Multi-Agent DDPG-Based Joint Beam Hopping and Resource Allocation for Cognitive GEO-LEO Satellite Networks

Quynh Tu Ngo, *Senior Member, IEEE,* Ying He, *Senior Member, IEEE,*
Beeshanga Jayawickrama, *Senior Member, IEEE,* Eryk Dutkiewicz, *Senior Member, IEEE*

***Abstract*—This letter presents a multi-agent deep reinforcement learning-based framework for joint beam hopping and resource allocation in the secondary LEO system within a cognitive GEO-LEO satellite network. Beam hopping is adopted to flexibly steer limited resources toward areas with high and dynamic traffic demand, improving spectral efficiency and load balancing. A centralized critic multi-agent deep deterministic policy gradient approach is proposed to optimize beam hopping patterns, bandwidth and power allocation, aiming to enhance system throughput and reduce service delay imbalance across LEO cells. Simulation results show that the proposed method outperforms benchmark approaches, achieving approximately 45% higher throughput and around 37% lower service delay under varying traffic and power conditions.**

***Index Terms*—Cognitive GEO-LEO satellite networks, beam hopping, resource allocation, multi-agent DRL, MADDPG.**

I. Introduction

Satellite communication has become a key enabler for global connectivity in the era of 5G and the upcoming 6G networks, particularly in extending coverage to remote and underserved regions. Low Earth orbit (LEO) constellations have gained significant momentum due to their low latency, high capacity, and global coverage capabilities. However, the rapid growth of LEO deployments introduces new challenges in managing limited spectrum resources efficiently. A promising solution to addressing spectrum scarcity is the cognitive GEO-LEO satellite architecture, in which LEO systems opportunistically utilize underused spectrum allocated to geostationary (GEO) satellites [1]. This spectrum-sharing strategy offers practical benefits but also introduces challenges due to the dynamic nature of spectrum availability and the non-uniform, time-varying traffic demands across LEO cells. In this context, beam hopping (BH) emerges as a compelling technique, allowing LEO satellites to dynamically allocate resources by steering beams toward high-demand areas. This not only maximizes spectral efficiency but also ensures more effective service delivery under the constraints of limited and opportunistic GEO spectrum access.

Recent advances in machine learning, particularly in deep reinforcement learning (DRL), have shown significant potential in addressing complex resource management challenges in satellite networks. For GEO satellite systems, a multi-agent DRL framework based on double deep Q-learning (DDQL) is proposed for BH pattern design and bandwidth allocation [2], while a DRL-powered genetic algorithm is developed to optimize BH patterns [3]. In the context of LEO satellite systems, a multi-agent value decomposition network with dueling DDQL framework is proposed enabling real-time BH decision through centralized training and distributed deployment of DRL agents [4]. Additionally, a multi-agent proximal policy optimization algorithm is introduced for joint beamforming and dynamic BH to support hybrid wide-spot beam coverage in LEO networks [5]. Regardless, these studies primarily focused on single-tier satellite systems and do not fully address the spectrum scarcity and interference challenges inherent in multi-tier architectures such as cognitive GEO-LEO networks. For BH in GEO-LEO satellite system, [6] proposed a distance-based BH strategy to minimize inter-beam interference between LEO cells. However, this approach lacks comprehensive resource management in both spatial and spectral domains, limiting its adaptability to the traffic demand of LEO networks.

In this letter, we propose a multi-agent deep deterministic policy gradient (MADDPG)-based framework for joint BH pattern design and resource allocation, including bandwidth and power, for the secondary LEO system in cognitive GEO-LEO satellite networks. In the proposed framework, clusters of LEO cells are modeled as agents that learn to make BH and resource allocation decisions in a distributed yet coordinated manner. Leveraging centralized training with decentralized execution, the approach efficiently handles both continuous and discrete actions, enabling dynamic adaptation to fluctuating traffic demands and power budgets. Simulation results demonstrate that our method outperforms benchmark approaches, achieving approximately 45% higher throughput and reducing service delay by around 37%.

II. System Model

Consider a cognitive GEO-LEO satellite network, as illustrated in Fig. 1. In this system, the GEO satellite system operates as the primary network, while the LEO satellite system functions as the secondary network. The GEO satellite serves its users through multiple narrow spot beams and employs an ι-frequency reuse scheme, where ι is the frequency reuse factor and $\iota \in \mathbb{N} | \iota \geq 1$. Let B_G denote the total spectrum

The authors are with the School of Electrical and Data Engineering, University of Technology Sydney, Sydney, Australia (Email: {quynhtu.ngo, ying.he, beeshanga.jayawickrama, eryk.dutkiewicz}@uts.edu.au).

This work has been supported by the SmartSat CRC, whose activities are funded by the Australian Government's CRC Program.

of GEO satellite. Under the ι-frequency reuse scheme, the spectrum is divided into ι sub-bands $\{\beta_1, \beta_2, \cdots, \beta_\iota\}$.

The LEO satellite is equipped with multi-beam steerable phased array antennas and maintains a radio environment map (REM) that records the activity status of GEO beams within its footprint [7]. The REM provides real-time information on GEO beam activity, enabling the LEO satellite to identify available spectrum in each time slot [7]. Leveraging this information, the LEO satellite utilizes BH techniques to efficiently serve its users.

Both GEO and LEO users are equipped with single antennas. The coverage area of a GEO spot beam, referred to as a GEO cell, is assumed to be circular with a radius of R_G. In contrast, the coverage area of a LEO spot beam, referred to as a LEO cell, is hexagonal with a radius of R_L, where $R_L < R_G$. The overall footprint of a LEO satellite is circular, with a radius of F_L. The number of LEO cells within this footprint is given by: $N = 2\pi F_L^2/3\sqrt{3}R_L^2$. Let K represent the maximum number of LEO beams serving the footprint area, where $K < N$. At a given time slot t, the LEO satellite receives user requests through a signaling beam. The service beam can then hop between LEO cells based on traffic demand and dynamically allocate beam resources.

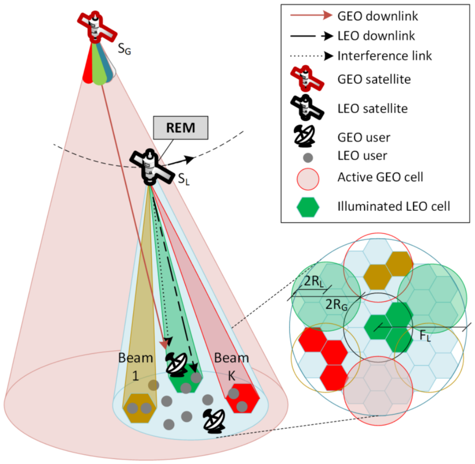

Fig. 1. An illustration of cognitive GEO-LEO beam hopping satellite networks with GEO satellite using 4-color frequency reuse scheme.

1) Interference Model: There are two types of interference considered in this system: interference from LEO beams to GEO users and LEO co-channel inter-beam interference.

The interference power caused by the service beam of LEO cell i to a GEO user U_G is given by:

$$I_{i,U_G} = \epsilon_i P_{L,i} G_{L,i} G_{U_G} L_{LU_G}^{-1}, \tag{1}$$

where $\epsilon_i \in \{0,1\}$ is a binary variable indicating whether LEO cell i is inside an overlapping GEO beam region, with $\epsilon_i = 1$ denoting the LEO cell i is in the region where GEO beams overlap. $P_{L,i}$ is the LEO beam i transmitting power. $G_{L,i}$ and G_{U_G} are the LEO satellite beam i transmitting antenna radiation pattern and GEO user receiving antenna radiation pattern, respectively. Both G_L and G_{U_G} are modeled as

$$G(\phi) = \eta \frac{4\pi A}{(c/f_c)^2} \left(\frac{J_1(\mu)}{2\mu} + \frac{36 J_3(\mu)}{\mu^3} \right)^2, \tag{2}$$

where ϕ is the off-boresight angle, η is the antenna efficiency, A is the antenna area, c is the speed of light, f_c is the carrier frequency, $J_1(\cdot)$ and $J_3(\cdot)$ are the first and third order Bessel functions, and $\mu = 2.07123 \sin(\phi)/\sin(\phi_{3dB})$ with ϕ_{3dB} is the off-boresight angle corresponding to the 3 dB beamwidth. L_{LU_G} is the free space propagation loss between LEO satellite and U_G, which is modeled as

$$L_{LU_G} = \frac{8\pi R_E}{c/f_c} \arcsin\left(\sqrt{\sin^2\left(\frac{l_{U_G}^a - l_L^a}{2}\right) + \cos(l_L^a)\cos(l_{U_G}^a)\sin^2\left(\frac{l_{U_G}^o - l_L^o}{2}\right)} \right), \tag{3}$$

where R_E is the Earth radius, $\{l_L^a, l_L^o\}$ and $\{l_{U_G}^a, l_{U_G}^o\}$ are the latitude and longitude of LEO satellite and U_G, respectively.

The LEO co-channel inter-beam interference experienced by a LEO user U_L within cell i, namely $I_{i,L}$, originates from all the other LEO beams serving LEO cells within the same GEO cell as cell i. These interfering LEO beams are illuminated in the same time slot and utilize the same sub-band as beam i. This interference power can be expressed as follows:

$$I_{i,L} = \sum_{j, j\neq i}^{M} \varepsilon_{ij} P_{L,j} G_{L,j} G_{U_L} L_{LU_L}^{-1}, \tag{4}$$

where $M = 2\pi R_G^2/3\sqrt{3}R_L^2$ is the number of LEO cells inside one GEO cell. $\varepsilon_{ij} \in \{0,1\}$ is a binary variable indicating the overlap band between beam i and beam j, with $\varepsilon_{ij} = 1$ denoting they use the same sub-band. $P_{L,j}$ is the LEO beam j transmitting power. $G_{L,j}$ and G_{U_L} are LEO satellite beam j transmitting antenna radiation pattern and LEO user receiving antenna radiation pattern, respectively. Both $G_{L,j}$ and G_{U_L} are modeled as in (2). L_{LU_L} denotes the path loss between LEO satellite and U_L, and is modeled similarly to (3).

III. Problem Formulation

As the secondary network, LEO satellite must optimize resource allocation: BH pattern, frequency band selection, and transmit power, based on GEO network activity from the REM. The goal is to maximize LEO throughput while minimizing service delay, ensuring fair and timely access for all traffic-demanding cells.

The LEO throughput is defined as the sum of the throughput across all LEO cells:

$$\Gamma_L = \sum_{i}^{N} \min\{S_i, D_i\}, \tag{5}$$

where S_i and D_i represent the traffic supply and demand for LEO cell i, respectively. The traffic supply for LEO cell i can be expressed as

$$S_i = \kappa_i B_i \log\left(1 + \frac{P_{L,i} G_{L,i} G_{U_L} |h_{LU_L}|^2 L_{LU_L}^{-1}}{I_{i,L} + \sigma_{U_L}^2}\right), \tag{6}$$

where $\kappa_i \in \{0,1\}$ is a binary variable indicating cell i illumination status, with $\kappa_i = 1$ denoting that cell i is illuminated. B_i represents the bandwidth for cell i. h_{LU_G} is the channel coefficient from LEO satellite to U_L following Shadowed Rician distribution, and $\sigma_{U_L}^2$ denotes the noise power at the LEO user.

Service delay refers to the amount of time a cell's requested data spends waiting before being served within one BH cycle. The traffic demand of cell i at slot t can be expressed as $D_{i,t} = \sum_{l=1}^{T_{BH}} \chi_{i,t}^{l}$, where $\chi_{i,t}^{l}$ is the number of packets that have been waiting in queue of cell i for l slots, and T_{BH} is the total number of slots in one BH cycle. Service delay for cell i is $\tau_i = \sum_{l=1}^{T_{BH}} l\chi_{i,t}^{l}/D_{i,t}$. The service delay imbalance metric between LEO cells can be expressed as

$$\Delta\tau = \max_{i}\left(\tau_i\right) - \min_{i}\left(\tau_i\right). \tag{7}$$

The optimization problem is formulated as:

$$\max_{\kappa_i, B_i, P_{L,i}} \quad -\alpha\Delta\tau + \left(1-\alpha\right)\Gamma_L, \tag{8a}$$

$$\text{s.t. } \sum_{i=1}^{N} \kappa_i \leq K, \tag{8b}$$

$$\sum_{i=1}^{N} P_{L,i} \leq P_L, \tag{8c}$$

$$B_i \subseteq \{\beta_1, \beta_2, \cdots, \beta_\iota\} \setminus \left\{\beta_j | A_{G,ij} = 1\right\}, \forall i, \tag{8d}$$

$$A_{L,ij} + A_{G,ij} \leq 1, \forall i, j, \tag{8e}$$

$$\sum_{i \in \mathcal{O}_G} I_{i,U_G} \leq I_{th}, \tag{8f}$$

where α is the weight balancing objectives, P_L is the total LEO transmit power, and I_{th} is the GEO user interference threshold. $A_{G,ij}$ and $A_{L,ij}$ are binary variables indicating whether GEO cell covering LEO cell i and LEO cell i, respectively, use sub-band β_j (1 if in use). The optimization jointly designs beam illumination, sub-band selection, and power allocation to maximize LEO throughput and reduce service delay imbalance across LEO cells. Constraint (8b) limits the number of active beams to K, and (8c) keeps total transmit power within budget. Constraints (8d) and (8e) avoid sub-band reuse in LEO cells overlapping with GEO cells, preventing interference. Constraint (8f) keeps LEO interference in overlapping GEO beam regions below I_{th}, where $\mathcal{O}_G = \{i | \bigcap_{g \in \mathcal{G}_i} C_g \neq \emptyset,\ |\mathcal{G}_i| \geq 2\}$ identifies LEO cells in these regions, $\mathcal{G}_i$ is the set of GEO cells covering LEO cell i, and C_g is the coverage area of GEO cell g.

The optimization problem in (8) is a mixed-integer nonlinear program (MINLP), NP-hard due to binary variables and non-convex interference coupling. To solve it, a MADDPG-based approach is used, jointly managing continuous (power) and discrete (beam/sub-band) decisions, enabling scalable and distributed control across LEO cells in dynamic conditions.

IV. Centralized Critic MADDPG for Joint Beam Hopping and Resource Allocation

MADDPG [8] is a DRL framework designed for decentralized multi-agent system, where each agent learns its own policy while interacting with others. To solve the optimization problem in (8), we cluster LEO cells based on their geographic locations within GEO cells, treating each cluster as an agent that jointly determines sub-band allocation, power distribution, and illumination status for its LEO cells. The objective is to maximize overall LEO system throughput while minimizing service delay imbalance. A key challenge stems from GEO user interference, which results from the aggregate effect of multiple LEO beams, making it hard for agents to assess their individual interference contributions. In standard MADDPG, agents rely only on local observations and individual rewards, making coordinated, interference-aware decisions difficult. To overcome this, we propose the Centralized Critic MADDPG (CC-MADDPG), which employs a single centralized critic that evaluates the joint interference impact on GEO users. This allows the system to penalize LEO cells when their combined actions exceed interference thresholds, encouraging more stable and interference-aware learning.

A. Problem Reformulation

The problem in (8) is reformulated as a multi-agent reinforcement learning problem, where each cluster of $N_b = 2\pi R_G^2/(3\sqrt{3}R_L^2)$ LEO beams is managed by a single agent. Consequently, the total number of agents is $N_a = N/N_b$. The problem is formulated follow Markov decision process, with its key components defined as follows:

State space: At the beginning of each time slot, every agent k obtains its observation (local state) from the environment, including the demand traffic of cells in cluster k – $D_k(t) = \{D_i(t)\}$, $i = 1, 2, \cdots, N_b$ and the activity status – $A_{G,kj}(t) = \{A_{G,ij}(t)\}$, $j = 1, 2, \cdots, \iota$ of the GEO beam covering cells in cluster k. These observations, $o_k(t) = \{D_k(t), A_{G,kj}(t)\}$, form the global state $s(t) \in \mathcal{S}$ at time slot t, where $\mathcal{S}$ represents the state space. The global state is defined as $s(t) = \left\{\{D_k(t), A_{G,kj}(t)\}\right\}$, $k = 1, 2, \cdots, N_a$.

Action space: Based on its observation, each agent determines its action, which includes whether to illuminate its beam, the sub-band to utilize, and the transmission power for its beam. The joint action is $a(t) = \{\{\{\kappa_i\}, \beta_j, \{P_{L,i}\}\}\}$, $a(t) \in \mathcal{A}$, where $\mathcal{A}$ denotes the action space.

Reward function: Upon taking action, each agent receives an immediate reward $r_k(t)$ defined as

$$r_k(t) = \underbrace{-\alpha\Delta\tau + \left(1-\alpha\right)\Gamma_L}_{base\ reward} \underbrace{- w_I \cdot \max\left(0, \sum_{i \in \mathcal{O}_G} I_{i,U_G} - I_{th}\right)}_{penalty\ term} \tag{9}$$

where w_I is the penalty weight. This imposes a penalty when the total interference on GEO users exceeds the threshold. The goal of each agent is to accumulate maximum reward overtime.

B. Proposed Centralized Critic MADDPG

The proposed CC-MADDPG framework, illustrated in Fig. 2, follows an actor-critic architecture consisting of N_a independent actor networks and a single centralized critic network. Each agent k has an actor network Π_{θ_k}, which maps its local observation o_k to an action a_k. The centralized critic Q_ϕ computes a single Q-value for the entire system by evaluating the joint action $a(t)$ based on the global state $s(t)$ and the actions of all agents. The training phase follows a centralized learning approach, where the critic has access to the full system state, while inference remains fully decentralized, with each agent selecting its action independently based only on its local observation.

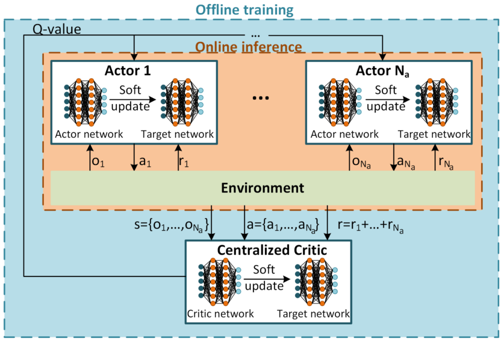

Fig. 2. Proposed centralized critic - MADDPG framework.

The environment evolves over time according to the transition probability:

$$s(t+1) \leftarrow \mathbb{P}\left(s(t+1)|s(t),a(t)\right), \tag{10}$$

where $\mathbb{P}(\cdot)$ denotes the probability distribution governing state transitions.

The centralized critic learns the action-value function for the joint policy $\Pi = (\Pi_1, \Pi_2, \cdots, \Pi_N)$:

$$Q^{\Pi}_{\phi}\left(s(t),a(t)\right) = \mathbb{E}\left[r(t) + \gamma \mathbb{E}_{a'(t+1)\sim\Pi} Q^{\Pi}_{\phi'}\left(s(t+1),a'(t+1)\right)\right], \tag{11}$$

where Q_ϕ is the critic network parameterized by ϕ, $r(t) = \sum_{k=1}^{N_a} r_k(t)$ is the global reward, γ is the discount factor, $a'(t+1) \sim \Pi$ are the next actions sampled from the policy, and ϕ' represents the parameters of the target critic network, which is periodically updated using a soft update strategy from ϕ. The critic network is trained by minimizing loss function:

$$L(\phi) = \mathbb{E}\left[\left(Q^{\Pi}_{\phi}\left(s(t),a(t)\right) - y(t)\right)^2\right], \tag{12a}$$

$$y(t) = r(t) + \gamma Q^{\Pi}_{\phi'}\left(s(t+1),a'(t+1)\right). \tag{12b}$$

The critic network parameters are updated using stochastic gradient descent:

$$\phi \leftarrow \phi - \lambda_Q \nabla_\phi L(\phi), \tag{13}$$

where λ_Q is the learning rate and $\nabla_\phi L(\phi)$ represents the gradient of the loss function with respect to ϕ.

Each agent k updates its actor policy Π_k using deterministic policy gradient theorem, maximizing the expected Q-value:

$$\nabla_{\theta_k} J(\Pi_k) = \mathbb{E}\left[\nabla_{\theta_k \Pi_k}\left(a_k|s_k\right)\nabla_{a_k} Q_\phi\left(s(t),a(t)\right)\right], \tag{14}$$

where $J(\Pi_k) = \mathbb{E}[Q_\phi\left(s(t),a(t)\right)]$ is the objective function and $\Pi_k\left(a_k|s_k\right)$ represents local policy network of agent k.

CC-MADDPG operates under a centralized training and decentralized inference paradigm. During training, performed offline on the ground, the centralized critic Q_ϕ learns the joint action-value function using the global state and the actions of all agents, as outlined in Algorithm 1.

Algorithm 1: CC-MADDPG: centralized training

```
Initialization: actor networks $\Pi_{\theta_k}$, critic network $Q_\phi$,
  target networks $\Pi'_{\theta_k}$ and $Q'_\phi$ as copies of the actors
  and critic, experience replay buffer;
for each episode do
    Initialize state $s(0) = \{o_1, \cdots, o_{N_a}\}$
    for each time step $t$ do
        Each agent $k$ selects action $a_k(t)$ using
          exploration, executes action $a_k(t)$, observes
          next state $o_k(t+1)$, receives reward $r_k(t)$
        Store joint $< s(t), a(t), r(t), s(t+1) >$ in
          replay buffer
        Sample mini-batch from replay buffer
        Update centralized critic $Q$ using (12), (13)
        Update actors $\Pi_k$ using (14)
        Update target networks using soft update
        Check stopping criteria (reward convergence)
        If converged, stop training
    end
end
```

C. Computational Complexity

The computational complexity of MADDPG [8] with N_a agents is $\mathcal{O}(N_a N_{ep} T \mathfrak{B}(|\mathcal{S}|\times|\mathcal{A}|+2(|\mathcal{S}|N_1+\sum_{l=1}^{L_A} N_l N_{l+1})+2(|\mathcal{A}|N_1+\sum_{l=1}^{L_C} N_l N_{l+1})))$ [9], where N_ep is the number of training episodes, T is the time steps per episode, $\mathfrak{B}$ is the mini-batch size, L_A and L_C are the number of hidden layers in the actor and critic networks, and N_l is the number of neurons in the l-th hidden layer. In contrast, the proposed CC-MADDPG employs only a single critic network, reducing the complexity to $\mathcal{O}(N_a N_{ep} T \mathfrak{B}(|\mathcal{S}| \times |\mathcal{A}| + 2(|\mathcal{S}|N_1 + \sum_{l=1}^{L_A} N_l N_{l+1})) + 2N_{ep} T \mathfrak{B}(|\mathcal{A}|N_1 + \sum_{l=1}^{L_C} N_l N_{l+1}))$.

V. Numerical Results

A. Simulation Setup

The environment simulator models the GEO satellite as Inmarsat-4F2 and the LEO satellite as Iridium-NEXT 914. The region of interest is the LEO footprint, a circular area with a radius of 2,350 km, centered at $(35°S, 150°E)$. Within this footprint, there are 17 GEO cells, each with a radius of 555 km, overlapping by 5% at the edges. The GEO satellite employs a 4-color frequency reuse scheme with the active beam pattern being randomly selected follow Poisson distribution with a mean of 5 beams per BH cycle. One BH cycle lasts 100 ms. LEO cells have a radius of 150 km, with a total of 261 cells within the footprint. These cells are grouped into 17 clusters, each managed by a separate agent. Each cluster contains either 15 or 16 LEO cells. Traffic demand in each LEO cell ranges from 50 to 200 kbps and follows a Poisson process with an arrival rate of 0.727. Data packets are queued for transmission and are dropped if not transmitted within 100 ms. Key simulation parameters and hyper-parameters of the proposed CC-MADDPG are listed in Table I. All actor networks share a common architecture comprising two hidden layers with 64 and 32 neurons, respectively. The critic network consists of two hidden layers with 512 and 256 neurons. All networks use Tanh activation function and are updated using the Adam optimizer every 100 time steps. At each time step, throughput, delay imbalance, and interference are normalized by their running averages before forming the reward.

TABLE I
KEY PARAMETERS USED IN SIMULATION

System Parameter	**Value**
Carrier frequency (Ka-band)	19 GHz
LEO satellite altitude	780 km
LEO antenna diameter	0.2 m
LEO antenna gain	38.5 dBi
LEO half power beamwidth	2.98°
GEO total bandwidth	800 kHz
GEO/LEO user antenna diameter	0.3 m
GEO/LEO user antenna gain	42.1 / 39.7 dBi
Noise power	-120 dBm/Hz
Shadowing ($\{m, b, \Omega\}$)	$\{10.1, 0.126, 0.835\}$
GEO user interference threshold	-115 dBW/MHz
DRL Hyper-parameter	**Value**
Training episodes	2100
Replay buffer size	100000
Mini batch size	64
Learning rate	0.001
Decaying rate	0.0001

B. Simulation Results

The convergence behavior of CC-MADDPG is illustrated in Fig. 3, which shows the critic loss under different learning rates and decay rates. The algorithm demonstrates improved convergence performance with a learning rate of 0.001 and a decay rate of 0.0001. Fig. 4 shows the traffic demand and supply in LEO cells after a BH cycle. LEO cells with zero traffic supply are located within active GEO cells, where spectrum access is restricted. Since GEO cells have priority in spectrum usage, these overlapping LEO cells are unable to transmit.

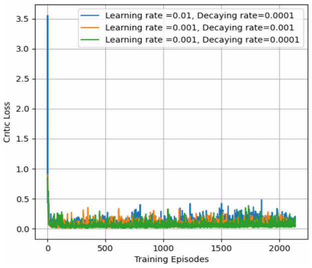

Fig. 3. Critic loss curve under different learning/decaying rates.

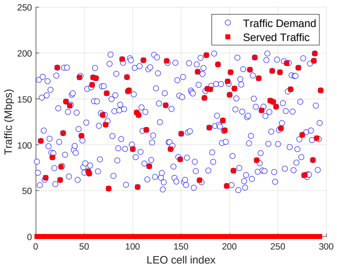

Fig. 4. LEO traffic demand versus supply after a BH cycle.

Figs. 5 and 6 show the performance comparison of the proposed CC-MADDPG against the following benchmarks:

- Greedy: choosing M beams that have largest traffic demand in each time slot for illumination [10].
- Beam hopping based on cell distance threshold [6].

Fig. 5 shows that the proposed CC-MADDPG algorithm consistently achieves the highest LEO system throughput under different transmit power budgets. With $\alpha = 0.5$, throughput is 61.76% higher than the Greedy algorithm [10], 45% higher than the other benchmark [6], 7.8% higher than using $\alpha = 0.7$, and 2.2% lower than using $\alpha = 0.3$. Fig. 6 shows that setting $\alpha = 0.7$, which places greater emphasis on balancing service delay, achieves the lowest average delay; the $\alpha = 0.5$ and 0.3 variants have 25% and 29.2% higher delay, while the benchmarks see delays 37% and 54.5% higher, respectively. Overall, simulation results demonstrate that the proposed method outperforms benchmark approaches on average, achieving approximately 45% higher throughput and reducing service delay by around 37%.

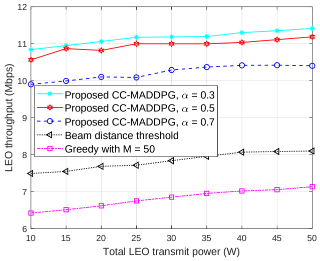

Fig. 5. LEO system throughput under varying transmit power budgets.

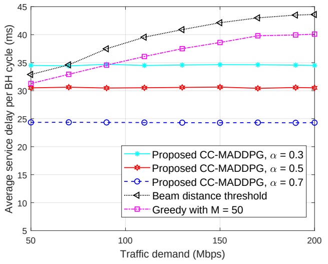

Fig. 6. Average service delay per BH cycle.

VI. CONCLUSION

This letter proposed a DRL-based joint BH and resource allocation framework for the secondary LEO system in a cognitive GEO-LEO satellite network. A centralized critic MADDPG approach was developed to optimize sub-band selection, BH pattern, and power allocation, aiming to maximize throughput and minimize service delay imbalance. By unifying beam activation and power control in a multi-agent DRL model, the framework enables coordinated optimization across clustered LEO cells. Simulation results show that the proposed method outperforms existing benchmarks in both throughput and delay under varying traffic and power conditions.

REFERENCES

[1] Q. T. Ngo, B. Jayawickrama, and et.al., "Optimizing spectrum sensing in cognitive GEO-LEO satellite networks: Overcoming challenges for effective spectrum utilization," *IEEE Veh. Technol. Mag.*, pp. 1–1, 2025.

[2] Z. Lin and et. al., "Dynamic beam pattern and bandwidth allocation based on multi-agent DRL for beam hopping satellite systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 3917–3930, 2022.

[3] X. Hu, L. Wang, Y. Wang, S. Xu, Z. Liu, and W. Wang, "Dynamic beam hopping for DVB-S2X GEO satellite: A DRL-powered GA approach," *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 808–812, 2022.

[4] M. Meng, B. Hu, S. Chen, and S. Kang, "Dynamic beam pattern based on cooperation multi-agent VDN-D3QN for LEO satellite communication system," *IEEE Trans. Green Commun. Netw.*, pp. 1–1, 2024.

[5] M. Meng, B. Hu, and et. al., "Joint beamforming and dynamic beam hopping based on MAPPO for LEO satellite communication system," *IEEE Wireless Commun. Lett.*, pp. 1–1, 2025.

[6] J. Tang, D. Bian, G. Li, J. Hu, and J. Cheng, "Resource allocation for LEO beam-hopping satellites in a spectrum sharing scenario," *IEEE Access*, vol. 9, pp. 56 468–56 478, 2021.

[7] Q. T. Ngo, B. Jayawickrama, Y. He, and E. Dutkiewicz, "A novel satellite-based REM construction in cognitive GEO-LEO satellite IoT networks," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 7532–7548, 2025.

[8] R. Lowe and et.al., "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[9] Q. T. Ngo and et.al., "A fast fuzzy DRL-based joint beam design and power allocation for multi-beam GEO-LEO coexisting satellite networks," *IEEE Trans. Wireless Commun.*, pp. 1–14, 2025.

[10] L. Chen, V. N. Ha, and et. al., "The next generation of beam hopping satellite systems: Dynamic beam illumination with selective precoding," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2666–2682, 2023.