# HIERARCHICAL DEEP REINFORCEMENT LEARNING FOR SPECTRUM RESOURCE OPTIMIZATION IN INTEGRATED TERRESTRIAL AND NON-TERRESTRIAL NETWORKS

Muhammad Ahmed Mohsin[1], Hassan Rizwan[2], Muhammad Umer[1], Sagnik Bhattacharya[1] , Ahsan Bilal[3] ,
John M. Cioffi[1]

[1]Stanford University, [2]University of California, [3]University of Oklahoma

muahmed@stanford.edu, hrizwan@email.com, mumer.bee20seecs@seecs.edu.pk, sagnikb@stanford.edu, ahsan.bilal-1@ou.edu, cioffi@stanford.edu

## INTRODUCTION

The relentless densification and heterogenization of wireless infrastructures—from dense terrestrial cellular grids and massive IoT deployments to burgeoning LEO satellite constellations—has intensified competition for the finite RF spectrum, especially as integrated terrestrial–non-terrestrial networks (TN–NTNs) link ground cells with thousands of satellites and high-altitude platforms. Projections foresee the active satellite count swelling from roughly 13 000 in 2024 to 33 000 by 2030, fueling a satellite-broadband market expansion from $9 billion in 2023 to $37 billion by 2034. Traditional spectrum-sharing schemes, whether centralized deep-RL or classical terrestrial-only methods, incur high signaling overhead, suffer from stale state information, and fail to account for the vertical heterogeneity and latency asymmetries inherent in TN–NTN coexistence. Flat multi-agent DRL approaches, in particular, become impractical as network layers and user density increase, due to prohibitive coordination costs and enlarged state–action spaces. Hierarchical reinforcement learning (HRL) offers a principled remedy by decomposing the global spectrum-allocation problem along the natural spatial and temporal tiers

## FUNDAMENTALS OF HRL

Deep reinforcement learning (DRL) extends Markov decision processes by training neural agents to interact with an environment over time—observing states, taking actions, receiving rewards, and aiming to maximize expected discounted rewards. Value-based methods (e.g., Deep Q-Networks) estimate action-values, while policy-based methods (e.g., REINFORCE, actor-critic) optimize the policy directly, with actor-critic combining both for greater stability. Key techniques like experience replay, target networks, and entropy regularization enhance performance in complex tasks. Hierarchical DRL further improves scalability by using a high-level controller to set sub-goals and lower-level controllers to execute them, making it ideal for complex systems like integrated terrestrial and non-terrestrial networks by reducing overhead and enabling adaptive, multi-scale control.

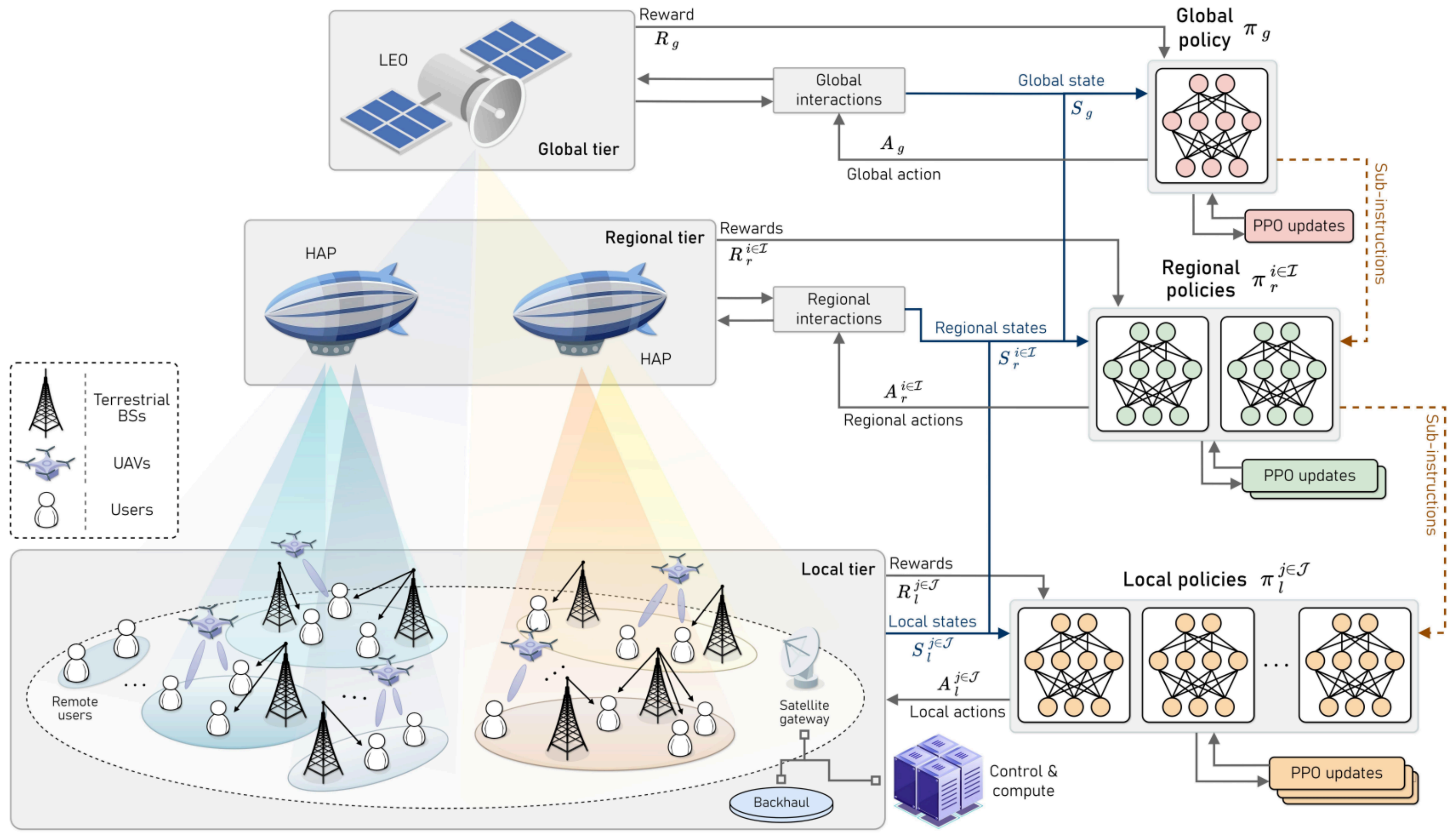


Fig. 2. System model and HDRL framework for spectrum sharing in integrated TN-NTNs.

## SYSTEM OVERVIEW

We consider an integrated terrestrial–non-terrestrial architecture in which a low-Earth-orbit (LEO) satellite, multiple high-altitude platforms (HAPs), unmanned-aerial-vehicle base stations (UAV-BSs), terrestrial base stations (TBSs), and ground users coexist. The LEO satellite employs a fixed multi-beam payload; each beam illuminates a distinct cell on the Earth's surface and delegates control of that cell to a dedicated HAP. Acting as regional hubs, HAPs relay data and control signalling between the satellite layer and the underlying terrestrial tier.

### GLOBAL TIER

The system's top layer is a low-Earth-orbit satellite carrying a fixed multi-beam payload that illuminates a finite number of ground cells. Each beam cell is served by a dedicated high-altitude platform (HAP), to which the satellite delegates both data traffic and control signaling.

$$B = \text{number of distinct satellite beams (ground cells)}$$

$$A_{\text{spec}} = \text{total available spectrum}$$

$$A_{\text{satellite}} = \{\, A_{\text{satellite},1}, \ldots, A_{\text{satellite},B} \}, \quad \sum_{b=1}^{B} A_{\text{satellite},b} = A_{\text{spec}}$$

### REGIONAL TIER

Each high-altitude platform partitions the spectral slice assigned by the satellite among its subordinate stations, which include both terrestrial base stations (TBSs) and UAV-based stations (UAV-BSs).

$$M_i = \text{number of subordinate stations under HAP } i$$

$$A_{\text{HAP}}^{(i)} = \{\, A_1^{(i)}, \ldots, A_{M_i}^{(i)} \}, \quad \sum_{m=1}^{M_i} A_m^{(i)} \;\le\; A_{\text{satellite},b(i)}$$

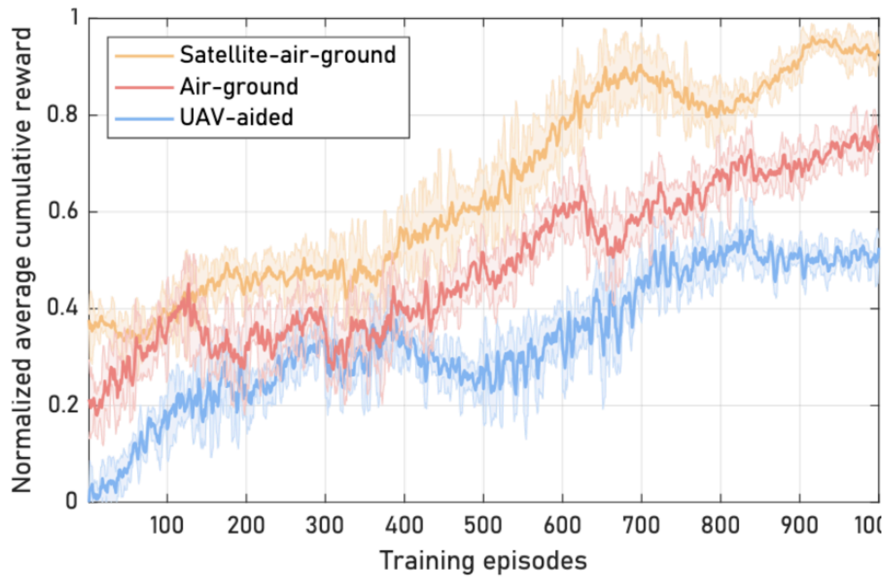
### LOCAL TIER

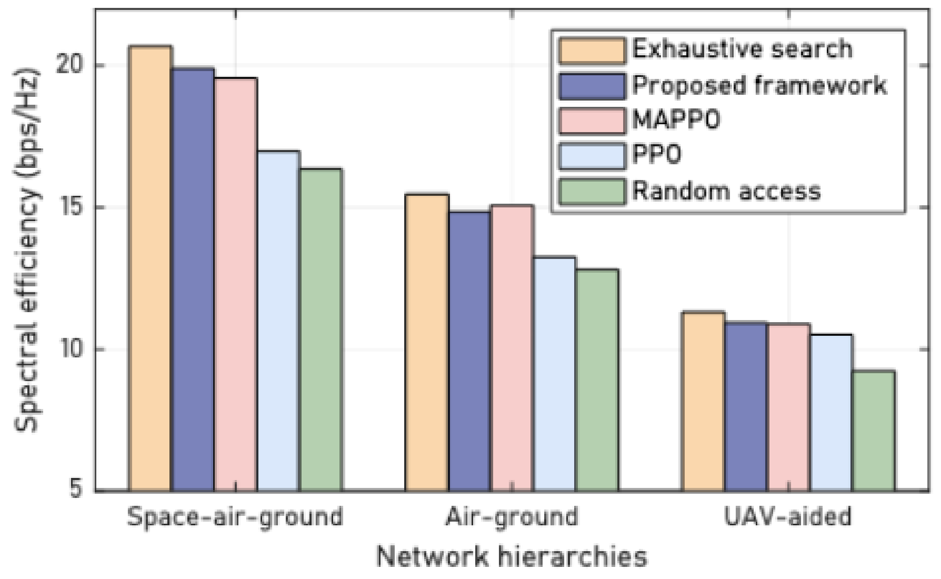
Within each HAP's footprint, individual stations assign specific sub-bands and power levels to user equipment, reacting to fast-fading and mobility.

$$a_{u,n}^{(i,m)} \in \{0,1\}, \quad \sum_{u} a_{u,n}^{(i,m)} \le 1 \quad \forall n$$

$$p_{u,n}^{(i,m)} \in [0,\ P_{\max}]$$

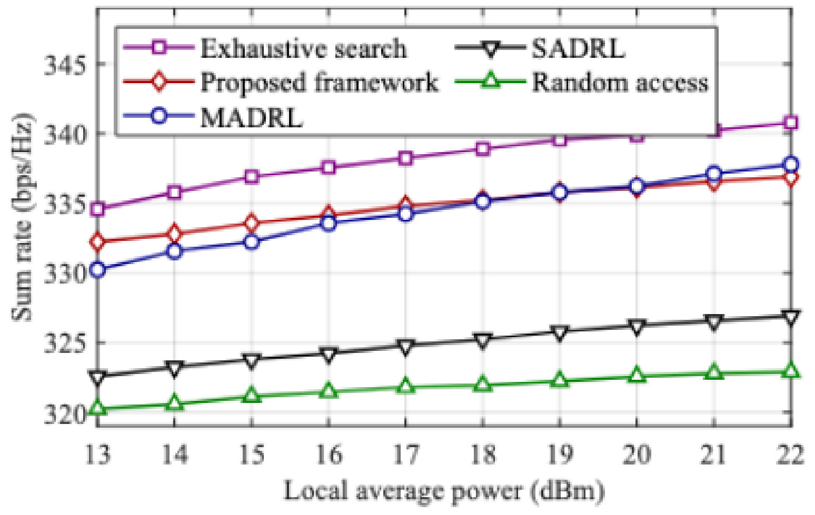
## RESULTS





Across all three hierarchies, normalized cumulative reward climbs sharply and reaches a stable plateau by about 700–800 episodes. Small oscillations persist due to changing channel conditions and evolving network states. Satellite–air–ground, air–ground, and UAV–aided setups all learn at similar speeds. Our HDRL framework consistently achieves spectral efficiency within 1.5 % of the exhaustive-search optimum across all configurations. MAPPO keeps pace in the simpler air–ground and UAV–aided cases but falls slightly behind in the full three-tier scenario.

Single-agent PPO and random access both underperform and show greater variability.



HDRL and MADRL deliver identical, steadily rising sum-rate curves as transmit power increases.

The exhaustive-search baseline maintains a constant ~5 bps/Hz advantage—about 1.5 % above HDRL—across the power range.Single-agent DRL and random access trail by at least 8 bps/Hz, highlighting the value of hierarchical coordination.

## CONCLUSION

The proposed HDRL framework decomposes spectrum sharing in TN–NTNs into global, regional, and local tiers, drastically cutting the effective action space compared to flat schemes. Simulations show HDRL converges up to 50× faster than exhaustive search, achieves 95 % of optimal spectral efficiency, outpaces MADRL by 3.75× in convergence, and boosts throughput by 5 %. Future work includes handling mobility-driven handovers, energy constraints, federated learning, and real-world TN–NTN testbed validation.

## REFERENCES