DRL-Based Scheduling With Support to Time-Varying Number of Active Users

Ingrid Nascimento, Silvia Lins and Aldebaro Klautau

***Abstract*— 5G use cases present current challenges that need to be addressed, such as big data generation, a large variety of services and devices. In this regard, Reinforcement learning (RL) is an important new tool for Radio Resource Scheduling. However, most works assume the number of users remains constant over time, which does not hold in realistic mobile network scenarios. In this work, a RL-based scheduler called RL-TANUS is evaluated in scenarios with diverse user traffic and a variable number of active users. Also, an analysis of the performance of RL-TANUS regarding throughput maximization is presented in comparison with scheduling baselines.**

***Keywords*— 5G, radio resource management, resource scheduling, reinforcement learning, deep learning.**

I. Introduction

Fifth generation (5G) communications embrace a variety of use cases such as unmanned vehicles, virtual reality, improved manufacturing operations, among others, which impose substantial challenges to mobile networks. 5G applications are often categorized as enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC), or massive Machine-Type Communications (mMTC), demanding more flexibility from the radio network to address the strict requirements in a cost-efficient manner. 5G New Radio (NR) standard improved radio flexibility by adding multi-numerology structure with different sub-carrier spacing (SCS), cyclic prefixes (CP), and symbol duration [1].

In addition, flexibility will be even more crucial with the emergency of new heterogeneous networks in future, which make Radio Resource Management (RRM) task more complex. Authors in [2] say that these 5G use-cases scenarios pose some challenges that RRM should handle for next wireless communications: (1) Massive connections between people and things will generate a huge demand over communication resources. (2) Big amount of information that requires efficient mining and big data exploitation in order to improve the design of resource management algorithms and, (3) Large amount of devices and services which impose stringent requirements related to Quality of Experience (QoE) and Quality of Service (QoS) which demand efficient RRM to provide appropriate resource to the target applications.

Artificial Inteligence (AI) is an ally to address previous mentioned challenges due to its capacity to improve resource efficiency, learn pattern from data and provide customized solutions [2]. Some recent work has shown the high applicability of Reinforcement Learning (RL) in the context of RRM, more specifically, in the Radio Resource Scheduling (RRS) tasks which is related to resource allocation among users, playing a fundamental role in the mobile cellular networks.

Due to the importance of RRS, RL is frequently considered to deal with scheduling tasks due to its self adaptivity to changing scenarios since conventional schedulers are unable to attend specific requirements of 5G applications considering its fixed metrics that do not easily adapt to different channel conditions, network operator demands and non-stationary traffic patterns [3].

Authors in [4] proposes a downlink scheduling framework which selects scheduling rules dynamically taking in consideration active users requirements and system conditions for QoS maximization. In [5], DRL is applied for packet scheduling and resource block allocation in the uplink service-oriented mmWave RAN in which multiple users and services are analyzed according to QoS requirements.

In order to add a fair evaluation of a RL-based scheduler, realistic scenarios should be provided with changing number of users during simulation under different traffic types, which is not commonly encountered in the literature. Then, there is a lack in the analysis of how a changing environment impacts the scheduler performance and its adaptability in the context of RRS tasks, which is under investigation in this work.

Our main contributions are related to (1) a scheduling RL-based strategy comparison with other scheduling baselines in efficiently allocate resources to users currently available in the scenario. (2) Also, this work provides a deeper evaluation of our Reinforcement Learning Time-varying Active Number of Users Scheduler (RL-TANUS), regarding agent behavior when different architectures for the RL-based scheduler are defined to work (3) in a changing scenario, with altering number of active users representing the mobility of users entering and leaving the base station (BS) coverage area. For the best of our knowledge, it is the first time that invalid action masking technique is investigated in the context of 5G RRS tasks. (4) In addition, the evaluation of the RL-TANUS scheduler is given considering different traffic patterns generated by the framework proposed in [6] that impacts environment dynamics.

II. System Model and Problem Statement

A. System description

We consider a downlink transmission with bandwidth of 100 MHz and using a carrier frequency $f_c = 60$ GHz, in

Ingrid Nascimento, Federal University of Pará, e-mail: ingrid.nascimento@itec.ufpa.br; Silvia Lins, Ericsson Research, Indaiatuba, Brazil, e-mail: silvia.lins@ericsson.com; Aldebaro Klautau, Federal University of Pará, Belém, Brazil, e-mail: aldebaro@ufpa.br. This work was supported in part by CAPES, CNPq and Ericsson Research.

which the total number of possible users is represented by $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$, and the BS serves a set of active users denoted by $\mathcal{U} = \{u | u \in \mathcal{N}\}$ of size U. The active users set size varies and it characterizes users entering and leaving the BS coverage area during the simulation execution generating a dynamic scenario.

The MIMO system is assumed as an analog architecture where, at the BS, there is a Uniform Planar Array (UPA) with N_t antenna elements, and each receiver uses an UPA with N_r antennas. Therefore, between the BS and a user, the MIMO channel is represented by an $\mathbf{H}$ matrix of $N_r \times N_t$. The codebooks are obtained from Discrete Fourier Transform (DFT) matrices and described by $\mathcal{C}_t = \{\bar{\mathbf{w}}_1, \cdots, \bar{\mathbf{w}}_{N_t}\}$ and $\mathcal{C}_r = \{\bar{\mathbf{f}}_1, \cdots, \bar{\mathbf{f}}_{N_r}\}$, being used at the transmitter and the receiver sides, respectively.

Regarding the beam selection modeling, $[p, q]$ is used to represent the chosen beam pair by a unique index $i \in \{1, 2, \cdots, M\}$, where $M \leq N_t N_r$. More detailed information about the communication system used can be found in [6].

The RL-TANUS scheduler aims to choose from the set of active users ($\mathcal{U}$), the User Equipment (UE) for packet transmission at each transmission time interval (TTI) t. There are three types of distinct users available in the set $\mathcal{N}$, which are a UAV, a CAR and pedestrians. The type of user from the set $\mathcal{U}$ determines the amount of data will be available for transmission, enabling in this way the differentiation of applications.

In addition, there are two different network load scenarios representing a heavy and light network traffic alternating among them in each 1000 steps. The heavy traffic represents a total throughput of 12 Gbps and light traffic is half of this value. Each user presents heavy and light traffic behavior. Also, users data traffic are defined as a Poisson processes with time-varying mean $\lambda_u[t]$ for a user $u \in \mathcal{U}$, more details about traffic generation can be found at [6].

The RL-TANUS would be able to deal with this constantly changing scenario in order to minimize packet loss and improve fair access of all users to resources.

B. Problem Statement

Our main objective is to find a policy that effectively schedules a user $u \in \mathcal{U}$ from a varying active users scenario such that maximize throughput and minimize packet loss for each scheduling period (TTI).

In our approach, all UEs have the same priority and must have data in the buffer to be transmitted. Only the selected UE can transmit and other active users have packets dropped or buffered. None previous selection stage of users is applied in order to explore adaptability of RL-TANUS scheduler in complex scenarios.

In a varying eligible users scenario, the main challenge is to provide consistent feature states as input to neural networks to indicate environment dynamics. Also, another constraint to be considered is the fixed input and output architecture of neural networks in the current ML implementation tools.

To deal with this complex problem, the RL-TANUS applies the invalid action masking approach to suitably learn a model able to select the appropriate user in order to maximize throughput with fair distribution of resources among active users in the scenario. The next section will highlight the details of the RL model design applied in the problem.

III. RL-TANUS Model

Figure 1 shows a generic scheme of the RL-TANUS scheduler and how it interacts with the radio environment. The problem is modeled as a Markov Decision Process (MDP) which consists of the following elements.

A. State space

The state space is composed by discrete values (from 0 to 2) that represents how many packets were transmitted $P_{\mathrm{t_x}}$, discarded P_{d} or buffered P_{b} for each active UE in each TTI t. These discrete numbers are defined according to a range of values that previous variables can assume. Also, the current choice of the RL-TANUS is added to the state space that will describe the environment dynamics. In general, the state space size will be of $1 + U^3$ which 1 represents the action taken by the scheduler.

B. Action space

At each TTI t, the RL-TANUS selects one user u from $\mathcal{U}$ set. In this regard, up to M users can be present in the scenario per TTI.

C. Reward

One of the metrics to evaluate the RL-TANUS is consider data that actually was transmitted by the scheduled user. The reward r at TTI t is the weighted sum of transmitted packages and discarded packages given by

$$r_{[\mathrm{t}]} = \frac{P_{\mathrm{t_x}[\mathrm{t}]} + 2P_{\mathrm{d}[\mathrm{t}]}}{P_{\mathrm{b}[\mathrm{t}]}}. \tag{1}$$

where $P_{\mathrm{t_x}[\mathrm{t}]}$, $P_{\mathrm{d}[\mathrm{t}]}$ and $P_{\mathrm{b}[\mathrm{t}]}$ corresponds, respectively, to transmitted, dropped and buffered packets of active users in the scenario at TTI t. In addition, penalties are applied when the RL scheduler selects same user consecutively.

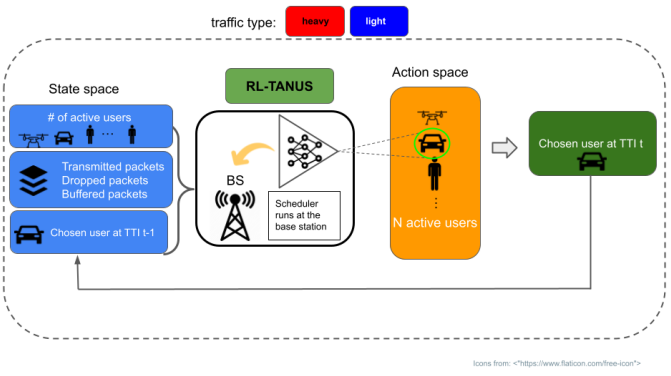

Fig. 1. Overview of the RL-TANUS. The environment can be under different traffic type.

IV. RL-TANUS Architectures

This section presents the RL-TANUS that is an actor-critic algorithm composed by two neural networks, one operates as critic which estimates the value function associated to each state encountered by the agent, and the another network is the actor or policy network, which updates the policy distribution in the direction suggested by the critic.

A. RL-TANUS-sorted-mask

In RL-TANUS-sorted-mask architecture we investigate the application of the invalid action masking technique in the context of RRS problem. The invalid action masking is usually applicable in games or contexts in which the discrete action space of different states usually have different sizes [7]. In our problem, M users compose the full discrete action space, however only a subset of this set, at each TTI, represents valid actions to be sampled by the scheduler.

In a policy gradient network, to differentiate valid actions, which represents active users in a given state, from the invalid ones, a mask is applied to "masking out" the logits of invalid actions. In this case, the logits of invalid actions are replaced by a large negative number, then when these values are submitted to a softmax layer to calculate a re-normalized probability distribution z, the resulting probability of choosing an inactive user will be virtually zero [7].

In addition, different arrangements of the state space will be tested in RL-TANUS architecture. In the sorted-mask architecture, the information of active users will be sorted and will be grouped in the top of neural network entry, as shown in Figure 2.

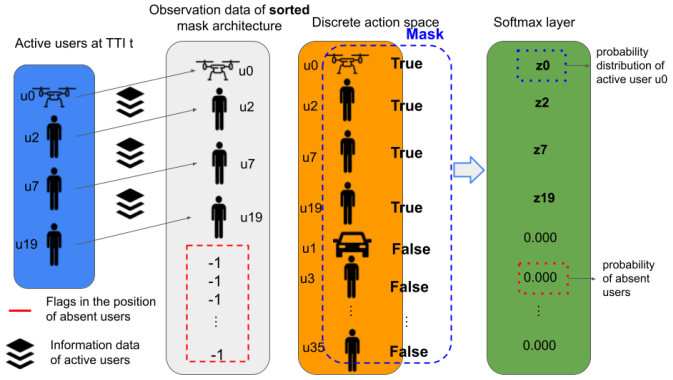

Fig. 2. RL-TANUS with sorted-mask architecture.

B. RL-TANUS-fixed-mask

In the RL-TANUS fixed-mask architecture is applied the same invalid action masking approach previously explained, then information of active users are placed in specific positions in the entry of the scheduler neural network, as shown in Figure 3.

The purpose of trying different types of architectures is the analysis of scheduler behavior for distinct configurations, testing if its learning efficiency and complexity is affected or improved.

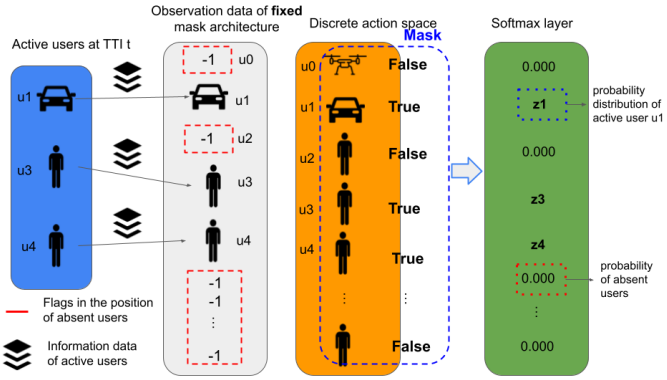

Fig. 3. RL-TANUS with fixed-mask architecture.

V. Experimental Results

In this section will be given some details about configurations of the agent scheduler and about training and testing process.

The data used in the training process is generated according to the framework described in [6]. A set of files containing the spatial information (position, orientation, acceleration, etc) of all moving objects (UAV, CAR and pedestrians) is used in the computation of the radio channels and parameters related to the telecommunication system, such as buffer size, etc as detailed in [6]. According to the amount of training episodes, the set of active users is determined in the beginning of the simulation for each TTI. In the test phase, a set of episodes not used during the training phase is applied to the agent for its performance evaluation.

A. Simulation settings

The RL-TANUS is composed by 2 critic network layers of 32 neurons each, as well as the actor network, composed by 2 layers of 32 neurons. The learning rate was set to 0.001 and ReLU activation function was applied. Different values for the number of steps to run in the environment per update, defined as n_{step}, impacted the results being considered $n_{step} = 25$ and $n_{step} = 30$ for sorted-mask and fixed-mask architecture, respectively.

For the experiments, we considered two additional schedulers, the B-Round Robin scheduler (B-RR) that chooses the user according to a sequential pattern (1-2-3-1-2-3,...), and the B-Random scheduler that selects a user randomly from the set of $\mathcal{U}$ active users. The simulation scenario presents $M = 5$ possible users.

For both architectures, 200 episodes were used in the training phase and the scheduler performance was tested for more than 100 thousand time steps. Figure 4 presents the accumulated rewards for the test dataset for RL-TANUS fixed-mask in comparison with baseline schedulers.

The results presented in Figure 4 can be further analyzed considering the maximum throughput achieved by the RL-TANUS fixed-mask in comparison with baselines, as can be seen in Figure 5. In this case, we can see that agent efficiently allocates users present in scenario in order to maximize

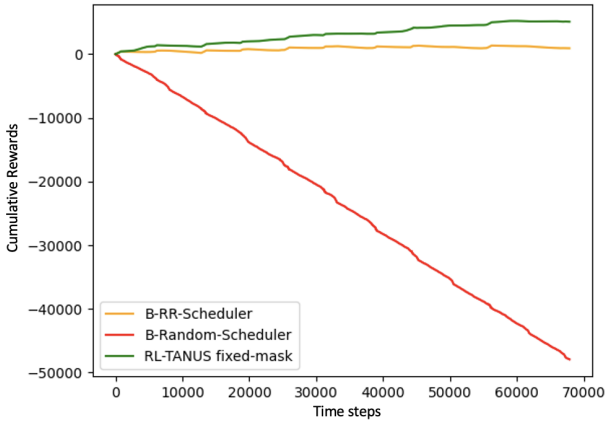

Fig. 4. RL-TANUS fixed-mask - cumulative rewards per time steps.

throughput. This result is directly related to a lower packet loss rate in all test episodes as shown in Figure 6.

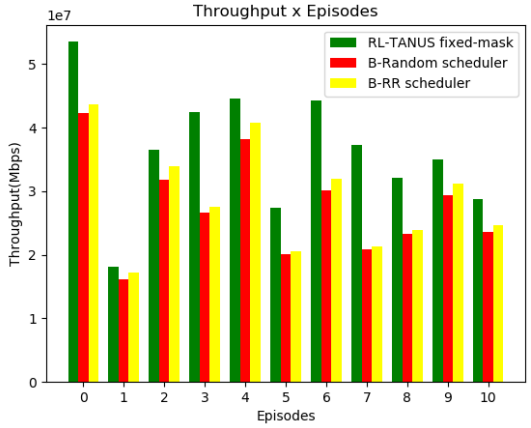

Fig. 5. RL-TANUS fixed-mask - Maximum throughput.

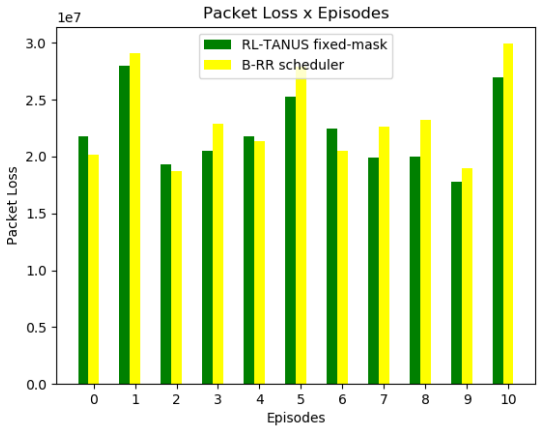

Fig. 6. RL-TANUS fixed-mask - Packet loss rate.

The performance of RL-TANUS sorted-mask architecture can be seen in Figure 7. Besides accumulated rewards, the agent achieved similar maximum throughput values to baselines as shown in Figure 8. Despite of better throughput achieved by RR baseline, it presents a higher packet loss rate in comparison to the agent as can be seen in Figure 9, which demonstrates that RL-TANUS sorted-mask presents greater goodput values.

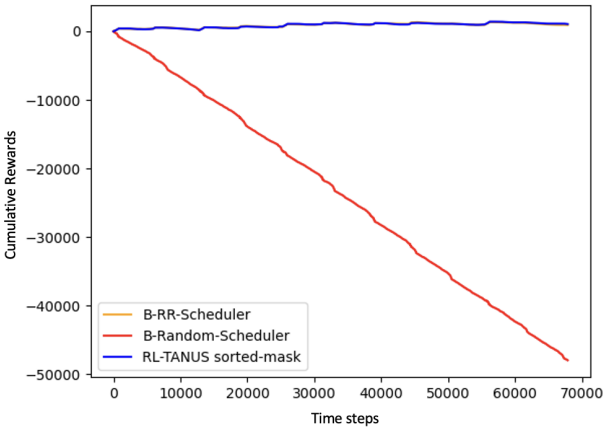

Fig. 7. RL-TANUS sorted-mask (that reaches the same results of the B-RR Scheduler) - Cumulative rewards per time steps.

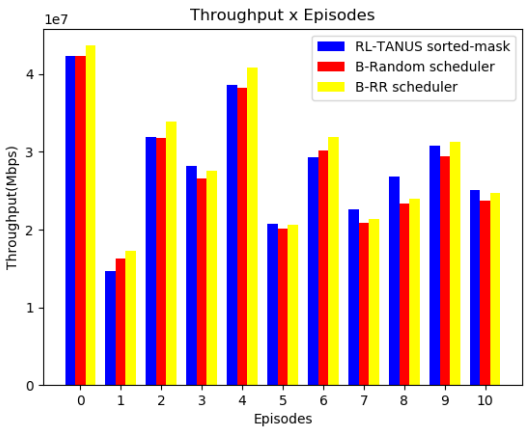

Fig. 8. RL-TANUS sorted-mask - Maximum throughput.

An additional analysis can be done regarding accumulated reward for both architectures. In Figure 10 can be seen that fixed-mask architecture presents a superior performance in all test episodes.

Its preeminence in relation to the sorted-mask architecture is demonstrated when throughput and packet loss values are analyzed in terms of users present in scenario during test phase. In Figure 11 we can see that fixed-mask achieved higher throughput values for UAV, as it presents bigger traffic than other users, and lower packet loss rate in comparison to the sorted-mask shown in Figure 12.

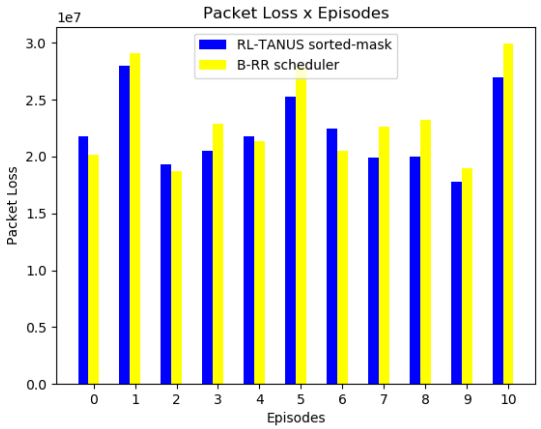

Fig. 9. RL-TANUS sorted-mask - Packet loss.

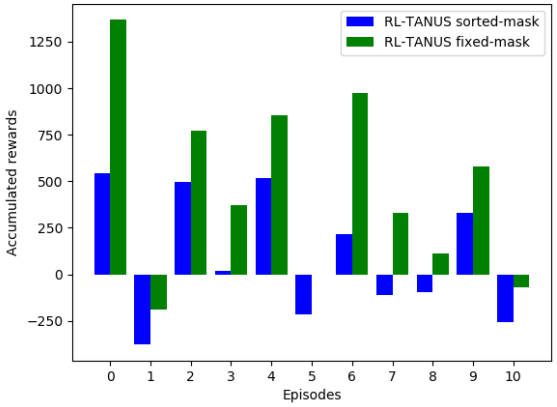

Fig. 10. Accumulated reward for sorted-mask and fixed-mask architectures during test phase.

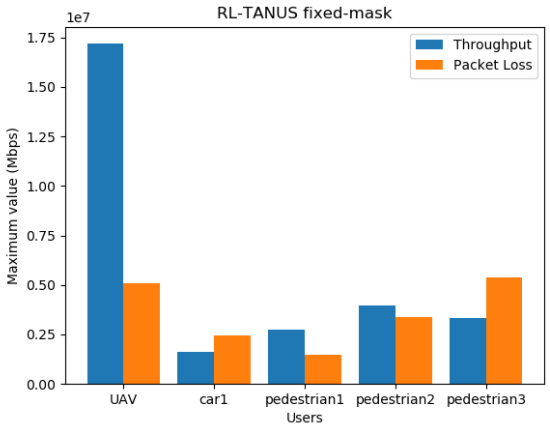

Fig. 11. Fixed-mask: Throughput and packet loss per users.

VI. Conclusions

In this work we introduced RL-TANUS scheduler, in the context of RRS tasks, in which the number of active users varies during the simulation presenting more realistic scenarios. Deeper analysis regarding some key performance indicators in the evaluated scenario were provided in order to bring useful insights about RL agent behavior. The experiments have shown that different definitions of RL-based scheduler architecture have significant impact in the agent performance, being the first time that invalid action masking technique is applied in the context of 5G RRS problem.

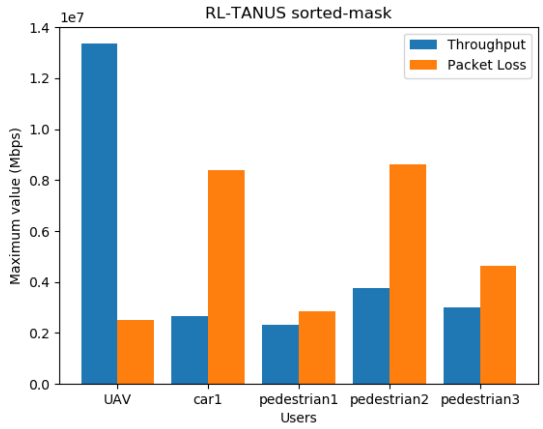

Fig. 12. Sorted-mask: Throughput and packet loss per users.

References

[1] Shao-Yu Lien, Shin-Lin Shieh, Yenming Huang, Borching Su, Yung-Lin Hsu, and Hung-Yu Wei. 5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access. *IEEE Communications Magazine*, 55(6):64–71, 2017.

[2] Mengting Lin and Youping Zhao. Artificial intelligence-empowered resource management for future wireless communications: A survey. *China Communications*, 17(3):58–77, 2020.

[3] Han Zhang, Wenzhong Li, Shaohua Gao, Xiaoliang Wang, and Baoliu Ye. Reles: A neural adaptive multipath scheduler based on deep reinforcement learning. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 1648–1656, 2019.

[4] Ioan-Sorin Comsa, Antonio De-Domenico, and Dimitri Ktenas. Qos-driven scheduling in 5g radio access networks - a reinforcement learning approach. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–7, 2017.

[5] Shuyi Shen, Ticao Zhang, Shiwen Mao, and Gee-Kung Chang. Drl-based channel and latency aware scheduling and resource allocation for multi-user millimeter-wave ran. In *2021 Optical Fiber Communications Conference and Exhibition (OFC)*, pages 1–3. IEEE, 2021.

[6] João Paulo Tavares Borges, Ailton Pinto De Oliveira, Felipe Henrique Bastos E Bastos, Daniel Takashi Né Do Nascimento Suzuki, Emerson Santos De Oliveira Junior, Lucas Matni Bezerra, Cleverson Veloso Nahum, Pedro dos Santos Batista, and Aldebaro Barreto Da Rocha Klautau Júnior. Reinforcement learning for scheduling and mimo beam selection using caviar simulations. In *2021 ITU Kaleidoscope: Connecting Physical and Virtual Worlds (ITU K)*, pages 1–7, 2021.

[7] Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. *arXiv preprint arXiv:2006.14171*, 2020.