DRL-based Channel Access in NR Unlicensed Spectrum for Downlink URLLC

Yan Liu*, Hui Zhou‡, Yansha Deng‡, and Arumugam Nallanathan†
*College of Electronic and Information Engineering, Tongji University, Shanghai, China
†School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
‡Department of Engineering, King's College London, UK

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10000882

Abstract—**To improve the capacity of cellular systems without additional expenses on licensed frequency bands, the 3rd Generation Partnership Project (3GPP) has proposed New Radio Unlicensed (NR-U). It should be noted that each node in NR-U has to perform the Listen-Before-Talk (LBT) operation before transmission to avoid collisions by other unlicensed radio access technologies (e.g., WiFi). Thus, packets transmissions are prone to delay due to the LBT channel access mechanism. How to achieve Ultra-Reliable and Low-Latency Communications (URLLC) requirements in NR-U networks under the coexistence with WiFi networks is of importance and extremely challenging. In this paper, we develop a novel deep reinforcement learning (DRL) framework to optimize the downlink URLLC transmission in the NR-U and WiFi coexistence system through dynamically adjusting the energy detection (ED) thresholds. Our results have shown that the NR-U system reliability has been improved significantly via the DRL compared to that without learning approaches, but with the sacrifice of WiFi system reliability. To address this, we redesigned the reward to take fairness into account, which guarantees the WiFi system reliability while improving the NR-U system reliability.**

Index Terms—**5G NR-U, WiFi, URLLC, Deep Reinforcement Learning, Channel Access**

I. Introduction

The increase in traffic and bandwidth requirements for new applications in the Fifth Generation (5G) and Beyond 5G (B5G) results in the shortage of the licensed spectrum. The unlicensed spectrum becomes a promising alternative considering its low cost, high flexibility, simplicity of deployment, and availability of bandwidth [1]. The 5G NR-Unlicensed (NR-U) [2] is one of the most promising techniques, which extends NR to unlicensed bands. However, the Ultra-Reliable and Low-Latency Communications (URLLC) service guarantee will require the re-design of channel access, transmission, and reception procedures in the NR-U system [3]. Meanwhile, the harmonious coexistence with other unlicensed radio access technologies (RATs), e.g., WiFi, must be guaranteed while providing good URLLC service for the NR-U system.

Specifically, each node in NR-U or WiFi should perform Listen-Before-Talk (LBT) or Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) before transmission to guarantee fairness and avoid collision among different nodes. The node senses the channel and compares the received power with a predetermined energy detection (ED) threshold to determine if the channel is idle. In other words, a node is allowed to transmit on a channel only if the energy level in the channel is less than the ED threshold for the duration of the Clear Channel Assessment (CCA) observation time [4]. Thus, the ED threshold can often impact the system performance substantially. In a high ED threshold setting, multiple nodes may regard the channel as idle and transmit at the same time, which leads to transmission failure and degrades the reliability; while in a low ED threshold setting, nodes may regard the channel as busy even there is no transmission, which leads to low channel utilization and increased the latency. In addition, NR-U and IEEE 802.11-based technologies adopt different ED threshold settings that result in heterogeneity of their sensing floors. This creates hidden and exposed nodes problems, which reduce the total network performance [5].

Motivated by the above, it is necessary to dynamically adjust the ED threshold values in the NR-U and WiFi coexistence system to achieve better URLLC performance. This is an untreated and challenging problem. The lack of prior knowledge at the NR-U gNodeB (gNB) and the WiFi access point (AP), regarding the stochastic traffic and unobservable channel statistics (e.g., random collisions, effects of LBT and CSMA/CA, and effects of physical radio including path-loss as well as fading), makes the problem more complex.

In this paper, we develop Deep Reinforcement Learning (DRL) approaches to solve the aforementioned problem and address the following fundamental questions: 1) how to jointly optimize ED thresholds configuration for both NR-U and WiFi systems while coexistence; 2) how to exactly and correctly model the realistic process of LBT and CSMA/CA mechanisms; 3) how to satisfy the URLLC requirements over unlicensed spectrum; 4) how to ensure the NR-U and WiFi fairness. The contributions are summarized as follows:

- We develop a DRL framework to maximize the number of successfully transmitted packets under the latency constraint in the NR-U and WiFi coexistence system via dynamic ED threshold values configuration optimization.
- In our framework, we practically simulate the random traffics, the realistic process of LBT and CSMA/CA mechanisms, the transmission latency check, and the collision detection. We use this generated simulation environment to train the DRL agents.
- Our results have shown that the reliability of the NR-

978-1-6654-3540-6/22 © 2022 IEEE

U system has been improved significantly by applying the DRL, but sacrificing the WiFi performance. To guarantee the WiFi performance while improving the NR-U performance, we redesign the reward function by taking the fairness into account. The results shown that the WiFi reliability is not sacrificed based on our reward design.

The remainder of this paper is organized as follows. Section II illustrates the system model. Section III describes the problem analysis and formulation. Section IV elaborates on the proposed DRL algorithm. Simulation results are illustrated in Section V. Finally, Section VI concludes the results.

II. System Model

We consider an indoor downlink (DL) transmission scenario defined by 3GPP in [2] for the NR-U and WiFi coexistence system in the 5 GHz band, which is located in a 120m × 80m area and a distance between two neighbors gNB/AP nodes of 40 m, as shown in Fig. 1. WiFi and NR-U systems share a single 20-MHz channel, and each of them deploys three small cells in a one-floor building. The gNB/AP nodes are mounted at a height of $H_{\rm gNB}/H_{\rm AP}$ on the ceiling and NR-U user equipments (UEs)/WiFi stations (STAs) are uniformly distributed in this layout with a height of $H_{\rm UE}/H_{\rm STA}$. Each UE/STA is connected to the closest gNB/AP, and each gNB/AP is associated with fifteen UEs/STAs.

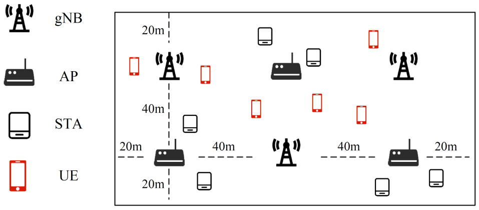


Fig. 1: NR-U and WiFi Coexistence Systems.

A. Network Model

We consider a flat Rayleigh small-scale fading channel, where the channel between two generic locations $x, y \in \mathbb{R}^3$ is assumed to follow $g(x,y) \sim \mathcal{CN}(0,1)$. All the channel gains are independent of each other, independent of the spatial locations, and identically distributed (i.i.d.). We also consider an indoor mixed office large-scale path-loss model [6] as

$$\begin{cases} \zeta_L(x,y) = 32.4 + 17.3\log_{10}(d_{\rm 3D}) + 20\log_{10}(f_{\rm c}), & P_{\rm LOS} \\ \zeta_N(x,y) = 32.4 + 31.9\log_{10}(d_{\rm 3D}) + 20\log_{10}(f_{\rm c}), & P_{\rm NLOS} \end{cases} \tag{1}$$

where $\zeta_L(x,y)$ and $\zeta_N(x,y)$ represent the large-scale path-loss under Line-Of-Sight (LOS) and Non-Line-Of-Sight (NLOS), respectively. In (1), $f_{\rm c}$ is the carrier frequency, $d_{\rm 3D}$ is the 3D distance between x and y, $P_{\rm LOS}$ is the LOS probability and $P_{\rm NLOS} = 1 - P_{\rm LOS}$ is the NLOS probability. The indoor mixed office LOS probability $P_{\rm LOS}$ is defined as

$$P_{\rm LOS} = \begin{cases} 1 & , d_{\rm 2D} < 1.2{\rm m} \\ \exp(\dfrac{d_{\rm 2D} - 1.2}{4.7}) & , 1.2{\rm m} \le d_{\rm 2D} \le 6.5{\rm m} \\ \exp(\dfrac{d_{\rm 2D} - 6.5}{32.6}) & , d_{\rm 2D} > 6.5{\rm m} \end{cases} \tag{2}$$

where $d_{\rm 2D}$ is the projection of $d_{\rm 3D}$ on the horizontal plane.

Then, the mean channel power gain could be derived as

$$h = g(10^{-(\zeta_L P_{LOS} + \zeta_N(1 - P_{LOS}))/10}), \tag{3}$$

where the spatial indices (x, y) are dropped for the brevity.

B. Chanel Access Schemes in Unlicensed Band

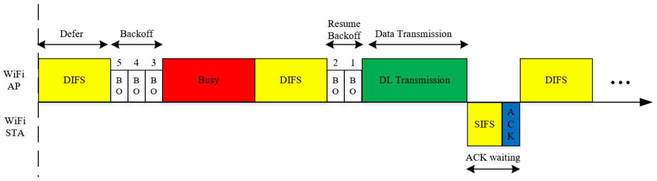


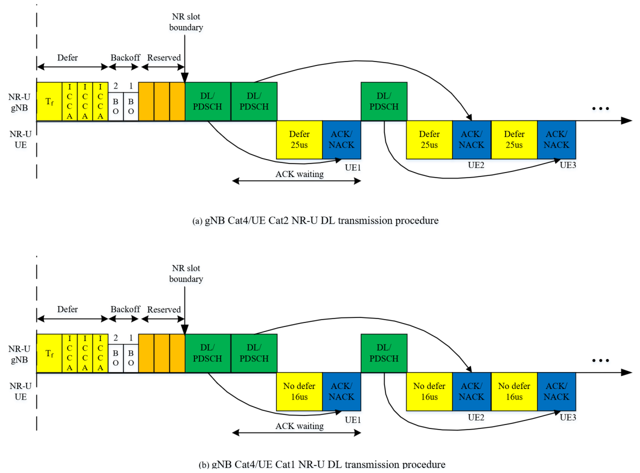
Fig. 2: CSMA/CA contention and frame structure in WiFi.

1) CSMA/CA in WiFi: As as illustrated in Fig. 2, if the channel has been continuously free over a Distributed Interframe Spacing (DIFS) interval, the AP transmits data immediately. Otherwise, the AP constantly monitors the channel until it is measured idle for a DIFS period, then selects a random backoff duration and counts down. The backoff procedure starts with a selection of a random "slotted" backoff interval BI = rand$[0, CW)$, where CW is the contention window. Then, the AP decrements the backoff counter each time the channel is sensed idle for a CCA slot duration $T_{\rm CCA}$. If the AP detects the channel busy, it suspends the backoff counter and continues to monitor the busy channel until it goes idle for a further DIFS period. The AP resumes the backoff counter when the channel is idle for a DIFS. Once the counter BI reaches zero, the AP transmits its packets. Any node that did not complete its countdown to zero in the current round, carries over the backoff value and resumes countdown in the next round. The destination STA, upon receiving the packets correctly, waits for a Short IFS (SIFS) interval immediately and transmits an Acknowledgment (ACK) feedback to the source AP in order to confirm the correct reception. When a transmission is lost, the contention window CW is doubled and applied for the retransmissions until it reaches a maximum value $CW_{\rm max}$. When a transmission is successful, contention window CW is reset to its minimum value $CW_{\rm min}$. When a maximum number of retransmissions is exhausted, the packets are dropped and CW is reset to its minimum value $CW_{\rm min}$.

2) LBT in NR-U: NR-U follows a LBT approach similar as CSMA/CA, i.e., a gNB is required to perform CCA to determine whether the channel is idle or not. Once the channel has been idle for $T_f = 16\mu s$, the gNB performs

CCA in d_i consecutive observation CCA slots, each lasting $T_{\text{CCA}} = 9\mu s$. If the channel is occupied in any of these slots, the process restarts, whereas if the channel is idle for all d_i slots, the gNB starts the backoff procedure by selecting a random number of observation slots $q \in \{0, 1, 2, ..., CW-1\}$. CCA is performed for each observation slot: if the channel is idle, q is decremented by one, otherwise the backoff contention is suspended and CCA is aborted. When $q = 0$, the gNB starts transmitting but this transmission cannot exceed the maximum channel occupancy time (MCOT). In summary, each channel access is preceded by a fixed (d_i) and random (q) number of slots. The latter depends on the current CW value, which is re-set to $CW_{\min}$ after a successful transmission and doubled (up to $CW_{\max}$) before each retransmission. In 5G NR-U, four LBT Categories have been defined [2].

- Category 1 (Cat1 LBT): Immediate transmission after a short switching gap of 16μs without performing LBT.
- Category 2 (Cat2 LBT): LBT without random backoff, where the CCA period is deterministic (e.g., 25μs).
- Category 3 (Cat3 LBT): LBT with random backoff and fixed contention window size.
- Category 4 (Cat4 LBT): LBT with random backoff with a contention window of variable size.



(a) gNB Cat4/UE Cat2 NR-U DL transmission procedure

(b) gNB Cat4/UE Cat1 NR-U DL transmission procedure

Fig. 3: LBT contention and frame structure in NR-U.

The Cat4 LBT is used for gNB to initiate a COT for data transmissions, while UE can use Cat2 LBT for specific signaling like feedback signals, as shown in Fig. 3 (a) (see details in [7]). Recently, NR-U has agreed to support Cat1 LBT, where the responding UEs are allowed to transmit without performing a CCA check if the gap between DL and UL transmissions is less than 16μs, as shown in Fig. 3 (b). This is one feature to benefit URLLC operations in NR-U.

3) Key differences between NR-U and WiFi:

a) Energy Detection Threshold: The sensitivity requirements of CCA for WiFi depend on two thresholds, CCA Carrier Sense (CCA-CS) threshold $\eta_{\text{wifi}}^{\text{cs}}$ and CCA Energy Detection (CCA-ED) threshold $\eta_{\text{wifi}}^{\text{ed}}$. CCA-CS refers to the capability of the receiver to detect and decode the preambles of other WiFi signals, and CCA-ED is used to detect the energy level of both WiFi and NR-U signals. When both of CCA-ED and CCA-CS levels are not higher than a set threshold, the CCA is indicated as idle. In contrast, in LBT, CCA-ED ($\eta_{\text{nru}}^{\text{ed}}$) is the only recommended sensing method for all signals.

b) Frame Structures: NR-U follows the same slotted structure as the NR, i.e., a node can only start data transmission at the spectrum slot boundaries (SSBs). Even though the backoff procedure may finish at any moment within the NR-U slot, the 3GPP specification does not regulate the behavior of the node until the next SSB. In the context of this, when the NR-U gNBs finish the backoff procedure and wait for the next SSB, if the WiFi APs started transmission, the WiFi APs could occupy the channel, which degrades the NR-U performance. To this end, one usual way is to send a reservation signal to occupy the channel, as shown in Fig. 3.

c) URLLC DL Transmission: For WiFi, an AP serves only one STA in a single TTI as shown in Fig. 2. In contrast, an NR-U gNB can serve multiple UEs in a single TTI and their respective ACKs may be received at different times, as shown as shown in Fig. 3. Besides, different from the LAA that supported a single DL/UL switching point within the COT, NR-U supports multiple DL/UL switching points within the COT [8, Sec. 7.6.2], which i) simplifies the feedback timings and ii) ensures channel availability in UL (in case a new LBT has to be done). This configuration is suitable for delay-sensitive URLLC traffic.

III. Problem Analysis and Formulation

In this section, we formulate the coexistence problem of NR-U and WiFi systems for DL URLLC transmission.

A. Energy detection

In NR-U and WiFi coexistence systems, each gNB and AP are required to sense the channel energy at each slot during the defer or back off periods. The sensed channel energy of gNB or AP i can be represented as

$$\begin{cases} \text{WiFi}_i^{\text{cs}} = \sum\limits_{k \in \mathcal{K}_w \setminus i} \alpha_k \text{P}_k h_{k,i}, & (4) \\ \text{WiFi}_i^{\text{ed}} = \sum\limits_{k \in \mathcal{K} \setminus i} \alpha_k P_k h_{k,i}, & (5) \\ \text{NR}_i^{\text{ed}} = \sum\limits_{k \in \mathcal{K} \setminus i} \alpha_k P_k h_{k,i}, & (6) \end{cases}$$

where $\mathcal{K}$ represents the set of all transmitters, $\mathcal{K}_w$ represents the set of all WiFi APs, α_k indicates whether the node k is transmitting or not, P_k represents the DL transmit power, and $h_{k,i}$ is the channel gain between node k and node i.

B. Signal to Interference and Noise Ratio (SINR)

When the gNB/ AP completes the defer and backoff period, the gNB/AP starts the data transmission. We model the decoding process via SINR, where the SINR of data transmission from transmitter i to the receiver j can be represented as

$$\mathrm{SINR}_{i,j} = \frac{P_i h_{i,j}}{\sum_{k\in\mathcal{K}\setminus i} \alpha_k P_k h_{k,j} + \sigma_n^2}, \tag{7}$$

where $h_{i,j}$ is the channel from the transmitter i to the receiver j, P_i represents the transmit power of transmitter i, and σ_n^2 is the power of the noise. At the receiver side, if $\mathrm{SINR}_{i,j}$ is less than a predefined SINR threshold γ_{th}, the incoming packets cannot be decoded successfully, and hence there will be no ACK returned to the corresponding transmitter.

C. Problem Formulation

We tackle problem of optimizing the NR-U and WiFi coexistence system configuration for DL URLLC defined by parameters $A^t = \left\{\eta_{\mathrm{nru}}^{\mathrm{ed},t}, \eta_{\mathrm{wifi}}^{\mathrm{ed},t}\right\}$ for NR-U and WiFi systems in each subframe t, respectively. In order to make this decision at the beginning of each subframe t, the central server accesses all prior history $U^{t'}$ in subframes $t' = 1, ...t-1$, consisting of the following variables: the number of successfully transmitted NR-U packets $N_{\mathrm{nru}}^{t'}$ and the number of successfully transmitted WiFi packets $N_{\mathrm{wifi}}^{t'}$. We denote $O^t = \left\{A^{t-1}, U^{t-1}, A^{t-2}, U^{t-2}, ..., A^1, U^1\right\}$ as the observed history of all such measurements and past actions. At each subframe, the NR-U and WiFi coexistence systems aim at maximizing the long-term objective R^t related to the average number of successfully transmitted packets under the latency constraint T_{cons} with respect to the stochastic policy π that maps the current observation history O^t to the probabilities of selecting each possible configuration A^t. This optimization problem (P1) can be formulated as

$$(\mathrm{P1}): \max_{\pi(A^t|O^t)} \quad \sum_{k=t}^{\infty} \gamma^{k-t} \mathbb{E}_\pi[N_{\mathrm{nru}}^k + N_{\mathrm{wifi}}^k] \tag{8}$$

$$s.t. \quad T_{\mathrm{late}} \leq T_{\mathrm{cons}}, \tag{9}$$

where $\gamma \in [0, 1)$ is the discount rate for the performance accrued in the future. We assume that the packet is dropped if it is not successfully transmitted under the latency constraint.

IV. Deep Reinforcement Learning Solutions

The problem (P1) is a Partially Observable Markov Decision Process (POMDP) problem. It is well known that RL is a promising technique to solve POMDP problems. However, traditional RL methods do not scale well with high dimensional state-action spaces. To overcome it, DRL, the combination of RL with the deep neural network (DNN), is proposed [9]. To this end, we resort to DRL in this work.

A. Reinforcement Learning (RL) Framework

We define $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $r \in \mathcal{R}$ as any state, action, and reward from the their corresponding sets, respectively. At the beginning of the tth subframe ($t \in \{0, 1, 2, ...\}$), the RL agent first observes the current state S^t corresponding to a set of previous observations $\left(O^t = \left\{U^{t-1}, U^{t-2}, ..., U^1\right\}\right)$ in order to select an specific action $A^t \in \mathcal{A}\left(S^t\right)$. Here, the action A^t corresponds to the NR-U and WiFi ED thresholds $\eta_{\mathrm{nru}}^{\mathrm{ed},t}$ and $\eta_{\mathrm{wifi}}^{\mathrm{ed},t}$, and the state S^t is a set of indices mapped to the current observed information $U^{t-1} = \left[N_{\mathrm{nru}}^{t-1}, N_{\mathrm{wifi}}^{t-1}\right]$, respectively.

With the knowledge of the state S^t, the RL agent chooses an action A^t from the set $\mathcal{A}$, which is a set of indices mapped to the set of available energy thresholds $\mathcal{F}_{\mathrm{ed}}$. Once an action A^t is performed, the RL agent will receive a corresponding reward R^{t+1}, and observe a new state S^{t+1}. The reward R^{t+1} indicates to what extent the executed action A^t can achieve the optimization goal, which is designed based on the new observed state S^{t+1}. As the optimization goal is to maximize the number of successfully transmitted packets in the NR-U and WiFi coexistence system, we define the reward R^{t+1} as a function that positively proportional to the observed number of successfully transmitted packets in the system as

$$R^{t+1} = (N_{\mathrm{nru}}^t + N_{\mathrm{wifi}}^t)/c_{\mathrm{suc}}, \tag{10}$$

where c_{suc} is a constant to normalize the reward function.

B. Deep Reinforcement Learning (DRL) Algorithm

We estimate the Q value via a DNN as a more effective but complicated function approximator, known as Deep Q-Network (DQN), which is a classic DRL algorithm. When updating the DQN algorithm, mini-batch samples are selected randomly from the experience memory as the input of the neural network, which breaks down the correlation among the training samples. By averaging the selected samples, the distribution of training samples can be smoothed, which avoids the training divergence.

The DQN agent parameterizes the action-state value function $Q(s, a)$ via a function $Q(s, a; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the weights matrix of a DNN with multiple layers. We consider the conventional fully-connected DNN, where the neurons between two adjacent layers are fully pairwise connected. The input of the DNN is given by the variables in the state S^t; the intermediate hidden layers are Rectifier Linear Units (ReLUs) by utilizing the function $f(x) = \max(0, x)$; while the output layer is composed of linear units, which are in one-to-one correspondence with all available actions in $\mathcal{A}$.

We consider ϵ-greedy approach to balance exploitation and exploration in the actor of the DRL agent, where ϵ is a positive real number and $\epsilon \leq 1$. In each subframe t, the agent randomly generates a probability p_ϵ^t to compare with ϵ. Then, with the probability ϵ, the algorithm randomly chooses an action from the remaining feasible actions to improve the estimate of the non-greedy action's value. With the probability $1 - \epsilon$, the exploitation is achieved by performing forward propagation of Q-function $Q(s, a; \boldsymbol{\theta})$ according to the observed state S^t. The weights matrix $\boldsymbol{\theta}$ is updated online along each training episode to avoid the complexities of eligibility traces, where a double deep Q-learning (DDQN) training principle [10] is applied to some extent reduce the substantial overestimations of value function. Accordingly, learning takes place over multiple training episodes, with each episode of duration N_{sub} periods. In each subframe, the parameter $\boldsymbol{\theta}$ is updated using RMSProp optimizer [11] as

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \lambda_{\mathrm{RMS}} \nabla L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t), \tag{11}$$

where $\lambda_{\mathrm{RMS}} \in (0,1]$ is RMSProp learning rate, and $\nabla L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t)$ is the gradient of the loss function $L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t)$ used to train the state-action value function. The gradient of the loss function is defined as

$$\nabla L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t) = \mathbb{E}_{S^i,A^i,R^{i+1},S^{i+1}}\Big[\big(R^{i+1}+\gamma \max_a Q(S^{i+1},a;\bar{\boldsymbol{\theta}}^t) - Q(S^i,A^i;\boldsymbol{\theta}^t)\big)\nabla_{\boldsymbol{\theta}} Q(S^i,A^i;\boldsymbol{\theta}^t)\Big]. \tag{12}$$

The expectation is taken over the so-called minibatch, which are randomly selected from previous samples $(S^i,A^i,S^{i+1},R^{i+1})$ for some $i \in t-M_r,...,t$, with M_r being the replay memory size [12]. When $t-M_r$ is negative, it is interpreted as including samples from the previous episode. Following DDQN, in (12), $\bar{\boldsymbol{\theta}}^t$ is a so-called target Q-network that is used to estimate the future value of the Q-function in the update rule and $\bar{\boldsymbol{\theta}}^t$ is periodically copied from the current value $\boldsymbol{\theta}^t$ and kept fixed for a number of episodes. The detailed DRL algorithm is presented in **Algorithm 1**.

C. Fairness between NR-U and WiFi systems

According to the NR-U fairness defined by 3GPP [2] that the NR-U network not degrading the WiFi network performance, compared to the case where two WiFi networks are deployed. Thus, to guarantee the performance of the WiFi system while improving the NR-U system, we replace the reward function in (10) by the redesign reward function as

$$R^{t+1} = \begin{cases} (N^t_{\mathrm{nru}} + N^t_{\mathrm{wii}})/c_{\mathrm{suc}} & , N^t_{\mathrm{wifi}} \geq \mathrm{N}^{\mathrm{thres}}_{\mathrm{wifi}} \\ 0 & N^t_{\mathrm{wifi}} < \mathrm{N}^{\mathrm{thres}}_{\mathrm{wifi}}, \end{cases} \tag{13}$$

where $\mathrm{N}^{\mathrm{thres}}_{\mathrm{wifi}}$ is the predefined successful packets threshold of the WiFi system, which can guide the agent to optimize the NR-U performance without sacrificing the WiFi performance.

V. Simulation Results

In this section, we examine the effectiveness of our proposed DRL algorithms via simulation. We adopt the standard network parameters listed in Table I, and hyperparameters for the DQN learning algorithm listed in Table II. We apply a Poisson arrival traffic, which is also known as FTP-3 recommended by 3GPP [13]. All testing performance results are obtained by averaging over 3000 episodes. The DQN is set with two hidden layers, each with 128 ReLU units.

TABLE I: Simulation Parameters

Parameters	Value	Parameters	Value
Height of gNB and AP	3m	Height of UE and STA	1m
File Size	32 bytes	SIFS	16μs
Defer Time	25μs	Contention Window	{4,8}
DL Transmit Power	23dBm	UL Transmit Power	18dBm
Noise Power	-104dBm	SCS	60 KHz
Mini-slot	36μs	Time Constraint	1ms
WiFi SINR Threshold	9dB	NR-U SINR Threshold	5.5dB
WiFi Rate	21.7Mbps	NR-U Rate	25.2Mbps
WiFi COT	2.080ms	NR-U COT	2ms
Preamble Detection Threshold	-82dBm	WiFi ED Threshold	-62dBm
NR-U ED Threshold	-72dBm	Packet Arrival Rate	5 packet/ms

TABLE II: Learning Hyperparameters

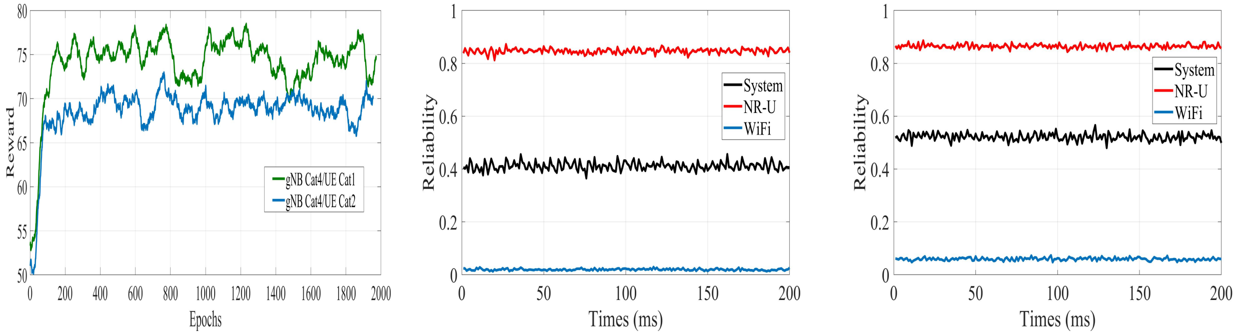
Hyperparameters	Value	Hyperparameters	Value
Learning rate λ_{RMS}	0.0001	Minimum exploration rate ϵ	0.1
Discount rate γ	0.5	Minibatch size	32
Replay Memory	10000	Target Q-network update frequency	1000

Algorithm 1: DRL-Based ED Thresholds Configuration

```
Input: : Action space $\mathcal{A}$ and Operation Iteration I.
Algorithm hyperparameters: learning rate
  $\lambda_{RMS} \in (0,1]$, discount rate $\gamma \in [0,1)$, $\epsilon$-greedy rate
  $\epsilon \in (0,1]$, target network update frequency $J$;
Initialization of replay memory $M$ to capacity $D$, the
  state-action value function $Q(S, A, \boldsymbol{\theta})$, the
  parameters of primary Q-network $\boldsymbol{\theta}$, and the target
  Q-network $\bar{\boldsymbol{\theta}}$;
for Iteration $\leftarrow 1$ to I do
    Initialization of $S^1$ by executing a random action
      $A^0$;
    for $t \leftarrow 1$ to T do
        if $p_\epsilon < \epsilon$ Then select a random action $A^t$
          from $\mathcal{A}$;
        else select $A^t = \arg\max_{a \in \mathcal{A}} Q(S^t, A^t, \boldsymbol{\theta})$.
        The BS broadcasts $A^t$ and backlogged UEs
          attempt communication in the $t$th subframe;
        The BS observes state $S^{t+1}$, and calculate the
          related reward $R^{t+1}$;
        Store transition $(S^t, A^t, R^{t+1}, S^{t+1})$ in replay
          memory $M$;
        Sample random minibatch of transitions
          $(S^t, A^t, R^{t+1}, S^{t+1})$ from replay memory $M$;
        Perform a gradient descent step and update
          parameters $\boldsymbol{\theta}$ for $Q(S^t, A^t, \boldsymbol{\theta})$ using (12);
        Update the parameter $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}$ of the target
          Q-network every $J$ steps.
    end
end
```

Fig. 4 (a) plots the convergence of reward for two schemes. The faster coverage of gNB Cat4/UE Cat1 scheme (at around 200 epochs) than gNB Cat4/UE Cat2 scheme (at around 500 epochs) is due to that Cat2 LBT is more dynamic and complicated. The reward of gNB Cat4/UE Cat1 scheme is larger than that of gNB Cat4/UE Cat2 scheme, due to that Cat2 LBT leads to long latency, thus degrading the reliability.

As shown in the Fig. 4 (b) and Table. III, we test the reliability performance of converged DRL trained model for gNB Cat4/UE Cat2 scheme and compares them with the performance of fixed ED threshold setting. We see that the system reliability achieves a performance gain at around +246.64%, and the NR-U reliability achieves up to +177.51% performance gain, compared with fixed ED threshold setting. However, this is achieved by sacrificing the WiFi system performance as shown in Table. III, where the reliability of WiFi drops from around 9.18% to around 2.25%. By redesigning the reward function to take the fairness into



(a) (b) (c)

Fig. 4: (a) Convergence of DRL (b) Reliability for gNB Cat4/UE Cat2 scheme (c) Reliability for gNB Cat4/UE Cat1 scheme

account, the reliability of WiFi increases from around 9.18% to around 9.54%.

TABLE III: Reliability for gNB Cat4/UE Cat2 Scheme

gNB Cat4/UE Cat2			
Reliability	NR-U + WiFi System	NR-U	WiFi
NRU+WiFi (Fix)	13.12%	30.37%	9.18%
NRU+WiFi (DRL)	45.48%	84.28%	2.25%
Gain (DRL)	+246.64%	+177.51%	-75.49%
NRU+WiFi (DRL+fair)	30.92%	69.98%	9.54%
Gain (DRL+fair)	+135.67%	+130.42%	+3.92%

Similarly, we test the reliability performance of converged trained model for gNB Cat4/UE Cat1 scheme in Fig. 4 (c) and Table. IV. First, we observe that both system and NR-U achieve the expected more than 100% reliability performance gain by applying the DRL. Then, we observe that both the system and NR-U reliability of gNB Cat4/UE Cat1 scheme is better than that of gNB Cat4/UE Cat2 scheme, due to that the UEs performing LBT process before sending feedback in gNB Cat4/UE Cat2 scheme will occupy channels for more time, thus leading to low reliability. It should be noted that for the gNB Cat4/UE Cat1 scheme, the performance gain of WiFi even increase to +118.11% based on our reward design.

TABLE IV: Reliability for gNB Cat4/UE Cat1 Scheme

gNB Cat4/UE Cat1			
Reliability	NR-U + WiFi System	NR-U	WiFi
NRU+WiFi (Fix)	13.61%	40.41 %	5.19 %
NRU+WiFi (DRL)	52.17%	86.79 %	5.89 %
Gain (DRL)	+283.32%	+114.77%	+13.48%
NRU+WiFi (DRL+fair)	51.85%	82.25 %	11.32 %
Gain (DRL+fair)	+280.96%	+103.54%	+118.11%

VI. Conclusion

In this paper, we first developed a novel DRL framework to optimize the ED threshold values configuration for attaining the long-term successfully transmitted packets under the latency constraint in the NR-U and WiFi coexistence system for DL URLLC. We then evaluated the reliability performance for the gNB Cat4/UE Cat2 scheme and the gNB Cat4/UE Cat1 scheme using our proposed learning framework, respectively. In addition, to guarantee the performance of the WiFi system while improving the NR-U system, we take the fairness into account in our proposed framework by redesigning the reward function. Our results have shown that the reliability of the NR-U system has been improved significantly for both schemes by applying the DRL and our redesign reward could guarantee the performance of the WiFi system while improving the NR-U system performance.

References

[1] P. Yang, L. Kong, and G. Chen, "Spectrum sharing for 5G/6G URLLC: Research frontiers and standards," *IEEE Commun. Standards Mag.*, vol. 5, no. 2, pp. 120–125, 2021.

[2] "Study on NR-based access to unlicensed spectrum; study on NR-based access to unlicensed spectrum (release 16)," *3GPP, TR 38.889 V16.0.0*, Dec. 2018.

[3] T.-K. Le, U. Salim, and F. Kaltenberger, "An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17," *IEEE Access*, vol. 9, pp. 433–444, Dec. 2020.

[4] R. Bajracharya, R. Shrestha, and H. Jung, "Future is unlicensed: Private 5G unlicensed network for connecting industries of future," *Sensors*, vol. 20, no. 10, p. 2774, May. 2020.

[5] L. B. Jiang and S. C. Liew, "Improving throughput and fairness by reducing exposed and hidden nodes in 802.11 networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 1, pp. 34–49, Jan. 2008.

[6] "Study on channel model for frequencies from 0.5 to 100 ghz," *3GPP, TR 38.901 v16.1.0*, Jan 2020.

[7] "Technical specification group radio access network; study on licensed-assisted access to unlicensed spectrum (release 13)," *3GPP, TR 36.889 V13.0.0*, Jun. 2015.

[8] "Draft report, chairman notes," *3GPP R1-180xxxx, 3GPP TSG RAN WG1 93 Meeting*, May. 2018.

[9] 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, pp. 529–533, 2015.

[10] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," *arXiv preprint arXiv:1509.06461*, Dec. 2015.

[11] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, Oct. 2012.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] "Study on RAN improvements for machine-type communications," *3GPP, TR 37.868 v11.0.0*, Sep. 2011.