URLLC Latency Minimization in Interweave CRNs Using Digital Twin and DRL Approach

Anal Paul[†], Keshav Singh[†], Chih-Peng Li[†], and Trung Q. Duong[‡]
[†]Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan
[‡]Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada
Emails: apaul@ieee.org, keshav.singh@mail.nsysu.edu.tw, cpli@faculty.nsysu.edu.tw, tduong@mun.ca

***Abstract*—In this paper, we present an innovative approach to spectrum management in cognitive radio networks (CRNs) aimed at serving ultra-reliable low-latency communication (URLLC) enabled secondary users (SUs). Unmanned aerial vehicles (UAVs) are deployed for accurate and reliable spectrum sensing (SS), enhancing cooperative spectrum sensing (CSS) effectiveness. A distinctive aspect of our methodology is the integration of digital twin (DT) technology, which, to our knowledge, has not been explored previously in the context of CRNs for bandwidth assignment to URLLC-enabled SUs. This integration facilitates more sophisticated and adaptive management of spectrum resources. Moreover, we propose a deep reinforcement learning (DRL) framework incorporating a modified proximal policy optimization (MPPO) algorithm. This algorithm is designed for better stability and convergence, outperforming the standard PPO in terms of faster convergence in the present URLLC transmission latency minimization process. Simulation results indicate that our proposed DT-based spectrum management and MPPO in CRNs result in a 27.89% increase in CRN's average throughput and a 39.94% reduction in transmission latency compared to the conventional equal resource allocation scheme.**

***Index Terms*—Cognitive radio networks, deep reinforcement learning, digital twin, spectrum sensing, ultra-reliable low-latency communication.**

I. Introduction

Wireless communication technologies proliferate and demand improved spectrum efficiency (SE) to meet the growing need for high data rates and widespread connectivity. Cognitive radio networks (CRNs) allow secondary users (SUs) to use the spectrum opportunistically, thus maximizing SE [1]. Yet, ultra-reliable low-latency communication (URLLC) integration into CRNs presents significant challenges due to its stringent reliability and latency demands [2]. Existing studies investigated self-organizing schemes for URLLC in device-to-device communications within CRNs, proposing resource allocation frameworks suitable for opportunistic spectrum access [3]. Chu *et al.* developed a method using a successive convex approximation to address probabilistic interference in CRNs [4]. These methods aimed to meet URLLC requirements and improve spectrum sharing. However, they often underperform in complex network scenarios and simplify practical implementation details. Research turned to digital twin (DT) integration with wireless networks, showing promise in resource optimization and prediction [5], [6], though DT's use in CRN spectrum management remains limited.

The dual tasks of spectrum sensing (SS) and communication in interweave CRNs become more challenging due to unpredictable primary user (PU) activity and variable channel state [7]. Mobile SUs introduce further complexity to spectrum allocation, complicating bandwidth assignment for intercell and intracell communication. The problem's non-convex nature and multiple constraints make finding closed-form solutions difficult. Deep reinforcement learning (DRL) is a capable approach to address these intricate resource allocation issues, noted for its adaptability and decision-making under uncertainty [6], [8]. The combination of DRL and DT could create a transformative strategy for CRNs, allowing them to continuously improve their policies without affecting the live network.

A. Motivations and Contributions of the Present Work

This work pioneers applying DT technology in interweave CRNs for spectrum management, particularly for bandwidth assignment in URLLC-enabled SUs. It relies on the advantages of unmanned aerial vehicles (UAVs) for SS and employs a DT framework on an edge server to enhance network visualization and resource allocation for mobile SUs. Addressing the gap in current literature, it explores the untouched potential of DT for interweaving CRNs' spectrum management and its significant role in enhancing SE. A DRL framework incorporating a modified proximal policy optimization (MPPO) algorithm is introduced to navigate the resource allocation complexities in non-convex optimization scenarios. The novel design of MPPO ensures better stability and convergence over the standard PPO algorithm [6], adapting specifically to the dynamic and intricate demands of URLLC-focused mobile SUs in CRNs.

In summary, the notable contributions of this work are:

1) A UAV-based SS methodology integrated with satellite service to enhance the reliability of CSS in CRNs.
2) The novel use of DT technology for spectrum management in CRNs, a first in the field for bandwidth assignment to URLLC-enabled SUs, provides a flexible platform for network experimentation and planning.
3) The MPPO algorithm, designed for resource allocation in interweave mobile CRNs to minimize transmission latency, is innovatively improved to offer better stability and convergence for URLLC.

II. System Model

Fig. 1 illustrates the proposed system model, which encompasses a single $\mathrm{PU_T}$, $\mathcal{I} = \{1, 2, \ldots, I\}$ aerial UAVs and $\mathcal{B} = \{1, 2, \ldots, B\}$ ground-based URLLC service enabled SUs. Notably, except for UAVs, every physical node in this system is equipped with a single antenna. Each UAV has

$\mathcal{K} = \{1, 2, \ldots, K\}$ number of antenna. UAVs are particularly favored for SS due to their potential advantage in obtaining a clear line-of-sight (LoS) from the PU_T. From SS aspects, all UAVs relay their local SS decision to a central satellite for obtaining a global sensing decision. If the global CSS determines that the PU_T is idle, the satellite distributes the available PU spectrum bands amongst the UAVs based on their traffic load requirements. An SU must be affiliated with a specific UAV to get URLLC service support.

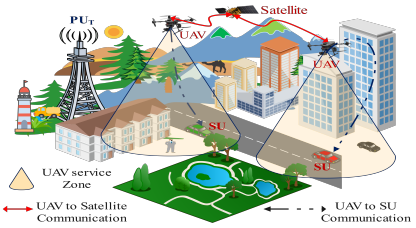


Fig. 1: UAV-aided interweave CRNs under DT framework

The entire network's DT underpins this spectrum allocation among UAVs to serve its respective affiliated URLLC-enabled SUs. DT at a mobile edge server hosts digital replicas of all physical entities in the CRN as p_q, where $q \in \mathcal{Q} = \{1, 2, \ldots, Q\}$. The set is defined as $p_q \in \{p_q^{\text{PU}_\text{T}} \cup p_q^{\text{SAT}} \cup p_q^{\text{UAV}\{\mathcal{A}\}} \cup p_q^{\text{SU}\{\mathcal{B}\}}\}$. It's imperative to note that these DTs emulate the operations of their physical counterparts. For the device associated with user p_q, its digital twin representation at a specific time point t is given by [9]

$$f_{DT}(p_q)[n] = \Theta\left(\mathcal{D}_n[n], \mathcal{S}_n[n], \mathcal{M}_n[n], \Delta\mathcal{S}_n(t+1)\right), \quad (1)$$

where, $\mathcal{D}_n$ denotes the accumulated data related to the physical device p_q, which includes configuration and past historical operational data. The term $\mathcal{S}_n[n]$ indicates the current operational state of the device p_q, which encapsulates time-variant multi-dimensional information. $\mathcal{M}_n$ represents the collection of behavior patterns for p_q that are defined by various behavioral dimensions. Lastly, $\Delta\mathcal{S}_n(t+1)$ provides the state transition of $\mathcal{S}_n[n]$ for the subsequent time slot $t+1$.

A. PU Spectrum sensing by UAVs

The channel gains from PU_T to the i-th UAV are represented by $\mathbf{h}_i \in \mathbb{C}^{1\times K}$. Taking into account the Doppler effect due to the UAV's motion, and in the context of additive white Gaussian noise (AWGN), the signal captured by the i-th UAV from PU_T at the discrete time index 'n' is:

$$y_i[n] = \mathbf{h}_i e^{j2\pi f_i n T_s} s[n] + n_i[n], \quad (2)$$

where $s[n]$ is the signal sent from PU_T with zero mean and a variance denoted as $\mathbb{E}[|s[n]|^2] = P_p$. Symbol f_i represents the Doppler frequency shift due to the UAV's motion. T_s is the sampling period and $n_i[n]$ is the AWGN at the i-th UAV, which also has zero mean and variance $\mathbb{E}[|n_i[n]|^2] = \wp_q$. The Doppler frequency for the i-th UAV is calculated as:

$$f_i[n] = \widehat{\nabla}_i[n] \cos(\varrho_i[n]) \cos(\varphi_i[n]) / \varsigma[n], \quad (3)$$

where $\widehat{\nabla}_i[n]$ is the magnitude of the relative velocity between the i-th UAV and the PU_T and $\varrho_i[n] \in [0, \pi/2]$ and $\varphi_i[n] \in [0, 2\pi)$ are the elevation and azimuth angles of departure (AoD) from the PU_T to the UAV, respectively. $\varsigma[n]$ denotes the signal's wavelength at discrete time index n. Considering Rician fading, the channel is expressed as:

$$\mathbf{h}_i[n] = \sqrt{(\text{PL})_i[n]} \sqrt{\frac{\kappa}{\kappa+1}} \mathbf{h}_i^{\text{LoS}}[n] + \sqrt{\frac{1}{\kappa+1}} \mathbf{h}_i^{\text{NLoS}}[n], \quad (4)$$

where κ is the Rician factor, denoting the power ratio of the LoS component to the scattered paths. The $(\text{PL})_i[n]$ is the path loss between PU_T and the i-th UAV and given by:

$$(\text{PL})_i[n] = 20 \log_{10}\left(\frac{4\pi f_c d_i[n]}{c}\right), \quad (5)$$

where f_c is the carrier frequency, c is the speed of light, and $d_i[n]$ is the distance between PU_T and the i-th UAV. The LoS component of the channel is determined as follows: $\mathbf{h}_i^{\text{LoS}}[n] = a_{\text{PU}_\text{T}}(\theta_i[n]) \mathbf{a}_{\text{UAV}}(\phi_i[n])$, where $a_{\text{PU}_\text{T}}(\theta_i[n])$ as the scalar response of the PU_T's antenna whose response value is set to 1 for simplicity, indicating a normalized response. While the steering vector $\mathbf{a}_{\text{UAV}}(\phi_i[n])$ is represented as:

$$\mathbf{a}_{\text{UAV}}(\phi_i[n]) = \left[1, e^{j\frac{2\pi\delta}{\varsigma}\sin\phi_i[n]}, .., e^{j\frac{2\pi\delta}{\varsigma}(K-1)\sin\phi_i[n]}\right]^{\text{T}}, \quad (6)$$

where δ represents the antenna separation distance. $\vartheta_v[n]$ represents the angle of arrival (AoA) of the signal from the PU_T to the i-th UAV at discrete time index n. T stands for transpose of the vector. The non-LoS (NLoS) component of the channel is modeled as follows:

$$\mathbf{h}_i^{\text{NLoS}}[n] = \left[1, e^{j\frac{2\pi\delta}{\varsigma[n]}\sin(\vartheta_v[n])}, .., e^{j\frac{2\pi\delta}{\varsigma[n]}(K-1)\sin(\vartheta_v[n])}\right]^{\text{T}}. \quad (7)$$

For n number of samples, which corresponds to a time duration of $\tilde{T} = n \times T_s$, the energy received E_i is calculated as $E_i = \sum_{n=0}^{N-1} |y_i[n]|^2$. The decision statistic, ζ_i, is thus: $\zeta_i = \frac{E_i}{\sigma^2}$. This statistic ζ_i is then gauged against a preset threshold λ:

$$\text{Decision}(\Phi) = \begin{cases} 1 \text{ ; i.e., PU is present} & \text{if } \zeta_i > \lambda \\ 0 \text{ ; i.e., PU is absent} & \text{if } \zeta_i \leqslant \lambda. \end{cases} \quad (8)$$

The threshold λ is established through analytical derivations, factoring in the desired detection and false alarm probabilities.

B. Probability of PU Detection and False Alarm

Each UAV makes a local SS decision and forwards its decision to the satellites for a global CSS decision.

1) Local SS decision at UAVs: The probability of a false alarm ($\mathbb{P}_{fa}$) signifies the likelihood that the decision statistic goes beyond the threshold in the absence of the PU, expressed as $\mathbb{P}_{fa} = \mathbb{P}(\zeta_i > \lambda | H_0)$. Given the Gaussian nature of the noise and accounting for the N samples, the distribution of ζ_i under H_0 is chi-squared with $2N$ degrees of freedom, the $\mathbb{P}_{fa}$ is:

$$\mathbb{P}_{fa} = 1 - F_{\chi^2_{2N}}\left(\frac{\lambda}{\sigma^2}\right), \quad (9)$$

where $F_{\chi^2_{2N}}$ signifies the cumulative distribution function (CDF) of the chi-squared (χ^2) distribution with $2N$ degrees

of freedom. This is defined as$F_{\chi^2_{2N}}(x) = \frac{1}{\Gamma(N)} \int_0^x e^{-t/2} t^{N-1} dt$. Here, $\Gamma(N)$ represents the gamma function. the probability of PU detection ($\mathbb{P}_d$) indicates the likelihood that the decision statistic surpasses the threshold when the PU is actively transmitting such that $\mathbb{P}_d = \mathbb{P}(\zeta_i > \lambda | H_1)$. When the PU is present, the distribution of ζ_i under H_1 (i.e., PU is present) has a non-central chi-squared distribution with $2N$ degrees of freedom and a non-centrality parameter λ_s:

$$\mathbb{P}_d = 1 - F_{\chi^2_{2N}}\left(\frac{\lambda}{\sigma^2}; \lambda_s\right), \tag{10}$$

where $\lambda_s = N\frac{P_p |h|^2}{\wp_q}$. The aforementioned equations in (9) aid in identifying the threshold λ for a given $\mathbb{P}_{fa}$, which is further utilized to compute the associated $\mathbb{P}_d$.

2) Global CSS decision at Satellite: Each UAV forwards its individual SS decision $\Phi \in \{0, 1\}$ to the satellite using a dedicated time division multiple access (TDMA) slot, which primarily experiences free space path loss. The global CSS decision at the satellite is based on a majority voting rule. Let $\bar{\Phi} = \sum_{i=1}^{I} \Phi_i$ represent the summation of local SS decisions, then the global CSS decision Φ_{global} will be expressed as:

$$\Phi_{\text{global}} = \begin{cases} 0 & \text{if } \bar{\Phi} < \frac{N}{2}, \\ 1 & \text{if } \bar{\Phi} \geqslant \frac{N}{2}. \end{cases} \tag{11}$$

C. Analysis of SU Mobility Model

Using our prior work in [10], we calculate the probability that the b-th SU is within the service area of UAV i, with service radius r, as the SU moves uniformly at velocity v_b in an urban grid of side ℓ. Denoting (x_i, y_i) and (x_b, y_b) as the 2D service footprint coordinates of i-th UAV and b-th SU, respectively, the probability $\mathbb{P}(I_b^i)$ is:

$$\mathbb{P}(I_b^i) = \frac{\pi r^2}{\ell^2} - \frac{8r^3}{3\ell^3} + \frac{r^4}{2\ell^4}. \tag{12}$$

III. Problem Formulation

In the present work, strategic bandwidth allocation to SUs is critical to satisfy service requirements while adhering to constraints like primary user receiver (PU_R) interference protection, UAV power limits, data rate minimums, block length caps, and URLLC latency thresholds. Our approach minimizes URLLC downlink transmission latency for the mobile SUs in CRNs. We utilize the Rician channel model, consider the Doppler effect, and apply a specific path loss model, noting that the URLLC vehicle receiver has a single antenna and operates under space division multiple access (SDMA).

A. Serving Bandwidth to Mobile SUs

The b-th SU receives bandwidth (W_b^i) from the total available PU spectrum ($W_{\max}$) through its affiliated i-th UAV. The satellite dynamically allocates bandwidth to UAVs based on their traffic demands, optimizing service quality and SE. Thus,

$$\sum\nolimits_{b=1}^{B} W_b^i \leqslant W_i^{\text{uav}} \leqslant W_{\max}, \forall i \in \mathcal{I}, \tag{13}$$

where W_i^{uav} is the bandwidth allocated to the i-th UAV.

B. Throughput calculation for URLLC-enabled vehicle

Let $\mathbf{g}_i^b$ represent the channel from the i-th UAV, with $i \in \mathcal{I}$, to the b-th URLLC-enabled SU vehicle. Given that $\mathbf{g}_i^b \in \mathbb{C}^{1\times K}$, we define the total time dedicated to serving the mobile SU as t. During this period, the registered i-th UAV serves for a duration of αt, while the remaining time, $(1-\alpha)t$, is served by other UAVs, denoted as j UAVs. Here, $0 < j < A$ and $0 < \alpha \leqslant 1$.

1) UAV to SU-vehicle channel model: The array response vector for the UAV, equipped with its multiple antennas, is defined as $\mathbf{a}(\phi) = \left[1, e^{j\frac{2\pi}{\lambda_w} d \sin(\phi)}, \ldots, e^{j\frac{2\pi}{\lambda_w}(K-1)d\sin(\phi)}\right]$, where ϕ represents the angle, which could be either AoD or AoA, λ_w denotes the signal's wavelength, d signifies the spacing between the antennas in the UAV's array. Integrating the LoS and NLoS components of the channel using AoD and AoA, we can express the channel as:

$$\mathbf{g}_i^b(\text{LoS}) = a_{\text{UAV}}(\phi_{\text{AoD}}) a_{\text{vehicle}}(\phi_{\text{AoA}})^*, \tag{14}$$

$$\mathbf{g}_i^b(\text{NLoS}) = \sum_{l=1}^{L} \beta_l a_{\text{UAV}}(\phi_{\text{AoD},l}) a_{\text{vehicle}}(\phi_{\text{AoA},l})^*, \tag{15}$$

where L is the number of multipath components and β_l represents the complex gain for the l-th scatterer. Upon combining the effects of pathloss, Rician fading, and the Doppler effect, the overall channel model becomes:

$$\mathbf{g}_i^b = \frac{1}{\sqrt{\text{PL}(d_{i,b})}} \left(\sqrt{\frac{\kappa}{\kappa+1}} \mathbf{g}_i^b(\text{LoS}) e^{j2\pi f_d n T_t} + \sqrt{\frac{1}{\kappa+1}} \mathbf{g}_i^b(\text{NLoS}) e^{j2\pi f_d n T_t} \right), \tag{16}$$

where $d_{i,b}$ denotes the distance between the i-th UAV and the b-th mobile SU, T_t is the transmission time, and h represents the altitude difference between the UAV and the SU. We define the elevation angle θ as $\theta = \tan^{-1}(h/d_{i,b})$. The path loss, $\text{PL}(d_{i,b})$, is described as:

$$\text{PL}(d_{i,b}) = \begin{cases} \text{PL}_{\text{LoS}}(d_{i,b}) & \text{if LoS is detected,} \\ \text{PL}_{\text{NLoS}}(d_{i,b}) & \text{otherwise.} \end{cases} \tag{17}$$

For the LoS scenario, $\text{PL}_{\text{LoS}}(d_{i,b}) = 20\log_{10}(4\pi d_{i,b} f_c / c) + \beta_{\text{LoS}}(h)$, where β_{LoS} is a height-dependent term for the UAV. For the NLoS scenario: $\text{PL}_{\text{NLoS}}(d_{i,b}) = \text{PL}_{\text{LoS}}(d_{i,b}) + \xi(\theta)$, where $\xi(\theta)$ is a correction factor contingent on the elevation angle, typically derived from empirical data in urban contexts.

We consider the total transmission slot $\bar{T}$. A fraction of this time, αt, is served by the i-th UAV, while the remaining time, $(1-\alpha)t$, is served by other UAVs in the set $J = \{1, 2, .., j, .., (A-1)\}$. We break the time t into n slots, such that $n \times T_t = \bar{T}$. For the i-th UAV:

$$\text{SINR}_i[n] = \frac{|\mathbf{g}_i^b|^2 P_i}{\sum_{j\neq i} |\mathbf{g}_j^b|^2 P_j + \sigma_b^2}, \tag{18}$$

where the term σ_b^2 represents the noise variance in the system and P_i denotes the transmission power of the UAV. Given the

bandwidth W_b^i and the probability in (12) for being UAV's services coverage, the URLLC rate for the i-th UAV is:

$$R_i[n] = W_b^i \mathbb{P}(I_b^i) \left(\log_2(1+\text{SINR}_i[n]) - \frac{\log_2(e)}{Q(\epsilon)} \sqrt{\frac{V}{B_i}} \right), \quad (19)$$

where $\sqrt{\frac{V}{B_i}}$ is a representation of the effective channel dispersion (V) normalized by the finite block length (FBL) B_i. Here, V is expressed as $V = 1-(1+\text{SINR}_i[n])^{-2}$ for AWGN channel in the high SNR regime. Symbol ϵ represents an acceptable packet error rate. The average rate over the duration αt is: $\bar{R}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} R_i[k]$. For the UAVs in set J:

$$R_j[n] = W_b^j \mathbb{P}(I_b^j) \left(\log_2(1+\text{SINR}_j[n]) - \frac{\log_2(e)}{Q(\epsilon)} \sqrt{\frac{V}{B_j}} \right) \quad (20)$$

The average rate for each UAV in set J to serve b-th SU across the duration $(1-\alpha)t$ is $\bar{R}_j = \frac{1}{n_j} \sum_{k=n_i+1}^{n} R_j[k]$. The combined average rate for all UAVs in set J is $\bar{R}_J = \frac{1}{|J|} \sum_{j \in J} \bar{R}_j$. The average data rate for URLLC over the entire time slot t is $\bar{R} = \alpha \bar{R}_i + (1-\alpha)\bar{R}_J$. Meanwhile, successful CRN transmissions occur primarily under non-interfering with PU conditions, which means no PU activity and no false alarms. This condition is modelled by $\mathbb{P}(H_0)(1-\mathbb{P}_f)$. The problematic scenario of undetected PU activity, $\mathbb{P}(H_1)(1-\mathbb{P}_d)$, is excluded from our analysis due to potential interference. Hence, our rate expression focuses solely on the non-interfering case:

$$\hat{R} = \mathbb{P}(H_0)(1-\mathbb{P}_f)\bar{R}. \quad (21)$$

C. Interference Protection for PU Receivers

PU_T transmissions may not always be detected, necessitating protection for PU_Rs due to their spectrum rights. We consider M number of immobile PU_Rs may encounter interference from UAVs due to wrong CSS outcomes. Therefore, we impose an interference protection to maintain PU_R's QoS as:

$$\mathfrak{I}_m = \mathbb{P}(H_1)(1-\mathbb{P}_d) \sum_{i=1}^{I} \mathbf{g}_m^i P_i, \quad (22)$$

where $\mathbf{g}_m^i$ represents the channel gain which is modeled similarly to UAV-SU channels.

D. Digital Twin and Computation Model

When transferring D_b amount of data to the b-th SU from the i-th UAV, our model calculates the required bandwidth as:

$$W_b^i[n] = \frac{D_b[n]}{\left(\log_2(1+\text{SINR}_i[n]) - \frac{\log_2(e)}{Q(\epsilon)} \sqrt{\frac{V}{B_i}} \right)} \quad (23)$$

However, the DT may not always precisely mimic the real-time channel state, which can result in inaccuracies in the SINR estimation. These inaccuracies, in turn, can affect the optimal bandwidth allocation. We quantify this discrepancy as follows: $\Delta W_b^i[n] = W_b^{i_\text{real}}[n] - W_b^{i_\text{DT}}[n]$. To compensate for this discrepancy and ensure efficient spectrum usage, the DT must adaptively adjust the bandwidth $W_b^{i_\text{DT}}[n]$ based on its predictive capabilities regarding future network demands and real-time feedback. Accordingly, the real data rate $\hat{R}_\text{real}[n]$ and the DT-estimated data rate $\hat{R}_\text{DT}[n]$ is used to adjust the anticipated latency time to serve the b-th Secondary User (SU). Thus, the total latency, including the adjustments made by the DT's estimations, is given by:

$$T_b[n] = \frac{D_b[n]}{\hat{R}_\text{real}[n]} + \frac{D_b[n]\hat{R}_\text{DT}[n]}{\hat{R}_\text{real}[n]\left(\hat{R}_\text{real}[n] - \hat{R}_\text{DT}[n]\right)} \quad (24)$$

E. Formulation of the Objective Function

Our primary goal is to minimize the transmission latency for URLLC-enabled SUs. To achieve this objective while considering system limitations and ensuring quality of service (QoS), the problem is mathematically formulated as follows:

$$\begin{aligned}
&\min_{\{W_b^i[n], P_i, \tilde{T}, \bar{T}, \mathbb{P}_d, \mathbb{P}_{fa}\}} T_b[n] \quad (25)\\
\text{s.t: } & C_1 : W_b^i[n] \leqslant W_i^\text{uav}, \textstyle\sum_i W_i^\text{uav} \leqslant W_\text{max}\ , \forall\{b,i\},\\
& C_2 : \mathfrak{I}_m \leqslant \mathfrak{I}_\text{thsld},\ \forall m,\\
& C_3 : \hat{R}_\text{real}[n] \geqslant R_\text{thsld}, \quad \hat{R}_\text{DT}[n] \geqslant R_\text{thsld},\\
& C_4 : T_b[n] \leqslant T_\text{thsld},\\
& C_5 : \text{SINR}_i[n] \geqslant 0,\\
& C_6 : \mathbb{P}_d \geqslant \mathbb{P}_d^\text{thsld}, \quad \mathbb{P}_{fa} \leqslant \mathbb{P}_{fa}^\text{thsld},\\
& C_7 : P_i \leqslant P_\text{max}, \forall i,\\
& C_8 : \tilde{T} + \bar{T} \leqslant T,
\end{aligned}$$

where the constraint C_1 enforces the bandwidth constraint. C_2 protects the PU_Rs' QoS by limiting the unwanted interference. C_3 ensures that the real and estimated data rates. C_4 ensures the total transmission latency doesn't surpass the acceptable limit. C_5 maintains the SINR quality. C_6 assures the CSS reliability. C_7 defines the UAV's maximum transmission power budget, and C_8 restricts the total time requirements for CSS ($\tilde{T}$) and CRN's communications ($\bar{T}$).

IV. DRL-based Solution with DT Enhancement

We use a DRL approach augmented with a DT utilizing a proposed MPPO algorithm. Our proposed MPPO is the advancement of the traditional PPO [6], [8]. Including the DT component helps estimate future traffic demands and channel conditions to inform the DRL agent, enabling proactive and adaptive strategy formulation for resource allocation, thus reducing latency and enhancing the reliability of services.

1) State Space with DT Predictions: We define the state space, $\mathcal{S}$, to integrate the real-time and predicted states provided by the DT. Each state $\mathbf{s}_t \in \mathcal{S}$ for time step t is expressed as a combination of the current observed state $\mathbf{s}_t^\text{obs}$ and the DT predicted state $\mathbf{s}_t^\text{DT}$, such that $\mathbf{s}_t = \{\mathbf{s}_t^\text{obs} \cup \mathbf{s}_t^\text{DT}\}$. The observed state at time t is given by $\mathbf{s}_t^\text{obs} = \{\mathbf{h}_i[n], \mathbf{g}_i^b[n], \mathbf{g}_m^i[n], \Phi_\text{global}[n], \mathbb{P}_d^\text{thsld}, \mathbb{P}_{fa}^\text{thsld}, P_\text{max}, W_i^\text{uav}, R_\text{thsld}, \mathfrak{I}_\text{thsld}, T_\text{thsld}, D_b[n]\}$, and the DT predicted state is $\mathbf{s}_t^\text{DT} = \Psi(\mathbf{s}_{t-1}, \mathbf{a}_{t-1}; \Theta_\text{DT})$, where Ψ is the DT's predictive function (i.e., $\Psi : \mathcal{S} \times \mathcal{A} \times \Theta \rightarrow \mathcal{S}$) parameterized by Θ_DT, and $\mathbf{a}_{t-1}$ is the action taken at the previous time step.

2) Adaptive Action Space: The DRL agent's action space, $\mathcal{A}$, is designed to be responsive to both the immediate and predicted state space. Actions $\mathbf{a}_t \in \mathcal{A}$ at each decision epoch are chosen considering the predictive insights from the DT:

$$\mathbf{a}_t = \{W_b^i[n], P_i[n], \tilde{T}, \bar{T}, \mathbb{P}_d[n], \mathbb{P}_{fa}[n]\}$$

A. Predictive Reward Function

The reward function $\mathscr{R}(\mathbf{s}_t, \mathbf{a}_t)$, crucial in the DRL agent's learning, is now influenced by both the immediate outcomes and predictive assessments from the DT:

$$\mathscr{R}(\mathbf{s}_t, \mathbf{a}_t) = \alpha_1 \Delta T_b[n] - \alpha_2 \sum^{J} \mathscr{I}(\mathrm{C}_j) + \alpha_3 \mathscr{R}_{\mathrm{DT}}(\mathbf{s}_t^{\mathrm{DT}}, \mathbf{a}_t) \quad (26)$$

where $\Delta T_b[n] = (T_{\mathrm{thsld}} - T_b[n])$, $\mathscr{R}_{\mathrm{DT}}$ represents the reward component based on DT's predictions, and α_3 adjusts the impact of the DT's predictive accuracy on the reward function. The function $\mathscr{I}(\mathrm{C}_j)$ is a binary indicator yielding 1 if the j-th constraint is violated and 0 otherwise.

B. Design of the Proposed MPPO-based DRL Algorithm

The actions $\mathbf{a}_t \in \mathcal{A}$, derived from the policy π_θ, parameterized by θ, are evaluated for their probabilities $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$. The main objective in MPPO is to optimize θ to maximize expected rewards while adhering to the strict constraints of URLLC. The gradient of the expected reward is given by:

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \hat{A}_t\right], \quad (27)$$

where $\hat{A}_t$ is an estimator that quantifies the advantage of taking action $\mathbf{a}_t$ in state $\mathbf{s}_t$. The MPPO objective function, accommodating these adaptations, is formulated as follows:

$$L^{\mathrm{MPPO}}(\theta) = \mathbb{E}\Bigg[\min\left(\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_t|\mathbf{s}_t)} \hat{A}_t, \mathrm{Clip}\left(\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_t|\mathbf{s}_t)}, \eth\right) \hat{A}_t\right) - \beta C(\mathbf{s}_t, \mathbf{a}_t)\Bigg] + \lambda_p \cdot \mathrm{KL}\left[\pi_{\theta_{\mathrm{old}}}(\cdot|\mathbf{s}_t) \| \pi_\theta(\cdot|\mathbf{s}_t)\right], \quad (28)$$

where $C(\mathbf{s}_t, \mathbf{a}_t)$ represents the cost of constraint violations, $\eth$ is the adaptive clipping parameter, and λ_p is the penalty coefficient for the KL divergence term [6]. These strategic adaptations ensure that MPPO is uniquely equipped to handle the complex requirements of URLLC in CRNs, making it a significantly more capable algorithm than standard PPO [6] in this context. By using this adapted objective, MPPO optimizes policy for resource distribution as outlined in **Algorithm 1**.

C. Computational Complexity Analysis of MPPO Algorithm

Denoting the number of matrix multiplications within the policy network's forward propagation as N_m, we express the complexity as $O\left(\mathcal{A} \times N_m \times N_l^{2.81}\right)$, where N_l is the neural network's layer count. The gradient complexity for the log probability of the policy network equates to $O\left(\mathcal{A} \times N_m \times N_l^3\right)$. Implementing variance reduction measures with overhead V_r yields a complexity of $O\left(V_r \times \mathcal{S} \times \mathcal{A} \times N_l^{2.81}\right)$. The environmental interactions, noted as C_e, contribute a complexity of $O\left(C_e + \frac{T \times \mathcal{S} \times \mathcal{A}}{\text{batch size}}\right)$, where T is the temporal scale of interaction.

Algorithm 1 Modified PPO algorithm for URLLC in CRNs.

Require: Initial policy parameters θ, value function parameters ϕ, learning rate ξ, adaptive clipping parameter $\eth$, penalty coefficient β, KL divergence coefficient λ_p, number of iterations N, convergence threshold δ_M, parameter convergence threshold δ_θ.

Initialize performance metric list $M = [\ ]$
for $n = 1$ to N **do**
  $\mathcal{T} \leftarrow \{\tau_i\}_{i=1}^m$ where $\tau_i = \{(\mathbf{s}_t, \mathbf{a}_t, \mathscr{R}(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}_{t=1}^{T_i}$ are trajectories collected under policy $\pi_{\theta_{\mathrm{old}}}$.
  for each trajectory $\tau_i \in \mathcal{T}$ **do**
    Calculate $\delta_t = \mathscr{R}(\mathbf{s}_t, \mathbf{a}_t) + \gamma V_\phi(\mathbf{s}_{t+1}) - V_\phi(\mathbf{s}_t)$
    Compute advantage $\hat{A}_t = \delta_t + (\gamma \lambda_p) \hat{A}_{t+1}$
    Assess constraint $C(\mathbf{s}_t, \mathbf{a}_t) = \max\{0, g(\mathbf{s}_t, \mathbf{a}_t) - c\}$
  end for
  Update θ by applying the MPPO gradient, incorporating KL divergence:

$$\nabla_\theta L^{\mathrm{MPPO}}(\theta) = \hat{\mathbb{E}}_t\Bigg[\min\Bigg(\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_t|\mathbf{s}_t)} \hat{A}_t, \mathrm{Clip}\left(\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_t|\mathbf{s}_t)}, 1-\eth, 1+\eth\right) \hat{A}_t\Bigg) - \beta C(\mathbf{s}_t, \mathbf{a}_t)\Bigg] + \lambda_p \cdot \mathrm{KL}\left[\pi_{\theta_{\mathrm{old}}}(\cdot|\mathbf{s}_t) \| \pi_\theta(\cdot|\mathbf{s}_t)\right]$$

  Update ϕ by minimizing the value function loss:

$$L_{\mathrm{VF}}^{\mathrm{MPPO}}(\phi) = \left(V_\phi(\mathbf{s}_t) - R_t\right)^2 \text{ where } R_t = \sum_{k=0}^{\infty} \gamma^k \mathscr{R}(\mathbf{s}_{t+k}, \mathbf{a}_{t+k})$$

  Dynamically adjust β to maintain constraint satisfaction:

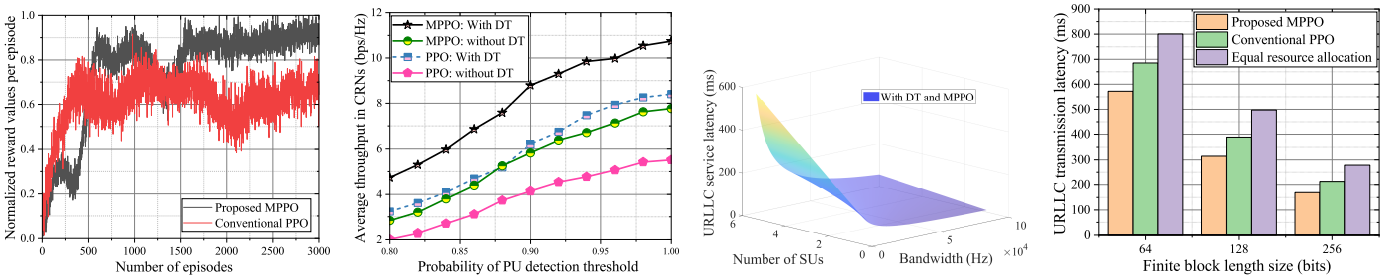
$$\beta \leftarrow \beta + \alpha_\beta \cdot \left(\hat{\mathbb{E}}_t[C(\mathbf{s}_t, \mathbf{a}_t)] - d\right)$$

  $M \leftarrow \mathrm{RecordMetric}(\pi_\theta)$, for instance, $M \leftarrow \hat{\mathbb{E}}_t[\mathscr{R}(\mathbf{s}_t, \mathbf{a}_t)]$.
  if $|M_n - M_{n-1}| < \delta_M$ **then** Break {Convergence based on M}
  if $\|\theta_n - \theta_{n-1}\|_2 < \delta_\theta$ **then** Break {Convergence based on θ}
end for
return π_θ, V_ϕ

V. Numerical Results and Analysis

This section presents a comprehensive evaluation of the proposed MPPO algorithm developed for URLLC services in CRNs. The parameters used in the simulation are as follows: number of UAVs $I = 3$, number of antennas in each UAU $K = 4$, number of URLLC service enabled SUs $B = 3$, distances (d_i) from $\mathrm{PU_T}$ to UAVs $(100, 5000)$ meters, the side length of squared service grid area $\ell = 200$ meter, UAV's service radius (r) is 50 meter, $f_c = 3$ GHz, $P_p = 0.1$ watt, $\mathbb{P}(H_0) = 0.7$, $\mathbb{P}(H_0) = 0.3$, $T = 200$ ms, $N = 200$, $W_{\max} = 30$ KHz, $P_{\max} = 1$ watt, noise $\sigma_b^2 = -120$ dBm, data segment size $D_b = 10$ Kb, $\epsilon = 10^{-5}$, $B_i = 256$ bits, $\mathbb{P}_d^{\mathrm{thsld}} = 0.95$, $\mathbb{P}_{fa}^{\mathrm{thsld}} = 0.05$, $R_{\mathrm{thsld}} = 1$ mbps, UAVs' mean velocity $\widehat{\mathbb{V}} = 40$ Km/h, SUs' mean velocity $v_b = 60$ Km/h, $\mathfrak{I}_{\mathrm{thsld}} = -120$ dBm, $T_{\mathrm{thsld}} = 100$ ms. Unless we specify a new value, the aforementioned values remain constant during simulations.

In our MPPO algorithm, we set ξ as 10^{-4}, while the penalty coefficient β starts at 0.01. The MPPO iterates $N = 3000$ times, with convergence thresholds δ_M and δ_θ set to 0.01 and 10^{-5}, respectively. We use a discount factor γ of 0.99 and a generalized advantage estimate (GAE) parameter λ_p of 0.95, with a clipping parameter $\eth$ of 0.2 and update rate α_β of 10^{-3}. The policy and value networks used three hidden layers, each comprising 128 neurons employing ReLU activation functions. The output layer utilizes a softmax activation for discrete actions and a linear activation for value predictions. These



(a) DRL convergence curve. (b) Analysis of DT impact. (c) Latency vs $W_{\max}$ and users. (d) Analysis of FBL in latency.

Fig. 2: An analysis of various performance metrics for the proposed work.

networks are optimized using the Adam optimizer.

Fig. 2a depicts the convergence of our MPPO algorithm, plotting episode count against normalized rewards from 0 to 1. The MPPO achieves quicker convergence by the 2000-episode mark and yields 34.64% higher rewards than conventional PPO by the 3000-th episode, demonstrating its efficiency. This superiority is attributed to MPPO's refined policy updates compared to PPO, potentially enabling more efficient exploration and exploitation of the learning environment.

Fig. 2b delineates the throughput efficacy of MPPO relative to the conventional PPO in the present interweave CRNs, utilizing DT technology across varying PU detection thresholds ($\mathbb{P}_d^{\text{thsld}}$). The data affirm that MPPO transcends PPO in throughput across the spectrum of $\mathbb{P}_d^{\text{thsld}}$ values. The incorporation of DT notably strengthens both algorithms, with MPPO achieving a marked throughput augmentation of 27.89% over PPO when $\mathbb{P}_d^{\text{thsld}} = 1$, underpinning the utilization of DT. Further analysis reveals that MPPO with DT increases the average throughput in CRNs by 38.39% compared to without involvement of DT technology. This enhancement is attributable to DT's capacity for real-time network mirroring and historical data analysis, facilitating refined resource allocation decisions conducive to optimized spectrum use and elevated throughput in CRNs.

Fig. 2c presents the latency behavior as a function of the maximum available bandwidth ($W_{\max}$) and the number of SUs under the application of DT and MPPO. The 3D surface plot shows that as the bandwidth allocation per user increases, there is a significant decrease in URLLC service latency, underscoring the bandwidth's impact on achieving lower latency. Notably, even with more SUs, the application of DT and MPPO maintains the latency below the 100 ms threshold when adequate bandwidth is provided.

Fig. 2d presents an analysis of URLLC transmission latency with respect to the FBL size, comparing the performance of the proposed MPPO algorithm, conventional PPO, and a baseline strategy employing equal resource allocation with DT. The graph indicates that the latency decreases for all strategies as the FBL size increases, which is expected since larger block sizes generally allow more efficient encoding schemes, reducing the time for successful transmission. Notably, the proposed MPPO with DT consistently achieves lower latencies across all block lengths than the conventional PPO with DT and the equal resource allocation strategy with DT. At an FBL of 256 bits, the MPPO with DT achieves a latency reduction of approximately 19.98% over PPO and approximately 38.94% over the equal resource allocation with DT.

VI. Conclusions

This work developed a UAV-assisted resource allocation framework for URLLC services in interweave CRNs, utilizing an MPPO-based DRL strategy for minimizing transmission latency. As confirmed by our simulations, integrating DT technology with the proposed MPPO exhibited superior latency reduction and throughput improvement compared to conventional PPO and equal allocation methods. At a block length of 256 bits, MPPO with DT achieved a latency reduction of 19.98% over PPO and 38.94% over equal allocation with DT. Furthermore, with a PU detection threshold set to 1, MPPO realized a throughput increase of 27.89% compared to conventional PPO. These findings highlight the benefits of incorporating DT into advanced DRL algorithms, suggesting substantial enhancements in CRN performance and the fulfillment of URLLC requirements.

References

[1] N. T. V. Khanh and T.-T. Nguyen, "Joint design of beamforming and antenna selection in short blocklength regime for URLLC in cognitive radio networks," *IEEE Access*, vol. 9, pp. 144 676–144 686, 2021.

[2] Q. Huang *et al.*, "Machine-learning-based cognitive spectrum assignment for 5G URLLC applications," *IEEE Netw.*, vol. 33, no. 4, pp. 30–35, 2019.

[3] P. K. Sharma *et al.*, "Cognitive D2D finite blocklength transmissions with the presence of time-selective interference," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 12 215–12 219, 2021.

[4] Z. Chu *et al.*, "Opportunistic spectrum sharing for D2D-based URLLC," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8995–9006, 2019.

[5] B. Li *et al.*, "FlexEdge: Digital twin-enabled task offloading for UAV-aided vehicular edge computing," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 11 086–11 091, 2023.

[6] W. Yu *et al.*, "Asynchronous hybrid reinforcement learning for latency and reliability optimization in the metaverse over wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2138–2157, 2023.

[7] A. Paul and S. P. Maity, "Outage analysis in cognitive radio networks with energy harvesting and Q-routing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6755–6765, 2020.

[8] K. K. Nguyen *et al.*, "RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 1086–1096, 2022.

[9] Y. Lu *et al.*, "Adaptive edge association for wireless digital twin networks in 6G," *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16 219–16 230, 2021.

[10] A. Paul *et al.*, "Spectrum sensing in cognitive vehicular networks for uniform mobility model," *IET Commun.*, vol. 13, no. 19, pp. 3127–3134, 2019.