Deep Reinforcement Learning-Based Channel and Power Allocation in Multibeam LEO Satellite Systems

Junrong Li, Fuzhou Peng, Xijun Wang, and Xiang Chen(✉)

School of Electronics and Information Engineering, Sun Yat-Sen University, Guangzhou, China
{lijr75,pengfzh}@mail2.sysu.edu.cn,
{wangxijun,chenxiang}@mail.sysu.edu.cn

Abstract. With the continuous growth in communication demand, improving the efficiency of resource allocation becomes crucial. Furthermore, flexible resource allocation for meeting the non-uniform and time-varying traffic demand has emerged as an important task in multi-beam satellite systems. To improve power utilization and meet dynamic traffic demand, this paper formulates an optimization objective that minimizes the trade-off between the unmet traffic demand and power consumption. This is realized by optimizing the allocation of channel and their power, while considering the impact of co-channel interference(CCI). We propose the deep reinforcement learning (DRL) technique to optimize resource allocation. Simulation comparisons between our proposed algorithm and benchmark schemes show its effectiveness in achieving a balance between power allocation and traffic demands. Notably, our algorithm outperforms others in terms of power consumption and meeting traffic demand.

Keywords: Multibeam satellite system · Dynamic resource allocation · Deep reinforcement learning

1 Introduction

Satellite communication is crucial in aviation, maritime, and rescue fields due to its extensive coverage and powerful communication capabilities [1]. With the continuous growth of traffic demand, some challenges, such as spectrum congestion and power limitations [2], have exacerbated. Therefore, optimizing resources allocation to improve resource efficiency are crucial. Furthermore, flexible resource allocation adapting to the distribution of traffic demand has become necessary for the future multibeam satellite systems [3]. Fortunately, the application of flexible payloads and multibeam provides opportunities for

This work was supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010158001.

X. Chen et al. (Eds.): IoTaaS 2023, LNICST 585, pp. 103–116, 2025.
https://doi.org/10.1007/978-3-031-70507-6_9

advanced resource allocation strategies [4]. These technology enables the flexibility of resource allocation and promote the efficiency of satellite communication systems. Therefore, numerous studies have focused on how to improve flexible and efficiency of resource allocation in satellite communication systems.

Many works have studied flexible resource allocation strategies to meet non-uniform traffic among beam in satellite systems. These strategies can be mainly divided into three types: bandwidth allocation [5], power allocation [6], and joint allocation of both bandwidth and power [7–9]. The authors [7] introduced a joint power and channel resource allocation scheme to match the asymmetric traffic demand among beam. The work [8] proposed a noval objective function that aims to closely match the non-uniform traffic demand and consider fairness among the beams. However, these studies mainly focus on meeting traffic demands alone, ignoring minimize the power usage. Considering that power consumption impacts satellite lifetime, minimization of power consumption becomes critical in satellite systems [10].

Some work [11–13] attempted to meet the traffic demand while minimizing the power consumption. More precisely, the authors in their work [11] proposed a multi-objective approach that not only achieved the satisfaction of traffic demand but also considered minimizing power consumption. To achieve this, a two-stage heuristic algorithm was employed to solve it. The work [12]shared the same optimization objective, but it applied a successive convex approximation algorithm. The work [13] focused on minimizing utilized power and bandwidth conditioned on satisfying the traffic demand, and it used a successive convex approximation approach to solve the non-convex problem. These iterative algorithms have a specific convergence time and overlook the correlation information among time-varying traffic, which might be not suitable for dynamic traffic. Compared with traditional static strategies, DRL can adapt to dynamic environments. Some DRL methods have been applied [14–16], effectively leveraging the inherent time and spatial correlations of the dynamic traffic demand. In their work [14], the authors optimized the allocation policy aimed to minimize both power consumption and unmet system demand. However, their approach simply allocated power for each beam, with only one channel for each beam, thereby overlooking the critical issue of CCI. The work [15] dynamically adjusted bandwidth allocation strategy adapting to time-varying traffic demand. The authors in the work [16] combined DRL and Simulated Annealing (SA) to adapt to uncertain and dynamic demand. However, their approach assumed equal power allocation for channel within the same beam and does not consider minimizing power consumption during the optimization process.

In this paper, we investigate the flexible and efficient resource allocation strategy for multibeam satellite system. In order to meet the time-varying and heterogeneous traffic demand while reducing power consumption, we formulate an optimization objective that aims to minimize both the unmet traffic demand and power consumption. This is realized by optimizing the allocation of channel and their power, while considering the impact of CCI. To solve this problem, we formulate it as a Markov decision process (MDP) and apply a model-free

DRL method specifically proximal strategy optimization (PPO) algorithm. In the simulation, we evaluate the proposed method and two baseline schemes within dynamic traffic demand. Our method shows its effectiveness in achieving the balance between power consumption and traffic demand.

The rest of the paper is structured as follows. Section 2 describes the system model and formulates the optimization problem. Section 3 formulates the problem as a Markov decision process. In Sect. 4, we present PPO based resource allocation. Section 5 presents the simulation results. Finally, Sect. 6 concludes the paper.

2 System Model and Problem Formulation

2.1 System Model

We focus on the downlink of a multibeam Low Earth Orbit (LEO) satellite system which consists of N_{b} beams ($\mathcal{N} = \{i | i = 1, 2, \ldots, N_{\mathrm{b}}\}$) and K channel ($\mathcal{K} = \{k | k = 1, 2, \ldots, K\}$). The total available bandwidth is denoted as B_{tot} and each channel bandwidth is B_{sc}, where $B_{\mathrm{sc}} = B_{\mathrm{tot}}/K$. The total bandwidth is reused by all the beams, i.e. the frequency reuse factor is 1. The system model is illustrated in Fig. 1. Assuming the LEO satellite uses earth fixed cells scenario [17], the area covered by the satellite beam remains fixed. Furthermore, we assume that there is a single super user in the center of each beam, which represents the aggregation of the overall beam demand. The dynamic time-varying requested traffic of all the beams is expressed as $D_i(t) = \{D_1(t), D_2(t), \ldots, D_{N_b}(t)\}$.

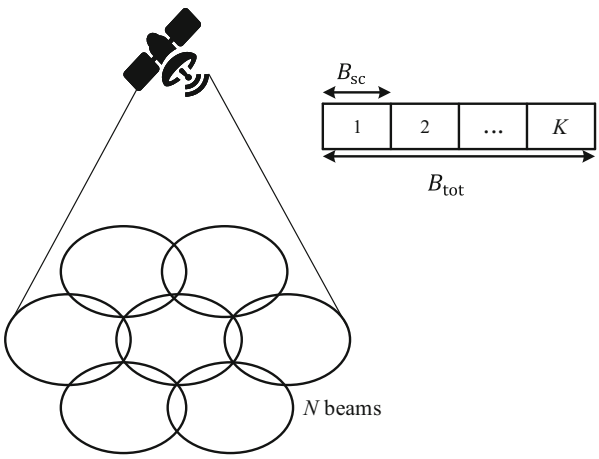

Fig. 1. An illustration of system model.

The channel coefficient from satellite to the user within the coverage of beam i is defined as:

$$h_i = \frac{a_i \sqrt{G_i G_r}}{4\pi f_n d_i / c}, \tag{1}$$

where G_i is the transmission antenna gain from satellite feed towards the beam i, G_i is calculated by $G_i = G_{\mathrm{MAX}} - \frac{12 G_{\mathrm{MAX}}}{\eta}\left(\frac{\theta_i}{70\pi}\right)^2$, and θ_i is the angle between the antenna axis of beam i and user. In addition, G_r is the receiver antenna gain of user, a_i is the small-scale fading model with Rician distribution, c and f_n are the light speed and the frequency of channel respectively, d_i is the distance between satellite and ground user within the beam i.

Considering the channel interference of different beams, the Signal-to-Interference plus Noise Ratio (SINR) of beam i on the channel k is given by

$$SINR_{i,k} = \frac{g_i \cdot x_{i,k} \cdot p_{i,k}}{\sum_{j=1, j\neq i}^{N_b} g_j \cdot x_{j,k} \cdot p_{j,k} + N_0 B_{sc}}, \tag{2}$$

where g_i is the channel gain, i.e., $g_i = |h_i|^2$. $x_{j,k}$ is a binary variable that indicates whether the channel k in the beam j is occupied or not, $x_{i,k} = 1$ indicates the channel is occupied. In addition, $p_{j,k}$ denotes the transmit power allocated to channel k in the beam j and N_0 is the noise power density.

According to Shannon's formula, the capacity of the channel k in the beam i is calculated as follows

$$C_{i,k} = B_{sc} \log_2 \left(1 + SINR_{i,k}\right). \tag{3}$$

The offered capacity of the beam i thus is given by:

$$C_i = \sum_{k=1}^{K} C_{i,k}. \tag{4}$$

2.2 Problem Formulation

In the multibeam satellite system, power and channel are dynamically allocated to meet dynamic demand. The objective of this paper is to satisfy time-varying and non-uniform traffic demand while optimizing power consumption. To evaluate the level of traffic satisfaction, the unmet system capacity (USC) [8] is adopted as a metric, which is the capacity not satisfied in the satellite system, defined by

$$USC = \sum_{i=1}^{N_b} \max\left(D_i - C_i, 0\right), \tag{5}$$

where D_i and C_i is the traffic demand and offered capacity of the beam i respectively. The proposed optimization problem is formulated as follows:

$$\begin{aligned} \min \quad & \sum_{t=1}^{T}\left[\sum_{i=1}^{N_b}\max\left(D_i - C_i, 0\right) + w\sum_{i=1}^{N_b}\sum_{k=1}^{K} p_{i,k}\right] \\ \text{s.t.} \quad & \text{C1: } x_{i,k} \in \{0,1\}, \forall i \in \mathcal{N}, \forall k \in \mathcal{K}, \forall t \in \{1,\ldots,T\}, \\ & \text{C2: } 0 \le p_{i,k} \le x_{i,k} P_{max}, \forall i \in \mathcal{N}, \forall k \in \mathcal{K}, \forall t \in \{1,\ldots,T\}. \end{aligned} \tag{6}$$

In problem (6), the optimization goal is formulated to minimize both the USC and the total power consumption, and the predefined weighted factor w represents the importance of the total power consumption compared to the USC. In constraint C1, $x_{i,k}$ is a binary variable that indicates whether a channel is occupied or not, with $x_{i,k} = 1$ indicating the channel is occupied, and $x_{i,k} = 0$ otherwise. Constraint C2 represents that the power is allocated to channel k only if the channel is occupied ($x_{i,k} = 1$), and the power allocated to each channel should be less than the maximum power limit P_{max}.

3 MDP Formulation

In this work, we consider the LEO satellite as the agent and formulate the resource allocation problem (6) as a Markov Decision Process (MDP). An MDP consists of states, actions, rewards, state transition probabilities. However, due to the difficulty in modeling the state transition probabilities, we adopt a model-free reinforcement learning algorithm. This type of algorithm learns directly from the interaction with the environment without requiring an explicit model of environment. Following is the MDP.

3.1 State

The state at time slot t is defined as

$$S_t = \{D_t, P_{t-1}, B_{t-1}\}. \tag{7}$$

Here, D_t represents the traffic demand requested by each beam at the current time t and is denoted as $D_t = \{D_i(t)\}_{i\in\mathcal{N}}$. The second component, P_{t-1}, is the power allocated to each channel at the previous time slot $t-1$. It is expressed as $P_{t-1} = \{p_{i,k}(t-1)\}_{i\in\mathcal{N},k\in\mathcal{K}}$. The final component of the state, B_{t-1}, denotes the satisfaction of traffic demand for each beam at the previous time slot $t-1$. This is given by $B_{t-1} = \{B_i(t-1)\}_{i\in\mathcal{N}}$, where $B_i(t-1)$ is computed as the ratio $C_i(t-1)/D_i(t-1)$.

3.2 Action

The agent needs to determine the allocation of channel and power at time slot t based on the current state. We can use the single variable $p_{i,k}(t)$ to indicate the allocation of channel and power at time slot t. More specifically, when $p_{i,k}(t) = 0$, it indicates that channel k is not allocated to beam i,i.e., $x_{i,k} = 0$. Conversely,

when $p_{i,k}(t) \neq 0$, it indicates that channel k is allocated to beam i, and the value of $p_{i,k}(t)$ determines the power allocated to that channel. Thus, the action taken by the agent at time slot t is defined as

$$A_t = \{p_{i,k}(t)\}_{i\in\mathcal{N},k\in\mathcal{K}}. \tag{8}$$

It should be noted that power values are within the range of 0 to P_{max}, and they are continuous.

3.3 Rewards

The reward function takes into account both the USC and power consumption. The reward received by the agent in time slot t is formulated as

$$R_t = \sum_{i=1}^{N_b} \max\left(D_i(t) - C_i(t), 0\right) + w \sum_{i=1}^{N_b}\sum_{k=1}^{K} \frac{p_{i,k}(t)}{P_{\max}}, \tag{9}$$

where $P_{\max}$ is the maximum power limit of each beam, and w controls the trade-off between the USC and power consumption. A higher value of w would place more emphasis on reducing the power consumption, while a lower value of w would prioritize reducing the USC.

4 PPO Based Resource Allocation

Algorithm 1. PPO based Resource Allocation

Initialize the value network $V_\omega(s)$, policy network π_θ with θ. Initialize $\lambda, \gamma, \epsilon$, total episode E, epoch N, time slot T
for episode $= 1, 2, \cdots, E$ **do**
  for time slot $= 1, 2, \cdots, T$ **do**
    Run old policy $\pi_{\theta_{old}}$ in environment
    Save the trajectory (s_t, a_t, r_t, s_{t+1})
  end for
  Estimate advantage $\hat{A}_t = \sum_{t'\geq t} (\lambda\gamma)^{t'-t}\delta_t$
  for epoch $= 1, 2, \cdots, N$ **do**
    Compute $L^{\text{critic}} = \frac{1}{2}\left(r_t + \gamma V_\omega(s_{t+1}) - V_\omega(s_t)\right)^2$
    Compute L^{clip} according to (10)
    Update ω by $\nabla_\omega L^{\text{critic}}$
    Update θ by $\nabla_\theta L^{\text{clip}}$
  end for
  $\theta_{old} \leftarrow \theta$
end for

In this section, we will use the PPO algorithm to optimize the channel and power allocation of the system based on the MDP model developed.PPO is a policy gradient-based reinforcement learning algorithm, which includes the policy network and the state-value network.

The policy network aims to learn the optimal policy to maximize long-term cumulative rewards. The objective function of policy network proposed by OpenAI [18] is shown as

$$L^{\text{clip}}(\theta) = \mathbb{E}_t \left[\min \left(p_t(\theta)\hat{A}_t, \text{clip}\left(p_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right], \tag{10}$$

where $p_t(\theta))$ is the policy probability ratio between the new policy and the old policy, which is defined as

$$p_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}. \tag{11}$$

The $\text{clip}\left(p_t(\theta), 1-\epsilon, 1+\epsilon\right)$ is used to limit the range of $p_t(\theta)$ to $(1-\epsilon, 1+\epsilon)$, where ϵ is a hyperparameter to control the clipping range.

The symbol $\hat{A}_t$ denotes an estimator of the advantage function at time slot t, which is calculated using a technique called Generalized Advantage Estimation (GAE) [19]. The estimator is written as

$$\hat{A}_t = \sum_{t' \geq t} (\lambda\gamma)^{t'-t}\delta_t. \tag{12}$$

where $\delta_t = r_t + \gamma V_\omega\left(s_{t+1}\right) - V_\omega(s_t)$ and $V_\omega(s_t)$ is a state-value function of policy at time slot t.

The objective of state-value network is defined as follows

$$L^{\text{critic}} = \frac{1}{2}\left(r_t + \gamma V_\omega(s_{t+1}) - V_\omega(s_t)\right)^2, \tag{13}$$

where $\delta_t = r_t + \gamma V_\omega\left(s_{t+1}\right) - V_\omega(s_t)$ and $V_\omega(s_t)$ is a state-value function of policy at time slot t, and γ is the discount factor.

Table 1. Simulation Parameters

Parameters	Values
Satellite altitude	1200km
Frequency band f_n	12GHz
Number of beams N_b	19
Beam radius	100km
Number of channels K	4
System Bandwidth B_{tot}	1000M
channel bandwidth B_{sc}	125M
Maximum power per channelP_{max}	8W
Noise power spectral density N_0	-174dBW/Hz
User antenna gain G_r	37.7dBi
Maximum satellite beam antenna gain	36.7dBi

PPO based resource allocation is presented in Algorithm 1. The algorithm consists of two stages: initialization and training. During the initialization stage, the parameters of the network and the parameters of the satellite resource allocation scenario are initialized. In each round of the training stage, the algorithm interacts with the environment for T time slots using the old policy and store the trajectory (s_t, a_t, r_t, s_{t+1}). Subsequently, the advantage function is estimated by the GAE method. Next, we calculates the loss functions of the policy network and the critic network, and updates the policy network and the critic network for N epochs using gradient-based methods. After multiple rounds of episode updates, the algorithm has converged, and we save the resource allocation policy for decision-making.

5 Simulation and Discussion

5.1 Simulation Setup

The simulation parameters [20] and training hyperparameter are shown in Table 1 and 2, respectively. The requested traffic demand is dynamically changing with time and space. The dynamic traffic model over time is based on the analysis and modeling of internet traffic in the Milan area. The average traffic demand of beams follows a sin curve with a period of 24 h [21]. In terms of spatial variability, the traffic demand among different beams is uneven at any given moment. Traffic demand in each beam follows a Gaussian distribution with a mean equal to the current average rate demand and a standard deviation is 0.2 times the the current average rate.

Table 2. Training Hyperparameters

Parameters	Values
Discount factor γ	0.9
Learning rate of actor	1e-4
Learning rate of critic	5e-3
Number of time slots per episode T	1024
Number of epochs per episode N	30
clip range ϵ	0.2
λ	0.9
Optimizer	Adam
Network Hidden sizes	[128,128]

5.2 Performance Evaluation Metrics

In order to evaluate the performance of the proposed resource allocation algorithm in the multi-beam satellite system, the following evaluation metrics have been defined:

1) USC: the total unsatisfied capacity across all beams, is defined by Eq.(5).
2) Due to the limited power resources of the satellite system, the total power consumption of the system is used as a metric, which is formulated as

$$PC = \sum_{i=1}^{N_b} \sum_{k=1}^{K} p_{i,k}. \tag{14}$$

To validate the performance of the proposed algorithm, this paper compares it with the following two algorithms:

1) Uniform power: a method that assigns equal power to all the channels with a frequency reuse factor of 1, i.e. $p_{i,k} = P_{\max}$.
2) SA: using SA algorithm for power allocation with the objective of minimizing the USC, with the same power allocation used as the starting point [8].

5.3 Performance Comparison

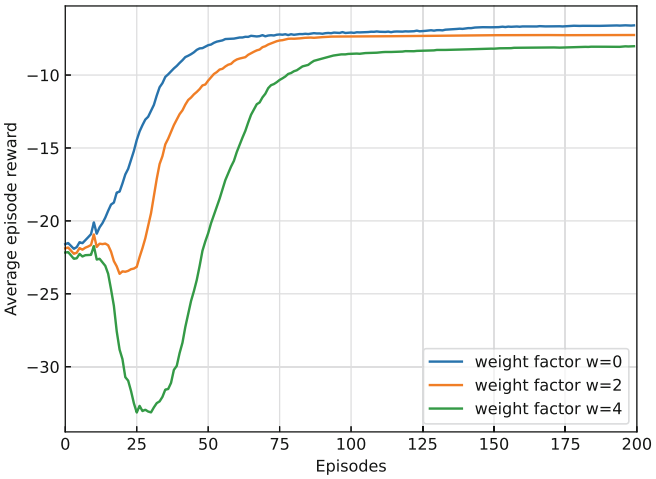

Fig. 2. The average reward with training process.

For the offline training of the DRL algorithm, the algorithm was trained for 200 episodes. Besides, each episode consisted of 1024 time slots, which were sampled from the traffic demand during a day. Figure 2 illustrates the change of the reward function during the training process of the algorithm with weight factor w set to $0, 2, 4$. Different w values adjust the weight of the power consumption and USC. For example, when w is set to 0, representing the case of minimizing USC. After 100 episodes of training, the algorithm has essentially converged. In the ideal scenario, the maximum value of the reward function is 0 with $w = 0$. However,

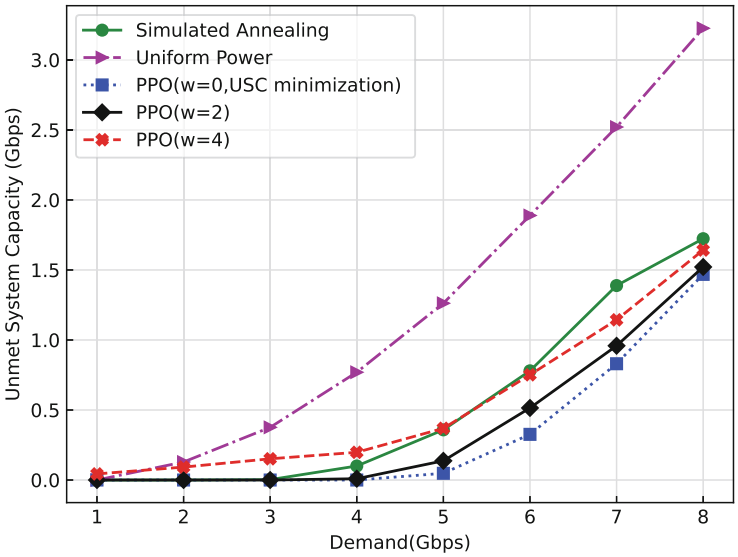

Fig. 3. USC versus traffic demand.

due to the existence of an upper limit on the maximum system capacity, not all traffic demand can be met. Therefore, when the algorithm converges, the reward is only a number close to 0.

After training the DRL agent, its performance was tested using dynamical traffic demands. Figure 3 and Fig. 4 show the USC and total power consumption of the proposed algorithm ($w = 0, 2, 4$), uniform power and SA algorithms at various total traffic demand level. The results from Fig. 3 indicate that when the total traffic demand is low (1, 2, 3 Gbps), the performance of SA and PPO algorithms ($w = 0, 2, 4$) is comparable, with USC tending towards 0, indicating that all traffic demand can be met. As the traffic demand increases, the PPO algorithm ($w = 0$, USC minimization) achieves the lowest USC. In comparison to $w = 0$, the power consumption decreases by approximately 10% with $w = 2$, while the increase in USC is minimal. This suggests that the PPO algorithm ($w = 2$) effectively balances USC and power consumption. The decrease in power consumption is due to considering the power consumption in the optimization objective. Meanwhile, the slight increase in USC can be explained by the Eq. (2), indicating that increasing the power beyond a threshold does not result in a significant improvement in SINR. Overall, the PPO algorithm with w of 0 and 2 outperforms the SA algorithm in both USC and power consumption. As for the uniform power, although it uses all the power, the power allocation cannot adapt to different channel conditions and traffic demand between beams, resulting in the higher USC.

Figure 5 illustrates the offered capacity per beam for different schemes under the total traffic demand of 6 Gbps. The uniform power scheme is unable to meet

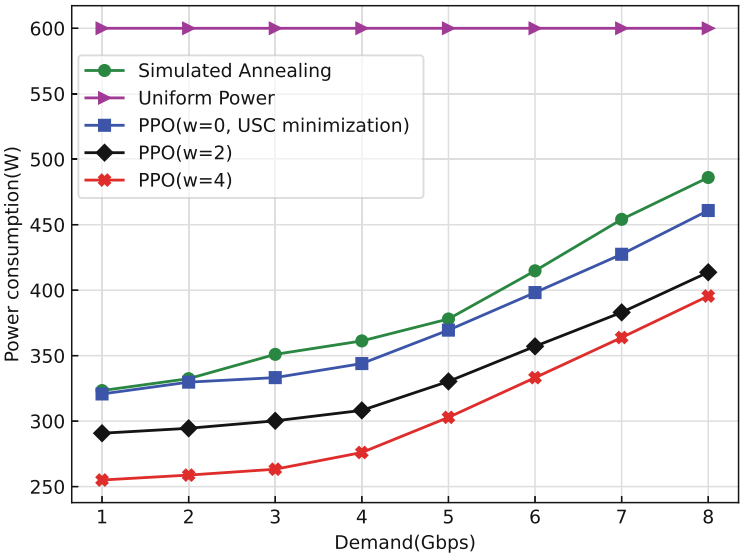

Fig. 4. power consumption versus traffic demand.

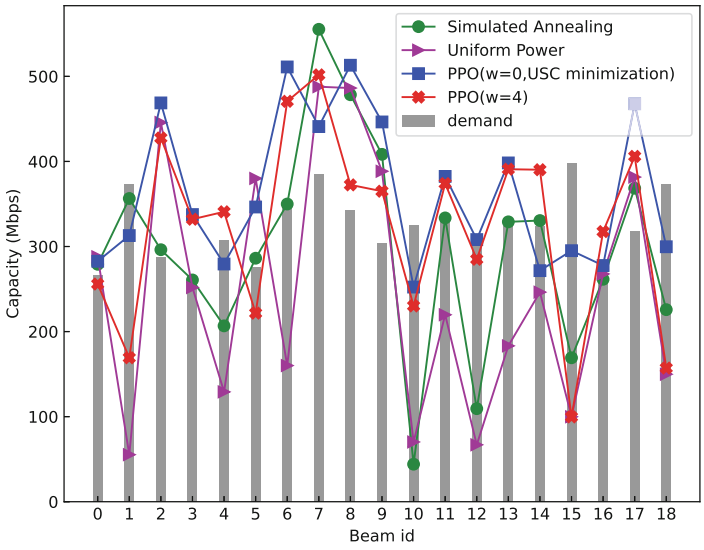

Fig. 5. The capacity provided by different schemes for each beam.

the traffic demand of each beam, while the PPO and SA algorithm can meet the demand of each beam as much as possible. Specifically, the PPO algorithm ($w = 0$) performs the best, while the SA and PPO ($w = 4$) algorithms show similar performance.

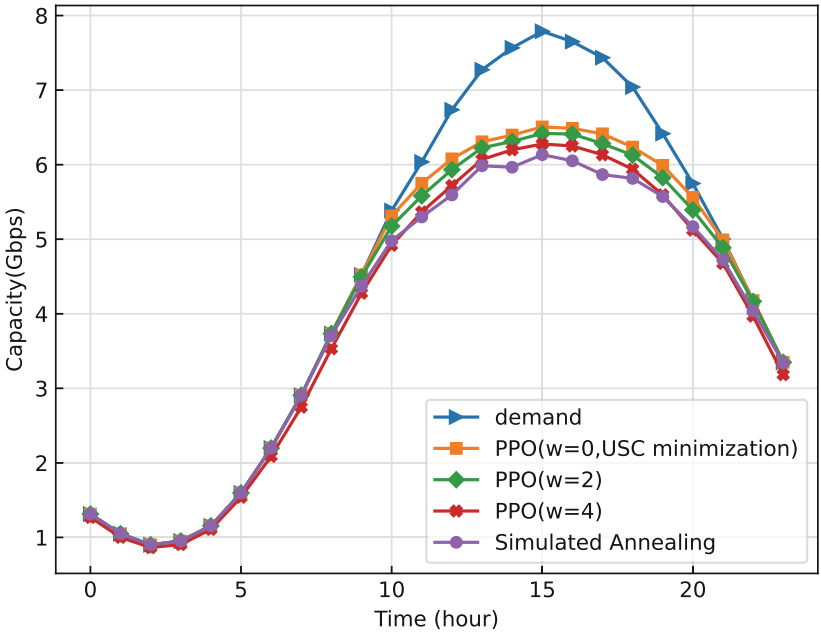

Fig. 6. System demand and provided capacity during a day.

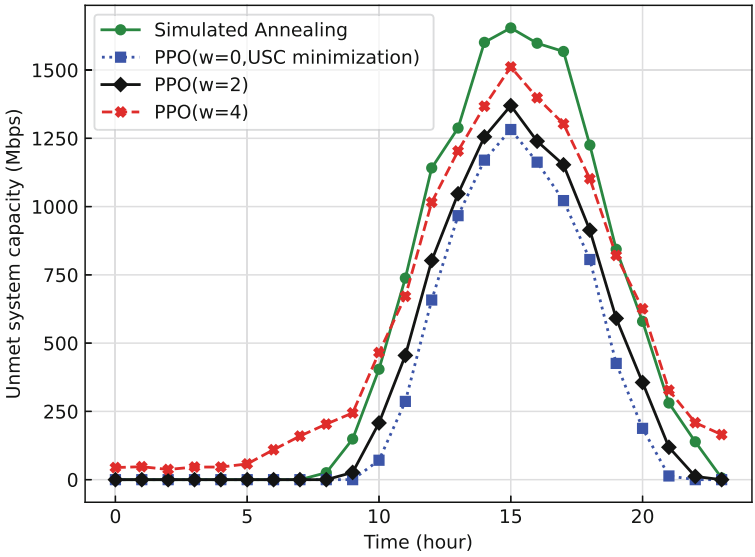

Fig. 7. Unmet system demand during a day.

To evaluate the performance of the proposed algorithm in response to time-varying traffic demands, Fig. 6 and Fig. 7 show the satisfied demand and USC during a day, respectively. In Fig. 6, we observe that the proposed algorithm can

dynamically adjust the capacity to meet changing demand. However, when the demand exceeds 5Gbps, some demand cannot be fully met because there is an upper limit for the capacity of system. From the Fig. 7, it can be seen that the USC of system is smaller when using the PPO algorithm with $w = 0$ and $w = 2$.

6 Conclusion

In this paper, we studied dynamic resource allocation in the multibeam LEO satellite system. In order to meet the dynamic traffic demand and reduce power consumption, we formulated an optimization objective aimed at minimizing both unmet traffic demand and power consumption. Subsequently, we applied PPO based algorithm to optimize resource allocation, which can learn dynamic characteristics of traffic and developing effective strategies. In the simulations, we employed different weight factors to adjust the trade-off between power consumption and unmet traffic demand in our proposed algorithm, resulting in diverse resource allocation schemes. Comparing these schemes with other benchmarks, our algorithm effectively can balance power consumption and unmet traffic demand and outperform the benchmarks in both power consumption and unmet system capacity.

References

1. Al-Hraishawi, H., Chougrani, H., Kisseleff, S., Lagunas, E., Chatzinotas, S.: A survey on nongeostationary satellite systems: the communication perspective. IEEE Commun. Surv. Tutor. **25**(1), 101–132 (2023). https://doi.org/10.1109/COMST.2022.3197695
2. Guan, Y., Geng, F., Saleh, J.H.: Review of high throughput satellites: Market disruptions, affordability-throughput map, and the cost per bit/second decision tree. IEEE Aerosp. Electron. Syst. Mag. **34**(5), 64–80 (2019). https://doi.org/10.1109/MAES.2019.2916506
3. Kisseleff, S., Lagunas, E., Abdu, T.S., Chatzinotas, S., Ottersten, B.: Radio resource management techniques for multibeam satellite systems. IEEE Commun. Lett. **25**(8), 2448–2452 (2021). https://doi.org/10.1109/LCOMM.2020.3033357
4. Kodheli, O., Lagunas, E., Maturo, N., Sharma, S.K., Shankar, B., Montoya, J.F.M., Duncan, J.C.M., Spano, D., Chatzinotas, S., Kisseleff, S., et al.: Satellite communications in the new space era: a survey and future challenges. IEEE Commun. Surv. Tutor. **23**(1), 70–109 (2020)
5. Hu, X., Liao, X., Liu, Z., Liu, S., Ding, X., Helaoui, M., Wang, W., Ghannouchi, F.M.: Multi-agent deep reinforcement learning-based flexible satellite payload for mobile terminals. IEEE Trans. Veh. Technol. **69**(9), 9849–9865 (2020). https://doi.org/10.1109/TVT.2020.3002983
6. Destounis, A., Panagopoulos, A.D.: Dynamic power allocation for broadband multi-beam satellite communication networks. IEEE Commun. Lett. **15**(4), 380–382 (2011). https://doi.org/10.1109/LCOMM.2011.020111.102201
7. Lei, J., Vazquez-Castro, M.A.: Joint power and carrier allocation for the multibeam satellite downlink with individual sinr constraints. In: 2010 IEEE International Conference on Communications. pp. 1–5. IEEE (2010)

8. Cocco, G., de Cola, T., Angelone, M., Katona, Z., Erl, S.: Radio resource management optimization of flexible satellite payloads for dvb-s2 systems. IEEE Trans. Broadcast. **64**(2), 266–280 (2018). https://doi.org/10.1109/TBC.2017.2755263
9. Zhao, D., Qin, H., Xin, N., Song, B.: Flexible resource management in high-throughput satellite communication systems: a two-stage machine learning framework. IEEE Trans. Commun. **71**(5), 2724–2739 (2023). https://doi.org/10.1109/TCOMM.2023.3255239
10. Gerard Maral, Michel Bousquet, Z.S.: Satellite Communications Systems: Systems, Techniques and Technology. Hoboken, NJ, USA: Wiley (2009)
11. Aravanis, A.I., Arapoglou, P.D., Danoy, G., Cottis, P.G., Ottersten, B.: Power allocation in multibeam satellite systems: a two-stage multi-objective optimization. IEEE Trans. Wirel. Commun. **14**(6), 3171–3182 (2015)
12. Efrem, C.N., Panagopoulos, A.D.: Dynamic energy-efficient power allocation in multibeam satellite systems. IEEE Wirel. Commun. Lett. **9**(2), 228–231 (2019)
13. Abdu, T.S., Kisseleff, S., Lagunas, E., Chatzinotas, S.: Flexible resource optimization for geo multibeam satellite communication system. IEEE Trans. Wireless Commun. **20**(12), 7888–7902 (2021)
14. Luis, J.J.G., Pachler, N., Guerster, M., del Portillo, I., Crawley, E., Cameron, B.: Artificial intelligence algorithms for power allocation in high throughput satellites: a comparison. In: 2020 IEEE aerospace conference. pp. 1–15. IEEE (2020)
15. Ma, S., Hu, X., Liao, X., Wang, W.: Deep reinforcement learning for dynamic bandwidth allocation in multi-beam satellite systems. In: 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). pp. 955–959 (2021). https://doi.org/10.1109/ICCCS52626.2021.9449160
16. Hu, X., Wang, Y., Liu, Z., Du, X., Wang, W., Ghannouchi, F.M.: Dynamic power allocation in high throughput satellite communications: a two-stage advanced heuristic learning approach. IEEE Trans. Veh. Technol. **72**(3), 3502–3516 (2023). https://doi.org/10.1109/TVT.2022.3218565
17. 3GPP: Solutions for nr to support non-terrestrial networks (ntn) v1.0.0 (release 16). Tech. Rep. TR 38.821, 3rd Generation Partnership Project (3GPP) (December 2019)
18. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
19. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015)
20. del Portillo, I., Cameron, B.G., Crawley, E.F.: A technical comparison of three low earth orbit satellite constellation systems to provide global broadband. Acta Astronaut. **159**, 123–135 (2019)
21. Andrew Fager: Analysis and modeling of internet usage (2017). https://www.kaggle.com/code/andrewfager/analysis-and-modeling-of-internet-usage/notebook