

# Multi-Agent Deep Reinforcement Learning Based Handover Strategy for LEO Satellite Networks

Chungnyeong Lee<sup>1</sup>, Inkyu Bang<sup>2</sup>, *Member, IEEE*, Taehoon Kim<sup>3</sup>, *Member, IEEE*,  
Howon Lee<sup>4</sup>, *Senior Member, IEEE*, Bang Chul Jung<sup>5</sup>, *Senior Member, IEEE*,  
and Seong Ho Chae<sup>6</sup>, *Member, IEEE*

**Abstract**—The high rotation speeds and mega-constellations of low earth orbit satellites (LEO SATs) cause the inter-satellite frequent handovers (HOs) problem which can lead to substantial performance degradation. This letter proposes a novel distributed multi-agent deep Q-network based SAT HO strategy for the LEO SAT networks to simultaneously minimize the number of HOs and maximize the throughputs and the visible times of UEs while satisfying the quality-of-service (QoS) constraints of all UEs. The proposed HO scheme allows UEs to independently and simultaneously perform the HO decision makings based on their own local information, which enables to immediately adapt to the dynamic changes of the LEO SAT network environments. The numerical results demonstrated that our proposed HO strategy achieves the lowest average HO rate and the highest achievable throughputs compared to other conventional HO strategies, while ensuring a higher QoS guarantee time ratio.

**Index Terms**—Low earth orbit satellites, handover strategy, multi-agent deep reinforcement learning.

## I. INTRODUCTION

THE low earth orbit (LEO) satellite (SAT) networks have recently garnered significant interests in providing ubiquitous connectivity with global coverage and seamless switching. The LEO SATs are able to provide several advantages such as low propagation delay and low energy consumption, but their high orbit rotation speed inevitably results in frequent inter-satellite handovers (HOs) during the service duration

Received 26 November 2024; revised 27 February 2025; accepted 22 March 2025. Date of publication 26 March 2025; date of current version 12 May 2025. This work was partly supported by Korea Research Institute for defense Technology planning and advancement(KRIT) grant funded by the Korea government(DAPA(Defense Acquisition Program Administration)) (KRIT-CT-22-047, Space-Layer Intelligent Communication Network Laboratory, 2022), by Innovative Human Resource Development for Local Intellectualization program(IITP-2025-RS-2020-II201741, 33%), and by ICAN(ICT Challenge and Advanced Network of HRD)(IITP-2025-RS-2022-00156326, 33%) through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT). The associate editor coordinating the review of this letter and approving it for publication was S. K. Jayaweera. (*Corresponding authors: Seong Ho Chae; Taehoon Kim.*)

Chungnyeong Lee is with the Department of IT Semiconductor Convergence Engineering, Tech University of Korea, Siheung 15073, South Korea (e-mail: lc9902130509@tukorea.ac.kr).

Inkyu Bang is with the Department of Intelligence Media Engineering, Hanbat National University, Daejeon 34158, South Korea (e-mail: ikbang@hanbat.ac.kr).

Taehoon Kim is with the Department of Computer Engineering, Hanbat National University, Daejeon 34158, South Korea (e-mail: thkim@hanbat.ac.kr).

Howon Lee and Bang Chul Jung are with the Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, South Korea (e-mail: howon@ajou.ac.kr; bcjung@ajou.ac.kr).

Seong Ho Chae is with the Department of Electronics Engineering, Tech University of Korea, Siheung 15073, South Korea (e-mail: shchae@tukorea.ac.kr).

Digital Object Identifier 10.1109/LCOMM.2025.3554818

of terrestrial user equipments (UEs) [1]. Such unnecessary frequent HO has been known to cause many problems such as throughput reduction, increased signaling overhead, and communication interruption, so careful design of the HO strategy is crucial for LEO SAT networks.

Recently, various HO strategies with different criteria and constraints have been proposed for LEO SATs networks [2], [3], [4], [5]. A network-flows graph based HO strategy for LEO SAT networks was proposed in [2], where the matching between the SATs and the user terminals (UTs) was determined by minimizing the product of costs and flows of the graph under the limited capacity and maximal flow of edges. The multiple forecast-based HO procedures to reduce HO delay and signaling cost for the multi-layer LEO SAT systems were proposed in [3], where the utility function-based HO optimization to reduce the dropping probability and increase the throughput was addressed. Wu et al. proposed the bipartite graph framework for the SAT HO and the potential game-based HO algorithm to maximize the benefits of mobile terminals in the LEO SAT networks [4]. The group HO strategy for massive UTs was proposed in LEO SAT networks [5]. However, the proposed HO strategies in [2], [3], [4], and [5] were designed in a centralized way, which requires the central controller and the global information, thus inevitably resulting in significant signaling overhead in dense LEO SAT networks.

There have also been some trials to relax such a limit of the centralized scheme, but only a few decentralized HO schemes have been proposed in LEO SAT networks so far [6], [7]. The multi-agent reinforcement learning (MARL) based LEO SAT HO strategy to minimize the average number of HOs while satisfying the load constraint of each SAT was proposed in [6], but it required each user to have access to all the action information of other users. Thus, it posed challenges for users to dynamically adapt to real-time changes in the environment. The distributed LEO SAT HO scheme based on successive deep Q-learning (SDQL) algorithm was proposed in [7], enabling each user to perform the SAT HO when HO triggered, based on the local information while considering HO delay, HO failure, quality-of-service (QoS) of users, and inter-SAT traffic balancing. However, it cannot successfully guarantee real-time responsiveness in multi-user environments because each user sequentially inputs all sub-states for its visible SATs into a fully connected Q-network and then sequentially takes an action based on all resulting output Q-values. It also did not consider the number of HOs even though it is an important aspect for the HO design.

To effectively adapt to the highly dynamic environments of LEO SATs networks, the HO decision should be made in real-time, independently, and simultaneously, rather than

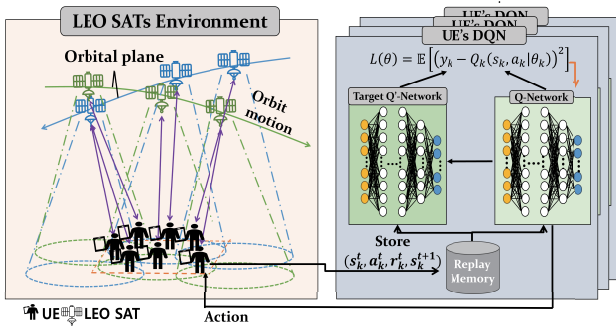


Fig. 1. System model with the proposed MADQN-based HO strategy.

sequentially, based on the local information. Thus, we propose a novel *distributed* satellite HO strategy based on multi-agent deep Q-network (MADQN) for LEO SAT networks. The contributions of this letter are summarized as follows.

- The proposed HO scheme enables UEs to independently and simultaneously make HO decisions based on their own local information to immediately adapt to the dynamic changes of the LEO SAT network environments. To this end, we formulate the multi-objective optimization problem to simultaneously minimize the number of HOs and maximize the throughputs and visible times of UEs, while satisfying the QoS constraints of all UEs.
- We transform it into a multi-agent deep reinforcement learning (MADRL) optimization problem and then propose the distributed HO algorithm based on MADQN, which considers the diverse characteristics of SAT networks such as the coverage of visible SATs, the visible time of SATs, the fading channels, the network interferences, and the load of SATs.
- Our numerical results demonstrate that the proposed HO strategy achieves the lowest average HO rate and the highest achievable throughputs compared to other conventional HO strategies, while ensuring a higher QoS guarantee time ratio.

## II. NETWORK MODEL AND PROBLEM FORMULATION

### A. Network Model

We consider the SAT HO decision-making problem during consecutive  $T$  time slots in LEO SAT networks, where total  $N$  LEO SATs orbiting over the pre-configured  $M$  orbital planes serve  $K$  UEs uniformly located on the ground with a finite area  $A$  of interest. We assume that each UE is within the coverage of at least one SAT during every time slot due to the sufficiently large coverage area of each LEO SAT. The considered scenario is illustrated in Fig. 1. Let us denote the sets of indices of  $N$  SATs,  $K$  UEs, and  $M$  orbital planes as  $\mathcal{N} \triangleq \{1, 2, 3, \dots, N\}$ ,  $\mathcal{K} \triangleq \{1, 2, 3, \dots, K\}$ , and  $\mathcal{M} \triangleq \{1, 2, \dots, M\}$ , respectively. We denote the number of SATs orbiting on the orbital plane  $m$  as  $N_m$ , then  $\sum_{m=1}^M N_m = N$  holds. The coverage area of each LEO SAT is assumed as a circle and each SAT is located above the center of its coverage area. All SATs are continuously orbiting along the pre-configured trajectories of orbital planes with the same orbital period. Accordingly, the footprints of all SATs on the Earth also move along with the satellite trajectory. Note that

the coverage area of each SAT can be partially overlapped according to the position of the SATs in different orbital planes.

Let us denote the indices set of the time slots as  $\mathcal{T} \triangleq \{1, \dots, T\}$ . The channels experience quasi-static fading during one time slot but vary from one time slot to another. The system bandwidth of each SAT is  $W$ . We assume that the UEs randomly move around within a finite area  $A$  at every time slot based on the random walk model [8]. Each UE is aware of its current position by using the global positioning system (GPS) and learns all covering SATs due to the predictable motion of LEO SATs [1]. Owing to high mobility of LEO SATs and random movement of UEs, the available SATs for each UE dynamically vary at every time slot. Thus, UEs are required to decide whether to be handed over to another visible SAT or not at every time slot. The bandwidth of each SAT is equally and orthogonally shared among all connected UEs.

### B. Handover Decision Making Problem

The coverage indicator between SAT  $n$  ( $\in \mathcal{N}$ ) and UE  $k$  ( $\in \mathcal{K}$ ) at time slot  $t$  is denoted by  $c_{k,n}^t$ , which is defined as 1 if UE  $k$  is within the coverage of SAT  $n$  at time slot  $t$ , and 0 otherwise. Thus, the set of SATs covering UE  $k$  at time slot  $t$  is denoted by  $\mathcal{C}_k^t = \{n \mid c_{k,n}^t = 1, n \in \mathcal{N}\}$ .

The connection between UE  $k$  and SAT  $n$  at time slot  $t$  is denoted by the indicator variable  $x_{k,n}^t$ , which is defined as 1 if UE  $k$  is connected to SAT  $n$ , and 0 otherwise. Then, the load of SAT  $n$  at time slot  $t$  can be expressed as  $l_n^t = \sum_{k \in \mathcal{K}} x_{k,n}^t$ . The UE  $k$  is able to obtain *local information* on the number of loads to the only covering SATs, i.e.,  $\forall n \in \mathcal{C}_k^t$ . Let us denote the connection between UE  $k$  and all SATs at time slot  $t$  as  $\mathbf{x}_k^t \triangleq [x_{k,1}^t, \dots, x_{k,N}^t]$ . Since UE  $k$  is allowed to be connected to only one SAT at each time slot, the HO of UE  $k$  occurs when  $\mathbf{x}_k^t \neq \mathbf{x}_k^{t+1}$ . We define the HO cost of UE  $k$  at time slot  $t$  as

$$\varphi_k^t = \begin{cases} \xi, & \text{if the HO occurs } (\mathbf{x}_k^t \neq \mathbf{x}_k^{t+1}), \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where  $\xi$  is the positive value. The channel between UE  $k$  and SAT  $n$  at time slot  $t$  is modeled as  $h_{k,n}^t = \left( \frac{c}{4\pi f_d^t d_{k,n}^t} \right) \sqrt{G_n G_k} \Psi_{k,n}^t$ , where  $c$  is the speed of light,  $f_d^t = \frac{v_t}{c} f_c^t$  is the Doppler effect with carrier frequency  $f_c^t$  and relative velocity  $v_t$  between SAT  $n$  and UE  $k$ ,  $G_n$  and  $G_k$  are antenna gains of SAT  $n$  and UE  $k$  respectively,  $d_{k,n}^t$  is the distance between SAT  $n$  and UE  $k$ , and  $\Psi_{k,n}^t$  represents the shadowed-Rice fading which incorporates both large-scale fading (due to shadowing) and small-scale fading (caused by multipath). Once UE  $k$  is connected to SAT  $n$  at time slot  $t$ , i.e.,  $x_{k,n}^t = 1$ , then the achievable throughput of UE  $k$  can be expressed as

$$R_{k,n}^t = \frac{W}{l_n^t} \log \left( 1 + \frac{P|h_{k,n}^t|^2}{\sum_{j \in \mathcal{I}_{k,n}^t} P|h_{k,j}^t|^2 + \frac{W}{l_n^t} N_0} \right), \quad (2)$$

where  $P$  is the transmit power of SAT,  $N_0$  is the noise power spectral density.  $\mathcal{I}_{k,n}^t = \{j \mid j \in \mathcal{C}_k^t \setminus n, l_j^t \geq 1\}$  represents the set of interfering SATs of UE  $k$  when UE  $k$  is associated with SAT  $n \in \mathcal{C}_k^t$ . We assume that each UE requires the throughput

requirement  $R_{k,n}^t \geq R_k^{\text{th}}$  to satisfy the target QoS. Note that load balancing and interference control are required to satisfy the throughput requirements.

Let us denote the remaining coverage time (i.e., visible time) of SAT  $n$  at UE  $k$  at time slot  $t$  as  $v_{k,n}^t$ . Then, the aim of this letter is to simultaneously minimize the HO cost and maximize both the throughput and the visible time under the constraint for QoS of each UE during  $T$  times slots, which can be formulated as

$$\mathbf{P}: \min_{\{x_{k,n}^t\}} \sum_{t=1}^T \sum_{\forall k \in \mathcal{K}} \varphi_k^t - \sum_{n \in \mathcal{C}_k^t} x_{k,n}^t (\omega_v v_{k,n}^t + \omega_R R_{k,n}^t), \quad (3)$$

$$\text{s. t. } x_{k,n}^t \in \{0, 1\}, \quad \forall n \in \mathcal{C}_k^t, \quad (4)$$

$$\sum_{n \in \mathcal{C}_k^t} x_{k,n}^t = 1, \quad \forall k \in \mathcal{K}, \quad (5)$$

$$\sum_{n \in \mathcal{C}_k^t} x_{k,n}^t R_{k,n}^t \geq R_k^{\text{th}}, \quad \forall k \in \mathcal{K}, \quad (6)$$

where  $\omega_v$  and  $\omega_R$  represent the scaling factors corresponding to the importance for the remaining visible time and throughput. The constraints (4) and (5) imply that each UE is associated with only one SAT in its coverage area for every time slot. Note that  $\mathcal{C}_k^t$  dynamically varies for each time slot due to the high rotation speed of the SATs and random mobility of UEs. Note that the above optimization problem is a combinatorial NP-hard integer optimization problem.

### III. PROPOSED MULTI-AGENT DEEP Q-NETWORK BASED HANDOVER SCHEME FOR LEO SATELLITE NETWORKS

In our optimization problem (3), the number of possible combinations is,  $(N^K)^T$ , and as  $T$  and  $K$ , the number of combinations grows exponentially. Therefore, it is necessary to transform this problem into a distributed approach. The MADRL presents a promising distributed method to effectively address such complex problems [11]. Therefore, in the following subsection, we reformulate optimization problem (3) into the MADRL framework.

#### A. Multi-Agent Deep Reinforcement Learning Framework

To address the optimization problem (3) based on MADRL, it is essential to convert the optimization problem into the decentralized partially observable Markov decision process (Dec-POMDP) framework. The Dec-POMDP consists of 7-tuple  $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z}, \mathcal{O}, \mathcal{K} \rangle$ :  $\mathcal{S}$  represents the set of states,  $\mathcal{A}$  represents the set of actions,  $\mathcal{P}(s'|s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  [10] represents the set of the conditional transition probabilities between states,  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  represents the reward function,  $\mathcal{Z}$  represents set of the locally observations,  $\mathcal{O}(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{Z}$  represents set of conditional observation probabilities, and  $\mathcal{K}$  is the number of agents. The optimization problem (3) can be transformed into the Dec-POMDP framework as follows.

1) *Agent*: Each UE  $k \in \mathcal{K}$  is an agent and all UEs interact with each other and the environment.

2) *State*: At each time slot  $t$ , the UEs engage in interaction with the environments, resulting in the update of its state.

In the following, we simply refer to the observation state as the state for brevity. The state of UE  $k$  is defined as<sup>1</sup>  $s_k^t = \langle \mathcal{C}_k^t, \mathcal{V}_k^t, h_{k,n}^t, l_n^t \rangle, \forall n \in \mathcal{C}_k^t$ .  $\mathcal{V}_k^t = \{v_{k,1}^t, v_{k,2}^t, \dots, v_{k,N}^t\}$  is the set of remaining coverage times (i.e., visible times) of all SATs for each UE  $k$  at time slot  $t$ . Note that the state consists of *current* and *local* information.

3) *Action*: The action of UE  $k$  in time slot  $t$  is denoted by  $a_k^t = n$  for  $\forall n \in \mathcal{C}_k^t$ . That is, UE  $k$  selects SAT  $n$  among the covering SATs at each time slot  $t$ .

4) *Reward*: The reward  $r_k^t$  of UE  $k$  at time slot  $t$  is defined as follows:

$$r_k^t = \begin{cases} -\zeta, & \text{if } R_{k,a_k^t}^t < R_k^{\text{th}}, \\ -\varphi_k^t, & \text{if } a_k^{t-1} \neq a_k^t, R_{k,a_k^t}^t \geq R_k^{\text{th}}, \\ \omega_v v_{k,a_k^t}^t + \omega_R R_{k,a_k^t}^t, & \text{otherwise,} \end{cases} \quad (7)$$

where  $-\zeta$  represents the penalty for the outage event that the QoS of UE  $k$  fails to meet its target QoS  $R_k^{\text{th}}$ , which is regardless of HO occurrence.  $-\varphi_k^t$  represents the penalty for the event of HO success, i.e., HO occurs and QoS requirement of UE  $k$  is satisfied.

#### B. Proposed Multi-Agent Deep Q-Network Algorithm

In this subsection, we propose a MADQN algorithm to find the optimal action-value function  $Q^{k*}(s, a)$  or the optimal policy  $\pi_k^*$ . The DQN is a multi-layered neural network to approximate the optimal action-value function  $Q^k(s, a; \theta_k) \approx Q^{k*}(s, a)$ , where  $\theta_k$  represents the weight of the neural networks. The action  $a_k$  is obtained according to the  $\epsilon$ -greedy policy. The transition tuples  $(s_k^t, a_k^t, r_k^t, s_k^{t+1})$  are stored in the replay memory  $D$  that can randomize over the data and removing correlations between the samples. The DQN improves the stability of learning by using a target network model  $Q^k(s, a; \theta_k^-)$  with weight  $\theta_k^-$ , which involves periodically updating the main network's weights  $\theta_k$  to update the target network. The loss function for DQN is given by  $L(\theta) = E[(y_k - Q_k(s_k, a_k | \theta_k))^2]$ , where  $y_k = r_k^t + \alpha \max_{a_{k,t+1}} Q_k(s_k^{t+1}, a_{k,t+1} | \theta_k^-)$  is the target value.

We utilize MADQN because it is compatible with our scenario that each UE independently utilizes local information and its decentralized training with decentralized execution structure of MADQN is well-suited for solving our problem effectively. In MADQN, each agent possesses an individual neural network that adheres to the structure of the DQN previously described, which is depicted in Fig. 1. Note that all agents are updated concurrently by interacting with each other through the environment, which ensures real-time performance. The pseudo code for MADQN is described in the Algorithm 1.

#### C. Analysis of the Proposed Method

1) *Complexity*: Let  $H, N_h, I_s$ , and  $I_o$  denote the number of training layers (input, hidden, and output layers), the number

<sup>1</sup>In satellite communications, unlike terrestrial communications, the long communication distance results in significant delay, which makes channel estimation a critical issue. In this letter, we assume that channel prediction can be successfully performed via Transformer-based method [9]. The state has a fixed-sized  $4 \times N$  matrix form and the informations about SATs not within the coverage of the UE are set to -1. The channel and load informations are obtained through the broadcasts of SATs.

**Algorithm 1** MADQN for HO Algorithm

---

**Initialization:**  
1 Initialize parameters: learning rate, exploration rate  $\epsilon$ , discount factor  $\alpha$  **for agent**  $k = 1, K$  **do**  
2 Initialize replay memory  $D$  to capacity  $N$   
3 Initialize action-value function  $Q_k(s_k, a_k|\theta_k)$  with random weight  $\theta_k$   
4 Initialize target network  $Q_k(s'_k, a'_k|\theta_k^-)$ ,  $\theta_k^- \leftarrow \theta_k$

**Learning :**  
5 **for**  $episode = 1, M$  **do**  
6  $s_1$ : Initialize the first state from  $s_k^t = \langle C_k^t, \mathcal{V}_k^t, h_{k,n}^t, l_n^t \rangle$ , for  $\forall k \in \mathcal{K}$   
7 **for**  $t = 1, T$  **do**  
8 Each agent (UE)  $k$  excute random action  $a_k^t$  with probability  $\epsilon$ , otherwise excute  $a_k^t = \arg \max_a Q_k(s_k, a_k|\theta_k)$   
9 Observe reward  $r_k^t$  and next state  $s_k^{t+1}$   
10 Store transition  $(s_k^t, a_k^t, r_k^t, s_k^{t+1})$  in  $D$   
11 **if** *it's time to update* **then**  
12 **for**  $k = 1, K$  **do**  
13 Randomly sample minibatch of transition  $(s_k^t, a_k^t, r_k^t, s_k^{t+1})$  from  $D$   
14 Define  $a_k^{max} = \max_{a_k} Q_k(s_k^{t+1}, a_k^{t+1}|\theta_k^-)$   
15  $y_k = \begin{cases} r_k^t & \text{for terminals } s_k^{t+1} \\ r_k^t + \alpha a_k^{max} & \text{otherwise} \end{cases}$   
16 Perform a gradient descent step on  $(y_k - Q_k(s_k, a_k|\theta_k))^2$   
17 **if**  $t \% \text{policy-update-freq} = 0$  **then**  
18  $\theta_k^- \leftarrow \theta_k$

---

of neurons in layer  $h$ , the size of the input layer, and the size of the output layer, respectively. The computational complexity of DQN algorithms for each agent can be expressed as  $C_{\text{DQN}} = \mathcal{O}(X)$ , where  $X = I_s N_1 + \sum_{h=2}^{H-1} N_{h-1} N_h + N_{H-1} I_o$ . For the training phase involving  $K$  agents,  $E$  episodes, and  $T$  time slots, the computational complexity of the proposed approach can be formulated as:  $C_{\text{MADQN}} = KET \times C_{\text{DQN}} = \mathcal{O}(KETX)$ .

2) *Convergence*: The convergence of the above proposed MADQN HO algorithm can be justified by demonstrating that conventional Q-learning converges to the optimal state under suitable learning rates and the neural networks effectively approximate the nonlinear Q-values. Given these properties, our MADQN-based methods are theoretically guaranteed to converge, ensuring stable and optimal solutions [11]. Furthermore, to empirically demonstrate convergence to optimality, we compare the performance of our proposed algorithm in a simplified environment scenario with that of the optimal algorithm, as simulated results will show.

#### IV. NUMERICAL RESULTS

In this section, we evaluate the performances of our proposed MADQN-based SAT HO strategy in terms of average HO rate, QoS guarantee time ratio, and moving average throughput, and show its superiority by comparing with the following benchmark HO strategies.

- 1) **Random SAT HO [12]**: Each UE randomly chooses one among its covering SATs for a HO.

TABLE I  
SIMULATION PARAMETERS

Parameters		Values
Orbital planes	Inclination	53°
	Count	2
Number of SATs per orbital plane		22
Altitude		550km
Rotation speed		7.59km/s
Radius of coverage area		55km
Satellite Tx power		150W
Satellite Tx max gain		50dBi
Bandwidth		250MHz
Carrier frequency		20GHz (Ka-band)
Throughput threshold $R^{th}$		10Mbps
Channel model		Rice [7]

- 2) **Maximum Channel Gain (MAX-CG) based SAT HO**: Each UE chooses the SAT with the best channel gain for a HO.
- 3) **Maximum visible time (MVT) based SAT HO [1]**: Each UE selects the SAT with the longest visible time for a HO.
- 4) **Load-Aware SAT HO based MADQN (LA-MADQN)**: Each UE selects the SAT for HO with the load information, where the state is defined as  $s_k^t = \langle C_k^t, l_n^t, \mathcal{V}_k^t \rangle$  and the reward  $r_k^t$  is defined as  $-\zeta$  for  $R_{k,a_k^t}^t < R_k^{th}$ ,  $-\varphi_k^t$  for  $a_k^{t-1} \neq a_k^t$  and  $R_{k,a_k^t}^t \geq R_k^{th}$ , and  $v_{k,a_k^t}^t$  otherwise, by following the similar MDP framework in [6].

We consider the scenario that  $K = 100$  UEs are uniformly distributed within a square with 100km side length and they are moving around with a speed of 60km/h based on the random walk model [8]. The detailed simulation parameters are summarized in TABLE I. The time step interval is set to 1 second and the total episode time is set to  $T = 90$  seconds. The remaining visible time and throughput importance scaling factors are set to  $\omega_v = 1$  and  $\omega_R = 1.3 \times 10^{-8}$ , respectively. The penalties  $-\zeta$  and  $-\varphi_k^t$  in the reward (7) are set to  $-20$  and  $-40$ , respectively. For the proposed MADQN-based SAT HO strategy, each DQN is set as follows: Each agent is equipped with a fully connected network comprising three hidden layers consisting of 256 neurons each. The ReLU function is employed as the activation function for each hidden layer. The capacity of the replay memory  $D$  is  $N = 1 \times 10^6$ , the discount factor is  $\alpha = 0.9$ , the learning rate is  $1 \times 10^{-3}$ , the batch size is 128, the number of training steps to update the target DQN is set to  $1 \times 10^2$ , the exploration rate  $\epsilon$  gradually decreases from 1 to  $1 \times 10^{-3}$ , and the Adam optimizer is used for training the DQN.

Fig. 2 compares the performances between the proposed algorithm and the optimal solution for a simplified scenario with  $K = 2$  UEs,  $N = 6$  LEO SATs, and  $T = 5$  time slots. The left sub-figure shows that the cumulative reward of the proposed algorithm converge to that of optimal solution as the training episodes increases, while the right sub-figure demonstrates that the proposed method, when executed after training, achieves the same throughput as the optimal solution, validating the optimality of our proposed algorithm.

Fig. 3 and Table II compare the proposed HO strategy with other benchmark HO strategies in terms of the QoS guarantee time ratio and the average HO rate, respectively. Note that the QoS guarantee time ratio is averaged over every 10 UEs and

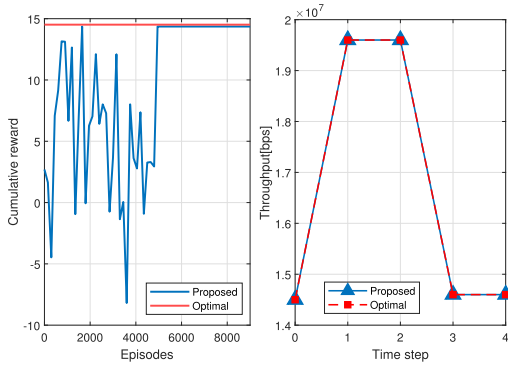


Fig. 2. Comparison of cumulative rewards in training episodes and throughputs for 5 time slots between the optimal solution and the proposed algorithm.

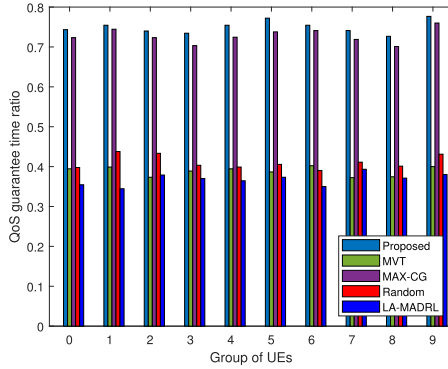


Fig. 3. Comparison of the QoS guarantee time ratio across groups of UEs.

TABLE II

AVERAGE HANDOVER RATES

Strategies	Random	MVT	MAX-CG	LA-MADQN	Proposed
Percentage	68%	64%	58%	44%	33%

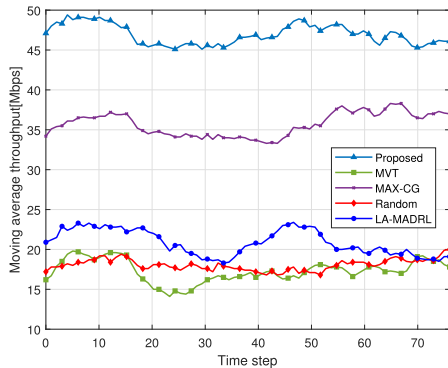


Fig. 4. Comparison of the moving average throughputs.

represented as grouped bars. The QoS guarantee time ratio is defined as the time portion that each UE satisfies the QoS requirements among the entire episode duration. The average HO rate is defined as the ratio of the total HO occurrence time to the entire episode duration, averaged across all UEs. Fig. 3 and Table II validate that our proposed MADQN-based HO algorithm significantly outperforms other algorithms in terms of the average HO rate, while maintaining a relatively long QoS guarantee time for all UEs compared to other HO strategies.

Fig. 4 compares the proposed HO strategy with other benchmark HO strategies in terms of the moving average

throughput. The moving average throughput is calculated by averaging achievable throughputs of all UEs over a window of 15 time steps. Coupled with the results from Table II, it is evident that our proposed algorithm effectively solves the optimization problem (3), achieving the highest moving average throughput and the lowest average HO rate compared to other benchmark algorithms.

## V. CONCLUSION

This letter has proposed a novel distributed MADQN based SAT HO strategy for LEO SAT networks to simultaneously minimize the number of HOs and maximize the throughputs of UEs while satisfying the QoS constraints of all UEs. In our proposed HO scheme, each UE independently and simultaneously performs the HO decision making based on its own local information for covering SATs, which enables to immediately adapt to the dynamic changes of the LEO SAT network environments. The numerical results demonstrated that our proposed HO strategy achieves the lowest average HO rate and the highest achievable throughputs compared to other conventional HO strategies, while ensuring a higher QoS guarantee time ratio. Exploring novel HO strategies for terrestrial-satellite coexisting networks, considering multi-path delay and computational cost optimization in large-scale networks, and investigating the fundamental trade-offs in advanced HO cost modeling including interference and load balancing would be an interesting future research topic.

## REFERENCES

- [1] 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, 4th Quart., 2006.
- [2] S. Zhang, A. Liu, C. Han, X. Ding, and X. Liang, "A network-flows-based satellite handover strategy for LEO satellite networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2669–2673, Dec. 2021.
- [3] Y. Li, W. Zhou, and S. Zhou, "Forecast based handover in an extensible multi-layer LEO mobile satellite system," *IEEE Access*, vol. 8, pp. 42768–42783, 2020.
- [4] Y. Wu, G. Hu, F. Jin, and J. Zu, "A satellite handover strategy based on the potential game in LEO satellite networks," *IEEE Access*, vol. 7, pp. 133641–133652, 2019.
- [5] L. Yang, X. Yang, and Z. Bu, "A group handover strategy for massive user terminals in LEO satellite networks," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2022, pp. 1–6.
- [6] S. He, T. Wang, and S. Wang, "Load-aware satellite handover strategy based on multi-agent reinforcement learning," in *Proc. IEEE Global Commun. Conf.*, Jan. 2020, pp. 1–6.
- [7] H. Liu, Y. Wang, and Y. Wang, "A successive deep Q-learning based distributed handover scheme for large-scale LEO satellite networks," in *Proc. IEEE 95th Veh. Technol. Conf.*, Jun. 2022, pp. 1–6.
- [8] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 2, pp. 95–107, Apr. 2020.
- [9] H. Jiang, M. Cui, D. W. K. Ng, and L. Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2717–2732, Sep. 2022.
- [10] X. Lin et al., "Intelligent adaptive MIMO transmission for non-stationary communication environment: A deep reinforcement learning approach," *IEEE Trans. Commun.*, early access, Jan. 14, 2025, doi: 10.1109/TCOMM.2025.3529263.
- [11] D.-D. Tran, S. K. Sharma, V. N. Ha, S. Chatzinotas, and I. Woungang, "Multi-agent DRL approach for energy-efficient resource allocation in URLLC-enabled grant-free NOMA systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1470–1486, 2023.
- [12] J.-H. Lee, C. Park, S. Park, and A. F. Molisch, "Handover protocol learning for LEO satellite networks: Access delay and collision minimization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7624–7637, Jul. 2024.