A Reinforcement Learning Based Resource Access Strategy for Satellite-Terrestrial Integrated Networks

Jiyun Qiu[1], Hao Zhang[2], Li Zhou[3], Penghui Hu[3], and Jian Wang[3(✉)]

[1] State Grid Shanghai Municipal Electrical Power Company, Shanghai 200002, China
[2] Global Energy Internet Research Institute Co. Ltd., Nanjing 210003, China
[3] School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
wangjnju@nju.edu.cn

Abstract. The satellite-terrestrial integrated network (STIN) has recently attracted considerable attention. The problem studied in this paper is how the access controller located at the ground station selects the best-joining satellite for multi-users, who are covered by multi-satellites with limited onboard resources and high-speed moving in STIN. This paper proposes a multi-objective satellite selection strategy for multi-user based on reinforcement learning. We adopt Q-learning to continuously confirm the optimal access choice in the continuous interactive learning with the environment. We consider the multi-parameters of the integrated satellite network, including satellites' elevation angle and coverage time for users and the available channel related to the overall capacity and the traffic load. Finally, a multi-LEO satellite system for multi-user is established in STK, based on which the access algorithm is implemented. Based on the simulation, we analyze the convergence of the algorithm, and the results show that the proposed access algorithm can improve selection efficiency and user satisfaction.

Keywords: LEO Satellite · Satellite-Terrestrial Integrated Networks · Access Algorithm · Multi-targets · SDN · Reinforcement Learning

1 Introduction

The Satellite-Terrestrial Integrated Network (STIN) is promised to provide land, sea, air, and space users with any time, global coverage, on-demand services, and safe and reliable information services [1]. The Low Earth Orbit (LEO) satellites based on STIN, such as OneWeb, SpaceX, and Telesat systems, have been providing broadband Internet access services for areas with underdeveloped telecommunication infrastructure [2].

However, STIN features a heterogeneous structure with wide-area distributed and highly dynamic nodes, limited onboard wireless resources, non-negligible delay, and severe fading. Moreover, moving at a rapid speed at a low altitude,

X. Jiang (Ed.): MLICOM 2022, LNICST 481, pp. 97–107, 2023.
https://doi.org/10.1007/978-3-031-30237-4_9

LEO satellites have a relatively short period of view, causing frequent handovers between the ground terminal and the satellites. To continue communication with the counterpart, the user has to switch among the covered satellites. Up to now, SpaceX has launched nearly 3108 satellites for Starlink [3]. In this scenario, it is increasingly common to see multi-satellites covering the same area simultaneously. Such a complex and dynamic environment greatly challenges wireless resource management.

In the current STIN access selection scheme, the single target access algorithm and multi-target weighting algorithm [4] are unsuitable for access decisions for dynamic networks and fail to ensure the QoS requirements of different users flexibly. Recently, AI (artificial intelligence) algorithms have been applied to communication. Especially, Q-learning was proposed to continuously confirm the optimal access choice in the continuous interactive learning with the environment.

This paper proposes an intelligent access resource strategy based on reinforcement learning for STIN, aiming to select the best-joining satellite for multiple users covered by multi-LEO satellites with limited onboard resources at high speed. The remainder of this paper is as follows. Section 2 illustrates the reference scenario and introduces the system model. Section 3 presents the proposed access algorithm based on reinforcement learning in detail. The proposed method is eventually validated through simulation in Sect. 4. Finally, the conclusions of this paper are drawn in Sect. 5.

2 System Model

2.1 Scenario

We introduce a general formalism for the general multi-star coverage scenario. The area covered by the satellite in the ground plane is as shown in Fig. 1(a). user_1 is simultaneously located under the signal coverage of the satellites LEO_1 and LEO_2. After the satellite returns the result, the ground control center sends it to the user to complete the satellite access. Supposing the network composed of n users and m satellites, we get :

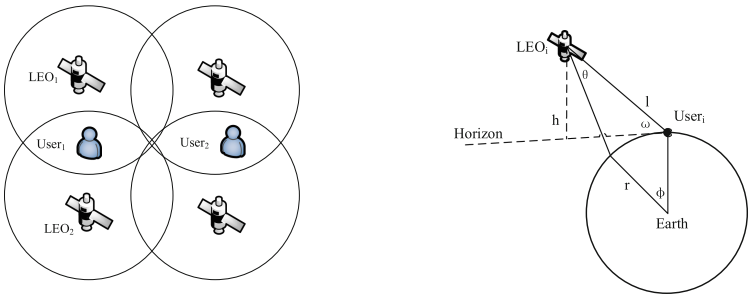
$$U = \{u_1, u_2, u_3, \ldots, u_n\} \tag{1}$$

$$S = \{s_1, s_2, s_3, \ldots, s_m\} \tag{2}$$

where U is the user set and S is satellite set. We assume multi-satellites covering the same area simultaneously. For example, if two satellites cover user_i , the set of satellites S_i that can serve the user is:

$$S_i = \{s_1, s_2\} \tag{3}$$

where $i \in \{1, 2, \ldots, n\}$.



(a) (b)

Fig. 1. Figure 1(a) shows the multi-star coverage model, and Fig. 1(b) shows the satellite coverage map of the earth . In Fig. 1(b), r represents the radius of the earth. ω represents the elevation angle of the satellite, which is the angle between the user and the satellite connection to the horizontal line, h is the height of the satellite relative to the ground, l represents the distance from the user to the satellite, and ϕ represents the satellite service area on the ground. θ represents the included angle of the satellite relative to the user.

2.2 Parameter Evaluation

We consider the multi-parameters of the integrated satellite network from the physical parameters of a single low-orbit satellite and the overall capacity of the traffic load, including the satellite elevation angle, coverage time, and the available channels.

Figure 2 shows the satellite-to-ground diagram, in which the ϕ is:

$$\varphi = \cos^{-1}\left[\frac{r}{r+h} \cdot \cos\omega\right] - \omega \tag{4}$$

Then we get the average radius of coverage area:

$$r' = r \cdot \sin\varphi \tag{5}$$

So size of the area is:

$$\mathrm{s} = 2\pi r^2 \cdot (1 - \cos\varphi) \tag{6}$$

Assuming satellites move around the earth in a uniform circular motion. The period T can be obtained as:

$$T = 2\pi\sqrt{\frac{(r+h)^3}{\mu}} \tag{7}$$

where $\mu = 398601.58\ \mathrm{km}^3/\mathrm{s}^2$ is the Kepler constant. Therefore, the coverage time of the satellite to the ground is:

$$T_s = \frac{2\varphi}{360} \cdot T \tag{8}$$

3 The Proposed Access Algorithm Based on Q-Learning

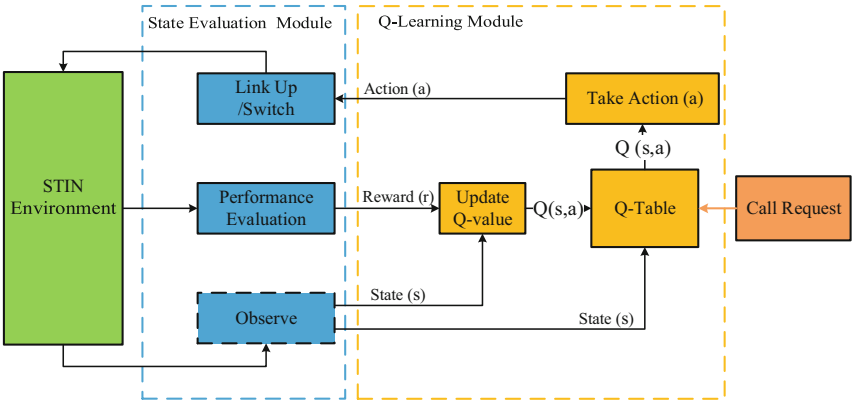


Fig. 2. Structure diagram of multi-satellite access scheme based on Q-learning.

3.1 Q-Learning Algorithm

Markov Decision Process (MDP) [5] is defined by a tuple (S, A, p, r) with explicit state transition properties. In the tuple, S represents states' finite set, A represents actions' finite set, p is a transition probability, and r represents the immediate reward obtained from state s to state s' after the execution of the action a. π is denoted as a "policy" that represents a mapping from a state to action. The goal of a time-infinite MDP is to maximize the expected discounted total reward or maximize the average reward:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma r_t(s_t, \pi(s_t))\right] \tag{9}$$

where $\gamma \in [0, 1]$ represents the discount factor, which determines the great significance of future rewards compared with the current reward. We aim to find an optimal policy $\pi\prime : \mathcal{S} \rightarrow \mathcal{A}$ and define value function $\mathcal{V}^{\pi} : \mathcal{S} \rightarrow \mathbb{R}$ that represents the expected value obtained by following policy π from each state $s \in \mathcal{S}$. The value function is:

$$\begin{aligned} \mathcal{V}^{\pi}(s) =& \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma r_t\left(s_t, a_t\right) | s_0 = s\right] \\ =& \mathbb{E}_{\pi}\left[r_t\left(s_t, a_t\right) + \gamma \mathcal{V}^{\pi}\left(s_{t+1}\right) | s_0 = s\right] \end{aligned} \tag{10}$$

As we need to find the optimal policy π^* , an optimal action at each state can be found through: $\mathcal{V}^*(s) = \max_{a_t}\{\mathbb{E}_\pi[r_t(s_t, a_t) + \gamma\mathcal{V}^\pi(s_{t+1})]\}$.

We define $\mathcal{Q}^*(s, a) \triangleq r_t(s_t, a_t) + \gamma\mathbb{E}_\pi[\mathcal{V}^\pi(s_{t+1})]$ as the optimal Q-function for all state-action pairs, then the optimal value function can be expressed as $\mathcal{V}^*(s) = \max_a\{\mathcal{Q}^*(s, a)\}$. For all state-action pairs, this can be done through iterative processes [6]:

$$\begin{aligned}\mathcal{Q}_{t+1}(s, a) =& \mathcal{Q}_t(s, a) \\ &+\alpha_t \left[r_t(s, a)+\gamma \max_{a'} \mathcal{Q}_t(s, a')-\mathcal{Q}_t(s, a)\right]\end{aligned} \tag{11}$$

The core idea behind this update is to find the Temporal Difference (TD) between the predicted Q-value.

3.2 Algorithm Structure

The overview of the proposed method is as shown in Fig. 2. State Evaluation Module is to collect the observed information of the STIN. And the Reinforcement Learning Module is the decision-making center to explore optimal access links by interacting with environmental information. The algorithm is shown in Algothrim 1. We denote $\mathcal{Q}_t^*(s, a)$ as the optimal Q-function at t. s^* and a^* is the corresponding state and action.

Algorithm 1. Q-learning-based resource access strategy for STIN

Input: For each (s, a), initialize the table entry $\mathcal{Q}(s, a)$ arbitrarily. Observe the current state s, initialize the learning rate α and the discount factor γ as Table 5.
When calls arrive **do**
1 Observe ω, t, and c in the STIN environment to get s as (12).
2 Generate random variable ρ, as (13):
if $0 \leq \rho \leq \varepsilon$ **then** Select random action a.
else Select the action $a = \arg\max Q(s, a)$.
3 Execute action a to get access
4 Obtain the immediate performance reward r as (15).
5 Update the Q-table entry:
$\mathcal{Q}_{t+1}(s, a) \leftarrow \mathcal{Q}_t(s, a)+\alpha_t\,[r_t(s, a)+\gamma \max_{a'} \mathcal{Q}_t(s, a')-\mathcal{Q}_t(s, a)]$
Until $|\mathcal{Q}_t^*(s, a) - \mathcal{Q}_{t+1}^*(s, a)| \leq 0.1$
Output: Access strategy $\pi^*(s) = \arg\max_a\{\mathcal{Q}_t^*(s, a)\}$

3.3 Q-Learning Based Access Resource Strategy Based Design

The proposed scheme designs the satellite network state as the state set, the alternative satellites as the action set, and the comprehensive network performance as the reward function of the selection strategy. The details are as follows:

Table 1. Parameter for Multi-satellite Environment in STK.

Parameter	Value
LEO number	$m = 48$
Orbit number	6
LEO number in per orbit	8
LEO height	550(km)
Orbit inclination	53°
Call arrival model	Poisson Distribution, $r \sim P(10)$
Number of calls arriving	Uniform Distribution, $n \sim U(5, 25), n \in \mathbb{N}$
Call duration	Exponential Distribution, $T \sim E(180)$(s)
The sampling period in STK	$T_s = 1$(min)
Channel capacity	240
Minimum services angle	$\omega_{\min} = 15^\circ$ or $\pi/12$
Elevation angle	$\omega \in [\pi/12, \pi/2]$
Cover time	$t \in [0, 11.94]$ (min) as (7) and (8)
Number of available channels	Uniform Distribution, $c \sim U(1, 240), c \in \mathbb{N}$

State Space. Three parameters, i.e., the satellite elevation angle ω, the coverage time t, and the number of available channels c, are considered as the state space of Q-learning. These parameters are selected based on signal strength, service continuity, and load balancing considerations. So the state space of Q-learning is: This paper considers the double-satellite coverage scenario. So the state space complete formula is as follows:

$$\mathrm{S}(\omega, \mathrm{t}, \mathrm{c}) = \{(\omega_1, t_1, c_1), (\omega_2, t_2, c_2)\} \tag{12}$$

Action Space. The action for the satellite access scenario is the set of satellites to be selected for access. In Fig. 1(a), the set of satellites covered by the user in the action space is as follows:

$$A_i = \{a_1, a_2\} \tag{13}$$

This paper adopts ϵ-greedy strategy, in which ϵ is the exploration probability. The system generates a random $\rho \in [0, 1]$ to determine whether to take the action with the maximum value or a random action according to ρ. The ϵ-greedy strategy is as follows:

$$a_\tau = \begin{cases} \arg\max Q(s, a), \varepsilon \leq \rho \leq 1 \\ random(A), 0 \leq \rho \leq \varepsilon \end{cases} \tag{14}$$

Reward Function. The observed QoS of the entire communication network is designed as the reward, including packet loss, jitter, and delay. Considering the comprehensive impact of the selection strategy on network performance, we define a utility function:

$$\mathrm{r}(\mathrm{s}, \mathrm{a}) = \alpha_\omega U_\omega(\omega^*) + \alpha_t U_t(t^*) + \alpha_c U_c(c) \tag{15}$$

Table 2. Influence of state parameters on performance indicators.

	Pack Loss	Delay Jitter	Delay
Satellite elevation	✓		✓
Coverage time		✓	
Load Balancing	✓	✓	

Table 3. Weight of the parameters affecting business in (15).

Parameters	α_w	α_t	α_c
Value	0.6	0.2	0.2

where $U_\omega(\omega^*), U_t(t^*)$, and $U_c(c)$ represent the satellite elevation angle, coverage time, and the benefit function of the available channel, respectively. α_ω, α_t, and α_c can be thought of as weights to the corresponding parameters. As shown in Table 3.

For the satellite elevation angle, the benefit function is:

$$U_\omega(\omega^*) = \sigma \left(\frac{\omega^* - \omega_{\min}}{\omega_{\min}} \right)^2 \tag{16}$$

where ω^* represents the current elevation angle, and $\omega_{\min}$ is the minimum angle that the system can provide services. $\sigma \in (0, 1)$ is a normalization parameter selected according to factors such as the geographical environment. This formula reflects that the larger the satellite elevation angle, the better the signal quality.

For the utility function of satellite coverage time, the definition is given as follows:

$$U_t(t^*) = \begin{cases} \mu \left(\frac{t_{\max}}{t_{\max} t^*} \right)^2, t_{\max} \neq t^* \\ 1, t_{\max} = t^* \end{cases} \tag{17}$$

where t^* represents the current coverage time, $t_{\max}$ is the longest satellite coverage time, and μ is a normalization parameter. This formula shows that the longer the coverage time, the better the communication quality of the user. For the load situation of the channel, we use change in the available channels before and after the action is taken to measure whether the action is beneficial for load balancing. And the function is defined as:

$$U_c(c^*) = \begin{cases} 0, \Delta c^* - \Delta c < 0 \\ 1, \Delta c^* - \Delta c > 0 \end{cases} \tag{18}$$

where c^* represents the current number of available channels. The difference in the number of available channels after the action selection measures whether the action benefits load balancing. If the difference is negative, the reward is 0. Otherwise, the reward is 1.

Finally, it is necessary to design the weights of α_ω, α_t, and α_c. We comprehensively consider delay, jitter, and packet loss rate as the QoS measure.

The effects of three state parameters ω, t, and c on these performances are in Table 2, which shows that pack loss is affected by both satellite elevation and the available channel. In contrast, the delay is only affected by the elevation. The value of the weights factor is as shown in Table 3.

4 Simulation and Result Analysis

4.1 Environment and the Parameters

We try to evaluate the availability of the algorithm in practical scenarios. The proposed access algorithm is simulated and verified. First, STK is used to build a low earth orbit (LEO) satellite to obtain satellite parameters, as shown in Table 1. After the parameters are obtained from the environment, they must be loaded into the reinforcement learning module for training through quantization processing. The specific quantization range is shown in Table 4.

Table 4. The actual parameter range corresponding to the quantized value in $U_\omega(\omega^*)$, $U_t(t^*)$, and $U_c(c^*)$ about the elevation angle, coverage time, and the number of available channels.

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Elevation Angle(ω^*)	$[15°, 33°)$	$[33°, 60°)$	$[60°, 90°)$			
Cover Time(t^*)	[0 s, 3.98 s)	[3.98 s, 7.96 s)	[7.96 s, 11.94 s)			
Channel Number(c^*)	$[0, 40)$	$[40, 80)$	$[80, 120)$	$[120, 160)$	$[160, 200)$	$[200, 240)$

The parameters of Q-Learning Module are as shown in Table 5.

Table 5. Parameter for Q-Learning Model.

Parameter	Description	Value
α	Learning rate	0.5
γ	Discount factor	0.8
ϵ	Probability choose to explore in the ϵ-greedy strategy	0.8
ρ	Decay coefficients for the probability of exploration	0.08
τ	Decay cycles for the probability of exploration	$10(s)$

4.2 Result Analysis

For this algorithm, we set the number of training rounds to 500. Then we analyze the impact of the access selection algorithm on communication performance. And the convergence process of the Q-learning algorithm model based on a real-time communication system is first analyzed. We use a single utility function

adopting the weighted sum of the satellite elevation angle, coverage time, and available channel for comparison. CLWA represents comprehensive weighting and static access algorithms in the following figures, and Q-Learning illustrates the proposed method.

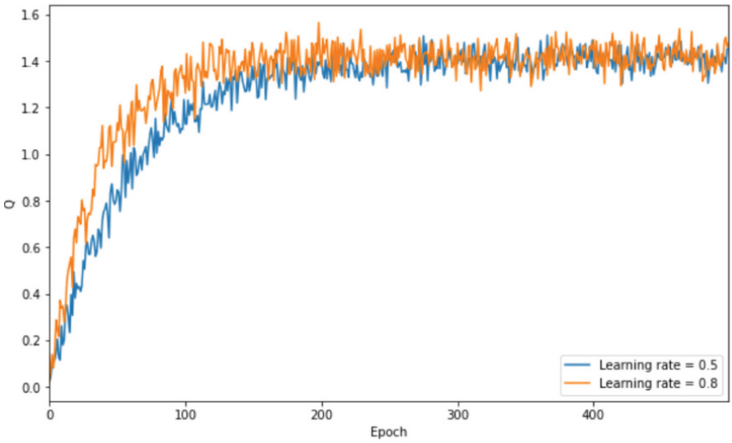


Fig. 3. Convergence Process of Q-learning.

Convergence Analysis. As the training progresses, Fig. 3 shows that the Q-learning algorithm is converging. It shows that the agent can obtain the optimal access strategy from the satellite elevation angle, coverage time and the number of available channels to explore the STIN environment. It also shows the difference in algorithm convergence when the learning rates α are 0.8 and 0.5, respectively. As demonstrated by the curve, when the learning rate is 0.8, the Q value changes faster and stabilizes earlier. α determines the learning ability. The larger the α, the faster the learning speed under the premise of convergence.

Successful Access Rate Analysis. We measure this performance with access probability, which refers to the number of calls successfully connected to the satellite to the total number of calls. The access probability is related to the access algorithm and the busyness of the network. As shown in Fig. 4, when the number of call arrivals per unit time increases from 5 to 25, the access probability of the curves corresponding to the two algorithms first remains close to 100%, then gradually decreases, and finally remains around 50%. It is because when the call arrival is relatively low, the network load is relatively small, the call requests of all users can be satisfied, and the access probability is 1. As the call arrival rate increases, the network load gradually increases. As demonstrated in Fig. 4, compared with the Q-Learning algorithm, the access probability of

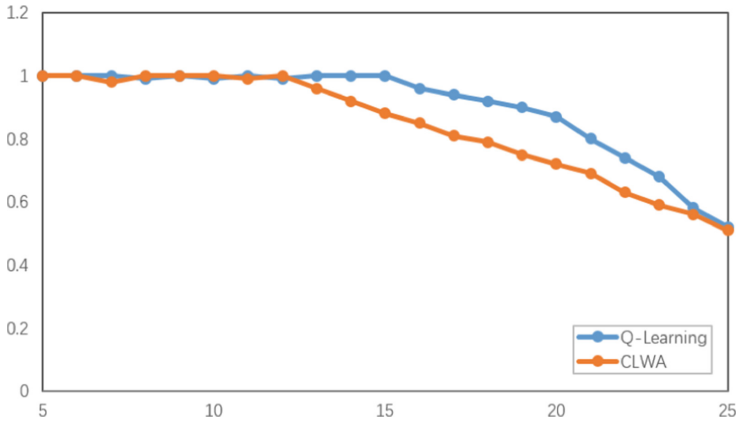


Fig. 4. Probability of complete call versus new call arrival rate.

the CLWA algorithm curve decreases first. Meanwhile, the access probability of the CLWA algorithm is lower than that of the Q-Learning algorithm, which indicates the proposed algorithm can improve the access probability of users, thereby providing higher communication quality and user satisfaction.

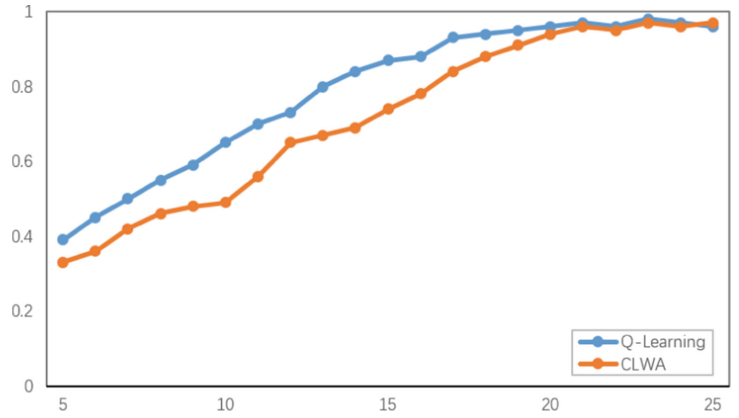


Fig. 5. Satellite channel utilization versus call arrival rate.

Network Resource Utilization Analysis. We consider the impact of this algorithm on the utilization of the entire network resources. Channel utilization refers to the difference between the successfully utilized channels and the channel capacity in STIN. As shown in Fig. 5, as the number of calls increases, the channel utilization rises and then tends to stabilize, and its value is close to 1.

When the call arrival rate is low, the network load is small, fewer channels are needed at this time, and the channel utilization rate is low. At the same time, it shows that the channel utilization rate of the CLWA algorithm is lower than that of the Q-Learning algorithm, and the time to reach the highest channel utilization rate is relatively late. It shows that the proposed Q-Learning-based access algorithm can better allocate the channel resources of the STIN and improve channel utilization.

5 Summary

This paper proposes a multi-objective integrated satellite access algorithm based on Q-learning for the satellite-terrestrial integrated network (STIN), aiming to select the optimal access satellite for multiple users covered by multi-LEO satellites with limited channel resources. We consider the multi-parameters, including the elevation angle of satellites, the coverage time, and the available channel related to the traffic load. According to the QoS requests, we design the access problem as a multi-objective optimal problem and adopt reinforcement learning to select the satellite. Finally, an LEO-based STIN is simulated in STK, and the proposed algorithm is implemented. Based on the results, we analyze the convergence of the algorithm and verify that the algorithm provides more efficient access selection by analyzing user satisfaction and network resource utilization.

Acknowledgement. This work is supported by the Science and Technology Project of State Grid Shanghai Municipal Electrical Power Company. (SGSHXT0 0YJJS2100140).

References

1. Liu, J., Shi, Y., Fadlullah, Z.M., Kato, N.: Space-air-ground integrated network: a survey. IEEE Commun. Surv. Tutor. **20**(4), 2714–2741 (2018). https://doi.org/10.1109/COMST.2018.2841996
2. del Portillo Barrios, I., Cameron, B., Crawley, E.: A technical comparison of three low earth orbit satellite constellation systems to provide global broadband. Acta Astronautica **159** (2019). https://doi.org/10.1016/j.actaastro.2019.03.040
3. Wikipedia: Starlink – Wikipedia, The Free Encyclopedia (2022). https://en.wikipedia.org/wiki/Starlink. Accessed 26 Aug 2022
4. Li, C., Zhang, Y., Hao, X., Huang, T.: Jointly optimized request dispatching and service placement for MEC in LEO network. China Commun. **17**(8), 199–208 (2020). https://doi.org/10.23919/JCC.2020.08.016
5. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 6382–6393. Curran Associates Inc., Red Hook (2017)
6. Luong, N.C., et al.: Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun. Surv. Tutor. **21**(4), 3133–3174 (2019). https://doi.org/10.1109/COMST.2019.2916583