Delay-Aware Dynamic Access Control for mMTC in Wireless Networks Using Deep Reinforcement Learning

Diego Pacheco-Paramo[a,b,*], Luis Tello-Oquendo[c]

[a] *reconoSER ID, Bogotá, Colombia*
[b] *Escuela de Ciencias Exactas e Ingeniería, Universidad Sergio Arboleda, Bogotá 111221, Colombia*
[c] *College of Engineering, Universidad Nacional de Chimborazo, Riobamba 060108, Ecuador*

Abstract

The success of the applications based on the Internet of Things (IoT) relies heavily on the ability to process large amounts of data with different Quality-of-Service (QoS) requirements. Access control remains an important issue in scenarios where massive Machine-Type Communications (mMTC) prevail, and as a consequence, several mechanisms such as Access Class Barring (ACB) have been designed aiming at reducing congestion. Although this mechanism can effectively increase the total number of User Equipments (UEs) that can access the system, it can also have an adverse effect on the delay, which may limit its usability in some scenarios. In this work, we propose a delay-aware double deep reinforcement learning mechanism that can dynamically adapt two parameters of the system in order to enhance the probability of successful access using ACB, while at the same time reducing the expected delay by modifying the Random Access Opportunity (RAO) periodicity. Results show that our system is able to accept a simultaneously massive number of machine-type and human-type UEs while at the same time reducing the mean delay when compared to previously known solutions. This mechanism can work adequately under varying load conditions and can be trained with real data traces, which facilitates its implementation in real scenarios.

Keywords: Delay, Double Deep Q-Learning, Access Class Barring, massive machine type communications

*Corresponding author
Email address: diego.pacheco@reconoserid.com (Diego Pacheco-Paramo)

1. Introduction

The Internet of Things (IoT) relies on the ability to gather, process, and analyze massive amounts of data in order to make the appropriate decisions at the right time. Computing systems have addressed this requirement through the development of robust architectures that support real-time analysis of massive data. From a radio access point of view, several technologies have been proposed which aim at providing extensive coverage, low power, and typically, low data rate transmissions. However, the characteristics of machine-type communications (MTC) [1] impose different demands than those that existed for human-to-human (H2H) communications, causing that current systems do not behave appropriately under these new conditions. Among the possible wireless networks that compete in the IoT market, cellular networks are one of the main options due in part to its extended coverage, existing infrastructure, and standardization efforts. However, it is still necessary to adapt some mechanisms to new traffic types such as MTC.
In cellular networks, high traffic loads can be controlled through the Access Class Barring (ACB) mechanism, which defines a barring rate that is broadcasted by the base station, delaying the access of a percentage of the active users (named UE herein). This mechanism can successfully control high loads of simultaneous UEs trying to access the medium, which results in fewer collisions and, therefore, in the long term, a higher amount of UEs accessing the network. However, this mechanism also produces a higher delay, which could be undesirable for specific applications. Therefore, ACB must be combined with other mechanisms that allow reducing its impact on delay. This can be done by varying the periodicity of random access opportunities (RAOs), which define the available slots where the devices can contend for network access. We propose a Delay-Aware Double Deep Q-learning mechanism that can adapt both the barring rate of ACB, in order to increase the successful access probability of UEs even under very high load scenarios, and also the RAO periodicity to reduce the impact on delay. These parameters are adapted dynamically, which allows the system to work successfully under different traffic conditions. Since reinforcement learning (RL) is a method that relies on data, we use traces obtained from a Telco to represent H2H traffic. Also, we rely on the fact that both the barring rate and the RAO periodicity can be modified, that is, our solution complies with the standards. The contributions of this paper can be summarized as follows:

- We model the access control problem as a Partially Observable Markov Decision Process (POMDP) and design an adaptive access control mechanism based on double deep RL that modifies the barring rate of ACB and the RAO periodicity simultaneously.
- We evaluate the system under simultaneous H2H and machine-to-machine (M2M) traffic, where the former is obtained from traces of a Telco, and the latter is modeled after the standards for heavy load scenarios.
- We compare our Delay-Aware Double Deep Q-Learning mechanism with two previous solutions, and show that our mechanism is able to reduce the mean access delay while maintaining full successful access probability.

The rest of the paper is organized as follows. In Section 2, we review the different solutions that have been proposed for handling the access control problem with mMTC. In Section 3, we detail the random access procedure and ACB scheme following the 3GPP standards closely. Then, in Section 4, we model the access control problem as a POMDP and describe our Delay-Aware Double Deep Q-Learning scheme. In Section 5, we evaluate our proposed scheme under different traffic conditions and compare it against two other solutions that are also dynamic through different key performance indicators (KPIs). Finally, Section 6 concludes the paper.

2. Related Work

There have been numerous research efforts devoted to optimizing the ACB barring factor for handling massive MTC (mMTC) connection attempts on the RACH through either static or dynamic approaches [2, 3, 4, 5, 6]. However, some studies [7, 8, 9] offer complex procedures, use questionable assumptions for getting high performance, or do not conform with network specifications (e.g., without considering the number of uplink grants or the updating period of notification information by the base station).

Duan et al. [10] presented an ACB scheme that calculates the optimal barring rate at each RAO based on the estimation of the number of contending UEs using preamble information (i.e., successful, unused). They also provide a scheme to dynamically select the number of available preambles allocated to MTC devices. The performance of the ACB schemes mentioned above is typically compared with that of idealized solutions that exploit the advantages of having full state information [5, 11, 10]. These full state information

solutions are impractical but provide an upper bound to the performance of the ACB scheme. In this paper, we use as a benchmark the idealized and full state information scheme presented by Duan et al. [10].

RL-based ACB schemes are suitable approaches to optimize the access control for wireless networks, and in particular, for cellular networks such as LTE-A and NB-IoT. El-Hameed and Elsayed [12] propose a Q-Learning mechanism that aims to assign preambles to H2H or M2M UEs according to the traffic intensity. However, their scheme requires that the system knows how much traffic per UE type is offered to assign access priorities, similarly as Extended Access Barring. Bello et al. [13] propose a Q-Learning approach in which each M2M UE has to learn when to transmit. This mechanism does not use ACB, and it is entirely decentralized. Bear in mind that decentralized access schemes do not conform to current LTE-A recommendations. Likewise, Yu et al. [14] propose a decentralized mechanism to optimize access control. In this case, the UEs can learn the features of different coexisting medium access control mechanisms, and adjust their transmissions using cognitive radio. Although this scenario is very promising, it heavily relies on the processing capacities of the terminals, which is not feasible in scenarios with low-power, low-processing devices such as those frequently found in IoT applications. A deep RL-based ACB scheme was proposed in [15]; it considered differentiated MTC services so that the high priority MTC UEs could transmit their data in a short time. Moon et al. [16] used Q-learning to adjust the barring factor by observing the access success rate. However, in these proposals is assumed that the base station has complete knowledge of the number of UEs contending for resources in the network, which is impractical.

In a previous work [17], we proposed a Double Deep Q-Learning (DDQL) solution that was able to adapt to dynamic traffic conditions through a single parameter, but suffered from high delays. In this work, we present a scheme that improves the access delay while maintaining the successful access rate concerning our previous proposal. Also, this mechanism can work properly only with the available information at the base station and can be integrated into current cellular systems.

3. Random Access Procedure and Access Control

The random access (RA) procedure is performed every time that a UE wants to switch from idle to connected mode. The UEs first acquire the network configuration parameters; then, they are subjected to the ACB scheme; finally, after passing the barring check, they perform the RA procedure.

The random access channel (RACH) is used to signal the connection request; it is allowed in predefined time/frequency resources, hereafter RAOs [18, 19]. The base station has a number of preambles (r) available for initial access to the network; these preambles [19, 20] are transmitted by the UEs for attempting the first access to the network. The *Master Information Block* (MIB) and the *System Information Blocks* (SIBs) are resources used by the base station to broadcast the configuration information periodically. In particular, the SIB2 includes some basic parameters, such as the periodicity of RAOs and the barring parameters.

Upon arrival, the UEs are subjected to the ACB scheme. The main goal of ACB is to redistribute the access requests of UEs through time; by doing so, the number of access requests per RAO is reduced. This fact helps to evade massive-synchronized accesses demands to the RACH, which might endanger the fulfillment of QoS objectives. UEs subjected to the ACB scheme must perform a barring check before initiating the RA procedure (i.e., before the transmission of their first preamble) as described in Algorithm 1 [21, 22].

Algorithm 1: ACB Scheme

repeat
  Set the mean barring time $T_{\text{ACB}}(n)$ and the barring rate $P_{\text{ACB}}(n)$ broadcast by the base station in the nth SIB2;
  Generate $\mathcal{U}[0,1)$ = a random number with uniform distribution between 0 and 1;
  if $\mathcal{U}[0,1) \leq P_{ACB}(n)$ **then**
    initiate the random access procedure;
  else
    Generate a new $\mathcal{U}[0,1)$;
    Set the barring time as

$$T_{\text{barring}} = [0.7 + 0.6\mathcal{U}[0,1)]\ T_{\text{ACB}}(n); \tag{1}$$

    wait for T_{barring};
  end
until *the random access procedure is initiated*;

UEs that succeed in the barring check are no longer subject to the ACB scheme and proceed to perform the RA procedure as follows.

A four-message handshake is performed in the contention-based random

access. In *Msg1*, a UE transmits a randomly chosen preamble from the preamble pool during one of the available RAOs. A preamble will be detected at the base station if it has not been chosen by more than one UE in the same RAO. Otherwise, a collision occurs. Then, the base station sends a random access response message, *Msg2*, which includes one uplink grant for each detected preamble. *Msg2* is used to assign time-frequency resources to the UEs for the transmission of *Msg3*. UEs wait for a predefined time window to receive the uplink grant. If the end of this window receives no uplink grant and the maximum number of access attempts has not been reached, the UEs wait for a random time and then perform a new access attempt. That is, they select a new preamble and transmit it at the next RAO. The UEs that receive an uplink grant send their connection request message, *Msg3*, using the resources specified by the base station. Finally, the base station responds to each *Msg3* transmission with a contention resolution message, *Msg4*. The interested reader is referred to [18, 21, 23, 24, 25] for further details.

4. Delay-Aware Double Deep Q-Learning Mechanism

In this section, we model the access control system of the base station as a POMDP, with its associated state space and actions. Then, we describe a Delay-Aware DDQL solution that can to modify two parameters that aim at increasing the successful access probability while reducing the mean access delay.

4.1. System Model

The system is composed of a single base station that provides access to both M2M and H2H UEs distributed under its coverage area. However, the base station does not differentiate between the two types of UEs. The base station can know the number of contending UEs by the number of preambles successfully received N_{psu} on each RAO. However, the number of received preambles might differ from the number of sent preambles due to collisions or transmission errors. Collisions occur when two or more UEs transmit the same preamble in the same RAO. In order to reduce collisions, our system dynamically adapts two parameters: P_{ACB} and T_{RAO}. There are 16 possible values for P_{ACB}=$\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1\}$ and six possible values for T_{RAO}=$\{1, 2, 3, 5, 10, 20\}$, where the latter are measured in ms. These values are set by the parameter *prach-ConfigIndex*[19] [21]. Let us remember that according to the specifications [21], the SIB2 messages are sent every

$80ms$, and therefore when T_{RAO}=5, there are 16 RAOs between two SIB2 messages, and when T_{RAO}=20, there are only four RAOs between two SIB2 messages, as can be seen in Fig. 1. Hence our action set is $\mathcal{A} = \{1, 2, 3, .., 96\}$, where each individual action a represents a combination of P_{ACB} and T_{RAO} values.

We have defined a state s as a set of variables that account for an accumulation of observations that occur between two SIB2 messages, since it is only through these messages that the base station can communicate with the UEs, and therefore it is only through these messages that changes can be exerted on their behaviour. Hence, the state s is defined as $s = \big(\overline{N_{psu}}, CV_{N_{psu}}, \Delta N_{psu}, P_{\text{ACB}}, T_{\text{RAO}}\big)$, where $\overline{N_{psu}}$ is the mean number of successfully received preambles among the RAOs available between two SIB2 messages. This means that if T_{RAO}=1, the mean will be calculated between 80 observations, and if T_{RAO}=20, the mean will be calculated between four observations. These values are discretized so $\overline{N_{psu}} \in \mathbb{N}$. However, since there are only $r = 54$ available preambles, $\overline{N_{psu}} \leq 54$. The value $CV_{N_{psu}}$ is the coefficient of variation of the number of successfully received preambles among the RAOs available between two SIB2 messages. These values are discretized so, $CV_{N_{psu}} \in \{0, 0.2, 0.4, 0.6, 0.8\}$. The value ΔN_{psu} represents the difference of $\overline{N_{psu}}$ between the current and the previous observation. This value only represents three variations: If the difference is positive, that is if the mean traffic is increasing, ΔN_{psu}=1. If the traffic is decreasing, ΔN_{psu}=2. If the mean traffic remains constant, ΔN_{psu}=3. The possible values for P_{ACB} and T_{RAO} were explained earlier, and refer to the values that were set by the base station in this period and that impact UEs during the whole set or RAOs. Therefore, the state space $\mathcal{S}$ is composed of 79200 states.

4.2. Double Deep Q-Learning (DDQL) Implementation

DDQL [26], aims at optimizing an objective function while representing the action values $Q(s, a)$ through a neural network [27]. Therefore, this solution replaces the Q tables that were used in traditional Q-Learning [28] with a neural network, which allows representing continuous state spaces or action sets, but more importantly, it can assign action values to previously unvisited states. In DDQL, the system aims to minimize a loss function defined by:

$$L = \frac{1}{2}\big(\mathcal{R} + \gamma \max_{a' \in \mathcal{A}} \big[Q(s', a', \theta^-)\big] - Q(s, a, \theta)\big)^2, \tag{2}$$

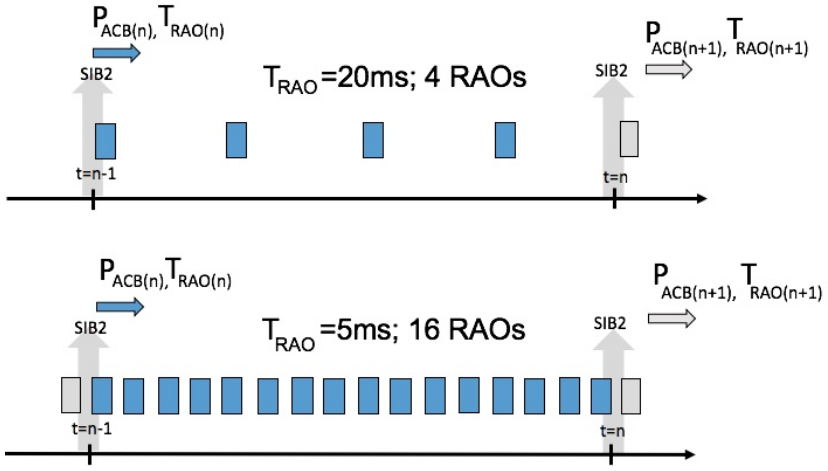

Figure 1: Random Access Oportunities for different T_{RAO}.

where $\mathcal{R}$ is the reward associated to the action a taken in state s, γ is the discount factor, which controls the impact of future rewards, $Q(s', a', \theta^-)$ represents the Q value for a given action a', state s' and neural network defined by the set θ^-, and $Q(s, a, \theta)$ represents the Q value for a given action a, taken in state s and a neural network defined by the set θ. The sets θ^- and θ represent the weights that define each neural network. DDQL uses two different neural networks aiming at reducing the inherent overestimations that result from calculating the future rewards of taking action a on a state s derived from the max operator seen in (2). This overestimation may result in finding a local optimum, or in divergence [26]. In DDQL, one neural network (θ^-) is used to evaluate the current policy, while the other is used to obtain the action that maximizes future rewards (θ). Therefore, the target for DDQL is represented by:

$$Y^{DDQL} = \mathcal{R} + \gamma\, Q(s', \max_{a' \in \mathcal{A}} [Q(s', a, \theta)], \theta'), \tag{3}$$

where the roles of each neural network are interchanged every τ iterations; also, experience replay is implemented [29]. The purpose of this technique is to break the dependency of consecutive experiences of the system with the environment by saving them on a buffer (of size ER) and then sampling them randomly for training. In our implementation, each neural network consists of a Multi-layer Perceptron with an input layer with five neurons

(one for each variable of the state space), N_L hidden layers of 5 neurons, and an output layer of 96 neurons (one for each action). The algorithm that describes the DDQL mechanism is shown in Algorithm 2.

Algorithm 2: Delay- Aware Double Deep QL Mechanism.

Controller: Double Deep Q-learning($\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma, \epsilon, \tau, ER, N_L$)
Input : $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{R}$ is the reward, γ is the discount factor, ϵ is the exploration probability, ER is the size of the experience replay, τ is the size of the target update and N_L is the number of hidden layers of the neural network
Local : Neural Network $\mathbf{Q}_A[s, a, \theta]$, state s, action a, weights θ
Local : Neural Network $\mathbf{Q}_B[s, a, \theta']$, state s, action a, weights θ'

```
repeat
    repeat
        if T_SIB2 then
            select P_acb and T_RAO with a from A using the ε−greedy approach
              and Q_A[s, a, θ];
            observe reward R(s, a', s') and state s';
            WRITE a,s,r,s' in buffer
        end
        s = s'
    until j = ER;
    Randomize buffer order
    k=0;
    repeat
        m=kτ+1;
        repeat
            target= r+γ Q_B[s', max_{a∈A}{Q_A[s', a, θ]}, θ'];
            WRITE target in buffer;
        until m = (k+1)τ;
        Q_B[s, a, θ']=Q_A[s, a, θ];
        Read buffer and train Q_A[s, a, θ] to reach target;
    until k = ER/τ;
until i = max RAO;
```

As mentioned earlier, one of the main characteristics of this system is that the base station is not able to know precisely how many contending UEs are at any given RAO. The base station must make its decisions based on the number of successfully received preambles, which might differ from the actual contending UEs. Given a number of UEs transmitting a preamble on a given RAO N_{pt}, and assuming that every preamble reaches the base station, the probability mass function of the number of successfully received preambles N_{psu} at the base station is given by [25]:

$$P_n(s) \triangleq \Pr(s \mid n) = \sum_{c=0}^{c_{\max}} P_n(s,c) \tag{4}$$

being

$$P_n(s,c) = \frac{r-(s-1+c)}{r} P_{n-1}(s-1,c) + \frac{s+1}{r} P_{n-1}(s+1,c-1) + \frac{c}{r} P_{n-1}(s,c), \tag{5}$$

where $P_0(0,0) = 1$, n is the number of contending UEs in a RAO (i.e., the UEs that transmit a preamble selected among the r available preambles with equal probability), s is the number of preambles selected by exactly one UE, c is the number of collided preambles, and $c_{\max} = \min\{r, \lfloor n/2 \rfloor\}$.

Following (4), we can obtain the number of received preambles with the highest probability for a given number of sent preambles, as shown in Fig. 2. It can be seen that while the number of sent preambles in a RAO is lower or equal to 10, it is more likely that the base station successfully receives as many preambles as were sent. However, when there are more than 10 preambles sent in a RAO, it is more likely to successfully receive less preambles than those that were sent. Collisions cause this; that is, we are not considering other causes such as transmission errors. Therefore, if we want to reduce the uncertainty on the base station associated with collisions, the number of UEs that transmit should be equal or lower than 10. Although it is not possible to guarantee this through ACB, our reward function promotes this behavior by penalizing the system every time that $\overline{N_{psu}} > 10$, that is, when it is evident that more than 10 UEs are contending for access.

The reward function $\mathcal{R}$ is described in Table 1. For a better understanding of the criteria used to define the reward function, it is necessary to define ranges for P_{ACB}, T_{RAO} and $\overline{N_{psu}}$. In the case of P_{ACB}, we have defined five ranges: *very low* when $P_{\text{ACB}} < 0.3$, *low* when $0.3 \leq P_{\text{ACB}} < 0.5$, *medium* when $0.5 \leq P_{\text{ACB}} < 0.7$, *high* when $0.7 \leq P_{\text{ACB}} < 1$ and *very high* when $P_{\text{ACB}} = 1$. In the case of T_{RAO} we have defined three ranges: *low* when $T_{\text{RAO}} < 3$, *medium* when $3 \leq T_{\text{RAO}} < 10$ and *high* when $T_{\text{RAO}} >= 10$. Finally, for $\overline{N_{psu}}$ we have defined four ranges: *low* when $\overline{N_{psu}} \leq 3$, *medium* when $3 < \overline{N_{psu}} < 7$, *high* when $7 \leq \overline{N_{psu}} \leq 10$ and *very high* when $N_{psu} > 10$. The first observation that must be done about $\mathcal{R}$ is that those action-state combinations that do not appear in Table 1 have reward $r = 0$. On the

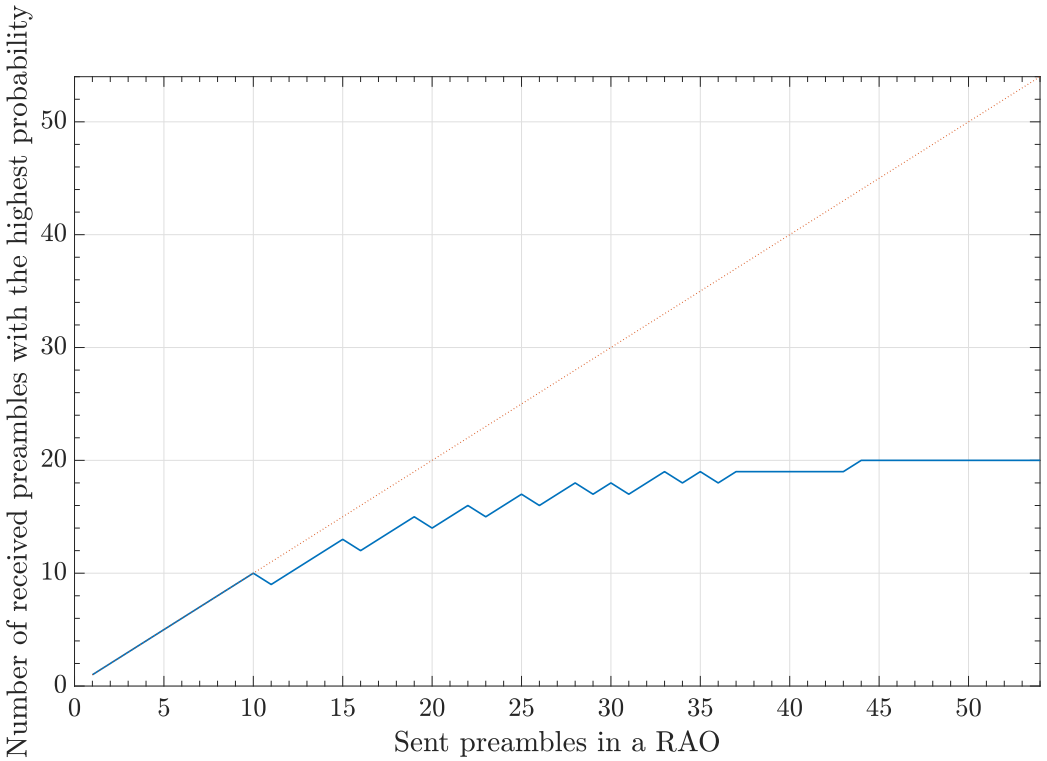

Figure 2: Number of received preambles with highest probability.

other hand, all the action-states combinations where T_{RAO} is *low* or *high* are avoided, and have a reward $r = -100$. In the former case, this is done to avoid having too many RAOs, which might result in a waste of resources in the base station. In the latter case, this is done to avoid increasing the mean delay. A *very high* value of P_{ACB} is desirable when there is *low* traffic in order to accept UEs with a very low delay. Therefore, we reward this action when $\overline{N_{psu}}$ is *low*. The amount of the reward depends on $CV_{N_{psu}}$ and ΔN_{psu}. The reward grows when there is a small variation and when the traffic does not grow. When $\overline{N_{psu}}$ is *medium* we use the same criteria, but in this case the rewards are smaller. If $\overline{N_{psu}}$ is *high*, then we only give a reward when there is very little variation and the traffic is decreasing. On the other hand, it is not convenient to accept all UEs when $\overline{N_{psu}}$ is *very high*, and therefore we assign negative values for r, based on $CV_{N_{psu}}$ and ΔN_{psu}. Since we want that the system adapts to dynamic traffic conditions, we have to set rewards when $P_{\text{ACB}} < 1$. The objective of reducing P_{ACB} is to maintain $\overline{N_{psu}}$ below

10, according to what was explained earlier. Therefore, we consider that the *medium* range of P_{ACB} is adequate for that purpose. Since P_{ACB} should not be lower than zero when $\overline{N_{psu}}$ is *low*, the reward remains zero in those cases. As $\overline{N_{psu}}$ grows to *medium*, we assign rewards depending on $CV_{N_{psu}}$ and ΔN_{psu}. The reward is higher if $CV_{N_{psu}}$ is low, and grows as ΔN_{psu} shrinks. If $\overline{N_{psu}}$ grows to *high*, the same criteria is used, although the rewards are lower. On the other hand, if $\overline{N_{psu}}$ is *very high*, then the assignment of P_{ACB} has not been successful, and therefore we try to avoid those states.

5. Experiments and Results

We consider a single base station which provides coverage for UEs of M2M and H2H communications simultaneously. We consider that M2M traffic follows the specifications [30], where it is stated that in high load scenarios, M2M traffic can show a bursty behavior associated to the simultaneous activation of, e.g., alarms, or event-driven wireless sensor networks. This behavior is represented by a Beta distribution (3,4) for 10 seconds. The intensity for a high load scenario is set in 30000 M2M UEs. In the case of H2H traffic, we use traces from the Telco Telecom Italia, which provided the data as part of a "Big Data Challenge" in 2014. Since this data is aggregated for periods of 10 minutes, we consider that H2H traffic has a constant intensity during this period. Also, because this data does not provide units, we use the observation provided in [31], where it is stated that the maximum load that can be provided by a base station is 55 EPS Radio Access Bearer setups per second. Based on this value, we normalize the H2H traffic from the traces. Unless otherwise stated, the values of the system are as appear in Table 2. In this work we evaluate the performance of the system through different KPIs: Mean Delay $E[M]$, Mean Number of preamble transmissions $E[k]$ and Probability of successful access P_s. These KPIs can be obtained individually for each type of UE, and have been evaluated in previous studies.
It is essential to make a distinction between the training and evaluation of the system. For training our Delay-Aware DDQL solution, we use 144 episodes, each one associated with the data obtained from the traces for H2H UEs for one day, which are aggregated in periods of 10 minutes. Therefore, on each of these episodes, the H2H traffic is constant, and the traces give its intensity. We have picked the trace obtained on November 1 from the most active cell on that day. In the case of M2M traffic, we use on each episode a beta distribution (3,4) with 30000 UEs. The parameter ϵ is set to 1 at the beginning of each episode, and then it is reduced linearly to 0. On the other hand, in the

Table 1: Reward Function $\mathcal{R}$

P_{ACB}	T_{RAO}	$\overline{N_{psu}}$	$CV_{N_{psu}}$	ΔN_{psu}	r
–	Low	–	–	–	-100
–	High	–	–	–	-100
Very High	Medium	Low	< 0.4	1	20
Very High	Medium	Low	< 0.4	2,3	80
Very High	Medium	Low	$>= 0.4$	1	20
Very High	Medium	Low	$>= 0.4$	2,3	40
Very High	Medium	Medium	< 0.4	1	20
Very High	Medium	Medium	< 0.4	2,3	60
Very High	Medium	Medium	$>= 0.4$	2,3	20
Very High	Medium	High	< 0.2	2	20
Very High	Medium	Very High	< 0.2	1	-100
Very High	Medium	Very High	< 0.2	2,3	-80
Very High	Medium	Very High	$>= 0.2$	1	-100
Very High	Medium	Very High	$>= 0.2$	2,3	-80
Medium	Medium	Medium	< 0.4	1	40
Medium	Medium	Medium	< 0.4	2	80
Medium	Medium	Medium	< 0.4	3	60
Medium	Medium	Medium	$>= 0.4$	1	40
Medium	Medium	Medium	$>= 0.4$	2,3	60
Medium	Medium	High	< 0.2	2	40
Medium	Medium	High	< 0.2	3	20
Medium	Medium	High	$>= 0.2$	2	20
Medium	Medium	Very High	< 0.2	1	-60
Medium	Medium	Very High	< 0.2	2,3	-40
Medium	Medium	Very High	$>= 0.2$	1	-60
Medium	Medium	Very High	$>= 0.2$	2,3	-40

Table 2: Default System Configuration for Evaluation Purposes

Parameter	**Setting**
PRACH Configuration Index	*prach-ConfigIndex* = 6
Periodicity of RAOs	5 ms
Subframe length	1 ms
Available preambles for contention-based random access	$r = 54$
Maximum number of preamble transmissions	*preambleTransMax* = 10
RAR window size	$W_{\text{RAR}} = 5$ subframes
Maximum number of uplink grants per subframe	$N_{\text{RAR}} = 3$
Maximum number of uplink grants per RAR window	$N_{UL} = W_{\text{RAR}} \times N_{\text{RAR}} = 15$
Preamble detection probability for the kth preamble transmission	$P_d = 1 - \frac{1}{e^k}$ [30]
Backoff Indicator	$BI = 20$ ms
Re-transmission probability for *Msg3* and *Msg4*	0.1
Maximum number of *Msg3* and *Msg4* transmissions	5
Preamble processing delay	2 subframes
Uplink grant processing delay	5 subframes
Connection request processing delay	4 subframes
Round-trip time (RTT) of *Msg3*	8 subframes
RTT of *Msg4*	5 subframes
Discount factor	$\gamma = 0.7$
Num. hidden layers (feedforward neural network)	$N_L = 10$
Update period (events) of the second neural network	$\tau = 100$
Buffer size for experience replay	$ER = 500$

evaluation phase, we use a single episode where H2H traffic is constant and has the maximum intensity, which is 55 user arrivals per second. For M2M traffic, we also use a beta distribution (3,4), although the intensity changes according to the experiment. Please notice that although training is done with 30000 M2M UEs, the experiments can be done with different values: 1000, 10000, 20000, 30000 and 40000 M2M UEs. For the evaluation of the Delay-Aware DDQL solution, the parameter ϵ is set to 0, that is, we use the best policy. In every result shown, we perform 100 independent experiments.

In Figs. 3 and 4 the variation of P_{ACB} and T_{RAO} can be seen for our

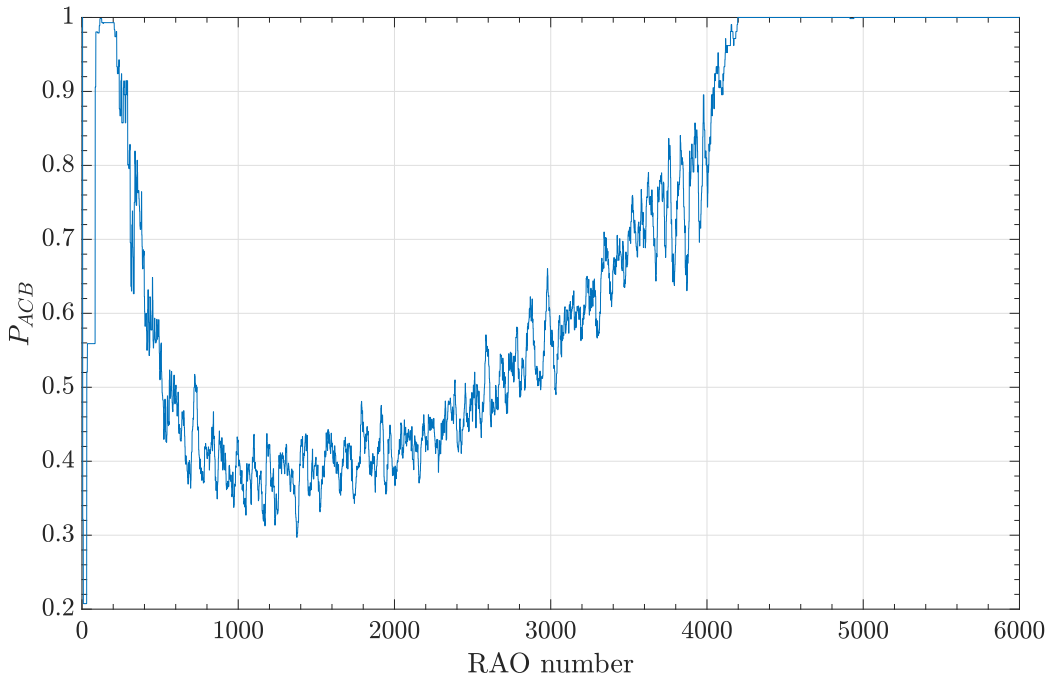

Figure 3: P_{ACB} variation for Delay Aware DDQL mechanism with 30000 M2M users.

Delay-Aware DDQL mechanism when there are 30000 M2M UEs and the Backoff Indicator is set to 120ms. It can be seen that at the beginning of the simulation, as M2M UEs begin to transmit, P_{ACB} suffers a significant reduction that sets its value close to 0, and then again to 1. This behavior occurs because when there are no UEs, T_{RAO} is 20, and when suddenly UEs begin to transmit, the system rapidly reacts to adapt to this new traffic conditions. After this, the system maintains T_{RAO} between 3 and 5. It should be noted that part of the dynamic adaptation of P_{ACB} and T_{RAO} consists of simultaneously reducing T_{RAO} with P_{ACB} in order to reduce collisions.

In Fig. 5, the performance of the Delay-Aware DDQL mechanism can be observed when there are 30000 M2M contending UEs. It can be seen that the system maintains the number of successful accesses below 10, according to the definition of our reward function. In fact, it is kept most of the time between 7 and 8, which is sufficient to keep the collision number below 6. Let us recall that the system can control only the number of UEs contending for the first time, and therefore we can see that the total number of transmitted preambles reaches almost 18.

In Table 3, we compare three solutions for the primary three KPIs. The

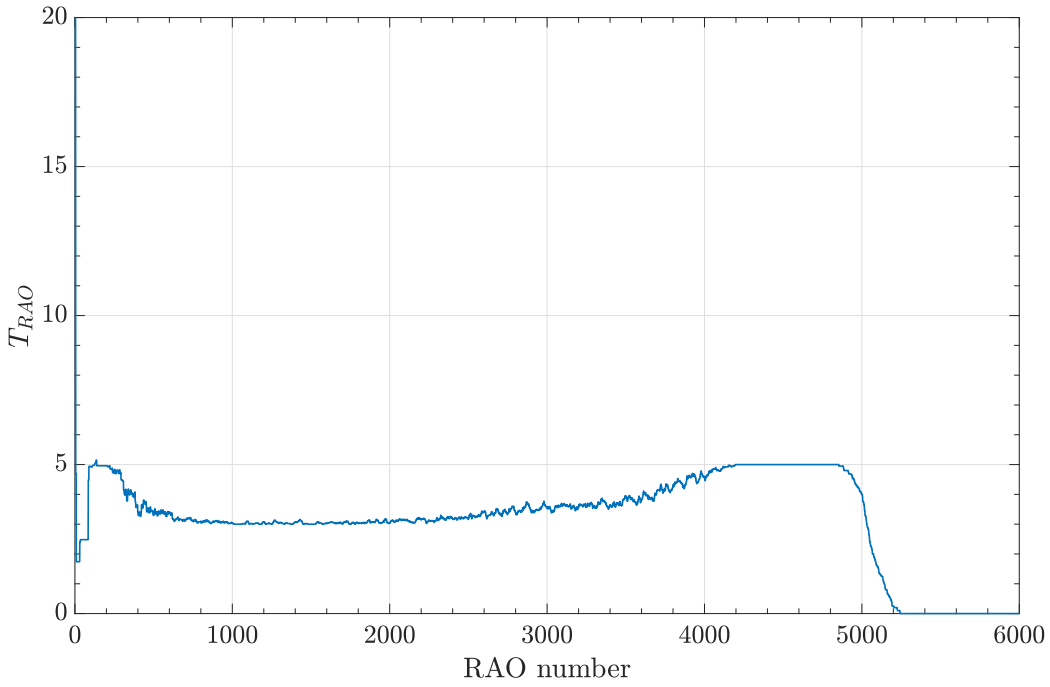

Figure 4: T_{RAO} variation for Delay Aware DDQL mechanism with 30000 M2M users.

first solution is our previously presented DDQL mechanism [17] that only adapts P_{ACB}. It can be seen that it shows the worst performance in terms of access delay, although it uses the lowest value of Backoff Indicator (BI=20 ms). The second solution is our proposed Delay-Aware DDQL mechanism, evaluated with a BI=120 ms. It can be seen that the total delay is considerably reduced (more than 2s) due to the variation of T_{RAO}. The last solution was the dynamic resource allocation (DRA) proposed by Duan et al. [10], and it also dynamically modifies P_{ACB} and the number of available preambles for MTC UEs. In order to compare this algorithm accurately, we increased the number of maximum retries in the system from 10 to 150. By doing this, and increasing the BI value to 960 ms, it is possible to reach a 100% probability of successful access. It can be seen that the mean delay is close to the one obtained by our presented solution, although the mean number of transmissions is a lot higher, which is detrimental for the energy consumption of MTC devices. The proposed Delay-Aware mechanism can maintain the mean number of transmissions below 2, although it is moderately higher than our previous DDQL mechanism. This is expected since, for the DDQL and Delay-Aware schemes, the maximum number of retries fol-

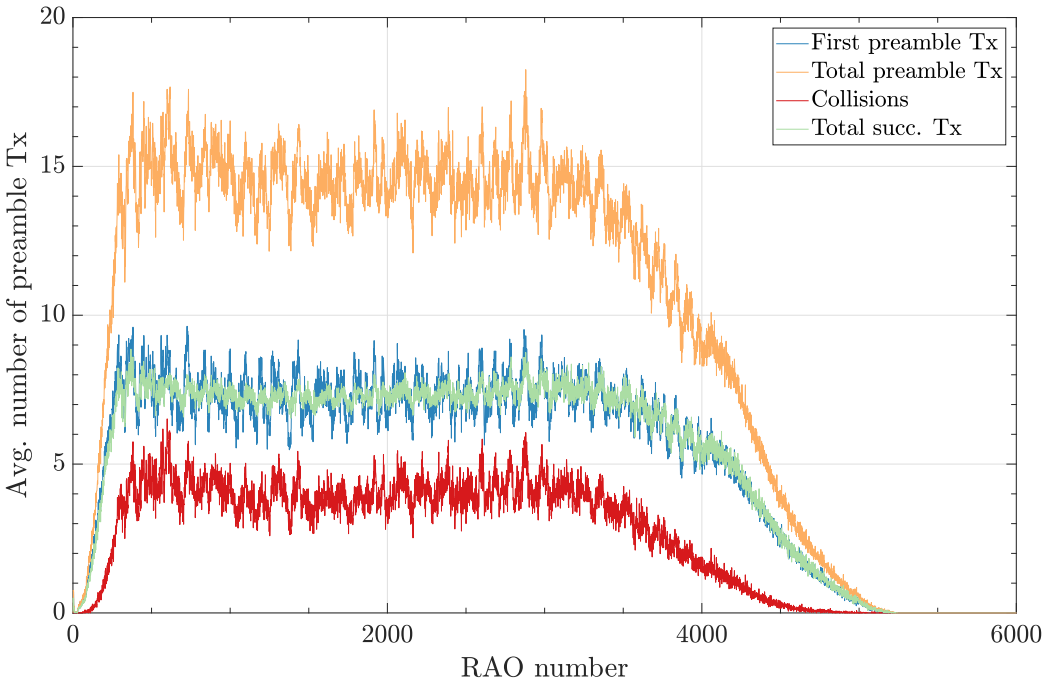

Figure 5: Preamble transmissions for Delay-Aware DDQL mechanism with 30000 M2M users.

lows the standard (i.e., 10 preamble retransmissions), and therefore a mean value as the one shown by [10] would result in a meager value of P_s.

Table 3: Performance of Access Control Mechanisms for 30 000 M2M UEs

KPI	*DDQL*	*Delay Aware DDQL*	*Duan et al. [10]*
$E(D)_{Total}$ (s)	5.87	3.43	3.66
$E(D)_{M2M}$ (s)	5.96	3.47	3.69
$E(D)_{H2H}$ (s)	2.98	1.65	2.18
$E(K)_{Total}$	1.77	1.89	8.38
$E(K)_{M2M}$	1.77	1.90	8.45
$E(K)_{H2H}$	1.63	1.72	5.36
$P_{sa-Total}$ (%)	100	100	100
P_{sa-M2M} (%)	100	100	100
P_{sa-H2H} (%)	100	100	100

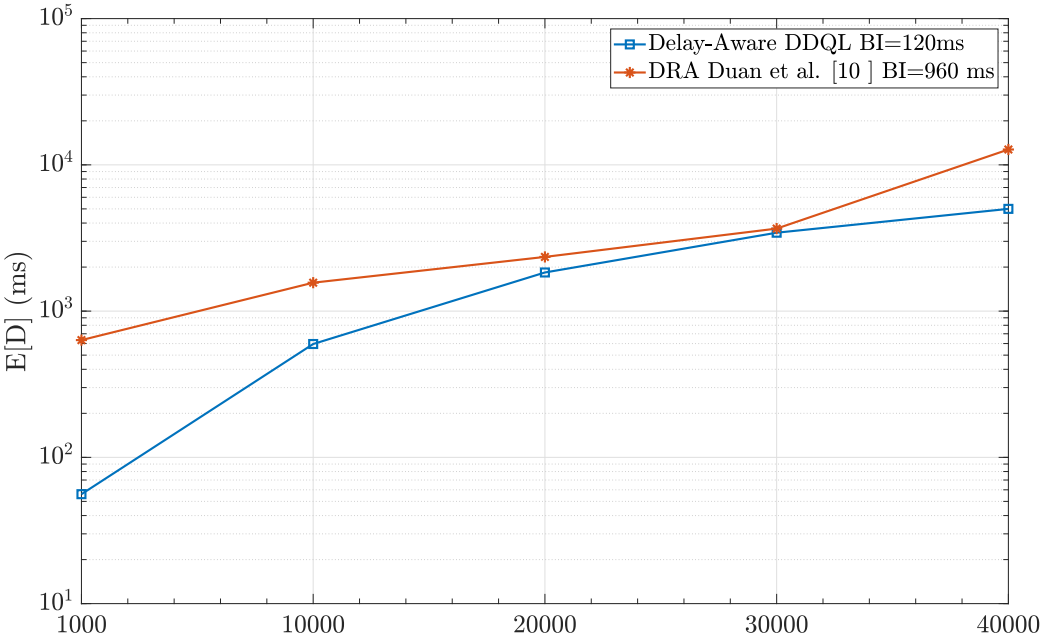

Figure 6: Mean Delay for Delay Aware DDQL mechanism and Duan et al. [10] for various M2M loads.

In Fig. 6, we compare our proposed solution against the DRA solution proposed in [10] as the M2M load varies from 1000 to 40000 UEs. For the Delay-Aware DDQL mechanism, we use a BI=120 ms and maintain the maximum number of retries is 10, as it is proposed in the 3GPP standards. For the DRA solution, we use a BI=960 ms and increase the maximum number of retries to 150. It can be seen that our proposed solution can maintain a lower delay than the DRA mechanism for every value of M2M UEs. Also, it can be seen that when there are only 1000 M2M UEs, the system can adapt in order to reduce the mean access delay to 56 ms. This value is increased to 5 s when there are 40000 M2M UEs, less than the mean access delay for our previous DDQL solution with 30000 UEs. The ability of the system to adapt to different loads shows that it is not necessary to retrain the system as the traffic changes.

In Fig.7, we compare the P_s for all UEs for the two previously mentioned mechanisms as the M2M load varies from 1000 to 40000 UEs. It can be seen that the DRA mechanism can maintain $P_s = 100\%$. This occurs because we have increased the maximum number of retries to 150, which is 15 times more than what is proposed in the standards. If we reduce this value, the

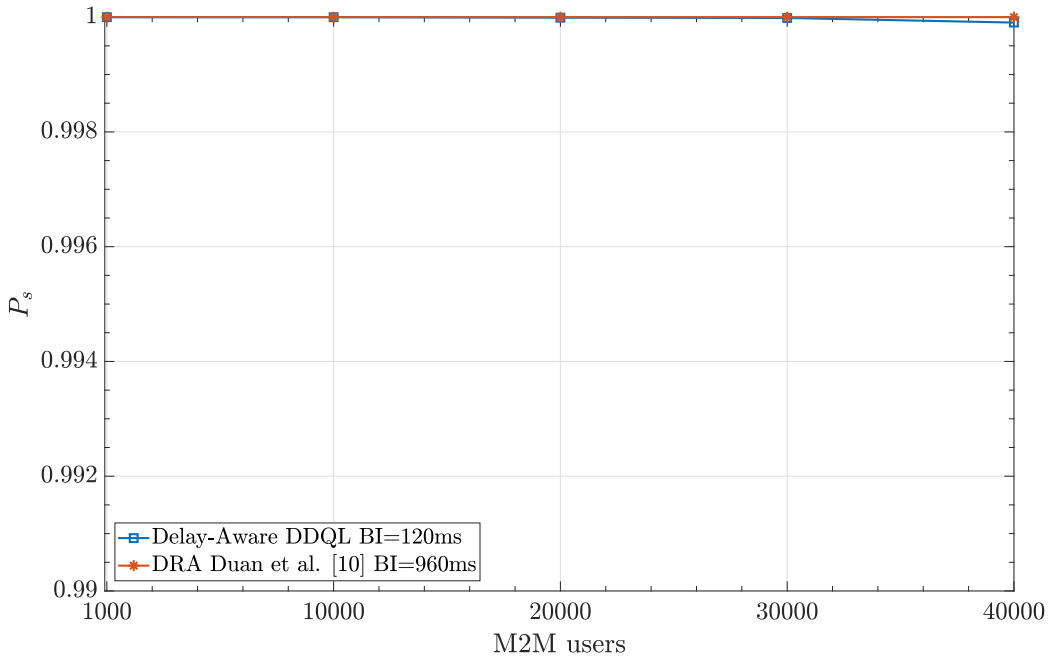

Figure 7: Probability of Successful Access for all users: DDQL mechanism and Duan et al. [10] for various M2M loads.

performance of this mechanism will suffer considerably. On the other hand, our proposed mechanism is not able to maintain P_s = 100% for the full range of M2M values tested. In fact, when there are 40000 M2M UEs, P_s = 99.99%, that is, we have a minimal error of around 0.01%. This is an acceptable value considering that we have increased the load 30% over the maximum load proposed in the standards.

In Fig. 8, we compare the mean number of transmissions for all UEs between our proposed Delay-Aware scheme and the DRA solution as the M2M load varies from 1000 to 40000 UEs. It can be seen that our solution maintains E[K] below 2, even when there are 40000 UEs. When there are 1000 UEs, E[K]=1.44, and when there are 40000 UEs, this value increases to 1.96. On the other hand, for the DRA mechanism, this value goes from 2.26 when there are 1000 M2M UEs to 26.89 when there are 40000 UEs. Having such a high number of transmissions can reduce the lifetime of IoT devices considerably, and therefore can directly affect the long-term performance of the applications.

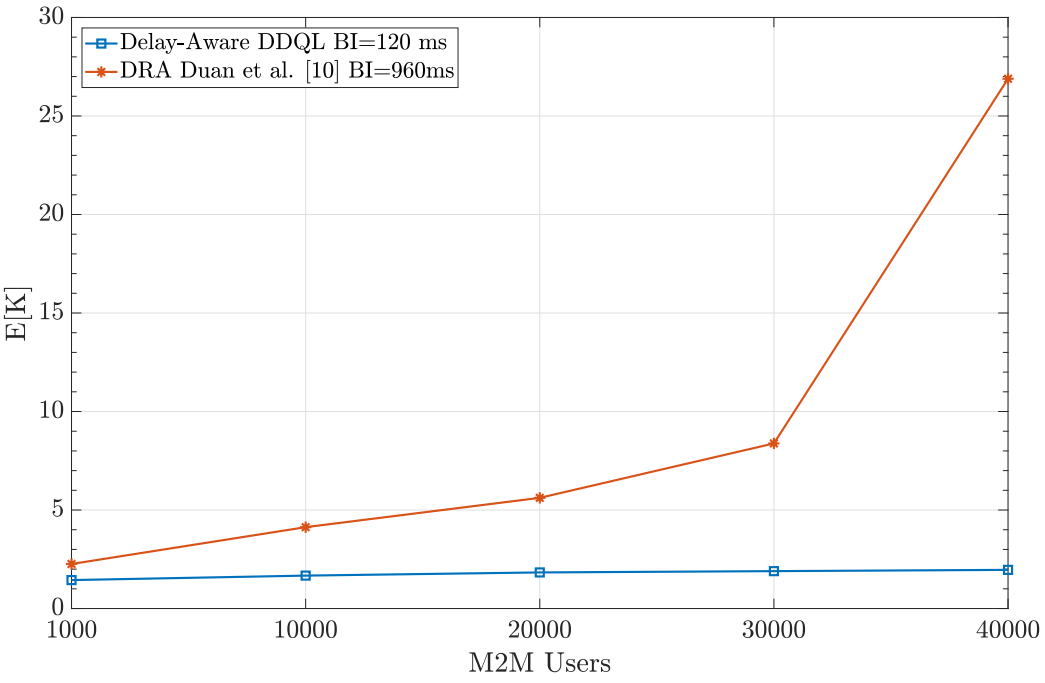

Figure 8: Mean number of transmissions for all users: DDQL mechanism and Duan et al. [10] for various M2M loads.

6. Conclusion

In this work, we proposed a Delay-Aware Double Deep Q-Learning mechanism for access control that dynamically adapts in order to allow successful mMTC access while reducing the mean access delay. This mechanism can coordinate the modification of two parameters (P_{ACB} and T_{RAO}) under different traffic conditions. We have evaluated our system when two types of traffic coexist: M2M and H2H communications. The former was modeled as bursty traffic according to the specifications. The latter was defined following traces from a Telco. Our proposed scheme shows a significant improvement over our previous Double Deep Q-Learning scheme in terms of delay, while it slightly increases the mean number of transmissions. Also, it performs better than a previously well-known dynamic solution. We have validated its performance under different traffic scenarios, showing its ability to perform in real conditions without retraining.

Acknowledgment

The work of Diego Pacheco-Paramo was in part supported by Universidad Sergio Arboleda under project Plataforma de datos para territorios inteligentes, IN.BG.086.19.014. The research of L. Tello-Oquendo was conducted under project CONV.2018-ING010, Universidad Nacional de Chimborazo.

References

[1] T. Taleb and A. Kunz, "Machine type communications in 3gpp networks: potential, challenges, and solutions," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 178–184, 2012.

[2] L. Ferdouse and A. Anpalagan, "A dynamic access class barring scheme to balance massive access requests among base stations over the cellular m2m networks," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2015, pp. 1283–1288.

[3] K. T. Ali, S. B. Rejeb, and Z. Choukair, "A dynamic access control scheme to balance massive access requests of differentiated m2m services in 5g/hetnets," in *2017 Sixth International Conference on Communications and Networking (ComNet)*. IEEE, 2017, pp. 1–6.

[4] N. Li, C. Cao, and C. Wang, "Dynamic resource allocation and access class barring scheme for delay-sensitive devices in machine to machine (m2m) communications," *Sensors*, vol. 17, no. 6, p. 1407, 2017.

[5] M. Tavana, A. Rahmati, and V. Shah-Mansouri, "Congestion control with adaptive access class barring for LTE M2M overload using Kalman filters," *Computer Networks*, vol. 141, pp. 222–233, 2018.

[6] L. Tello-Oquendo, J.-R. Vidal, V. Pla, and L. Guijarro, "Dynamic access class barring parameter tuning in lte-a networks with massive m2m traffic," in *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*. IEEE, 2018, pp. 1–8.

[7] H. Kim, S. s. Lee, and S. Lee, "Dynamic extended access barring for improved M2M communication in LTE-A networks," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2017, pp. 2742–2747.

[8] R.-H. Hwang, C.-F. Huang, H.-W. Lin, and J.-J. Wu, "Uplink access control for machine-type communications in LTE-A networks," *Personal and Ubiquitous Computing*, vol. 20, no. 6, pp. 851–862, Nov 2016.

[9] C. M. Chou, C. Y. Huang, and C.-Y. Chiu, "Loading prediction and barring controls for machine type communication," in *2013 IEEE International Conference on Communications (ICC)*. IEEE, June 2013, pp. 5168–5172.

[10] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9847–9861, 2016.

[11] Z. Wang and V. W. Wong, "Optimal access class barring for stationary machine type communication devices with timing advance information," *IEEE Transactions on Wireless communications*, vol. 14, no. 10, pp. 5374–5387, 2015.

[12] A. S. El-Hameed and K. M. Elsayed, "A Q-learning approach for machine-type communication random access in LTE-Advanced," *Telecommunication Systems*, pp. 1–17, 2018.

[13] L. M. Bello, P. Mitchell, D. Grace, and T. Mickus, "Q-learning Based Random Access with Collision free RACH Interactions for Cellular M2M," in *Next Generation Mobile Applications, Services and Technologies, 2015 9th International Conference on*. IEEE, 2015, pp. 78–83.

[14] Y. Yu, T. Wang, and S. C. Liew, "Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–7.

[15] Z. Chen and D. B. Smith, "Heterogeneous machine-type communications in cellular networks: Random access optimization by deep reinforcement learning," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[16] J. Moon and Y. Lim, "A reinforcement learning approach to access management in wireless cellular networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–17, 2017.

[17] D. Pacheco-Paramo, L. Tello-Oquendo, V. Pla, and J. Martinez-Bauset, "Deep reinforcement learning mechanism for dynamic access control in wireless networks handling mmtc," *Ad Hoc Networks*, vol. 94, p. 101939, 2019.

[18] 3GPP, *TS 36.321, Medium Access Control (MAC) Protocol Specification*, Sept 2012.

[19] ——, *TS 36.211, Physical Channels and Modulation*, Dec 2014.

[20] D. Chu, "Polyphase codes with good periodic correlation properties (corresp.)," *IEEE Transactions on information theory*, vol. 18, no. 4, pp. 531–532, 1972.

[21] 3GPP, *TS 36.331, Radio Resource Control (RRC), Protocol specification*, Sep 2017.

[22] ——, *TS 22.011, V15.1.0, Service Accessibility*, June 2017.

[23] ——, *TR 36.912, Feasibility study for Further Advancements for E-UTRA*, Apr 2011.

[24] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2018.

[25] L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, J. Martinez-Bauset, V. Casares-Giner, and L. Guijarro, "Efficient random access channel evaluation and load estimation in LTE-A with massive MTC," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1998–2002, Feb 2019.

[26] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-Learning," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, February 2015, pp. 2094–2100.

[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[28] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[29] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine learning*, vol. 8, no. 3-4, pp. 293–321, 1992.

[30] 3GPP, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Sept 2011.

[31] Nokia, "Mobile Broadband solutions for Mass Events," Nokia, Tech. Rep., 2014.

Diego Pacheco-Paramo received the BS degree in electronics engineering from Universidad de los Andes, Bogotá, Colombia, in 2004. He received the MS and PhD degrees from Universitat Politécnica de Valéncia, Spain, in 2009 and 2013, respectively. From 2012 to 2013, he was a visitor researcher in the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, USA. From 2014 to 2015, he was a postdoctoral researcher in the Laboratory of Information, Networking and Communication Sciences (LINCS) as a member of Télécom ParisTech in Paris, France. From 2016 to 2019 he was Assistant Professor at Universidad Sergio Arboleda in Bogotá, Colombia. Currently he is CRO at reconoSER ID.

Luis Tello-Oquendo received the electronic and computer engineering degree (Hons.) from Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador, in 2010, the M.Sc. degree in telecommunication technologies, systems, and networks, and the Ph.D. degree (Cum Laude) in telecommunications from Universitat Politécnica de Valéncia (UPV), Spain, in 2013 and 2018, respectively. From 2013 to 2018 he was Graduate Research Assistant with the Broadband Internetworking Research Group, UPV. From 2016 to 2017 he was a Research Scholar with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA, USA. He is currently an Associate Professor with the Universidad Nacional de Chimborazo. His research interest include MTC, wireless SDN, 5G and beyond cellular systems, IoT, machine learning. He is a member of the IEEE and ACM. He received the Best Academic Record Award from the Escuela Técnica Superior de Ingenieros de Telecomunicación, UPV, in 2013.