Double-Threshold-Based Massive Random Access Protocol for Heterogeneous MTC Networks

Yuncong Xie[1,2], Pinyi Ren[1,2(✉)], and Dongyang Xu[1,2]

[1] School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China
qy1z29@stu.xjtu.edu.cn , pyren@mail.xjtu.edu.cn , xudongyang@xjtu.edu.cn
[2] Shaanxi Smart Networks and Ubiquitous Access Research Center, Xi'an 710049, People's Republic of China

Abstract. In this paper, we consider the uplink transmission of a heterogeneous MTC network with massive number of delay-insensitive terminals and URLLC terminals coexistence, and propose a novel double-threshold-based massive random access (DT-MRA) protocol. On the one hand, the proposed DT-MRA protocol allows partial delay-insensitive terminals temporarily preempting URLLC spectrum channels to alleviate the traffic overload, via adaptively tuning access control parameters including ACB factor and traffic offloading factor. On the other hand, a grant-free access mechanism with packet repetition is applied to support the stringent QoS requirements of URLLC. Then, we formulate an optimization problem to maximize the access throughput of delay-insensitive terminals, while satisfying the QoS requirements of each URLLC terminal. Based on the statistical traffic load information, an optimal access control strategy is also developed to solve this optimization problem. Simulation results validate the theoretical analyses and demonstrate the effectiveness of our proposed DT-MRA protocol, in terms of improving access throughput.

Keywords: Random access protocol · Massive machine-type communications (mMTC) · Ultra-reliable and low-latency communications (URLLC)

1 Introduction

The fifth generation (5G) cellular networks is envisioned to support seamless data exchanges among machine-type devices (MTDs), with minimum or no human intervention. Based on the application scenario and design requirement,

This research work is supported in part by the National Science and Technology Major Project under Grant No. 2018ZX03001003-004, and in part by the Fundamental Research Funds for the Central Universities.

B. Li et al. (Eds.): IoTaaS 2020, LNICST 346, pp. 493–509, 2021.
https://doi.org/10.1007/978-3-030-67514-1_40

5G machine-type communication services fall into two main categories [1]. One is the massive machine-type communications (mMTC) with massive connectivity and delay-insensitive traffic. The other is the ultra-reliable and low-latency communications (URLLC), which has a stringent Quality-of-Service (QoS) requirement in end-to-end (E2E) latency bound (e.g., 1 ms) and packet loss probability (e.g., no more than 10^{-5}). With the development of wireless technologies, there appears some emerging use cases with mMTC and URLLC services coexistence. Take the industrial internet-of-things (IoT) as an example, there exists a massive number of delay-insensitive sensors to monitor the factory environment, such as temperature and humidity. Apart from this, there also exists a relatively small number of URLLC sensors to perform mission-critical tasks, such as remote control and alarming [2]. Obviously, how to design the radio access protocol of heterogeneous MTC networks with diverse services is a challenging issue.

In this paper, we focus on the uplink transmission of heterogeneous MTC networks with massive delay-insensitive and URLLC terminals coexistence. Traditionally, LTE cellular networks use the grant-based uplink transmission protocol. Under this protocol, each terminal needs to carry out multiple control signaling exchanges before the uplink data transmission, such as four-step handshaking, scheduling request and scheduling grant [3]. Obviously, the grant-based transmission protocol is not suitable to MTC terminals with sporadic and small packet arrivals, which can be explained in the following two aspects: On the one hand, the massive access of delay-insensitive MTC terminals at the same time will lead to a serious radio access network (RAN) congestion and low access efficiency. In view of this, some congestion mitigating mechanisms, such as access class barring (ACB) [4], group paging [5], and traffic offloading [6], are developed to sustain the successful operation of massive access. On the other hand, the connection-oriented RA procedure induces excessive control signaling overhead and causes a large access delay, which is conflicting with the ultra-low E2E latency requirement of URLLC terminals. To overcome this bottleneck, the authors in [6] proposed a grant-free uplink access scheme with repetition coding, where each URLLC terminal can directly transmits its packets to the BS in a contention-based manner without connection-establishment, and transmits multiple copies of the same packet to achieve target reliability within the given E2E latency bound.

The major contributions of this paper are summarized as follows: Firstly, we propose a novel double-threshold-based massive random access (DT-MRA) protocol, which allows partial delay-insensitive terminals temporarily preempting URLLC spectrum channels to alleviate the traffic overload caused by massive access. Then, an optimization problem is formulated to maximize the access throughput of delay-insensitive terminals, while satisfying the QoS requirements of each URLLC terminal. Based on the statistical traffic load information, an optimal access control strategy is also developed to solve this optimization problem. Simulation results validate the theoretical analyses and demonstrate the effectiveness of our proposed DT-MRA protocol, in terms of improving access throughput.

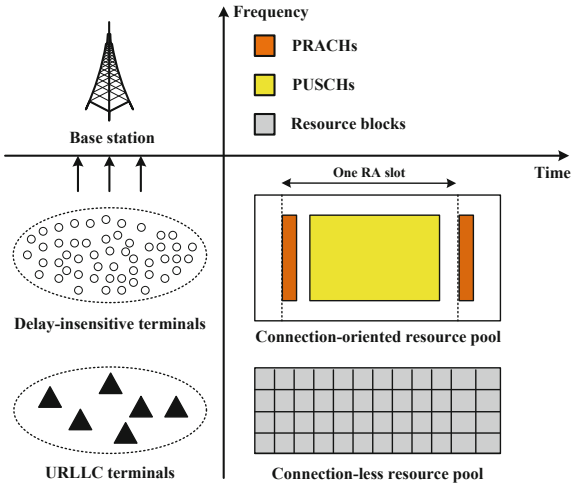

Fig. 1. Graphical representation of the heterogeneous MTC netwotk with delay-insensitive and URLLC terminals coexistence.

2 System Model

As depicted in Fig. 1, we consider a heavy-loaded MTC network consisting of one BS, K delay-insensitive terminals and L URLLC terminals. Time is discretized into slots with a duration of τ, which is the basic time unit of the system. In particular, the QoS requirement of each URLLC terminal is characterized by an E2E latency bound T_{th} and the corresponding maximal allowable packet loss probability ϕ_{th}.

2.1 Time-Frequency Resource Configuration

With the consideration of differentiated characteristics between delay-insensitive and URLLC terminals, the BS prepares two independent spectrum resource pools that equipped with different RA procedures and time-frequency resource configurations. One is called as connection-oriented resource pool, which utilizes the traditional LTE-based RA procedure to support the massive access of delay-insensitive terminals, i.e., each terminal randomly selects one of M available preambles and transmits to the BS over the physical random access channel (PRACH). Note that the PRACH is a certain time-frequency resource blocks (RBs) that appear periodically [7]. Denote the interval between two consecutive PRACHs as one RA slot, whose duration is T_{RA} and contains I_{RA} consecutive slots. The RA procedure is successful when a preamble is only selected by one

terminal, and then BS will assigns dedicated physical uplink shared channel (PUSCH) for this terminal to complete the uplink data transmission.[1]

The other is called as connection-less resource pool, which utilizes a grant-free RA procedure to reduce the control signaling overhead and satisfy the ultra-low E2E requirement of URLLC, i.e., each terminal can directly transmits its packets to BS in a contention-based manner. In the connection-less resource pool, the total bandwidth is equally divided into N orthogonal channels, such that there exists NI_{RA} RBs within each RA slot[2]. Due to the small payload characteristic of delay-insensitive and URLLC terminals, we assume that only one RB is required to complete the uplink data transmission. For simplicity, we further assume that transmission errors can only occur in the channel collision, i.e., the impact of noise and other channel imperfections are negligible.

2.2 Packet Arrival Process

To simplify the packet arrival process of delay-insensitive terminal, we assume that newly packet arrivals will only take place at the beginning of each RA slot. Moreover, each delay-insensitive terminal has an infinite-size queue buffer to store newly arrival packets and the number of newly arrival packets per RA slot follows the Poisson distribution with a parameter of $\lambda \in (0, 1)$. One access request will be generated when a new packet is arrived, and each delay-insensitive terminal can sustain at most one ongoing access request, regardless of the number of packets in the queue buffer. Therefore, the data transmission of each delay-insensitive terminal can be modeled as a $Geo/G/1/1$ queue [8].

Due to the sporadic packet arrival characteristic, the newly packet arrival of each URLLC terminal can be modeled as a poisson process with exponentially distributed inter-arrival time, and denote the average number of new arrival packets per slot as $\mu \in (0, 1)$. Since the packet inter-arrival interval in typical URLLC applications is much longer than the E2E delay bound T_{th} (e.g., 1 ms) [9], it is reasonable to assume that the queueing delay is zero because the packets will be immediately dropped once the E2E delay bound expires.

3 Protocol Description and Performance Analysis

3.1 DT-MRA Protocol Description

To effectively support the massive access characteristic of delay-insensitive terminals and the stringent QoS requirement of URLLC terminals, we propose a novel double-threshold-based massive random access (DT-MRA) protocol and the detail is described as follows: On the one hand, the URLLC terminals can

[1] This conference paper only focus on the first step of connection-oriented RA procedure, i.e, preamble transmission in the PRACH, and assume that the number of PUSCHs is sufficiently large to complete the data transmission.

[2] Note that one RB in the connection-less resource pool is defined as a time-frequency block with one channel bandwidth and one slot duration.

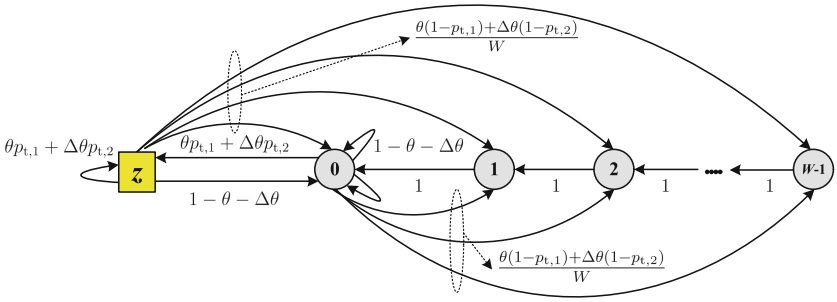

Fig. 2. State transition diagram of each delay-insensitive terminal.

only utilize the connection-less resource pool for the grant-free data transmission to achieve an ultra-low E2E latency. Meanwhile, a packet repetition mechanism is also applied to guarantee the ultra-high transmission reliability, which allows each URLLC terminal transmits the same packet Q_{th} times in consecutive slots, via randomly selects one out of N orthogonal channels at each slot. Based on the time-frequency configuration of connection-less resource pool, we have $Q_{\text{th}} = \lfloor T_{\text{th}}/\tau \rfloor$ to satisfy the E2E latency bound of URLLC.

On the other hand, each delay-insensitive terminal can either perform LTE-based RA procedure in the connection-oriented resource pool, or temporarily utilizes one RB in the connection-oriented resource pool to perform grant-free RA procedure. To realize the function mentioned above, a double-threshold ACB scheme is developed for massive access of delay-insensitive terminals. At the beginning of each RA slot, the BS will broadcast two access control thresholds, including the ACB factor $\theta \in (0, 1]$ and the traffic offloading factor $\Delta\theta \in [0, 1)$. Then, each delay-insensitive terminal generates a random number $\delta \in [0, 1]$ in a non-cooperative manner, and performs the ACB check via comparing it with the given two access control thresholds:

1. If $\delta \leq \theta$, the terminal will attempts to perform the LTE-based RA produce in connection-oriented resource pool, via randomly selects one out of M preambles.
2. If $\theta < \delta \leq \theta + \Delta\theta$, the terminal will performs the grant-free RA produce in connection-less resource pool, via randomly selects one out of NI_{RA} RBs.
3. Otherwise, the terminal is rejected by the ACB check mechanism temporarily and reattempt to initiate access at the beginning of next RA slot.

3.2 Transmission Performance Analysis

In this subsection, we will analyze the transmission performance of each delay-insensitive and URLLC terminal. As depicted in Fig. 2, the transmission behaviour of each delay-insensitive terminal can be modeled as a discrete-time Markov process. At each RA slot t, denote $p_{t,1}$ and $p_{t,2}$ as the successful transmission probability in connection-oriented resource pool and connection-less

resource pool, respectively. The data transmission phase is initially in state z, and remain in state z when it passes the ACB check and successfully transmitted with a probability of $\theta p_{\mathrm{t},1} + \Delta\theta p_{\mathrm{t},2}$. When the transmission collision occurs, it randomly selects a integer value between 0 and $W-1$, where W is the Uniform Backoff (UB) window size in unit of RA slot, and moving to state $s \in \{0, 1, ..., W-1\}$ with equal probability $\frac{\theta(1-p_{\mathrm{t},1})+\Delta\theta(1-p_{\mathrm{t},2})}{W}$, and counts down at each RA slot until it reaches state 0. Otherwise, it shifts to state 0 that implies the terminal is rejected by the ACB check temporarily and reattempt to initiate access at the beginning of next RA slot.

Based on the discussion above, the steady-state probability distribution in Fig. 2 can be obtained as

$$\begin{cases} \pi_z = \left(\dfrac{(W-1)(\theta(1-p_1)+\Delta\theta(1-p_2))}{2(\theta p_1 + \Delta\theta p_2)} + \dfrac{1}{\theta p_1 + \Delta\theta p_2} \right)^{-1}, \\ \pi_0 = \dfrac{1-(\theta p_1 + \Delta\theta p_2)}{\theta p_1 + \Delta\theta p_2}\pi_z, \\ \pi_s = \dfrac{(W-s)(\theta(1-p_1)+\Delta\theta(1-p_2))}{W(\theta p_1 + \Delta\theta p_2)}\pi_z, \ s = 1, 2, ..., W-1, \end{cases} \tag{1}$$

where $p_1 = \lim_{t\to\infty} p_{\mathrm{t},1}$ and $p_2 = \lim_{t\to\infty} p_{\mathrm{t},2}$ is the steady-state success transmission probability in connection-oriented and connection-less resource pools, respectively.

For each delay-insensitive terminal, denote the nonempty probability of data queue as ρ. If one particular terminal performs RA procedure in the connection-oriented resource pool, via randomly selects one of M available preambles. The data transmission is successful only when the behavior of other $K-1$ delay-insensitive terminals satisfies the following conditionals: 1) the data queue is empty, and the corresponding probability is $1-\rho$, 2) the data queue is nonempty, but not transmitting in the connection-oriented resource pool, the corresponding probability is $\rho\big((1-\theta)(\pi_0+\pi_z)+\sum_{j=1}^{W-1}\pi_j\big)$, and 3) the data transmission occurs in the connection-oriented resource pool, but not select the same preamble as the terminal of interest, the corresponding probability is $\rho\theta(1-\frac{1}{M})(\pi_0+\pi_z)$. Based on the analysis above, the mathematical expression of p_1 can be obtained as

$$\begin{aligned} p_1 &\triangleq \binom{M}{1}\frac{1}{M}\left((1-\rho)+\rho\Big((1-\theta)(\pi_0+\pi_z)+\sum_{j=1}^{W-1}\pi_j+\theta(1-\frac{1}{M})(\pi_0+\pi_z)\Big)\right)^{K-1} \\ &= \left(1-\rho\frac{\theta}{M}(\pi_0+\pi_z)\right)^{K-1} = \left(1-\frac{\rho\theta\pi_z}{M(\theta p_1+\Delta\theta p_2)}\right)^{K-1}, \end{aligned} \tag{2}$$

where the closed-form expression of ρ is given by [8]

$$\rho = \frac{\lambda}{\lambda+\pi_z}, \tag{3}$$

since the value K is large, we have $K-1 \approx K$ and $(1-x)^K \approx e^{-Kx}$ for $0 < x < 1$. By combining (1) and (3), the mathematical expression of (2) can

be approximated as

$$p_1 \overset{\text{when } K \text{ is large}}{\approx} \exp\left(-K\frac{\rho\theta\pi_z}{M(\theta p_1+\Delta\theta p_2)}\right) = \exp\left(-\frac{K}{M}\frac{1}{\theta p_1+\Delta\theta p_2}\frac{\lambda\theta}{1+\lambda\left(\frac{1}{\theta p_1+\Delta\theta p_2}\left(1+\frac{W-1}{2}(\theta+\Delta\theta)\right)-\frac{W-1}{2}\right)}\right). \tag{4}$$

To evaluate the access performance of all delay-insensitive terminals, we are interested in the average access throughput $\bar{\lambda}_{\text{out}}$ of delay-insensitive traffic, which is defined as the average number of delay-insensitive terminals that successful access to the BS per RA slot. Based on the characteristic of $Geo/G/1/1$ queue model, the mathematical expression of $\bar{\lambda}_{\text{out}}$ can be written as

$$\bar{\lambda}_{\text{out}} = \tilde{\lambda}(1-\rho), \tag{5}$$

where $\tilde{\lambda} = K\lambda$ is the aggregate data arrival rate of delay-insensitive terminals, the mathematical expression of π_z and ρ are given by (1) and (3), respectively. By substituting (1) and (3) into (5), the average access throughput of delay-insensitive traffic is obtained as

$$\bar{\lambda}_{\text{out}} = \frac{K\lambda}{1+\lambda\left(\frac{(W-1)(\theta(1-p_1)+\Delta\theta(1-p_2))}{2(\theta p_1+\Delta\theta p_2)}+\frac{1}{\theta p_1+\Delta\theta p_2}\right)}. \tag{6}$$

In the following, we analyzes the transmission performance of URLLC: As an analytical tool that indicates the degree of reliability loss, the packet loss probability of each URLLC terminal is defined as the probability that all Q_{th} packets are not successfully transmitted within the E2E latency bound T_{th}. Therefore, the packet loss probability of each URLLC terminal is given by $\phi = p_{\text{c}}^{Q_{\text{th}}}$, where p_{c} is the collision probability of each individual packet. Since the other $L-1$ URLLC terminals and K delay-insensitive terminals might collided with the packet of interest, denote the corresponding probability as p_{c1} and p_{c2}, respectively. Therefore, the mathematical expression of packet loss probability can be rewritten as

$$\phi = \left(1-(1-p_{\text{c1}})(1-p_{\text{c2}})\right)^{Q_{\text{th}}}. \tag{7}$$

In general, the probability that one URLLC terminal has packets to transmit is $p_{\text{ra}} \triangleq 1-e^{-\mu}$. For each of other $L-1$ URLLC terminals, the probability that its generated packets not collided with the packet of interest, i.e., randomly selects one out of other $N-1$ orthogonal channels, is obtained as $p_{\text{access}} \triangleq \binom{N-1}{1}/\binom{N}{1} = \frac{N-1}{N}$. Therefore, the mathematical expression of p_{c1} is given by

$$p_{\text{c1}} = 1-\left((1-p_{\text{ra}})+p_{\text{ra}}p_{\text{access}}\right)^{L-1} = 1-\left(\frac{e^{-\mu}+N-1}{N}\right)^{L-1}. \tag{8}$$

Apart from this, to avoid the collision with the packet of interest, the behavior of K delay-insensitive terminals should also satisfies one of the following conditions:

1) the data queue of delay-insensitive terminal is empty with probability $1-\rho$, 2) the data queue of delay-insensitive terminal is nonempty, but not transmitting in the connection-less resource pool, the corresponding probability is $\rho\big((1-\Delta\theta)(\pi_0+\pi_z)+\sum_{s=1}^{W-1}\pi_s\big)$, and 3) the data transmission of delay-insensitive terminal occurs in the connection-less resource pool, but not select the same RB as the packet of interest, the corresponding probability is $\rho\Delta\theta(1-\frac{1}{NI_{RA}})(\pi_0+\pi_z)$. Therefore, the mathematical expression of p_{c2} is given by

$$\begin{aligned} p_{c2} &= 1-\left[(1-\rho)+\rho\Big((1-\Delta\theta)(\pi_0+\pi_z)+\sum_{s=1}^{W-1}\pi_s+\Delta\theta\Big(1-\frac{1}{NI_{RA}}\Big)(\pi_0+\pi_z)\Big)\right]^K \\ &= 1-\left[(1-\rho)+\rho\Big(\Big(1-\frac{\Delta\theta}{NI_{RA}}\Big)(\pi_0+\pi_z)+\sum_{s=1}^{W-1}\pi_s\Big)\right]^K, \end{aligned} \tag{9}$$

substituting (8) and (9) into (7), the packet loss probability of each URLLC terminal can be rewritten as

$$\begin{aligned} \phi &= \left[1-\left((1-\rho)+\rho\Big((1-\frac{\Delta\theta}{NI_{RA}})(\pi_0+\pi_z)+\sum_{s=1}^{W-1}\pi_s\Big)\right)^K\left(\frac{e^{-\mu}+N-1}{N}\right)^{L-1}\right]^{Q_{\text{th}}} \\ &\triangleq \left[\omega\Big(1-\frac{\rho\Delta\theta\pi_z}{NI_{RA}(\theta p_1+\Delta\theta p_2)}\Big)^K\right]^{Q_{\text{th}}}, \end{aligned} \tag{10}$$

where $\omega=\big(\frac{e^{-\mu}+N-1}{N}\big)^{L-1}$, and since the value K is large, we have the approximation $(1-x)^K\approx e^{-Kx}$ for $0<x<1$. By combining (1) and (3), the mathematical expression of (10) can be approximated as

$$\begin{aligned} \phi &\overset{\text{when } K \text{ is large}}{\approx} \left[\omega\exp\Big(-K\frac{\rho\Delta\theta\pi_z}{NI_{RA}(\theta p_1+\Delta\theta p_2)}\Big)\right]^{Q_{\text{th}}} \\ &= \left[\omega\exp\left(-\frac{K}{NI_{RA}}\frac{1}{\theta p_1+\Delta\theta p_2}\frac{\lambda\Delta\theta}{1+\lambda\big(\frac{1}{\theta p_1+\Delta\theta p_2}\big(1+\frac{W-1}{2}(\theta+\Delta\theta)\big)-\frac{W-1}{2}\big)}\right)\right]^{Q_{\text{th}}}, \end{aligned} \tag{11}$$

Similar to the theoretical derivations above, the mathematical expression of p_2 is also obtained as

$$\begin{aligned} p_2 &= \left[(1-\rho)+\rho\Big(\Big(1-\frac{\Delta\theta}{NI_{RA}}\Big)(\pi_0+\pi_z)+\sum_{s=1}^{W-1}\pi_s\Big)\right]^{K-1}\left(\frac{e^{-\mu}+N-1}{N}\right)^L \\ &= \left[1-\frac{\rho\Delta\theta\pi_z}{NI_{RA}(\theta p_1+\Delta\theta p_2)}\right]^{K-1}\left(\frac{e^{-\mu}+N-1}{N}\right)^L. \end{aligned} \tag{12}$$

4 Optimal Access Control Strategy Design

4.1 Problem Formulation

In this subsection, we formulate an optimization problem to maximize the average access throughput of delay-insensitive terminals, while meeting the QoS requirement of each URLLC terminal. Therefore, the optimization problem can be mathematically written as

$$(\mathbf{P1}) \max_{\theta,\Delta\theta,W} \bar{\lambda}_{\text{out}} \tag{13}$$

$$\text{s.t.} \quad \phi \le \phi_{\text{th}}, \tag{14}$$

$$\theta \in (0,1], \ \Delta\theta \in [0,1), \tag{15}$$

$$\theta + \Delta\theta \le 1, \tag{16}$$

$$W \in \{1,2,...,W_{\max}\}, \tag{17}$$

where (14) implies the constraint in packet loss probability of each URLLC terminal, and (15) (16) (17) denote constraints in the values of ACB factor θ, traffic offloading factor $\Delta\theta$ and UB window size W, respectively.

Based on the observation in (6), we can draw the conclusion that $\bar{\lambda}_{\text{out}}$ is crucially determined by the value of p_1 and p_2, which is a function of θ, $\Delta\theta$ and W. By applying $K-1 \approx K$ and $L \approx L-1$ to (12), the expression of p_2 can be approximated as

$$p_2 \approx \left[1 - \frac{\rho\Delta\theta\pi_z}{NI_{RA}(\theta p_1 + \Delta\theta p_2)}\right]^K \left(\frac{e^{-\mu}+N-1}{N}\right)^{L-1}, \tag{18}$$

and then we have $p_2 \approx 1-\phi^{1/Q_{\text{th}}} \ge 1-{\phi_{\text{th}}}^{1/Q_{\text{th}}}$. Plugging (4) and (12) into (13), and performing some manipulations, the problem (P1) can be converted to the following equivalent optimization problem:

$$(\mathbf{P2}) \max_{p_1,p_2} \bar{\lambda}_{\text{out}} = -Mp_1 \ln p_1 - NI_{RA}p_2 \ln \frac{p_2}{\omega} \tag{19}$$

$$\text{s.t.} \ \ 0 < p_1 \le 1, \tag{20}$$

$$1 - \phi_{\text{th}}^{1/Q_{\text{th}}} \le p_2 \le 1, \tag{21}$$

4.2 Optimal Access Control Strategy

In this subsection, the optimal access control strategy is developed to obtain the maximal average access throughput $\bar{\lambda}_{\max} = \max_{\theta,\Delta\theta,W} \bar{\lambda}_{\text{out}}$ and the corresponding optimal setting of $\theta^*, \Delta\theta^*$ and W^*. Since $\bar{\lambda}_{\text{out}}$ is determined by the value of p_1 and p_2, which is a function of the ACB factor θ, the traffic offloading factor $\Delta\theta$ and the UB window size W, as shown in (4) and (12), respectively. Therefore, the general idea of solving optimization problem is given as follows: In the first step, we aim to solve the optimization problem (P2), and find the optimal setting

of p_1 and p_2 that maximizing the access throughput $\bar{\lambda}_{\text{out}}$, which is denoted as p_1^* and p_2^*. In the second step, we aim to find the corresponding optimal solution $(\theta^*, \Delta\theta^*, W^*)$ under given values of p_1^* and p_2^*. Based on the discussion above, we have the following theorem that characterizes the optimal solution:

Theorem 1. In the optimization problem (P2), the global optimal solution is $(p_1^*, p_2^*) = (e^{-1}, 1-\phi_{\text{th}}^{1/Q_{\text{th}}})$, and the corresponding maximum access throughput is $\bar{\lambda}_{\max} = Me^{-1} + NI_{RA}(1-\phi_{\text{th}}^{1/Q_{\text{th}}}) \ln \frac{\omega}{1-\phi_{\text{th}}^{1/Q_{\text{th}}}}$, where $\omega = \left(\frac{e^{-\mu}+N-1}{N}\right)^{L-1}$. Furthermore, the corresponding optimal setting of $(\theta^*, \Delta\theta^*, W^*)$ in the optimization problem (P1) should together satisfy

$$\frac{\Delta\theta^*}{\theta^*} = -\frac{NI_{RA}}{M} \ln \frac{1-\phi_{\text{th}}^{1/Q_{\text{th}}}}{\omega} \triangleq \alpha, \tag{22}$$

$$W^* = 2\frac{\frac{K\theta^*}{M} - \frac{1}{\lambda}\left(\theta^* e^{-1} + \Delta\theta^*(1-\phi_{\text{th}}^{1/Q_{\text{th}}})\right) - 1}{\theta^*(1-e^{-1}) + \Delta\theta^* \phi_{\text{th}}^{1/Q_{\text{th}}}} + 1, \tag{23}$$

$$\theta^* \in (0,1],\ \Delta\theta^* \in [0,1),\ \theta^* + \Delta\theta^* \leq 1. \tag{24}$$

Proof. In the optimization problem (P2), $\bar{\lambda}_{\text{out}}$ is a convex function of (p_1, p_2). For $0 < p_1 \leq 1$, we have $\frac{\partial}{\partial p_1}\bar{\lambda}_{\text{out}} \geq 0$ when $p_1 \in (0, e^{-1}]$ and $\frac{\partial}{\partial p_1}\bar{\lambda}_{\text{out}} < 0$ when $p_1 \in (e^{-1}, 1]$. For $1-\phi_{\text{th}}^{1/Q_{\text{th}}} \leq p_2 \leq 1$, we have $\frac{\partial}{\partial p_2}\bar{\lambda}_{\text{out}} < 0$ since $\omega e^{-1} < e^{-1} < 1-\phi_{\text{th}}^{1/Q_{\text{th}}}$. Therefore, the combination of $(p_1^*, p_2^*) = (e^{-1}, 1-\phi_{\text{th}}^{1/Q_{\text{th}}})$ is the global optimal solution of (P2), and the corresponding maximum access throughput is $\bar{\lambda}_{\max} = -Mp_1^* \ln p_1^* - NI_{RA}p_2^* \ln \frac{p_2^*}{\omega} = Me^{-1} + NI_{RA}(1-\phi_{\text{th}}^{1/Q_{\text{th}}}) \ln \frac{\omega}{1-\phi_{\text{th}}^{1/Q_{\text{th}}}}$. Furthermore, the mathematical expressions in (22) (23) and (24) can be obtained via substituting $(p_1^*, p_2^*) = (e^{-1}, 1-\phi_{\text{th}}^{1/Q_{\text{th}}})$ into (4) and (12), respectively. This completes the proof.

Remark 1. According to the constraint shown in (17), the value of UB window size W should not be less than 1, i.e., $W \geq 1$. Therefore, when $\frac{K\theta^*}{M} - \frac{1}{\lambda}\left(\theta^* e^{-1} + \Delta\theta^*(1-\phi_{\text{th}}^{1/Q_{\text{th}}})\right) < 1$, i.e., $\lambda < \frac{\theta^* e^{-1} + \Delta\theta^*(1-\phi_{\text{th}}^{1/Q_{\text{th}}})}{\frac{K\theta^*}{M}-1} = \frac{e^{-1}+\alpha(1-\phi_{\text{th}}^{1/Q_{\text{th}}})}{\frac{K}{M}-\frac{1}{\theta^*}} < \frac{e^{-1}+\alpha(1-\phi_{\text{th}}^{1/Q_{\text{th}}})}{\frac{K}{M}-1-\alpha} \triangleq \lambda_H$, the maximal access throughput $\bar{\lambda}_{\max}$ cannot be achieved since (23) does not hold for any combination of $\theta^*, \Delta\theta^*$ and W^*. That is to say, the maximal access throughput $\bar{\lambda}_{\max}$ can be achieved when $\lambda \geq \lambda_H$, and the corresponding optimal setting of $(\theta^*, \Delta\theta^*)$ and W^* can be adaptively tuning based on (22) (23) and (24).

Since the maximal access throughput $\bar{\lambda}_{\max}$ cannot be achieved when $\lambda < \lambda_H$, the following theorem characterizes the optimal setting of $(\theta^*, \Delta\theta^*, W^*)$ that maximizes the access throughput $\bar{\lambda}_{\text{out}}$ when $\lambda < \lambda_H$:

Theorem 2. On the one hand, the optimal setting of $(\theta^*, \Delta\theta^*, W^*)$ when $\frac{e^{-1}}{\frac{K}{M}-1} \triangleq \lambda_L \leq \lambda < \lambda_H$ should together satisfy the constraint that $\theta^* + \Delta\theta^* = 1$

and $W^* = 1$. On the other hand, the optimal setting of $(\theta^*, \Delta\theta^*, W^*)$ when $\lambda < \lambda_L$ has the unique solution of $\theta^* = 1$, $\Delta\theta^* = 0$ and $W^* = 1$. Denote the maximal access throughput when $\lambda_L \leq \lambda < \lambda_H$ and $\lambda < \lambda_L$ as $\bar{\lambda}'_{\max}$ and $\bar{\lambda}''_{\max}$, respectively. It is proved that $\bar{\lambda}'_{\max}$ and $\bar{\lambda}''_{\max}$ are monotonic increasing functions of λ, and $\bar{\lambda}''_{\max} < Me^{-1} \leq \bar{\lambda}'_{\max} < \bar{\lambda}_{\max}$.

Proof. Based on the idea of reduction to absurdity, assuming that $p_1^* = e^{-1}$ can be obtained when $\lambda < \lambda_L$, then substituting $p_1^* = e^{-1}$ and $\Delta\theta = 0$ into (4), we can get the following mathematical expression that

$$\frac{1}{\theta} + \frac{W-1}{2}(1 - e^{-1}) = \frac{K}{M} - \frac{e^{-1}}{\lambda}, \tag{25}$$

it can be observed that the (25) does not hold when $\frac{K}{M} - \frac{e^{-1}}{\lambda} < 1$, i.e., $\lambda < \frac{e^{-1}}{\frac{K}{M}-1} \triangleq \lambda_L$, which is conflict with previous assumptions. By performing some manipulations that $\frac{\partial}{\partial\lambda}\bar{\lambda}_{\text{out}} > 0$ and $\frac{\partial}{\partial p_1}\bar{\lambda}_{\text{out}} < 0$ for $p_1 \in (e^{-1}, 1)$, we can draw the conclusion that $p_1 > e^{-1}$ is founded when $\lambda < \lambda_L$, and the optimal value of p_1 is $p_1^* = e^{-1}$ when $\lambda \geq \lambda_L$. Similarly, we can also prove that the optimal value of p_2 is $p_2^* = 1 - \phi_{\text{th}}^{1/Q_{\text{th}}}$ if and only if $\lambda \geq \lambda_H$.

Based on the discussion above, we can draw the conclusion that $p_1^* = e^{-1}$ and $p_2^* > 1 - \phi_{\text{th}}^{1/Q_{\text{th}}}$, substituting it into (4) and (12), we can get the following expression that

$$\frac{\Delta\theta^*}{\theta^*} = -\frac{NI_{RA}}{M}\ln\frac{p_2^*}{\omega} \triangleq u(p_2^*), \tag{26}$$

$$\frac{W^*-1}{2} = \frac{\frac{K\theta^*}{M} - \frac{1}{\lambda}\left(\theta^* e^{-1} + \Delta\theta^* p_2^*\right) - 1}{\theta^*(1-e^{-1}) + \Delta\theta^*(1-p_2^*)}, \tag{27}$$

it can be observed that the (27) is hold if and only if $W = 1$, since $\lambda < \lambda_H$. Thus, the optimal setting of $(\theta^*, \Delta\theta^*)$ can be mathematically written as

$$\theta^* = \frac{1}{\frac{K}{M} - \frac{e^{-1}+p_2^* u(p_2^*)}{\lambda}}, \Delta\theta^* = \frac{u(p_2^*)}{\frac{K}{M} - \frac{e^{-1}+p_2^* u(p_2^*)}{\lambda}}, \tag{28}$$

and the corresponding maximal network throughput can be obtained as

$$\bar{\lambda}'_{\max} = Me^{-1} + NI_{RA}p_2^* \ln\frac{\omega}{p_2^*}, \tag{29}$$

since $\bar{\lambda}'_{\max}$ and $\theta^* + \Delta\theta^*$ is monotonically decreasing with p_2^*, with the constraint that $\theta^* + \Delta\theta^* \leq 1$, we can draw the conclusion that the value of $\bar{\lambda}'_{\max}$ can be maximized when $\theta^* + \Delta\theta^* = 1$. Substituting $\theta^* + \Delta\theta^* = 1$ into (28) and performing some manipulations, we can get the following expression that

$$\frac{K}{M} - \frac{e^{-1} + p_2^* u(p_2^*)}{\lambda} = 1 + u(p_2^*), \tag{30}$$

which has an unique root of p_2^* that maximizes the value of $\bar{\lambda}'_{\max}$.

Similar to the theoretical analysis above, we have $\theta^* = 1$, $\Delta\theta^* = 0$ and $W^* = 1$ when $\lambda < \lambda_L$, and the corresponding maximal network throughput is $\bar{\lambda}''_{\max} = -Mp_1^* \ln p_1^*$. Substituting $\theta^* = 1$, $\Delta\theta^* = 0$ and $W^* = 1$ into (4) and performing some manipulations, we can get the following mathematical expression that

$$\frac{K}{M} + (1 + \frac{1}{\lambda} p_1^*) \ln p_1^* = 0, \tag{31}$$

which has an unique root of p_1^* that maximizes the value of $\bar{\lambda}''_{\max}$. This completes the proof.

Remark 2. According to Theorem 1 and Theorem 2, we can define the following traffic load regions based on the value of λ_H and λ_L. Then, we can obtain the corresponding optimal access control strategy via comparing the values of λ, λ_H and λ_L.

1. **High Traffic Load Region** (when $\lambda \geq \lambda_H$), in which the maximal access throughput is $\bar{\lambda}_{\max} = Me^{-1} + NI_{RA}(1 - \phi_{\text{th}}^{1/Q_{\text{th}}}) \ln \frac{\omega}{1-\phi_{\text{th}}^{1/Q_{\text{th}}}}$, and the corresponding optimal setting of $(\theta^*, \Delta\theta^*)$ and W^* can be adaptively tuning based on (22) (23) and (24). For example, the optimal setting of $(\theta^*, \Delta\theta^*)$ when $W = 1$, which is denote as $\theta^*|_{W=1}$ and $\Delta\theta^*|_{W=1}$, can be written as

$$\theta^*|_{W=1} = \frac{1}{\frac{K}{M} - \frac{e^{-1} + \alpha(1-\phi_{\text{th}}^{1/Q_{\text{th}}})}{\lambda}}, \tag{32}$$

$$\Delta\theta^*|_{W=1} = \frac{\alpha}{\frac{K}{M} - \frac{e^{-1} + \alpha(1-\phi_{\text{th}}^{1/Q_{\text{th}}})}{\lambda}}, \tag{33}$$

Similarly, the optimal UB window size when $\theta + \Delta\theta = 1$, which is denote as $W^*|_{\theta+\Delta\theta=1}$, can be written as

$$W^*|_{\theta+\Delta\theta=1} = 2\frac{\frac{K}{M} - \frac{e^{-1} + \alpha(1-\phi_{\text{th}}^{1/Q_{\text{th}}})}{\lambda} - 1 - \alpha}{1 - e^{-1} + \alpha\phi_{\text{th}}^{1/Q_{\text{th}}}} + 1. \tag{34}$$

2. **Medium Traffic Load Region** (when $\lambda_L \leq \lambda < \lambda_H$), in which the maximal access throughput is $\bar{\lambda}'_{\max} = Me^{-1} + NI_{RA}p_2^* \ln \frac{\omega}{p_2^*} < \bar{\lambda}_{\max}$, where p_2^* is the unique solution of the following mathematical equation

$$\frac{K}{M} - \frac{e^{-1} + p_2^* u(p_2^*)}{\lambda} = 1 + u(p_2^*), \tag{35}$$

where $u(p_2^*) = -\frac{NI_{RA}}{M} \ln \frac{p_2^*}{\omega}$, the optimal setting of $(\theta^*, \Delta\theta^*, W^*)$ cannot be adaptively tuned, and has an unique solution of $\theta^* = \frac{1}{\frac{K}{M} - \frac{e^{-1} + p_2^* u(p_2^*)}{\lambda}}$, $\Delta\theta^* = \frac{u(p_2^*)}{\frac{K}{M} - \frac{e^{-1} + p_2^* u(p_2^*)}{\lambda}}$ and $W^* = 1$, respectively.

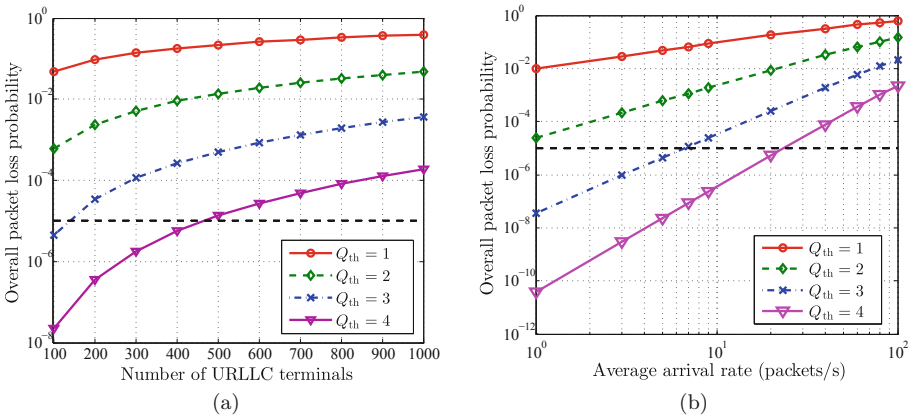

Fig. 3. Transmission performance of URLLC without traffic offloading mechanism (i.e., when $\Delta\theta = 0$): (a) Overall packet loss probability ϕ vs. Number of URLLC terminals L, where the number of channels in the connection-less resource pool $N = 10$ and the average arrival rate of each URLLC terminal is $\tilde{\mu} = 5$ (packets/s). (b) Overall packet loss probability ϕ vs. Average arrival rate $\tilde{\mu}$, where $L = 100$ and $N = 10$.

3. **Low Traffic Load Region** (when $\lambda < \lambda_L$), in which the maximal access throughput is $\bar{\lambda}''_{\max} = -Mp_1^* \ln p_1^* < Me^{-1}$, where p_1^* is the unique solution of the following mathematical equation

$$\frac{K}{M} + (1 + \frac{1}{\lambda} p_1^*) \ln p_1^* = 0, \tag{36}$$

the corresponding optimal setting of $(\theta^*, \Delta\theta^*, W^*)$ cannot be adaptively tuned, and has an unique solution of $\theta^* = 1$, $\Delta\theta^* = 0$ and $W^* = 1$, respectively.

5 Simulation Results

In this section, numerical and simulation results are provided to validate the theoretical analyses and demonstrate the performance gain of our proposed DT-MRA protocol. Consider the uplink transmission of a heterogeneous MTC network with $K = 10000$ delay-insensitive terminals and $L = 100$ URLLC terminals coexistence. The QoS requirement of each URLLC terminal is characterized by an E2E latency bound $T_{\rm th} = 1$ ms and the corresponding maximal allowable packet loss probability $\phi_{\rm th} = 10^{-5}$. To support multiple transmission attempts of URLLC, we consider a 5G NR mini-slot structure with unit slot length of $\tau = 0.25$ ms (e.g., each mini-slot contains 7 OFDM symbols, and the subcarrier spacing is $W_0 = 30$ kHz [10]), and thus each URLLC terminal can transmit same packet $Q_{\rm th} = 4$ times within the E2E latency bound. In the connection-oriented resource pool, there are $M = 54$ available preambles and the duration of each RA

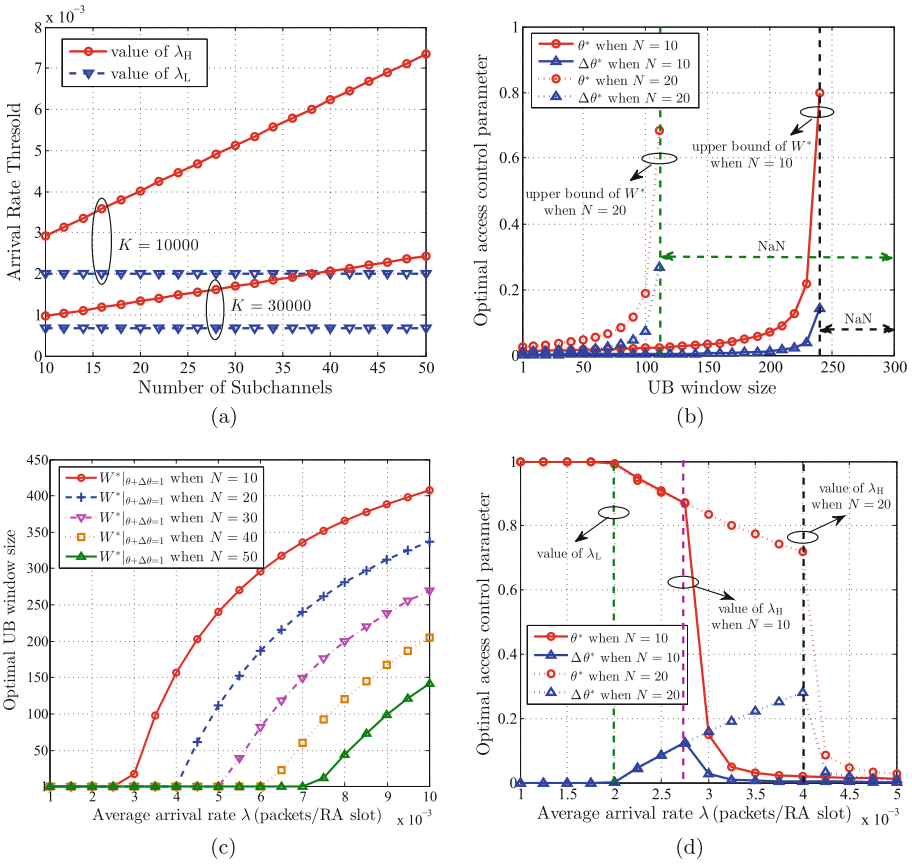

Fig. 4. Illustration of the optimal access control strategy, where the arrival rate of URLLC is $\mu = 10^{-3}$ (packets/slot) and $Q_{\text{th}} = 4$: (a) the value of arrival rate thresholds (λ_H, λ_L) vs. the number of channels in the connection-less resource pool N. (b) the optimal access control parameters $(\theta^*, \Delta\theta^*)$ vs. the UB window size W. (c) the optimal UB window size W^* when $\theta + \Delta\theta = 1$ vs. the average arrival rate of each delay-insensitive terminal λ (packets/RA slot). (d) the optimal access control parameters $(\theta^*, \Delta\theta^*)$ when $W = 1$ vs. the value of λ (packets/RA slot).

slot is $T_{RA} = 5\,\text{ms}$, which contains $I_{RA} = T_{RA}/\tau = 20$ consecutive slots. Moreover, the maximum allowable UB window size is $W_{\max} = 401$. To achieve reliable numerical results, Monte-Carlo simulations are implemented that all statistical results are averaged over 10^7 RA slots.

As depicted in Fig. 3, numerical results are provided to illustrate the overall packet loss probability ϕ of URLLC when $\Delta\theta = 0$, i.e., the connection-less resource pool is only utilized by URLLC terminals for the grant-free data transmission. In this scenario, the value of ϕ is obtained via substituting $\Delta\theta = 0$ into (11). On the one hand, Fig. 3(a) shows how the overall packet loss probability

varies with the total number of URLLC terminals L and the degree of packet repetition Q_{th}. Since the E2E latency bound $T_{\text{th}} = 1\,\text{ms}$ is given, the value of $Q_{\text{th}} = \lfloor T_{\text{th}}/\tau \rfloor$ is determined by the frame structure. For example, $Q_{\text{th}} = 2$ implies that the slot length is set to $\tau = 0.5\,\text{ms}$, and then we can further obtain that the average number of new arrival packets per slot is $\mu = \tau\tilde{\mu} = 2.5 \times 10^{-3}$. From the curves in Fig. 3, we can observe that the overall packet loss probability decreases substantially with the degree of packet repetition increases, which validates the viewpoint that the packet repetition mechanism in our proposed DT-MRA protocol is helpful to guarantee the ultra-high transmission reliability of each URLLC terminal.

Figure 4 illustrates how to design the optimal access control strategy, i.e., find the optimal setting of access control parameters $(\theta^*, \Delta\theta^*, W^*)$, based on the statistical traffic load information (λ, μ). In practical networks, the statistical traffic load information can be obtained via designing traffic prediction algorithms, such as the work in [11]. In this conference paper, we assume that the statistical traffic load information can be perfectly estimated without error. Figure 4(a) shows how the value of arrival rate thresholds (λ_H, λ_L) changes as the number of channels in the connection-less resource pool varies. It is not hard to observe that the value of λ_L keeps constant when the value of N changes, and the value of λ_H increases when the value of N increases, which comply well with the theoretical derivations in Theorem 2. Figure 4(b) illustrates how the optimal setting of access control parameters $(\theta^*, \Delta\theta^*)$ and UB window size W^* in high traffic load region adaptively tuning based on (22) (23) and (24), where $\lambda = 5 \times 10^{-3} > \lambda_H$ and the upper bound of W^* can be obtained via substituting $\theta^* + \Delta\theta^* = 1$ into (22) and (23). As a supplementary to Fig. 4(b), Fig. 4(c) shows how the upper bound of W^* changes with the value of λ varies. By comparing with Fig. 4(a), we can observe that the upper bound of W^* can be larger than 1 only when $\lambda > \lambda_H$, and can only be equal to 1 when $\lambda \leq \lambda_H$, which comply well with the conclusions in Remark 2. Moreover, Fig. 4(d) shows how the optimal access control parameters $(\theta^*, \Delta\theta^*)$ when $W^* = 1$ changes with the value of λ varies, we can clearly observe that $(\theta^*, \Delta\theta^*) = (1, 0)$ when $\lambda \leq \lambda_L$ (i.e., low traffic load region), $\theta^* + \Delta\theta^* = 1$ and $\Delta\theta^* > 0$ when $\lambda_L < \lambda < \lambda_H$ (i.e., medium traffic load region), and $\theta^* + \Delta\theta^* \leq 1$ when $\lambda \geq \lambda_H$ (i.e., high traffic load region). Note that the values of θ^* and $\Delta\theta^*$ when $W^* = 1$ are given by (32) and (33), respectively.

Figure 5(a) illustrates how the average access throughput of delay-insensitive terminals $\bar{\lambda}_{\text{out}}$ varies with number of channels N in the connection-less resource pool, under different value of data arrival rate λ. Specifically, the relationship between $\bar{\lambda}_{\text{out}}$ and N has been discussed in the Theorem 1 and Theorem 2. Based on the observation in the Fig. 5(a), there exists a prefect match between the simulation results and theoretical analysis. Moreover, we can see that the value of $\bar{\lambda}_{\text{out}}$ is not vary with N when $\lambda < \lambda_H$, and linearly increases with N thanks to the impact of traffic offloading. Note that the value of λ_H is also linearly increases with N. As depicted in Fig. 5(b), we demonstrate the effectiveness of our proposed DT-MRA protocol, in terms of improving the access throughput.

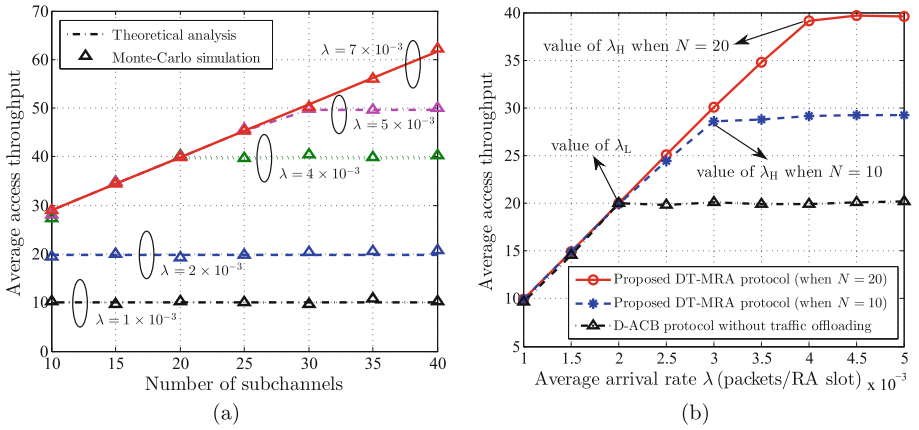

Fig. 5. Monte-Carlo simulation results to validate the theoretical analysis, where the arrival rate of URLLC is $\mu = 10^{-3}$ (packets/slot) and $Q_{\text{th}} = 4$: (a) Average access throughput of delay-insensitive terminals $\bar{\lambda}_{\text{out}}$ vs. the number of channels N. (b) Performance comparison between our proposed DT-MRA protocol and benchmark protocol, in terms of average access throughput $\bar{\lambda}_{\text{out}}$.

To provide comparable numerical results, the D-ACB protocol proposed in [12] without traffic offloading mechanism is specified as a benchmark scheme. On the one hand, we can see that the maximum value of $\bar{\lambda}_{\text{out}}$ is linearly increases with λ when $\lambda < \lambda_H$, and stay constant when $\lambda \geq \lambda_H$. This is due to the value of $\bar{\lambda}_{\max}$ is a constant with λ varies, and the value of $\bar{\lambda}'_{\max}$ and $\bar{\lambda}''_{\max}$ is linearly increases with λ, which has been proved in the Theorem 1 and Theorem 2. On the other hand, when $\lambda > \lambda_L$, our proposed DT-MRA protocol can provide a remarkable performance gain with compare to the benchmark scheme, thanks to the traffic offloading mechanism. Seen from a different perspective, our proposed DT-MRA protocol also has the advantage of low complexity, due to the optimal access control parameters $(\theta^*, \Delta\theta^*, W^*)$ can be determined without the need of real-time load estimation per RA slot.

6 Conclusions

In this paper, we considered a heterogeneous MTC network with massive number of delay-insensitive and URLLC terminals coexistence, and proposed a novel double-threshold-based massive random access (DT-MRA) protocol. On the one hand, the proposed DT-MRA protocol allows partial delay-insensitive terminals temporarily preempting URLLC spectrum channels to alleviate the traffic overload, via adaptively tuning access control parameters including ACB factor and traffic offloading factor. On the other hand, a grant-free access mechanism with packet repetition was applied to support the stringent QoS requirements of URLLC. Then, an optimization problem was formulated to maximize the access throughput of delay-insensitive terminals, while satisfying the QoS requirements

of each URLLC terminal. Based on the statistical traffic load information of each terminal, an optimal access control strategy was also developed to solve this optimization problem. Simulation results validated the theoretical analyses and demonstrated the effectiveness of our proposed DT-MRA protocol, in terms of improving access throughput. In future works, we will further investigate the issue of joint random access and resource allocation strategy design under a practical case that the number of PUSCHs is limited, which is still a open issue in the MTC networks with delay-insensitive and URLLC terminals coexistence.

References

1. Jayawickrama, B.A., He, Y., Dutkiewicz, E., Mueck, M.D.: Scalable spectrum access system for massive machine type communication. IEEE Netw. **32**(3), 154–160 (2018)
2. Li, X., Li, D., Wan, J., Liu, C., Imran, M.: Adaptive transmission optimization in SDN-based industrial internet of things with edge computing. IEEE Internet Things J. **5**(3), 1351–1360 (2018)
3. 3GPP TR 36.881 v1.1.0, Study on latency reduction techniques for LTE, June 2016
4. Zhan, W., Sun, X., Li, Y., Tian, F., Wang, H.: Optimal group paging frequency for machine-to-machine communications in LTE networks with contention resolution. IEEE Internet Things J. **6**(6), 10534–10545 (2019)
5. Lin, Y., Huang, J., Fan, C., Chen, W.: Local authentication and access control scheme in M2M communications with computation offloading. IEEE Internet Things J. **5**(4), 3209–3219 (2018)
6. Singh, B., Tirkkonen, O., Li, Z., Uusitalo, M.A.: Contention-based access for ultra-reliable low latency uplink transmissions. IEEE Wireless Commun. Lett. **7**(2), 182–185 (2018)
7. 3GPP TS 36.321 V12.5.0, Evolved universal terrestrial radio access (E-UTRA); Medium access control (MAC) protocol specification, April 2015
8. Gross, D., Harris, C.M.: Fundamentals of Queueing Theory. Wiley, Hoboken (1998)
9. Sun, C., She, C., Yang, C., Quek, T.Q.S., Li, Y., Vucetic, B.: Optimizing resource Allocation in the short blocklength regime for ultra-reliable and low-latency communications. IEEE Trans. Wireless Commun. **18**(1), 402–415 (2019)
10. Sachs, J., Wikstrom, G., Dudda, T., Baldemair, R., Kittichokechai, K.: 5G radio network design for ultra-reliable low-latency communication. IEEE Netw. **32**(2), 24–31 (2018)
11. Xu, Y., Yin, F., Xu, W., Lin, J., Cui, S.: Wireless traffic prediction with scalable gaussian process: framework, algorithms, and verification. IEEE J. Sel. Areas Commun. **37**(6), 1291–1306 (2019)
12. Duan, S., Shah-Mansouri, V., Wang, Z., Wong, V.W.S.: D-ACB: adaptive congestion control algorithm for bursty M2M traffic in LTE networks. IEEE Trans. Veh. Technol. **65**(12), 9847–9861 (2016)