Load-Aware Satellite Handover Strategy Based on Multi-Agent Reinforcement Learning

Shuxin He*, Tianyu Wang*†, and Shaowei Wang*
*School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
†National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
Email: MG1723062@smail.nju.edu.cn, {tianyu.alex.wang, wangsw}@nju.edu.cn

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322449

***Abstract*—Low Earth orbit (LEO) satellites play an important role to realize personal global communication in future mobile communication networks, where terrestrial users can be covered by multiple satellites due to densely deployed satellites in the constellation. Since the speed of LEO satellites is much higher than that of mobile users, it yields a large amount of satellite handovers, which causes heavy signaling overhead. Also, terrestrial users need to compete for satellite channels while they can only obtain partial information of the satellite system from their individual views. Thus, a distributed satellite handover strategy is required to balance satellite load to avoid network congestion, while at the same time, maintain low signalling overhead. In this paper, we propose a novel satellite handover strategy based on multi-agent reinforcement learning that aims to minimize average satellite handovers while satisfying the load constraint of each satellite. Simulation results show that the proposed strategy outperforms the local handover strategies based on basic criteria in terms of average satellite handover and user blocking rate.**

***Index Terms*—Channel utilization efficiency, low earth orbit satellite network, multi-agent reinforcement learning, satellite handover.**

I. Introduction

In recent years, the increasing demand for high-data-rate and real-time applications of mobile users has accelerated the development of mobile communication, and massive capacity and connectivity of the network are required. It has been widely recognized that low Earth orbit (LEO) satellite networks shall provide high quality services with global coverage in the future mobile communication systems. LEO satellite constellation typically operates at the altitude of 500-1500 km with the period of less than 2 hours [1]. In contrast to geostationary earth orbit satellite communication, LEO satellite communication has the advantages of low propagation delay and low energy consumption [2].

However, LEO satellite networks have disadvantage in mobility management compared with geostationary satellite systems [3]. Specifically, since the satellites rotation speed is much higher than typical user speed, terrestrial users have to be handed over to other satellites multiple times within a call duration. A satellite handover is performed when the serving satellite is below a minimum elevation angle relative for the corresponding user and the connection is transferred to another visible satellite. The handover process has a significant impact on the communication quality.

This work was partially supported by the National Natural Science Foundation of China (61801208,61671233,61931023,U1936202), the Jiangsu Science Foundation (BK20170650), and the open research fund of National Mobile Communications Research Laboratory (2019D02).

The basic satellite handover criteria for selecting the next serving satellite includes, the remaining visible time, available satellite channels and elevation angle, which affect the number of handovers, network load and the service quality, respectively. The number of available satellite channels is applied as the basic criterion in [6, 7]. The methods of handover requests estimation and handover queuing have been developed to reserve resources to ensure low drop blocking probability and forced termination probability. In [6], the multimedia traffic is differentiated into two types and then the satellite handover requests are addressed based on the queue state of each traffic type. In [7], dynamic Doppler-based handover prioritization scheme is proposed, which utilizes Doppler shift monitoring to estimate the actual number of handover request and the actual time of occurrence to overcome the protracted reservation of resources.

The elevation angle is applied as the basic criterion in [8, 9]. In [8], the authors propose a novel hard handover scheme based on the dual satellite diversity (the common area between contiguous satellites), which uses two different elevation angle thresholds to select the highest and the second highest satellites to reduce the signalling overhead. In [9], a hybrid channel adaptive algorithm is proposed based on the concept of normal and critical conditions and only makes use of dual satellite diversity under critical channel conditions.

Some other studies consider two or more criteria for satellite handover. In [10], the three handover criteria are considered for the satellite selection of the new calls and handover calls, which provides users with low forced termination probability while fulfilling the QoS limitations even for heavy traffic conditions. In [11], a graph-based satellite handover framework is proposed for LEO satellite communication networks, and then the satellite handover optimization problem is formulated as the longest/shortest path of the graph for a specific user.

Most of the studies either consider only one handover criterion for a specific optimization goal, or give an overall solution taking three criteria into consideration from the perspective of a single user. However, users can only obtain partial information of the satellite system relative to themselves without a central controller. Besides, since the channel budget of a satellite is limited, the competition for available channels between users covered by the same satellite may cause highly imbalanced satellite load. Therefore, it is essential to design a distributed satellite handover strategy which takes into account the real-time resource competition of the users.

Multi-agent reinforcement learning (MARL) enables a group of agents to learn by interacting with their dynamic environment [12]. The agents observe the current state, take joint actions based on partial view of the system and get the relative reward [13]. In this paper, we propose a novel satellite handover strategy based on MARL to optimize the average satellite handovers and reduce blocking rate of users in a LEO satellite network. The main contributions are summarized as follows:

- We introduce the LEO satellite system model and formulate an optimization problem aiming to minimize the satellite handovers while satisfying the load constraint of each satellite. The handover minimization problem is then translated into an MARL-based problem from the learning perspective.
- A multi-agent Q-learning algorithm is developed based on the users' real-time knowledge of satellite network, in which a combined Boltzman exploration and ϵ-greedy action selection policy is proposed.
- The performance of our proposed MARL-based satellite handover strategy is analyzed and simulation results show that the proposed strategy achieves superior performance in terms of the average number of handovers and user blocking rate, as compared to the existing strategies developed by basic handover criteria.

The rest of the paper is organized as follows. In Section II, we introduce the LEO satellite system model and formulate an optimization problem for average handovers minimization, which is then translated into an MARL-based problem. In Section III, we solve the problem by using a multi-agent Q-learning algorithm. Simulation results are analyzed in Section IV and conclusions are drawn in Section V.

II. System Model

A. LEO Satellite Handover

We consider the satellite handover problem during a specific time period T, as shown in Fig. 1. The set of satellites is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$. K users, the set of which is denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$, are uniformly distributed on the surface of Earth. The elevation angle between user k and satellite n, denoted by $\theta_{k,n}$, can be acquired based on the location information of the users and their covering satellites. The minimum elevation angle, denoted by θ_0, is a limitation to guarantee the link quality of user communication. Therefore, the serving elevation angle constraint are given by

$$\theta_{k,n} \geq \theta_0, \forall k \in \mathcal{K}, \forall n \in \mathcal{N}. \tag{1}$$

The coverage of N satellites for K users can be viewed as a coverage graph. The covering time periods of the satellites are overlapped. When a satellite moves away from or enters the vision field of a user in the system, the satellite coverage graph changes to a new state. Therefore, the whole considered period T can be divided into U different coverage sections, denoted by $[t_0, t_1), [t_1, t_2), \ldots, [t_u, t_{u+1}), \ldots, (t_{U-1}, t_U]$. The

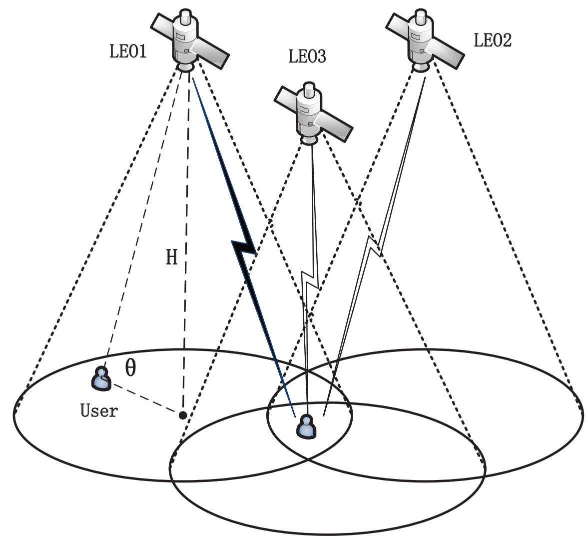

Fig. 1. A satellite handover scenario

intersection points, denoted by $\{t_1, t_2, t_3, \ldots, t_{U-1}\}$, represents the state changing times of satellite coverage. During each coverage section, the satellite coverage graph remains unchanged.

Since each user is covered by more than one satellite during each coverage section, we introduce $c_{k,n}^u$ as the coverage indicator between satellite n and user k during $[t_u, t_{u+1})$, which is defined as

$$c_{k,n}^u = \begin{cases} 1 & \text{user } k \text{ is covered by satellite } n, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The covering satellites set of user k during $[t_u, t_{u+1})$ can be given by

$$\mathcal{I}_k^u = \{n \mid c_{k,n}^u = 1, n \in \mathcal{N}\}, \forall k \in \mathcal{K}. \tag{3}$$

Then we introduce $x_{k,n}^u$ to indicate whether user k is served by satellite n during $[t_u, t_{u+1})$, given by

$$x_{k,n}^u = \begin{cases} 1 & \text{user } k \text{ is served by satellite } n, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We assume that the total bandwidth of each satellite is divided into L channels with equal bandwidth and each user can use only one channel during the transmission of one satellite. The channel budget constraint is then given by

$$\sum_{k \in \mathcal{K}} x_{k,n}^u \leq L, \forall n \in \mathcal{N}. \tag{5}$$

The problem of the satellite handover can be viewed as an optimization problem which selects the satellite from $\mathcal{I}_k^u$ for serving user k during each coverage section. We denote by HO_k as the satellite handovers of user k. If the service relation between satellite n and user k changes with the coverage section changing from $[t_u, t_{u+1})$ to $[t_{u+1}, t_{u+2})$, i.e. $x_{k,n}^u \neq x_{k,n}^{u+1}$, HO_k increases by 1. The average number of user handovers in the system is then given by

$$\overline{HO} = \frac{\sum_{k \in \mathcal{K}} HO_k}{K}. \tag{6}$$

We aim to optimize the service association indicator $x_{k,n}^u$ during a time period T to minimize the average number of handovers while improving channel utilization efficiency in the LEO satellite system. Then the optimization problem is formulated as follows:

$$\max_{\{x_{k,n}^u\}} \overline{HO} \tag{7a}$$

$$\text{s.t. } x_{k,n}^u \in \{0,1\}, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{7b}$$

$$\sum_{k\in\mathcal{K}} x_{k,n}^u \le L, \forall n \in \mathcal{N}, \tag{7c}$$

$$\theta_{k,n} \ge \theta_0, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{7d}$$

(7b) indicates two states of user association between user k and satellite n, (7c) is the constraint of total channels of a satellite. and (7d) is the minimum elevation angle constraint, where θ_0 is predefined.

B. Problem Formulation

Problem (7) is a combinatorial integer optimization problem, which is NP-hard in general. Here, we transform the problem into an MARL-based optimization problem based on stochastic game. The LEO satellite handover optimization problem is essentially a general-sum K-agents game since the agents have both cooperative and competitive relations in the system. The key definitions of MARL are given as follows:

(Definition 1) Agent: Users $k \in \mathcal{K}$ that take actions at each step and cause coverage state transitions.

(Definition 2) State: $s_t^k =< c_t^{k,n}, L_t^n, v_t^{k,n} >$ represents the state of k-th agent at time t, which consists of the covering satellites $c_t^{k,n}$, the available channels of the satellites L_t^n and the remaining visible time of the satellites $v_t^{k,n}$, respectively.

(Definition 3) Action: a_t^k denotes the action of k-th agent, and $a_t^k = x_t^{k,n}$. $x_t^{k,n}$ indicates whether user k is served by satellite n at time t.

(Definition 4) Reward: $r_t^k(s_t^k, a_t)$ denotes the reward of k-th agent. It is used to describe the instantaneous reward after action a_t is executed at state s_t^k. Note that a_t is a joint action of all agents at time t. We assume that the agents do not know other agents' reward functions but they can obtain other agents' actions. We define the reward function in three different cases:

- User k selects a satellite covering the user, but not serving the user, the immediate handover occurs, i.e. $c_t^{k,n} = 1, x_t^{k,n} = 0$;
- User k selects a satellite covering the user and serving the user, but the satellite is overloaded, i.e. $c_t^{k,n} = 1, x_t^{k,n} = 1, L_t^n < \sum_{k\in\mathcal{K}} x_t^{k,n}$;
- User k selects a satellite covering the user and serving the user, and the satellite channels is sufficient for its serving users, i.e. $c_t^{k,n} = 1, x_t^{k,n} = 1, L_t^n \ge \sum_{k\in\mathcal{K}} x_t^{k,n}$.

The reward function is then defined as

$$r_t^k(s_t^k, a_t) = \begin{cases} -20, & \text{if } c_t^{k,n} = 1, x_t^{k,n} = 0, \\ -10, & \text{if } c_t^{k,n} = 1, x_t^{k,n} = 1, \\ & \quad L_t^n < \sum_{k\in\mathcal{K}} x_t^{k,n}, \\ v_t^{k,n}, & \text{if } c_t^{k,n} = 1, x_t^{k,n} = 1, \\ & \quad L_t^n \ge \sum_{k\in\mathcal{K}} x_t^{k,n}. \end{cases} \tag{8}$$

TABLE I

MULTI-AGENT Q-LEARNING ALGORITHM

Algorithm

Initialize:
$t = 0, s^k = s_0^k =< c_0^{k,n}, L_0^n, v_0^{k,n} >$;
$Q^k(s_0^k, \mathbf{a}_0) = 0, \forall k \in \mathcal{K}, s \in \mathcal{S}$;
while $t < T$
  do
  for $k \in \mathcal{K}$
    for $s \in \mathcal{S}$
      Agent k:
        observe $s_t^k =< c_t^{k,n}, L_t^n, v_t^{k,n} >$;
        choose action a_t^k based on policy $\pi^k(s_t^k)$;
        obtain $\{a_t^1, \ldots, a_t^{k-1}, a_t^{k+1}, \ldots, a_t^K\}$;
        observe the reward $r^k(s_t^k, \mathbf{a}_t)$ and s_{t+1};
        update the Q-values $Q_t^k(s_t^k, \mathbf{a}_t)$ by Eq. (11);
    end for
  end for
  $t = t + 1$;
end while

The reward value is a positive integer only when the immediate handover would not occur and the channel budget of the satellite is sufficient. Otherwise, it is a negative integer. In particular, the positive integer is equal to the remaining visible time, since the longer the remaining visible time is the less possible a handover would occur in the coming future.

The expected accumulative reward of agent k is given by

$$v^k(s, \pi) = \sum_{t=0}^{T} \gamma \mathrm{E}\{r_t^k | s_0 = s, \pi\}, \tag{9}$$

where s_0 is the initial state and γ is the discounted factor. The objective of agent k is to find an optimal policy π_k^* that maximizes the expected cumulative reward, i.e.,

$$\pi_k^* = \arg\max_{\pi} v^k(s, \pi). \tag{10}$$

Hence, the objective of the LEO satellite handover optimization problem that minimizes the average handovers is equivalent to Eq. (10), which searches for an optimal policy to maximize the expected cumulative reward overtime.

III. MULTI-AGENT Q-LEARNING ALGORITHM

Q-Learning algorithm is a classic and effective reinforcement learning algorithm. We extend the Q-learning to the multi-agent cases and employ a multi-agent Q-learning algorithm for the general-sum K-agents game [13, 14]. The Q-function of agent k is denoted by $Q^k(s, a^1, a^2, \ldots, a^K)$, and then the Q-values of agent k is updated as follows:

$$\begin{aligned} Q_{t+1}^k(s_t, \mathbf{a}_t) = (1-\alpha) Q_t^k(s_t, \mathbf{a}_t) + \alpha \left[r^k(s_t, \mathbf{a}_t) + \right. \\ \left. \gamma \cdot \pi^1(s_{t+1}) \ldots, \pi^K(s_{t+1}) \cdot Q_t^k(s_{t+1}) \right], \end{aligned} \tag{11}$$

where $\mathbf{a}_t = (a_t^1, a_t^2, \ldots, a_t^K)$ is the joint action at time t.

The multi-agent Q-learning algorithm is given in Table I. At each time step t, agent k observes the state s_t^k, select an action a_t^k based on the policy $\pi^k(s_t^k)$, and the actions of other

TABLE II
BOLTZMAN EXPLORATION COMBINED WITH ϵ-GREEDY POLICY

Algorithm
Input: $Q^k(s_t^k, \mathbf{a}_t), \tau, \epsilon$;
calculate $\pi_t(\mathbf{a}_t)$ by Eq. (14);
generate a random number $\xi \in (0, 1)$;
if $\xi < \epsilon$;
calculate $\pi_t(\mathbf{a}_t)$ by Eq. (14);
$a_t^* = \arg\max_{\mathbf{a}_t} \pi_t(\mathbf{a}_t)$;
else
$a_t^* = \arg\max_{\mathbf{a}_t} Q^k(s_t^k, \mathbf{a}_t)$;
Output: $\mathbf{a}_t^*$;

agents are simultaneously decides. Then the immediate reward $r^k(s_t^k a_t)$ is obtained. The Q-tables of agent k is then updated by Eq. (11). An optimal equilibrium policy is obtained when the expected rewards of the strategy is maximized for any agent k, i.e.,

$$(\pi_*^1, \pi_*^2, \dots, \pi_*^K) = \arg\max_{\mathbf{a}} Q^k(s, \mathbf{a}), \forall k \in \mathcal{K}. \quad (12)$$

Balancing exploitation and exploration in reinforcement learning is very important when designing an action selection policy. Exploitation means that the agents choose the best action based on the current Q-values, also called a greedy policy. Exploration means that the agents try more actions that are not exploited so far to explore a larger action space. We introduce the ϵ-greedy action selection policy and extend it into a combined Boltzman exploration and ϵ-greedy policy according to [15]. Let a_t^* be the selected action at time t. Given an exploration parameter $\epsilon \in [0, 1)$, we have

$$a_t^* = \begin{cases} \text{random action from } \mathcal{A}_t & \text{if } \xi < \epsilon, \\ \arg\max_{a_t} Q^k(s_t^k, \mathbf{a}_t) & \text{otherwise,} \end{cases} \quad (13)$$

where ξ is a uniform random variable.

Table II shows the combined Boltzman exploration and ϵ-greedy policy, in which the selection probability of an action, denoted by $\pi_t(\mathbf{a}_t)$, is weighted by its related Q-value, i.e,

$$\pi_t(\mathbf{a}_t) = \frac{e^{\left(\frac{Q^k(s_t^k, \mathbf{a}_t)}{\tau}\right)}}{\sum_{\mathbf{a}_t} e^{\left(\frac{Q^k(s_t^k, \mathbf{a}_t)}{\tau}\right)}}, \quad (14)$$

where τ is a temperature parameter. Although this policy is viewed as a random action selection policy, we can see that the agents have more possibility to choose good actions due to the property of the probability function. Therefore, we combine Boltzman exploration process with ϵ-greedy policy for a better balance of the exploitation and exploration.

IV. SIMULATION RESULTS

In this section, we evaluate the performance of our proposed MARL-based satellite handover strategy and compare with the following handover strategies:

- **Maximum Visible Time (MVT)-Based Satellite Handover Strategy:** Users select the next satellite with the longest remaining visible time when the elevation angle of the serving satellite reduces to the minimum value. This strategy provides the lower bound of the satellite handovers.
- **Maximum Available Channels (MAC)-Based Satellite Handover Strategy:** Users select the next satellite with the least load.
- **Graph-Based Weighted (GBW)-Based Satellite Handover Strategy:** The strategy models the satellites' coverage for a user as a time graph [11]. We calculate the shortest/longest path from one node to another to achieve different objectives. The weight of the remaining visible time and the available satellite channels are set as by $w_1 = 0.7$ and $w_2 = 0.3$.

TABLE III
SIMULATION PARAMETERS

$a = 6680$ km	semimajor axis of the orbit
$b = 0.6$	eccentricity of the orbit
$c = 45°$	inclination of the orbit
$d = 25°$	right ascension of ascending node of the orbit
$e = 22.5°$	argument of perigee of the orbit
$f = 17.5°$	true anomaly of the orbit
$H = 800$ km	Height of a satellite
$v = 780$ km/h	Satellite rotation speed
$N = 48$	Number of satellites
$W = 10$ MHz	Bandwidth budget of a satellite
$\theta_0 = 10°$	Minimum elevation angle
$T = 600$ s	Satellite rotation time

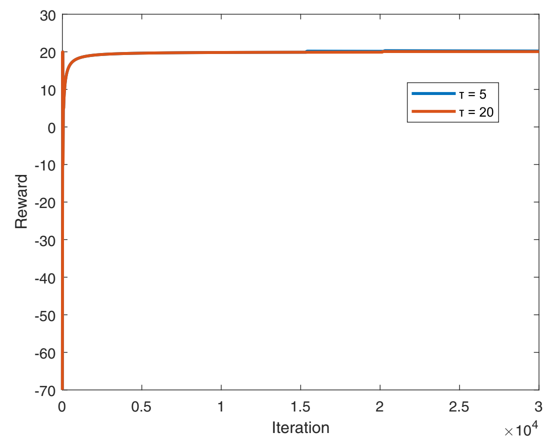

Fig. 2. Reward as a function of the iteration number with $K = 6$ and $L = 5$.

We consider a square area with 2 km side length, where users are uniformly distributed within the area. To construct the LEO satellite system, the semimajor axis, eccentricity, inclination, the right ascension of ascending node (RAAN), argument of perigee and true anomaly are set as $a = 6680$km, $b = 0.6, c = 45°, d = 25°, e = 22.5°$ and $f = 17.5$, respectively. We consider a satellite rotation period $T = 600$ s. The bandwidth budget of each satellite is $W = 10$ MHz. The minimum elevation angle is set as $\theta_0 = 10°$. The simulation parameters are summarized in Table III.

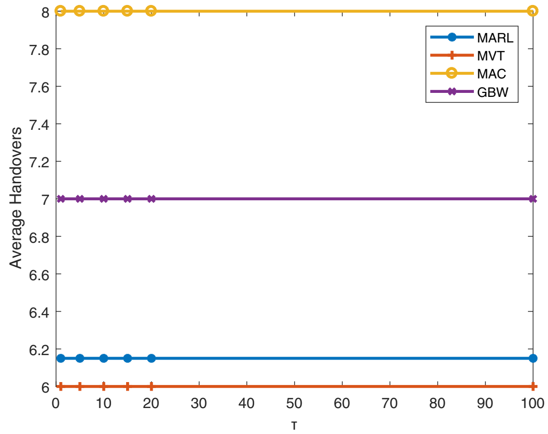

Fig. 3. Average number of handovers as a function of the action selection policy parameter τ with $K = 6$ and $L = 5$.

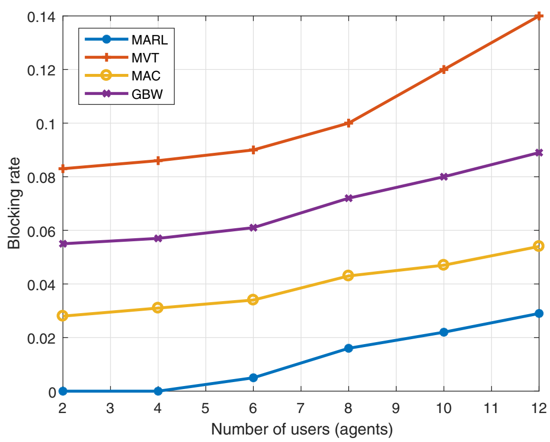

Fig. 4. Blocking rate as a function of the number of users with $\epsilon = 0.1, \tau = 10$ and $L = 5$.

In Fig. 2, we show the reward as a function of the number of iterations, where the number of users is $K = 6$ and the number of channels is $L = 5$. As can be seen, the average reward converges to about 20 with merely small fluctuation. The reason is that they select more good actions based on the exploration function (14) when using the combined policy, while the agents select completely random actions during exploration when using ϵ-greedy policy.

In Fig. 3, we show the average handovers as a function of the action selection policy parameter τ, where the number of users is $K = 6$ and the number of channels is $L = 5$. It can be seen that the changes of τ has no impact on the final performance of satellite handover. This demonstrates that the proposed multi-agent Q-learning algorithm always converges to the same equilibrium when the action selection parameters are set within a reasonable range. Besides, we can observe that our proposed strategy performs better than the MAC-based and GBW-based strategies by a 23% and 12% reduction of the average handovers, respectively. Also, the proposed strategy achieves a close performance to the MVT-based strategy, merely yielding a 4% increase of the average handovers.

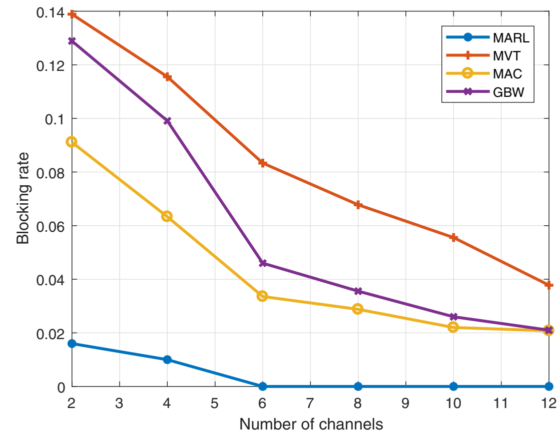

Fig. 5. Blocking rate as a function of the number of channels with $\epsilon = 0.1, \tau = 10$ and $K = 6$.

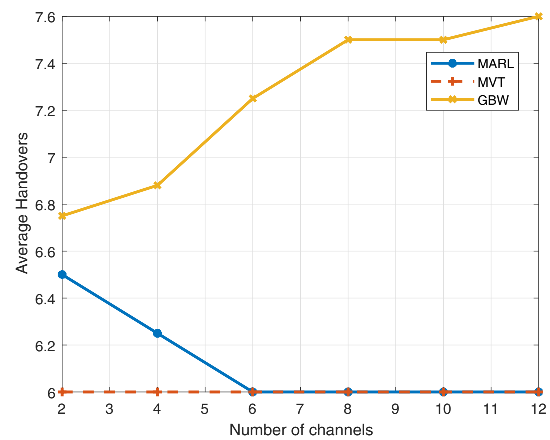

Fig. 6. Average number of handovers as a function of the number of channels with $\epsilon = 0.1, \tau = 10$ and $K = 6$.

In Fig. 4, we show the user blocking rate as a function of the number of users, where the system parameters are $\epsilon = 0.1, \tau = 10$ and $L = 5$. As the number of users increases, we can see that the user blocking rate increases because a larger percentage of users would be dropped when competing for the channel resource. We can also see that when the number of users is below 6, the blocking rate increases slowly. And when the number of users exceeds 6, the blocking rate increases with a larger speed. Besides, our proposed strategy outperforms the MVT, MAC and GBW-based strategies by a $50\% - 100\%$ reduction of blocking rate.

In Fig. 5, we show the blocking rate as a function of the number of channels, where the system parameters are $\epsilon = 0.1, \tau = 10$ and $K = 6$. We can see that the blocking rate drops as the number of channels increases since more users can be served with a larger channel budget. Moreover, our proposed strategy outperforms the MVT, MAC and GBW-based strategies by a $75\% - 100\%$ reduction of blocking rate. From Figs. 3, 4 and 5, we show that the proposed strategy can significantly improves the blocking rate performance at the cost of only a small increase in handover signaling.

In Fig. 6, we show the average satellite handovers as a function of the number of channels, where the system parameters are $\epsilon = 0.1, \tau = 10$ and $K = 6$. As we can see, the average number of handovers of the GBW-based strategy increases as the number of channels increases, while the average handovers of our proposed strategy decreases to the lower bound. In the MARL process, even if a satellite has the longest visible time for an agent, the reward of the joint action that take the signalling overhead and satellite load into consideration can still be negative. Therefore, when the channel budget is not sufficient to serve all visible users, the agents may select the satellite with the second longest visible time to avoid low channel utilization efficiency. When the channel budget is sufficient, the agents would not need to sacrifice the handover performance to guarantee load balance. The proposed strategy outperforms the GBW-based strategy by a 4%-20% reduction of handovers.

V. Conclusions

In this paper, we have investigated the satellite handover problem in an LEO satellite network. An optimization problem was formulated to minimize the average number of handovers subject to the load constraint of each satellite. The problem was then transformed into an MARL problem. We have solved the problem by using a distributed Q-learning algorithm, where a combined Boltzman exploration and ϵ-greedy policy is employed to better balance exploitation and exploration. Simulation results have shown that the proposed satellite handover strategy not only reduces the average number of handovers compared to the baseline strategies, but also hugely decreases the user blocking rate, which implies better channel utilization of the entire system.

References

[1] P. Chitre and F. Yegenoglu, "Next-Generation Satellite Networks: Architectures and Implementations," *IEEE Commun. Mag.*, vol. 37, no. 3, pp. 30-36, Mar. 1999.

[2] A. Jamalipour and T. Tung, "The Role of Satellites in Global IT: Trends and Implications," *IEEE Pers. Commun.*, vol. 8, no. 3, pp. 5-11, Jun. 2001.

[3] P. K. Chowdhury, M. Atiquzzaman and W. Ivancic, "Handover schemes in satellite networks: state-of-the-art and future research directions," *IEEE Commun. Surveys Tuts.*, vol. 8, no. 4, pp. 2-14, Feb. 2006.

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

[5] S. Karapantazis and F.-N. Pavlidou, "Dynamic time-based handover management in LEO satellite systems," *Electron. Lett.*, vol. 43, no. 5, pp. 57-58, Mar. 2007.

[6] S. Karapantazis and F.-N. Pavlidou, "QoS handover management for multimedia LEO satellite networks," *Telecommun. Syst.*, vol. 32, no. 4, pp. 225-245, May 2012.

[7] E. Papapetrou and F.-N. Pavlidou, "Analytic study of Doppler-based handover management in LEO satellite systems," *IEEE Trans. Aero. Electron. Syst.*, vol. 41, no. 3, pp. 830-839, Jul. 2005.

[8] M. Gkizeli, R. Tafazolli and B. G. Evans, "Modeling handover in mobile satellite diversity based systems," in *Proc. IEEE 54th Veh. Technol. Conf.*, Atlantic City, USA, pp. 131-135, Oct. 2001.

[9] M. Gkizeli, R. Tafazolli and B. G. Evans, "Hybrid channel adaptive handover scheme for non-GEO satellite diversity based systems," *IEEE Commun. Lett.*, vol. 5, no. 7, pp. 284-286, Aug. 2006.

[10] E. Papapetrou, S. Karapantazis, G. Dimitriadis and F.-N. Pavlidou, "Satellite Handover Techniques for LEO Networks, *Int. J. Satellite Commun. and Net.*, vol. 22, no. 2, pp. 231-245, Apr. 2004.

[11] Z. Wu, F. Jin, J. Luo, Y. Fu, J. Shan and G. Hu, "A Graph-Based Satellite Handover Framework for LEO Satellite Communication Networks," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1547-1550, Aug. 2016.

[12] L. Busoniu, R. Babuska and B. De Schutter, "Multi-agent Reinforcement Learning: An Overview," *Innovations in Multi-Agent Systems and Applications*, vol. 310, pp. 183-221, Sept. 2019.

[13] J. Hu and M. P. Wellman, "Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm," *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 242-250, Jul. 1998.

[14] J. Hu and M. P. Wellman, "Nash Q-Learning for General-Sum Stochastic Games," *J. Machine Learning Research*, vol. 7, Aug. 2004

[15] A. D. Tijsma, M. M. Drugan and M. A. Wiering,"Comparing exploration strategies for Q-learning in random stochastic mazes," *IEEE Symposium Series on Computational Intelligence*, Athens, Greece, Dec. 2016.