A Multi-Agent Deep Reinforcement Learning-Based Handover Scheme for Mega-Constellation Under Dynamic Propagation Conditions

Haotian Liu, Yichen Wang, *Member, IEEE*, Peixuan Li, and Julian Cheng, *Fellow, IEEE*

***Abstract*— With the rapidly increasing number of satellites, the handover scheme design is critically important for the low Earth orbit (LEO) satellite networks, especially for the mega-constellations that include massive number of LEO satellites. However, the existing handover schemes for LEO satellite networks are designed based on the static propagation conditions, which cannot satisfy the dynamic feature of communication environment caused by the mobility of LEO satellites and users. To address this issue, a centralized adaptive intelligent handover scheme for mega-constellations is proposed, where the dynamics of the propagation conditions and limited LEO satellite capacity are taken into considerations. Specifically, we first use a three-state Markov model to characterize the dynamically varying propagation conditions between satellites and users. Then, the Loo model is employed to describe the dynamic land mobile satellite channels. By considering the user transmission rate requirement and the load-balancing demand of satellites, we design the user utility function and formulate an optimization problem that aims to maximize the overall long-term utility of the network. To reduce the handover decision-making complexity, a multi-agent successive hysteretic deep Q-learning algorithm is developed and it can efficiently solve the formulated problem by reducing the state and action space. To reduce the signaling overhead and the computation complexity of the proposed centralized handover scheme brought to the control center, a distributed intelligent handover scheme is further developed, where each user is enabled to independently make the handover decision only based on the local information. Simulation results show that both the proposed centralized and distributed approaches can efficiently improve the network performance over the existing schemes.**

***Index Terms*— Satellite communication, low earth orbit (LEO) satellite, mega-constellation, satellite handover, multi-agent deep reinforcement learning, propagation condition.**

Manuscript received 23 February 2023; revised 5 September 2023 and 9 May 2024; accepted 22 May 2024. Date of publication 6 June 2024; date of current version 11 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271383 and Grant 62071373 and in part by the Innovation Capability Support Program of Shaanxi under Grant 2021TD-08. An earlier version of this paper was presented in part at the IEEE Vehicular Technology Conference (VTC)-Spring, Helsinki, Finland, June 2022 [DOI: 10.1109/VTC2022-Spring54318.2022.9860376]. The associate editor coordinating the review of this article and approving it for publication was M. C. Vuran. *(Corresponding author: Yichen Wang.)*

Haotian Liu, Yichen Wang, and Peixuan Li are with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: lht1998@stu.xjtu.edu.cn; wangyichen0819@mail.xjtu.edu.cn; lpx7835995@stu.xjtu.edu.cn).

Julian Cheng is with the School of Engineering, The University of British Columbia, Kelowna, BC V1V 1V7, Canada (e-mail: julian.cheng@ubc.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3407358.

Digital Object Identifier 10.1109/TWC.2024.3407358

I. Introduction

In THE past few decades, the booming development of the wireless applications has triggered an increasing demands for reliable and wide-area network connectivity. However, limited by the cell coverage of terrestrial wireless networks, massively deploying base stations is not a feasible way to meet such demands, especially in the rural and ocean areas. Thanks to the wide coverage and high throughput capabilities, satellite communication is a powerful approach to address the abovementioned issue. Compared with the geostationary Earth orbit (GEO) and medium Earth orbit (MEO) satellites, the low Earth orbit (LEO) satellites have the lower orbital altitude and thus can efficiently reduce the propagation delay, decrease the transmit power, and provide lower launch and deployment cost. Consequently, LEO satellite communication has been widely accepted as an efficient approach to achieve seamless global communications and will play an important role in the future sixth-generation (6G) systems [1], [2], [3].

Although LEO satellite communication can offer improved global coverage and communication quality, it still has several challenges, where one of the most important challenges is the frequent handover. Different from the terrestrial networks with the pre-deployed fixed base stations, LEO satellites move fast [4], [5]. For example, the moving speed of the satellite can achieve 7 km/s in the Iridium system [6]. As a consequence, the connections between users and satellites show highly dynamic feature such that the handover has to be frequently performed during the user's service duration. Moreover, in recent years, several companies have launched the plans to construct the LEO mega-constellations, such as SpaceX and OneWeb [7], [8]. Compared with the traditional LEO networks, the number of visible satellites for ground users increases significantly in the LEO mega-constellations, which will result in more frequent and complex satellite handovers [9]. Consequently, the handover scheme design is critically important to the mega-constellation based LEO satellite networks.

In the past few years, several handover schemes have been developed to reduce the impact of handover on the system performance. Depending on whether the connected satellite changes, the handovers between ground users and satellites can be broadly classified into beam handovers and satellite handovers [10]. Specifically, a beam handover occurs when the user moves from one beam to another of the same satellite,

while a satellite handover occurs when the existing connection of one satellite is transferred to another satellite. Early research studies towards the handover schemes for LEO satellite networks mainly focused on the beam handover since the number of satellites in the traditional LEO satellite networks was limited [11], [12], [13]. As the LEO mega-constellations will greatly increase the frequency of satellite handovers, recent works mainly focused on the satellite handover scheme design. A number of graph-based satellite handover schemes have been developed [14], [15], [16]. In [14], the authors proposed a basic graph-based satellite handover framework, where each covering period of a satellite was regarded as the node and the user's handover between satellites was treated as a directed edge. In this way, the handover process was viewed as finding an optimal path in the constructed graph. Authors in [15] adopted the graph-based framework to the multiple-input multiple-output (MIMO) system where gateway stations were required to switch to different satellites in order to avoid the waste of power resources. In [16], the priorities of different users were taken into account and a handover model of the network-flows was built.

Although the graph-based satellite handover schemes can achieve good performance, it is challenging to apply them into the mega-constellations because the scale of the constructed graphs increases rapidly as the number of satellites grows. To address this issue, machine learning, especially reinforcement learning (RL), is regarded as an efficient approach. As a state-of-the-art technology, RL has been widely employed in the satellite communications, such as beam hopping [17], [18], [19], power control [20], [21], [22] and routing design [23], [24], [25]. Instead of trying to find the optimal solution directly, RL enables each agent to autonomously interact with a dynamic environment in real time and learn the optimal solution by itself [26] such that it shows great superiority when the system is in the highly dynamic environment. Moreover, once the learning phase is completed, an optimal strategy can be quickly obtained with a low complexity. In recent years, several works have applied the RL methods into the satellite handover designs [27], [28], [29], [30]. Authors in [27] proposed a user's quality of experience (QoE) driven intelligent handover scheme which includes two parts, namely the handover factors modeling and the RL-based handover decision. The handover factors modeling aimed to determine the handover factors including the spatial relationship, available channel and relay overhead. The handover decision determined an optimal satellite to perform the handover based on the derived factors. Authors [28] proposed a satellite handover strategy aiming to minimize the average times of handover while satisfying the load constraint of each satellite. In [29], a centralized-training-and-distributed-execution scheme was developed where a trainer node deployed in the backhaul network was responsible for training the environment parameters. Once the training stage is finished, the trained environmental parameters are disseminated from the trainer node to users and each user makes handover decisions individually based on the obtained parameters. Authors in [30] took the freshness of information into consideration and proposed an age-oriented satellite access control strategy which aimed to minimize the long-term peak age-of-information. Besides the abovementioned RL-based works, the multi-layer architecture and software-defined network technology were also adopted in the handover mechanism designs [31], [32].

Although many handover schemes have been proposed for satellite communications networks, all existing handover schemes are designed under the static propagation conditions. However, due to the movement of users and satellites, the propagation conditions between users and satellites show highly dynamic features. To be more specific, land mobile satellite (LMS) users often operate under the cluttered circumstances. With the movement of ground users and satellites, the links between users and satellites are likely to be shadowed or blocked by obstacles, such as trees and buildings, which will result in a serious degradation in the channel quality and cause the outage of the system [33]. In this case, to ensure the quality-of-service (QoS) of users, it is necessary to switch the original shadowed or blocked link to a new link which is in line-of-sight (LOS) condition. Nevertheless, the changes of the propagation conditions are stochastic and cannot be accurately predicted, which are not sufficiently cognized by the existing handover schemes. Moreover, it is more challenging to consider the dynamic propagation conditions in the mega-constellations since the huge number of satellites will significantly increase the signaling overhead and computation complexity for obtaining and estimating the propagation conditions. Consequently, there is an urgent need to design a dynamic propagation conditions oriented handover scheme in the mega-constellation, which can fully recognize and exploit the dynamic of the propagation conditions.

To achieve the above goal, we propose a centralized adaptive intelligent satellite handover scheme for LEO mega-constellations under dynamic propagation conditions. Specifically, we first employ a three-state Markov model to characterize the dynamic propagation conditions and adopt the Loo model to describe the stochastic channel of the LMS link. Then, the utility function for each user is designed, where the user transmission rate requirement and the load-balancing demand of satellites are jointly considered. Based on the utility function, an optimization problem that aims to maximize the overall long-term utility of the network is formulated. To solve the formulated problem, we develop a multi-agent successive hysteretic deep Q-learning (MA-SHDQL) algorithm which can efficiently solve the problem with a low complexity. To reduce the signaling overhead and the computation complexity of the proposed centralized handover scheme, we further propose a distributed intelligent handover scheme, where each user is enabled to independently make the handover decision only based on the local information. Simulation results show that both the proposed centralized and distributed schemes can efficiently improve the network performance over the existing schemes. The main contributions of this paper are summarized as follows.

1) Different from the existing handover schemes where only the static propagation condition is considered, the dynamic propagation conditions between users and satellites are integrated into the proposed scheme design.

Specifically, we first employ a three-state Markov model to characterize the typical propagation conditions including the LOS, shadowed, and blocked propagation states as well as the probabilistic transitions among the three states. Then, the Loo model is adopted to describe the time-varying LMS channels. In this way, the stochastic nature of the dynamic propagation conditions between users and satellites are sufficiently characterized and also put significant challenges in the satellite handover scheme design.

2) Based on the characterization of dynamic propagation conditions between users and satellites, we design the frame structure and propose a centralized adaptive intelligent handover mechanism for LEO mega-constellations, where the handover delay and handover failure probability are jointly considered. Then, we design the utility function for each user by taking the throughput of users and load-balancing demand of satellites into consideration. Moreover, we further formulate an optimization problem that aims at maximizing the long-term network overall utilities to optimize the performance of the proposed scheme.
3) Although the formulated optimization problem can be characterized by a single-agent Markov decision process (MDP) framework and solved by the fundamental RL method, huge state and action space will be incurred, which will significantly increase the computation complexity and thus might not be appropriate for the LEO mega-constellations. To solve this problem, we convert the single-agent MDP framework to a cooperative multi-agent MDP framework, where the space and action space are efficiently reduced. Based on the constructed cooperative multi-agent MDP framework, we develop a MA-SHDQL algorithm, which can efficiently solve the formulated problem with relatively low complexity. In this way, the handover decisions can be intelligently determined by cognizing the dynamic propagation environments between users and satellites.
4) As the LEO mega-constellation includes a large number of satellites, the centralized handover scheme might incur heavy signaling overhead and computation burden to the control center. To address this issue, we further propose a distributed intelligent handover scheme, where each user can independently determine its own handover decisions only based on its local information. Due to the distributed manner, the complexity of the developed distributed scheme will not increase as the numbers of users and satellites grow. Consequently, the proposed distributed scheme is promising to be applied in the LEO satellite network with massive users and satellites.

The rest of this paper is organized as follows. Section II presents the system model with dynamic propagation conditions. In Section III, we formulate the handover problem and proposed a MA-SHDQL algorithm. In Section IV, a distributed handover scheme is further developed. Simulation results are given in Section V and the paper is concluded in Section VI.

II. System Model

A. Network Model

We consider a LEO mega-constellation, which consists of M LEO satellites indexed by $\mathcal{M} = \{1, 2, \cdots, M\}$ and N ground mobile users indexed by $\mathcal{N} = \{1, 2, \cdots, N\}$. The service arrival of all ground users follows a Poisson process with arrival rate λ and average duration T_m. A hybrid wide-spot beam (HWSB) coverage scheme is employed [6]. Compared with the traditional spot beam coverage scheme where the footprints of beams on the earth move along with the satellite trajectory, HWSB coverage scheme enables the narrow bandwidth spot beams to steer to users such that the beam handover is not needed. In this paper, we call the users serviced by the corresponding narrow bandwidth spot beams as the beams' *targeted users*. For the users that are not serviced by the corresponding narrow bandwidth spot beams but receive the signals of the beams, we call them as the *interfered users* of the beams. Due to the limited on-board resource of satellites, it is assumed that each satellite can only provide at most $C_{\max}$ narrow bandwidth spot beams at the same time, which implies that each satellite can service at most $C_{\max}$ users simultaneously. A control center is deployed to collect the global information of the system and make handover decisions for users.[1] We assume that the time is divided into slots with a duration of ΔT and the system remains static during each time slot and changes instantaneously from one slot to another.

B. Dynamic Propagation Conditions

In this paper, we integrate the dynamics of propagation conditions into the handover scheme design. We assume that all users are under a cluttered circumstance such that the LMS channel quality may vary significantly with the user's surroundings. Moreover, due to the movement of mobile users and satellites, the surroundings of users may change over time, resulting in the dynamics of propagation conditions. In this paper, we model the dynamics of propagation conditions as a three-state first-order Markov chain [33]. As shown in Fig. 1, the propagation condition includes three states: LOS, shadowed and blocked. The Markov chain is characterized by the transition probability matrix $\mathbf{P}(\theta_{i,j}(t))$ and the state probability vector $\mathbf{W}(\theta_{i,j}(t))$. It is noted that the values of elements in $\mathbf{P}(\theta_{i,j}(t))$ and $\mathbf{W}(\theta_{i,j}(t))$ are determined by the elevation angle $\theta_{i,j}(t)$, which is defined as the angle between the local horizontal plane of user i and the direction towards satellite j in slot t, since the probability of the LMS channels being shadowed or blocked is directly related to the elevation angle of the satellite. According to [33], we define the propagation condition frame (PCF) that includes T_F slots. We assume that the propagation condition remains unchanged during each PCF and changes to other propagation conditions according to $\mathbf{P}(\theta_{i,j}(t))$ and $\mathbf{W}(\theta_{i,j}(t))$ at the beginning of each PCF.

[1] In realistic global LEO satellite networks, multiple control centers are usually deployed and each control center is responsible for coordinating and controlling the handovers of users in a given area. In this paper, we mainly focus on one control center since the control centers make handover decisions for its assigned users independently.

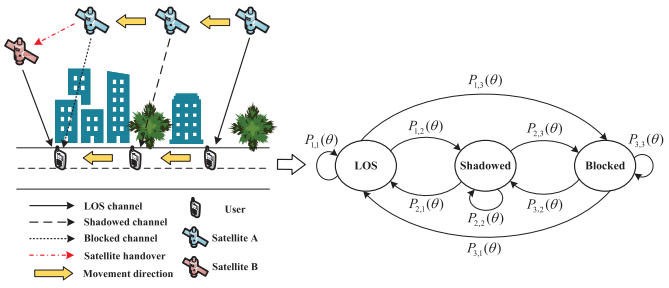

Fig. 1. The dynamic of the propagation condition in the LEO satellite networks and the corresponding Markov model.

C. Channel Model

Based on the abovementioned dynamic propagation conditions, the channel model can be developed. Suppose that the satellite j ($j \in \mathcal{M}$) uses a narrow bandwidth spot beam j_q to service its targeted user q ($q \in \mathcal{N}$), then the overall channel power gain from the beam j_q to the interfered user i ($i \in \mathcal{N}$, $i \neq q$) in slot t can be written as [34]

$$Q_{i,j_q}(t) = L_{i,j}(t) \cdot G^S_{i,j_q}(t) \cdot G^T \cdot h_{i,j}(t) \tag{1}$$

where

- $L_{i,j}(t) = (\frac{c}{4\pi f_I d_{i,j}(t)})^2$ is the free space loss (FSL) between user i and satellite j in slot t with c, f_I and $d_{i,j}(t)$ denoting the light speed, carrier frequency and distance between user i and satellite j in slot t, respectively.
- $G^S_{i,j_q}(t)$ is the satellite antenna gain from beam j_q to user i in slot t and can be expressed as [35], [36], [37]

$$G^S_{i,j_q}(t) = G^{\max}_{j_q}(t) \left(\frac{J_1(\mu_{i,j_q}(t))}{2\mu_{i,j_q}(t)} + 36\frac{J_3(\mu_{i,j_q}(t))}{\mu^3_{i,j_q}(t)} \right)^2 \tag{2}$$

where $G^{\max}_{j_q}(t)$ is the maximum satellite antenna gain of beam j_q in slot t, $J_n(\cdot)$ is the Bessel function of first kind and n-th order [38], and $\mu_{i,j_q}(t) = 2.07123 \sin\varphi_{i,j_q}(t)/\sin\varphi_{3\text{dB}}$ with $\varphi_{i,j_q}(t)$ and $\varphi_{3\text{dB}}$ denoting the off-axis angle from beam j_q to user i in slot t and one-sided half-power beamwidth of beams, respectively. According to [39] and [40], satellites in the mega-constellation with HWSB coverage scheme can adjust the power of the narrow bandwidth spot beams to compensate for the variations of FSL such that the received power of the targeted user will not vary significantly. In this paper, we also compensate the variation of the signal FSL by adjusting the power of transmitted signal. However, we treat the adjustment of the transmission power as the equivalent adjustment of the satellite antenna gain for the convenience of expression. Consequently, we suppose that the maximum satellite antenna $G^{\max}_{j_q}(t)$ can be adaptively adjusted according to the FSL between the targeted user q and the satellite j while keeping the power of the narrow bandwidth spot beam fixed. Specifically, $G^{\max}_{j_q}(t)$ can be written as $G^{\max}_{j_q}(t) = K_G/L_{q,j}(t)$, where K_G is a coefficient determined by the capability of satellite antenna. Note that, although adjusting the maximum satellite antenna gain can compensate for the variation in FSL for the targeted users, it also incurs the heavier interference to the interfered users and thus puts greater challenge on the scheme design. In addition, considering that the altitude of satellites is much larger than the size of narrow bandwidth spot beams, we treat the footprints of beams as circles [34] and thus $\mu_{i,j_q}(t)$ can be expressed as $\mu_{i,j_q}(t) = 2.07123 \cdot d_{i,q}(t)/R$, where $d_{i,q}(t)$ is the distance between the user i and the targeted user q in slot t and R is the radius of beams. Apparently, the satellite antenna gain from beam j_q to the targeted user q is larger than that to other users and thus we have $G^S_{i,j_q}(t) \leq G^S_{q,j_q}(t) = G^{\max}_{j_q}(t)$.
- G_T is the terminal antenna gain of users. All users are assumed to employ omnidirectional antennas such that G_T is a constant.
- $h_{i,j}(t)$ is a random variable that characterizes both the shadowing and the multipath fading. Based on the Loo model, the probability density function (PDF) of $h_{i,j}(t)$ is given by [41]

$$f_{h_{i,j}(t)}(x) = \frac{1}{2b_{i,j}(t)\sqrt{2\pi n_{i,j}(t)}} \int_0^{+\infty} \frac{1}{y} I_0\left(\frac{y\sqrt{x}}{b_{i,j}(t)}\right) \times \exp\left[-\frac{(\ln y - m_{i,j}(t))^2}{2n_{i,j}(t)} - \frac{x+y^2}{2b_{i,j}(t)}\right] dy \tag{3}$$

where $m_{i,j}(t)$ and $n_{i,j}(t)$ denote the average signal gain and the variance of the LOS components in slot t, respectively, the multipath component in slot t follows a Rayleigh distribution characterized by its average power $b_{i,j}(t)$, and $I_0(\cdot)$ is the modified Bessel function of zeroth order. It is worth noting that, as shown in Table II, the parameters in (3) are jointly determined by the propagation condition $c_{i,j}(t)$ described in Section II-B and the elevation angle $\theta_{i,j}(t)$ between user i and satellite j in slot t.

Similarly, the overall channel power gain from the beam j_q to the targeted user q in slot t can be written as $Q_{q,j_q}(t) = L_{q,j}(t) \cdot G^{\max}_{j_q}(t) \cdot G^T \cdot h_{q,j}(t)$.

Compared with the regular channel models that adopt a static PDF to characterize the channel gain in the given environment [41], [42], [43], the adopted Markov channel model employs a group of PDFs to characterize the channel gain in the given environment and selects the best one according to the propagation condition. Consequently, the adopted Markov channel model can more accurately characterize the LMS channel. However, the PDF of the channel gain under the Markov channel model depends on the propagation condition and the elevation angle between the user and satellite, which shows the time-varying feature and thus puts greater challenge on the handover scheme design.

D. Signal Model

We denote the satellite decision of user i in slot t by vector the $\mathbf{x}_i(t) = [x_{i,1}(t), x_{i,2}(t), \cdots, x_{i,M}(t)]^T$, where $x_{i,j}(t) \in \{0,1\}$. To be more specific, $x_{i,j}(t) = 1$ implies

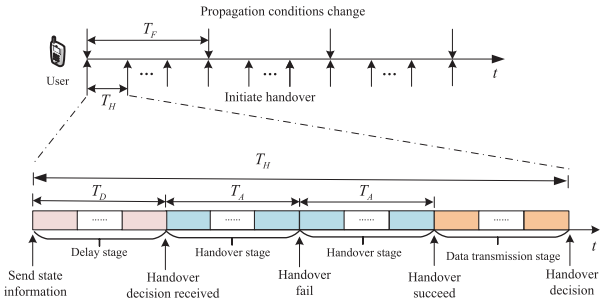

Fig. 2. The handover mechanism in the dynamic propagation conditions.

that user i chooses satellite j for handover in slot t and $x_{i,j}(t) = 0$ represents that user i does not choose satellite j for handover in slot t. We also denote by vector $\mathbf{g}_{i,q}(t) = \left[Q_{i,1_q}(t), Q_{i,2_q}(t), \cdots, Q_{i,M_q}(t)\right]^T$ the channel power gain of beams from M satellites to user i in slot t where the beams are supposed to be steered to the targeted user q. Then, the received signal of user i in slot t is given by

$$y_i(t) = \underbrace{\sqrt{P\mathbf{x}_i(t)^T\mathbf{g}_{i,i}(t)}s_i}_{\text{Desired Signal}} + \underbrace{\sum_{q=1,q\neq i}^{N}\sqrt{P\mathbf{x}_q(t)^T\mathbf{g}_{i,q}(t)}s_q}_{\text{Interference Signal}} + n \tag{4}$$

where P is the transmit power, s_i is the transmitted signal to user i with $\mathbb{E}\{||s_i||^2\} = 1$, and n is the additive white Gaussian noise (AWGN) with average power σ^2. According to (4), we can calculate the signal interference noise ratio (SINR) of user i in slot t as

$$\text{SINR}_i(t) = \frac{\mathbf{x}_i(t)^T\mathbf{g}_{i,i}(t)P}{\sum\limits_{q=1,q\neq i}^{N}\mathbf{x}_q(t)^T\mathbf{g}_{i,q}(t)P + \sigma^2}. \tag{5}$$

Thus, the transmission rate of user i in slot t is determined by

$$R_i(t) = B\log(1 + \text{SINR}_i(t)) \tag{6}$$

where B is the bandwidth of the beam.

III. MA-SHDQL Algorithm Based Handover Scheme Design

A. Frame Structure and Handover Mechanism Descriptions

To enable users to track the highly dynamic propagation conditions and avoid the transmission outage, we define the handover decision frame (HDF) which includes T_H slots such that the system has to make the handover decision for each user every T_H slots. To be more specific, as shown in Fig. 2, we divide the HDF into three parts, namely delay stage, handover stage and data transmission stage. The delay stage includes T_D slots used for the signaling exchange between the users, the satellites, and the control center. During the delay stage, the connection relationships between the users and the satellites will not change since the users have not received the new handover decisions. After the delay stage, all users will receive the corresponding handover decisions and then the following handover and data transmission stages begin. The durations of the handover and data transmission stages depend on the SINR of the user in each HDF. For the handover stage, it includes T_A slots used for the signaling exchanges between the user and the satellite and onboard processing of the satellite. However, the handover may be failed due to the outage caused by the poor channel quality or the severe interference. In that case, the handover stage has to be re-performed during the next T_A slots until the end of the current HDF. The outage probability of user i in slot t is determined by

$$P_i^o(t) = \Pr\{R_i(t) < R_{\min}\} \tag{7}$$

where $R_{\min}$ denotes the minimum transmission rate requirement. Based on (1)-(7), it is clear that the outage probability of each user is not only determined by itself, but also by other users who interfere with it. Once the handover of one user succeeds, the user will access the satellite and enter the data transmission stage including the remaining slots of the current HDF. Based on the above descriptions, the proposed handover mechanism can be summarized as follows:

- At the beginning of each HDF, users in service update their state information and send the updated information to the control center.
- During the next T_D slots, i.e., the delay stage, the user state information is sent to the control center via corresponding connected satellites. Then, the control center makes the handover decisions for users based on the collected user state information. Finally, the handover decisions are sent back to the users. During the delay stage, the connection relationships between the users and the satellites will not change. Specifically, if the user has accessed the current satellite successfully in the last HDF, the user will keep connected with the satellite in the delay stage of the current HDF. If the user has not accessed the current satellite successfully in the last HDF, the user will keep accessing the satellite in the delay stage.
- After the delay stage of each HDF, which can be denoted by $\mathcal{H} = \{T_D, T_H + T_D, 2T_H + T_D, \cdots\}$, all users in service will perform the corresponding handover actions based on the received handover decisions.
- For the users that decide not to perform the handover in the current HDF, they will keep the connection relationships with the currently selected satellites. Specifically, for the users that have successfully accessed the satellite, the handover stage is no longer needed and the remaining $T_H - T_D$ slots of the current HDF can be used for data transmission. For the users that have not accessed the satellite, the handover stage is needed.
- For the users that decide to perform the handover in the current HDF, the next T_A slots in the HDF will be consumed for the handover stage. However, the handover may fail according to (7). Once the handover fails, another handover stage including T_A slots will be initiated.
- The maximum allowed number of handover stages in the current HDF is $\lfloor (T_H - T_D)/T_A \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. If the user successfully accesses the satellite after k ($0 \leq k \leq \lfloor (T_H - T_D)/T_A \rfloor$) handover stages, the

remaining $(T_H - T_D - kTA)$ slots of the current HDF will be used for data transmission.

B. Optimization Problem Formulation

We treat the transmission between users and satellites as a trade, where satellites provide data transmission service for users and users have to pay for occupying the limited beam resources of satellites. We suppose that the size of each data packet is S_p and users can get a payment B_p once one data packet is successfully transmitted. Denote by $r_{i,j}(t) \in \{0,1\}$ the connection status between user i and satellite j in slot t. $r_{i,j}(t) = 1$ means user i has accessed to the satellite j successfully in slot t and the data transmission stage has started, while $r_{i,j}(t) = 0$ means user i has not accessed to the satellite j successfully in slot t. Moreover, we adopt $r_i(t) = \sum_{j=1}^{M} r_{i,j}(t)$ to denote that whether user i has accessed one satellite successfully. Since each user can only choose one satellite for service simultaneously, we have $r_i(t) \in \{0,1\}$. Consequently, the payment that user i can get in slot t is given by

$$G_i^P(t) = B_P \cdot D_i(t) \tag{8}$$

where $D_i(t)$ is the number of data packets transmitted to user i in slot t. $D_i(t)$ is given by

$$D_i(t) = \begin{cases} 0, & \text{if } (r_i(t)=0) \cup (R_i(t) < R_{\min}) \\ \left\lfloor \dfrac{R_i(t)\cdot \Delta T}{S_p} \right\rfloor, & \text{if } (r_i(t)=1) \cap (R_i(t) \geq R_{\min}). \end{cases} \tag{9}$$

Once satellite j receives the handover request from user i, it will steer the spot beam to user i during the handover stage as well as the data transmission stage. In this process, user i has to pay the cost for occupying the satellite's beam. Therefore, the cost that user i needs to pay for occupying satellite j's spot beam in slot t is given by

$$G_{i,j}^C(t) = \begin{cases} B_C \cdot U_j^C\left(n_j(t)\right), & \text{if } (r_{i,j}(t)=0) \cap (t \in \mathcal{H}) \\ G_{i,j}^C(t-1), & \text{otherwise} \end{cases} \tag{10}$$

where B_C is the minimum cost for occupying a spot beam, $n_j(t) = \sum_{i=1}^{N} x_{i,j}(t)$ represents the number of occupied spot beams of satellite j in slot t, and $U_j^C(\cdot)$ is a cost function. As $U_j^C(\cdot)$ should be an increasing function of the number of occupied spot beams, the Sigmoid-like function is adopted and $U_j^C(\cdot)$ is written as

$$U_j^C(x) = 1 + \frac{K_C - 1}{1 + \exp\left[-\dfrac{10}{C_{\max}}\left(x - \dfrac{C_{\max}}{2}\right)\right]} \tag{11}$$

where K_C is the upper-limit of $U_j^C(x)$. Apparently, the cost in each slot remains unchanged if the user keeps connected with the current satellite or the current HDF has not finished. Moreover, when the handover happens and the user chooses a new satellite to access, the fewer the number of available spot beams of the satellite has, the more the cost will be paid.

Based on (8)-(11), the utility function of user i in slot t is represented by

$$G_i(t) = G_i^P(t) - \sum_{j=1}^{M} x_{i,j}(t) G_{i,j}^C(t). \tag{12}$$

Then, the optimization problem that aims to maximize the system overall long-term utilities can be formulated as

$$(\text{OP}) \quad \max_{\mathbf{x}(t)} \sum_{t=0}^{\infty} \sum_{i=1}^{N} G_i(t) \tag{13}$$

$$\text{s.t.} \quad \sum_{j=1}^{M} x_{i,j}(t) \leq 1, \quad \forall t, \ \forall i \in \mathcal{N} \tag{14}$$

$$\sum_{i=1}^{N} x_{i,j}(t) \leq C_{\max}, \quad \forall t, \ \forall j \in \mathcal{M} \tag{15}$$

$$x_{i,j}(t) = x_{i,j}(t-1), \forall t \notin \mathcal{H}, \quad \forall i \in \mathcal{N}, \ \forall j \in \mathcal{M} \tag{16}$$

$$x_{i,j}(t) \in \{0,1\}, \quad \forall t, \forall i \in \mathcal{N}, \ \forall j \in \mathcal{M} \tag{17}$$

where $\mathbf{x}(t) = [\mathbf{x}_1(t), \mathbf{x}_2(t), \cdots, \mathbf{x}_N(t)]$ denotes the handover decisions of all users in slot t. Constraints (14) and (15) indicate that each user can only choose at most one satellite for handover and each satellite can service at most $C_{\max}$ users in each slot, respectively. Constraint (16) implies that the handover decision will be performed by the user after the delay stage of each HDF and remain unchanged during the remaining slots of the HDF. We can observe from the formulated problem (OP) that the handover decision $\mathbf{x}(t)$ should be adaptive to the dynamical network environment to obtain the maximum system overall long-term utilities. Consequently, conventional optimization algorithms may not be feasible to solve the formulated optimization problem. To address this issue, we first give a brief introduction to the basic deep Q-learning (DQL) algorithm. Then, we present the general single-agent MDP framework of the developed optimization problem. After that, we convert it to a multi-agent version to reduce the state and action space. Finally, a MA-SHDQL algorithm is proposed based on the constructed multi-agent MDP framework.

C. The DQL Preliminaries

The DQL algorithm, namely the deep neural network based Q-learning algorithm, is one of the most well-known and widely-used RL algorithms, which allows the agent to learn the optimal policy π^* by evaluating the action-values of the state-action pairs [44]. Specifically, we denote by $\mathcal{S}$ and $\mathcal{A}$ the state space and the action space, respectively. In each time step, the agent obtains a state s ($s \in \mathcal{S}$) from the environment and take an action a ($a \in \mathcal{A}$) according to its policy $\pi(a|s)$. Then, the environment will transform into a new state s' and the agent will obtain a reward r according to the probability $p(s', r|s, a)$. Based on the above discussions, the action-value $Q^\pi(s,a)$ that represents the expected cumulative long-term reward when the agent takes action a in state s given the policy π can be denoted by

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{\tau=1}^{+\infty} \gamma^\tau R^{(k+\tau)} \mid S^{(k)} = s, A^{(k)} = a\right] \tag{18}$$

where γ is the discount factor, $R^{(k)}$ is the reward obtained in step k, $S^{(k)}$ and $A^{(k)}$ denote the state and the action taken in step k, respectively. The action-value $Q^{\pi}(s,a)$ satisfies a recursive relationship between the action-value of the current state-action pair $Q^{\pi}(s,a)$ and that of its possible successor state-action pairs $Q^{\pi}(s',a')$, which can be denoted by [26]

$$Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p\left(s' \mid s,a\right) \left(\sum_{a' \in \mathcal{A}} \pi\left(a' \mid s'\right) Q^{\pi}\left(s',a'\right) \right) \tag{19}$$

where $r(s,a) = \sum_r r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$ and $p(s' \mid s,a) = \sum_r p(s', r \mid s, a)$ denote the expected reward of taking action a in state s and the probability of transition into state s' after taking action a in state s, respectively. (19) is also known as the Bellman equation and shows that it is possible to estimate the action-value of one state-action pair through the action-values of other state-action pairs. Moreover, there are two widely-used technologies in DQL algorithm, namely the experience replay and the target network technology. The agent that employs the experience replay technology will randomly sample a group of experiences from an experience pool to update the neural network in each step. The agent that adopts the target network technology will deploy two neural networks with same structure in the training process, namely the evaluation network with parameter vector $\boldsymbol{\omega}$ and the target network with parameter vector $\boldsymbol{\omega}_{\text{target}}$. Both the experience replay and target network technologies are used to improve the stability of DQL algorithm. Based on the above discussions, the loss function of the neural network is defined as

$$L(\boldsymbol{\omega}) = \frac{1}{|\mathcal{D}|} \sum_{(s,a,r,s') \in \mathcal{D}} \left(r + \gamma \max_{a' \in \mathcal{A}} Q\left(s', a'; \boldsymbol{\omega}_{\text{target}}\right) - Q(s,a;\boldsymbol{\omega}) \right)^2 \tag{20}$$

where $|\mathcal{D}|$ is the number of experiences in the mini-batch $\mathcal{D}$, $Q(s,a;\boldsymbol{\omega})$ and $Q\left(s,a;\boldsymbol{\omega}_{\text{target}}\right)$ denote the estimated action-values with the evaluation network and the target network, respectively. Based on the loss function defined in (20), the stochastic gradient descent (SGD) is used to update the parameters $\boldsymbol{\omega}$ in each step. In this way, the estimated action-value of the network $Q(s,a;\boldsymbol{\omega})$ will gradually converge to the actual optimal action-value and thus optimal policy can be obtained.

D. The Single-Agent MDP Framework

As we mentioned before, the handover is triggered at the beginning of each HDF. Thus, we treat each HDF as one step and a single-agent MDP framework can be established to describe the handover process as shown in Fig. 3.

1) State: The state of step k is determined by[2]

$$S^{(k)} = \left[S_{i,j}^{(k)} \right]_{N \times M} \tag{21}$$

[2]Actually, the control center only needs to collect the information of its assigned users and their visible satellites to form the state. In this paper, for the convenience of expression and maintaining the dimension of the state stable, we denote the state by the information of all users and satellites in the network.

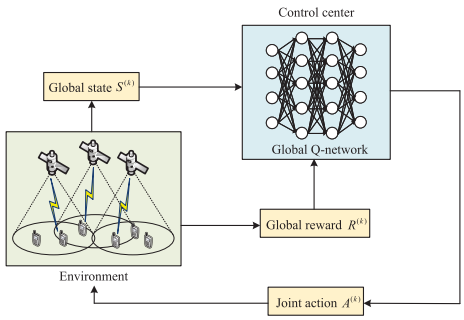

Fig. 3. The handover process in the single-agent MDP framework.

where

$$S_{i,j}^{(k)} = \left[\mathbf{L}_i \;\; \theta_{i,j}^{(k)} \;\; \theta_{i,j}^{v,(k)} \;\; \theta_{i,j}^{a,(k)} \;\; c_{i,j}^{(k)} \;\; G_{i,j}^{C,(k)} \;\; r_{i,j}^{(k)} \right] \tag{22}$$

$\mathbf{L}_i = [\varphi_i, \varsigma_i]$ is the position of user i with φ_i and ς_i denoting the longitude and latitude of user i, $\theta_{i,j}^{(k)} = \theta_{i,j}(kT_H)$ is the elevation angle between user i and satellite j in step k, $\theta_{i,j}^{v,(k)} = \theta_{i,j}^{(k)} - \theta_{i,j}^{(k-1)}$ and $\theta_{i,j}^{a,(k)} = \theta_{i,j}^{v,(k)} - \theta_{i,j}^{v,(k-1)}$ are used to characterize the dynamic of the satellite, $c_{i,j}^{(k)} = c_{i,j}(kT_H)$ is the propagation condition between user i and satellite j in step k, $G_{i,j}^{C,(k)} = G_{i,j}^{C}(kT_H)$ denotes the expected cost for user i to occupy the beam resources of satellite j in step k, and $r_{i,j}^{(k)} = r_{i,j}(kT_H)$ is the connection status between user i and satellite j in step k.

2) Action: The action includes the handover decisions of all users and can be written as

$$A^{(k)} = \left[x_{i,j}^{(k)} \right]_{N \times M} \tag{23}$$

where $x_{i,j}^{(k)}$ denotes the satellite handover decision between user i and satellite j in step k. Since the user will performed the handover decision made by the control center after the delay stage with the duration of T_D slots, we have $x_{i,j}^{(k)} = x_{i,j}(kT_H + T_D)$. In addition, the action $A^{(k)}$ needs to satisfy the constraints given by (14)-(17).

3) Reward: The reward of the system in step k is designed as the average utilities of all users that obtained in the last HDF and is given by

$$R^{(k)} = \frac{1}{T_H} \sum_{\tau=(k-1)T_H}^{kT_H} \frac{1}{N(\tau)} \sum_{i=1}^{N} G_i(t) \tag{24}$$

where $N(t)$ is the number of active users, i.e., the users in service in slot t.

Based on the above discussions, we have established the general MDP framework for the formulated optimization problem. However, the state and action space grow exponentially as the numbers of users and satellites increase, which makes it impractical to be implemented in the mega-constellation. Consequently, we convert the original single-agent MDP framework into a multi-agent version, which can efficiently reduce the state and action space. Note that, the multi-agent MDP framework is the reformulation of the single-agent MDP framework and the handover decisions are still made by the control center.

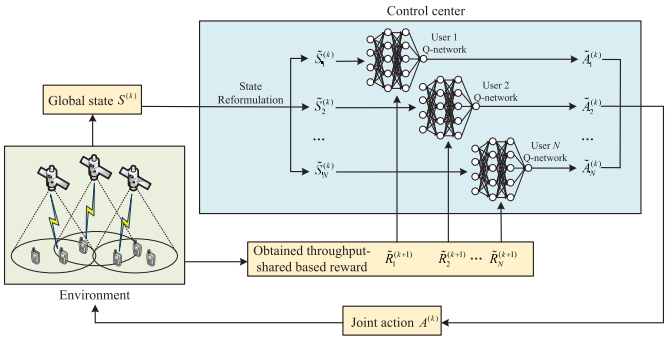

Fig. 4. The handover process in the cooperative multi-agent MDP framework.

E. The Establishment of Cooperative Multi-Agent MDP Framework

In the multi-agent MDP framework, each agent corresponds to a user and is responsible to make the handover decision for the user. Since each agent only needs to make the handover decision of one user, the state and action space can be efficiently reduced. However, as other learning agents will be the unpredictable elements of the environment, it is challenging to determine the joint actions that are optimal to the system. Consequently, cooperations should be integrated into the multi-agent MDP framework to motivate the agents to cooperate with each other and achieve the global optimal result. Since global state with shared reward [45], [46] and Q-learning algorithm variants [47] are two widely used methods to facilitate the cooperation in the multi-agent MDP framework, we also adopt them in this paper and convert the single-agent based MDP framework into a multi-agent version. As shown in Fig. 4, the established cooperative multi-agent MDP framework for user i ($i \in \mathcal{N}$) is given as follows.

1) State: Due to the existence of the interference between users, it is reasonable to build the state of each user using the information not only from itself but also from the interfering users. As the mutual interference between distanced users can be ignored due to the limited coverage of spot beams according to (1)-(3), we first define the neighborhood set for user i ($i \in \mathcal{N}$), which is denoted by

$$\mathcal{I}_i = \{i' \in \mathcal{N}, i' \neq i \mid d_{i,i'} < D\} \tag{25}$$

where D is the neighborhood distance threshold. When the distance between user i and user i' is larger than D, the spot beam steered to user i will incur little interference on user i' and vice versa. Thus, only the users in neighborhood set $\mathcal{I}_i$ will affect the handover decision of user i. Consequently, the state of user i in step k can be expressed as

$$\tilde{S}_i^{(k)} = \left[\tilde{S}_{i,j}^{(k)}\right]_{1\times M} \tag{26}$$

where $\tilde{S}_{i,j}^{(k)}$ is the substate of satellite j for user i. Specifically, the substate $\tilde{S}_{i,j}^{(k)}$ includes the local information $L_{i,j}^{(k)}$ and the neighborhood information $H_{i,j}^{(k)}$ of satellite j to user i in step k, which can be denoted by

$$\tilde{S}_{i,j}^{(k)} = \left[L_{i,j}^{(k)} \;\; H_{i,j}^{(k)}\right]. \tag{27}$$

The designs of $L_{i,j}^{(k)}$ and $H_{i,j}^{(k)}$ are described as follows.

- **Local information:** The local information of the user is defined as the information of the visible satellites (i.e., the satellites whose elevation angles are greater than the minimum visible elevation angle $\theta_{\min}$) to the user. Specifically, the local information $L_{i,j}^{(k)}$ is used to evaluate the value of satellite j to user i in step k, ignoring the existence of other users. In this case, the value of a satellite can be evaluated based on three factors. The first factor is the channel power gain from satellite j to user i. As it is impossible to estimate the accurate channel power gain in next HDF, we estimate the average channel power gain in step k by using the propagation condition $c_{i,j}^{(k)}$ and elevation angle $\theta_{i,j}^{(k)}$ based on (1)-(3). The second factor is the expected cost $G_{i,j}^{C,(k)}$ that user i has to pay for accessing satellite j in each slot in step k. The last factor is the connected status $r_{i,j}^{(k)}$ between user i and satellite j in step k. Therefore, the local information of satellite j to user i in step k can be denoted by

$$L_{i,j}^{(k)} = \left[\theta_{i,j}^{(k)} \;\; \theta_{i,j}^{v,(k)} \;\; \theta_{i,j}^{a,(k)} \;\; c_{i,j}^{(k)} \;\; G_{i,j}^{C,(k)} \;\; r_{i,j}^{(k)}\right]. \tag{28}$$

- **Neighborhood information:** The neighborhood information of the user is defined as the information of the visible satellites of the user to its neighbors. Specifically, the neighborhood information $H_{i,j}^{(k)}$ is used to estimate the interference of satellite j to user i's neighbors in step k. For each neighbor of user i, two factors are adopted to estimate the interference. The first one is the distance between neighbor i' and user i, and the other one is the average channel power gain, which can be estimated by the propagation condition $c_{i',j}^{(k)}$ and elevation angle $\theta_{i',j}^{(k)}$. Thus, the neighborhood information of user i to satellite j in step k can be denoted by

$$H_{i,j}^{(k)} = \left[H_{i,j}^{i',(k)}\right]_{1\times|\mathcal{I}_i|} \tag{29}$$

where $|\mathcal{I}_i|$ is the number of users in user i's neighborhood set $\mathcal{I}_i$ and $H_{i,j}^{i',(k)}$ is the neighborhood information of neighbor i' to user i and satellite j in step k. $H_{i,j}^{i',(k)}$ can be expressed as

$$H_{i,j}^{i',(k)} = \left[d_{i,i'} \;\; \theta_{i',j}^{(k)} \;\; \theta_{i',j}^{v,(k)} \;\; \theta_{i',j}^{a,(k)} \;\; c_{i',j}^{(k)}\right] \tag{30}$$

where $d_{i,i'}$ is the distance from neighbor i' to user i.

2) Action: In the multi-agent framework, the action is the handover decision of the user and thus the action space is determined by

$$\tilde{A}_i^{(k)} = \left[x_{i,j}^{(k)}\right]_{1\times M}. \tag{31}$$

3) Reward: Due to the interference between users and their neighbors, the throughput of one user in each slot is jointly determined by the actions of its own and neighbors. Consequently, we design a shared-throughput based reward function to facilitate the cooperation between users. In slot t, the shared-throughput of user i is the average number of data

packets sent to user i and its neighbors who have successfully accessed the satellite, which can be denoted by

$$\overline{D}_i(t) = \frac{D_i(t) + \sum_{i' \in \mathcal{I}_i} D_{i'}(t)}{1 + \sum_{i' \in \mathcal{I}_i} r_{i'}(t)} \tag{32}$$

where $r_i(t)$ denotes the access status of user i in slot t. Then, the shared-throughput based reward function of user i in slot t is given by

$$\tilde{R}_i(t) = B_P \cdot \overline{D}_i(t) - \sum_{j=1}^{M} x_{i,j}(t) G_{i,j}^{C}(t). \tag{33}$$

Moreover, the reward of user i in step k can be written as

$$\tilde{R}_i^{(k)} = \frac{1}{T_H} \sum_{\tau=(k-1)T_H}^{kT_H} \tilde{R}_i(t). \tag{34}$$

F. The Proposed MA-SHDQL Algorithm

Based on the above discussions, we have established the cooperative multi-agent MDP framework. To deal with the non-stationary environment of users, we propose a MA-SHDQL algorithm. As shown in Fig. 4, in the control center, each user has an independent deep neural network called Q-network which is responsible to map the action a in a state s to its estimated action-value $Q(s, a)$. The mapping process is completely implemented by the network without any artificial restrictions. The algorithm can be divided into decision phase and training phase.

1) The Decision Phase: The handover decision is made based on the state information with the Q-network in this phase. In step k, user i updates the state information $\tilde{S}_i^{(k)}$ according to (26)-(30). In general, the conventional DQL algorithm will take $\tilde{S}_i^{(k)}$ as the input of the Q-network, and then the Q-network will output the estimated action-values of all satellites. However, we notice that the number of satellites that satisfy the elevation angle requirement for handover is significantly smaller than the total number of satellites in the constellation, and the channel power gains between users and different satellites are independent. Consequently, it is reasonable to estimate the action-values of different satellites with the Q-network in a successive way. To be specific, in step k, the estimated action-value for handover to satellite j for user i can be denoted by

$$\tilde{Q}_i\left(\tilde{S}_{i,j}^{(k)}\right) = \begin{cases} -\infty, & \text{if } \theta_{i,j}^{(k)} < \theta_{\min} \\ \mathcal{Q}_i(\tilde{S}_{i,j}^{(k)}; \boldsymbol{\omega}), & \text{otherwise} \end{cases} \tag{35}$$

where $\theta_{\min}$ is the minimum elevation angle of the visible satellite and $\mathcal{Q}_i(\,\cdot\,;\boldsymbol{\omega})$ represents the Q-network of user i with parameters $\boldsymbol{\omega}$. It is noted that the network $\mathcal{Q}_i(\,\cdot\,;\boldsymbol{\omega})$ is only fed by the substate of one satellite at each time and outputs the estimated action-value. Thus, the size of the input layer is equal to the dimensions of the substate and the size of the output layer is one. Since the motion patterns of satellites in the constellation are the same, we can evaluate the values of different satellites with the same network. Therefore, we can estimate the action-values of accessible satellites by input their substates into Q-network successively, and a Q-table that contains the action-values of all satellites for user i in step k can be determined by $\left[\tilde{Q}_i\left(\tilde{S}_{i,j}^{(k)}\right)\right]_{1\times M}$. Then, the estimated optimal satellite for user i to perform handover in step k can be expressed as $j^* = \arg\max_{j\in\mathcal{M}} \tilde{Q}_i\left(\tilde{S}_{i,j}^{(k)}\right)$. In order to avoid the local optimal handover decision and fully explore the action space, the widely used ϵ-greedy explore strategy is adopted, where each user will choose the estimated optimal satellite j^* with the probability of $1-\epsilon$ or select a satellite from the accessible satellites randomly with the probability of ϵ. Particularly, the ϵ-greedy explore strategy can be expressed as

$$\tilde{A}_i^{(k)} = \begin{cases} j^*, & \text{with probability } 1-\epsilon \\ \text{randomly choose}, & \text{with probability } \epsilon. \end{cases} \tag{36}$$

2) The Training Phase: After users make handover decisions, they will obtain rewards from the environment and update the parameters of their Q-network. To be specific, it is assumed that user i chooses the satellite j in step k according to (36) and then obtains the reward $\tilde{R}_i^{(k+1)}$ according to (34). Then, user i will obtain an experience of satellite j for the step $k+1$, which is denoted by $\left(\tilde{S}_{i,j}^{(k)}, \tilde{A}_i^{(k)}, \tilde{R}_i^{(k+1)}, \tilde{S}_{i,j}^{(k+1)}\right)$. User i will store the generated experience into its memory pool and then randomly samples B previous experiences from it, which is denoted by $\left(\tilde{S}_i^{(z)}, \tilde{A}_i^{(z)}, \tilde{R}_i^{(z)}, \tilde{S}_i^{\prime(z)}\right)$ $(z = 1, 2, \cdots B)$. The temporal difference error of the experience z of user i is given by

$$\tilde{\delta}_i^{(z)} = \tilde{R}_i^{(z)} + \gamma \mathcal{Q}_i\left(\tilde{S}_i^{\prime(z)}; \boldsymbol{\omega}_{\text{target}}\right) - \mathcal{Q}_i\left(\tilde{S}_i^{(z)}; \boldsymbol{\omega}\right) \tag{37}$$

where γ is the discount factor and $\mathcal{Q}_i(\,\cdot\,;\boldsymbol{\omega}_{\text{target}})$ represents the target network of user i which has the same structure with the Q-network $\mathcal{Q}_i(\,\cdot\,;\boldsymbol{\omega})$.

In general, the loss function is defined as the mean square temporal difference errors of all sampled experiences. However, in the multi-agent framework, one user may receive a small reward because of the bad actions of other users even if the user has chosen an optimal action. To address this issue, the hysteretic Q-learning [48] is employed which introduces two different learning rates α_1 and α_2 $(0 < \alpha_1 < \alpha_2 < 1)$. The learning rate α_1 is used when the received reward is smaller than the expected reward, i.e., $\tilde{\delta}_i^{(z)} < 0$, and α_2 is used when the received reward is larger than the expected reward. i.e., $\tilde{\delta}_i^{(z)} \geq 0$. In this way, each user is encouraged to attach less importance to the small reward and will estimate the action-value optimistically. In this paper, we integrate the hysteretic Q-learning into the loss function design [49] and thus the loss function can be expressed as

$$\tilde{L}_i(\boldsymbol{\omega}) = \frac{1}{B} \sum_{z=1}^{B} \beta_i^{(z)} \cdot \left(\tilde{\delta}_i^{(z)}\right)^2 \tag{38}$$

where

$$\beta_i^{(z)} = \begin{cases} 1, & \text{if } \tilde{\delta}_i^{(z)} \geq 0 \\ \beta, & \text{if } \tilde{\delta}_i^{(z)} < 0 \end{cases} \tag{39}$$

Algorithm 1 MA-HSDQL Algorithm

Initialization:
Initialize learning rate α, scale factor β, discount factor γ, evaluation network $\mathcal{Q}_i(\cdot;\boldsymbol{\omega})$ and target network $\mathcal{Q}_i(\cdot;\boldsymbol{\omega}_{\text{target}})$ $(\forall i \in \mathcal{N})$, termination time T
while $t < T$ **do**
  if $t \in \mathcal{H}$ **then**
    Obtain the global state $S^{(k)}$ based on (21) and (22)
    Convert the global state $S^{(k)}$ into the states of each user $\tilde{S}_i^{(k)}$ based on (26)-(30)
    Calculate the Q-value of each satellite for each user $\tilde{Q}_i\left(\tilde{S}_{i,j}^{(k)}\right)$ based on (35)
    Select the handover satellites for each user with ϵ-greedy explore strategy based on (36)
    if $t > 0$ **then**
      Obtain the reward of each user based on (32)-(34)
      Store the experience of each user $\left(\tilde{S}_{i,j}^{(k-1)}, \tilde{A}_i^{(k-1)}, \tilde{R}_i^{(k)}, \tilde{S}_{i,j}^{(k)}\right)$ into its memory pool
      Sample a batch of experiences $\left(\tilde{S}_i^{(z)}, \tilde{A}_i^{(z)}, \tilde{R}_i^{(z)}, \tilde{S}_i^{\prime(z)}\right)$ $(z = 1, 2, \cdots B)$ for each user
      Calculate the loss function of each user $\tilde{L}_i(\boldsymbol{\omega})$ based on (38) and (39)
      Update the network $\mathcal{Q}_i(\cdot\,;\boldsymbol{\omega})$ of each user based on the calculated loss function $\tilde{L}_i(\boldsymbol{\omega})$
      Every C steps, update the target network $\mathcal{Q}_i(\cdot\,;\boldsymbol{\omega}_{\text{target}})$ of each user
    end
  end
  $t \leftarrow t + 1$
end

and $\beta \leq 1$ is the scale factor and is scheduled to increase gradually with the improvement of users' strategies. In this way, users are motivated to optimistically battle the negative updates in the early training stage. Moreover, β is also added to the substate of satellites $\tilde{S}_{i,j}^{(k)}$ such that the user can distinguish the experiences generated in different time and improves the stability of learning process. The SGD is used and the Q-network parameters $\boldsymbol{\omega}$ is updated to minimize the loss function $\tilde{L}_i(\boldsymbol{\omega})$ in each step while the target network parameters $\boldsymbol{\omega}_{\text{target}}$ will be copied from the Q-network parameters $\boldsymbol{\omega}$ every C steps. The detailed algorithm is given in Algorithm 1.

Remark 1: The developed MA-SHDQL algorithm builds a general RL-based centralized handover framework for massive LEO satellite networks, which can adapt to various environments while keeping the training and decision processes unchanged, and thus shows good scalability. Specifically, when the environment changes, we only need to add the features representing the new environment into the user state information and the agent can update the policy based on the new environment. For example, the other types of fading such as the rain attenuation can be considered in (1). In that case, the information about the fading such as the rainfall rate and the rain type should be added in (28) and (30) such that the agent can learn the relationship between the new added fading features and the obtained reward by interacting with the environment and thus the optimal policy can be achieved.

Remark 2: The developed MA-SHDQL algorithm is designed for the dynamic environments and the training process can be regarded as the process that the agents learn the dynamic characteristics of the environment. Therefore, as long as the dynamic characteristic of the environment does not show significant change, the training process of the MA-SHDQL algorithm does not need to be re-performed. Moreover, the situations that might cause the change of the dynamic characteristic of the environment such as the changes of the user geographical distribution and the channel model in different weather conditions will not affect the feasibility of the proposed scheme since we can pre-train a set of agent policies corresponding to various situations in the deployment stage of LEO satellite networks. When the dynamic characteristic of the environment changes significantly, the agent can choose one policy from the pre-trained policy set to match the current environment. In this way, the re-training time caused by the change of the environment dynamic characteristic can be avoided.

G. The Complexity Analysis

The complexity of the developed MA-SHDQL algorithm is mainly determined by the complexity of the neural network. Since we deployed one fully-connected neural network for each agent, the complexity of the MA-SHDQL algorithm for evaluating the value of one satellite can be denoted by [18], [29]

$$\Psi_{\text{MA-SHDQL}}^{\text{Network}} = O\left(\tilde{S} \cdot W_1 + \sum_{l=1}^{L-1} W_l W_{l+1}\right) \tag{40}$$

where $\tilde{S} = |\tilde{S}_{i,j}^{(k)}|$ is the dimension of the substate of satellite j for user i in step k according to (27)-(30), W_l denotes the input size of the l-th layer, W_{l+1} denotes the output size of the l layer and L is the total number of layers of the constructed neural network. As the agents corresponding to different users can perform the decision phase in a parallel way [18], the complexity of the decision phase of the MA-SHDQL algorithm is determined by the user with the highest complexity (i.e., the user with the most visible satellites) and can be written as

$$\Psi_{\text{MA-SHDQL}}^{D} = O\left(\hat{M} \times \left(\tilde{S} \cdot W_1 + \sum_{l=1}^{L-1} W_l W_{l+1}\right)\right) \tag{41}$$

where $\hat{M}$ is the maximum number of user visible satellites. Accordingly, the complexity of the training phase is determined by the structure of the neural network and the number of sampled experiences for each training step. Consequently, the complexity for one training step of each user in MA-SHDQL algorithm can be denoted by

$$\Psi_{\text{MA-SHDQL}}^{T} = O\left(B \times \left(\tilde{S} \cdot W_1 + \sum_{l=1}^{L-1} W_l W_{l+1}\right)\right). \tag{42}$$

IV. SDQL Algorithm Based Distributed Handover Scheme Design

Based on the above discussions, we have developed a MA-SHDQL algorithm based centralized handover scheme where users cooperate with each other to achieve the global optimal actions. However, the cooperation between the users requires the control center to collect the global information and perform the developed algorithm, which may lead to

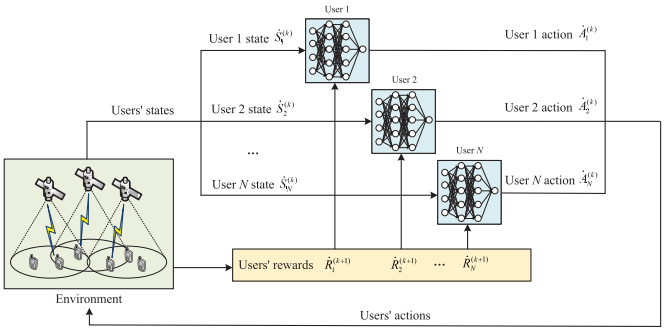

Fig. 5. The handover process in the distributed multi-agent MDP framework.

heavy signaling overheads and computation burden to the control center, especially for the LEO mega-constellations with numerous satellites and users. To address this problem, we further propose a distributed handover scheme where there is no cooperation between users and each user makes the handover decision independently only based on the local information.

The frame structure and the handover mechanism proposed in Section III-A are reserved in the distributed handover scheme. The only difference is that the users in service do not send their current information to the control center at the beginning of each HDF. Instead, they make the handover decisions by themselves based on their local information. Consequently, the delay stage is not included in the frame structure of the distributed handover scheme and thus $T_D = 0$. As the cooperation between users is not needed, the cooperative multi-agent MDP framework developed in Section III-E does not work in the distributed handover scheme. Thus, we construct a new MDP framework for the distributed handover scheme as shown in Fig. 5, which is described as follows.

A. State

In the distributed handover scheme, each user builds its state only based on its local information. The definition of the local information in the distributed scheme is the same with that in the centralized scheme, namely the information of the visible satellites to the user. Specifically, the state of user i in step k is given by

$$\dot{S}_i^{(k)} = \left[\dot{S}_{i,j}^{(k)}\right]_{1\times M} \tag{43}$$

where $\dot{S}_{i,j}^{(k)}$ is the substate of satellite j and can be expressed as

$$\dot{S}_{i,j}^{(k)} = \left[\theta_{i,j}^{(k)} \quad \theta_{i,j}^{v,(k)} \quad \theta_{i,j}^{a,(k)} \quad c_{i,j}^{(k)} \quad G_{i,j}^{C,(k)} \quad r_{i,j}^{(k)}\right]. \tag{44}$$

B. Action

The action space in the distributed scheme is the same with $\tilde{A}_i^{(k)}$ in (31), i.e.,

$$\dot{A}_i^{(k)} = \left[x_{i,j}^{(k)}\right]_{1\times M}. \tag{45}$$

C. Reward

In the distributed scheme, since each user cannot obtain the information of its neighbors, the shared-throughput based reward function in (33) is infeasible. Instead, we design the reward function based on the user's own throughput, which is equal to the utility function of the user in (12). Specifically, in the distributed scheme, the reward function of user i in step k can be expressed as

$$\dot{R}_i^{(k)} = \frac{1}{T_H}\sum_{\tau=(k-1)T_H}^{kT_H} G_i(t). \tag{46}$$

Based on the above discussions, we have established the MDP framework for the distributed scheme. Apparently, due to the lack of the global information, the coordination between different users is difficult to be implemented and the hysteretic Q-learning approach cannot be employed. However, we can still estimate the action-values of different satellites with the same network successively since the channel power gains of different satellites are still independent. Therefore, we develop a successive deep Q-learning (SDQL) algorithm to make the handover decision in the distributed scheme.

The developed algorithm can also be divided into the decision phase and training phase. In the decision phase, each user makes their handover decision according to its own state. Particularly, it is assumed that user i obtains its state $\dot{S}_i^{(k)}$ in step k based on (43). By inputting the substates $\dot{S}_{i,j}^{(k)}$ into its Q-network successively, a Q-table denoted by $\left[\dot{Q}_i\left(\dot{S}_{i,j}^{(k)}\right)\right]_{1\times M}$ is obtained, where $\dot{Q}_i\left(\dot{S}_{i,j}^{(k)}\right)$ represents the estimated action-value of the handover to satellite j for user i in step k under the given state $\dot{S}_{i,j}^{(k)}$. Moreover, $\dot{Q}_i\left(\dot{S}_{i,j}^{(k)}\right)$ can be expressed as

$$\dot{Q}_i\left(\dot{S}_{i,j}^{(k)}\right) = \begin{cases} -\infty, & \text{if } \theta_{i,j}^{(k)} < \theta_{\min} \\ \dot{\mathcal{Q}}_i(\dot{S}_{i,j}^{(k)};\boldsymbol{\omega}), & \text{otherwise} \end{cases} \tag{47}$$

where $\dot{\mathcal{Q}}_i(\cdot\,;\boldsymbol{\omega})$ denotes the Q-network of user i with parameters $\boldsymbol{\omega}$. Note that, in the distributed scheme, the Q-network is deployed in the user's terminal instead of the control center. $\dot{\mathcal{Q}}_i(\cdot\,;\boldsymbol{\omega})$ is fed by the substate of one satellite $\dot{S}_{i,j}^{(k)}$ at a time and outputs the corresponding estimated action-value of the satellite. Based on the obtained Q-table $\left[\dot{Q}_i\left(\dot{S}_{i,j}^{(k)}\right)\right]_{1\times M}$, the optimal satellite for handover is determined by $j^* = \arg\max_{j\in\mathcal{M}} \dot{Q}_i\left(\dot{S}_{i,j}^{(k)}\right)$ and the user will also choose the handover satellite with ϵ-greedy explore strategy as described in Section III-F.

In the training phase, each user trains its own network based on the received reward from the environment. Suppose that user i makes the handover to satellite j in step k, then it will receive the reward $\dot{R}_i^{(k+1)}$ in step $k+1$ according to (46) and an experience is generated which is denoted by $\left(\dot{S}_{i,j}^{(k)}, \dot{A}_i^{(k)}, \dot{R}_i^{(k+1)}, \dot{S}_{i,j}^{(k+1)}\right)$. The generated experience is stored into the memory pool and a batch of experiences $\left(\dot{S}_i^{(z)}, \dot{A}_i^{(z)}, \dot{R}_i^{(z)}, \dot{S}_i^{\prime(z)}\right)$ $(z = 1, 2, \cdots B)$ are sampled to

TABLE I
SIMULATION PARAMETERS

Notation	Value	Notation	Value
Satellite antenna coefficient K_G	-150 dBi [50], [54]	Average call duration T_m	3 min [27]
Antenna gain of users G_T	0 dBi [50]	Duration of a slot ΔT	10 ms [31]
Radius of spot beams R	15 km [39]	Duration of the handover stage T_A	100 ms [27], [32]
Neighborhood distance threshold D	37.5 km	Duration of a HDF T_H	1 s
Transmit power of spotbeam P	16 dBW [54]	Duration of a PCF T_F	3 s [33]
Bandwidth of users B	2 MHz [50]	Size of a packet S_P	1000 bits [56]
Minimal transmission rate $R_{\min}$	2 Mbps [55]	Minimum paid fee B_P	1
Carrier frequency f_I	20 GHz [50]	Minimum charging fee B_C	5
Noise power spectral density	-173 dBm/Hz [31]	Upper limit of charging coefficient K_C	5
Arrival rate λ	0.1 s^{-1} [27]	Simulation time	5 min [4]

calculate the loss function which is given by

$$\dot{L}_i(\boldsymbol{\omega}) = \frac{1}{B}\sum_{z=1}^{B}\left(\dot{R}_i^{(z)} + \gamma \dot{\mathcal{Q}}_i\left(\dot{S}_i^{\prime(z)}; \boldsymbol{\omega}_{\text{target}}\right) - \dot{\mathcal{Q}}_i\left(\dot{S}_i^{(z)}; \boldsymbol{\omega}\right)\right)^2. \tag{48}$$

The SGD is also adopted to update the network parameters $\boldsymbol{\omega}$ in order to minimize the loss function $\dot{L}_i(\boldsymbol{\omega})$ and the target network parameters $\boldsymbol{\omega}_{\text{target}}$ is updated to equal to the evaluation network parameters $\boldsymbol{\omega}$ every C steps.

Remark 3: Similar with the MA-SHDQL algorithm, the developed SDQL algorithm also shows good scalability as it can also adapt to various environments only with the minor modifications of the user state information given by (44).

Remark 4: Similar with the MA-SHDQL algorithm, the training process of the SDQL algorithm does not need to be re-performed as long as the dynamic characteristic of the environment does not show significant change. In addition, the situations that might cause the change of the environment dynamic characteristic such as the changes of the user geographical distribution and the channel model in different weather conditions also will not affect the feasibility of the proposed scheme as the agent policies corresponding to different situations can be pre-trained.

Since the SDQL algorithm also adopts a fully-connected neural network to evaluate the values of satellites, the complexity of SDQL algorithm for evaluating the value of one satellite can be denoted by

$$\Psi_{\text{SDQL}}^{\text{Network}} = O\left(\dot{S}\cdot W_1 + \sum_{l=1}^{L-1} W_l W_{l+1}\right) \tag{49}$$

where $\dot{S} = |\dot{S}_{i,j}^{(k)}|$ is the dimension of the substate of satellite j for user i in step k in SDQL algorithm according to (44). Accordingly, the decision complexity and the training complexity of the SDQL algorithm can be denoted by

$$\Psi_{\text{SDQL}}^{D} = O\left(\hat{M}\times\left(\dot{S}\cdot W_1 + \sum_{l=1}^{L-1} W_l W_{l+1}\right)\right) \tag{50}$$

and

$$\Psi_{\text{SDQL}}^{T} = O\left(B\times\left(\dot{S}\cdot W_1 + \sum_{l=1}^{L-1} W_l W_{l+1}\right)\right) \tag{51}$$

respectively.

V. SIMULATION RESULTS

A. Simulation Parameters

To evaluate the performance of the proposed schemes, we construct a OneWeb-like constellation, where 720 satellites are uniformly deployed in 18 planes. The altitude and the inclination angle of each plane are set to 1200 km and 90°, respectively. The number of available narrow bandwidth spot beams of each satellite $C_{\max} = 50$. The minimum elevation of visual satellites $\theta_{\min} = 20°$. To increase the access stability, a minimum accessible elevation θ_A is introduced and is set to $\theta_A = 25°$. In each handover decision, the users can only choose the satellites whose elevation is higher than θ_A for handover. 50 users are placed randomly according to a uniform distribution on a square area with the length of 200 km centered on (40°N,116°E). Since the speed of the satellite is much faster than that of ground users, we ignore the speed of users. Nevertheless, all users will move along with the rotation of the earth. Each user has its own neighborhood set $\mathcal{I}_i$ ($i \in \mathcal{N}$). Note that, the number of neighbors of each user may be different, which will cause that the lengths of the inputs to the algorithm are distinct. To maintain a stable input for the algorithm, we keep the number of neighbors in $\mathcal{I}_i$ as a constant c. In this paper, c is set to be the expected number of neighbors and can be calculated as $c = \mathbb{E}\left[|\mathcal{I}_i|\right] = \left\lceil \rho \cdot \pi D^2 \right\rceil$, where $\lceil\cdot\rceil$ is the ceiling function, ρ is the density of users and D is the neighborhood distance threshold. For the users whose numbers of neighbors are larger than c, we select c nearest neighbors to construct $\mathcal{I}_i$. On the other hand, for the users whose numbers of neighbors are less than c, we fill the idle elements with zeros to keep a fixed length of inputs. A fully connected network which contains three hidden layers is adopted for each user and the numbers of neurons for the three hidden layers are 200, 150 and 60, respectively. ReLU function is used as the activation function. The learning rate $\alpha = 1\times10^{-5}$ and the discount rate $\gamma = 0.7$. The size of the memory pool is 128 and the size of mini-batch $B = 32$. The exploration rate ϵ is initially set to 1 and gradually decreases to 0 while the scale factor β is set to 0.2 at the beginning and gradually increases to 1. The interval to train the target network $C = 100$. The rest parameters are provided in Table I [31], [50].

A group of Markov model parameters in [33] is selected to characterize the dynamic of propagation condition of LMS channels. As we mentioned before, the probability that the channel is shadowed or blocked is directly related to the elevation. Consequently, we divide four intervals according

TABLE II

AVERAGE LOO MODEL PARAMETERS FOR DIFFERENT ELEVATIONS AND STATES. $\delta = 20\log_{10}(e^m)$, $\Psi = 20\log_{10}(e^{\sqrt{n}})$, $\text{MP} = 10\log_{10}(2b)$

Elevation intervals	State 1: Line-of-sight			State 2: Shadowed			State 3: Blocked		
	δ(dB)	Ψ(dB)	MP(dB)	δ(dB)	Ψ(dB)	MP(dB)	δ(dB)	Ψ(dB)	MP(dB)
(20°,40°]	-0.3	0.73	-15.9	-8.0	4.5	-19.2	-24.4	4.5	-19.0
(40°,60°]	-0.35	0.26	-16.0	-6.3	1.4	-13.0	-15.2	5.0	-24.8
(60°,80°]	-0.5	1.0	-19.0	-5.6	1.2	-10.0	-12.3	4.1	-16.0
(80°,90°]	-0.25	0.87	-21.7	-6.6	2.3	-13.0	-11.0	8.75	-24.2

TABLE III

MARKOV CHAIN MATRICES $\mathbf{P}(\theta)$ AND $\mathbf{W}(\theta)$

Elevation intervals	$\mathbf{P}(\theta)$			$\mathbf{W}(\theta)$
(20°,40°]	0.8628	0.0737	0.0635	0.4000
	0.1247	0.8214	0.0539	0.2667
	0.0648	0.0546	0.8806	0.3333
(40°,60°]	0.8681	0.0952	0.0367	0.4546
	0.1300	0.8429	0.0271	0.3636
	0.0701	0.0761	0.8538	0.1818
(60°,80°]	0.8763	0.0724	0.0513	0.4666
	0.1382	0.8201	0.0417	0.2667
	0.0783	0.0533	0.8684	0.2667
(80°,90°]	0.8870	0.0562	0.0568	0.5000
	0.1489	0.8039	0.0472	0.2000
	0.0890	0.0371	0.8739	0.3000

to the elevation angle of satellites, which correspond to the elevation of (20°, 40°], (40°, 60°], (60°, 80°] and (80°, 90°], respectively. Each interval has a group of corresponding Markov model parameters and channel parameters, which describe the transition between propagation conditions and the effects of shadowing as well as multipath fading, respectively. The detailed parameters are provided in the Table II and Table III.

B. Signaling Overhead and Delay Analysis

According to the handover mechanism proposed in Section III-A, in the proposed centralized scheme, the delay stage including T_D slots is employed for the signaling exchanges between the users, the satellites, and the control center. In this section, we will determine the value of T_D by evaluating the signaling overhead and the delay of the handover process. Specifically, the handover process includes the propagation delay, the transmission delay and the processing delay, which are evaluated as follows.

1) The Propagation Delay: The handover process of the centralized scheme includes four ground-satellite propagation delays, namely the propagation delay from the user to the satellite, from the satellite to the control center, from the control center to the satellite, and from the satellite to the user. The ground-satellite propagation delay is determined by the distance between the user (the control center can be viewed as a special user) and the satellite. Based on geometry and triangular transformations, the maximum distance allowed for the communication between the user and the visible satellite can be denoted by $D_{\max} = \left(R_E^2 \cdot \sin^2\theta_{\min} + (H_S^2 + 2H_S R_E)\right)^{\frac{1}{2}} - R_E \cdot \sin\theta_{\min}$, where $R_E = 6400$ km and $H_S = 1200$ km denote the satellite's altitude and the radius of earth, respectively. Then, the maximum ground-satellite distance can be calculated as $D_{\max} = 2457.7$ km and thus the maximum ground-satellite propagation delay can be calculated by $\tau^P_{\max} = D_{\max}/c = 8.2$ ms. Consequently, the maximum overall propagation delay of the centralized scheme is $\tau^{\text{Propagation}} = 4\tau^P_{\max} = 32.8$ ms.

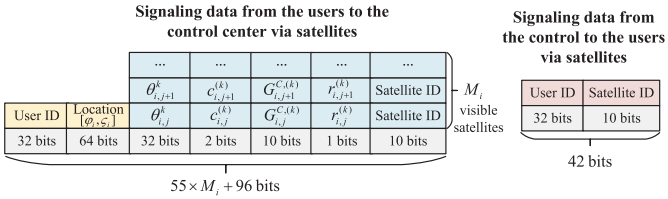

Fig. 6. The signaling data volume.

2) The Transmission Delay: Similar with the propagation delay, the handover process of the centralized scheme also includes four ground-satellite transmission delays. The transmission delay from the user to the satellite and from the satellite to the control center correspond to the transmission of the user state information, while the transmission delay from the control center to the satellite and from the satellite to the user correspond to the transmission of the handover decision made by the control center. For the user signaling data, each active user sends the state information given by (22) to the control center in the handover process according to the developed MA-SHDAL algorithm. As the elements $\theta^{v,(k)}_{i,j}$ and $\theta^{a,(k)}_{i,j}$ can be calculated by using the historical elevation angle, they do not need to be reported. Moreover, the identities (IDs) of the satellites and the users should also be reported. For the handover decision, the user ID and handover satellite ID should be reported. Therefore, for user i with M_i visible satellites, the overall signaling data volume is shown in Fig. 6. The elements $\theta^{(k)}_{i,j}$, φ_i, and ς_i are continuous variables and denoted by 32-bit floating numbers. The elements $c^{(k)}_{i,j}$, $G^{C,(k)}_{i,j}$, $r^{(k)}_{i,j}$, and IDs of the user and satellites have finite values and thus the numbers of bits required to represent them can be written as $\lceil \log_2 V \rceil$, where V is the number of possible values of the element [51]. Based on the above analysis, we can estimate the signaling data transmission delay of the proposed centralized scheme. In order to clarify the feasibility of the proposed centralized scheme, we estimate the transmission delay in a realistic scenario, where the maximum number of user visible satellites $\hat{M}$ is 30 and each satellite can serve 1000 users simultaneously in the constructed OneWeb-like constellation. As a result, the maximum signaling data volumes from the user to the satellite, from the satellite to the control center, from the control center to the satellite, and from the satellite to the user are $(55 \times 30 + 96) = 1746$ bits,

$1000 \times 1746 = 1.746 \times 10^6$ bits, $1000 \times 42 = 4.2 \times 10^4$ bits, and 42 bits, respectively. We assume that the transmission rate between the user and the satellite is 10 Mbps and the transmission rate between the satellite and the control center is 1000 Mbps [52], [53]. In this way, the maximum transmission delay of the proposed centralized scheme can be calculated by $\tau^{\text{Transmission}} = 1746/(10 \times 10^6) + 1746 \times 10^3/(1000 \times 10^6) + 42 \times 10^3/(1000 \times 10^6) + 42/(10 \times 10^6) = 2.0$ ms.

3) The Processing Delay: The processing delay of the centralized handover scheme mainly depends on the complexity of the proposed MA-SHDQL algorithm. As the training phase can be performed offline [18], [24], we only focus on the processing delay of the decision phase. It is assumed that all calculated values are represented by k-bits floating numbers. Therefore, according to the complexity analysis in Section III-G and the simulation parameters in Section V-A, the complexity of the MA-SHDQL algorithm is given by $\Psi^D_{\text{MA-SHDQL}} \approx 1.5 \times 10^6 \times k^2$. To evaluate the corresponding processing delay, we refer to the processing delay in [18], where the author constructed a fully-connected neural network with the complexity of $1.5 \times 10^5 \times k^2$ and the corresponding processing delay is 508 μs. Since the complexity of the proposed MA-SHDQL algorithm is about 10 times higher than that in [18], the processing delay of the proposed MA-SHDQL algorithm is also about 10 times longer than that in [18]. Consequently, the processing delay of the developed MA-SHDQL algorithm is $\tau^{\text{Processing}} = 10 \times 508$ μs $= 5.08$ ms.

Based on the above discussions, we can calculate the overall delay of the proposed centralized handover scheme, which is determined by $\tau^{\text{Centralized}} = \tau^{\text{Propagation}} + \tau^{\text{Transmission}} + \tau^{\text{Processing}} \approx 40$ ms. That is to say, in the worst case, after the user sends its information to the control center at the beginning of each HDF, it will receive the handover decision after 40 ms, i.e., about 4 slots later. Therefore, we set the duration of the delay stage $T_D = 4$ slot.

C. Performance Analysis

For performance comparison, we compare the proposed centralized scheme and distributed scheme with three different satellite handover schemes as follows. It is noted that the delay stage is not included in the distributed scheme and the comparison schemes since users do not need to send their information to the control center to obtain the handover decisions under these schemes.

- **Maximum Elevation in Best Propagation Conditions (ME-BPC) Scheme:** In ME-BPC scheme, the user first determines the priorities of accessible satellites based on their propagation conditions in each handover decision, where the LOS satellite has the highest priority, the shadowed satellite has a lower priority and the blocked satellite has the lowest priority. Then, the user will select the satellite with the maximum elevation from the satellites with the highest priority to perform the handover.
- **Minimum Cost (MC) Scheme:** In each handover decision, according to the occupied spot beams and the connection status of accessible satellites, each user first calculates the cost that will be paid in each slot if it decides to perform the handover to the corresponding satellite based on (10). Then, the user will select the satellite with the minimum cost to perform the handover. Note that, if there are multiple satellites with the minimum cost, the user will select one randomly from them.
- **Random Scheme:** In this method, the user will randomly select an accessible satellite in each handover decision.

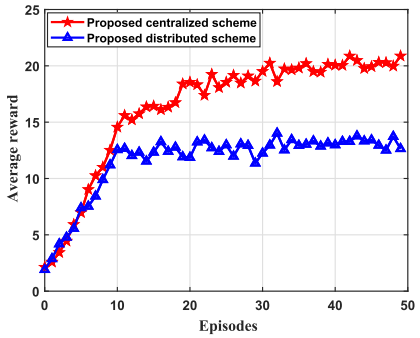

Fig. 7. The training processes of the proposed schemes.

Figure 7 shows the training processes of the centralized and distributed schemes, where the first 10 episodes is the training episodes. As we adopt the adaptive ϵ-greedy exploration strategy, where the exploration rate ϵ is initially set to 1 and gradually decreases to 0 within the training episodes [18], [29], the system performances improve significantly in the first 10 episodes. We can also observe from Fig. 7 that the system performances of the centralized scheme still improve after the training episodes. The reason is that the proposed centralized scheme is based on the multi-agent MDP framework, where each agent needs to collaborate with other agents to achieve the global optimal joint actions. As a result, the learning process will continue after the training episodes until the policies of all agents converge, which takes about 30 episodes as shown in Fig. 7. It is noted that once the training is finished, the training process is no longer required. Moreover, the the training can be executed offline such that the long on-line training time can be avoided [18], [24], [25], [30].

Figure 8 shows the performances of the average reward, average transmission rate, average cost and average handover delay[3] versus the length of user distribution area. Clearly, with the increase of the length of user distribution area, the density of users decreases and the distance between users becomes larger, resulting in less interference caused by the neighbor users. Therefore, the performances of all schemes are improved with the increase of the length of users' distribution area. Moreover, we can observe from Figs. 8(a), 8(b) and 8(d) that the proposed schemes outperform the comparison schemes in terms of the average reward, average and average handover delay. The superiority of the centralized scheme mainly comes from that each user can learn to cooperate with other users

[3]For user i, we denote by N_i^{Handover} the overall number of times that user i performs the handover to a new satellite in the simulation. In addition, we denote by T_i^{Handover} the overall number of slots that user i spends for the handover stages in the simulation. Therefore, average handover delay (AHD) can be calculated by $\text{AHD} = \Delta T \cdot \sum_{i \in \mathcal{N}} T_i^{\text{Handover}} / \sum_{i \in \mathcal{N}} N_i^{\text{Handover}}$.

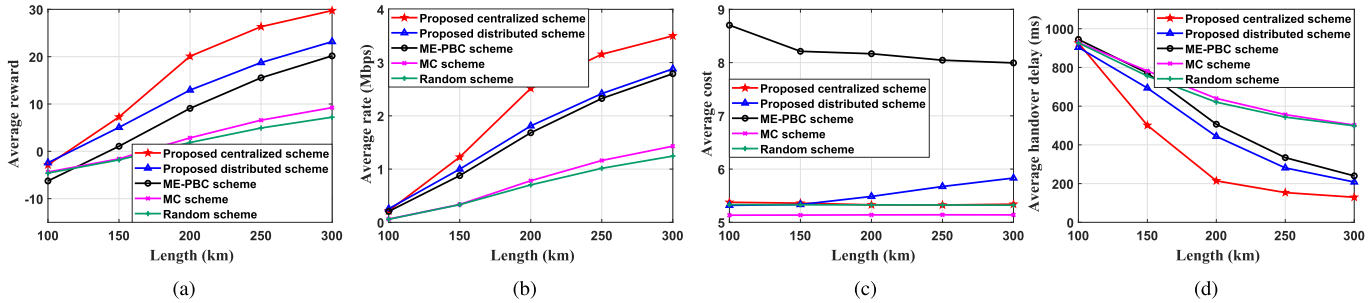

Fig. 8. Performance evaluations versus the length of user distribution area. (a) Average reward. (b) Average rate. (c) Average cost. (d) Average handover delay.

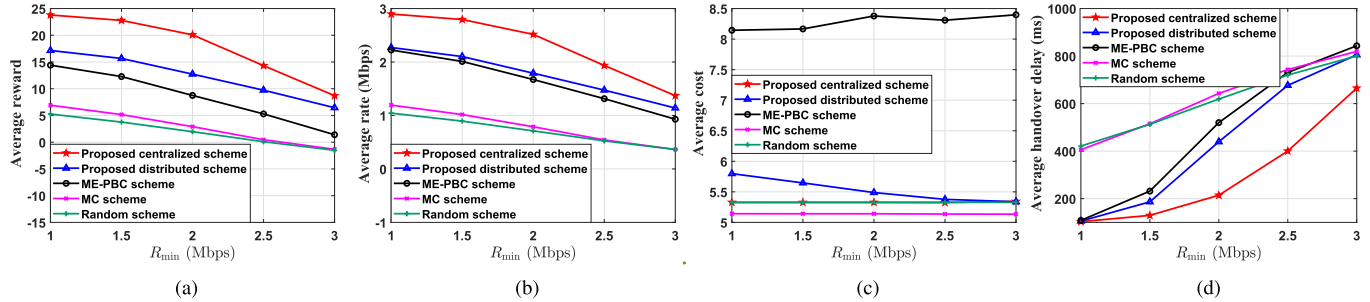

Fig. 9. Performance evaluations versus the minimum transmission rate requirement. (a) Average reward. (b) Average rate. (c) Average cost. (d) Average handover delay.

through the proposed MA-SHDQL algorithm. Thus, each user can make a good trade-off between the obtained reward and the interference caused to its neighbors to maximize the utility of the system. For the distributed scheme, although users cannot learn to cooperate due to the lack of global information, it can still achieve a high transmission rate and a low cost through the interaction with the environment, resulting in a better reward compared with the comparison schemes. In addition, it can be observed from Fig. 8(c) that the average costs of the proposed distributed schemes increase with the increase of the length of the user distribution area. It implies that users tend to access the same satellite in the proposed distributed schemes since the cost is determined by the number of occupied spot beams of the satellite.

Figure 9 depicts the performances of the proposed schemes and the comparison schemes versus the minimum transmission rate requirement. According to (7) and (9), the outage probability of the system is an increasing function of the minimum transmission rate requirement $R_{\min}$. As a result, for each user, the number of times that the user fails to perform the handover and transmit the data packets increases with the increasing $R_{\min}$. Therefore, the average transmission rate decreases and the average handover delay increases as shown in Figs. 9(b) and 9(d), which cause the decrease of the reward as shown in Fig. 9(a). However, the proposed centralized and distributed schemes still outperform the comparison schemes, verifying the superiority of our proposed schemes on improving the system performance.

Figure 10 plots the performances of the system versus the number of satellites in the constellation for different schemes. We change the number of satellites in the constellation by setting different numbers of satellites in each plane while keeping the number of planes fixed as 18. It can be observed that, with the increase of the number of satellites, the proposed centralized scheme shows significant performance superiority. The reason is that users in the centralized scheme can fully exploit the diversity of satellites based on the obtained neighborhood information. Specifically, the increasing number of satellites gives the users more opportunities to select the satellite which can not only provide a good reward but also can efficiently suppress the interference caused to their neighbors. However, the proposed distributed scheme and other comparison schemes lead users to make handover decisions selfishly without considering the interference caused to other users. Consequently, the performances under the proposed distributed scheme and comparison schemes are not improved significantly as shown in Figs. 10(a), 10(b) and 10(d). Moreover, the increasing number of satellites facilitates the inter-satellite load balancing such that the average cost of all schemes decreases, which can be observed from Fig. 10(c).

Figure 11 shows the performances of the proposed schemes and the comparison schemes versus the number of users in the network, where the number of spot beams for each satellite and the length of user distribution area are set to be to 25 and 300 km, respectively. We can observe from Fig. 11 that the performances in terms of average reward, rate, cost, and handover delay of all schemes degrade as the number of users increases. However, the proposed centralized and distributed schemes outperform the other comparison schemes significantly under the heavy traffic scenario because the proposed schemes can not only provide high transmission rate to users, but also realize efficient load-balancing between satellites. We can also observe that the average reward and rate achieved by the developed distributed scheme are getting closed to those of the proposed centralized scheme as the number of users increases. The reason for this phenomenon is that when the number of users becomes large, the overall interference caused by one user to its neighbors tends to be static no matter which satellite

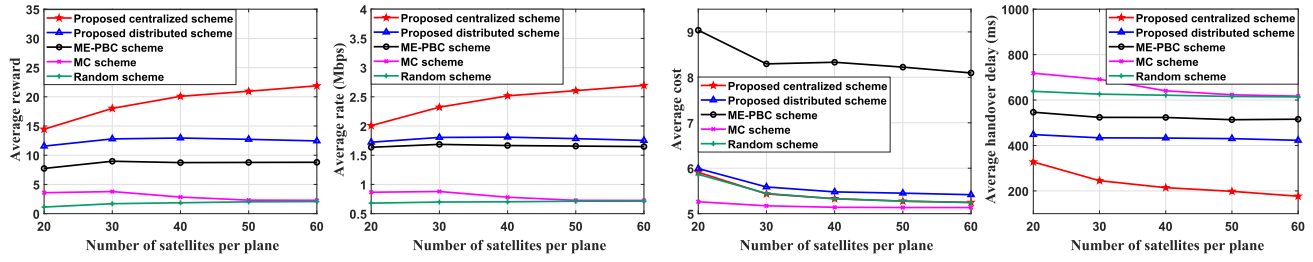

Fig. 10. Performance evaluations versus the number of satellites. (a) Average reward. (b) Average rate. (c) Average cost. (d) Average handover delay.

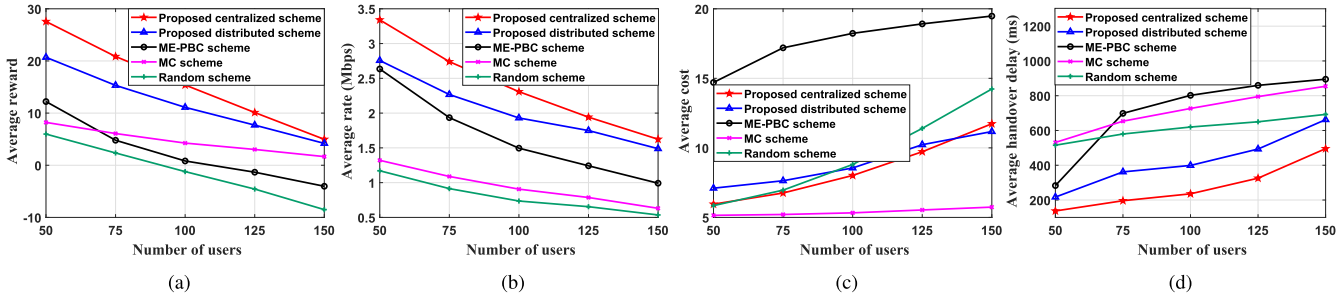

Fig. 11. Performance evaluations versus the number of users. (a) Average reward. (b) Average rate. (c) Average cost. (d) Average handover delay.

the user chooses to access, which leads to a stable interference pattern between users. In this way, although the centralized scheme enables the cooperation between neighbor users, each user tends to maximize its own throughput under the stable interference pattern, which is similar with the reward design for the developed distributed scheme.

VI. Conclusion

In this paper, a centralized adaptive intelligent handover scheme for mega-constellation was proposed, where the dynamics of propagation conditions and the limited satellite capacity were considered. Specifically, a Markov model was adopted to characterize the dynamic propagation conditions and a Loo model was employed to characterize the LMS channels. Then, the user utility function that considers the user transmission rate requirement and the load-balancing demand of satellites was developed and an optimization problem that aims to maximize the overall long-term utility function was formulated. To solve the formulated problem, we developed a low-complexity MA-SHDQL algorithm. Moreover, to reduce the signaling overhead and the computation complexity brought to the control center of the proposed centralized scheme, a distributed intelligent handover scheme was further developed, where each user makes the handover decision independently based on the local information. Simulation results showed that both the proposed centralized and distributed schemes can efficiently improve the network performance over the existing schemes.

References

[1] 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.

[2] X. Fang, W. Feng, T. Wei, Y. Chen, N. Ge, and C.-X. Wang, "5G embraces satellites for 6G ubiquitous IoT: Basic models for integrated satellite terrestrial networks," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 14399–14417, Sep. 2021.

[3] X. Lin, S. Cioni, G. Charbit, N. Chuberre, S. Hellsten, and J. Boutillon, "On the path to 6G: Embracing the next wave of low earth orbit satellite access," *IEEE Commun. Mag.*, vol. 59, no. 12, pp. 36–42, Dec. 2021.

[4] A. Al-Hourani, "Session duration between handovers in dense LEO satellite networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2810–2814, Dec. 2021.

[5] I. Ali, N. Al-Dhahir, and J. E. Hershey, "Predicting the visibility of LEO satellites," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 35, no. 4, pp. 1183–1190, Feb. 1999.

[6] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband LEO satellite communications: Architectures and key technologies," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 55–61, Apr. 2019.

[7] B. Al Homssi et al., "Next generation mega satellite networks for access equality: Opportunities, challenges, and performance," *IEEE Commun. Mag.*, vol. 60, no. 4, pp. 18–24, Apr. 2022.

[8] O. B. Osoro and E. J. Oughton, "A techno-economic framework for satellite networks applied to low earth orbit constellations: Assessing starlink, OneWeb and kuiper," *IEEE Access*, vol. 9, pp. 141611–141625, 2021.

[9] 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.*, Helsinki, Finland, Jun. 2022, pp. 1–6.

[10] 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.

[11] E. D. Re, R. Fantacci, and G. Giambene, "Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 2, pp. 397–405, Feb. 1995.

[12] 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.

[13] E. Del Re, R. Fantacci, and G. Giambene, "Handover queuing strategies with dynamic and fixed channel allocation techniques in low earth orbit mobile satellite systems," *IEEE Trans. Commun.*, vol. 47, no. 1, pp. 89–102, Jan. 1999.

[14] 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.

[15] L. Feng, Y. Liu, L. Wu, Z. Zhang, and J. Dang, "A satellite handover strategy based on MIMO technology in LEO satellite networks," *IEEE Commun. Lett.*, vol. 24, no. 7, pp. 1505–1509, Jul. 2020.

[16] 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.

[17] X. Hu, Y. Zhang, X. Liao, Z. Liu, W. Wang, and F. M. Ghannouchi, "Dynamic beam hopping method based on multi-objective deep reinforcement learning for next generation satellite broadband systems," *IEEE Trans. Broadcast.*, vol. 66, no. 3, pp. 630–646, Sep. 2020.

[18] Z. Lin, Z. Ni, L. Kuang, C. Jiang, and Z. Huang, "Dynamic beam pattern and bandwidth allocation based on multi-agent deep reinforcement learning for beam hopping satellite systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 3917–3930, Apr. 2022.

[19] G. Xu, F. Tan, Y. Ran, Y. Zhao, and J. Luo, "Joint beam-hopping scheduling and coverage control in multibeam satellite systems," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 267–271, Feb. 2023.

[20] H. Tsuchida et al., "Efficient power control for satellite-borne batteries using Q-learning in low-earth-orbit satellite constellations," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 809–812, Jun. 2020.

[21] J. Huang, Y. Yang, L. Yin, D. He, and Q. Yan, "Deep reinforcement learning-based power allocation for rate-splitting multiple access in 6G LEO satellite communication system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2185–2189, Oct. 2022.

[22] X. Li, H. Zhang, W. Li, and K. Long, "Multi-agent DRL for user association and power control in terrestrial-satellite network," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–5.

[23] K. Tsai, L. Fan, L. Wang, R. Lent, and Z. Han, "Multi-commodity flow routing for large-scale LEO satellite networks using deep reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Austin, TX, USA, Apr. 2022, pp. 626–631.

[24] P. Zuo, C. Wang, Z. Wei, Z. Li, H. Zhao, and H. Jiang, "Deep reinforcement learning based load balancing routing for LEO satellite network," in *Proc. IEEE 95th Veh. Technol. Conf.*, Helsinki, Finland, Jun. 2022, pp. 1–6.

[25] J. Liu, B. Zhao, Q. Xin, J. Su, and W. Ou, "DRL-ER: An intelligent energy-aware routing protocol with guaranteed delay bounds in satellite mega-constellations," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 4, pp. 2872–2884, Oct. 2021.

[26] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* Cambridge, MA, USA: MIT Press, 1998.

[27] H. Xu, D. Li, M. Liu, G. Han, W. Huang, and C. Xu, "QoE-driven intelligent handover for user-centric mobile satellite networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10127–10139, Sep. 2020.

[28] S. He, T. Wang, and S. Wang, "Load-aware satellite handover strategy based on multi-agent reinforcement learning," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, Jan. 2020, pp. 1–6.

[29] Y. Cao, S.-Y. Lien, and Y.-C. Liang, "Deep reinforcement learning for multi-user access control in non-terrestrial networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1605–1619, Mar. 2021.

[30] Y. Cai, S. Wu, J. Luo, J. Jiao, N. Zhang, and Q. Zhang, "Age-oriented access control in GEO/LEO heterogeneous network for marine IoRT: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 24919–24932, Dec. 2022.

[31] 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.

[32] B. Yang, Y. Wu, X. Chu, and G. Song, "Seamless handover in software-defined satellite networking," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1768–1771, Sep. 2016.

[33] F. P. Fontan, M. Vazquez-Castro, C. E. Cabado, J. P. Garcia, and E. Kubista, "Statistical modeling of the LMS channel," *IEEE Trans. Veh. Technol.*, vol. 50, no. 6, pp. 1549–1567, Nov. 2001.

[34] X. Yan, H. Xiao, K. An, G. Zheng, and S. Chatzinotas, "Ergodic capacity of NOMA-based uplink satellite networks with randomly deployed users," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3343–3350, Sep. 2020.

[35] J. Arnau, D. Christopoulos, S. Chatzinotas, C. Mosquera, and B. Ottersten, "Performance of the multibeam satellite return link with correlated rain attenuation," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6286–6299, Nov. 2014.

[36] I. Ahmad, K. D. Nguyen, N. Letzepis, G. Lechner, and V. Joroughi, "Zero-forcing precoding with partial CSI in multibeam high throughput satellite systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1410–1420, Feb. 2021.

[37] E. Zedini, A. Kammoun, and M.-S. Alouini, "Performance of multibeam very high throughput satellite systems based on FSO feeder links with HPA nonlinearity," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5908–5923, Sep. 2020.

[38] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York, NY, USA: Academic, 2007.

[39] M. Albulet, "SpaceX V-band non-geostationary satellite system: Attachment A: Technical information to supplement schedule S," Federal Commun. Commission, Washington, DC, USA, Tech. Rep. SAT-LOA-20170301-00027, 2017.

[40] Q. W. Pan, J. E. Allnutt, and C. Tsui, "Evaluation of diversity and power control techniques for satellite communication systems in tropical and equatorial rain climates," *IEEE Trans. Antennas Propag.*, vol. 56, no. 10, pp. 3293–3301, Oct. 2008.

[41] C. Loo, "A statistical model for a land mobile satellite link," *IEEE Trans. Veh. Technol.*, vol. VT-34, no. 3, pp. 122–127, Aug. 1985.

[42] A. Abdi, W. C. Lau, M. Alouini, and M. Kaveh, "A new simple model for land mobile satellite channels: First- and second-order statistics," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 519–528, May 2003.

[43] G. Corazza and F. Vatalaro, "A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 738–742, Aug. 1994.

[44] V. Mnih et al., "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[45] A. K. Agogino and K. Tumer, "A multiagent approach to managing air traffic flow," *Auton. Agents Multi-Agent Syst.*, vol. 24, pp. 1–25, Jan. 2012.

[46] T. Kagan and A. Agogino, "Distributed agent-based air traffic flow management," in *Proc. Auton. Agent Multi. Agent Syst. (AAMAS)*, Honolulu, HI, USA, Jan. 2007, p. 255.

[47] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems," *Knowl. Eng. Rev.*, vol. 27, no. 1, pp. 1–31, Feb. 2012.

[48] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic Q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams," in *Proc. IEEE Int. Conf. Intell. Rob. Syst. (IROS)*, San Diego, CA, USA, Oct. 2007, pp. 64–69.

[49] P. Xiang, H. Shan, M. Wang, Z. Xiang, and Z. Zhu, "Multi-agent RL enables decentralized spectrum access in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10750–10762, Oct. 2021.

[50] B. Deng, C. Jiang, L. Kuang, N. Ge, S. Guo, and S. Zhao, "Resource allocation of multibeam communication satellite systems in sparse networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

[51] Z. Lu, C. Zhong, and M. C. Gursoy, "Dynamic channel access and power control in wireless interference networks via multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1588–1601, Feb. 2022.

[52] S. Liu et al., "LEO satellite constellations for 5G and beyond: How will they reshape vertical domains?" *IEEE Commun. Mag.*, vol. 59, no. 7, pp. 30–36, Jul. 2021.

[53] I. del Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta Astronautica*, vol. 159, pp. 123–135, Jun. 2019.

[54] Z. Lin, Z. Ni, L. Kuang, C. Jiang, and Z. Huang, "Multi-satellite beam hopping based on load balancing and interference avoidance for NGSO satellite communication systems," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 282–295, Jan. 2023.

[55] A. Wang, L. Lei, E. Lagunas, A. I. Pérez-Neira, S. Chatzinotas, and B. Ottersten, "Joint optimization of beam-hopping design and NOMA-assisted transmission for flexible satellite systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8846–8858, Oct. 2022.

[56] X. Liu, X. Yan, Z. Jiang, C. Li, and Y. Yang, "A low-complexity routing algorithm based on load balancing for LEO satellite networks," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Boston, MA, USA, Sep. 2015, pp. 1–5.

Haotian Liu received the B.S. degree in information engineering from Xi'an Jiaotong University, China, in 2021, where he is currently pursuing the M.S. degree in information and communications engineering. His research interests include large-scale LEO satellite communication, satellite handover, beam hopping, resource allocation, and reinforcement learning.

Yichen Wang (Member, IEEE) received the B.S. degree in information engineering and the Ph.D. degree in information and communications engineering from Xi'an Jiaotong University, China, in 2007 and 2013, respectively.

From August 2014 to August 2015, he was a Visiting Scholar with the Signal and Information Group, Department of Electrical and Computer Engineering, University of Maryland, College Park, USA. He is currently a Professor with the Information and Communications Engineering School and the Vice Director of the Institute of Wireless Communications and Shaanxi Smart Network and Ubiquitous Access Center, Xi'an Jiaotong University. He has published more than 100 technical papers in international journals and conferences. His current research interests include 5G/B5G/6G technologies, random access for massive networks, ultra-reliable low-latency communications, security-related technologies for wireless networks, integrated space and terrestrial information networks, statistical QoS provisioning technology for wireless networks, and resource allocation over wireless networks. He is a member of the IEEE Communications Society and the IEEE Vehicular Technology Society. He also serves and has served as the Technical Program Committee Member for many world-renowned conferences, including IEEE GLOBECOM, ICC, WCNC, VTC, and PIMRC. He received the Best Paper Award from ICCCS in 2023, the Best Letter Award from IEICE Communications Society in 2010, and the Exemplary Reviewer Award from IEEE COMMUNICATIONS LETTERS in 2014. He served as the TPC Co-Chair for the IEEE VTC'16-Spring Workshop on User-Centric Networking for 5G and Beyond and the Track Co-Chair for the Cloud Communications and Networking of CHINACOM'17. He is currently serving as an Editor for *KSII Transactions on Internet and Information Systems*.

Peixuan Li received the B.S. degree in information engineering from Xi'an Jiaotong University, China, in 2022, where she is currently pursuing the M.S. degree in information and communications engineering. Her research interests include large-scale LEO satellite communication, resource allocation, edge-cloud collaboration, and semantic communication.

Julian Cheng (Fellow, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 1995, the M.Sc. (Eng.) degree in mathematics and engineering from Queen's University, Kingston, ON, Canada, in 1997, and the Ph.D. degree in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2003. He was with Bell Northern Research and NORTEL Networks. He is currently a Full Professor with the School of Engineering, Faculty of Applied Science, The University of British Columbia, Kelowna, BC, Canada. His research interests include machine learning and deep learning for wireless communications, wireless optical technology, and quantum communications. He is a Registered Professional Engineer in British Columbia, Canada. He served as the President for Canadian Society of Information Theory from 2017 to 2021. He was the Co-Chair of the 12th Canadian Workshop on Information Theory in 2011, the 28th Biennial Symposium on Communications in 2016, and the General Co-Chair of the 2021 and 2024 IEEE Communication Theory Workshop. He is the Chair of the Radio Communications Technical Committee of the IEEE Communications Society. He was a past Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE ACCESS; and an Area Editor of IEEE TRANSACTIONS ON COMMUNICATIONS. He served as a Guest Editor for the Special Issues on Optical Wireless Communications and Positioning and Sensing Over Wireless Networks of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.