Computer Science and Information Systems 23(1):277–297 https://doi.org/10.2298/CSIS250923008L

Adaptive Bandwidth Allocation via Uncertainty-Constrained Deep Reinforcement Learning

Li Wei, Wu Yong and Yan Dong

Suzhou Suneng Group Co., LTD
215004 Suzhou, China
{liwei_sz,wuy11,yandong_sz}@js.sgcc.com.cn

Abstract. With the rapid growth of network services, traditional static bandwidth allocation schemes can no longer meet the demands of multi-user, dynamic, and QoS-sensitive applications. Ensuring both efficiency and stability in bandwidth allocation remains a significant challenge, especially under high variability and uncertainty conditions. To address this, we propose a novel algorithm named Uncertainty-Constrained Stability-aware Deep Reinforcement Learning (UCS-DRL) for dynamic bandwidth allocation. UCS-DRL adopts a dual-policy architecture: a task policy that learns optimal bandwidth allocation decisions, and a stability policy guided by uncertainty-aware value estimation to identify and mitigate potential risky or unstable behaviors during deployment. Furthermore, the framework incorporates a curiosity-driven exploration mechanism based on Random Network Distillation, which enhances exploration efficiency by encouraging the agent to visit informative and under-explored states. Experimental results show that UCS-DRL achieves high bandwidth utilization and service quality while reducing policy volatility and risky actions, balancing performance and robustness in dynamic bandwidth allocation.

keywords: Dynamic Resource Allocation, Reinforcement Learning, Uncertainty Estimation, Stability-aware Control, Dual-policy Framework.

1. Introduction

In the context of the rapid development of smart substations, power systems are placing higher demands on the efficiency of bandwidth allocation and the intelligence of scheduling in communication networks[1]. The variety of services deployed within substations is continuously increasing[2]: on one hand, applications such as remote monitoring, relay protection signal transmission, and real-time equipment status reporting require stable and continuous bandwidth, with high sensitivity to transmission delay and reliability; on the other hand, event-driven services such as fault alarms and emergency command dispatch often generate large volumes of data in a short time, exhibiting strong time-criticality and transient high-bandwidth consumption. With emerging applications such as 4K/8K high-definition video[3], VR/AR-assisted inspection, and remote interactive maintenance, network traffic volume and dynamics are further intensified. Under such an environment characterized by concurrent multi-service operations and significantly different traffic patterns, how to achieve efficient, precise, and dynamic bandwidth allocation—while ensuring quality of service for diverse applications and overall system reliability—has become a key challenge for communication networks in smart substations[4].

Traditional bandwidth allocation strategies exhibit significant limitations in such complex service environments[5]. However, they cannot adapt in real time to competition among services and frequent traffic fluctuations, often leading to inefficient utilization or imbalanced allocation of resources. When sudden traffic surges occur, these methods usually fail to adjust promptly, resulting in congestion or interruption of critical services; conversely, in low-load scenarios, they often lead to bandwidth redundancy and waste. Meanwhile, modern networks, particularly smart substations, impose multiple constraints on bandwidth scheduling[6], such as those of latency, throughput, packet loss, and fairness, which are often conflicting. Traditional methods typically optimize a single objective, making achieving comprehensive trade-offs across multiple metrics difficult. More critically, real-world networks are subject to diverse uncertainties, including sudden traffic demands, link quality fluctuations, and external disturbances, and traditional strategies often fail to maintain stability under abnormal conditions. This shortcoming is particularly severe in critical infrastructures such as intelligent substations, where failure of allocation strategies at crucial moments may directly disrupt the stable and continuous operation of the power grid[7].

Recently, reinforcement learning (RL) has been widely applied in network scheduling and resource management[8]. It gradually optimizes strategies through interaction with the environment and has strong adaptability and generalization capabilities. In particular, policy gradient methods (such as PPO[9]) and value iteration methods (such as DQN[10]) demonstrate good learning capabilities in complex state spaces. However, most existing RL methods generally neglect to model the stability and uncertainty of policy deployment[11], leading to potential uncontrollable decisions by the trained policy when faced with unexpected scenarios such as sudden traffic spikes or abnormal states, thereby impacting service stability and network stability.

To address these challenges, we propose Uncertainty-Constrained Stability-aware Deep Reinforcement Learning (UCS-DRL) that explicitly incorporates uncertainty constraints to ensure stable and reliable policy deployment. UCS-DRL constructs a dual-module collaborative training framework comprising a task and stability policies. In UCS-DRL, the task policy employs the Deep Deterministic Policy Gradient(DDPG) method to achieve efficient bandwidth utilization. It features a pluggable interface to accommodate other high-performance policy structures. The stability policy incorporates a multi-scenario uncertainty bounding mechanism that integrates optimistic, pessimistic, and most likely behavioral scenarios. UCS-DRL enables a more comprehensive characterization of the potential fluctuation range in policy outputs and allows for proactive identification of extreme risk boundaries before action execution. Additionally, UCS-DRL combines a curiosity mechanism to expand the policy exploration space and enhance generalization capabilities in unknown states. Dynamically deploying UCS-DRL in the controller effectively improves throughput and service reliability while controlling end-to-end latency and deployment risks, achieving stable, efficient, and intelligent bandwidth resource allocation. Experimental results show that our method outperforms baselines in key metrics such as throughput, latency reduction, and adaptation capability, demonstrating superior performance and practical engineering value.

Our contributions are summarized as follows:

- We propose the UCS-DRL that achieves efficient and stable dynamic bandwidth allocation through a collaborative training framework of task and stability policies under uncertainty constraints.
- We design a pluggable task policy module to enhance resource allocation efficiency while maintaining training stability and convergence performance.
- We construct a stability policy based on multi-scenario uncertainty modeling that considers optimistic, pessimistic, and most likely behavioral estimates, balancing constraint and adaptability in policy output space.
- We combine internal policy uncertainty evaluation with external exploration signals from random network distillation to improve environmental awareness, policy generalization, and deployment robustness under distribution shifts.

The remainder of this paper is organized as follows: Related works are given in Section 2. The problem is modeled in Section 3. The details of UCS-DRL are explained in Section 4. The experiments and results are presented and analyzed in Section 5. Eventually, the paper ends in Section 6 with conclusions.

2. Related Work

2.1. Traditional Bandwidth Scheduling Methods

In early network bandwidth scheduling, allocation strategies were primarily based on static rules that divided bandwidth by user type, service priority, service level, or traffic weight[12]. For example, static scheduling mechanisms such as Weighted Round Robin (WRR)[13] and Weighted Fair Queuing (WFQ)[14] are widely used in traditional network devices, offering advantages such as simple implementation, low computational complexity, and controllable scheduling overhead. However, these methods heavily rely on manually designed rules, making it challenging to maintain efficient resource utilization and service quality assurance in complex business models and dynamic network conditions. To overcome the shortcomings of static methods in adapting to environmental changes, researchers have proposed a series of dynamic scheduling algorithms based on heuristic ideas. For example, Shortest Path First (SPF)[15] and Fastest Bandwidth Lowest Delay[16] strategies are applied in the joint selection of paths and bandwidth to improve overall network throughput and task completion efficiency. Meanwhile, fuzzy logic control has also been introduced into network resource scheduling, enhancing the decision-making flexibility of schedulers in uncertain environments. A typical example is the Fuzzy Bandwidth and Delay Guaranteed Routing Algorithm(FBDRA)[17] models bandwidth requirements and delay tolerance through a fuzzy inference system, thereby improving the satisfaction rate of QoS metrics while maintaining system stability. Although these heuristic methods demonstrate good practicality and performance in small-to-medium-sized network environments, they exhibit poor generalization capabilities and scalability when faced with real-world scenarios such as high-dimensional state spaces, heterogeneous multi-service requirements, and complex topological dynamic changes[18]. This has provided research motivation for developing more intelligent and adaptive scheduling algorithms.

2.2. Traffic Prediction and Proactive Resource Allocation

To address the traffic dynamics caused by concurrent multi-service operations in smart substations, researchers increasingly focused on proactive resource allocation methods[19] based on traffic prediction in recent years. Unlike traditional scheduling mechanisms that rely on real-time status feedback, proactive scheduling anticipates future traffic trends over a certain period and preemptively reserves bandwidth. It adjusts resources, enhancing the capability to guarantee service quality for critical applications. For instance, periodically reported equipment status data and scheduled video inspection tasks exhibit strong temporal regularity, making them suitable for high-precision traffic prediction through time series modeling[20]. For unexpected events such as fault alarms and protective actions, their occurrence probabilities and data volumes can be predicted by combining historical fault patterns with operational conditions, using statistical learning or machine learning methods. Previous studies have employed traditional time-series analysis methods such as ARIMA and Hidden Markov Models (HMM) for industrial communication traffic prediction. Recently, deep learning models including Long Short-Term Memory (LSTM)[21], Gated Recurrent Units (GRU), and Transformer[22] have demonstrated superior capabilities in capturing complex nonlinear traffic characteristics. Nevertheless, most existing methods still rely on offline training, making them difficult to adapt to the dynamic changes of substation operating environments. Moreover, they seldom consider the impact of prediction errors on the robustness of resource allocation, which limits their practical effectiveness in real-world systems.

2.3. Reinforcement Learning

With the development of deep reinforcement learning technology, its application in intelligent network control has received increasing attention[23]. Reinforcement learning has the ability to "learn through interaction," enabling it to gradually approach optimal strategies through trial-and-error exploration mechanisms[24]. It is naturally suited to network scheduling problems characterized by high-dimensional state spaces, complex system dynamics, and sparse rewards. In existing research, value-based methods such as DQN (Deep Q Network) have been widely used in discrete bandwidth allocation and routing selection scenarios[25]. For example, by encoding path and bandwidth allocation actions, a Q-network is trained to obtain the policy with the maximum cumulative reward, thereby optimizing system throughput or load balancing effects. Researchers have adopted policy gradient methods such as PPO (Proximal Policy Optimization) for scheduling scenarios with continuous actions to achieve refined resource control and rapid response adjustments[26]. In addition, to improve exploration efficiency and generalization capabilities, some studies have combined intrinsic incentive mechanisms such as Random Network Distillation(RND)[27] and Intrinsic Curiosity Module(ICM)[28] to guide agents to focus on insufficiently explored state regions and avoid falling into local optima. In multi-objective scheduling scenarios, some studies have introduced multi-task reinforcement learning architectures, incorporating multiple metrics such as bandwidth utilization, latency, service quality, and stability risks into a unified reward function to improve the system's overall performance[29].

In real-world network deployments, ensuring policy safety and robustness is critical. Safe reinforcement learning addresses this through constraint-based optimization and

uncertainty-aware methods[11]. Representative approaches like Penalized Proximal Policy Optimization (P3O)[30] formalize safety as cost constraints, while Robust RL frameworks [31,32] enhance robustness by explicitly accounting for model uncertainty or adversarial disturbances. Building on these foundations, our work introduces an uncertainty-constrained mechanism to jointly optimize performance and safety in network control.

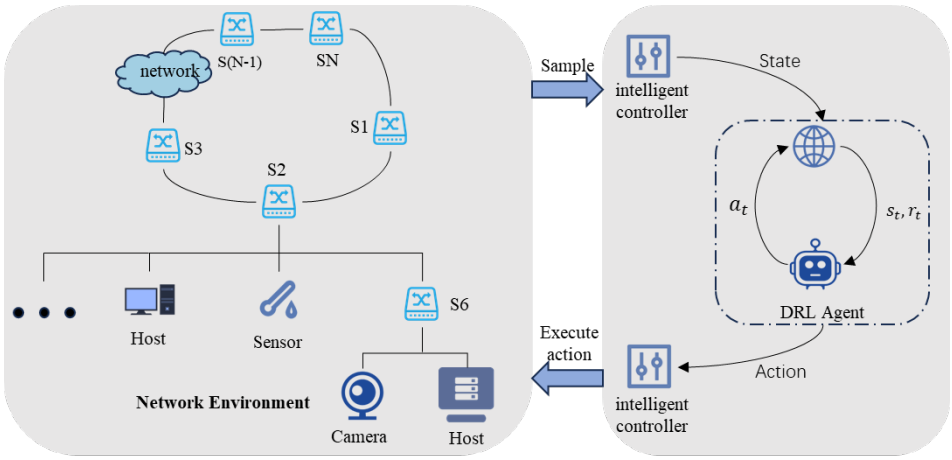

Fig. 1. Network topology: a ring backbone with tree-structured access networks, highlighting one representative access node and its connected end devices

3. System Model and Problem Formulation

3.1. System Model

We investigate the problem of dynamic bandwidth allocation at the access layer in smart substation communication networks. Considering that highly reliable ring-based backbone networks are widely adopted in practical engineering to ensure network connectivity, while actual bandwidth contention primarily occurs within the tree-structured subnetworks connected to nodes on the ring, we focus on the local bandwidth scheduling scenario under a single node. In this architecture, a communication node on the ring connects to multiple end devices—such as high-definition video cameras, protection relays, condition monitoring sensors, and local monitoring hosts—via local access links. These devices share the upstream link bandwidth of the node, forming a typical "multi-source to single-sink" traffic aggregation pattern.

To more realistically simulate such scenarios, we construct the network topology shown in Fig. 1 in a simulation environment. The topology consists of a ring-based backbone network, with several end-user nodes connected under each switching node, forming a ring-tree hybrid structure. We focus on one representative node in the network, connecting multiple end devices — including high-priority protection devices, high-bandwidth

video surveillance equipment, medium-priority status sensors, and a local monitoring host capable of data aggregation and command dispatch. These devices share an upstream link via a local switch to access the ring network, with the total link bandwidth set to $B_{total} = 100$ Mbps. Due to significant differences among services in delay sensitivity, data volume, and transmission continuity, the absence of an effective scheduling mechanism may lead to delays or packet loss for critical services due to bandwidth contention.

During the simulation, the system periodically collects real-time status information of each traffic flow, including transmission rate, queue length, end-to-end delay, and packet loss, serving as input for the bandwidth allocation policy. This sensing mechanism is implemented by deploying lightweight monitoring modules at local nodes, without requiring global topology information, thus meeting substations' localized and low-overhead scheduling requirements. Based on these local state feedbacks, the scheduler dynamically adjusts the bandwidth allocation for each traffic flow to ensure quality of service for high-priority applications while improving overall resource utilization efficiency. At each time step, the intelligent controller can obtain the following network state information: each user's bandwidth request volume $d_i(t)$; each link's delay $q_l^{delay}(t)$; the user's actual bandwidth usage in the previous cycle $u_i(t-1)$; and the user's service priority p_i. These observations are obtained through port queues, forwarding table statistics, or probe mechanisms via the switching device interface, with specific calculations as follows:
Bandwidth request estimation: Use the queue rate to estimate user i instantaneous bandwidth request:

$$d_i(t) \approx \frac{\Delta Q_i(t)}{\Delta t} = \frac{Q_i(t) - Q_i(t-\Delta t)}{\Delta t}\,. \tag{1}$$

Link delay estimation: Use the controller to actively detect the delay of link l:

$$q_l^{delay}(t) = \frac{1}{2} \cdot RTT_l(t), \tag{2}$$

where $RTT_l(t)$ measured by ICMP or probe packets.

Estimated actual bandwidth usage: The user's actual bandwidth usage in the previous cycle is estimated based on the byte change recorded in the forwarding table:

$$u_i(t-1) = \frac{\Delta Bytes_i(t-1)}{\Delta t}\,. \tag{3}$$

Through the above system design, we establish a simulation network environment that supports multi-user access, features dynamic link states, and reflects realistic bandwidth scheduling requirements. This environment serves as a foundational platform for training and evaluating reinforcement learning algorithms. Based on the typical "ring-tree" hybrid topology of smart substations, this model fully incorporates the service heterogeneity, traffic burstiness, and delay sensitivity present in practical communication scenarios, effectively capturing the complexity and uncertainty inherent in bandwidth scheduling at the access layer of power systems.

3.2. Problem Formulation

Based on the aforementioned system model, we formalize the bandwidth allocation task as a reinforcement learning problem and models it as a Markov Decision Process (MDP)[33].

At each discrete time step *t*, the controller needs to make a round of bandwidth allocation decisions based on the current network state to achieve optimal resource scheduling while ensuring network stability and user service quality. Let the set of terminal users be $U = 1, 2, ..., N$.

State space S: The system state at time step t is s_t, which includes user bandwidth requirements, link quality, historical usage, and priority, i.e.:

$$s_t = \{d(t), q(t), u(t), p\} \ . \tag{4}$$

Action space A: The controller selects bandwidth allocation actions based on the current state.

$$a_t = [a_1(t), a_2(t), ..., a_N(t)], \quad s.t. \sum_{i=1}^{N} b_i(t) \leq B_{total}, b_i(t) \geq 0 \, . \tag{5}$$

Reward function R: Measures the effectiveness of current bandwidth allocation, satisfies user bandwidth requirements (satisfaction), maintains good link quality, and utilizes bandwidth resources as efficiently as possible (resource utilization):

$$R_t = \lambda_1 \cdot R_{th}(t) - \lambda_2 \cdot R_{lat}(t) + \lambda_3 \cdot R_{ada}(t) \, , \tag{6}$$

where:

$$R_{th}(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i(t)}{d_i(t) + \varepsilon} \tag{7}$$

represents throughput score (positive). A value of this indicator closer to 1 denotes that it better meets user's need.

$$R_{lat}(t) = \frac{1}{L} \sum_{i=1}^{L} \frac{q_l^{delay}(t)}{q^{max}} \tag{8}$$

indicates delay penalty (negative), where L represents the number of links that are actively probed by the controller. This item is a value between 0 and 1. The higher the value, the more severe the delay.

$$R_{ada} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(a_i(t) \in A_i^{stable}(t)) \tag{9}$$

denotes the adaptation score (positive). A higher score indicates better adaptation, which is defined by two key behaviors: bandwidth allocation aligned with actual link capacity, and allocation actions remaining within predefined policy confidence boundaries under dynamic network conditions.

State transition function P: determined by dynamic changes in the network, including link load fluctuations, traffic release, and sudden congestion, and exhibits a certain degree of uncertainty.

Discount factor $\gamma \in (0, 1)$: indicates the discount for future rewards.

The ultimate objective is to solve for that optimal policy that maximizes the desired long-term cumulative discount reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum \gamma^t r_t\right] \, . \tag{10}$$

The complexity of this problem lies in the fact that the user demand changes rapidly, the link state fluctuates frequently, and the traditional static bandwidth allocation policy is difficult to adapt. The introduction of a reinforcement learning framework can utilize the continuous interaction process between the controller and the environment to autonomously learn adaptive dynamic strategies and achieve fine-grained and efficient allocation of bandwidth resources.

4. Methodology

4.1. Overview

To address dynamic bandwidth allocation challenges, we propose UCS-DRL, a stability-aware deep reinforcement learning algorithm based on uncertainty constraints, designed for intelligent and robust scheduling of network resources. The algorithm adopts a dual-policy coordination framework, integrating a task policy for performance optimization with a stability policy for risk-constrained decision-making. By jointly evaluating task rewards and behavioral risks, UCS-DRL ensures high network performance while enhancing deployment robustness. The task policy leverages high-performance reinforcement learning architectures to generate efficient bandwidth allocation sequences. The stability policy employs a multi-scenario uncertainty bounding mechanism[34], adapted from project risk analysis, to estimate policy behavior under optimistic, pessimistic, and most likely conditions - ensuring performance remains within stable operational margins. To mitigate over-constraint and encourage exploration of informative states, we integrate a curiosity-driven exploration mechanism that enriches the data space without compromising stability. Guided by the task policy, the stability policy, enhanced with uncertainty-aware risk screening and curiosity exploration, serves as the final decision module. The two policies are functionally complementary and trained asynchronously, forming the core of UCS-DRL. The algorithm's workflow is illustrated in Fig. 2, and its pseudocode is presented in Algorithm 1.

4.2. Task Policy

A performance-oriented efficient policy optimization mechanism. We use the deep deterministic policy gradient algorithm (DDPG)[27] as the core policy framework to achieve dynamic perception and response control of network status. DDPG is an actor-critic-based reinforcement learning algorithm suitable for high-dimensional, continuous control tasks, making it ideal for bandwidth allocation with adjustable granularity.

For modeling, we encode the current network state as a vector s_t, including key variables such as user demand, link quality, previous allocation records, and user priority. The actor network takes s_t as input and outputs a continuous action vector a_t, i.e., the bandwidth allocation recommendation. To ensure the feasibility of actions, a softmax operation is applied and multiplied by the total bandwidth B_{total} to satisfy the bandwidth constraint.

DDPG improves training stability via target networks and experience replay. The critic network is updated using the Bellman equation to estimate the Q-value function:

$$Q(s_t, a_t) = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}))\,, \tag{11}$$

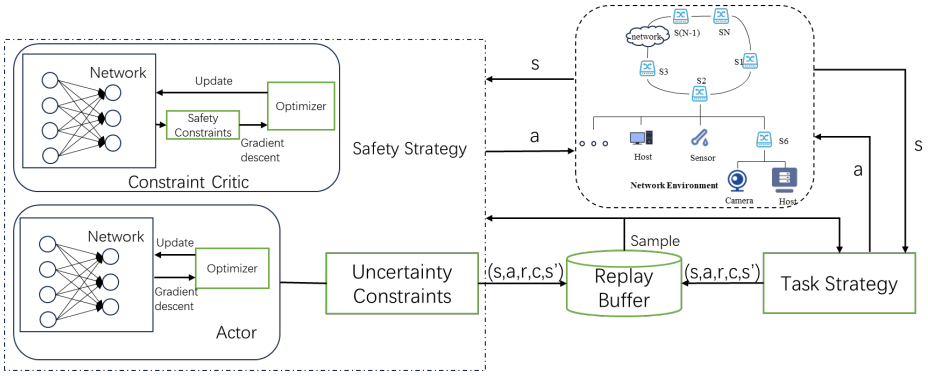

Fig. 2. Illustration of overall architecture of UCS-DRL that primarily consists of four components: the environment,the task policy utilizing the off-policy algorithm, the stability policy employing the actor-critic framework, and the data space

where $\mu'(s)$ and $Q'(s,a)$ are the target actor and critic networks, respectively, and r_t is the immediate reward based on bandwidth matching and link stability. The actor network is optimized by maximizing the expected Q-value gradient:

$$\nabla_{\theta^\mu} J \approx \mathbb{E}\left[\nabla_a Q(s,a|\theta^Q)|_{a=\mu(s)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)\right] . \tag{12}$$

UCS-DRL effectively enhances the performance of bandwidth allocation strategies while maintaining action continuity.

4.3. Stability Policy

Stability-oriented value constraint policy mechanism. After the reinforcement learning agent has constructed a policy with high task performance but possible stability risks, the next step is to further guide it to generate a policy with higher stability while ensuring performance. In the bandwidth allocation problem proposed in this paper, stability is reflected in the resource allocation process by avoiding key link overload and ensuring the service continuity of high-priority tasks. Therefore, we introduce a stability policy framework to impose soft stability constraints during the training process to constrain the direction of policy updates. Specifically, the stability policy acts as an auxiliary critic that estimates the long-term risk of state-action pairs, penalizing decisions that could lead to future instability (e.g., persistent congestion or service interruption). The penalty term is dynamically weighted and incorporated into the policy gradient update, allowing the agent to continuously trade off between maximizing task rewards and minimizing stability violations during training. The stability policy framework is designed based on a value function optimization architecture, utilizing a dual-network structure: the actor network outputs deterministic actions, while the evaluator network assesses the value of state-action pairs. Additionally, we introduce a constrained evaluator network[35] that integrates real-time stability costs into the traditional value function, i.e.:

$$Q_{rc} = r + \gamma \max_{a'} Q_{rc}(s', a') - c\,, \tag{13}$$

where r is the immediate reward, c is the immediate cost incurred due to violations of bandwidth constraints (such as overload, conflicts, or unmet priorities), and γ is the discount factor. The constraint value function influences the actor's policy through gradient updates, enabling it to perceive and avoid potential stability risks during the decision-making process. During training, both the constraint commentator network and the target network adopt soft update strategies to ensure training stability. The agent policy is optimized based on the gradient of the constraint value function, thereby achieving a resource allocation policy that balances efficiency and stability. The resulting policy effectively controls the expansion direction of the data space, making the reinforcement learning model more controllable and stable during bandwidth scheduling, particularly suitable for dynamic and complex network environments.

4.4. Uncertainty Constraints and Diversity

In reinforcement learning-driven bandwidth allocation problems, uncertainty in the environment and policy may lead to risky behaviors such as resource conflicts, performance degradation, or policy overfitting. To address this, we introduce an uncertainty constraint mechanism that combines the multi-scenario uncertainty bounding method in project management to dynamically evaluate the lower bound threshold of task performance, thereby balancing the agent's pursuit of performance while constructing a secure data space. The mechanism bounds performance uncertainty by integrating optimistic, pessimistic, and most-likely scenario evaluations, providing a probabilistic safeguard against over-optimistic policy updates. Specifically, we use the Beta distribution model to calculate the task performance threshold μ_r, which is updated every K episodes during training. Based on the maximum, minimum, and average values of episodic returns within that cycle, we derive three scenario-based estimates: T_o(optimistic) reflects the maximum achievable throughput under ideal network conditions such as minimal interference and full resource availability; T_p(pessimistic) represents the minimum sustainable performance under worst-case disturbances including congestion or traffic bursts; T_m(most likely) captures the expected performance under typical operational loads and serves as the baseline for stable allocation. These estimates are then used to compute the following formula:

$$\mu_r = \frac{T_o + 4T_m + T_p}{6}\,. \tag{14}$$

This dynamic threshold is used to filter samples in the experience pool, controlling the quality of the data space and guiding the policy to maintain high bandwidth allocation efficiency under stability constraints. To alleviate the inhibition of policy exploration capabilities caused by excessive constraints, we introduce an intrinsic motivation mechanism to encourage agents to actively explore unknown state spaces.There have been many excellent results in exploration, such as the curiosity-driven mechanism ICM [36] and the never give up (NGU) [37]. The way of intrinsic reward-driven exploration in reinforcement learning mainly realizes these exploration algorithms. In our experiments, we employ Random Network Distillation (RND) [38] to score states for novelty. RND consists of a fixed set of random feature extraction networks $f(s)$ and a learnable network $\hat{f}(s)$, where the prediction error represents the unfamiliarity of the state:

$$r_t^{\text{curiosity}} = \|f(s_t) - \hat{f}(s_t)\|^2\,. \tag{15}$$

Curiosity rewards are linearly weighted into task rewards to form the final optimization goal:

$$r_t^{\text{final}} = r_t^{\text{task}} + \lambda_{\text{curiosity}} \cdot r_t^{\text{curiosity}}. \tag{16}$$

The task policy dominates sampling and updates the policy network, while the stability policy performs offline constraint screening on behavior samples to form candidate strategies for deployment. The curiosity module is embedded in the task policy in the form of rewards, encouraging agents to move beyond their normal behavior areas and form a healthy data flow cycle. Ultimately, the controller will use only strategies that meet both performance and stability requirements for actual network bandwidth allocation tasks.

Algorithm 1 Uncertainty-Constrained Stable Deep Reinforcement Learning

1: **Input:** Network topology E
Initialize constraint critic network $Q_c(s, a|\theta^{Q_c})$ and actor $\mu(s|\theta^{\mu})$
Initialize target constraint network Q_c' and μ'
Replay buffer R, the task policy π_{task}, the stability policy π_{stable}, the list $ep\_min\_return\_list$, the task performance constraint threshold ep_min_return in Algorithm 2
2: **Output:**π_{stable}
3: **for** $episode$ = 1,N **do**
4: Initialize a random process $\mathcal{N}$ for action exploration
5: Receive initial observation state s_0
6: **for** t = 1,T **do**
7: Select action $a_t = \mu(s_t|\theta^{\mu}) + \mathcal{N}_t$ according to the current policy π_{stable} and exploration noise
8: Execute action a_t and observe reward r_t, cost c_t, next state s_{t+1}
9: Calculate curiosity rewards: $r_t^{\text{curiosity}} = \|f(s_t) - \hat{f}(s_t)\|^2$
10: Calculate total rewards: $r_{total} = r_t + \beta \cdot r_{curiosity}$
11: **if** $r_t \geq ep\_min\_return$ **then**
12: $\Pi_{stable} = \Pi_{stable} \cup (s_t, a_t, r_{total}, c_t, s_{t+1})$
13: **end if**
14: Set $y_i = r_{total} + \gamma * Q_c'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q_c'}) - c_i$
15: Update constraint critic by minimizing the loss: $L = \frac{1}{N}\sum_i \left(y_i - Q_c\left(s_i, a_i \mid \theta^{Q_c}\right)\right)^2$
16: Update the actor policy π_{stable} using the sampled policy gradient
17: Update the target network:
$\theta^{Q_c'} = \tau\theta^{Q_c} + (1-\tau)\theta^{Q_c'}$
$\theta^{\mu'} = \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$
18: **end for**
19: **end for**

In the Algorithm 1, it's crucial to highlight that we utilize the task performance threshold calculated with uncertainty as a screening criterion (Line 9-11) for the target data space. Additionally, we incorporate the immediate cost as a soft stability constraint within the expected Q value in the stability policy(Line 12). This process culminates in the generation of a stability data space that takes task performance considerations into account.

Algorithm 2 Uncertainty Constraint Mechanism

Input: The list $ep\_min\_return\_list$ consisting of the minimum per-episode returns of K episodes is used as input
Output: The expected minimum return per episode ep_min_return as a task performance constraint for the next K episodes
Sort the list $ep\_min\_return\_list$ in ascending order in place
$T_p = ep\_min\_return\_list[0]$
$T_o = ep\_min\_return\_list[-1]$
$T_m = \text{avg}(ep\_min\_return\_list)$
$ep\_min\_return = (T_p + T_m \times 4 + T_o)/6$

5. Experiments and Results

5.1. Experimental Setups

To validate the effectiveness and robustness of the proposed uncertainty-constrained stable deep reinforcement learning bandwidth allocation algorithm (UCS-DRL) in real-world network scenarios, we conducted systematic simulation experiments. All experiments were conducted on a unified hardware platform and simulation environment to ensure fairness in comparison and reproducibility of the results. The experiments were conducted in two phases: training and deployment. During the training phase, we conducted training in cycles of 100 steps per round, totaling 1,000 training cycles, and recorded algorithm performance changes every 10 rounds to observe convergence trends. During the deployment phase, we selected the converged strategies from the algorithms, conducted independent tests, recorded the average metric results per round, and statistically analyzed the final stable performance to demonstrate practical deployability.

In terms of hyperparameter settings, the discount factor γ is set to 0.98, and the batch size is set to 128. The soft update coefficient τ of the target network is set to 0.005, while the learning rate of the actor network is set to 3×10^{-4} and that of the critic network is set to 3×10^{-3}. Additionally, the number of hidden layers for both the actor and critic networks is 2. A curiosity exploration mechanism based on Random Network Distillation (RND) is introduced to assist in expanding the training data space for the policy, with the exploration reward ratio set to 0.01.

5.2. Baselines

In order to comprehensively evaluate the performance of the UCS-DRL algorithm, we have selected three representative comparison methods, covering static resource allocation strategies, fuzzy logic control methods, and typical deep reinforcement learning algorithms. These methods are widely used in bandwidth scheduling and QoS routing optimization and other related research fields, and have a good representative comparison.
Static:A static scheduling policy with fixed proportional bandwidth allocation.
DDPG[27]: A classical deep reinforcement learning method.
IFRA-GLB[39]: A method based on fuzzy logic and reinforcement learning.
UCS-DRL: An innovative algorithmic framework proposed in this paper.

5.3. Results and Analysis

Reward Curve Analysis. Fig. 3 illustrates the cumulative reward curves over training episodes for the proposed UCS-DRL method and two baseline algorithms, DDPG and IFRA-GLB. In the initial stages of training, all methods exhibit substantial reward fluctuations, which is attributed to the agent's exploratory behavior in an uncertain environment. As training progresses, however, the proposed UCS-DRL approach demonstrates a significantly faster convergence rate and achieves a higher asymptotic reward plateau compared to the baselines. This superior performance highlights the effectiveness of incorporating uncertainty-based constraints into the decision-making process, enabling the agent to avoid high-variance or sub-optimal actions early in training and thus accelerate learning.Moreover, the stability of the reward curve in the later episodes aligns closely with the stabilization of key network performance metrics such as throughput and latency, indicating that the designed reward function accurately reflects real-world system performance. The shaded regions around each curve represent the standard deviation across multiple runs, further confirming that UCS-DRL not only achieves higher average rewards but also maintains greater consistency and robustness during training. In contrast, both DDPG and IFRA-GLB show slower convergence and lower final performance, underscoring the advantage of the uncertainty-aware framework in guiding more efficient and reliable policy optimization.

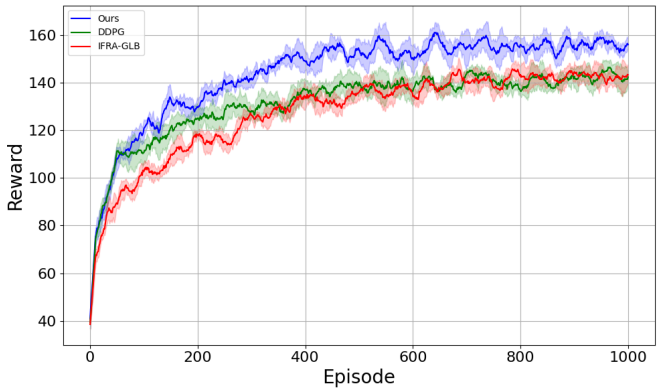

Fig. 3. Reward curves of different algorithms during training

Throughput performance. UCS-DRL demonstrates stable and continuous performance improvement throughout the training process. As shown in Fig. 4, the average network throughput increases steadily from an initial 20 Mbps to a final stabilized value of 94.5 Mbps, achieving approximately 94.5 percent of the total link capacity of 100 Mbps. This result significantly outperforms the baseline methods. DDPG and IFRA-GLB stabilize at around 82 and 86 Mbps, respectively. At the beginning of training, the agent's policy is suboptimal due to lack of experience. This leads to inefficient bandwidth allocation, underutilized links, and low throughput. However, UCS-DRL exhibits faster convergence and smaller fluctuations in the later stages compared to the benchmark algorithms.

This behavior indicates that the curiosity-driven exploration mechanism effectively encourages the agent to explore underutilized state action spaces, helping to avoid premature convergence to local optima. The task oriented policy framework guides the DDPG based actor critic structure to prioritize high impact actions. These include allocating sufficient bandwidth to high priority protection flows while opportunistically serving high bandwidth video streams. This guidance accelerates policy optimization. The integration of uncertainty aware risk control further prevents over allocation to bursty or unreliable flows, maintaining high link utilization without inducing congestion. The sustained high throughput achieved by UCS-DRL reflects its ability to dynamically balance exploration and exploitation, efficiently utilize available bandwidth, and adapt to time-varying traffic patterns. This makes it particularly suitable for smart substation networks, where both high resource efficiency and service reliability are critical.

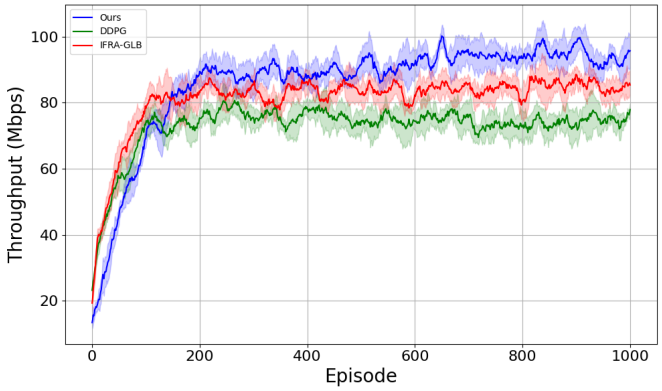

Fig. 4. Throughput performance comparison under different methods

Latency performance. UCS-DRL effectively avoids over-scheduling of highly loaded links by introducing uncertainty constraints and a risk-aware control mechanism. As illustrated in Fig. 5, the average latency of UCS-DRL decreases steadily from an initial 100 ms to 25 ms during the early training phase. Eventually, it converges to approximately 22.3 ms with minimal fluctuation. This smooth convergence indicates that the agent quickly learns an efficient bandwidth allocation policy that prioritizes delay-sensitive services while maintaining system stability under uncertain traffic conditions. In contrast, DDPG and IFRA-GLB exhibit significant oscillations in latency during the late stages of training. DDPG, which relies solely on reward shaping without explicit risk modeling, tends to over-exploit high-throughput paths, leading to temporary congestion and bursty latency spikes. IFRA-GLB, although it considers global load balancing, fails to account for traffic uncertainty and sudden bursts (e.g., fault alarms or video stream initiation), resulting in unstable performance and delayed response to dynamic changes. Moreover, UCS-DRL's uncertainty-aware design enables proactive bandwidth reservation for critical services based on predicted traffic patterns and confidence intervals. This not only prevents buffer overflow in high-priority queues but also reduces unnecessary waiting

time for time-sensitive data, which are essential for mission-critical applications in smart substations.

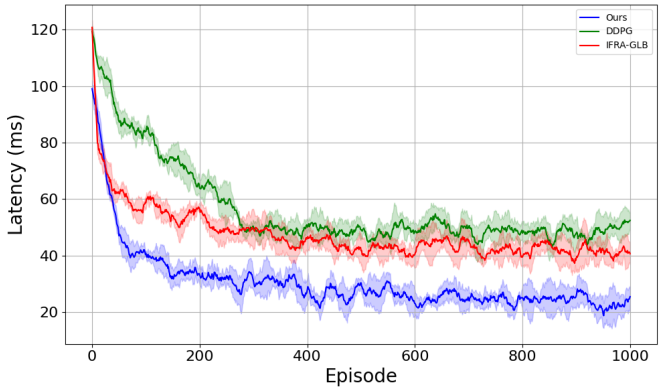

Fig. 5. Latency performance comparison under different methods

Adaptation performance. A higher Adaptation Rate indicates better system capability in dynamically adjusting to network conditions while adhering to stability and reliability constraints. These include avoiding high packet loss links, respecting bandwidth limits, and preventing congestion. UCS-DRL achieves outstanding performance in this metric. By incorporating a multi-scenario uncertainty bounding mechanism and a risk lower bound filtering operation during policy evaluation, UCS-DRL effectively suppresses actions with high uncertainty. This ensures that the selected actions are not only efficient but also compliant with operational stability requirements. As shown in Fig. 6, UCS-DRL maintains an average Adaptation Rate of approximately 97.5% in the test deployment phase, significantly outperforming DDPG (87.9%) and IFRA-GLB (91.4%). Further analysis reveals that the embedded risk control mechanism not only enhances the stability of the learned policy during deployment but also guides the training process toward the "high benefit low risk" region of the solution space. This demonstrates the effective synergy between reinforcement learning and stability-aware constraints. Unlike DDPG, which often generates aggressive actions under uncertainty, UCS-DRL prioritizes reliable and feasible decisions, making it well suited for mission-critical power communication environments.

5.4. Ablation study

Ablation study analysis on the curiosity-driven module reveals that removing the RND module leads to significant performance degradation, manifested through reduced exploration efficiency, slower convergence speed, and deteriorated policy quality. As shown in Fig. 7, without RND, the model exhibits delayed exploration progress during early training (0-400 episodes), with insufficient state-space coverage causing the policy to converge to local optima. While the complete model stabilizes after approximately 400 episodes, the ablated version fails to achieve comparable performance even after 1000 episodes,

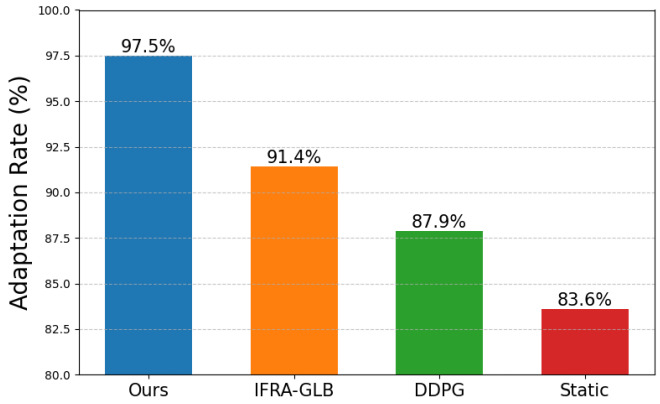

Fig. 6. Adaptation rate comparison in the deployment phase

demonstrating higher decision variance and lower adaptability. These results verify that in dynamic bandwidth allocation scenarios, the RND module effectively promotes comprehensive exploration of the network state space through its state novelty scoring mechanism, thereby preventing the resource allocation strategy from becoming overly conservative due to stability constraints. The prediction error-driven intrinsic reward provides essential exploratory momentum for optimizing bandwidth allocation strategies. By discovering non-typical resource allocation patterns, it significantly enhances the policy's adaptability and robustness in complex network environments, establishing itself as a key component in maintaining an efficient data flow cycle for bandwidth allocation.

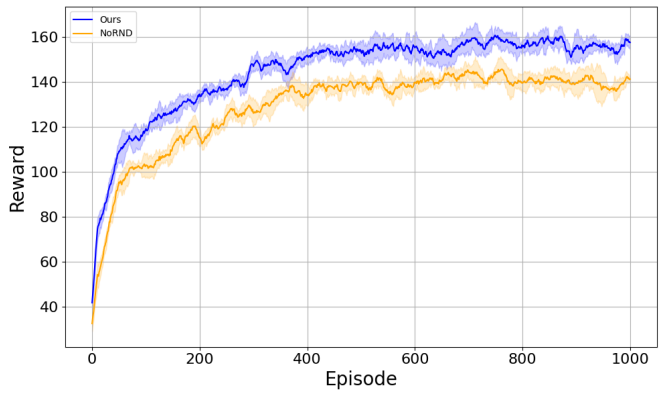

Fig. 7. Performance comparison of ablation studies

5.5. Deployment Evaluation Results

To further evaluate the generalization and stability of the trained policies, we conduct a comprehensive evaluation in the test phase, where the learned models are applied to previously unseen traffic scenarios. The performance is assessed across four key metrics: throughput, average latency, bandwidth utilization, and adaptation rate. Results are summarized in Table 1. Our proposed UCS-DRL algorithm outperforms all baseline methods across all metrics. It achieves a throughput of 91.3 Mbps and a bandwidth utilization of 91.3%, demonstrating efficient resource scheduling under diverse traffic patterns. More importantly, UCS-DRL achieves an adaptation rate of 95.2%, significantly higher than DDPG (87.9%) and IFRA-GLB (91.4%), indicating its strong ability to avoid risky actions—such as overloading links or violating bandwidth limits—while maintaining high performance. In terms of latency, UCS-DRL achieves the lowest average delay of 23.7 ms, which is critical for time-sensitive services in smart substations. This reflects its effectiveness in prioritizing high-priority flows and avoiding congestion through proactive bandwidth allocation. Fig. 8 presents a bar chart comparing the performance of all methods across the four evaluation metrics in the test phase. While static policy methods achieve moderate results in known scenarios, they lack adaptability and fail to respond effectively to dynamic and unseen conditions. In contrast, learning-based methods, particularly UCS-DRL, exhibit superior generalization and robustness due to its uncertainty-aware design and risk-constrained learning framework. These results confirm that UCS-DRL not only learns an efficient scheduling policy during training but also generalizes well to new scenarios, achieving a favorable balance between performance and stability. This makes it highly suitable for intelligent bandwidth management in complex and dynamic substation communication environments.

The above experimental results show that UCS-DRL has superior performance in terms of throughput efficiency, service delay control and adaptability, which proves the adaptability, stability and engineering practicability of the proposed algorithm in the complex dynamic network environment. Its task-stability dual-policy cooperative mechanism and uncertainty-aware design jointly construct a set of bandwidth allocation policy framework with learning capability, stability control, and stable output capability, which has good potential for promotion in future networks.

Table 1. Comparison of Algorithm Performance

Algorithm	Throughput (Mbps)	Latency (ms)	Utilization (%)	Adaptation (%)
Ours	**91.3±1.2**	**18.6±0.8**	**87.9±0.9**	**97.5±0.6**
IFRA-GLB	78.4±2.1	26.7±1.5	74.1±1.7	91.4±1.2
DDPG	83.7±1.8	22.3±1.2	81.5±1.4	87.9±1.5
Static	86.1±0.9	20.1±0.7	84.2±0.8	83.6±1.8

6. Conclusion

In this paper, we address the key challenge of balancing high performance with operational stability in dynamic bandwidth allocation by proposing a stable deep reinforce-

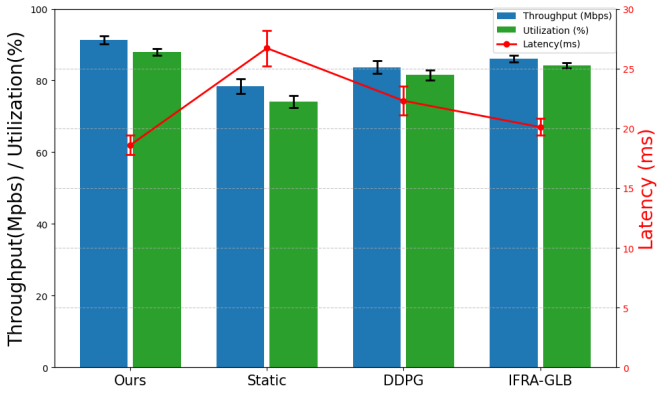

Fig. 8. Algorithm performance comparison in the deployment phase

ment learning algorithm. Our method employs a pluggable dual-policy framework: a performance-oriented task policy that maximizes bandwidth utilization and allocation efficiency, and a stability policy that constrains high-risk decisions via multi-scenario uncertainty modeling (optimistic, pessimistic, and most likely) and a value-based conservative critic, thereby directly mitigating the instability risks inherent in RL-based control. To enhance exploration and avoid local optima, we integrate a curiosity-driven mechanism that guides the agent toward unfamiliar states in early training, improving policy generalization. The reward function jointly optimizes throughput, delay, and stability adaptation rate, enabling balanced learning between performance and deployment robustness.

We evaluate our approach in a simulated multi-user environment. Extensive experiments demonstrate that our method outperforms baselines in both task performance and stability maintenance, significantly reducing policy violations during resource allocation. Ablation studies confirm the critical contributions of the curiosity module and uncertainty estimation mechanism. Importantly, our framework imposes minimal constraints on the base policy and is highly compatible with diverse task settings. We also acknowledge limitations including simulation-based validation and focus on single-node scheduling. Future work will explore multi-agent reinforcement learning for cross-node coordination and test the method on physical network hardware and more complex topologies.

Acknowledgments. This work was supported by Science and Technology Project of State Grid Jiangsu Electric Power Co., Ltd., and Science and Technology Project of Suzhou Suneng Group Co., Ltd. under Grant SGSZSNJTKJJS2500979.

References

1. José Gaspar, Tiago Cruz, Chan-Tong Lam, and Paulo Simões. Smart substation communications and cybersecurity: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(4):2456–2493, 2023.
2. Weiti Lv. Research on network application automation system based on computer artificial intelligence technology. In *2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pages 1934–1938. IEEE, 2023.

3. Xiaodong Qiao, Naichen Yan, and Lu Zhang. Design and implementation of substation monitoring system based on harmonyos and broadband-narrowband wireless communication technology. In *2025 2nd International Conference on Smart Grid and Artificial Intelligence (SGAI)*, pages 1484–1488. IEEE, 2025.
4. R Mohandas and D John Aravindhar. An intelligent dynamic bandwidth allocation method to support quality of service in internet of things. *International Journal of Computing*, 20(2):254–261, 2021.
5. G Senthilkumar, KN Madhusudhan, Y Jeyasheela, and P Ajitha. A novel blockchain enabled resource allocation and task offloading strategy in cloud computing environment. *Automatika*, 65(3):973–982, 2024.
6. Nurshazlina Suhaimy, Nurul Asyikin Mohamed Radzi, Wan Siti Halimatul Munirah Wan Ahmad, Kaiyisah Hanis Mohd Azmi, and MA Hannan. Current and future communication solutions for smart grids: A review. *IEEE Access*, 10:43639–43668, 2022.
7. Lan-Huong Nguyen, Van-Linh Nguyen, Ren-Hung Hwang, Jian-Jhih Kuo, Yu-Wen Chen, Chien-Chung Huang, and Ping-I Pan. Toward secured smart grid 2.0: Exploring security threats, protection models, and challenges. *IEEE Communications Surveys & Tutorials*, 27(4):2581–2620, 2025.
8. Hoa Tran-Dang, Sanjay Bhardwaj, Tariq Rahim, Arslan Musaddiq, and Dong-Seong Kim. Reinforcement learning based resource management for fog computing environment: Literature review, challenges, and open issues. *Journal of Communications and Networks*, 24(1):83–98, 2022.
9. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, pages 1–12, 2017.
10. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
11. Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, 2024.
12. Liqing Liu, Xiaoming Yuan, Decheng Chen, Ning Zhang, Haifeng Sun, and Amir Taherkordi. Multi-user dynamic computation offloading and resource allocation in 5g mec heterogeneous networks with static and dynamic subchannels. *IEEE Transactions on Vehicular Technology*, 72(11):14924–14938, 2023.
13. Ahmed Mohammed, Nor Fadzilah Abdullah, Sameer Alani, Othman S Alheety, Mohammed Mudhafar Shaker, Mohammed Ayad Saad, and Sarmad Nozad Mahmood. Weighted round robin scheduling algorithms in mobile ad hoc network. In *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–5. IEEE, 2021.
14. Mays A Mawlood and Dhari Ali Mahmood. Performance analysis of weighted fair queuing (wfq) scheduler algorithm through efficient resource allocation in network traffic modeling. *Journal of Communications Software and Systems*, 20(3):266–277, 2024.
15. Mohit Mittal, Celestine Iwendi, Suleman Khan, and Abdul Rehman Javed. Analysis of security and energy efficiency for shortest route discovery in low-energy adaptive clustering hierarchy protocol using levenberg-marquardt neural network and gated recurrent unit for intrusion detection system. *Transactions on Emerging Telecommunications Technologies*, 32(6):1–16, 2021.
16. Inayat Ali, Seungwoo Hong, and Taesik Cheung. Quality of service and congestion control in software-defined networking using policy-based routing. *Applied Sciences*, 14(19):1–13, 2024.
17. Jianhu Gong and Hamed Nazari. A fuzzy bandwidth and delay guaranteed routing algorithm for performance enhancement of video conference over mpls networks. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7079–7090, 2023.
18. Yingnan Deng. A reinforcement learning approach to traffic scheduling in complex data center topologies. *Journal of Computer Technology and Software*, 4(3):1–6, 2025.

19. Torana Kamble, Sanjivani Deokar, Vinod S Wadne, Devendra P Gadekar, Hrishikesh Bhanudas Vanjari, and Purva Mange. Predictive resource allocation strategies for cloud computing environments using machine learning. *Journal of Electrical Systems*, 19(2):68–77, 2023.
20. Bo Wu and Yifan Hu. Analysis of substation joint safety control system and model based on multi-source heterogeneous data fusion. *IEEE Access*, 11:35281–35297, 2023.
21. Wei Sun, Pengyu Li, Zhi Liu, Xue Xue, Qiyue Li, Haiyan Zhang, and Junbo Wang. Lstm based link quality confidence interval boundary prediction for wireless communication in smart grid. *Computing*, 103(2):251–269, 2021.
22. Yang Chen, Jia Hao, Yu Peng, and Hongyan Xia. Transformer-based performance prediction and proactive resource allocation for cloud-native microservices. *Cluster Computing*, 28(9):568–590, 2025.
23. Linqiang Huang, Miao Ye, Xingsi Xue, Yong Wang, Hongbing Qiu, and Xiaofang Deng. Intelligent routing method based on dueling dqn reinforcement learning and network traffic state prediction in sdn. *Wireless Networks*, 30(5):4507–4525, 2024.
24. Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International Conference on Machine Learning*, pages 16828–16847. PMLR, 2022.
25. Jiawei Xu, Yufeng Wang, Bo Zhang, and Jianhua Ma. A graph reinforcement learning based sdn routing path selection for optimizing long-term revenue. *Future Generation Computer Systems*, 150:412–423, 2024.
26. Abdelhak Bentaleb, Mehmet N Akcay, May Lim, Ali C Begen, and Roger Zimmermann. Bob: Bandwidth prediction for real-time communications using heuristic and reinforcement learning. *IEEE Transactions on Multimedia*, 25:6930–6945, 2022.
27. David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 387–395, Bejing, China, 22–24 Jun 2014. PMLR.
28. Hongliang Zeng, Ping Zhang, Fang Li, Chubin Lin, and Junkang Zhou. Ahegc: Adaptive hindsight experience replay with goal-amended curiosity module for robot control. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16602–16615, 2024.
29. Yuqing Cheng, Zhiying Cao, Xiuguo Zhang, Qilei Cao, and Dezhen Zhang. Multi objective dynamic task scheduling optimization algorithm based on deep reinforcement learning. *The Journal of Supercomputing*, 80(5):6917–6945, 2024.
30. Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Penalized proximal policy optimization for safe reinforcement learning. In *International Joint Conference on Artificial Intelligence*, pages 1–7, 2022.
31. Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.
32. Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:16082–16097, 2022.
33. Davide Maran, Pierriccardo Olivieri, Francesco Emanuele Stradi, Giuseppe Urso, Nicola Gatti, and Marcello Restelli. Online markov decision processes configuration with continuous decision space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14315–14322, 2024.
34. Abdullah Al Hayajneh, Hasnain Nizam Thakur, and Kutub Thakur. The evolution of information security strategies: A comprehensive investigation of infosec risk assessment in the contemporary information era. *Computer and Information Science*, 16(4):1–20, 2023.
35. Claire Chen, Shuze Liu, and Shangtong Zhang. Efficient policy evaluation with safety constraint for reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, pages 1–22, 2025.

36. Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 488–489, 2017.
37. Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*, pages 1–28.
38. Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, pages 1–17, 2019.
39. Zhiqun Wang, Zikai Jin, Zhen Yang, Wenchao Zhao, and Mahdi Mir. An intelligent fuzzy reinforcement learning-based routing algorithm with guaranteed latency and bandwidth in sdn: Application of video conferencing services. *Egyptian Informatics Journal*, 27:1–11, 2024.

Li Wei is an Engineer at Suzhou Suneng Group Co., LTD of China. He has rich experience on power grid technology and application. His research interests include power system technology, substation network, and flow control.

Wu Yong is an Engineer at Suzhou Suneng Group Co., LTD of China. He has rich experience on communication technology and application. His research interests include communication technology, substation network, electric power information network Quality of Service, and flow control.

Yan Dong a Senior Engineer at Suzhou Suneng Group Co., LTD of China. His research interests include intelligent operation and maintenance, information and communication, electric power information network Quality of Service, network security, and artificial intelligence.

Received: September 23, 2025; Accepted: December 14, 2025.