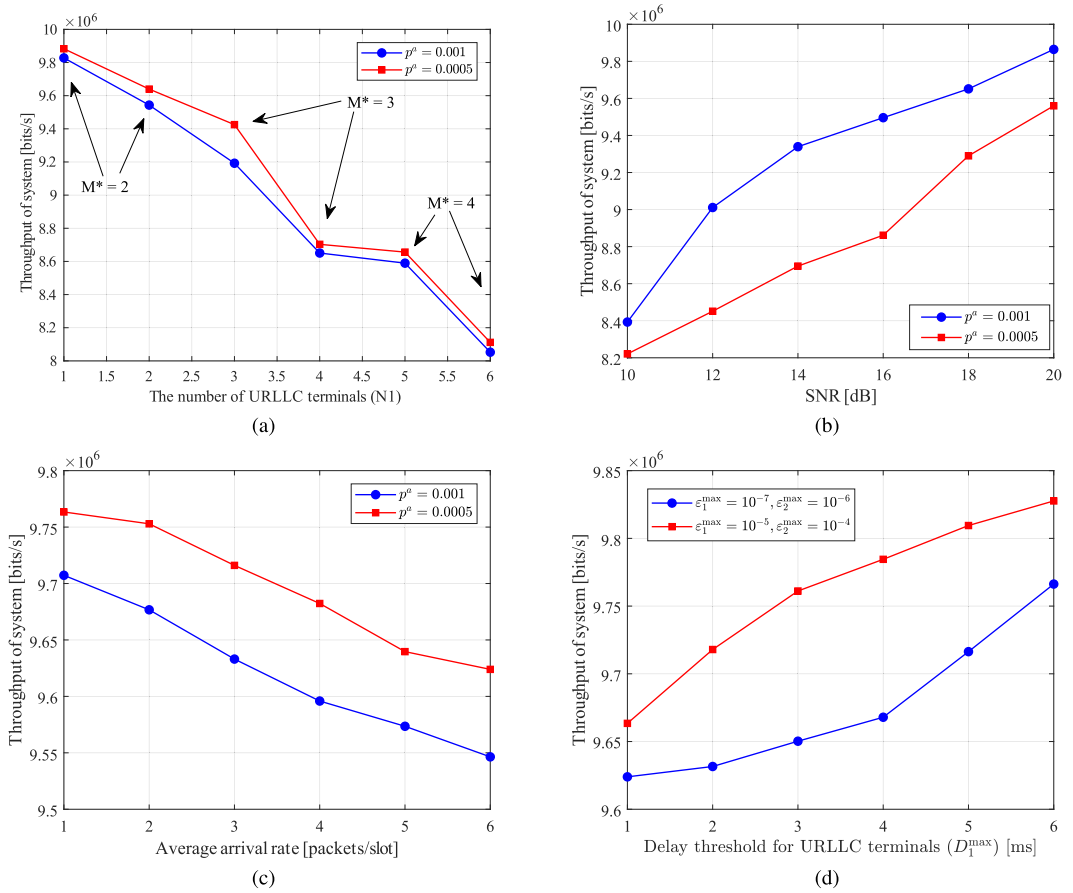

FIGURE 7. Throughput of the network achieved by the proposed M-ALOHA algorithm in different cases. (a) Different p^a and N_1, SNR = 15 dB, average arrival rate = 1 packets/slot, $D_1^{\max} = 1$ ms, $\varepsilon_1^{\max} = 10^{-7}$, $D_2^{\max} = 100$ ms, $\varepsilon_2^{\max} = 10^{-5}$. (b) Different p^a and SNR, $N_1 = 3$, average arrival rate = 1 packets/slot, $D_1^{\max} = 1$ ms, $\varepsilon_1^{\max} = 10^{-7}$, $D_2^{\max} = 100$ ms, $\varepsilon_2^{\max} = 10^{-5}$. (c) Different p^a and average arrival rate, $N_1 = 3$, SNR = 15 dB, $D_1^{\max} = 1$ ms, $\varepsilon_1^{\max} = 10^{-7}$, $D_2^{\max} = 100$ ms, $\varepsilon_2^{\max} = 10^{-5}$. (d) Different $\varepsilon_1^{\max}$, $\varepsilon_2^{\max}$ and $D_1^{\max}$, $D_2^{\max} = 100$ ms, SNR = 15 dB, $N_1 = 3$, average arrival rate = 1 packets/slot, $p^a = 0.001$.

is going to flatten out over the next 100 iterations. It shows that the BOMV-GWO algorithm we adopted to solve COP is a convergence algorithm. Fig. 6(b) shows the access probabilities for both URLLC-type and mMTC-type terminals obtained by the BOMV-GWO algorithm. Obviously, it can be seen that the access probabilities for URLLC-type terminals (i.e., Terminal 1-3) are much larger than mMTC-type terminals (i.e., Terminal 4-9). Large access probabilities imply more opportunities to transmit data packets. It is indicated that our proposed M-ALOHA algorithm satisfies the ultra-stringent QoS of URLLC-type terminals by making them have much more transmission opportunities. Besides, the optimum system parameters are obtained as follows: $B^* = 6.67$ MHz, $M^* = 3$, and $R_i^{s*} = 1$ packets/slot. It indicates that a moderate service rate is enough for sporadic arrivals. Compared with the grant-based algorithm that assigns a dedicated channel for URLLC-type terminal, our grant-free algorithm accommodates more mMTC-type terminals.

Fig. 7 shows the key factors affecting the throughput of the constructed MTC network including that the number of URLLC-type terminals (i.e., N_1), the value of average SNR, the average arrival rate and the delay threshold for URLLC-type terminals (i.e., $W_1^{\max}$). The other parameters are set as follows: the MMOO arrival parameter $q^a = 0.9$, the number of mMTC-type terminals $N_2 = 6$. The bandwidth for a sub-channel, the number of sub-channels and the packets transmission rate are optimized by the BOMV-GWO algorithm we proposed.

Fig. 7(a) shows the throughput performance of our proposed M-ALOHA algorithm in the case of different MMOO arrival parameter p^a and the number of URLLC-type terminals (i.e., N_1). The number of mMTC-type terminals (i.e., N_2) is set to be 6. The number of sub-channels (i.e., M) in Fig. 7(a) is obtained by the BOMV-GWO algorithm. As observed from Fig. 7(a), the throughput of the network declines as the number of URLLC-type terminals increases. There are two reasons for this phenomenon. Firstly, for supporting more URLLC-type terminals, and achieving the stringent delay-QoS requirement and the ultra reliability for URLLCs, it is a reasonable cost of throughput for obtaining

this benefit. Secondly, from the perspective of formula (33), the key factors that affect the throughput are the bandwidth for a sub-channel (i.e., B), the successful transmission probability for each terminal, and the packets transmission rate. Obviously, as the number of URLLC-type terminals grows, the number of sub-channels that is needed also increases. For a given total system bandwidth, the bandwidth for each sub-channel will be decreased. Besides, the increase of the number of terminals will lead to more collisions, thus a smaller probability of successful transmission of each terminal. In addition, this result also indicates that our proposed grant-free access algorithm can adjust the number of sub-channels for accommodating more URLLC terminals.

Fig. 7(b) shows the performance of our proposed M-ALOHA algorithm in the case of different MMOO arrival parameters p^a and the value of SNR. It is obviously that the throughput of system increases along with the value of SNR. It is reasonable because the increase of SNR will decrease the block error rate (BLER), the probability of successful transmission of each terminal will be increased.

Fig. 7(c) shows the performance of our proposed M-ALOHA algorithm in the case of different MMOO arrival parameters p^a and the average arrival rate. It can be seen that the throughput of system declines with the average arrival rate increases. It is reasonable because the higher the average arrival rate is, the more difficult to satisfy the requirements of delay and reliability QoS. Thus the set of the feasible solutions of (34) becomes smaller. Therefore, the throughput of system becomes lower.

Fig. 7(d) shows the performance of our M-ALOHA algorithm in the case of different probability of delay violation and the delay threshold of URLLC terminals. It can be seen that the throughput of system increases when the delay threshold of URLLC terminals (i.e., $W_1^{\max}$) increases for both different QoS requirement (i.e., the value of $\varepsilon_1^{\max}$ and $\varepsilon_2^{\max}$). According to formula (8), a smaller delay threshold means that a higher delay-QoS requirement. Obviously, it is easier to search for the feasible solutions when the $W_1^{\max}$ is higher, which matches the fact that it is easier to guarantee the delay-QoS requirement when the delay threshold is bigger. Then, the throughput of the system increases. Additionally, a smaller value of $\varepsilon_1^{\max}$ and $\varepsilon_2^{\max}$ means that a higher requirement of reliability. It can be obviously seen that the throughput of system become lower when the reliability constrains become stricter (i.e. when $\varepsilon_1^{\max}$ and $\varepsilon_2^{\max}$ become smaller). It is reasonable because the feasible solution set becomes larger as its reliability constraints become looser. As a result, a higher throughput of system is achieved.

VI. CONCLUSION

In this paper, we proposed the multi-channel ALOHA-type (M-ALOHA) grant-free access algorithm for an uplink MTC network with delay-tolerant mMTC-type terminals and URLLC-type terminals co-existence. we construct a statistical service model characterizing the transmission rate of each terminal with joint consideration of the features of M-ALOHA access scheme, short packet transmissions and frequency-selective fading channel. Then, based on the service-martingales theory, we derived the delay-bound violation probability for both the cases where the saturated state and unsaturated state are available at a terminal. The simulation results showed that our martingales-based delay analysis was accurate. Finally, we formulated our M-ALOHA algorithm as a system throughput maximization problem with multiple constraints including the martingales-based statistical delay-QoS constraint and the constraint of the system total bandwidth. The COP was handled via the proposed BOMV-GWO algorithm. The optimum access probability for each terminal and the optimum parameters of system design were all obtained. The simulation results showed that our M-ALOHA algorithm could meet the highly stringent delay and reliability QoS of URLLCs via providing much more access opportunities to URLLC-type terminals. Additionally, simulation results showed that the M-ALOHA algorithm could support more terminals in the system. In the future, we desire to study the active terminals estimation algorithm and the grant-free NOMA algorithm in MTC network.

REFERENCES

[1] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Optimal cross-layer resource allocation for critical MTC traffic in mixed LTE networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5944–5956, Jun. 2019.

[2] S. Ali, N. Rajatheva, and W. Saad, "Fast uplink grant for machine type communications: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 97–103, Mar. 2019.

[3] L. Zhao, X. Chi, L. Qian, and W. Chen, "Analysis on latency-bounded reliability for adaptive grant-free access with multipackets reception (MPR) in URLLCs," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 892–895, May 2019.

[4] 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.

[5] M. Gharbieh, H. ElSawy, H.-C. Yang, A. Bader, and M.-S. Alouini, "Spatiotemporal model for uplink IoT traffic: Scheduling and random access paradox," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8357–8372, Dec. 2018.

[6] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System level analysis of uplink grant-free transmission for URLLC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Singapore, Dec. 2017, pp. 1–6.

[7] N.-T. Nguyen, B.-H. Liu, S.-I. Chu, and H.-Z. Weng, "Challenges, designs, and performances of a distributed algorithm for minimum-latency of data-aggregation in multi-channel WSNs," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 1, pp. 192–205, Mar. 2019.

[8] O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, "Multi-channel random access with replications," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2538–2542.

[9] J.-B. Seo, H. Jin, and B. C. Jung, "Multichannel uplink NOMA random access: Selection diversity and bistability," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1515–1519, Sep. 2019.

[10] J. Choi, "An effective capacity-based approach to multi-channel low-latency wireless communications," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2476–2486, Mar. 2019.

[11] L. Zhao, X. Chi, and S. Yang, "Optimal ALOHA-like random access with heterogeneous QoS guarantees for multi-packet reception aided visible light communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7872–7884, Nov. 2016.

[12] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 630–643, May 2003.

[13] F. Poloczek and F. Ciucu, "Service-martingales: Theory and applications to the delay analysis of random access protocols," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr. 2015, pp. 945–953.
[14] T. Liu, L. Sun, R. Chen, F. Shu, X. Zhou, and Z. Han, "Martingale theory-based optimal task allocation in heterogeneous vehicular networks," *IEEE Access*, vol. 7, pp. 122354–122366, May 2019.
[15] Y. Hu, H. Li, Z. Chang, R. Hou, and Z. Han, "End-to-end backlog and delay bound analysis using martingale for Internet of vehicles," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Helsinki, Finland, Sep. 2017, pp. 98–103.
[16] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge Univ. Press, 2015.
[17] 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.
[18] L. Zhang and Y.-C. Liang, "Average throughput analysis and optimization in cooperative IoT networks with short packet communication," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11549–11562, Dec. 2018.
[19] K. Hammad, A. Moubayed, A. Shami, and S. Primak, "Analytical approximation of packet delay jitter in simple queues," *IEEE Wireless Commun. Lett.*, vol. 5, no. 6, pp. 564–567, Dec. 2016.
[20] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.
[21] L. Li, L. Sun, W. Kang, J. Guo, C. Han, and S. Li, "Fuzzy multilevel image thresholding based on modified discrete grey wolf optimizer and local information aggregation," *IEEE Access*, vol. 4, pp. 6438–6450, Sep. 2016.
[22] Z. Zhou, R. Ratasuk, N. Mangalvedhe, and A. Ghosh, "Resource allocation for uplink grant-free ultra-reliable and low latency communications," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Proto, Portugal, Jun. 2018, pp. 1–5.

RUIZHE QI was born in Jilin, China, in 1997. She received the B.S. degree from the Department of Communications Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China, in 2018. She is currently pursuing the M.S. degree with the College of Communication Engineering, Jilin University, Changchun, China. Her research interests include random access algorithms, ultrareliable and low-latency communications, and delay-quality of service (QoS) guarantees.

XUEFEN CHI received the B.Eng. degree in applied physics from the Beijing University of Posts and Telecommunications, Beijing, China, in 1984, and the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1990 and 2003, respectively. She was a Visiting Scholar with the Department of Computer Science, Loughborough University, Loughborough, U.K., in 2007, and the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in 2015. She is currently a Professor with the Department of Communications Engineering, Jilin University, Changchun. Her current research interests include machine type communications, indoor visible light communications, random access algorithms, delay-quality of service (QoS) guarantees, and queuing theory and its applications.

LINLIN ZHAO received the B.Eng., M.S., and Ph.D. degrees from the Department of Communications Engineering, Jilin University, Changchun, China, in 2009, 2012, and 2017, respectively. From 2017 to 2019, she held a postdoctoral position at the Department of Communications Engineering, Jilin University. She joined Jilin University, in 2019. She is currently a Postdoctoral Research Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. Her current research interests include throughput optimal random access algorithms, resource allocation schemes, and delay and reliability analysis and optimization, especially for reliability analysis of ultrareliable low-latency communications. She was a recipient of the Best Ph.D. Thesis Award of Jilin University, in 2017, and acquired the Macau Young Scholars Program, in 2019.

WANTING YANG was born in Jilin, China, in 1996. She received the B.S. degree from the Department of Communications Engineering, Jilin University, Changchun, China, in 2018. She is currently pursuing the M.S. degree with the College of Communication Engineering, Jilin University. Her research interests include wireless video transmission, ultrareliable and low-latency communications, and modeling and performance of 5G wireless radio networks.