Stability and Satisfaction Index Optimization for Beam Allocation in Mega LEO Constellation

Wenduo Pei
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China
wenduo@gs.zzu.edu.cn

Ruijie Zhu
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China
zhuruijie@zzu.edu.cn

Jingbo Wei
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China
weijingbo@gs.zzu.edu.cn

Yudong Zhang
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China
zhang_zyd@163.com

Wenchao Zhang
School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China
zhangwenchao9066@163.com

Chao Xi
Satellite Communications Business Division
Space Star Technology CO., LTD
Beijing, China
xichaofh@163.com

Bo Yang
Satellite Communications Business Division
Space Star Technology CO., LTD
Beijing, China
bagoiyb@126.com

Abstract—**We propose a D3QN-PER based approach to solve complex beam allocation problem in mega LEO satellite constellation. Simulation results show that this method can improve communication stability and user demand satisfaction rate.**

Keywords—Mega LEO constellation, beam allocation, deep reinforcement learning

I. Introduction

Satellite communication can cover oceans, deserts and other areas which are difficult to be served by terrestrial networks. Compared with geostationary earth orbit (GEO) satellites, low earth orbit (LEO) satellites are characterized by short transmission delay, low propagation loss and high bandwidth. The LEO satellite communication system has gradually moved towards large-scale and multi-functional integration, such as Oneweb and Starlink [1]-[4]. Most LEO constellations are composed of multi-beam satellites. The satellite providing multiple high-gain spot beams and more flexible resource allocation to improve satellite service quality.

Beam allocation is a key part of LEO constellation networking. Different from the single satellite beam allocation scenario, the high density and high dynamic of mega LEO constellation lead to multiple coverage and frequent connection handover for users [5]-[7]. Since the increasing available satellites allows users to choose beams from different satellite resource pools, dynamic access satellite selections should be considered additionally in the beam allocation problem. Moreover, co-frequency interference will deteriorate signal reception in adjacent user clusters [8]. How to avoid redundant handovers and minimize the impact of interference is the goal of designing a beam allocation scheme.

Several existing researches on selecting access satellites for users have been investigated. A graph-based handover framework that takes remaining service time, satellite load and elevation as reference factors is proposed to generate handover schemes with different preferences [9]. The access satellite selection is modeled as a weighted matching problem and a greedy algorithm is proposed to solve the optimization problem in [10], which reduces handovers and interruptions during communication. The performance of traditional heuristic algorithms will be greatly challenged with the rapid increase of satellites and users. [11] uses a deep reinforcement learning based algorithm to reduce the relay delay in the mobile user scenario. To reduce the interference between different LEO constellations, [12] deploys the DRL model on the user terminal to select satellites independently.

Resource allocation of multi-beam satellites has been intensively studied by many researchers to satisfy the diversified demands of users. The problem of beam arrangement and channel allocation for LEO constellation is proposed in [13], and a static user clustering with frequency allocation scheme based on heuristic algorithm is given. [14] considers the different delay tolerance of areas which the satellite serves and the real-time bandwidth resources of satellite, DRL approach is used to allocate beam bandwidth to satisfy user demands while reducing bandwidth consumption. For the resource allocation of beam-hopping satellites, [15] takes the freedom of beams in time, space and bandwidth into account, the beam pattern design and bandwidth allocation scheme based on multi-agent DRL is proposed to reduce communication delay and improve the system throughput.

In fact, the solutions of access satellite allocation problem and frequency band assignment problem are interrelated and jointly determine the result of beam allocation, few of existing studies consider these two parts comprehensively. In this paper, we formulate a multi-objective optimization problem to jointly solve the two sub-problems by transforming satellites resource pools into a beam allocation matrix. Then we apply a D3QN-PER based beam allocation (DP-BA) approach to tackle the optimization problem. After obtaining the allocation scheme, each satellite emits beams to the user clusters it serves with the specific frequency band. Simulation results show that the proposed method can generate beam allocation scheme with long connection duration and high user demand satisfaction rate.

II. System Model

In our beam allocation scenario, the mega LEO constellation is modeled as set $S = \{s_1, s_2, \dots, s_N\}$, where N

This work was supported in part by the National Natural Science Foundation of China under Grant 62001422, 62036010, and Henan Scientific and Technology Innovation Talents No. 22HASTIT016.

is the total number of multi-beam satellites. We select an area on the earth, users distributed in this area can be represented as set $U = \{u_1, u_2, \dots, u_V\}$, and all users in the area are divided into M user clusters, the set of user clusters is denoted as $C = \{c_1, c_2, \dots, c_M\}$. We consider the allocation scheme in a time period $T = \{t_1, t_2, \dots., t_K\}$, which t_i represents the ith time slot, and the length of each time slot is du. As shown in Fig. 1, a user cluster can be covered by multiple satellites simultaneously, and the coverage time of a satellite to the user clusters is only a few minutes, the user clusters need to switch between different satellites to ensure continuous communication.

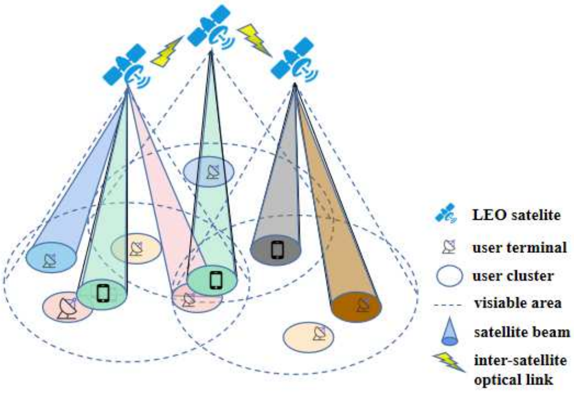

Fig. 1 The structure of mega LEO constellation network

We presume that the total bandwidth of each satellite is B, and B is divided into n chunks. The frequency band set of each satellite is described as $F = \{f_1, f_2, \dots, f_n\}$. Frequency multiplexing technology is used in satellite-ground communication. We assume that each frequency band can reuse r times for a satellite, the reuse factor equals to r accordingly. Fig. 2 shows an example which $n = 4$ and $r = 2$. Each satellite beam serves a user cluster with a specific frequency band f_i , $i \in n$. Therefore, a single satellite can emit $n \times r$ beams at most in a time slot.

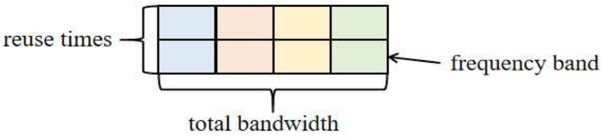

Fig. 2 The frequency band of LEO satellite

For a user cluster, if the access satellite and frequency band are determined, the beam will stare the cluster in t through phased array antenna. When solving the beam allocation problem, there are some constraints need to be satisfied:

- Visibility constraint: We denote $Vis_{i,j}^t$ as visible relationship between satellite i and user cluster j. It can be defined as:

$$Vis_{i,j}^t = \begin{cases} 1, & C_j \text{ is covered by } S_i \\ 0, & otherwise \end{cases} \tag{1}$$

  If a beam is assigned to a user cluster, the cluster must be visible to the satellite which the beam belongs to.

- Connection constraint: We use $Sat_{i,j}^t$ and $Fre_{k,j}^t$ to represent whether the satellite s_i and frequency band f_k are allocated to user cluster c_j in time slot t respectively, they can be formulated as:

$$Sat_{i,j}^t = \begin{cases} 1, & s_i \text{ is allocated to } c_j \\ 0, & otherwise \end{cases} \tag{2}$$

$$Fre_{k,j}^t = \begin{cases} 1, & f_k \text{ is allocated to } c_j \\ 0, & otherwise \end{cases} \tag{3}$$

  To simplify the model, we assume that each user cluster has the same demand and can be assigned at most one satellite and one frequency band during t.

- Interference constraint: Two user clusters whose distance is less than threshold μ cannot be assigned the same frequency band of same satellite to avoid interference, so we define the user clusters interference set $I_j = \{c_1, c_2, \dots, c_m\}$, and $c_m \in C$ represents the cluster have interference constraint with cluster c_j.

In this paper, we achieve beam allocation for user clusters through joint selection of access satellites and frequency bands. As shown in Fig. 3, we reconstruct the satellites resource pools into a beam allocation matrix, and combine $Sat_{i,j}^t$ and $Fre_{k,j}^t$ into $Conn_{i,j,k}^t$, which can be expressed as:

$$Conn_{i,j,k}^t = \begin{cases} 1, & c_j \text{ connects with } s_i \text{ at } f_k \\ 0, & otherwise \end{cases} \tag{4}$$

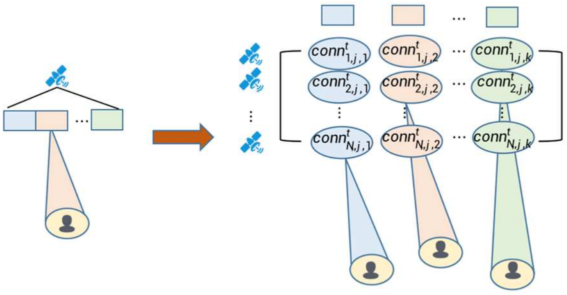

Fig. 3 The beam allocation matrix in mega LEO constellation

Then we define the connection duration time between the user cluster c_j and satellite s_i in the whole time period T as Dur_j^i, HO_j is handover times of c_j in T. Based on the constraints described above, we model the beam allocation multi-objective joint optimization problem as follow:

$$opt. \max_{Conn_{i,j,k}^t} \begin{cases} \dfrac{\sum_{j=1}^{M} \sum_{i=1}^{N} Dur_j^i / HO_j}{M} \\ \dfrac{\sum_{t=1}^{T} \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{n} Conn_{i,j,k}^t}{M * T} \end{cases} \tag{5}$$

$$\text{s.t. } C1: \sum_{k=1}^{n} \sum_{i=1}^{N} Conn_{i,j,k}^t \leq 1, \forall j \in C, \forall t \in T$$

$$C2: Conn_{i,j,k}^t + Conn_{i,x,k}^t \leq 1, \forall j \in C, \forall x \in I_j, \forall t \in T$$

$$C3: \sum_{j=1}^{M} Conn_{i,j,k}^t \leq r, \forall i \in S, \forall t \in T, \forall k \in F$$

$$C4: Conn_{i,j,k}^t \in [0,1], \forall i \in S, \forall t \in T, \forall k \in F, \forall j \in C.$$

III. The Proposed DRL Based Beam Allocation

A. Clustering Users by a Greedy Algorithm

The locations of user clusters need to be determined before allocating beams to them. In the selected area, we cluster all users in U into multiple user clusters according to users' position, and a cluster can be served by a single beam and vice versa. Without considering the number of users in the cluster, dividing users into fewer clusters can save beam resource, but it is a minimum clique cover problem which is NP-hard [16].

It's not the focus of our paper, so we propose a greedy algorithm to obtain relatively few user clusters, the solution may be not the optimal.

First, we construct an undirected graph G, with each user $u_i \in U$ as a node i. If the distance between two users u_i and u_j less than a threshold μ, then they meet the condition that they can coexist in a same cluster, and there is an edge between the two nodes i and j. Then we find the node n_{max} with the highest degrees in G, form a cluster which contains n_{max} and other nodes connected with it, then delete them from G, repeat above process until each node belongs to a cluster, the specific process is shown in Algorithm 1.

Algorithm 1. *A Greedy algorithm for user clustering*

```
Input: users position, beam coverage radius, distance threshold $\mu$
Output: center point list
Initialize: node set, edge set
For i=0, M do:
    For j=i, M do:
        Calculate $Dis_j^i$ between $u_i$ and $u_j$
        If $Dis_i^i < \mu$:
            Append (i, j) to edge set
    End For
End For
Construct undirected graph G through node set and edge set
While node set $\neq \emptyset$:
    Find $n_{max}$ with highest degrees in G and nodes connect with $n_{max}$
    Append the position of $n_{max}$ to center point list
    Delete $n_{max}$ and nodes connect with $n_{max}$ from node set
    Update node set, edge set, G
End While
Return center point list
```

B. Deep Reinforcement Learning

With the powerful perception and excellent decision-making ability, Deep Reinforcement Learning (DRL) is widely applied in solving communication resource allocation problems [17]-[19]. DRL algorithm mainly contains three key factors: state, action and reward. The DRL agent observes the environment to get a high-dimensional observation and extract current state s_t from the observation and takes a specific action a_t according to s_t, the environment will send the agent a reward r_t as the feedback after the action a_t is performed. Through continuous interaction with the environment, the agent can store a series of states, actions and rewards, and constantly update the policy based on these information until the policy reaches the optimum [20].

To obtain the optimal beam allocation scheme, the proposed DP-BA algorithm is based on Dueling Double Deep Q-Network with Prioritized Experience Replay (D3QN-PER) which belongs to DQN algorithm family, compared with Natural DQN (NDQN), D3QN-PER solves the problems such as sparse high rewards and overestimated Q-value [21], accelerates the convergence. After the agent put state s into neural network, different from Q-network of NDQN, dueling network disassembles the network output into a state value and a sequence of Advantage, which can help the agent distinguish the influence of actions and states on Q value. The action-value function of Dueling DQN can be defined as:

$$Q(s,a,\theta) = V(s;\theta) + A(s,\mathrm{a};\theta) - \frac{\sum A(s,a,\theta)}{|A|} \qquad (6)$$

V represents the state value and A represents action advantage, θ is eval network parameter, then reconstruct V and A into a Q-value sequence, the action a_t with highest Q-value will be selected to interact with environment and the agent will get a reward r_t and the next state s_{t+1},then a experience e_t: (s_t, a_t, r_t, s_{t+1}) will be stored in the replay buffer. D3QN integrates the algorithm idea of double DQN on the basis of dueling network. By the interaction between eval network and target network, the overestimation of Q value is avoided, we can describe the target value of D3QN as:

$$y_t^{D3QN} = r_{t+1} + \gamma Q(s_{t+1}, argmax_a Q(s_{t+1}, a, \theta); \varphi) \quad (7)$$

φ is parameter of target network. When training the network model, the agent sample m experiences from the buffer to update the policy, Prioritized Experience Reply (PER) uses the TD-error δ to measure the value of experience e_t, and calculates the priority value of e_t by Equation (5), PER makes the experiences with higher priority are more likely to be sampled to improve convergence speed. ϵ is a tiny constant to ensure the experience with lowest priority also has the probability to be selected from the buffer.

$$Priority_i = \frac{|\delta + \epsilon|}{\sum|\delta + \epsilon|} \qquad (8)$$

the MSE loss function can be calculate as;

$$L_\theta = (y_t^{D3QN} - Q(s_j, a_j, \theta))^2 \qquad (9)$$

C. The proposed DP-BA Algorithm

Beam allocation of mega LEO constellation is a complex sequential decision problem. We need to transform it into a multi-objective optimization problem based on DRL to apply DP-BA algorithm. The environment state, action space and reward function need to be defined for algorithm application.

State: We decouple the beam allocation problem in period T into the sub-problem in each time slot t .Since the visible satellites of each user cluster may be different, we define the union of the visible satellites of all user clusters in t as set $S_u^t = \{v_1, v_2, \dots, v_L\}$, for satellite $v_i \in S_u$ we represent its local properties for user cluster c_j by a vector $z_{i,j}^t = (rt_{i,j}^t, ls_{i,j}^t, rb_{i,j}^t, uf_{i,j}^t)$. $rt_{i,j}^t$ represents the remaining coverage time slots for v_i to c_j, if v_i is not visible to c_j, $rt_{i,j}^t = 0$, which is calculated by the satellite ephemeris data. We use a binary variable $ls_{i,j}^t$ to denote whether v_i connects with c_j in $t-1$. The remaining frequency chunks of v_i in t is defined as $rb_{i,j}^t$, and $uf_{i,j}^t$ is a vector can be expressed as $uf_{i,j}^t = [bf_{i,j,1}^t, bf_{i,j,2}^t, \dots, bf_{i,j,n}^t]$, $bf_{i,j,k}^t \in [0,1]$ is used to judge whether f_k of v_i can be allocated to c_j, if the allocation violates the limitation of frequency reuse times or the interference constraint, $bf_{i,j,k}^t = 0$, otherwise 1. So the state of c_j in t is defined as:

$$state_j^t = \{z_{1,j}^t, z_{2,j}^t, \dots, z_{L,j}^t\} \qquad (10)$$

Action: The beam allocation problem we proposed contains satellite allocation and frequency band allocation simultaneously, so we define $Act1_j^t = \{a_{s_1}, a_{s_2}, \dots, a_{s_L}\}$ and $Act2_j^t = \{a_{b_1}, a_{b_2}, \dots, a_{b_n}\}$ to represent the action space of satellite selection and frequency band respectively. Then the total action space of c_j in t can be obtained as:

$$Act_j^t = \{a_1, a_2, \dots, a_{L*n}\} \qquad (11)$$

and $a_m = \left(a_{s_i}, a_{b_j}\right), a_m \in Act_j^t, a_{s_i} \in Act1_j^t, a_{b_j} \in Act2_j^t$, even the action is a combination of two sub-actions.

Reward: Our objective of beam allocation scheme is to increase connection duration between user clusters and satellites while improving the user cluster demand satisfaction rate. We define $Ava_{i,j,k}^t$ and the reward function $reward$ as follow:

$$Ava_{i,j,k}^t \begin{cases} 1, & if\ rt_{i,j}^t > 0\ and \sum_{k=1}^{n} bf_{i,j,k}^t > 0 \\ 0, & else \end{cases} \tag{12}$$

$$reward = \begin{cases} \alpha \dfrac{rt_{i,j}^t}{rt_{max}} + \beta ls_{i,j}^t + \gamma \dfrac{rb_{i,j}^t}{mn}, & if Ava_{i,j,k}^t = 1 \\ P, & else \end{cases} \tag{13}$$

$Ava_{i,j,k}^t$ is used to reflect whether v_i and f_k can be allocated to c_j. Three weight factors α, β, γ are positive values between 0 and 1, because we want to maximize the average connection duration between user cluster and satellite by α and β, and we reduce the possibility of interference by encourage allocate the satellite with more remaining frequency chunks to user cluster. P is a negative number used to penalize unavailable allocation.

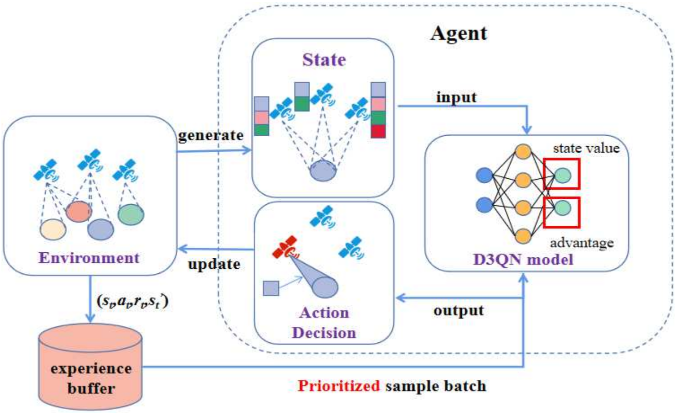

Fig. 4 The model architecture of DP-BA

As shown in Fig. 4, the agent extracts the visible satellites feather sequence from the environment and generate state s, then puts the state into D3QN model to get a sequence of Q-value, ξ-greedy strategy is applied to select satellite and frequency band from the action space, after the action a is performed the reward r will be obtained, the agent updates the state and gets new state s' ,then stores the experience into replay buffer, samples the experience from buffer to update network parameter until loss converges, the detailed procedures of DP-BA algorithm is shown in Algorithm. 2.

IV. Simulation Result

In order to evaluate the performance of the proposed DP-BA algorithm in mega LEO constellation, the simulation is based on the first shell of Starlink Phase 1, and in the ground segment, we set a rectangular simulation area which latitude and longitude are range from 25 to 45 and 105 to 125. To verify the superiority and robustness of our algorithm, 1000, 2000 and 3000 users are randomly generated in the simulation area and divided into 190, 280 and 350 user clusters respectively by Algorithm. 1. Table.1 shows the detailed parameter settings of LEO constellation and the D3QN model in our simulation. We assume that the total bandwidth B of a single satellite is divided into 4 chunks and each frequency band reuse factor is set to be 2. One objective of our algorithm is to optimize the average connection time, which is represented by the ratio of the time a user cluster is served to the handover times between different beams. Another objective is the average demand satisfaction rate of user clusters which defined as the ratio of the number of user clusters which get beams to the total user clusters, two comparison algorithms are listed as follows:

Algorithm 2. *The Training Process of DP-BA Algorithm*

```
Initialization: D with capacity $N_D$, network parameter $\boldsymbol{\theta_e}$, $\boldsymbol{\theta_t}$
For episode=0, E do:
  For t=0, SN do:
    Obtain the union of visible satellites of all user clusters $S_v^t$
    For j=0, M do:
      Generate the current state $State_j^t$.
      Agent puts $State_j^t$ into Q-network ($\boldsymbol{\theta_e}$) and chooses the action $a_j^t$ according to $\varepsilon$-greedy strategy.
      Interact with the environment through $a_j^t$, update the environment and get next state $State_{j+1}^t$.
      Obtain the reward $r_j^t$ according to (13).
      Store experience ($State_j^t, a_j^t, r_j^t, State_{j+1}^t$) to $D$.
    End For
    Samples transitions ($State_i^t, a_i^t, r_i^t, State_{i+1}^t$) through PER from $D$ and input them Q-network ($\boldsymbol{\theta_t}$) for training.
    Set $y_i^t = r_{i+1}^t + \gamma Q(state_{i+1}^t, \max_{a_{i+1}^t} Q(state_{i+1}^t, a_{i+1}^t, \theta_t); \theta_e)$
    Update $\boldsymbol{\theta_e}$ by executing a gradient descent on $L(\theta_e)$ (9)
    Set $\boldsymbol{\theta_t}$=$\boldsymbol{\theta_e}$ every $C$ steps.
  End For
End For
```

Entropy Weighting Algorithm (EWA): Calculate the weights of the attributes mentioned in the *state* formulation of every visible satellite of user cluster through entropy and assign the satellite and frequency band with the highest weighted value to the user cluster.

First-fit Selection Algorithm (FSA): Assign the first visible satellite with available frequency band to the user cluster.

TABLE I. Simulation Parameters

PARAMETERS	VALUES
LEO CONSTELLATION PARAMETERS	
Total number of satellites N:	1584
Number of orbit planes O:	72
Number of satellites per plane:	22
Satellite altitude H:	550km
Minimum elevation EL:	25°
Number of frequency bands n:	4
Frequency reuse factor r:	2
Time period T:	86400s
Number of time slots K:	1440
D3QN-PER PARAMETERS	
Training Episode:	100
Learning rate:	0.001
Size of replay buffer N_D:	1000
Target net update frequency C:	200
Mini-batch size m:	128
Discount factor γ:	0.9

In our simulation, we first record the change of the total reward in the training process of Natural DQN and our algorithm respectively to prove that our algorithm is more effective in solving beam allocation problem, the comparison result is shown in Fig. 5:

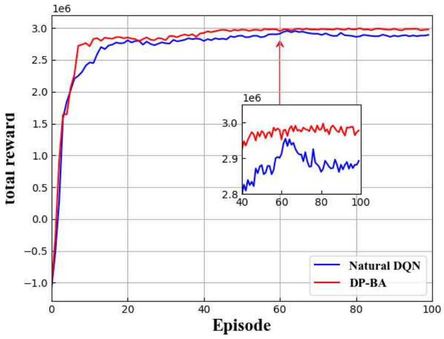

Fig. 5 The comparison of total reward convergence

NDQN converges after around the 70 episodes, while DP-BA algorithm converges within 50 episodes. Our algorithm converges more quickly and the final total reward is higher.

Then, we compare our algorithm with EWA and FSA on the average demand satisfaction rate and average connection time between satellite and user cluster respectively.

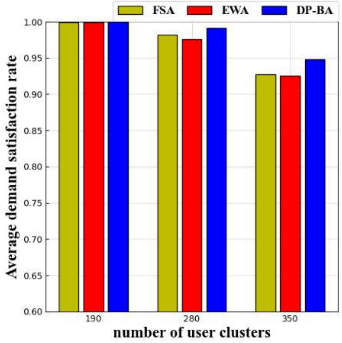

Fig. 6 The comparison of average demand satisfaction rate

Fig. 6 shows that three algorithms can allocate beams for almost all user clusters when the demand in the experimental area is low and as the number of user clusters in the region increases, some user clusters cannot obtain resources due to limited beam resources. However, compared with the other two algorithms, our algorithm allocates beams to more user clusters and leads to higher demand satisfaction rate.

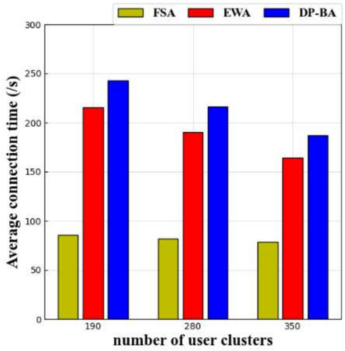

Fig. 7 The comparison of average connection time

As shown in Fig. 7, our algorithm performs better in terms of the average connection time between the beam and the user cluster. Compared with the entropy weighting algorithm, our algorithm increases the connection time by about 12%-15%. Longer connection time means fewer handovers, which improves the stability of communication and reduces signaling overhead.

V. Conclusion

In this paper, we divide the beam allocation problem of the mega LEO constellation into two sub-problems: access satellite allocation and frequency band allocation for user clusters, then formulate the two sub-problems as a multi-objective joint optimization problem and propose a DP-BA algorithm to solve it. After experimental verification, our algorithm performs better than entropy weighting algorithm and first-fit selection algorithm in mega LEO constellation in terms of connection duration and demand satisfaction rate.

References

[1] Chen et al, "System integration of terrestrial mobile communication and satellite communication —the trends, challenges and key technologies in B5G and 6G," China Communications, pp. 156-171, 2020.

[2] R. Zhu et al, “Load-Balanced Virtual Network Embedding Based on Deep Reinforcement Learning for 6G Regional Satellite Networks,” TVT, 2023.

[3] R. Ding et al, "5G Integrated Satellite Communication Systems: Architectures, Air Interface, and Standardization" WCSP, pp. 702-707, 2020.

[4] P. Wang et al, "Mega-Constellation Design for Integrated Satellite-Terrestrial Networks for Global Seamless Connectivity", IWCL, vol. 11, no. 8, pp. 1669-1673, 2022.

[5] H. Xie et al, "LEO Mega-Constellations for 6G Global Coverage: Challenges and Opportunities", IEEE Access, vol. 9, pp. 164223-164244, 2021.

[6] Y. Lee et al, "Connectivity Analysis of Mega-Constellation Satellite Networks With Optical Intersatellite Links", TAES, vol. 57, no. 6, pp. 4213-4226, 2021.

[7] D. Cui Qin, et al. "QoE-Aware Intelligent Satellite Constellation Design in Satellite Internet of Things," IoTj, pp. 4855-4867, 2021.

[8] E. Kang et al, "Link Budget Analysis of Low Earth Orbit Satellites Considering Antenna Patterns and Wave Propagation in Interference Situations," ISAP, pp. 511-512, 2022.

[9] Z. Wu et al, "A Graph-Based Satellite Handover Framework for LEO Satellite Communication Networks," ICL, pp. 1547-1550, 2016.

[10] S. Zhang et al, "A Multi-objective Satellite Handover Strategy Based on Entropy in LEO Satellite Communications", ICCC, 723-728. 2020.

[11] H. Xu et al, "QoE-Driven Intelligent Handover for User-Centric Mobile Satellite Networks," TVT, vol. 69, pp. 10127-10139, 2020.

[12] J. Wang et al, "Deep Reinforcement Learning-based Satellite Handover Scheme for Satellite Communications," WCSP, 1-6, 2021.

[13] N. Pachler et al. "Static beam placement and frequency plan algorithms for LEO constellations." ISCN, 2020.

[14] Y. He et al "Efficient Resource Allocation for Multi-Beam Satellite-Terrestrial Vehicular Networks: A Multi-Agent Actor-Critic Method With Attention Mechanism," TITS, vol. 23, pp. 2727-2738, 2022.

[15] Z. Lin et al. "Dynamic Beam Pattern and Bandwidth Allocation Based on Multi-Agent Deep Reinforcement Learning for Beam Hopping Satellite Systems," TVT, pp. 3917-3930, 2022.

[16] J. Mark Keil et al, "Approximating the minimum clique cover and other hard problems in subtree filament graphs", Discrete Applied Mathematics, pp. 1983-1995, 2006.

[17] R. Zhu et al, “DRL Based Deadline-Driven Advance Reservation Allocation in EONs for Cloud-Edge Computing,” IoTj, pp. 21444-21457, 2022.

[18] R. Zhu et al, "Deep Reinforcement Learning Based Virtual Network Embedding for 6G Satellite Networks," OECC, pp. 1-3, 2021.

[19] R. Zhu et al, "DRL-Based Deadline-Driven Advance Reservation Allocation in EONs for Cloud–Edge Computing," IoTj, pp. 21444-21457, 2022.

[20] R. Zhu et al., “Auto-learning Communication Reinforcement Learning for Multi-intersection Traffic Light Control,” KBS, 2023.

[21] J. Yan et al, "Dueling-DDQN Based Virtual Machine Placement Algorithm for Cloud Computing Systems," ICCC, pp. 294-299, 2021.