Dynamic Channel Reservation Strategy Based on DQN Algorithm for Multi-Service LEO Satellite Communication System

Zongwang Li, Zhuochen Xie, and Xuwen Liang

***Abstract*—In this letter, a dynamic channel reservation (DCR) strategy based on deep Q network (DQN) is proposed for multi-service low earth orbit (LEO) satellite communication system. We develop a novel modeling method to represent DCR problem of multiple services as a reinforcement learning (RL) task. Based on this model, we calculate the influence of current channel allocation results on future environment and take it as one of the factors for channel allocation decision. Moreover, a corresponding neural network is designed as a decision evaluator to provide an end to end mapping of decision to its value, which effectively avoids the influence of artificial preconditions. Simulation results show that the proposed strategy can improve the overall quality of service (QOS) of the system.**

***Index Terms*—Multi-service LEO satellite, dynamic channel reservation strategy, deep Q network.**

I. Introduction

THE INTEGRATION of the satellite communication system and 5G, complementing each other's strengths and weaknesses, together forming a seamless global integrated land, air, and space integrated communication network, which is an important direction for future communication's development [1]. In contrast to GEO systems, LEO satellites move quickly around the earth, resulting in frequent handing over among beams and satellites [2]. With the diversification of service types and the growth of traffic, channel reservation problem in LEO satellite system has become more complicated. On the one hand, it is necessary to consider the priority of different service types while reserving the channel. On the other hand, the effect of the current allocation result on the subsequent environment needs to be analyzed when allocating channels.

Channel reservation strategy is divided to fixed channel reservation (FCR) and DCR. Guaranteed handover strategy is first proposed in [3]. The number of reserved channels in this algorithm can't be changed dynamically, resulting in low channel utilization. In [4], a grey model based DCR strategy is proposed to dynamically adjust the channel reservation number by predicting whether the calls need to handover. But it mainly focuses on effective utilization of reserved channels in a single service in stead of multiple services.

In [5], a DCR strategy for multiple services is proposed. It sets up different access and reservation mechanisms according to the type of services, and each mechanism works based on adaptive probability. Genetic algorithms is used to calculate the optimal access threshold for users of different levels in [6] and [7]. Based on handover forecast, a multi-beam joint resource allocation scheme is proposed in [8]. However, those strategies rely on artificial prior conditions when calculating related parameters. For example, in [6], the access threshold of new calls for high-priority services is set to be greater than that of handover calls for low-priority services in every case, which limits the flexibility of the strategy and in turn reduces the system performance.

In addition, due to limited bandwidth and the fact that a call does not end immediately, the impact of current channel allocation result on future environment is more obvious as traffic increases. However, existing strategies pay less attention to such impact, and they usually focus on selecting the appropriate channel allocation result for the current call request, which limits the strategy's performance in the long run.

In this letter, we propose a DCR strategy based on DQN. For the first time, the deep reinforcement learning algorithm is applied to DCR problem of the multi-service LEO satellite system, which improves the overall QOS of the system. The main contributions of this letter can be summarized as the follows.

1) A novel modeling approach is proposed for DCR problem based on RL, we calculate the impact of current channel allocation result on future environment to maximize the long-term performance of the strategy.

2) A neural network-based decision evaluator is designed to provide an end-to-end mapping between decisions and their value, which avoids the restriction of strategy flexibility brought by artificial prior conditions.

3) We improved the methods of samples storage and parameters update in DQN according to the particularity of satellite communication. And the performance of our proposed strategy is verified through computer simulations.

II. System Model

A. Description of the Problem

In this letter, we mainly discuss how allocate channels for incoming calls rather than handover mechanism. From this perspective, switching between satellites is actually switching between different satellite beams. To simplify the complexity

Manuscript received September 24, 2020; revised November 8, 2020; accepted December 1, 2020. Date of publication December 8, 2020; date of current version April 9, 2021. This work was supported by the Natural Science Foundation of China under Grant 91738201. The associate editor coordinating the review of this article and approving it for publication was K. Adachi. *(Corresponding author: Zhuochen Xie.)*

Zongwang Li is with the Shanghai Engineering Center for Microsatellites, Chinese Academy of Sciences, Shanghai 201210, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lizw@microsate.com).

Zhuochen Xie and Xunwen Liang are with the Shanghai Engineering Center for Microsatellites, Chinese Academy of Sciences, Shanghai 201210, China (e-mail: xiezc.ac@hotmail.com; 18217631362@163.com).

Digital Object Identifier 10.1109/LWC.2020.3043073

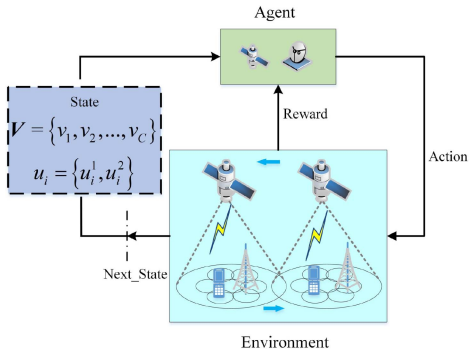

Fig. 1. RL model of DCR problem.

of channel reservation, we only discuss the beam handovers in this section. Assuming that a satellite has C beams, the total bandwidth of each beam is B, it can provide services for K service types, and the priority weight vector of the services is defined as $\boldsymbol{W} = [w_1, w_2, \ldots, w_K]$. We classify services according to the following characteristics [5]: $B_{\max}$, $B_{\min}$, B_{avg}, $T_{\max}$, $T_{\min}$, T_{avg}. Which refer to the maximum, minimum, average required bandwidth and the maximum, minimum, average occupation time in turn. We define the bandwidth utilization matrix (BUM) of the satellite as $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c, \ldots, \boldsymbol{v}_C\}$, where $\boldsymbol{v}_c$ is BUM of the beam c, $\boldsymbol{v}_c \in \mathbb{R}^{M_c \times 4}$, expressed in (1).

$$\boldsymbol{v}_c = \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ \cdots & \cdots & \cdots & \cdots \\ v_{m1} & v_{m2} & v_{m3} & v_{m4} \\ \cdots & \cdots & \cdots & \cdots \\ v_{M1} & v_{M2} & v_{M3} & v_{M4} \end{bmatrix} \tag{1}$$

where M_c is the number of calls being serviced in the beam c, $v_{m1} \in \{1, 2, 3, \ldots, K\}$ represents the service type (ST), $v_{m2} \in \{0, 1\}$ represents the call type (CT) (0 is the new call,1 is the handover call), v_{m3} is the occupied bandwidth and v_{m4} is the time the call has been serviced.

Let the current call request information be $\boldsymbol{u}_i = \{u_i^1, u_i^2\}$, where $u_i^1 \in \{1, 2, 3, \ldots, K\}$ represents the ST, $u_i^2 \in \{0, 1\}$ represents the CT. For each call request, the strategy has three processing decisions, namely access with maximum request bandwidth (AMARB), access with minimum request bandwidth (AMIRB) and refuse (R). We can process call request with decision AMIRB and decision R to reserve channels.

Obviously, satellite can not admit or reserve channels for every call request due to the limited bandwidth. We have to find an optimal strategy from the perspective of long-term profit. It is crucial to calculate the impact of current channel allocation result on future environment effectively. Hence, we try to solve this problem by RL theory.

B. Model of the Problem

As shown in Fig. 1, we formulate DCR problem as a task of RL. The model is mainly made up of agent, state variables *S*, action variables *A* and reward variables *R*, which aims to find an optimal dynamic channel reservation strategy with maximum long-term performance gain. The elements of this model are set as follows.

1) *Agent:* Agent is the access control center in the system. It may be located in the gateway of the ground station, or in the payload of the satellite. As the processing capacity of the satellites increases, the initial access delay will be lower.

2) *State Variables:* State is an abstract representation of the environment, which is the basis for agent to make decision. Therefore, state should contain two kinds of information, namely the usage of satellite's bandwidth resource BUM and the call request information $\boldsymbol{u}_i$. Let $\boldsymbol{V}_i$ be the BUM of the system when $\boldsymbol{u}_i$ arrives, then the state variable s_i can be defined as in (2).

$$s_i = \{\boldsymbol{V}_i, \boldsymbol{u}_i\} \tag{2}$$

3) *Action Variables:* Action is the channel allocation result of agent to current call request. According to the analysis in part A, there are three kinds of processing decisions, namely AMARB, AMIRB and R. We use a indicator parameter to represent action a_i for u_i, which is shown in (3).

$$a_i = \begin{cases} 0, & \text{AMARB} \\ 1, & \text{AMIRB} \\ 2, & \text{R} \end{cases} \tag{3}$$

4) *Reward Variables:* Reward is the feedback of environment to agent. It indicates which action is more favorable for a certain state when only considering current call request. Consequently, AMARB should obtain the highest reward, and AMIRB should obtain the second highest reward while R should obtain a negative reward. We set the original reward r_i^o of the action in (4).

$$r_i^o = \begin{cases} r_0^o, & a_i = 0 \\ r_1^o, & a_i = 1 \\ r_2^o, & a_i = 2 \end{cases} \tag{4}$$

where $r_0^o > r_1^o > 0 > r_2^o$. Moreover, ST and CT of $\boldsymbol{u}_i$ have different effects on the performance of strategy. The action reward r_i of $\boldsymbol{u}_i$ belongs to ST $= k$ is set in (5). α reflects the priority of handover connection relative to new connection, usually $\alpha > 0.5$.

$$r_i = \begin{cases} (1-\alpha) \cdot r_i^o \cdot \omega_k, & u_i^{n+1} = 0 \\ \alpha \cdot r_i^o \cdot \omega_k, & u_i^{n+1} = 1 \end{cases} \tag{5}$$

As discussed in part A, we can not process every call request with the action of highest rewards due to limited bandwidth. Therefore, our goal is to find a strategy that can maximize the cumulative reward. We define the value of the action a_i under the strategy π in (6) based on Bellman Equation.

$$\begin{aligned} q_\pi(s, a) &= E_\pi(r_i + \gamma r_{i+1} + \gamma^2 r_{i+2} + \cdots | s = s_i, a = a_i) \\ &= r_i + \gamma \sum_{s' \in S} P_{ss'} \sum_{a' \in A} \pi(a'|s') q_\pi(s', a') \end{aligned} \tag{6}$$

where, $\sum_{s' \in S} P_{ss'} \sum_{a' \in A} \pi(a'|s') q_\pi(s', a')$ is future reward that measures the impact of current channel allocation result on future environment, γ is a discount factor which is used to adjust the importance of future reward, $P_{ss'}$ is the probability of state s changing to state s', $\pi(s', a')$ is the probability of

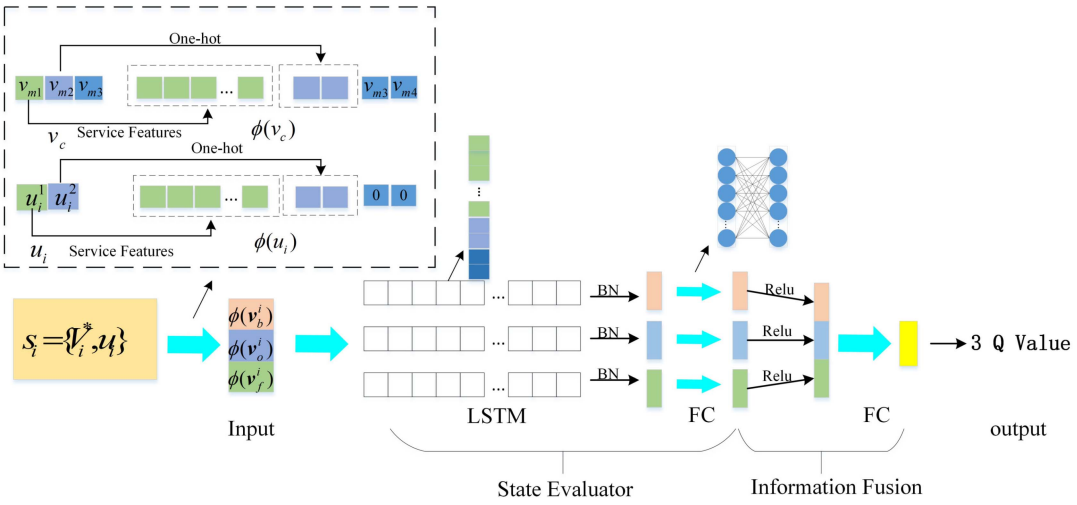

Fig. 2. Reformulation of state variables and structure of Q network.

selecting a' when state is s'. Let π_* be the optimal strategy, it should be defined as

$$\pi_*(a|s) = \begin{cases} 1, & a = \arg\max_{a \in A} q_*(s,a) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where, $q_*(s_i, a_i)$ is the maximum value of action can be expressed in (8).

$$q_*(s_i,a_i) = r_i + \gamma \sum_{s_{i+1} \in S} P_{s_i s_{i+1}} \max_{a_{i+1}} q_*(s_{i+1}, a_{i+1}) \tag{8}$$

We can acquire the optimal strategy through eq. (8). Unfortunately, the set of state variables is infinite. We can't get the optimal strategy through traditional RL algorithm due to huge amount of calculation. Moreove, it would limit the flexibility of strategy if we artificially divide these states into limited categories. DQN algorithm [9] can effectively solve such problems.

III. Dynamic Channel Reservation Strategy Based on DQN

In this section, a decision evaluation based on neural network is designed to provide an end to end mapping of channel allocation decision to its value. We improve DQN to find the optimal strategy according to the characteristic of satellite communication.

A. Decision Evaluator

A neural network called Q network in DQN is used to map the action in a certain state to its value, namely: $(s_i, a_i) \rightarrow Q(s_i, a_i|\theta)$, where θ is the parameter of the Q network. This mapping process is completely implemented by a neural network without any artificial restrictions. Thus, it is more flexible to make decision through comparing the Q value of optional actions.

In fact, for a state variable s_i, the information of BUM required only relates to the original beam and the surrounding beams of u_i. In order to avoid the additional complexity caused by user location, we simplify the surrounding beams into the forward beam to be entered and the backward beam which is opposite to the direction of the user relative to the satellite movement. The compressed BUM is expressed as $\boldsymbol{V}_i^* = \{\boldsymbol{v}_o^i, \boldsymbol{v}_f^i, \boldsymbol{v}_b^i\}$, where $\boldsymbol{v}_o^i, \boldsymbol{v}_f^i$, and $\boldsymbol{v}_b^i$ are the BUM of original beam, forward beam and backward beam of $\boldsymbol{u}_i$.

To enable the state variables to better reflect the communication environment, the BUM $\boldsymbol{V}_i^*$ and the connection information $\boldsymbol{u}_i$ are reformulated according to the service characteristics as shown in Fig. 2. The service type v_{m1} and u_i^1 are converted into service characteristic values, and the call type v_{m2} and u_i^2 are encoded by one-hot encoding. We set the occupied bandwidth and the time has been serviced of call request information $\boldsymbol{u}_i$ to be 0. During the evaluation, treat $\phi(\boldsymbol{u}_i)$ as an element in $\phi(\boldsymbol{v}_o^i)$. The reconstructed state variable can be expressed in (9).

$$\phi(s_i) = \left\{\phi\left(\boldsymbol{v}_o^i\right), \phi\left(\boldsymbol{v}_b^i\right), \phi\left(\boldsymbol{v}_f^i\right)\right\} \tag{9}$$

As shown in Fig. 2, the structure of the Q network is divided into two parts, the state evaluation (SE) and the information fusion (IF). The input of Q network is $\phi(s_i)$ with 3 elements while the output of Q network are 3 actions' Q value. In SE, we use 3 Long Short-Term Memory (LSTM) layers to analyze each element in $\phi(s_i)$ separately. In IF, we use 2 fully connected (FC) layers to fuse the analysis results and convert them to output.

B. Q Network Update and Application

The key to the proposed strategy is whether the decision evaluator based on the Q network is accurate. We update the parameters of Q network based on DQN algorithm.

There is a special scene when the remaining bandwidth in a beam is less than the minimum required bandwidth of the call request $\boldsymbol{u}_x$, the agent will directly reject the call request without decision evolution. For this reason, state s_x is defined as an invalid state. We ignore the invalid state when we store experience pairs $\{\phi(\boldsymbol{s}_i), a_i, r_i, \phi(\boldsymbol{s}_{i+1})\}$ in the memory pool D, where *null* is mark of state s_{i+1}.

$$null = \begin{cases} 1, & s_{i+1} \text{ is invalid state} \\ 0, & s_{i+1} \text{ is not invalid state} \end{cases} \tag{10}$$

As shown in (11), the loss function $L(\theta)$ is defined as the error of the target Q value and Q value, where the target value y_i of the state variable s_i is expressed in (12), $\hat{Q}$ is the target Q network which has same architecture as the Q network and its parameter is θ^-, η is a constant, used to measure the Q value of the invalid state.

$$L(\theta) = E\Big[(y_i - Q(\phi(s_i), a_i; \theta))^2\Big] \tag{11}$$

$$y_i = \begin{cases} r_i + \eta, & null = 1 \\ r_i + \gamma \cdot \max_{a_{i+1}} \hat{Q}\big(\phi(s_{i+1}), a_{i+1}; \theta^-\big), & null = 0 \end{cases} \tag{12}$$

The main implementation process of improved DQN is illustrated Algorithm 1, which is divided into the initialization stage and the training stage. The initialization stage includes the initialization of scene parameters and model parameters. In the training stage, by continuously updating the parameters of the Q network, we obtain an approximately optimal decision evaluator. In particular, during training state, when we make decision, $\varepsilon - greedy$ strategy is used to avoid local optimum.

In application, we select the decision with the maximum Q value according to the evaluation results. Due to the previous training process, network can provide an approximately optimal mapping relationship. We only needs to input the reconstructed state into the Q network for one forward calculation to get the Q value of actions, which reduces the calculation cost of the agent.

Algorithm 1 The Process of Improved DQN

Initialization

Initialize the bandwidth utilization matrix : $V \in \varnothing$
Initialize Q network with random parameter θ, target network $\hat{Q}$ with parameter $\theta = \theta^-$, memory pool $D \in \varnothing$

Training

for request step $i = 1 : N$ **do**
  Obverse $\boldsymbol{V}_i$ and get state variable s_i
  if s_i is invalid state **then**
    set $a_i = 2$
  else
    reformulate s_i into $\phi(s_i)$
    with probability ε select $a_i \in A$
    otherwise select $a_i = \arg\max_{a \in A} Q(\phi(s_i), a; \theta)$
    update V_i and obtain r_i
    get s_{i+1} and reformulate it into $\phi(s_{i+1})$
    if s_{i+1} is invalid state **then**
      set $null = 1$
    else
      set $null = 0$
    end if
    store $\{\phi(s_i), a_i, r_i, \phi(s_{i+1}), null\}$ in D
  end if
  sample a batch of experience tuples
  calculate error $L(\theta)$ through eq.(13) and eq.(14)
  perform a gradient descent step on $L(\theta)$ with respect to the network patameter θ
  update θ^- with $\theta^- = \theta$ every G step
end for

IV. Simulation Results

In this section, we present the simulation results of proposed strategy what we called DCR-DQN and compare it with no priority strategy (NPS), FCR strategy and probability based DCR strategy (PDR) [5].

A. Performance Evaluation Function

To measure the performance of channel reservation strategy, the overall quality of service P_{Qos1} is defined as the performance evaluation function [6], which is shown in (13).

$$P_{Qos1} = (1-\alpha)\sum_{k}^{K}\left(1 - P_b^k\right)\cdot\omega_k + \alpha\sum_{k}^{K}\left(1 - P_d^k\right)\cdot\omega_k \tag{13}$$

where P_b^k is the new connection blocking rate of service type k, P_d^k is the handover connection dropping rate of the service type k, $\omega_k \in \boldsymbol{W}$ is the priority weight of service type k.

It is obvious that P_{Qos1} only measures the performance of successful communication, and does not consider the users' demand for communication speed. Therefore, we develop P_{Qos2} to measure users' satisfaction with the communication speed, which is expressed in (14).

$$P_{Qos2} = \sum_{k}^{K} P_m^k \cdot \omega_k \tag{14}$$

where P_m^k is the proportion of calls belong to service type k that communicate with maximum request bandwidth.

TABLE I
Simulation Parameters Table

Altitude of the satellite	$H = 780km$
Number of satellites	$N = 5$
Number of beams of a satellite	$C = 48$
Width of spot beam	$L = 425km$
Maximum residence time of users in each spot beam	$57s$
Channel capacity	$B = 30Mbps$
New connection arrival $\left(\text{request} \cdot s^{-1}\right)$	$\lambda \in [0.01, 0.2]$

B. Simulation Environment

Assuming that users are uniformly distributed in each spot beam, new call request arrival rate of service type I and service type II follow an independent Poisson distribution, and the communication duration of various services follow a negative exponential distribution with an average value of T_m. Referring to [10], we simulate channel reservation scenarios for 5 LEO satellites based on the basic handover mechanism, that is, due to the movement of the satellite, the handover only occurs when the location of the call is not within the coverage of the beam or satellite. The parameters of simulation scene as shown in Table I. In addition, the relevant parameters of strategy are shown in Table II. In order to better compare the performance of the three strategies in the condition of multiple services, the value of each service feature is set in Table III.

Fig. 3 shows the simulation results of P_{Qos1}, which measures the ability of the strategy to ensure the success rate

TABLE II
STRATEGY PARAMETERS TABLE

Parameters	Value
Priority weights	$W = [0.8, 0.2]$, $\alpha = 0.8$
Original rewards	$r_0^o = 1$, $r_1^o = 0.5$, $r_2^o = -1$
Total time of satellite running	$T = 12h$
Discount factor	$\gamma = 0.99$
Learning rate	0.01
Memory pool capacity	50000
Batch size	128
Exploration factor	$\varepsilon \in [0.05, 0.9]$
Input size of each LSTM	10
Hidden size of each LSTM	16
input size of first FC	3×16
output size of first FC	8
input size of second FC	8
output size of second FC	3

TABLE III
SERVICE FEATURE TABLE

Type	$B_{\text{avg}}/kb \cdot s^{-1}$	$B_{\min}/kb \cdot s^{-1}$	$B_{\max}/kb \cdot s^{-1}$	T_m/s	$T_{\min}/s$	$T_{\max}/s$
Service I-1	30	30	30	180	60	6000
Service I-2	256	256	256	300	60	1800
Service I-3	3000	1000	6000	600	300	18000
Service II-1	10	5	20	30	10	120
Service II-2	256	64	512	180	30	360
Service II-3	5000	1000	10000	120	30	1200

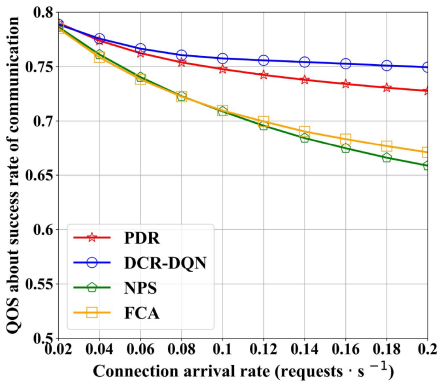

Fig. 3. The overall quality of service P_{Qos1}.

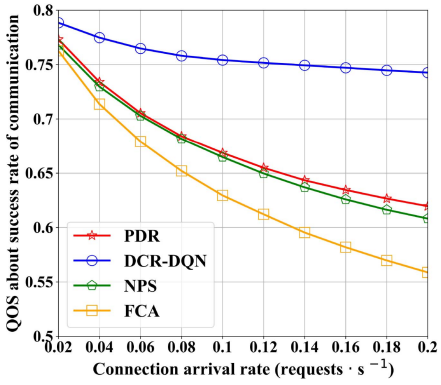

Fig. 4. The overall quality of service P_{Qos2}.

of communication from the perspective of the overall user. Fig. 4 shows the simulation results of P_{Qos2}, which measures the user satisfaction with communication speed. We can easily note that DCR-DQN achieves better performance both in P_{Qos1} and P_{Qos2} over other strategies especially in high connection arrival rate. NPS has a worst performance in P_{Qos1} since it does not reserve channels for higher priority calls. Due to the inability to dynamically adjust the reserved channels, FCR (10% of total channels reserved) has a worst performance in P_{Qos2}. When the connection arrival rate is low, the bandwidth resources are sufficient, and the performance of PDR and DCR-DQN is similar. However, as the user arrival rate increases, the performance of DCR-DQN is gradually better than that of PDR. This improvement is due to the fact that DCR-DQN proves the effectiveness of our proposed strategy. It is noted that the channel reservation of DCR-DQN is more flexible with end to end decision evaluator.

V. CONCLUSION

In this letter, we propose a dynamic channel reservation strategy based on improved DQN for multi-service DCR problem in LEO satellite communication system, which improves the overall quality of service. A novel modeling approach is proposed and an end to end mapping is established. Simulation results proved the effectiveness of proposed strategy. In the future, we will study more complex application scenarios such as how to choose satellite to handover under multi-satellite coverage. We hope our work will stimulate the further studies on resource allocation approaches for multi-service satellite.

REFERENCES

[1] "Solutions for NR to support non-terrestrial networks (NTN)," 3GPP, Sophia Antipolis, France, Rep. TS 38.821, Jan. 2020. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/38821.htm

[2] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[3] G. Maral, J. Restrepo, E. del Re, R. Fantacci, and G. Giambene, "Performance analysis for a guaranteed handover service in an LEO constellation with a 'satellite-fixed cell' system," *IEEE Trans. Veh. Technol.*, vol. 47, no. 4, pp. 1200–1214, Nov. 1998.

[4] Q. Zou and L. Zhu, "Dynamic channel allocation strategy of satellite communication systems based on grey prediction," in *Proc. Int. Symp. Netw. Comput. Commun. (ISNCC)*, 2019, pp. 1–5.

[5] H. Fei, Z. Li-Dong, and W. Shi-Qi, "A novel probability-based handoff strategy for multimedia LEO satellite communications," *J. Electron. Sci. Technol.*, vol. 5, no. 1, pp. 7–12, 2007.

[6] J. Zhou, X. Ye, Y. Pan, F. Xiao, and L. Sun, "Dynamic channel reservation scheme based on priorities in LEO satellite systems," *J. Syst. Eng. Electron.*, vol. 26, no. 1, pp. 1–9, Feb. 2015.

[7] J. Wang, L. Sun, J. Zhou, and C. Han, "A dynamic channel reservation strategy based on priorities of multi-traffic and multi-user in LEO satellite networks," *J. Circuits Syst. Comput.*, vol. 29, no. 05, 2020, Art. no. 2050082. [Online]. Available: https://doi.org/10.1142/S0218126620500826

[8] Y. Li, S. Wang, and W. Zhou, "A novel dynamic resource optimization method in LEO-MSS downlink with multi-service based on handover forecasting," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, 2019, pp. 809–814.

[9] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[10] I. D. Moscholios, V. G. Vassilakis, N. C. Sagias, and M. D. Logothetis, "On channel sharing policies in LEO mobile satellite systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1628–1640, Aug. 2018.