Hierarchical Multi-Objective Deep Reinforcement Learning for Packet Duplication in Multi-Connectivity for URLLC

Qiyang Zhao[†], Stefano Paris[†], Teemu Veijalainen[‡], and Samad Ali[*]
Nokia Bell Labs, [†]Paris, France, [‡]Espoo, Finland, [*]Oulu, Finland
[*]University of Oulu, Oulu, Finland
Email: {qiyang.zhao, stefano.paris, teemu.veijalainen, samad.ali}@nokia-bell-labs.com, samad.ali@oulu.fi

***Abstract* — In this paper, machine learning solutions have been investigated to improve the decision of packet duplication in a multi-connectivity cellular network to optimize the satisfaction of delay and reliability in 5G. A multi-agent deep reinforcement learning scheme with sequential actor-critic model has been developed to improve the decision of packet duplication from observations of radio environment including channel state, interference and load. A multi-objective reward function has been developed to minimize the transmission delay, error rate and maximize satisfaction of the URLLC targets. System-level simulations have been carried out in a heterogeneous network by utilizing dual connectivity between macro and small cells. Our deep reinforcement learning scheme is shown to prioritize packet duplication to the UE where it gains from lower queueing and interference. Comparing with standard 5G multi-connectivity, it reduces the overall packet error rate and delay, with increased satisfaction rate of URLLC targets. Furthermore, it improves the network throughput and resource efficiency in dynamic user traffic with lower redundancy.**

Keywords – Multi-Connectivity, Packet Duplication, Deep Reinforcement Learning

I. INTRODUCTION

Dual-Connectivity (DC) has been introduced in 3GPP LTE system to enable the connection of a User Equipment (UE) to two Base Stations (BSs) in order to improve network capacity and throughput by splitting the data across two separate connections. 5G New Radio (NR) system has introduced the support of Ultra-Reliability Low-Latency Communications (URLLC) with a target of six nines (99.9999%) packet transmission reliability at a latency of 1 ms. This is used to support industrial and time sensitive applications [1]. In order to meet this target, 3GPP Release-15 proposes to exploit DC for packet duplication to improve reliably by utilizing spatial diversity. It also extends DC to four connections (named as legs), including two component carriers (CC) on each of the two BSs (named as gNBs) [2]. The gNB and UE can transmit up to four copies of the same packet payload. This can potentially reduce the packet error rate (PER) if secondary legs provide better channel quality and lower interference, or reduce packet delay if secondary cells have more resources and lower buffered traffic [3]. On the other hand, packet duplication generates redundant data transmissions which reduce the spectral efficiency of the 5G system [4]. In the situation where the secondary cell has limited resources or is highly loaded, packet duplication can significantly increase delay to the connected UEs. Moreover, transmitting duplicated packets can increase interference to the adjacent UEs allocated with the same resources, particularly when the density of gNBs and UEs is high. Therefore, to optimize reliability and delay for multiple UEs with the best spectral efficiency is challenging for using packet duplication in multi-connectivity. Since the network performance is affected by multi-leg transmission at each UE, we will have to deal with a combinatorial optimization problem if it is solved in a centralized manner. On the other hand, if the duplication decision is made locally at each UE, then a multi-agent optimization problem must be solved, which requires careful coordination between all the agents.

In this paper, we study the challenge on how to effectively utilize packet duplication to maximize the Quality of Service (QoS) satisfaction of delay and reliability of multiple UEs in the network, according to the dynamics of radio and traffic environment. In the Release-15 specification [5], the Cell-Range Extension (CRE) offset is used in HetNet to enable DC for UEs in an area where the Reference Signal Received Power (RSRP) differences of macro and small cells are within a certain range. Moreover, a DC range parameter has been introduced to control the amount of UEs using packet duplication. With this parameter, a UE can duplicate packets only if the following condition is satisfied:

$$RSRP_{macro} + DC_{Range} > RSRP_{small} + CRE \quad (1)$$

In this context, packet duplication is enabled to UEs only if the macro cell RSRP is high enough with respect to the small cell RSRP. This disables DC to UEs where the leg channel quality is poor, which reduces unnecessary duplication. An improved scheme is proposed in [6], which duplicates packet only when the first attempt of transmission on the primary leg fails, This is based on the assumption that primary leg is more reliable. However, these schemes have constraints in the following aspects: 1) The impacts of interference, traffic load, scheduled resources are not considered, which largely affects the QoS; 2) The QoS target changes with different services, and modeling the correlation between QoS and RSRP offset is difficult; 3) The channel and traffic fluctuate in time, where the threshold based scheme cannot capture the long term QoS impact of duplication; 4) The interactions of multiple UE transmission affects the network level QoS, which is not considered.

In order to cope with these problems, we propose a Deep Reinforcement Learning (DRL) [7] solution to decide packet duplication by observing the long term impact of signal strength, interference, traffic, and load with respect to delay and reliability. We design a sequential actor-critic deep neural network model that predicts the potential best decision policy of packet duplication, under the states that represents different system and environment conditions. A reward function of delay and reliability is defined with respect to the targeted value of URLLC services. The model is trained in the network by combining the rewards of multiple UEs taking actions at the

same time, such that the model maximizes the system-wide QoS rather than an individual UE.

The rest of this paper is organized as follows. Section II discusses the related work. Section III illustrates the system model of packet duplication in multi-connectivity. Section IV describes our deep reinforcement learning solution. Section V Section IV presents the system level simulation and discusses the results. We finally conclude this paper in Section VI.

II. Related Work

Machine learning (ML) has been extensively investigated to solve multi-agent interactive problems in wireless communications. Reinforcement Learning (RL) is recognized as an effective model-free approach to provide optimal decision policies from trial-and-error experience. Multi-Armed Bandit (MAB) is a simple distributed RL algorithm that selects action with largest utility and observes reward from the system [9]. The utility of each action is modeled as a random variable with unknown mean, and estimated using the upper bound of confidence interval computed from the observed outcomes of actions taken during time. MAB converges fast in a static scenario. However, the changes of environment will force the agent to explore other actions until sufficiently large iterations are performed to change the decision policy. This is unrealistic in radio network where the channel, traffic, mobility are constantly changing.

Conventional RL consider the problem as Markov Decision Process (MDP), which explores the optimal state transitions that leads to maximum accumulated rewards. Q-learning is a well-known RL, where a finite set of states is used to model the environment [6]. An action changes the system into a new state, and a Q value is computed for each action based on the accumulated reward of current state, discounted by the best action of the next state. In this context, Q-learning is an off-policy approach that searches for actions maximizing future rewards. However, the states for modeling radio environment can be large and continuous. The Q table is unable to capture complex state transitions, and the exploration is challenging.

Deep learning (DL) has been exploited as an effective approach to model complex environments. It can be used to predict the channel quality, traffic, mobility based on the past observations [10]. Deep neural network (DNN) is a well-known DL, that can use back propagation algorithm to optimize the hyper-parameters, which minimizes the loss of predicted outcome. However, it is difficult to capture all possible features affecting the QoS performance in radio network, because there are many underlying operations in lower layers, and interactive behavior from multiple users. To train a DNN model predicting the QoS requires a large number of radio parameter collection, and it is difficult to guarantee the accuracy due to the complex radio environment.

III. System Model

In this paper, we consider a HetNet scenario where UEs with DC can be connected to macro and small cells. The UEs located in the area satisfying RSRP criteria (1) is enabled with DC, which is shown in the grey area in Fig. 1.

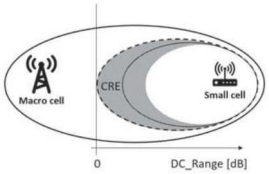

Fig. 1. Example DC range configuration (reproduced from [6])

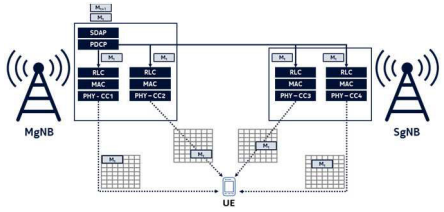

Fig. 2. PDCP duplication in Multi-Connectivty Archtiecture

The UE is connected to a Primary Cell (PCell) having the strongest RSRP. In downlink the Master gNB (MgNB) hosts the SDAP layer receiving packets from the core network and passing them to the PDCP layer, which controls the duplication of packets. The MgNB activates a set of Secondary Cells (SCell) hosted at a Secondary gNB (SgNB). The MgNB and SgNB are connected through the Xn interface, which is used to forward duplicated PDCP packets from the MgNB to the SgNB. A similar procedure exists in the uplink where the duplicated packets are transmitted by the UE to multiple cells. The general architecture for downlink transmissions is illustrated in Fig. 2.

IV. Deep Reinforcement Learning

We propose a ML agent implemented at the PDCP layer to decide which secondary legs to use to duplicate and transmit the packet for the UE within the DC range. Our ML agent predicts the impacts of channel and traffic conditions as well as interference and congestion on the QoS experienced by a packet transmission on the available legs. Based on the predicted QoS, the ML agent decide the best combination of legs to use to duplicate the PDCP packet.

A. Sequential Actor-Critic Model

The ML agent has the potential to model radio environment with DL and multi-user interactions through RL. We propose here a DRL model to combine states, actions, rewards obtained from multiple users to learn the joint optimal decision policy. A high-level architecture is illustrated in Fig. 3.

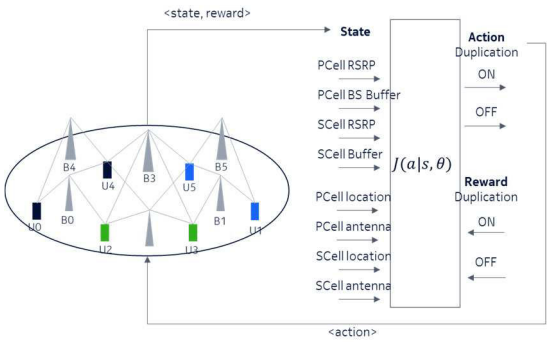

Fig. 3. Deep Reinforcement Learning for Packet Duplication

The DL model takes as input the radio states s_i of each cell. The state includes the RSRP measured at the UE, the buffered traffic in the cell, the location of the gNB, and the antenna direction of the leg. It generates the optimized probability of selecting an action by turning on or off packet duplication under the state condition $J(a|s,\theta)$, which is used by the PDCP layer to duplicate packets. When the packet is acknowledged by the receiver as delivered or lost, the model collects the delay D and error rate P and compute a reward from a function that defines the optimization objective. The delay measured at the PDCP layer includes both the packet queueing delay and the transmission time over physical channel. The error rate is reported by the UE, which is computed from the measured error rate of the received transport blocks during the packet transmission. For duplicated transmissions to UE u, the probability P_u is the defined as the joint error probability while the delay D_u is the minimum delay across multiple legs:

$$P_u = 1 - \prod_{l \in L_u} BLER_l \left(\frac{p_l g_l}{N_0}\right), \tag{2}$$

$$D_u = \min\left(\sum_{l \in L_u} \alpha_l \, log\left(1 + \frac{p_l g_l}{N_0}\right)\right)^{-1}, \tag{3}$$

where $p_l g_l$ are transmit power, channel/antenna gain of leg l, N_0 is thermal noise, α is the transmission efficiency which depends on the modulation and coding scheme selected by the link adaptation algorithm.

We observe that state variables like RSRP and load change during the packet transmission due to the dynamics of the radio channel and user traffic. Similarly, the gNB location and antenna direction relative to the UE position change when the UE is moving across different cells. Thus the state changes can be affected by the environment in addition to the action. In order to capture such behaviour, we use a sequential model to predict the variation of channel and traffic during next packet transmission, which improves accuracy of reward prediction.

A Long Short-Term Memory (LSTM) neural network is used for $J(\theta)$ that uses multiple observed states of current and x past packets to infer the decision for the next packet. The probability of taking an action on UE u for packet n is:

$$P(a)_{t,u} = J(a_{n,u}|s_{n,u}, \dots s_{n-x,u}, \theta) \tag{4}$$

In order to model the interaction between multiple UEs and learn the set of individual UE decisions that affects the network performance, we propose a hierarchical actor-critic RL. It contains a critic predicting the long-term accumulated reward of each action, and an actor predicting the optimal decision policy. The critic uses a value-based RL model $V(a_i|s_i;\theta_v)$ that indicates the impact of an individual UE action on the overall network performance. This is achieved by combining the rewards observed from multiple UEs sending their packets at the same TTI, with Stochastic Gradient Descend (SGD) applied to optimize θ_v, as follows:

$$\theta_v \leftarrow \theta_v - d \sum_{u_i \in u} \left(r_{s_i,a_i,u_i} - V(a_i|s_i;\theta_v)\right)^2. \tag{5}$$

This way the model maximizes the cumulative performance of the entire network rather than the performance of a single UE. For each packet transmission, the critic computes a time-difference (TD) value between the output of the value function $V(a_i|s_i;\theta_v)$ and the observed reward of a UE r_{s_i,a_i} as indicated in equation (6). This indicates the difference between UE individual performance and the network target

$$\sigma_{s_i,a_i} = r_{s_i,a_i} - V(a_i|s_i;\theta_v). \tag{6}$$

The actor uses policy RL model $J(a|s_i,\theta_j)$ to indicate the probability of duplicating a packet on a selected leg. The TD value is used to optimize θ_j, such that $J(a|s_i,\theta_j)$ converges to an action that maximizes the value function $V(a_i|s_i;\theta_v)$ and the long-term reward of the network. The optimization of the parameter θ_j is achieved though the policy gradient update as follows:

$$\theta_j = \theta_j + \sigma_{s_i,a_i} \log J(a|s_i,\theta_j). \tag{7}$$

The sequential actor-critic model can be implemented in a semi-centralized architecture. The network periodically collects a batch of samples composed of the triple {state, action, reward} from multiple gNBs and UEs and train the critic that predicts the network performance of each action. The TD values are sent to each gNB, which trains the actor. After that the gNB and UE can use the local actor to decide duplication for each packet, without communication with the network.

B. Multi-Objective Reward Function

The reward function is a key component of reinforcement learning since it defines the objective of the decision policy, which eventually steers the operational point of the controlled system. Conventional rewards are usually defined to optimize a single performance metric. However, in a 5G network there are multiple services from different users which require different level of delay and reliability targets. In this context, a novel multi-objective reward function has been designed to capture both the delay and reliability targets. The objective is to allow operators to easily tune the reward function, such that the ML model can optimize the performance in the targeted range. Indeed, delay and reliability are two completely different metrics. In packet duplication, a higher reliability (lower error probability) can be achieved by assigning more resources from multiple legs. However, this causes more redundant traffic to the cells, which increases the system load and eventually the delay experienced by data connections. The ML algorithm is supposed to maximize the satisfaction rate of delay and reliability target for all the users in the network. To this end, we developed negative log functions to map the metrics in configured targets on reward.

The reward of PER is defined as follows:

$$r_p = \begin{cases} 1 & P < P_1 \\ -k_p \log P & \text{otherwise} \end{cases} \tag{8}$$

P_1 is defined as the targeted PER of a 5G service, which enables the DRL to minimize PER towards the targeted level. It can be derived as follows

$$P_1 = \exp\left(-\left(k_p\right)^{-1}\right). \tag{9}$$

The reward of packet delay is defined as follows:

$$r_d = \begin{cases} 1 & D < D_1 \\ 0 & D > D_2 \\ -k_{d_1} \log k_{d_2} D & \text{otherwise} \end{cases} \tag{10}$$

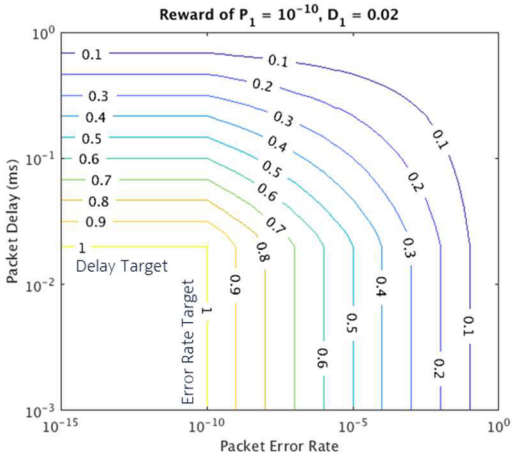

Fig. 4. Reward function of packet error rate and delay

D_1 is defined as the target delay of a 5G service, which enables the DRL to minimize delay towards the targeted level, when it is below the upper bound D_2. They can be derived as

$$\begin{cases} D_1 = \left(k_{d_2}\right)^{-1} \cdot \exp\left(-\left(k_{d_1}\right)^{-1}\right) \\ D_2 = \left(k_{d_2}\right)^{-1} \end{cases} \tag{11}$$

The final reward function combining PER and delay is as follows:

$$R = r_p \cdot r_d \,. \tag{12}$$

Fig. 4 illustrates the contour plot of the reward function (12) when target error rate and delay are set equal to $P_1 = 10^{-10}$ and $D_1 = 0.02$ ms, respectively. It can be observed that the reward stays at 1 when PER and delay is below P_1 and D_1, respectively. The reward decreases exponentially as each QoS metric increases to 1. This forces the DRL agent to select an action if it receives delay or error rate closer to the target and stops exploration when the target is achieved.

V. Performance Evaluation

In this section, a system level simulation of the proposed sequential actor-critic DRL with multi-objective reward function is demonstrated. The scenario is based on a HetNet defined in [11] and previously utilized in [5], [6]. It comprises 7 macro gNBs with 3 sectors each, and 4 clustered small cells per macro cell sector. The macro and small cell layers operate at separated frequency of 2 GHz and 3.5 GHz, respectively, with 10 MHz bandwidth for each. The cell selection of each UE is based on RSRP measurement, with a 15 dB CRE and 10 dB DC range to enable DC between macro and small cells. There are 288 UEs randomly placed in the converge area. The offered traffic to the network is FTP3, with an arrival rate of 100 packets/s of each UE in downlink. We change the packet size from 25 to 600 bytes to evaluate performance at different traffic levels. The key parameters are listed in Table I.

The network layout is illustrated in Fig. 5. Note that the UE location is changed randomly at each simulation round, and the performance metrics are averaged across multiple random drops of the UEs. The proposed solution is compared with standard where PDCP duplication is enabled for UE in DC range.

Table I. System Simualtion Parameters

Parameters	Values
Number of macro cells	7, with 3 sectors each
Number of small cell	84, with 4 cells each macro sector
Number of UEs	288, uniformly placed
Carrier frequency	Macro: 2 GHz, Small: 3.5 GHz
Bandwidth	Macro: 10 MHz, Small: 10 MHz
Traffic model	FTP, 100 packets/s, 25 – 600 Kbytes
Cell range extension	15 dB (small to macro)
Dual Connectivity range	10 dB (macro to small)
Simulation length	21000 TTIs by 10 rounds
Height	Macro: 32 m, Small: 10 m, UE: 1.5 m
Transmit power	Macro: 46 dBm, Small: 30 dBm
Channel model	Urban macro (UMa) and micro (UMi)
Subcarrier spacing	15 kHz

We first investigate the performance of PER and delay distributions on different users. The packet size is set to 600 bytes with UE traffic at 60 MB/s.

In the PER evaluation we set the lower bound PER $P_1 = \{10^{-10}, 10^{-5}\}$, which corresponds to $k_p = \{0.1, 0.2\}$ according to (9). Fig. 6 illustrates the cumulative density function (CDF) of the PER. It can be observed that the DRL schemes achieve significant lower PER than no duplication due to enhanced SINR from two legs, and lower PER than full duplication due to reduced transmission on low quality legs which causes interference. With reward configuration of $P_1 = 10^{-10}$, it

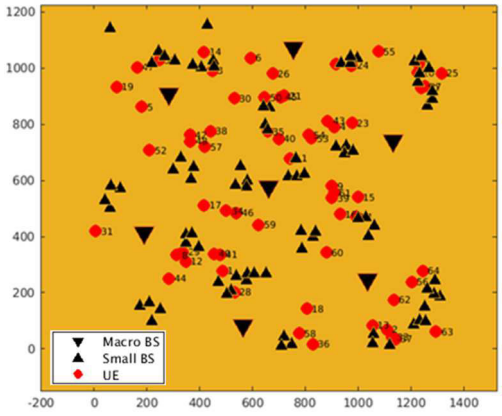

Fig. 5. Example heterogeneous network layout

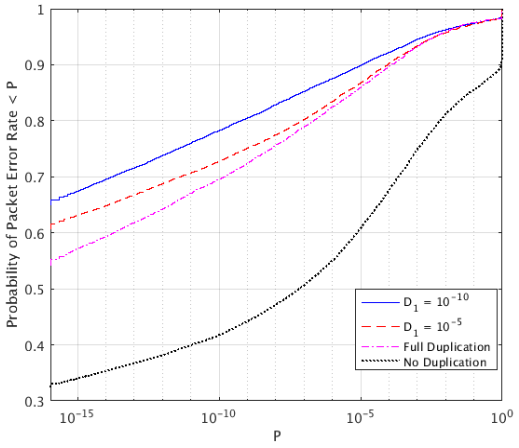

Fig. 6. Packet Error Rate Distribution

provides lower PER compared to $P_1 = 10^{-5}$. The DRL increases duplication to further reduce PER below 10^{-5}, such that more users are satisfied with lower PER.

In the delay evaluation, we set the lower bound of packet delay at $D_1 = \{0.5, 10\}$ ms and we fix the upper bound at $D_2 = 50$ ms. This corresponds to $k_{d_1} = \{0.5, 1.5\}$ and $k_{d_2} = 0.02$, according to (11). Fig. 7 shows the CDF of packet delay. It can be observed that the DRL schemes achieve lower delay than no duplication by gaining from a leg that has lower buffered data and higher SINR. It also reduces the delay above 1 ms than full duplication, which is achieved from reduced redundant traffic that creates queueing for other users. Moreover, the CDF of packets above 10 ms shows that the DRL with reward target of $D_1 = 0.5$ reduces the delay by half compared to $D_1 = 10$, and 10 times compared to no and full duplication.

We can conclude from the above analysis of the empirical distributions of the PER and delay that the proposed DRL scheme provides QoS gain compared to full and no duplication. With configured reward function, a higher PER satisfaction or lower delay level is achieved with respect to the targeted QoS.

We investigate the performance of DRL under various network traffic conditions and compare it against baseline schemes that enable or disable packet duplications for UEs in the DC range. The traffic is tuned by changing the packet size while keeping the same arrival rate, as detailed in TABLE I. The DRL scheme is evaluated with parameters $P_1 = \{10^{-10}, 10^{-5}\}$ and $D_1 = \{0.5, 10\}$ for the reward function.

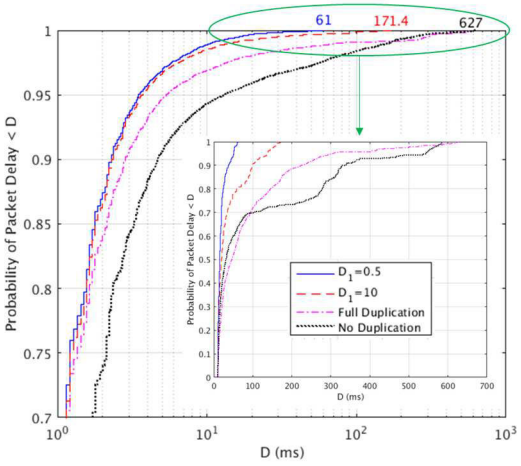

Fig. 7. Packet Delay Distribution

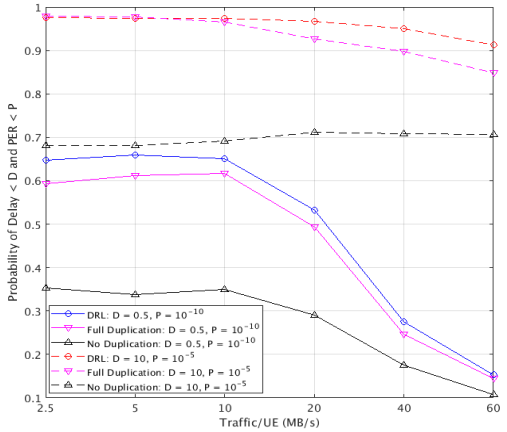

Fig. 8. Probaiblity of PER < P and Delay < D ms

Fig. 8 presents the percentage of user satisfaction of 2 sets of delay and PER target: (0.5 ms, 10^{-10}) and (0.5 ms, 10^{-5}). The DRL schemes provide significantly higher satisfaction than no duplication. With the target of delay at 0.5 ms and PER at 10^{-10}, it provides higher satisfaction than full duplication. On the other hand, with a relaxed target of delay at 10 ms and PER at 10^{-5}, a similar satisfaction probability is achieved. This shows that DRL further optimizes the ultra-low delay and PER over standard duplication in DC, by reducing interference and redundant load from unnecessary duplication that provides no gains to the performance.

Fig. 9 shows that our DRL schemes significantly reduce delay when the network traffic increases above 40 MB/s/UE. No duplication has higher delay over all because of limited resources and lower SINR provided in a single cell. The delay of full duplication is lower, but it increases significantly at higher traffic. This is because the redundant traffic in SCells causing higher queueing delay to the adjacent UEs, which overloads the network. On the other hand, the DRL scheme selects the packet to duplicate only when the SCell has lower buffered bytes, and the UE receives higher SINR. It gives the priority of duplication to the UEs that has delay gain from SCells, which avoids overloading the network.

The performance of network throughput measured by the rate of delivered packets is shown in Fig. 10. It can be observed that the DRL scheme maximizes the throughput under different traffic conditions. The throughput of full duplication largely decreases when traffic becomes higher than 20 MB/s/UE. The DRL schemes limit the throughput drop by duplicating only a subset of packets, thus reducing congestion. With lower delay and PER target configured, the throughput gets higher. This aligns with the delay performance in Fig. 9 that redundant traffic and interference is effectively controlled.

Fig. 11 presents the percentage of allocated resources for packet transmission. It can be observed that DRL decreases the resource usage assigned for duplicated packets when network traffic is above 20 MB/s/UE. This reduces redundant load in the

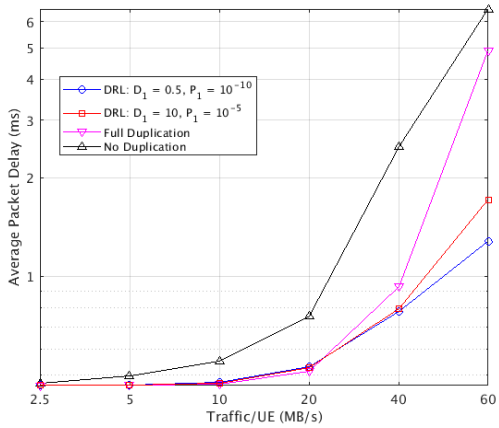

Fig. 9. Average Packet Delay

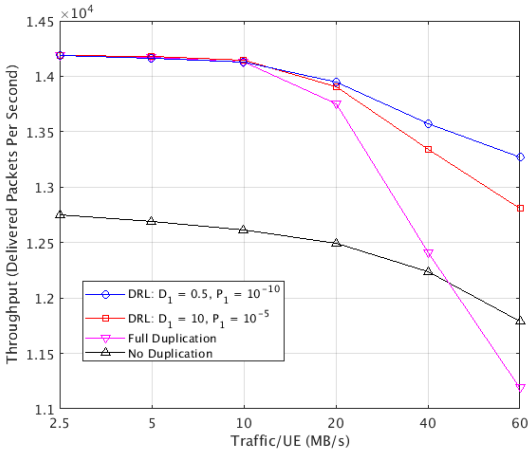

Fig. 10. Network throughput (delivered packets per second)

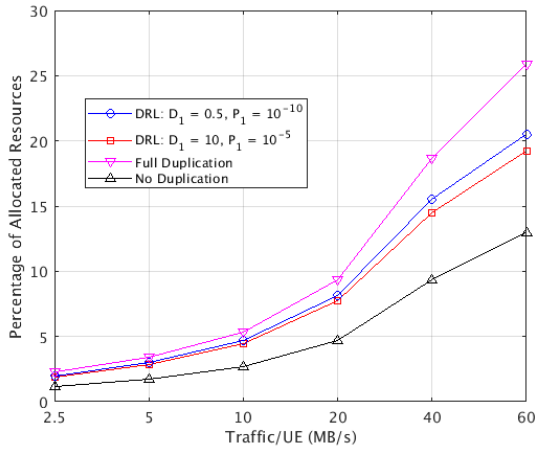

Fig. 11. Percentage of Allocated Resources

network, which permits to minimize delay and maximize throughput. The reward configuration with targeted delay 0.5 ms and PER 10^{-10} has 10% higher duplication probability than that with 10 ms and 10^{-5}. This is because DRL attempts to further reduce PER and delay by duplicating traffic at lightly loaded SCells, and it does not degrade network throughput as shown in Fig. 10. Furthermore, the DRL schemes is shown to reduce more than 25% to 33% redundant duplicated packets than full duplication, while providing higher delay and PER satisfaction probability as shown in Fig. 8. This proves that DRL largely improves the network resource efficiency with targeted QoS provided.

VI. Conclusion

In this paper, we have presented a deep reinforcement learning solution to autonomously optimize the decision of packet duplication in 5G multi-connectivity scenarios. An actor-critic reinforcement learning algorithm is proposed to coordinate multiple UEs duplication decisions to maximize long-term network level QoS performance. A sequential deep neural network is used to predict the optimal decisions by observing the state of the radio channel and traffic load of the cells where UEs are connected to. A multi-objective reward function is developed to maximize the delay and reliability satisfaction probability according to the targets of URLLC.

System-level simulation in a standard 5G HetNet shows that our proposed deep reinforcement learning solution largely reduces the overall packet error rate and delay at different traffic compared to standard dual connectivity. Furthermore, it largely improves network throughput and resource efficiency, with higher QoS target satisfaction provided. The solution can be applied for multiple services by configuring different QoS targets in the reward function, and for UE connecting to multiple base stations, component carriers and frequency bands.

Acknowledgement

This work was supported by the Academy of Finland 6Genesis Flagship (grant 318927), and in part by 5G-FORCE project.

References

[1] 3GPP, “NR; NR ad NG-RAN overall description; Stage 2 (Release 15),” TSG-RAN, Tech. Spec. 38.300 V15.5.0, Mar. 2019.

[2] 3GPP, “E-UTRA and NR; Multi-connectivity; Stage 2 (Release 15),” TSG-RAN, Tech. Spec. 37.340 V15.5.0, Mar. 2019.

[3] J. Rao and S. Vrzic, “Packet duplication for URLLC in 5G dual connectivity architecture,” in Proc. IEEE WCNC, Barcelona, Spain, Apr. 2018.

[4] N. Mahmood et al., “On the resource utilization of multi-connectivity transmission for URLLC services in 5G New Radio,” in Proc. IEEE WCNC, Marrakech, Morocco, Apr. 2019.

[5] N. Mahmood et al., “Reliability oriented dual connectivity for URLLC services in 5G New Radio,” in Proc. ISWCS, Lisbon, Portugal, Aug. 2018.

[6] M. Centenaro, D. Laselva, J. Steiner, K. Pedersen and P. Mogensen, "Resource-Efficient Dual Connectivity for Ultra-Reliable Low-Latency Communication," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020, pp. 1-5.

[7] R. Li et al., "Deep Reinforcement Learning for Resource Management in Network Slicing," in IEEE Access, vol. 6, pp. 74429-74441, 2018, doi: 10.1109/ACCESS.2018.2881964.

[8] R. S. Sutton, and A. G. Barto, “Reinforcement Learning: An Introduction”, (2nd ed.), The MIT Press, Cambridge, Massachusetts, London, England.

[9] Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R. and Auer, P., 2011, October. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In International Conference on Algorithmic Learning Theory (pp. 189-203). Springer, Berlin, Heidelberg.

[10] C. Zhang, P. Patras and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2224-2287, thirdquarter 2019, doi: 10.1109/COMST.2019.2904897.

[11] 3GPP, “Small cell enhancements for E-UTRA and E-UTRAN – physical layer aspects (Release 12),” TSG-RAN, Tech. Rep. 36.872 V12.1.0, Dec. 2013.