EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Reinforcement Learning-Based Robust Resource Scheduling for Dynamic LEO Satellite Networks

N.N. Song[1], Y. Z. Li[1,*] , Y. K. Zhao[1], J. Li[1]and W. Li[1]

[1] State Grid Shanxi Electric Power Company Limited, Lvliang Power Supply Branch, Lvliang, 033000,China

Abstract

INTRODUCTION: Low Earth Orbit (LEO) satellite communication extends Narrowband Internet of Things (NB-IoT) coverage for global 6G IoT services. However, long propagation delays and high mobility cause outdated channel state information (CSI), degrading system performance.
OBJECTIVES: This study aims to design a robust resource scheduling strategy that mitigates CSI outdatedness, improves resource utilization, and supports large-scale IoT connectivity in dynamic LEO satellite environments.
METHODS: We propose a reinforcement learning–based robust scheduling framework. The Chebyshev inequality transforms probabilistic signal-to-noise ratio (SNR) constraints into deterministic bounds using statistical moments. A multi-objective optimization problem is formulated to maximize served terminals and minimize resource fragmentation. The Proximal Policy Optimization (PPO) algorithm enables intelligent allocation across network slices under dynamic conditions.
RESULTS: Simulation results demonstrate that the proposed approach achieves higher scheduling success rates and better resource utilization compared with baseline methods. The reinforcement learning agent adapts effectively to environmental variations, maintaining stable performance even under severe CSI outdatedness.
CONCLUSION: The robust reinforcement learning–based scheduling provides an effective solution for NB-IoT over LEO satellites, enhancing reliability and scalability of future 6G global IoT networks.

Keywords: Low Earth Orbit satellite (LEO), NB-IoT, Robust Resource Scheduling, PPO

Received on 14 October 2025, accepted on 13 April 2026, published on 20 April 2026

Copyright © 2026 N. N Song *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetsis.10581

1. Introduction

With the rapid advancement of information technology, sixth-generation (6G) mobile communication systems are moving toward the dual visions of global coverage and ubiquitous connectivity[1-3]. As a key enabler of IoT, low-power wide-area networks, such as NB-IoT, have already achieved significant progress in terrestrial networks[4-6]. To overcome the inherent coverage limitations of terrestrial networks, it is essential to construct an integrated space–air–ground communication infrastructure to realize the 6G vision[7]. In this context, LEO satellite communication systems, with their advantages of low latency and reduced path loss, have emerged as a pivotal technology to complement terrestrial NB-IoT networks and support global IoT services in the 6G era[8].

Despite these advantages, emerging LEO-IoT systems face significant challenges due to constrained satellite resources and highly dynamic traffic conditions. The fundamental limitations of current research approaches

*Corresponding author. Email: lyz753951852@126.com

include: (1) Limited scalability due to fixed spectrum resources and increasing IoT device density, where traditional static allocation methods cannot adapt to dynamic traffic patterns; (2) Inadequate handling of channel state information (CSI) aging caused by high satellite mobility (7.5 km/s) and long propagation delays (5-10 ms), leading to significant mismatch between scheduled and actual channel conditions [9]; (3) Lack of robust mechanisms to guarantee quality-of-service (QoS) under uncertainty, as most existing methods rely on perfect CSI assumptions that are unrealistic in LEO scenarios [10]. Kodheli et al. addressed uplink resource limitations by modeling physical resource block allocation as a two-dimensional knapsack problem and solving it with a heuristic algorithm, achieving approximately a 25% improvement in spectral efficiency[11]. From a fairness perspective, Maity proposed a network-slicing-based fairness-aware scheduler that combines weighted greedy algorithms with simulated annealing to balance resource allocation among services with heterogeneous QoS requirements[12]. Huang et al. tackled the coupled challenges of task offloading and resource allocation by introducing mobile edge computing, constructing a multi-layer game model, and applying deep reinforcement learning, thereby improving computational efficiency while reducing latency[13]. Researchers devised weighted greedy scheduling and simulated annealing algorithms based on priority to determine bandwidth allocation and duration according to traffic demand[14].

Resource optimization under dynamic traffic conditions has also been extensively studied. Gou proposed a joint optimization algorithm for resource allocation and task scheduling, achieving a trade-off between energy efficiency and latency through dynamic weight adjustment[15]. Considering the time-varying nature of traffic loads, Lin designed a hybrid scheduling strategy that combines resource reservation with dynamic adjustment for bursty traffic scenarios, leveraging predictive models to guide allocation and enhance system adaptability[16]. Furthermore, Chu investigated channel phase mismatch in multi-beam LEO-IoT networks, modeling power control and beamforming jointly as a non-convex optimization problem to minimize total transmit power[17]. Xu develops a multi-agent deep reinforcement learning approach with a hybrid action space for task offloading and resource allocation in space–air–ground integrated networks[18]. Considering multiple satellites, cloud servers and UAV platforms, the approach aims to reduce energy usage and latency. Researchers addressed uplink rate prediction errors for massive environmental monitoring terminals, proposing an uncertainty-aware time-slot rescheduling framework based on Monte Carlo sampling, dynamically adjusting link assignments within a rolling horizon[19]. Experimental results across eight typical coverage scenarios demonstrate that this approach can improve effective uplink throughput by 10–15%, highlighting the robustness of the proposed scheduling framework against rate fluctuations.

Recent advancements in robust optimization for satellite communications have demonstrated the importance of uncertainty quantification. For CSI aging mitigation, Zhang et al. proposed an autoregressive channel prediction model combined with Kalman filtering to compensate for outdated channel estimates in high-speed LEO scenarios [20]. In the domain of deep reinforcement learning applications to resource allocation, the Proximal Policy Optimization (PPO) algorithm has shown superior performance in handling continuous action spaces and ensuring training stability compared to traditional policy gradient methods [21]. Existing studies typically rely on channel prediction, conservative margin designs, or fast feedback/adaptive mechanisms. However, these approaches face inherent limitations in highly dynamic low Earth orbit channels, making it challenging to allocate resources and ensure reliability without heavily relying on accurate CSI predictions [22].

This paper makes the following contributions: (1) Robust Optimization Framework: We address QoS degradation caused by outdated CSI by establishing a time-varying channel model based on the first-order autoregressive AR(1) process. The key challenge is that satellite-side resource allocation decisions are made based on terminal-reported channel state information, while actual transmission occurs under evolved channel conditions, creating a mismatch between assumed and true channel states. To tackle this, we transform probabilistic signal-to-noise ratio (SNR) constraints into deterministic conservative bounds using Chebyshev's inequality. This transformation explicitly establishes the quantitative relationship between safety margins and confidence levels, relying only on channel mean and variance without requiring higher-order statistics or long-sequence prediction, making it suitable for practical onboard scheduler implementation. (2) Reinforcement Learning-Based Scheduling Algorithm: We propose a Proximal Policy Optimization (PPO)-based deep reinforcement learning framework that learns optimal resource scheduling policies in dynamic LEO satellite environments. The PPO algorithm employs a clipped surrogate objective function to ensure stable policy updates while preventing large policy deviations during training. The actor-critic architecture features shared convolutional layers for efficient state representation extraction, processing high-dimensional information including real-time channel conditions, terminal service requirements, and resource utilization status. The advantage estimation mechanism, implemented through Generalized Advantage Estimation (GAE), enables the agent to effectively balance the dual objectives of maximizing the number of served terminals while minimizing resource fragmentation. This learning-based approach adapts to dynamic beam coverage patterns as LEO satellites traverse their orbits, where the number of terminals, their spatial distribution, and traffic demands within a single beam footprint vary continuously. (3) Performance Validation and Analysis: Through extensive simulations based on Iridium satellite constellation parameters and 3GPP Release 17 NB-IoT specifications,

we demonstrate that the proposed approach achieves superior performance compared to baseline methods across multiple metrics including scheduling success rates, resource utilization efficiency, and convergence speed. The results validate the effectiveness of combining robust optimization with deep reinforcement learning for LEO satellite NB-IoT resource scheduling.

2. System Model and Problem Formulation

Fig.1 shows the resource–scheduling scenario for a LEO satellite IoT system. At time t_0, an IoT terminal initiates contention-based random access by sending Msg_1 and, in the uplink, reports the corresponding channel state information CSI_0. At $t_0+\tau_1$, the satellite replies with the random-access response Msg_2; at $t_0+\tau_1+\tau_2$, the terminal transmits Msg_3; and at $t_0+\tau_1+\tau_2+\tau_3$, the satellite sends the contention-resolution message Msg_4 to determine the UE that wins access. The total RA delay is therefore $t_{\mathrm{RA}}=\tau_1+\tau_2+\tau_3$.After the satellite obtains CSI_0 from the RA stage and the terminal's service request, it proceeds to resource scheduling. When the allocation decision is finally made at time $t_1=t_0+\Delta t$, the channel has already evolved to CSI_1, where $\Delta t=t_{\mathrm{RA}}+t_{\mathrm{RS}}$ and t_{RS} accounts for processing, scheduling, and queuing delays. These accumulated delays cause a mismatch between the CSI used for decision making and the actual CSI during transmission, which can lead to noticeable performance differences.

LEO satellite NB-IoT system resource scheduling needs to simultaneously consider two dimensions: service efficiency and resource utilization. Service efficiency refers to the number of terminals successfully served by the system, while resource utilization reflects the degree of effective allocation of system resources. For these two objectives, the following optimization problems have been constructed:

Optimization objective 1: Maximize the number of served terminals. Define a binary vector $\mathbf{v}=\{v_1,v_2,\ldots,v_N\}$, where $v_i=1$ indicates that terminal i successfully accesses the NB-IoT network, and $v_i=0$ indicates unsuccessful access. The service efficiency objective function is:

$$\max \sum_{N}^{i=1} v_i, \tag{1}$$

this function directly quantifies the total number of terminals successfully served by the system; a larger value indicates stronger system service capability.

Optimization objective 2: Minimize resource fragmentation waste. LEO satellite systems have limited spectrum resources, and resource fragmentation significantly reduces overall system performance. Define the resource waste function: $\mathrm{Waste}(i,S_i^{SC})$ represents the resource waste allocated to user i on a specific subcarrier set S_i^{SC}, with the specific formula as follows:

$$\begin{aligned} Waste(i,S_i^{SC}) = &\sum_{j\in K-S_{sc}^i}\left(\left(T_{sc}^i+\left(N_{RU}^i\times N_i^{slot}\times N_{rep}^i\right)\right)-S_j\right)^+ \\ &+\sum_{j\in S_{sc}^i}\left(T_{sc}^i-S_j\right)^+ \end{aligned} \tag{2}$$

where $(x)^+=\max(x,0)$. Therefore, the resource utilization optimization objective is:

$$\min \sum_{i=1}^{N}\mathrm{Waste}(i,S_i^{SC}). \tag{3}$$

where N represents the total number of IoT terminals, W represents the system frequency domain resources, and T_{sc} represents the time of current resource allocation. This optimization objective design pursues both system service capacity and optimizes resource utilization efficiency, improving overall system performance by reducing resource fragmentation.

Considering the characteristics of low Earth orbit satellite NB-IoT systems, the constraint conditions mainly consist of two categories: resource physical constraints and business QoS constraints.

Resource exclusivity constraint: The same resource unit cannot be concurrently assigned to multiple terminals:

$$C_1:\sum_{i=1}^{N} o_{b,w,i}\le 1,\forall b=1,2,...,B,\forall w=1,2,...,W \tag{4}$$

where $O_{b,w,i}$ indicates whether subcarrier b in subframe w is allocated to terminal i.

Channel quality constraint: When users transmit data, the SNR requirement must be satisfied, expressed as:

$$C_2:SNR_{dB}(i)=10\log_{10}\left(\frac{p_i\cdot\left|h_i^{t+\Delta t}\right|^2}{N_i^{sc}BN_0}\right)\ge SNR_{dB}^{req}(MCS_i,BER_i) \tag{5}$$

where p_i is the terminal transmission power, B represents the subcarrier bandwidth, N_0 represents the noise power, and $SNR_{dB}^{req}(MCS_i,BER_i)$ represents the required signal-to-noise ratio threshold for a given modulation and coding scheme MCS_i and target bit error rate.

Regarding the CSI over time issue: Due to the high-speed motion and long propagation delay of LEO satellites, there exists a difference between the obtained channel state and the actual channel state during transmission:

$$h_i^{t+\Delta t}=\rho_i(\Delta t)h_i^t+\sqrt{1-\rho_i(\Delta t)^2}\,\varepsilon_i, \tag{6}$$

where $\rho_i(\Delta t)$ is the channel time correlation coefficient, and ε_i is the channel estimation error. This uncertainty makes constraint C2 probabilistic, which is a key issue that needs to be addressed in this chapter.

Business QoS constraints include: 1) Delay constraint: Terminal data transmission must be completed within the maximum allowed delay:

$$C_3: T_{sc}^i + (N_i^{RU} \times \frac{N_i^{slot}}{2} \times N_i^{rep}) \le T_i^{req} + d_i \quad (7)$$

where T_{sc}^i is the data open transmission delay, T_i^{req} is the required transmission delay, and d_i is the maximum tolerable delay.

Reliability constraint: The transmission success rate must reach the specified requirement:

$$C_4: 1-(1-P_i^s)^{N_{rep}^i} \ge R_i \quad (8)$$

where $P_i^s = (1-BER_i)^{D_i}$ represents the probability that data D_i is successfully transmitted in a single transmission, and $1-(1-P_i^s)^{N_i^{rep}}$ is the overall success rate after N_i^{rep} repetitions.

Data volume constraint: Allocated resources need to satisfy terminal transmission requirements:

$$C_5: N_i^{RU} = \begin{cases} \left\lceil \frac{D_i}{r(MCS_i)\times 16} \right\rceil, if\ N_i^{sc} = 1 \\ \left\lceil \frac{D_i}{r(MCS_i)\times 24} \right\rceil, otherwise \end{cases} \quad (9)$$

where $r(MCS_i)$ is the data rate of MCS_i.

Combining the optimization objectives and constraint conditions, the complete optimization problem for intra-slice resource scheduling in low Earth orbit satellite NB-IoT systems is:

$$\max_{N_{sc}^i, N_{RU}^i, N_{rep}^i, T_{sc}^i, S_{sc}^i, MCS_i} \left(\sum_{i=1}^{N} v_i \Big/ N \right) + \left(1 - \sum_{i=1}^{N} Waste(i, S_i^{sc}) \Big/ \left(W \cdot T_{sc} \right) \right)$$

$$C_1: \sum_{i=1}^{N} o_{b,w,i} \le 1, \forall b = 1,2,...,B, \forall w = 1,2,...,W$$

$$C_2: SNR_{dB}(i) = 10\log_{10}\left(\frac{p_i \cdot \left| h_i^{t+\Delta t} \right|^2}{N_i^{sc} BN_0} \right) \ge SNR_{dB}^{req}(MCS_i, BER_i)$$

$$C_3: T_{sc}^i + (N_i^{RU} \times \frac{N_i^{slot}}{2} \times N_i^{rep}) \le T_i^{req} + d_i$$

$$C_4: 1-(1-P_i^s)^{N_{rep}^i} \ge R_i$$

$$C_5: N_i^{RU} = \begin{cases} \left\lceil \frac{D_i}{r(MCS_i)\times 16} \right\rceil, if\ N_i^{sc} = 1 \\ \left\lceil \frac{D_i}{r(MCS_i)\times 24} \right\rceil, otherwise \end{cases} \quad (10)$$

where C1 denotes the resource exclusivity constraint, ensuring that each resource unit can only be allocated to one terminal at a time. C2 represents the channel quality constraint, which requires that the SNR during transmission must meet the minimum threshold. C3 is the latency constraint, guaranteeing that the data transmission is completed within the specified delay limit. C4 corresponds to the reliability constraint, which enforces that the transmission success probability satisfies the predefined reliability requirement. Finally, C5 is the data volume constraint, ensuring that the allocated resources are sufficient to meet the terminal's data transmission demand.

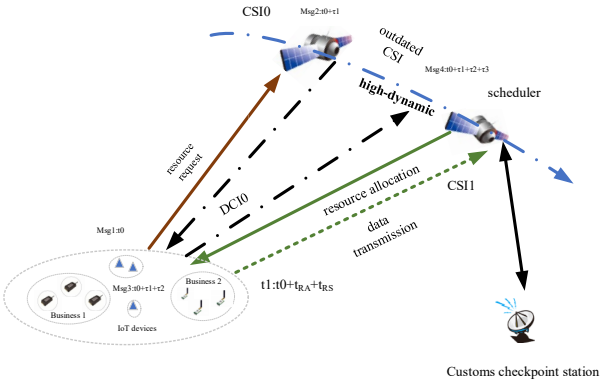

Figure 1. System Model

3. Resource Scheduling Optimization Approach Based on Deep Reinforcement Learning

The random constraint C2 in the optimization problem needs to be transformed into a deterministic constraint in order to be solved. This section uses the Chebyshev inequality to transform it into a deterministic constraint.

To facilitate subsequent mathematical processing, the SNR constraint is transformed into a linear form. After substituting the channel model, the original expression can be written as:

$$SNR_{dB}(i) = 10\log_{10}\left(\frac{p_i \cdot \left| h_i^{t+\Delta t} \right|^2}{N_i^{sc} BN_0} \right) \ge SNR_{dB}^{req}(MCS_i, BER_i) \quad (11)$$

By transformation, we obtain:

$$\frac{p_i \cdot |h_i^{t+\Delta t}|^2}{N_i^{sc} BN_0} \ge SNR_i^{req} \quad (12)$$

where $SNR_i^{req} = 10^{SNR_{dB}^{req}/10}$ is the linear-domain SNR requirement.

Considering the time-varying channel model:

$$\left| h_i^{t+\Delta t} \right|^2 = \left| \rho_i(\Delta t) h_i^t + \sqrt{1-\rho_i(\Delta t)^2}\, \varepsilon_i \right|^2 \quad (13)$$

which can be expanded as:

$$|h_i^{t+\Delta t}|^2 = |h_i^t|^2 |\rho|^2 + (1-|\rho|^2)|\varepsilon_i|^2 + 2\sqrt{1-|\rho|^2}\, Re\{\rho^* h_i^t \varepsilon_i^*\}$$

(14)

Due to the randomness of channel states, the deterministic constraint is no longer applicable. Therefore, it is reformulated into a probabilistic constraint:

$$P\{p_i \cdot |h_i^{t+\Delta t}|^2 \geq SNR_i^{req} N_i^{sc} BN_0\} \geq 1-\eta \quad (15)$$

where η denotes the predefined violation probability($0 < \eta < 1$).

Let the random variable be defined as $Z = |h_i^{t+\Delta t}|^2$. Based on the properties of complex Gaussian distribution, its statistical characteristics can be obtained as:

$$E[Z] = |\rho|^2 |h_i^t|^2 + (1-|\rho|^2)\sigma^2 \quad (16)$$

$$Var[Z] = (1-|\rho|^2)^2 \sigma^4 + 2|\rho|^2 (1-|\rho|^2) |h_i^t|^2 \sigma^2 \quad (17)$$

According to the Chebyshev inequality:

$$P\{|X-\mu| \geq k\sigma\} \leq \frac{1}{k^2} \quad (18)$$

Applying this to our problem and letting $k = \sqrt{1/\eta}$, the deterministic equivalent constraint C2' can be obtained as:

$$\text{C2':} \quad p_i \cdot \left(E[Z] - \frac{\sqrt{Var[Z]}}{\sqrt{\eta}} \right) \geq SNR_i^{req} N_i^{sc} BN_0 . \quad (19)$$

The physical meaning of this transformation is that the average channel gain is reduced by a safety margin related to the variance. The margin is determined by the required reliability level.

Based on the above derivation, the complete deterministic optimization problem can be formulated as:

$$\max_{N_{sc}^i, N_{RU}^i, N_{rep}^i, T_{sc}^i, S_{sc}^i, MCS_i} \quad \frac{\sum_{i=1}^{N} v_i}{N} + \left(1 - \frac{\sum_{i=1}^{N} Waste(i, S_i^{SC})}{W \cdot T_{sc}} \right) \quad (20)$$

$$\text{s.t.} \qquad C1, C2', C3, C4, C5.$$

Due to the high-dimensional non convex nature of the problem and strict constraints, traditional optimization methods are difficult to efficiently solve in the dynamic environment of LEO satellites. This article proposes a solution method based on PPO.

The state space should contain all information required by the agent at each decision moment. Define the system state vector $S_t : S_t = [h_t, r_t, q_t]$, where $h_t = [h_1(t), \ldots, h_N(t)]$ represents the channel state information at the current moment, r_t represents the system resource utilization state, and $q_t = [(D_i, R_i, d_i)]_{i=1}^N$ represents the QoS requirement set of terminals.

The action space defines the specific actions that the agent can take at each moment. In this study, the action space design is based on the decision variables of the optimization problem, $A_t = \{N_{sc}^i, N_{RU}^i, T_{rep}^i, T_{sc}^i, S_{sc}^i, MCS_i\}$.

The reward function is designed as:

$$R_t = \lambda_1 \left(\sum_{i=1}^{N} v_i \Big/ N \right) + \lambda_2 \left(1 - \sum_{i=1}^{N} Waste(i, S_i^{SC}) \Big/ \left(W \cdot T_{sc} \right) \right) - \lambda_3 \sum_{c \in \{C1 \sim C5\}} Penalty(c) \quad (21)$$

where the first term represents the number of terminals successfully served by the system, the second term quantifies the degree of resource waste, and the third term Penalty(c) represents the penalty for violating constraint conditions, which is a piecewise function that equals 1 if constraints are violated and 0 if constraints are not violated.

The core of the PPO algorithm is to limit the policy update step size through a policy clipping mechanism, balancing exploration and exploitation, and improving training stability. The core objective function of the algorithm can be expressed as:

$$L^{PPO}(\theta) = \mathsf{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)] \quad (22)$$

where the policy probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$ reflects the degree of change between new and old policies. The clipping parameter ε limits the policy update step size to ensure training stability. The advantage function $\hat{A}_t$ serves as a baseline for policy gradients to reduce estimation variance.

In terms of network architecture design, this study constructs a policy network $\pi_\theta(a \mid s)$ and a value network $V_\phi(s)$. The policy network is responsible for generating six-dimensional decision variables in the action space, while the value network is used to evaluate the state value function. The advantage function calculation adopts the Generalized Advantage Estimation (GAE) method:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \ldots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \quad (23)$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (24)$$

where γ is the discount factor and λ is the GAE parameter. Based on this, the algorithm updates parameters by alternately optimizing the policy network and value network, ultimately achieving the resource scheduling optimization objective. This design of the PPO algorithm ensures training stability and effectively solves the resource scheduling optimization problem constructed in this paper. Algorithm 1 below presents the specific procedure of the PPO-based resource scheduling algorithm.

Algorithm 1 PPO-based Robust Resource Scheduling Algorithm

Initialize: Initialize policy network $\pi(s;\theta_\pi)$ and value network $V(s;\theta_v)$,set PPO parameters $(\gamma,\lambda,\varepsilon_{clip},\alpha)$

for epoch = 1 to MAX EPOCHS:
  for episode = 1 to NUM EPISODES:
    while not done:
      for each terminal i:
        Calculate robust lower bound according to Eq.(20) using Chebyshev inequality
        Select action based on current policy $\alpha \sim \pi(s;\theta_\pi)$; execute action a, observe new state s', reward r, termination flag done
        Store data
  for iteration = 1 to NUM ITERATIONS:
    Use GAE to calculate advantage function A(s, a), update policy network, update value network, perform gradient updates
    Evaluate current policy and check convergence
  Return optimal resource allocation strategy

4. Numerical Results

The simulation system is based on Iridium LEO satellite parameters and 3GPP R17 standard[23], and the PPO algorithm parameters are detailed in Table 1.

Table 1. System and Training Parameters

Parameter	Value
Satellite Orbital Altitude	780 km
Number of Terminals	1000
Channel Temporal Correlation Coefficient	0.85
Channel Estimation Error Variance	0.25
Carrier Frequency	1616MHz
System Bandwidth	180kMz
Subcarrier Spacing	15 kHz/3.75 kHz
Background Noise Power Spectral Density	-174 dBm/Hz
Training Episodes	120
Episodes per Training Round	8
Discount Factor	0.99
GAE Parameter	0.95
PPO Clipping Parameter	0.2
Learning Rate	1e-4
Batch Size	64
Confidence Level	0.95
Network Structure	[256,128,64]

As observed in Fig.2, although the proposed robust algorithm exhibits a lower initial reward value, it demonstrates rapid learning capability and surpasses the non-robust algorithm after approximately 20 training episodes. This phenomenon indicates that by incorporating the Chebyshev inequality to handle CSI uncertainty, the algorithm initially adopts a conservative strategy but quickly adapts to the characteristics of the system. During the mid-training phase (episodes 40-80), the robust algorithm's reward value continues to rise and stabilizes, while the non-robust algorithm shows slow performance improvement. In the late training phase, both algorithms reach steady states; however, the robust algorithm achieves a final convergence value approximately 25% higher than its counterpart, demonstrating the effectiveness of the robust optimization approach in addressing CSI uncertainty.

As shown in Fig.3, in the low error region (variance less than 0.1), the non-robust algorithm outperforms the robust algorithm. This occurs because the robust algorithm incorporates probabilistic guarantee constraints through the Chebyshev inequality, reserving certain performance margins to handle channel uncertainty. Under conditions of small channel errors, this conservative resource allocation strategy actually limits the system capacity. The non-robust algorithm, which relies entirely on current CSI for decision-making, can utilize system resources more effectively, thereby achieving higher service success rates under low error conditions. As the channel error variance increases, The non-robust algorithm's performance deteriorates rapidly, primarily due to its complete dependence on estimated CSI for resource allocation. When substantial discrepancies exist between the actual channel state at transmission time and the estimated values, transmissions that originally satisfied SNR requirements may experience interruptions[24]. In contrast, the robust algorithm transforms probabilistic constraints into deterministic ones through the Chebyshev inequality, incorporating both the expectation and variance characteristics of channel gains during resource allocation. This probability-guaranteed design ensures that the system maintains stable service quality with high probability even when channel states change.

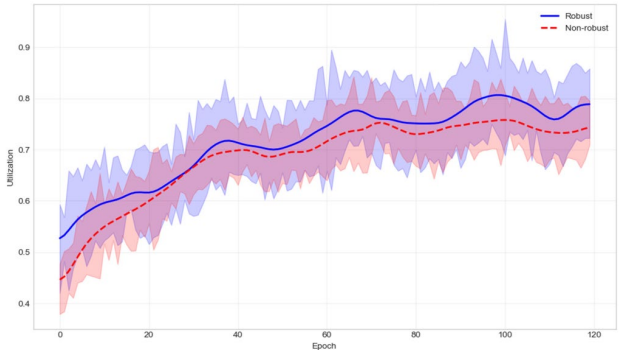

Figure 2. Comparison of Reward over Training Episodes

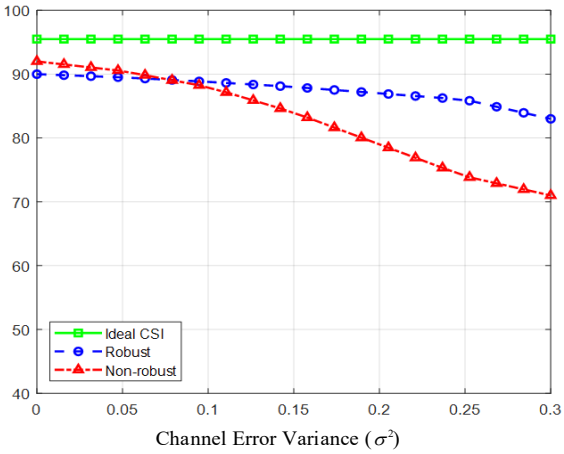

Figure 3. Effect of Channel Estimation Error Variance on Successfully Scheduled Terminal Count

As illustrated in Fig.4, this paper compares the average transmission delay of the proposed robust algorithm, non-robust algorithm, and random scheduling algorithm under varying numbers of terminals. The results show that as the number of terminals increases, the average delay of all three algorithms exhibits an upward trend, yet the magnitude of increase differs significantly. The robust algorithm consistently maintains the lowest average delay through reasonable resource reservation and scheduling optimization. When the number of terminals increases from 100 to 1000, the robust algorithm's delay increases only from 120 to 135ms, representing an increase of approximately 12.5%. In comparison, the non-robust algorithm's delay increases from 110ms to 140ms, reaching an increase of 27.3%, while the random scheduling algorithm performs worst with delay increasing from 155ms to 185ms. This result validates that reinforcement learning-based scheduling strategies can effectively optimize resource allocation and reduce queueing waiting time.

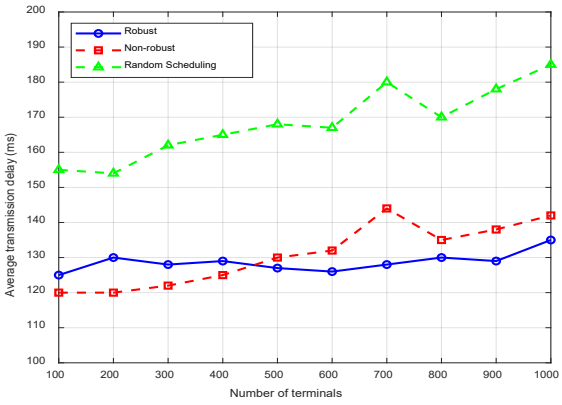

Figure 4. Effect of Channel Error Variance on Resource Utilization Rate

5. Conclusion

This paper addresses the critical challenge of CSI outdatedness in LEO satellite NB-IoT systems by proposing a reinforcement learning-based robust resource scheduling framework. The method integrates an AR(1) channel prediction model with Chebyshev inequality-based constraint transformation to convert stochastic channel conditions into tractable deterministic constraints, and employs a PPO-based algorithm to intelligently solve the multi-objective resource scheduling problem under channel uncertainty. Simulation results demonstrate that the proposed method significantly outperforms traditional non-robust scheduling algorithms in scheduling success rate and resource utilization efficiency, particularly under high-mobility scenarios with severe Doppler effects, while maintaining low computational complexity suitable for real-time operations. Statistical analysis confirms these improvements are robust across various system configurations and channel conditions. However, several limitations merit acknowledgment. The AR(1) channel model may inadequately capture dynamics in scenarios with severe shadowing or multipath effects, and scalability remains a concern for mega-constellation deployments. Additionally, the trained policy exhibits limited generalization when transferred across significantly different orbital configurations without fine-tuning. The current study does not encompass comprehensive evaluation under extreme operational conditions such as dense satellite constellations, sudden large-scale terminal access events, or severe weather-induced channel perturbations, which are essential for precisely delineating robustness boundaries. Furthermore, practical deployment faces challenges in regional adaptation, as the Chebyshev-based constraint transformation requires accurate SNR statistical moment estimation while PPO training demands extensive high-quality samples across diverse scenarios. Regional variations in channel environments and terminal distributions necessitate customized data collection and parameter calibration strategies that require further investigation. Looking forward, several promising directions emerge from this work. Future research should explore hybrid approaches combining physics-informed neural networks with statistical models to improve prediction accuracy under diverse propagation conditions, including severe weather scenarios. Extending the framework to multi-agent reinforcement learning would enable coordinated resource allocation across satellite constellations in mega-constellation deployments. Incorporating meta-learning and transfer learning techniques could enable rapid policy adaptation across different orbital parameters, geographical regions, and traffic patterns with minimal retraining overhead. Developing online statistical estimation algorithms for continuously updating channel moment information would enhance practical deployability. Comprehensive stress testing under extreme operational conditions, including flash crowd terminal access scenarios and dense constellation dynamics, should be conducted to identify failure modes and establish operational boundaries. Collaboration with satellite operators for multi-regional field trials would provide invaluable insights into practical deployment challenges, enable validation with real-world telemetry data, and facilitate development of automated calibration procedures for region-specific adaptation. In conclusion, this work establishes a promising foundation for intelligent, robust resource scheduling in next-generation LEO satellite IoT systems, opening new pathways toward fully autonomous satellite network management while clearly identifying the critical steps required for transitioning from research prototype to operational deployment.

Acknowledgements

This work was supported by the Research and Application of Power Emergency Repair Communication Technology Based on the Integration of Beidou-3 and Tiantong Satellites under Grant 5205J0240001.

References

[1] Hassebo A. The Road to 6G, Vision, Drivers, Trends, and Challenges. In: IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC); 2022; Las Vegas, NV, USA. IEEE; 2022. p. 1112-1116.

[2] Iqbal M, Abdullah AYM, Shabnam F. An Application Based Comparative Study of LPWAN Technologies for IoT Environment. In: IEEE Region 10 Symposium (TENSYMP); 2020; Dhaka, Bangladesh. IEEE; 2020. p. 1857-1860.

[3] Azari MM, Solanki S, Chatzinotas S, Kodheli O, Sallouha H, Colpaert A, Mendoza Montoya JF, Pollin S, Haqiqatnejad A, Mostaani A, Lagunas E, Ottersten B. Evolution of Non-Terrestrial Networks From 5G to 6G: A Survey. IEEE Commun Surv Tutor. 2022; 24(4):2633-2672.

[4] Mangalvedhe N, Ratasuk R, Ghosh A. NB-IoT deployment study for low power wide area cellular IoT. In: IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); 2016; Valencia, Spain. IEEE; 2016. p. 1-6.

[5] Charbit G, Medles A, Jose P, et al. Satellite and Cellular Networks Integration - A System Overview. In: Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit); 2021; Porto, Portugal. IEEE; 2021. p. 118-123.

[6] Azari MM, Solanki S, Chatzinotas S, et al. Evolution of non-terrestrial networks from 5G to 6G: A survey. IEEE Commun Surv Tutor. 2022; 24(4):2633-2672.

[7] Gou L, Bian D, Nie Y, Zhang G, Zhou H, Shi Y, Zhang L. Hierarchical Resource Management for Mega-LEO Satellite Constellation. Sensors. 2025; 25(3):902.

[8] Kodheli O, Lagunas E, Maturo N, et al. Satellite communications in the new space era: A survey and future challenges. IEEE Commun Surv Tutor. 2020; 23(1):70-109.

[9] Y. Liu, X. Xie, and Z. Wang. CSI feedback with model-driven deep learning of massive MIMO systems. IEEE Communications Letters, 2020; 24(3): 518-522.

[10] T. Zhang, W. Chen, F. Yang, et al. Letaief. Robust transmission design for intelligent reflecting surface-aided secure communication systems with imperfect cascaded CSI. IEEE Transactions on Wireless Communications. 2022; 21(4): 2487-2501.

[11] Kodheli O, Andrenacci S, Maturo N, Chatzinotas S, Zimmer F. Resource Allocation Approach for Differential Doppler Reduction in NB-IoT over LEO Satellite. In: Advanced Satellite Multimedia Systems Conference and the 15th Signal Processing for Space Communications Workshop (ASMS/SPSC); 2018; Berlin, Germany. IEEE; 2018. p. 1-8.

[12] Maity I, Chougrani H, Chatzinotas S. Fairness-Aware Inter-Slice Scheduler for IoT Services Over Satellite. IEEE Open J Commun Soc. 2023; 4:3040-3050.

[13] Huang C, Chen G, Xiao P, Xiao Y, Han Z, Chambers JA. Joint Offloading and Resource Allocation for Hybrid Cloud and Edge Computing in SAGINs: A Decision Assisted Hybrid Action Space Deep Reinforcement Learning Approach. IEEE J Sel Areas Commun. 2024; 42(5):1029-1043.

[14] Maity I, Chougrani H, Chatzinotas S. Fairness-Aware Inter-Slice Scheduler for IoT Services Over Satellite. IEEE Open J Commun Soc. 2023; 4: 3040-3050.

[15] Gou L, Bian D, Nie Y, Zhang G, Zhou H, Shi Y, Zhang L. Hierarchical Resource Management for Mega-LEO Satellite Constellation. Sensors. 2025; 25(3):902.

[16] Lin Z, Ni Z, Kuang L, Jiang C, Huang Z. Satellite-Terrestrial Coordinated Multi-Satellite Beam Hopping Scheduling Based on Multi-Agent Deep Reinforcement Learning. IEEE Trans Wirel Commun. 2024; 23(8):10091-10103.

[17] Chu J, Chen X, Zhong C, et al. Robust Design for NOMA-Based Multibeam LEO Satellite Internet of Things. IEEE Internet Things J. 2021; 8(3):1959-1970.

[18] Xu H, Chen X, Huang X, et al. Uncertainty-Aware Scheduling for Effective Data Collection from Environmental IoT Devices through LEO Satellites. Future Gener Comput Syst. 2025; 166(5): 107656-107672.

[19] Lee J, Park C, Park S, Molisch AF. Handover Protocol Learning for LEO Satellite Networks: Access Delay and Collision Minimization. IEEE Trans Wirel Commun. 2024; 23(7):7624-7637.

[20] J. Zhang, Y. Wei, E. Björnson, et al. Performance analysis and optimization of MIMO-NOMA downlink with imperfect CSI. IEEE Transactions on Communications. 2020; 68(5): 2982-2996.

[21] J. Schulman, F. Wolski, P. Dhariwal, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[22] A. Guidotti, A. Vanelli-Coralli, M. Conti, et al. Architectures and key technical challenges for 5G systems incorporating satellites. IEEE Transactions on Vehicular Technology. 2019; 68(3): 2624-2639.

[23] Merias P. Study on Narrow-Band Internet of Things (NB-IoT)/enhanced Machine Type Communication (eMTC) support for Non-Terrestrial Networks (NTN). In: 3GPP TSG RAN, editor. Proceedings of the 3GPP TSG RAN Meeting RP-86; 2019 Dec 9-12; Sitges, Spain. Sophia Antipolis: 3GPP; 2019. (Doc. RP-193235).

[24] Lin Z, Lin M, Champagne B, et al. Robust Hybrid Beamforming for Satellite-Terrestrial Integrated Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 8792-8796