Dynamic Medium Access in Clustered NOMA IoT Networks based on Reinforcement Learning

Abhishek Kumar*, Jorge Martinez-Bauset†, and Frank Y. Li*
*Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway
†Departamento de Comunicaciones, Universitat Politècnica de València (UPV), València 46022, Spain
Email: {abhishek.kumar, frank.li}@uia.no; jmartinez@upv.es

***Abstract*—The burgeoning growth of massive machine-type communication or massive Internet of things traffic in beyond fifth-generation communication systems calls for novel solutions for medium access. In this paper, we propose a reinforcement learning (RL)-based random access scheme for non-orthogonal multiple access (NOMA)-enabled uplink transmissions and investigate how RL can potentially bring benefits for medium access under various network configurations and traffic load conditions. Aiming to mitigate network congestion and enhance overall network performance, we introduce access control and compare network performance of the RL-based access scheme with two benchmark schemes. For performance assessment, we define system and cluster throughput as the key performance metrics. This study provides valuable insights on the feasibility and efficacy of RL-based access control mechanisms across low and high device density scenarios.**

***Index Terms*—NOMA, reinforcement learning, uplink traffic, concurrent transmissions, performance evaluation.**

I. Introduction

The advent towards the next-generation communication systems, particularly in the realm of massive machine-type communication (mMTC) or massive Internet of things (mIoT), heralds a transformative era in wireless networks. In such systems, a myriad of devices, sensors, and machines, at a density of $10^6 \sim 10^7$ devices per square kilometer, may be connected to the network, requiring intelligent multiple medium access mechanisms to maximize radio resource utilization [1].

To facilitate multi-user transmissions, non-orthogonal multiple access (NOMA) emerges as a highly efficient multiple access mechanism as it allows concurrent transmissions share the same radio resource (time and frequency) [2]. In contrast to conventional orthogonal multiple access (OMA) that relies on dedicated resource allocation. NOMA receivers deploy successive interference cancellation (SIC) to disentangle concurrent signals originated from multiple users and potentially deem all transmissions as successful. This distinguishing feature positions NOMA as a promising candidate for multiple access, particularly in the landscape of mMTC/mIoT applications.

While NOMA exhibits advantages over OMA concerning spectral efficiency and cumulative rate, the scalability and complexity of user pairing in NOMA pose challenges. On the other hand, leveraging spatial diversity through clustering and employing multiple antennas at a base station (BS) have resulted in a reduced number of concurrent transmissions based on the same radio resource. Nevertheless, dealing with a substantial number of contending devices for medium access calls for novel approaches. In scenarios with lower device density, Markov modeling has been proven as a powerful tool for performance analysis and prediction, as demonstrated in [3] [4]. However, implementing NOMA and accommodating a large number of devices is not an easy task. To further optimize radio resource utilization in this context, there is a growing interest in exploring innovative techniques, such as incorporating machine learning instead of conventional Markov modelling for medium access. This avenue seeks to leverage intelligent learning and decision-making algorithms to dynamically adapt and optimize radio resource management for mMTC/mIoT applications.

In recent years, there has been a discernible upswing in endeavors to explore reinforcement learning (RL) algorithms in cluster-based uplink NOMA transmissions. In [5], an RL-based access scheme was proposed for resource allocation, addressing factors such as the number of users, channel gains, and transmit power levels within multi-constrained cluster uplink NOMA IoT networks. Additionally, a Q-learning (QL)-based random access method for NOMA MTC networks was proposed in [6]. Their method incorporates considerations such as short-range clustering, a feedback-based reward mechanism, and an adaptive frame structure. Furthermore, the study in [7] introduced a QL scheme designed to minimize random access channel collisions by determining the most suitable random access slot for each resource-constrained MTC device. In the context of grant-free NOMA systems, [8] introduced a deep RL algorithm to enhance throughput. Moreover, [9] proposed two distributed QL methods for managing bursty mMTC traffic in a uplink grant-free NOMA scenario.

However, a prevailing assumption in a majority of existing studies is that devices exhibit saturated traffic, meaning that they always have at least one packet ready for transmission. Despite this focus on individual device traffic, there has been limited attention given to the intricacies of concurrent transmissions from multiple clusters. Furthermore, none of the above studies delved into the exploration of collisions caused by transmissions from different clusters. To address these issues, the focus in this study is centered on the development of a robust RL algorithm aimed to efficiently utilize radio resources in the dynamic context of packet arrivals and access control, encompassing scenarios that involve both intra- and inter-cluster transmissions. In this paper, we conduct a comprehensive evaluation of RL-based uplink data transmissions,

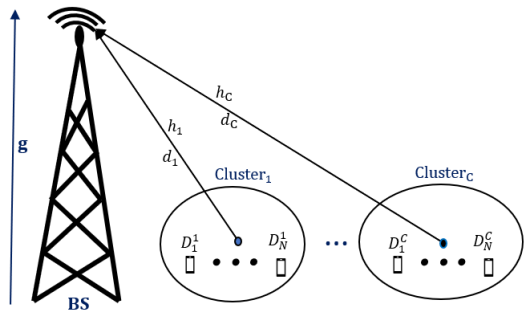

Fig. 1. Network scenario: Cluster-based concurrent uplink transmissions in NOMA-enabled network. When can RL help?

examining their performance across diverse network configurations. Our investigation seeks to determine the circumstances under which RL appears to be beneficial. Specifically, we explore whether the introduction of RL becomes imperative in scenarios characterized by a low or high device density.

The rest of the paper is organized as follows. After outlining network scenario and transmission principle in Sec. II, Sec. III describes the proposed RL-based access scheme. Then Sec. IV presents simulation results based on three access schemes, followed by Sec. V that concludes the paper.

II. Network Scenario and Access Principle

In this section, we illustrate the envisaged network scenario and explain the data transmission principle in this network.

A. Network Scenario

Consider a static NOMA-enabled IoT network that consists of a BS serving multiple devices randomly distributed within the cell coverage. Devices are grouped into multiple clusters[1] and two or more clusters are covered by a single beam. All devices in this network are battery-powered, equipped with a single antenna, and have identical transmit power. As depicted in Fig. 1, devices within clusters adhere to the NOMA transmission principle to transmit their packets to the BS, which contains M antennas. The height of these antennas is g meters above the ground. A cluster c_i $(i = 1, \ldots, C)$, with its center positioned at a distance of d_i meters away from the BS, comprises N_i devices that are uniformly distributed within a certain radius from the cluster center.

By allowing concurrent transmissions from devices in clusters, the total received signal y at the BS is obtained by

$$y = \sum_{i=1}^{C} \sum_{j=1}^{N_i} \mathbf{H}_j^i x_j^i + \tau, \tag{1}$$

where the symbols x_j^i, $\mathbf{H}_j^i$, and τ represent the transmitted signal by the j-th device from the i-th cluster, the complex channel gain vector between the j-th device from the i-th cluster and the BS, and the additive noise, respectively. The additive noise present in the channel follows a Gaussian distribution with zero mean and variance σ^2, denoted as $\tau \sim \mathcal{CN}\left(0, \sigma^2\right)$. The complex channel gain between devices and the BS $\mathbf{H}_j^i \sim \mathcal{CN}\left(\mathbf{0}_M, \mathbf{I}_M\right)$ follows Rayleigh fading with zero mean complex Gaussian distribution. The signal transmitted by device D_j^i, denoted as x_j^i, is expressed by

$$x_j^i = \sqrt{P} s_j^i \,, \tag{2}$$

where P is the transmit power for device D_j^i in cluster c_i, s_j^i is the transmit data signal for device D_j^i with unit variance. The path loss is calculated as $128 + 37.6 \log_{10} d$, where d is the distance (in kilometers) between a device and the BS.

[1]A *cluster* confines a group of devices that are located in the vicinity of each other and labeled with the same cluster identity.

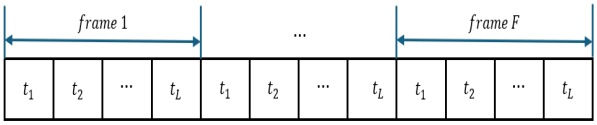

Fig. 2. Illustration of the frame structure for data transmission.

B. Transmission Principle

To accommodate data transmission in this network, the uplink transmissions are structured into frames and each frame is composed of L time-slots of equal duration, as shown in Fig. 2. Active[2] devices can transmit a packet once per frame and they are granted the autonomy to choose independently a random slot within the frame *with equal probability* for their data transmission. Although this flexibility can mitigate potential collisions among concurrent transmissions, collisions may nevertheless occur if multiple devices select the same slot in a frame and a collision may not be resolvable despite SIC.

Given that the behavior of devices for slot selection is random and independent, the number of devices selecting the same slot in a frame may vary from 0 to $\sum_{i=1}^{C} N_i$. When only a single packet occupies a time-lot in a frame, its successful decoding by the BS depends on the channel condition. Specifically, when the signal-to-noise ratio (SNR) is sufficiently large, the packet will be successfully received by the BS; otherwise, the transmission fails.

For multi-user detection, the BS follows the SIC principle to decode signals when multiple packets collide in the same time-slot. Typically, SIC employs an algorithm that proceeds by selecting to decode first the signal with the highest signal-to-interference-plus-noise ratio (SINR), where the interference corresponds to the signals transmitted by other packets in the same time-slot. When the SINR is larger than a given threshold[3], the signal is decoded and subtracted from the received signal aggregate in the same time-slot. The decoding process proceeds with the next signal with the highest SINR until no more signal can be decoded.

C. Introducing Access Control for Performance Optimization

Given a limited number of uplink resources, it is crucial for a BS to advise a mechanism that leads to an appropriate number of contending devices in order to maximize the number of successful packet transmissions. To this end, we investigate access control (with and without the help of RL), based on which the BS periodically broadcasts an access probability to

[2]A device that has at least one packet to transmit is regarded to be active.

[3]This threshold is 10 dB based on real-life experiments performed in [10].

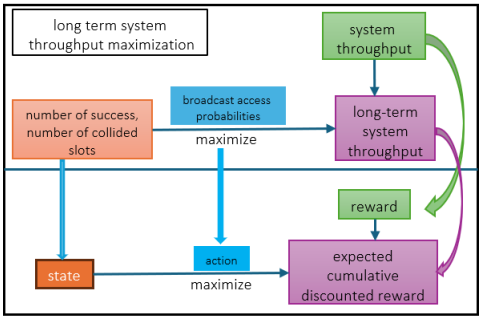

Fig. 3. The proposed RL-based access scheme for throughput maximization.

all devices. This access control mechanism follows the basic principle of the access class barring scheme proposed by the 3rd generation partnership project (3GPP) [11].

Let the BS periodically broadcast an access probability θ_i to all devices in cluster c_i according to its estimated traffic load and performance objective. Based on the received access probability, active devices follow the principle of immediate first transmission for medium access [3]. In the beginning of each frame, active devices decide whether to initiate a transmission or not. This decision-making process is guided by the broadcasted access probability. An active device in cluster c_i will proceed to select a time-slot in the current frame with probability θ_i, or it will differ its access until the next frame with probability $1-\theta_i$. No carrier sensing is performed before a packet transmission. This mechanism seeks to optimize the performance of devices for uplink NOMA-enabled transmissions by efficiently utilizing the available radio resources and minimizing collisions.

D. Performance Metric: Cluster and System Throughput

The performance of this network is evaluated in terms of cluster throughput and system throughput. The cluster throughput for cluster c_i, denoted by γ_i, is defined as the average number of packets successfully transmitted per frame. The system throughput, denoted by γ, is defined as the total number of packets successfully transmitted per frame across the entire network, and it is given by $\gamma = \sum_{i=1}^{C} \gamma_i$.

III. Proposed RL-based Access Scheme

In this section, we propose an RL-based access scheme tailored to NOMA-enabled uplink transmissions *where the BS functions as an agent* in the context of reinforcement learning. Based on the received reward, the agent takes actions on adjusting access probabilities that determine the fraction of active devices from each cluster that will transmit in a frame.

A. RL-based Access Scheme Overview

Empowered by RL capabilities, the BS is able to learn and adapt its strategies dynamically for performance optimization based on its interactions with the environment and the feedback it receives from the network. By leveraging RL, the BS can continuously refine its access probability decisions over time, ultimately leading to improved network efficiency and enhanced resource utilization. As shown in Fig. 3, the decision-making components include state, action, reward, and an algorithm for effective access control, as presented below.

1) State: A state $s_t \in \mathbb{S}$ is defined by a tuple of two concatenated components N_S and N_C, and it is expressed as

$$s_t = (N_S, N_C)\,. \tag{3}$$

N_S and N_C stand for the total count of successful transmission(s) with $N_S \in [0, 1, \ldots, \sum_{i=1}^{C} N_i]$ and the total count of collided slot(s) that have unresolved collisions[4] in a frame with $N_C \in [0, 1, \ldots, L]$, respectively. Accordingly, the set of possible combinations of states $\mathbb{S}$ can be expressed as:

$$\mathbb{S} = \left\{(0,0),(0,1),\ldots,(0,L);(1,0),\ldots,(1,L);\ldots;\left(\sum_{i=1}^{C} N_i, 0\right)\right\}. \tag{4}$$

2) Action: An action undertaken by the agent at time t, a_t, is an element selected from the action space $\mathbb{A}$ where $a_t \in \mathbb{A}$. a_t is described as a set of access probabilities to different clusters and it is defined as:

$$a_t = (\theta_1, \theta_2, \ldots, \theta_C)\,, \tag{5}$$

where θ_i $(i = 1, 2, \ldots, C)$ belongs to the interval $[0, 1]$. With a granularity level of ζ between two adjacent values, the set of possible action combinations $\mathbb{A}$ is expressed as:

$$\mathbb{A} = \left\{ \begin{matrix} [0,0,\ldots,0], & [0,0,\ldots,\zeta], & \ldots, & [0,0,\ldots,1], \\ [0,\zeta,\ldots,0], & [0,\zeta,\ldots,\zeta], & \ldots, & [0,\zeta,\ldots,1], \\ \vdots & \vdots & \ldots, & \vdots \\ [1,1,\ldots,0], & [1,1,\ldots,\zeta], & \ldots, & [1,1,\ldots,1] \end{matrix} \right\}. \tag{6}$$

By selecting an appropriate action, the BS can adjust the level of restrictions on the fraction of active devices in each cluster that may transmit in a frame.

3) Reward: The instantaneous reward $r_{t+1} \in \mathbb{R}$ in our scheme is formulated in terms of N_S and N_C observed at time $t+1$. This dual component provides the agent clear feedback about the outcome of the transmissions that occurred during the frame starting at time t. The reward r_{t+1} is defined as

$$r_{t+1} = N_S + P_o\,, \tag{7}$$

where P_o represents the payoff, which is expressed as:

$$P_o = \begin{cases} -\psi_1 & \text{if } N_C = L \\ \psi_1 & \text{if } 0 \le N_C < L \quad \text{and} \quad N_S > 0 \\ -\psi_2 & \text{if } 0 \le N_C < L \quad \text{and} \quad N_S = 0 \end{cases} \tag{8}$$

where ψ_1 and ψ_2 are integer constants. Typically, ψ_1 and ψ_2 should be configured as $\psi_1 < \psi_2$. *The objective of the proposed reward function is to maximize the number of packets that can be successfully decoded per frame.*

4) Exploration versus exploitation: In the realm of RL, an agent needs to strike a balance between exploring new options (exploration) and exploiting the current best-known option (exploitation) in order to optimize cumulative rewards over time. In this study, we embrace the concept of *epsilon* (ϵ)*-greedy exploration* to find this balance. In the initial phase,

[4]Unresolved collisions occur when two or more devices transmit in the same time-slot, and after following the SIC decoding algorithm one or more packets cannot be successfully received by the BS.

when exploration is prioritized at a high rate, we set ϵ to ϵ_{max}. As the learning process proceeds over time, we implement an *exponential update mechanism* for ϵ. (The definition of the update mechanism will be given in Sec. IV.) This mechanism systematically reduces the value of ϵ from ϵ_{max} to ϵ_{min}, guiding the agent towards more focused exploitation.

In order to achieve adaptable and effective access control, we incorporate the QL algorithm into our RL-based access scheme. The QL algorithm involves an iterative process where the BS updates its Q-values based on the observed states, actions, and associated rewards. The expected cumulative discounted reward, denoted as $Q^{\pi}(s_t, a_t)$, captures the essence of a state–action pair (s_t, a_t) under a given policy π. This expectation is formally expressed as:

$$Q^{\pi}(s_t, a_t) \triangleq E\left[R_t | s_t, a_t, \pi\right] , \tag{9}$$

where R_t is the long-term cumulative discounted reward and it is defined as

$$R_t = \sum_{q=0}^{\infty} \delta^q r_{t+q+1} , \tag{10}$$

where δ is a discount factor given as $\delta \in [0, 1]$.

The policy π establishes a mapping relationship from state s_t to action a_t. When applying QL for our policy exploration, an optimal policy π^* that maximizes the value of a state-action pair $Q^{\pi}(s, a)$ needs to be identified.

$$\pi^* = \arg\max_{\pi} Q^{\pi}(s, a) \quad \forall s, a . \tag{11}$$

For a given state–action pair (s_t, a_t) and the corresponding reward r_{t+1}, the value of a state-action pair undergoes an update based on the QL update rule [12]:

$$\begin{aligned} Q(s_t, a_t) \leftarrow\ & Q(s_t, a_t) \\ & + \alpha \left[r_{t+1} + \delta \max_{a_{t+1} \in \mathbb{A}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] , \end{aligned} \tag{12}$$

where $\alpha \in (0, 1]$ denotes the learning rate that governs the importance given to new information during a learning process.

IV. Performance Evaluation

In this section, we first present the simulation setup and then discuss the performance evaluation of the proposed scheme.

A. Simulation Configuration and Training

1) Simulation configuration: Consider a single-cell network depicted in Fig. 1, where the BS covers two distinct clusters ($C = 2$). The two clusters c_1 and c_2 are located at a distance of $d_1 = 450$ and $d_2 = 900$ meters away from the BS, respectively. Each cluster has a radius of 25 meters and devices inside the clusters are uniformly distributed. The BS is equipped with single antenna ($M = 1$) and the antenna height is configured as $g = 30$ meters. In this network, packets arrive to inactive devices following a Bernoulli distribution with an arrival rate λ ranging from 0.1 to 1. For other parameters used in our simulations, refer to [4] for details.

2) Access schemes for comparison: For system performance assessment, we configure a network as $N_1 = N_2 = N = \{2, 8\}$ devices per cluster and $L = \{1, 4\}$ slot(s) per frame. For packet decoding in the presence of concurrent transmissions, SIC is employed. If the transmission fails, the device retries in the subsequent frame until a successful transmission is achieved. To evaluate the performance of the proposed RL-based access scheme, two benchmark schemes (Schemes 1 and 2) are defined.

- **Scheme 1:** *Without access control (wac)*. With $\theta_i = 1$, *wac* devices transmit their packets without performing any channel or network status check. An active device initiates a transmission by randomly selecting a time-slot.
- **Scheme 2:** *Optimal access control (oac)*. For any given traffic load and network configuration, the *oac* scheme *exhaustively* searches among all possible combinations of θ_i where $\theta_i \in [0, \zeta, 2\zeta, \ldots, 1]$ to find out the optimal θ_i value for each cluster so that highest system throughput is achieved.
- **Scheme 3:** *RL-based access control (rac)*. With $\theta_i \in [0, \zeta, 2\zeta, \ldots, 1]$, optimal throughput in *rac* is achieved, by *dynamically adjusting the θ_i values by the BS based on the QL algorithm* presented above. In Scheme 3, the number of unresolvable collision(s) and the number of successful transmissions occurred in a frame, N_c and N_s, will affect the reward function, as specified in (8) and (7).

In both *oac* and *rac*, active devices follow the access principle presented in Sec. II for data transmission based on the latest θ_i value(s) they receive from the BS.

3) Training for RL-based access: The training stages of the QL-based algorithm are outlined in Algorithm 1, where the environment represents the network that is operated based on the proposed RL-based access scheme. During the training phase, a total number of E episodes are executed, with each episode comprising T training steps (frames). For each episode, the environment needs to be reset, where the values of (N_S, N_C) are configured as zero (i.e., $s_{t=0} = (0, 0)$). At each step, these values are updated based on the action $a_t = (\theta_1, \theta_2)$ taken by the BS through an ϵ-greedy policy. The selected (θ_1, θ_2) values are then broadcasted to devices within clusters.

The environment computes a reward (r_{t+1}) for the BS based on parameters such as (N_S, N_C). Devices achieving a successful transmission in c_i contribute to throughput γ_i calculation. The exponential update mechanism for ϵ used in our algorithm is defined as $\epsilon_{new} = \max(\epsilon_{cur} \times (\epsilon_{min}/\epsilon_{max})^{1/E}, \epsilon_{min})$, where ϵ_{new}, ϵ_{cur}, ϵ_{min}, and ϵ_{max} are the updated, current, minimum, and maximum value of the exploration rate, respectively. Moreover, we configure α and δ as 0.01 and 0.1; ϵ_{min} and ϵ_{max} as 0.001 and 1; ζ, ψ_1, and ψ_2 as 0.1, 10, and 20; E and T as 1000 and 10000, respectively.

B. Performance Evaluation: Single Slot Per Frame

Considering first a scenario with a single slot per frame, we present performance comparison with two device density levels as traffic intensity λ varies ranging from 0.1 to 1 packet per frame.

1) Performance comparison for $N = 2$ and $L = 1$: With this configuration, each cluster may accommodate a maximum

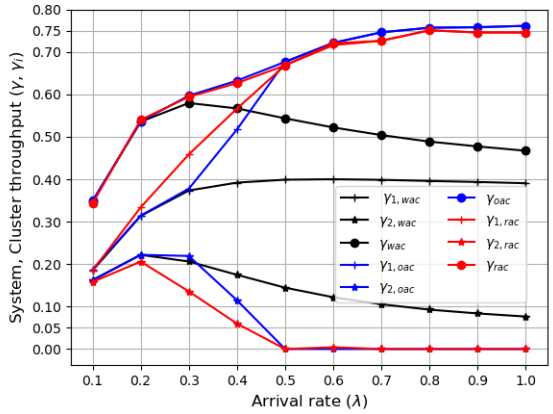

Fig. 4. System, cluster throughput with λ for $N = 2$, and $L = 1$.

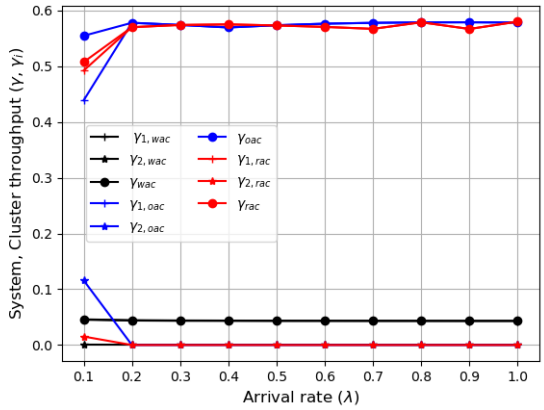

Fig. 5. System, cluster throughput with λ for $N = 8$, and $L = 1$.

Algorithm 1 Training algorithm for dynamic access control

1: Input: Episodes E, Steps per episode T, Time slots L, Learning rate α, Discount factor δ, and Set of access probabilities $\mathbb{A}$
2: Initialize QL agent
3: Set ϵ to ϵ_{max}
4: **for** episode in $1, 2, \ldots, E$ **do**
5: Reset the environment and observe $s_{t=0} = (0, 0)$
6: **for** step in $1, 2, \ldots, T$ **do**
7: Get current state s_t from the environment.
8: Choose a_t using ϵ-greedy strategy.
9: Take a step in the environment and obtain r_{t+1} using (7).
10: Update $Q(s_t, a_t)$ using (12).
11: Update s_{t+1}.
12: **end for**
13: Update ϵ following the exponential update mechanism.
14: **end for**

number of two devices. In other words, up to 4 devices may contend for a single radio resource for their data transmission.

As depicted in Fig. 4, both system and cell throughput ascend when traffic load λ is low and this trend applies to for all three schemes. However, a noticeable decline in system throughput is discerned with the *wac* scheme after a certain traffic load. This is due to the fact that as the ratio between the number of total active devices and the number of radio resource increases, more unresolvable collisions occur, particularly affecting devices farther away from the BS.

When the *oac* or *rac* scheme is employed, the BS optimizes system throughput by searching optimal θ_i values, either exhaustively (*oac*) or via QL (*rac*). As such, highest system throughput may be achieved at the expense of of imposing a lower access probability to devices in cluster c_2.

On the other hand, in the *rac* scheme, the BS aims to both maximize system throughput and at the same time minimize unresolvable collisions. Observe that, in the reward function defined in (7), a high number of unresolvable collisions is severely penalized. Then, as the number of active devices escalates beyond the total number of time-slots in a frame, the values of ψ_1 and ψ_2 are strategically selected such that the reward is increased. These selections lead to a higher access penalty being imposed to devices in cluster c_2 than the one imposed to these devices by the *oac* scheme.

2) Performance comparison for $N = 8$ and $L = 1$: With this configuration, the number of devices competing for the same single resource is much higher, leading to a congested network even at a very low arrival rate, e.g., when $\lambda = 0.1$. In other words, to maximize throughput, access control is beneficial even under low traffic loads.

As evident in Fig. 5, the system throughput of the proposed RL-based scheme is always higher than that of the conventional NOMA transmission. Furthermore, the system performance of the *oac* and *rac* schemes remains near-constant, exhibiting identical behavior across the range of λ.

However, distinctions prevail. The execution of the *oac* scheme relies on the knowledge of all system information including number of devices in each cluster and transmission outcomes. In contrast, the *rac* scheme does not need to acquire such information as it only observes the result of actions at states and adapts automatically its behavior according to the reward function.

Another phenomenon we observe in Fig. 5 is the unfairness problem when an access scheme is designed solely to *maximize system throughput*. As shown in the figure, the system throughput is (almost) identical as the cluster throughput of cluster c_1, whereas the contributions to system throughput from cluster c_2 are negligible. This result raises an unfairness concern when quality of service for all clusters needs to be addressed. In other words, when access fairness among different clusters is also a design goal, the reward function needs to be reformulated by taking fairness into account.

C. Performance Evaluation: Four Slots Per Frame

Let us now configure the number of time slots per frame as $L = 4$.

1) Performance comparison for $N = 2$ and $L = 4$: With this configuration, there are a sufficient number of radio resources for data transmission, even with the highest arrival rate as $\lambda = 1$. As shown in Fig. 6, both system and cluster throughput increase with λ, and the performance remains nearly identical across all three schemes. This result can be attributed to the fact that the average number of active devices per cluster is either less than or equal to the total available slots. Consequently, each active device receives minimal intra-

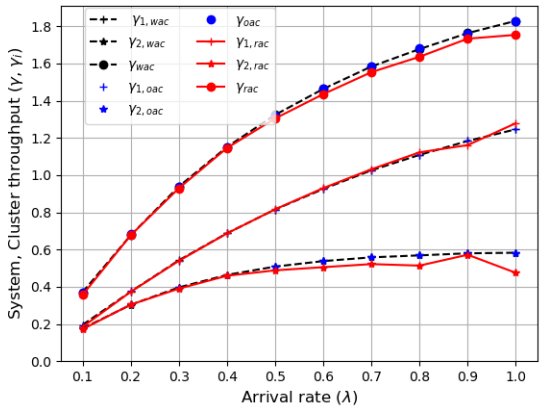

Fig. 6. System, cluster throughput with λ for $N = 2$, and $L = 4$.

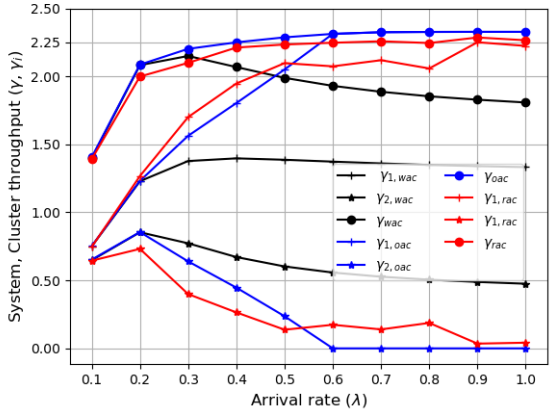

Fig. 7. System, cluster throughput with λ for $N = 8$, and $L = 4$.

and/or inter-cluster interference from other devices and SIC can resolve almost all collisions.

2) Performance comparison for $N = 8$ and $L = 4$: Finally, we observe in Fig. 7 the performance of the three schemes for a network with high device density and multiple radio resource. As depicted in this figure, the system throughput of *wac* exhibits an initial increase, then followed by a gradual descent already at a light traffic load (when $\lambda = 0.3$).

On the other hand, the other two schemes lead to much higher system throughput and much later saturation point as traffic load increases. Note that the system throughput achieved by the *rac* scheme is close to the optimal one achieved by the *oac* scheme. However, *rac* adopts a different access control policy than the one followed by *oac*. From low to medium load, the *rac* access control policy penalizes the access of cluster c_2 devices more than in the *oac* scheme. However, from medium to high load, an opposite behavior is observed.

V. Concluding Remarks

This paper presents a comprehensive study on NOMA enabled transmissions for IoT uplink traffic when devices form clusters and the BS deploys an RL-based access control scheme that aims at maximizing system throughput. To assess system performance when RL is deployed, we compare the RL-based access scheme with two benchmark schemes. While one of the benchmark schemes does not apply access control, the other one achieves optimal access control by exhaustive search. Through extensive simulations under various network configurations and traffic load conditions, we find that access control serves as an effective mechanism to maximize system throughput, particularly under heavy load conditions. Also, in scenarios where the deployment of access control leads to improved system performance, RL is able to find appropriate access control probabilities that attain a system throughput close to the optimal one. Moreover, we observe that the goal of maximizing system throughput is achieved at the expense of sacrificing access fairness among devices from distinct clusters. When access fairness among different clusters is regarded as an additional design goal, the reward function deployed by the RL algorithm must be redesigned.

Acknowledgment

The research leading to these results has received funding from the Norway (NO) Grants 2014-2021, under project contract no. 42/2021, RO-NO-2019-0499. The work of Jorge Martinez-Bauset was supported by Grant PID2021-123168NB-I00 under MCIN/AEI/10.13039/501100011033 and ERDF *A way of making Europe*.

References

[1] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and nonterrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 1117–1174, 2nd Quart., 2022.

[2] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to nonorthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[3] T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3192–3206, May 2021.

[4] A. Kumar, J. Martinez-Bauset, F. Y. Li, C. Florea and O. A. Dobre, "Understanding inter- and intra-cluster concurrent transmissions for IoT uplink traffic in MIMO-NOMA networks: A DTMC analysis," *IEEE Internet Things J.*, early access, Dec. 2023, doi: 10.1109/JIOT.2023.3341613.

[5] W. Ahsan, W. Yi, Z. Qin, Y. Liu and A. Nallanathan, "Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5083–5098, Aug. 2021.

[6] M. V. da Silva, S. Montejo-Sánchez, R. D. Souza, H. Alves and T. Abrão, "D2D assisted Q-learning random access for NOMA-based MTC networks," *IEEE Access*, vol. 10, pp. 30694–30706, Mar. 2022.

[7] S. K. Sharma and X. Wang, "Collaborative distributed Q-learning for RACH congestion minimization in cellular IoT networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, Apr. 2019.

[8] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6369–6379, Jul. 2020.

[9] J. Liu, Z. Shi, S. Zhang and N. Kato, "Distributed Q-learning aided uplink grant-free NOMA for massive machine-type communications," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2029–2041, Jul. 2021.

[10] Y. Qi, X. Zhang, and M. Vaezi, "Over-the-air implementation of NOMA: New experiments and future directions," *IEEE Access*, vol. 9, pp. 135828–135844, Sep. 2021.

[11] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the accurate performance evaluation of the LTEA random access procedure and the access class barring scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, Dec. 2017.

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