IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 6, 2021, accepted October 20, 2021, date of publication October 26, 2021, date of current version November 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123213

Packet Delivery Maximization Using Deep Reinforcement Learning-Based Transmission Scheduling for Industrial Cognitive Radio Systems

PHAM DUY THANH[1], TRAN NHUT KHAI HOAN[2], HOANG THI HUONG GIANG[1], AND INSOO KOO[1]

[1]Department of Electrical, Electronic and Computer Engineering, University of Ulsan (UOU), Ulsan 44610, South Korea
[2]Department of Electronics and Telecommunication Engineering, College of Engineering Technology, Can Tho University, Can Tho 900000, Vietnam

Corresponding author: Insoo Koo (iskoo@ulsan.ac.kr)

This work was supported in part by the National Research Foundation of Korea through the Korean Government Ministry of Science and ICT (MSIT) under Grant 2021R1A2B5B01001721.

ABSTRACT The performance of data aggregation in industrial wireless communications can be degraded by environmental interference on Industrial Scientific Medical (ISM) channels. In this paper, cognitive radio (CR) was applied to enable devices to share primary channels with the aim of enhancing the transmission performance of the WirelessHART network. We considered a linear convergecast system, where the packets generated at each device were routed to the gateway (GW) through the aid of neighboring devices. The solar-powered cognitive access points (CAPs) were deployed to improve the network performance by opportunistically allocating the primary channels to the devices for data transmissions. Firstly, we formulate the scheduling problem of long-term throughput maximization as a framework of a Markov decision process with the constraints of the minimum delay, the number of required ISM channels, and the harvested energy at the CAPs. Then, we propose a deep reinforcement learning-based scheduling scheme to optimally assign multiple ISM and primary channels to the field devices in each superframe. The simulation results confirmed the superiority of the proposed scheme compared to existing methods.

INDEX TERMS WirelessHART, cognitive radio, Markov decision process, industrial scientific medical.

I. INTRODUCTION

Wireless technologies have been considered a promising alternative for automotive control systems, industrial and factory automation, and other interconnected embedded systems [1], [2]. They offer several advantages over traditional wired communication systems, such as fewer infrastructure requirements, reduced connector trouble, and simplicity for future upgrading [3], [4]. On the other hand, there have been concerns regarding the network latency and reliability, which hampered the deployment rate owing to the stringent communication requirements in industrial control applications. Thus, the control performance might be deteriorated significantly because of increasing latency, jitter, and packet loss rate.

WirelessHART [5], the first open wireless communication standard designed for industrial process monitoring was introduced to address these issues. In particular, WirelessHART uses a tightly integrated medium access and networking layer for multi-hop multipath routing based on multi-channel TDMA, in which centralized resource allocation is implemented to guarantee network performance. The WirelessHART architecture was developed for the wireless mesh networking protocol by leveraging time diversity, path diversity and frequency diversity to support advanced process monitoring and control applications. WirelessHART networks require multiple sensor nodes to report data of their measurements to the controller periodically for supervisory control. Aggregating data from multiple sources to a single destination is a many-to-one transmission paradigm whose corresponding networking primitive is called convergecast. The major difference between industrial WirelessHART networks and wireless sensor networks lies in the characteristics of flows. In wireless sensor networks, the traffics are usually generated with a random or unpredictable generating time, which leads to the challenge of capturing the delay of each specific data packet. In many industrial applications, data are

The associate editor coordinating the review of this manuscript and approving it for publication was Ghufran Ahmed.

generated periodically [6]. For instance, in industrial wireless monitoring scenarios, sensors are usually configured with certain sampling periods to periodically perform measurements of some external signals to report data to processing and decisioning units. In such cases, the packets delivered in each discrete-time period are given. This feature can help the network manager to make fine-grained scheduling decisions for the network performance optimization in terms of delay and spectrum efficiency.

In CSMA-based ISM band protocols, the devices verify the absence of other traffic before transmitting on a shared transmission medium to avoid the collisions with other devices. However, the ISM bands are only narrow portions of the frequency spectrum reserved internationally for industrial, scientific and medical purposes. Therefore, equipments operating in ISM bands have to tolerate high interference generated by other ISM applications. There is no regulatory protection from ISM devices operating in the ISM bands. Furthermore, in recent years, the increasing use of microelectronics devices as well as the attraction of unlicensed use has been leading to an overload in ISM bands [7].

Meanwhile, according to a Federal Communications Commission spectrum policy task force report [8], utilization of the licensed spectrum varies between 15% and 80%. A new communication paradigm, i.e., dynamic spectrum access whose key enabling technology is referred to as Cognitive Radio (CR), was recently proposed to tackle spectrum inefficiency issues. CR is a form of wireless communication in which a radio can sense the surrounding environment and automatically alter its characteristics such as power, frequency, modulation, and other operating parameters to dynamically reuse whatever spectrum is available. On the other hand, CR is regarded as a promising technology to improve the spectrum utilization of wireless users via heterogeneous wireless sensor networks by enabling secondary users to share the spectrum with primary users [9]. Variant functionalities of CR including spectrum sensing, spectrum management, spectrum sharing, and spectrum mobility, have been well investigated in literature [10]. CR technology was applied to enhance the reliability in wireless industrial networks [11]–[13]. Particularly, devices can detect and avoid interference by integrating CR principles into the lower layers of the industrial wireless sensor networks, which opens the possibility of utilizing additional radio spectrum channels.

To reduce the overload in ISM bands as well as obtain the better licensed spectrum utilization, a CR-based scheme [14] was proposed by utilizing vacant licensed frequency for data transmissions. The licensed frequency blocks are regarded as primary channels such as television broadcast, digital television broadcast bands [15], or cellular frequency bands [16]. However, to access a licensed frequency channel, CR network must ensure that its transmission does not impact on the quality of service (QoS) of licensed network. For instance, according to IEEE 802.22, the acceptable probability of interference with the primary networks should be less than 0.1 [17]. The reliability of using licensed bands is highly dependent on the spectrum sensing, and the secondary users decide their transmissions according to the sensing results. Thus, designing a scheme of switching between ISM and licensed channels for wireless devices in industrial networks are needed to be intensively investigated, such that the transmissions are assigned with higher reliability among the ISM/licensed channels.

A. RELATED WORKS

Several study efforts have focused on multi-channel convergecast protocols [18], [19]. Zhang *et al.* [18] proposed joint link scheduling and channel assignment approaches for both cases of single-packet buffering and multiple-packet buffering constraints in a linear convergecast topology. The latency-optimal link scheduling problem was investigated for a tree-routing topology with and without a restriction on the number of channels [19]. Although the solutions proposed in these studies can optimize the latency and channels in the convergecast operation, the system performance is still degraded remarkably by interference, such as noise or other devices that affect the connectivity and induce low reliability on the ISM channels. Some techniques have been directly applied to improve the convergecast reliability, such as allowing retransmissions [20], [21] or constructing multiple routing choices [22]. Nevertheless, these methods might only enhance the convergecast reliability to some extent but generally cannot maximize the reliability under stringent latency constraints.

In addition, Yunhuan *et al.* [23] studied the cognitive radio-based interference tackling scheme to obtain the best available channel set for direct sequence spread spectrum/channel hopping transmission link. Lyu *et al.* [24] proposed a redundant transmission approach in industrial cyber-physical systems by exploring spectrum opportunities in licensed channels to guarantee transmission reliability for state estimation. With the advancement of artificial intelligence (AI) algorithms, especially deep reinforcement learning (DRL), several studies to obtain efficient resource scheduling in industrial scenarios have been proposed [25], [26]. Specifically, the authors in [25] proposed a green resource allocation framework for the industrial internet of things under 5G heterogeneous networks, while the reinforcement learning schemes were developed in [26] to maintain the aggregated interference from both upstream and downstream transmissions to the desired value. As a result, these DRL-based schemes are proven to efficiently deal with the dynamics of environments. Furthermore, high-dimensional problems in practical scenarios can be solved by using DRL, which might be a big challenge for conventional reinforcement learning. Among the aforementioned literature, most studies focused on designing a transmission schedule for devices to either increase the transmission reliability of devices or optimize channels and latency in the convergecast operation. On the other hand, only a few studies examined ways of improving the resistance of the WirelessHART convergecast network to interference by integrating CR principles into a WirelessHART protocol standard [23], [24].

TABLE 1. Table of literature summary.

References	Advances/Key contributions	Loopholes/Drawbacks
[18], [19]	Optimal link and latency scheduling was obtained for the convergecast system	Dynamic spectrum access in cognitive radio and energy harvesting were not considered
[20]–[22]	The reliability can be obtained by applying retransmissions or routing protocols	The network reliability might not be optimized in the stringent latency constraints, and energy harvesting was not considered
[11]–[13], [23], [24]	The network reliability was enhanced by exploring spectrum opportunities in licensed bands	Spectrum sensing imperfection and energy harvesting were not taken into account
[26]	Reinforcement learning algorithms were designed to manage the aggregated interference generated by multiple wireless regional area networks	The Q-learning scheme requires high memory requirement for storing the lookup table with high dimensional problems. Furthermore, the energy harvesting was not considered.

Table 1 shows the literature summary for industrial wireless sensor network.

Along with challenges in spectrum management, energy-efficient utilization is one of the main concerns in wireless communications. Energy harvesting can ensure energy autonomy by renowned renewable energy, such as radio frequency power [27] and solar power [28], [29], to recharge the limited-capacity battery of the devices. Among the different types of existing renewable energy, solar power, which is harvested directly from sunlight, is considered the most effective energy resources, even though the density of solar energy is strongly dependent on the environmental conditions. Therefore, the efficient utilization of solar energy harvested from ambient environment needs to be investigated intensively to improve the performance of the WirelessHART convergecast system. Nevertheless, the energy harvesting distribution is difficult to obtain in practice to devise an energy-efficient approach for devices in industrial cognitive radio networks. Consequently, designing the transmission schedule to enhance the performance of energy harvesting-powered WirelessHART convergecast systems with an unknown distribution of energy arrivals is the primary motivation of this paper.

B. MAIN CONTRIBUTIONS

To the best of the authors' knowledge, this paper is the first attempt to formulate the joint ISM/primary channels schedule for the transmission of the devices in the WirelessHART linear convergecast network with solar energy harvesting. Specifically, we focused on the joint ISM/primary channel allocation scheme for a linear convergecast system, in which the primary channels are exploited opportunistically to improve the long-term throughput considering the interference on ISM channels. Moreover, by taking the limited ISM channels and dynamics of primary channels into account, this study developed a deep reinforcement learning-based scheduling scheme to efficiently schedule the transmission of devices under the constraints of harvested energy, buffering capacity, minimum latency, and the number of required ISM channels. The main contributions of this paper can be summarized as follows:

- We first investigate an energy-harvesting linear convergecast model that contains field devices with sensing data needed to send to the GW. The solar-powered cognitive access points (CAPs) are deployed to determine the availability of the primary channels. The constraints of single-buffer capability in devices, the limited energy harvesting in CAPs, minimum latency, and the number of required ISM channels for the scheduling are considered.
- Secondly, long-term throughput maximization is formulated as a framework of the Markov decision process (MDP). Subsequently, the deep Q-learning scheduling scheme is proposed to achieve an optimal policy of the MDP problem. Thereby, the agent (i.e., the GW) can interact directly with the environment and learn the optimal scheduling via trial-and-error. As a result, the field devices can be scheduled with the proper ISM/primary channels and time slots through each superframe by using the proposed approach.

The remainder of this paper is organized as follows. Section II presents the network model. Next, we present the joint ISM channel, device and data flow scheduling in Section III. Section IV outlines the proposed deep reinforcement learning approach. Subsequently, the joint time and ISM/primary channel scheduling and sub-schedule extraction are given in Section V. We discuss simulation results in Section VI. Finally, this work is concluded in Section VII.

II. NETWORK MODEL

A. BRIEF OVERVIEW OF WirelessHART

A WirelessHART system contains the following basic components: (a) field devices connected to process equipment; (b) gateways that are responsible for communication between field devices and host applications; and (c) a network manager which provides network configuration, system health monitoring, routing table managing and communication scheduling for all nodes. WirelessHART is a complete wireless mesh networking protocol based on low-power radios using the IEEE 802.15.4-2006 standard that supports 16 channels in the 2.4 GHz license-free ISM band with the total data rate of up to 250 kbits/s. To minimize the influence of noise in the channels with high interference levels (e.g. due to the coexistence with IEEE 802.11), channel blacklisting is utilized by considering the wireless channel

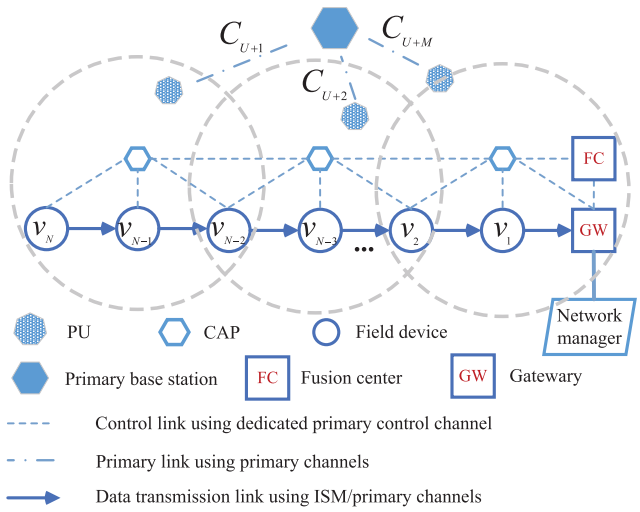

FIGURE 1. Linear convergecast system model.

quality [30]–[32]. The MAC layer enables channel-hopping at each slot boundary to arbitrate and coordinate network communications. The transmission is based on the TDMA protocol where each time slot has a fixed duration of 10 ms for transmitting a small packet with a maximum size of 133 bytes and receiving an associated acknowledgment. To appropriately establish the global transmission schedule, WirelessHART supports multiple superframes for data communications. A superframe is comprised of multiple time slots, where the network manager determines the number of time slots. In a superframe, each time slot can be assigned one or more links [33], [34].

B. COGNITIVE RADIO-ASSISTED LINEAR CONVERGECAST MODEL

Figure 1 presents the considered linear convergecast network. The industrial wirelessHART topology was modeled as a graph $G = (V, E)$, in which vertices in $V = \{v_o, v_1, \ldots, v_N\}$ denotes the network devices, and the edges in E represent communication links (device pairs). There is a set of N field devices, denoted by $\mathbb{N} = \{v_1, v_2, \ldots, v_N\}$, in the network and a gateway (GW) denoted by v_o. In this paper, the field devices and the GW are powered by grid energy. For simplicity, the terms "device" and "field device" are used interchangeably throughout this paper. The TDMA transmission protocol was adopted, in which time is synchronized and slotted with the standard duration of 10 ms, enabling exactly one packet transmission and its corresponding acknowledgement. In the linear convergecast network, each field device generates one data packet at the beginning of a convergecast operation (i.e. at the start of each superframe) and transmits it to the GW. This kind of convergecast is used for periodic data collection in WirelessHART. Each device has a single-packet buffering capacity. The field device has a half-duplex capability from which it can either transmit or receive a packet at a time slot. Furthermore, each device is only scheduled on one channel at a given time slot. Channel hopping is carried out in a time slot basis and parallel transmissions can be scheduled concurrently in different channels.

In this article, we consider the interference constraint of the ISM channels on each link (e.g., the interference levels with other devices using IEEE 802.11 standard). For the GW to receive a data flow from a device, it must be successfully transmitted via all links routed to the GW. Hence, the successful transmission rate of each link should be enhanced such that the GW can obtain as many packets as possible through each superframe. For this reason, the CR technique is exploited such that the devices can opportunistically switch to primary channels to achieve more reliable data transmissions because each device-to-device link can be assigned additional primary channels to improve the transmission reliability. In addition, it was assumed that the signaling information among the GW, CAPs, and devices could be exchanged securely and reliably using a common control channel made up of one dedicated primary channel or a couple of dedicated primary channels depending on the network implementation. This work does not focus on guaranteeing the common control channels for CRNs because they have been well-studied [35], [36]. In the present study, the terms "common control channel" and "dedicated primary control channel" are used interchangeably.

According to the IEEE 802.15.4-2006 standard, the 2.4 GHz license-free ISM band is divided into a set of U ($U = 16$) ISM channels, denoted by $\mathbb{U} = \{C_1, \ldots, C_u, \ldots . C_U\}$ where C_u represents the ISM channel u. We consider a set of M primary channels, denoted by $\mathbb{M} = \{C_{U+1}, \ldots, C_{U+m}, \ldots, C_{U+M}\}$, where C_{U+m} is the primary channel m, as shown in Figure 1. A primary network comprises a primary base station (PBS) and multiple primary users (PUs). PBS and PUs have the licensed right to utilize M primary channels while the devices can opportunistically share the primary channels to transmit their packets. We assume that K CAPs are sequentially connected to the GW by which one CAP is connected to the GW and each CAP is connected to its next CAP leading towards the GW. The communication on this connection can be done via a dedicated primary control channel, such that the sensing results and signaling information can be shared among CAPs and the GW.

In our industrial network setting, the CAPs are deployed to assist the utilization of primary channels for the transmissions of the devices. However, advancement in recent wireless technology has triggered the device demand of running on independent fixed power sources, and green communication becomes the utmost importance nowadays. This can be accomplished via harvesting energy from the surrounding environment such as solar. Moreover, providing wired power in an industrial wireless network may be hard in some circumstances or even extremely costly when retrofitting extra devices in buildings. Energy-harvesting can offer power autonomy to wireless devices, which can bring simpler deployment and long-term energy supply. Unfortunately, with solar energy harvesting, the energy arrival can be

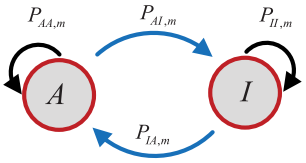

FIGURE 2. Activity model of primary channel *m*.

substantially affected by the ambient practical environment, thus the randomness of energy arrival is taken into account when developing a resource allocation scheme. Although the system suffers from the uncertainty of the harvested energy, the benefit of solar-powered wireless networks has been well investigated to improve the network performance [37]. Furthermore, solar cells are proven to perform well under indoor light conditions for these applications [38]. Therefore, this work is designed for indoor industrial sensor networks, where the illuminance level might be low and solar energy is hard to collect than outside environments. To facilitate the harvesting, the energy-harvesting devices (i.e., Access points) can be placed in high light-intensity locations (e.g., the roof-tops of the building) for higher harvesting capacity. For the above reasons, allowing opportunistically use of primary channels with the help of the solar-powered CAPs can increase the number of packets received by the GW. Therefore, it results in the higher reliability compared to only-ISM-band utilization where the ISM channels have a low transmission quality.

In this paper, CAPs are assumed to be placed in the roof-tops of the buildings so they can harvest solar energy for their operation while the devices and the GW are powered by grid energy. Each CAP is used to supervise the region of the groups of K_s field devices. In addition, CAPs can also perform the cooperative spectrum sensing on several assigned primary channels at the beginning of each superframe to check whether the primary channels are free or not. Subsequently, local sensing results at CAPs are sent to a fusion center, denoted by FC, where the global sensing results are determined and sent back to the CAPs to broadcast them to the devices through the control channel. In this paper, the network manager was assumed to be integrated into the GW. Hence, the GW is responsible for scheduling all devices in each superframe.

C. ACTIVITY MODEL OF MULTIPLE PRIMARY CHANNELS

In practice, modeling the activity of PUs on each primary channel is sophisticated and may not properly be investigated because the secondary system is working independently from the primary system. Furthermore, it is challenging to obtain prior knowledge of the existence of ambient primary users in the network. Therefore, instead of investigating PUs' activities on a primary channel, researches in the literature have been focusing on modeling primary channel behavior (i.e. busy when the channel is occupied by PUs or free when the channel is vacant) [39]. Besides, the primary channel behavior is often considered as a 2-state Discrete-Time Markov Chain process, in which the transition probabilities between two states (busy or free) were well-studied in the literature [40]. Thereby, the activities of multiple primary users were already represented in the primary channel behavior. Hence, we do not investigate the impact of the activity of the PUs as well as the number of PUs on the network performance. Instead, we focus on the spectrum occupation properties of the PUs on the licensed channels by modeling the primary channel behavior as a 2-state discrete-Time Markov Chain process, in which that state of the primary channel in a superframe can be represented by "busy" or "free." Thereby, these states can indicate the occupation of the PUs on the primary channels in each superframe. Thus, by using the transition probabilities of the Markov Chain process, the GW can update the probabilities that the primary channels are free in each superframe to assign the proper primary channels to the devices, which results in the performance improvement of the WirelessHART network.

We consider the cognitive system with M uncorrelated primary channels. The state of each primary channel (either vacant or occupied by the primary users) is assumed to be unchanged within a cognitive frame and can be changed between two consecutive cognitive frames. In this paper, it is supposed that the cognitive frame has the same length as each superframe. In the superframe, the state of the primary channel is denoted as either A or I, which represent the hypothesis that the primary channel is "active" (i.e., busy) or "inactive" (i.e., free), respectively. Furthermore, this paper assumes that the state transition probability of each primary channel between two adjacent cognitive frames follows a discrete-time Markov chain model, as depicted in Figure 2. $P_{xy,m}\,|x, y \in \{A, I\}$ refers to the state transition probability of primary channel m from state x in a current frame to state y in the next frame. The interference with the primary users may happen according to the dynamic behavior of the licensed channels. We assume that a packet is dropped when transmitted on a primary channel if and only if it is actually busy (i.e., the transmission collision between the devices and primary network occurs). Furthermore, we do not focus on the channel switching delay, which was well investigated in [41].

D. SENSING IMPERFECTION

In this paper, the sensing error of the CAPs was taken into account. At the start of a superframe, the CAPs perform the cooperative spectrum sensing on the primary channels assigned by the GW and then send the local results to the FC to make a global decision [42]. The global sensing result is denoted as $\mathbf{H}[\tau] = [H_1[\tau], H_2[\tau], \ldots, H_M[\tau]]$, in which $H_m[\tau] \in \{A, I\}$ indicates the status (active or inactive) of primary channel m in superframe τ. Nevertheless, the sensing error is inevitable in the wireless channel, particularly in cooperative spectrum sensing. Two metrics that represent the sensing performance are false alarm probability, $P_{f,m} = \Pr(H_m[\tau] = A\,|I)$, and detection probability, $P_{d,m} = \Pr(H_m[\tau] = A\,|A)$. The former represents the probability that the channel m is sensed as "active," but it is actually

''inactive.'' The latter indicates the probability that the channel is sensed correctly as ''active.''

Generally, the performance of the WirelessHART system can be lowered by the values of false alarm probability and misdetection probability that represents the channel is actually ''active'' but is sensed as ''inactive.'' More particularly, the false alarm results in the missing opportunity for the devices to use the primary channel because the devices trust the sensing outcome ''active'' and will not utilize the assigned primary channel. On the other hand, the misdetection event leads to transmission collision on primary channel m between the devices and PUs. For instance, when the false alarm event happens, the devices will not use the primary channel m for transmissions, which can result in a low successful transmission probability to the GW by using the ISM channels. In case of misdetection, the devices will transmit data on primary channel m because the sensing outcome is ''inactive''; however, the primary channel is actually ''active,'' leading to the transmission collision between the devices and PUs on the primary channel. In this paper, the probabilities of all primary channels are updated by the GW at the end of each superframe. Given the maximally allowable collision probability between the devices and PUs, the detection probability, $P_{d,m}$, can be maintained to be greater than a threshold, ς, by modifying sensing parameters to protect the PU communications on the primary channels [17], [43].

E. ENERGY HARVESTING

Each CAP is powered by a rechargeable battery that can be recharged by a solar energy harvester. Let E_B be the battery capacity of a CAP, which is assumed to be limited. Herein, the harvested energy in superframe τ of each CAP, denoted as $E^h[\tau]$, is finite, in which $E^h[\tau] \in \{E^{h,1}, E^{h,2}, \ldots, E^{h,\xi}\}$; $0 \leq E^{h,z} < E_B$, and $z \in \{1, 2, \ldots, \xi\}$, and is assumed to follow a Poisson distribution with a mean harvested energy, $E^{h,mean}$. The harvesting modeling and the impact of daytime and nighttime on solar harvesting performance were well studied [44], [45]. In literature [44], the empirical measurements were conducted to model the energy harvesting for solar-powered wireless devices and the results verified that the harvested energy highly depends on properties such as harvesting time, light intensity, and deployment operating environment. As a result, the Poisson distribution model for solar-powered energy harvesting can achieve a near fit with the practical measurements. Therefore, we adopt the Poisson distribution model for solar-energy harvesting in this paper. In the real-time implementation, the mean value of harvested energy, $E^{h,mean}$, may change according to daytime, nighttime, or different time intervals in a day. However, the system can also measure the value of $E^{h,mean}$ to update the policy in every different time intervals. Thus, the proposed scheme can work efficiently with the considered energy harvesting model.

F. PROBLEM FORMULATION

There are N data flows during a convergecast operation, in which the data flow $p_n | n \in \{1, 2, \ldots, N\}$ is defined as the data packet generated by device v_n. In this paper, the throughput of the network (i.e., reward), which is defined as the total number of successfully received packets at the GW in superframe τ, can be described as

$$R[\tau] = \sum_{n=1}^{N} R_n[\tau], \tag{1}$$

where

$$R_n[\tau] = \begin{cases} 1 & p_n \text{ is succesfully received by GW} \\ 0 & \text{otherwise} \end{cases}$$

represents the result indicator of the transmitted packet p_n in superframe τ. Some ISM channels might be blacklisted to protect wireless services that share a fixed portion of the ISM band, so the number of ISM channels available for use by WirelessHART might be restricted to less than 16. Therefore, efficient ISM channel utilization is critical for scheduling. The scheduling length was also considered, in which the scheduling is made to finish a convergecast with a minimum number of time slots. Thus, the CR technique is leveraged to enhance the performance of the WirelessHART system by opportunistically using the free primary channels. Obviously, the network will achieve better immediate throughput in the current superframe if the GW assigns more assisted primary channels to the CAPs for sensing and utilizing. On the other hand, the CAPs may lack of energy for use in future superframes due to the limits of battery capacity and energy harvesting capability. Therefore, the trade-off between the number of primary channels for sensing at the beginning of each superframe and the maximum long-term throughput is critical.

$I_m[\tau] \in \{0, 1\}$ is denoted as the sensing indicator of primary channel m in superframe τ. If it is selected to be sensed, $I_m[\tau] = 1$, and otherwise $I_m[\tau] = 0$. In addition, let E_s denote the amount of energy required to sense each primary channel, and the term $\sum_{m=1}^{M} E_s I_m[\tau]$ represents the total amount of sensing energy required in superframe τ, which may change due to the dynamics of primary channels.

Considering the foregoing analysis, the objective was to find the optimal joint ISM/primary channel assignment to all devices for maximizing the throughput of the WirelessHART in the long-term operation under the constraints of energy harvesting capacity, radio frequency resource, minimum size of the superframe, and buffering capacity at the devices. Therefore, the problem formulation can be expressed as follows:

$$\begin{aligned} &\max_{\mathbf{S}[\tau], \mathbf{S}_\mathrm{D}[\tau]} \left(\sum_{\tau=1}^{\infty} R[\tau] \right) \\ &s.t. \sum_{m=1}^{M} E_s I_m[\tau] \leq E_{\max} \\ &\quad N_{sl} \text{ and } N_{\mathrm{ISM}} \text{ are minimized} \\ &\quad \mathbf{S} \text{ and } \mathbf{S}_\mathrm{D} \text{ satisfy buffer constraints,} \end{aligned} \tag{2}$$

where $E_{\max}$ is the maximum amount of energy required for sensing at each CAP.

$$\mathbf{S} = \begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,N_{sl}} \\ C_{2,1} & C_{2,2} & \dots & \dots \\ \vdots & \vdots & \dots & \dots \\ C_{l_{\max},1} & \dots & \dots & C_{l_{\max},N_{sl}} \end{bmatrix}$$

represents the joint time and ISM/primary channel schedule for superframe τ, where $C_{i,t} = u \cup m | u \in \{1, 2, \dots, U\}, m \in \{U+1, \dots, U+M\}$ is the ISM/primary channel assigned for the link i of the slot t. $l_{\max}$ is the maximum number of parallel links assigned in a time slot of each superframe.

$$\mathbf{S}_{\mathrm{D}} = \begin{bmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,N_{sl}} \\ v_{2,1} & v_{2,2} & \dots & \dots \\ \vdots & \vdots & \dots & \dots \\ v_{l_{\max},1} & \dots & \dots & v_{l_{\max},N_{sl}} \end{bmatrix}$$

is the device scheduling (i.e. assignment for transmitting devices of links) for superframe τ, in which $v_{x,t} = n \in \{1, 2, \dots, N\}$ denotes that the device n is assigned to transmit data in time slot index t. N_{sl} denotes the number of slots in the superframe. N_{ISM} represents the total maximum number of ISM channels assigned in a superframe (i.e. the maximum number of parallel transmissions using ISM channels in a time slot of the scheduling $\mathbf{S}$). By defining the proper $\mathbf{S}$ and $\mathbf{S}_{\mathrm{D}}$, we allow multiple parallel transmissions on ISM/primary channels in each time slot to improve the latency and the data transmission performance of the system.

The optimization in this paper is affected considerably by the number of selected primary channels, which are used to replace the ISM channel for the transmissions of the devices. When the number of primary channels is large, the energy required for sensing of CAPs will also be increased. This may deteriorate performance of the system due to the energy-constrained issue of CAPs. Thus, the primary channels should be used properly with an energy-efficient sensing manner according to the dynamic activity of the primary users on their licensed channels and the limit harvested energy of CAPs. However, it is difficult to directly obtain the solution for the problem (2) due to the dynamics of the primary channels and the complexity of the joint time slot and ISM/primary channel allocation for all devices. Therefore, problem (2) can be decomposed into three processes: joint ISM channel and data flow allocation process, primary channel allocation process, and joint time and ISM/primary channel scheduling process. The main idea is that, the ISM channels will be scheduled offline first with a minimum number of ISM channels and time slots in the superframe. Subsequently, the primary channels will be allocated according to the dynamics of the primary channels and the remaining energy of the CAPs in each superframe.

In particular, in the joint ISM channel, device and data flow scheduling process, the GW determines the ISM channel scheduling, device scheduling, and data flow scheduling, which are respectively denoted by $\mathbf{S}_{\mathrm{ISM}}$, $\mathbf{S}_{\mathrm{D}}$ and $\mathbf{S}_{\mathrm{DF}}$, in which only ISM channels are assigned to transmit the respective data flows for all the devices. The objective of this process is to determine $\mathbf{S}_{\mathrm{ISM}}$, $\mathbf{S}_{\mathrm{D}}$, and $\mathbf{S}_{\mathrm{DF}}$ with a minimum number of required ISM channels and time slots, which is expressed as

$$\begin{aligned} &\min_{\mathbf{S}_{\mathrm{ISM}},\mathbf{S}_{\mathrm{D}},\mathbf{S}_{\mathrm{DF}}} N_{\mathrm{ISM}} \text{ and } \min_{\mathbf{S}_{\mathrm{ISM}},\mathbf{S}_{\mathrm{D}},\mathbf{S}_{\mathrm{DF}}} N_{sl} \\ &s.t.\ \mathbf{S}_{\mathrm{ISM}}, \mathbf{S}_{\mathrm{D}}, \text{ and } \mathbf{S}_{\mathrm{DF}} \text{ satisfy buffer constraints,} \end{aligned} \tag{3}$$

where

$$\mathbf{S}_{\mathrm{ISM}} = \begin{bmatrix} C^{\mathrm{ISM}}_{1,1} & C^{\mathrm{ISM}}_{1,2} & \dots & C^{\mathrm{ISM}}_{1,N_{sl}} \\ C^{\mathrm{ISM}}_{2,1} & C^{\mathrm{ISM}}_{2,2} & \dots & \dots \\ \vdots & \vdots & \dots & \dots \\ C^{\mathrm{ISM}}_{N_{\mathrm{ISM}},1} & \dots & \dots & C^{\mathrm{ISM}}_{N_{\mathrm{ISM}},N_{sl}} \end{bmatrix}$$

is the ISM channel scheduling, where $C^{\mathrm{ISM}}_{i,t} = u | u \in \{1, 2, \dots, U\}$ indicates that the device n is assigned to transmit data on ISM channel i in time slot index t.

$$\mathbf{S}_{\mathrm{DF}} = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,N_{sl}} \\ p_{2,1} & p_{2,2} & \dots & \dots \\ \vdots & \vdots & \dots & \dots \\ p_{N_{\mathrm{ISM}},1} & \dots & \dots & p_{N_{\mathrm{ISM}},N_{sl}} \end{bmatrix}$$

is the data flow scheduling, in which $p_{i,t} = n | n \in \{1, 2, \dots, N\}$ denotes that packet n is transmitted on ISM channel i in time slot index t.

It is noted that the joint ISM channel, device and data flow allocation process are determined off-line by the GW according to the system parameters, which will be presented in Section III. These are, then disseminated to all field devices to store in their local memory storage. Furthermore, once the logical ISM channels are assigned in $\mathbf{S}_{\mathrm{ISM}}$, they can be mapped easily to the actual ISM channels for real-time convergecast operation. After defining $\mathbf{S}_{\mathrm{ISM}}$, $\mathbf{S}_{\mathrm{D}}$, and $\mathbf{S}_{\mathrm{DF}}$, the second process, called the primary channel allocation process, will be implemented based on the dynamics of primary channel activity. In the second process, the primary channel allocation, $\mathbf{A}$, is determined, in which the primary channels are allocated to the data flows through each superframe based on the system state and predefined $\mathbf{S}_{\mathrm{ISM}}$ and $\mathbf{S}_{\mathrm{DF}}$ using deep reinforcement learning as follows:

$$\begin{aligned} &\max_{\mathbf{A}[\tau]} \left(\sum_{\tau=1}^{\infty} R[\tau] \right) \\ &s.t. \sum_{m=1}^{M} E_s I_m[\tau] \le E_{\max}, \end{aligned} \tag{4}$$

where $\mathbf{A}[\tau] = [A_1[\tau], A_2[\tau], \dots, A_N[\tau]]$ represents the primary channel assignment for data flows in superframe τ, with $A_n[\tau] \in \{0, 1, 2, \dots, M\} | n \in \{1, 2, \dots, N\}$ denotes the assigned primary channel for data flow n. $A_n[\tau] = 0$ indicates that the data flow n is not allocated to any primary channel. In the third process, the joint time and ISM/primary

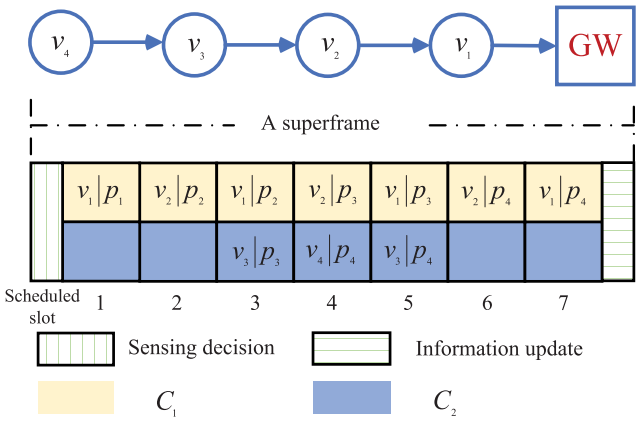

FIGURE 3. Example of a joint ISM channel, device, and data flow allocation (N = 4).

channel scheduling, **S**, is made by each device after receiving the corresponding global sensing results **H** (broadcasted by CAPs) such that only the primary channels sensed to be free according to **A**, are used to replace the ISM channels based on $\mathbf{S}_{\text{ISM}}$. With joint ISM channel, device and data flow scheduling, each data flow may be assigned on different ISM channels and devices in a superframe. However, when once a data flow is assigned to a primary channel, all the links associated to that data flow will be assigned to the same primary channel in the superframe.

Overall, to solve problem (2), the solution for problem (3) is first found through off-line scheduling to obtain $\mathbf{S}_{\text{ISM}}$, $\mathbf{S}_{\text{D}}$, and $\mathbf{S}_{\text{DF}}$. Subsequently, we leverage the deep reinforcement learning to deal with problem (4) by directly interacting with the environment to learn the optimal scheduling for each system state.

III. JOINT ISM CHANNEL, DEVICE AND DATA FLOW SCHEDULING

At the beginning of each superframe, each node generates a new data packet for forwarding to the GW. The objective is to efficiently schedule in a superframe for all devices to transmit their packets to the GW. Accordingly, this section investigates joint ISM channel, device, and data flow scheduling that requires a minimum number of ISM channels and time slots. Each device is allocated to transmit a data flow on an ISM channel with a time slot index, as shown in Figure 3. Because each device has a single-packet buffering capacity, a device with a data packet in its buffer needs to be scheduled for transmission before receiving a new packet. For example, in the first time slot of the scheduling in Figure 3, the only device v_1 can transmit its data packet (i.e., p_1) to its destination (i.e., the GW) because at this time slot every device have their own data flow generated by themselves, and there is no device with an empty buffer at the beginning of the first time slot. In the second time slot, v_1 has an empty buffer, so v_2 is assigned to transmit p_2 to its destination (i.e. v_1).

The reliability of each link (v_i, v_j) on each ISM channel, which is defined as the successful packet reception ratio

Algorithm 1 Joint ISM Channel, Device, and Data Flow Scheduling

Input: N, $G = (V, E)$.
Output: $\mathbf{S}_{\text{ISM}}$, $\mathbf{S}_{\text{D}}$, and $\mathbf{S}_{\text{DF}}$.
$\Delta_n = 0 \,|\forall n \in V\,;\, \Delta' = 0$.
for $t = 1 : 2N - 1$ **do** //
  $i_{\text{ISM}} = 1$
  if $t \mod 2 == 1$ **then**
    $\mathbf{S}_{\text{ISM}}(i_{\text{ISM}}, t) = 1$.
    $\mathbf{S}_{\text{D}}(i_{\text{ISM}}, t) = i_{\text{ISM}}$.
    $\mathbf{S}_{\text{DF}}(i_{\text{ISM}}, t) = \Delta' + 1$.
    $i_{\text{ISM}} = i_{\text{ISM}} + 1$.
  end if
  for each v_n scheduled in $\mathbf{S}_{\text{D}}$ of time slot $t - 1$ **do**
    if $(n + 1 \le N) \cap (\Delta_{n+1} < N - (n + 1) + 1)$ **then**
      $\mathbf{S}_{\text{ISM}}(i_{\text{ISM}}, t) = i_{\text{ISM}}$.
      $\mathbf{S}_{\text{D}}(i_{\text{ISM}}, t) = n + 1$.
      $\Delta_{n+1} = \Delta_{n+1} + 1$.
      if $t \mod 2 == 0$ **then**
        $\mathbf{S}_{\text{DF}}(i_{\text{ISM}}, t) = \mathbf{S}_{\text{DF}}(i_{\text{ISM}}, t - 1) + 1$.
      else
        $\mathbf{S}_{\text{DF}}(i_{\text{ISM}}, t) = \mathbf{S}_{\text{DF}}(i_{\text{ISM}} - 1, t) + 1$.
      end if
    end if
    $i_{\text{ISM}} = i_{\text{ISM}} + 1$.
  end for
end for

(i.e., successful transmission probability on ISM channel u), is denoted as ρ_u^{ij}. In this paper, we consider the constraint of interference on the ISM channels in each link. In a convergecast operation, each data flow needs to be successfully transmitted via all links that are routed to the GW. Thus, the successful packet reception ratio on the ISM channels becomes relatively low if the size of the network (i.e., the total number of field devices) is large. To increase a number of packets received at the GW, the primary channels are exploited such that the devices can switch to currently the free channels to achieve more reliable transmissions. By adopting the jointly optimal convergecast time and channel scheme reported elsewhere in [18], the design of the joint ISM channel, device and data flow scheduling to obtain the minimum number of time slots and ISM channels is expressed in **Algorithm 1**. In the algorithm, Δ_n denotes the number of packets that field device v_n has transmitted since the beginning of a convergecast operation. The number of time slots required for the single-buffer linear convergecast is $2N - 1$. Meanwhile, the minimum number of required ISM channels to complete the convergecast in $2N - 1$ slots is $\frac{1}{2}N$ [18]. Note that $\mathbf{S}_{\text{ISM}}$, $\mathbf{S}_{\text{D}}$, and$\mathbf{S}_{\text{DF}}$ will be used to generate the joint time and ISM/primary channel scheduling, which is presented in Section V.

In this paper, we do not investigate the mutual interference management between the cognitive network and the primary

network because the methods of mitigating interference in CRNs have been well investigated in the literature [46]. Instead, we focus on reducing the impact of interference generated by the nearby devices using the same ISM bands on the packet delivery process of the field devices in WirelessHART. Therefore, we aim to design a spectrum allocation scheme, in which the field devices can properly switch to primary channels when they are sensed and estimated to be more reliable than ISM channels by using CR.

In essence, CR is mainly used to solve the issue of spectrum under-utilization. However, allowing cognitive radios to opportunistically share the licensed channels cannot always guarantee the higher reliability than using ISM channels due to the imperfect sensing characteristic of the realistic scenarios. The performance of a CR network can be degraded significantly if the sensing engine of the cognitive radios induces a lot of faults in detecting the state of the primary channels. Therefore, the spectrum sensing techniques play an important role to guarantee the reliability of the spectrum sharing with two key metrics, *probability of detection* (P_d) and *probability of false alarm* (P_f), which is discussed in Section II. D. In this paper we do not focus on designing spectrum sensing algorithms. The values of P_d and P_f are set to guarantee an acceptable level of interference with the primary networks, where the value of P_d should be greater than 0.9 as studied in [17]. Given these values, we design a scheme to optimally use the primary channels for the transmissions of the devices with estimating the probability that the primary channels are free in each superframe. Consequently, our proposed algorithm allows the GW to learn the optimal transmission policy on ISM/primary channels for the devices in WirelessHART systems, in which a primary channel is only selected when its estimated reliability is higher than that of ISM channels.

IV. DEEP Q-LEARNING APPROACH FOR PRIMARY CHANNEL ALLOCATION

In this section, the primary channel allocation problem in (4) is reformulated as the framework of a MDP. Generally, the MDP problem can be solved using the value iteration-based dynamic programming in a partially observable Markov decision process (POMDP) algorithm [47]. On the other hand, the POMDP solution requires high formulation and computational cost, reducing the system performance in practice. Another popular approach to the MDP problem is the Q-learning algorithm, where the agent (i.e., the GW) is able to learn the optimal policy by regularly interacting with the working environment. By taking an action at a given state, the agent makes the environment transit to another state. The agent receives the corresponding reward according to the quality of the action taken. In that way, the agent can maximize the cumulative reward by interacting with the environment on a trial-and-error basis. However, the Q-learning method is unsuitable for the problems with high-dimensional state and action spaces. Therefore, deep Q-learning was adopted to solve the MDP problem, in which a deep neural network, represented by a weight vector, was used to approximate the Q-value of each state-action pair. Consequently, a deep learning scheme is considered an effective approach for the MDP problem, where the complexity is degraded significantly and a nearly optimal solution can be acquired.

A. MARKOV DECISION PROCESS

Herein, the primary channel allocation problem in (4) is reformulated as an MDP framework based on the decision-making model. We first define the state and action spaces of the MDP framework. The state space of the system is denoted as $\mathbb{S}$, in which each state of the system at superframe τ is composed of the remaining energy of CAPs and the belief of primary channels, as follows:

$$s[\tau] = \left(\mathbf{E}^{rm}[\tau], \mathbf{b}[\tau]\right), \tag{5}$$

where $\mathbf{E}^{rm}[\tau] = \left[E_1^{rm}[\tau], E_2^{rm}[\tau], \ldots, E_K^{rm}[\tau]\right]$ is the energy vector including the remaining energy of CAPs at the beginning of superframe τ; $\mathbf{b}[\tau] = [b_1[\tau], b_2[\tau], \ldots, b_M[\tau]]$ represents the vector of the probabilities that the primary channels are inactive (i.e., free) in superframe τ. The values of these probabilities are updated by using equations (8-10) at the end of each superframe according to the result of each action taken by the GW (i.e., the GW successfully/unsuccessfully receives the assigned packets on the primary channels). Thereby, the GW can estimate the belief of the primary channels to decide the channel allocation for the devices in the next superframe.

Based on the system state, the GW, which is considered the learning agent, is in charge of selecting an action. Particularly, the GW makes the primary channel allocation, in which the primary channels are assigned to the data flows such that the number of successfully received packets is maximized over the long run. Hence, the action space of the system can be denoted as follows:

$$\begin{aligned}\mathbf{a}[\tau] &= \mathbf{A}[\tau] \\ &= [A_1[\tau], A_2[\tau], \ldots, A_N[\tau]] \in \mathbb{A},\end{aligned} \tag{6}$$

where $A_n \in \{0, 1, 2, \ldots, M\} | n \in \{1, 2, \ldots, N\}$ is the primary channel allocation for data flow n, which is described in Section II.F.

The operation of the system in a superframe can be described as follows. At the start of a superframe τ, the agent observes the system state and decides an action $\mathbf{a}[\tau]$. The agent then forwards it to CAPs through the dedicated primary control channel. The CAPs sense the primary channels based on $\mathbf{a}[\tau]$ and sends the local sensing results to the FC to decide the global sensing results. Subsequently, the global sensing results $\mathbf{H}[\tau] = [H_1[\tau], H_2[\tau], \ldots, H_M[\tau]]$, where $H_m[\tau] \in \{I, A, NA\}$, made by the FC, will distribute to the APs and the GW. The notation I and A denote the ''inactive'' and ''active'' state of primary channel m, respectively, while NA indicates that the primary channel m is not assigned to be used in superframe τ.

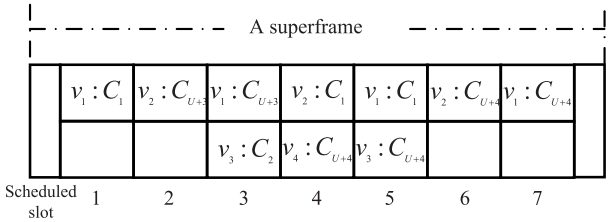

FIGURE 4. Example of the joint time and ISM/primary scheduling $\mathbf{S}[\tau]$ with $\mathbf{a}[\tau] = [0, 3, 1, 4]$ and $\mathbf{H}[\tau] = [A, NA, I, I]$.

Subsequently, the CAPs broadcast $\mathbf{a}[\tau]$ and $\mathbf{H}[\tau]$ to the devices for their joint time and ISM/primary scheduling, $\mathbf{S}[\tau]$. Note that the primary channels assigned in $\mathbf{a}$ will not be used by the devices if the global sensing results shows the active state of the primary channels. This means that the devices can only use the primary channels that are currently sensed to be free in each superframe. Figure 4 gives an example of a joint time and ISM/primary channel scheduling, given the joint ISM and data flow allocation in Figure 3, where $\mathbf{a} = [0, 3, 0, 4]$ and $\mathbf{H}[\tau] = [A, NA, I, I]$. From the figure, primary channel 1 is assigned for the data flow 3 in $\mathbf{a}$, but three links of data flow 3 are finally allocated to the channel ISM in the joint time and ISM/primary channel scheduling $\mathbf{S}[\tau]$ because the global sensing result of primary channel 1 is "active." The links of data flows 2 and 4 are assigned successfully to primary channels 3 and 4, respectively, because the sensing results are "inactive."

After determining $\mathbf{S}[\tau]$, each device produces the sub-scheduling $\mathbf{S}^{sub}[\tau]$ for itself, in which each device is set to one of the possible states, such as "transmit," "receive," or "sleep" in the time slots of the superframe. As a result, the devices perform their transmissions in the corresponding time slot index based on $\mathbf{S}^{sub}[\tau]$. At the end of a superframe, the GW receives an immediate reward, $R[\tau]$, which is defined as the packets received at the GW in the current superframe τ and is calculated using equation (1). At the end of a superframe, the GW updates the remaining energy information reported by the CAPs and the belief of the primary channels. This action makes the system transfer from state $s[\tau]$ to another state $s[\tau+1]$, which is updated at the end of each superframe as follows. The energy level at each CAP in the next superframe can be expressed as

$$E_k^{rm}[\tau+1] = \min\left(E_k^{rm}[\tau] - E_b - \sum_{m=1}^{M} E_s I_m[\tau] + E_k^h[\tau], E_B\right), \quad (7)$$

where E_b represents the broadcasting energy of each CAP for broadcasting the scheduling information (i.e. the global sensing results and primary channel assignment) to the devices. $E_k^h[\tau]$ represents the total amount of harvested energy of the AP_k during superframe τ, and

$$I_m[\tau] = \begin{cases} 0 & \text{if } H_m[\tau] = NA \\ 1 & \text{otherwise} \end{cases}$$

is the sensing indicator of primary channel m in superframe τ. In case $H_m[\tau] = I$, the devices then use primary channel m for their data transmissions, and the GW successfully receives and decodes the data flow transmitted on primary channel m at the end of superframe τ. The belief of the primary channel m is then updated by

$$b_m[\tau+1] = P_{II,m}. \quad (8)$$

In case $H_m[\tau] = I$, the devices then use primary channel m for their data transmissions, but the GW unsuccessfully receives and decodes the data flow transmitted on primary channel m at the end of superframe τ. The belief of primary channel m is updated by

$$b_m[\tau+1] = P_{AI,m}. \quad (9)$$

In case $H_m[\tau] = A$, the devices then do not use primary channel m for their data transmission, then the belief of the primary channel m is updated by

$$b_m[\tau+1] = \frac{b_m[\tau]P_{f,m}P_{II,m} + (1-b_m[\tau])P_{d,m}P_{AI,m}}{b_m[\tau]P_{f,m} + (1-b_m[\tau])P_{d,m}}. \quad (10)$$

For example, in Figure 4, if the channel C_{U+3} is sensed to be "active," i.e. $H_3[\tau] = A$, then v_2 and v_1 will use channel C_1 in time slot indices 2 and 3, as defined in Figure 3, for the current superframe. On the other hand, if $H_m[\tau] = NA$, the CAPs did not sense the status of the primary channel m. Hence, the updated belief of the channel m in this case is

$$b_m[\tau+1] = b_m[\tau]P_{II,m} + (1-b_m[\tau])P_{AI,m}. \quad (11)$$

This work aims to generate the joint time and channel scheduling policy to maximize the long-term reward from the current superframe. Accordingly, the proper primary channel allocation is required in each superframe to maximize the total discounted reward. We define the state−action value function as expected sum of rewards when the system is in state s and action $\mathbf{a} \in \mathbb{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{|\mathbb{A}|}\}$, as follows:

$$Q(s, \mathbf{a}) = \mathbb{E}\left[\sum_{i=\tau}^{\infty} \gamma^{i-\tau} R[\tau] \,|\, s[\tau] = s, \mathbf{a}[\tau] = \mathbf{a}\right], \quad (12)$$

where γ is the discount factor, and $\mathbb{E}[.]$ represents the expectation operator. The goal is to find the optimal action, $\mathbf{a}^*$, in the current superframe to maximize the Q-value function, as follows

$$\mathbf{a}^* = \arg\max_{\mathbf{a}\in\mathbb{A}} \{Q(s, \mathbf{a})\}. \quad (13)$$

Using the Q-learning algorithm, the agent calculates the Q-value in each step (i.e., each superframe) and stores it in a Q-table to find the optimal solution. The simplest form of updating the state-action value function can be given as

$$Q(s, \mathbf{a}) = Q(s, \mathbf{a}) + \alpha\left[R + \gamma \max_{\mathbf{a}'\in\mathbb{A}} Q(s', \mathbf{a}') - Q(s, \mathbf{a})\right], \quad (14)$$

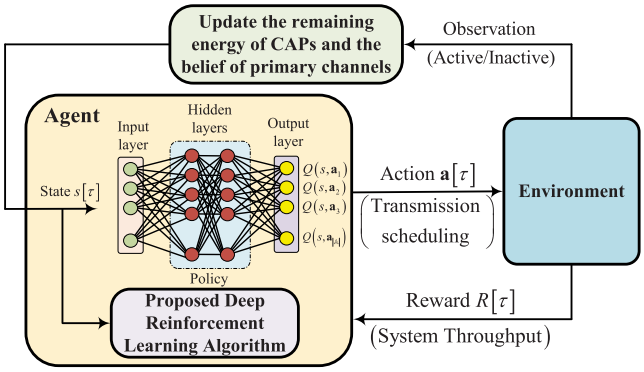

FIGURE 5. Structure of the proposed deep reinforcement learning for transmission scheduling.

where $\alpha \in (0, 1)$ is the learning rate; s' and $\mathbf{a}'$ represent the next state and action, respectively; R is the immediate reward that the GW receives at the end of the current superframe. With the appropriate configuration, the Q-learning can offer the optimal value function after the training phase, from which the agent can choose the optimal action in each superframe. Nevertheless, the traditional Q-learning method might face with the wide variance in the function approximation when system size increases, making the scheme converge to a locally optimal policy. Therefore, we investigate a method to approximate the Q-value function, which is called deep Q-learning. Specifically, a neural network was constructed with a vector of weight to approximate the Q-value function, denoted by $Q(s, \mathbf{a}, \mathbf{w})$, such that the proposed scheme can be applied effectively in large-size systems.

B. DEEP Q-LEARNING BASED SOLUTION

This section presents the proposed DQL algorithm to solve the problem of the MDP. DQL is a combination of a value-based approach and a neural network. Herein, the feed-forward neural network (FNN) was employed to approximate the Q-value function of each action according to a given state, named a Q-network. The network was composed of an input layer, multiple hidden layers, and an output layer, as illustrated in Figure 5, in which, the input of FNN is defined as the system state s while the output is the Q-value of any state-action pair. The input layer contains $(K + M)$ neuron units representing the elements of each state. Each hidden layer is a fully connected layer that includes a finite number of neuron units where the rectified linear unit function is utilized as a nonlinear activation function. The output vector of the hidden layers can be expressed by

$$\mathbf{y} = \max(0, \mathbf{w}.\mathbf{s} + \mathbf{u}), \tag{15}$$

where $\mathbf{w}$ and $\mathbf{u}$ are the weight and bias parameters, respectively. The output layer of the FNN is a vector with the size of $|\mathbb{A}|$, which matches the output values of the last hidden layer to the estimated Q-value of each state-action pair by applying the linear action function. During training, the network parameters were modified to minimize the loss function

Algorithm 2 Training Process of Deep Q-Learning Algorithm

Input: $U, M, N, K, E_b, E_s, E_B, \alpha, \delta, \gamma, P_{AI}, P_{II}, P_{d,m}, P_{f,m}, d_\varepsilon, \varepsilon_{\min}$.
Output: Q-network parameter $\mathbf{w}$.
Initialize parameter $\mathbf{w}$, $\mathbf{w}'$.
Initialize exploration rate ε.
Initialize replay memory D.
while not converged **do**
  Initialize a random action $s \in \mathbb{S}$
  for each superframe $\tau = 1, 2 \ldots .T$ **do**
    Observe the current state $s[\tau]$.
    Select an action for current step:
    $$\mathbf{a}[\tau] = \begin{cases} \arg\max_{\mathbf{a}[\tau]\in\mathbb{A}} Q(s[\tau], \mathbf{a}[\tau], \mathbf{w}) & \text{w.p. } 1-\varepsilon \\ \text{any action } \mathbf{a}[\tau] \in \mathbb{A} & \text{otherwise.} \end{cases}$$
    Perform the chosen action $\mathbf{a}[\tau]$, obtain the reward $R[\tau]$, and the next state s'.
    Store the transition $\langle s[\tau], \mathbf{a}[\tau], R[\tau], s' \rangle$ in replay memory D.
    Randomly sample the mini batches, $\langle s_j, \mathbf{a}_j, R_j, s_{j+1} \rangle$ from replay memory D.
    for j in mini-batches size **do**
      Calculate the current Q-value $Q(s_j, \mathbf{a}_j, \mathbf{w})$.
      Calculate the target Q-value:
      $$Q_{target} = \begin{cases} R_j & \text{final } s_{j+1} \\ R_j + \gamma \max_{\mathbf{a}'\in\mathbb{A}} Q(s_{j+1}, \mathbf{a}', \mathbf{w}') & \text{otherwise} \end{cases}$$
    end for
    Update Q-network parameter $\mathbf{w}$.
    Update next state s'.
    Update exploration rate $\varepsilon = \max(\varepsilon \times d_\varepsilon, \varepsilon_{\min})$.
  end for
  Copy network parameter from $\mathbf{w} \rightarrow \mathbf{w}'$.
end while

defined as the mean square error between the current value and the target Q-value, as follows:

$$L(\mathbf{w}) = E\left[\left(R + \gamma \max_{\mathbf{a}'\in\mathbb{A}} Q(s', \mathbf{a}', \mathbf{w}) - Q(s, \mathbf{a}, \mathbf{w})\right)^2\right], \tag{16}$$

where $R + \gamma \max_{\mathbf{a}'\in\mathbb{A}} Q(s', \mathbf{a}', \mathbf{w})$ denotes the target Q-value. Two well-known methods were also adopted, namely experience replay [48] and fixed target network [49] to remove the oscillation caused by the data correlations between consecutive transitions in Q-function approximation. Another neural network with network weight $\mathbf{w}'$ was used to calculate the target Q-value while the network parameters remained unchanged during some training iterations. In the experience-replay technique, the transition tuples $(s, \mathbf{a}, R, s')$ are stored in the replay memory, D, in which the mini batches are randomly selected to train the Q-network to increase sample

Algorithm 3 Joint Time and ISM/Primary Channel Scheduling

Input: N_{ISM}, $\mathbf{S}_{\text{ISM}}$, $\mathbf{S}_{\text{DF}}$, $\mathbf{a}$, and $\mathbf{H}$.
Output: Scheduling $\mathbf{S}$.
$\mathbf{S} = []$.
for $t = 1 : 2N - 1$ **do**
  for $u = 1 : N_{\text{ISM}}$ **do**
    $n = \mathbf{S}_{\text{DF}}(u, t)$.
    if n is not empty **then**
      if $A_n \neq 0 \cap H_{A_n} ==$ "I" **then**
        $\mathbf{S}(u, t) = U + A_n$. // Primary channel allocation
      else
        $\mathbf{S}(u, t) = \mathbf{S}_{\text{ISM}}(u, t)$. // ISM channel allocation
      end if
    end if
  end for
end for

efficiency as follows:

$$L(\mathbf{w}) = E_D\left[\left(R + \gamma \max_{\mathbf{a}' \in \mathbb{A}} Q(s', \mathbf{a}', \mathbf{w}') - Q(s, \mathbf{a}, \mathbf{w})\right)^2\right]. \quad (17)$$

The target network parameters are repetitively replaced by those of Q-network in several training steps. The temporal difference (TD) error between the current Q-value and the target value was calculated by using the following equation:

$$\delta = R + \gamma \max_{\mathbf{a}' \in \mathbb{A}} Q(s', \mathbf{a}', \mathbf{w}') - Q(s, \mathbf{a}, \mathbf{w}). \quad (18)$$

Using the stochastic gradient descent to minimize the loss function in the direction of gradient, the weight parameter $\mathbf{w}$ can be updated as

$$\mathbf{w} = \mathbf{w} + \alpha\delta\nabla_{\mathbf{w}} Q(s, \mathbf{a}, \mathbf{w}). \quad (19)$$

During the training phase, the agent selects an action $\mathbf{a}$ at the beginning of each superframe according to an $\varepsilon-$greedy policy, in which $0 \leq \varepsilon \leq 1$ represents the exploration rate. The exploration rate ε decays over each time step at the rate of d_ε. The training is repeated until convergence. **Algorithm 2** outlines the proposed deep Q-learning procedure.

V. JOINT TIME AND ISM/PRIMARY CHANNEL SCHEDULING AND SUB-SCHEDULE EXTRACTION

This section presents the way the field devices generate the joint time and ISM/primary channel scheduling $\mathbf{S}[\tau]$ and the sub-scheduling $\mathbf{S}^{sub}[\tau]$ when receiving $\mathbf{a}[\tau]$ and $\mathbf{H}[\tau]$. **Algorithm 3** describes the joint time and ISM/primary channel scheduling. In $\mathbf{S}[\tau]$, the ISM/primary channels are assigned to data transmissions with the specific time slot index. They then need to generate the sub-scheduling $\mathbf{S}^{sub}[\tau]$ for itself based on the generated $\mathbf{S}[\tau]$ and $\mathbf{S}_{\text{D}}$ in which the sub-scheduling shows the assigned state for each device in each time slot of the whole superframe τ. At each time slot in a superframe, each device can operate in three states: transmit (Tr), receive (Re), and sleep (Sl). The sub-scheduling of device v_n, which is denoted by $\mathbf{S}_n^{sub}[\tau]$, is a matrix with the size of a $2 \times 2N - 1$, in which the first row indicates the state of the device v_n; the second row shows the allocated channel. **Algorithm 4** outlines the procedure for generating the sub-scheduling of each device. Figure 6 gives an example of the sub-scheduling generations of the device v_1 and v_2.

Algorithm 4 Extraction for Sub-Scheduling of Device v_n

Input: N_{ISM}, $\mathbf{S}_{\text{D}}$ and $\mathbf{S}$.
Output: Sub-scheduling $\mathbf{S}_n^{sub}$.
$\mathbf{S} = []$.
for $t = 1 : 2N - 1$ **do**
  for $u = 1 : N_{\text{ISM}}$ **do**
    if $\mathbf{S}_{\text{D}}(u, t) == n$ **then**
      $\mathbf{S}_n^{sub}(1, t) = Tr$.
      $\mathbf{S}_n^{sub}(2, t) = \mathbf{S}(u, t)$.
    else if $\mathbf{S}_{\text{D}}(u, t) == n + 1$ **then**
      $\mathbf{S}_n^{sub}(1, t) = Re$.
      $\mathbf{S}_n^{sub}(2, t) = \mathbf{S}(u, t)$.
    else
      $\mathbf{S}_n^{sub}(1, t) = Sl$.
    end if
  end for
end for

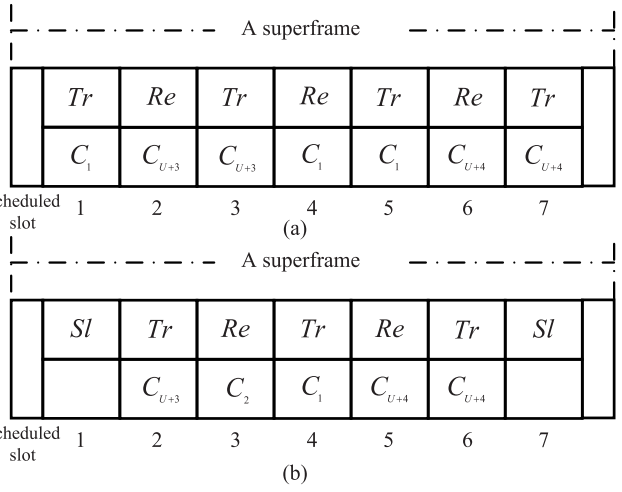

FIGURE 6. An example of the sub-scheduling generation of device v_1 (a) and device v_2 (b), based on the example of Figure 4.

The process of convergecast operation using the proposed scheduling is summarized as follows. At first, the ISM channel, device and data flow scheduling, i.e. $\mathbf{S}_{\text{ISM}}$, $\mathbf{S}_{\text{D}}$, and $\mathbf{S}_{\text{DF}}$, respectively, are designed offline by the GW and are stored locally in each device. These tables of scheduling are fixed in every superframe of the operation. At the beginning of each superframe, GW assigns primary channels to devices by generating a vector of the primary channel assignment, $\mathbf{a}$, and sends it to CAPs for cooperative spectrum sensing. Subsequently, the CAPs locally sense the assigned primary channel and send the local results to FC. Next, FC will produce the global sensing result $\mathbf{H}$ and send it back to CAPs. CAPs

broadcast **a** and **H** through the dedicated primary control channel. Based on **a** and **H**, each device generates a joint time and ISM/primary channel scheduling, **S**, by itself using **Algorithm 3** and sub-scheduling, $\mathbf{S}_n^{sub}[\tau]$, using **Algorithm 4**. Finally, devices will transmit their packet according to the assigned channel and slot obtained in $\mathbf{S}_n^{sub}[\tau]$. At the end of each superframe, the GW will update the remaining energy of the CAPs and the belief of the primary channels for the next superframe scheduling.

VI. SIMULATION RESULTS

In this section, we summarized performance of the proposed scheme in comparison with a baseline scheme [18], in which a myopic optimization is adopted for primary channel assignment [43] and a random scheme through a numerical simulation using Python 3.7 with TensorFlow deep learning libraries. For the baseline scheme, ISM channel scheduling was implemented using the algorithm in [18]. The system performed the primary channel assignment with the largest amount of sensing energy in each superframe. For the random scheme, the action of primary channel assignment was taken randomly. The numerical simulation results can demonstrate the efficiency of the proposed transmission scheduling under various network parameters.

We consider the small-scale network area, where the interference generated by PUs transmissions on the primary channels is assumed to be large enough during forwarding time of the devices, such that a received packet cannot be successfully decoded by the receiving side (i.e., field devices and the GW) due to the high interference from the PUs. Therefore, a packet is successfully received at the GW on the primary channels in a superframe if and only if no any collision generated by the PUs. The simulation included four field devices and four primary channels. The battery capacity in each CAP, E_B, was set to 20 μJ. The broadcasting energy was $E_b = 3\mu J$ and the sensing energy for each primary channel was $E_s = 2\mu J$. The neural network has four layers: an input layer, two hidden layers with 64 nodes each, and an output layer. The learning rate was $\alpha = 1.5 \times 10^{-2}$. The state transition probability from the "active" to "inactive" state of each primary channel was set to 0.2 and 0.8 from the "inactive" to "inactive" state. The values of detection probability and false alarm probability were 0.9 and 0.1, respectively, from reference [17]. The ReLU function and the linear function were used as an activation function for the hidden layers and the output layer of the DQN, respectively. Furthermore, an adaptive optimization algorithm (i.e. the Adam optimizer) was used to update the weights of the Q-network periodically. The sizes of the replay memory and minibatch were set to 3000 and 300, respectively. The initial exploration rate was set to 1; the decay rate was 0.9999, and the minimum exploration rate was 0.02. The mean value of harvested energy was $E^{h,mean} = 5\mu J$ and each CAP was assumed to manage two field devices. The successful packet reception ratio of a link on each ISM channel was assumed to be identical, i.e. $\rho_u^{ij} = \rho_u = 0.7$. The Q-network was trained over 200 episodes, each of which

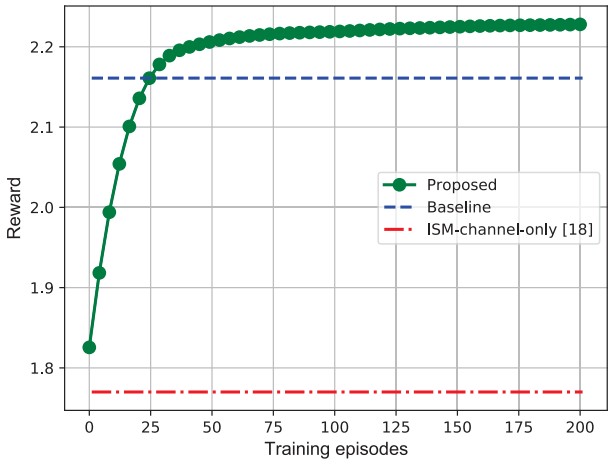

FIGURE 7. Convergence behavior of the proposed method.

contained 4×10^3 superframes. The simulation results were obtained by averaging 10^5 superframes.

We first examined the convergence rate of the proposed algorithm with the increment of training episodes in Figure 7. In the simulation, the ISM-channel-only scheme was implemented by merely using the ISM channels, and the optimal scheduling was obtained using the algorithm in [18]. For the proposed scheme, we regularly calculated the average value of the rewards received in a number of superframes in each episode to plot a curve with less fluctuation during the training phase, as shown in Figure 7. In this paper, the Q-network keeps training until it meets the convergence condition ($\leq$0.001) or reaches the maximum number of predefined training episodes. As a consequence, the throughput of the system using the proposed scheme converged to an optimal value after approximately 100 episodes. On the other hand, the baseline and ISM-channel-only schemes offer a lower reward at 2.16 and 1.7 (received packets), respectively. The reason is that the baseline scheme always maximizes the current reward regardless of the status of the CAPs battery and the primary channels in each superframe. Consequently, it would have insufficient energy for future utilization because the harvested energy and battery capacity are limited. This leads to overlooking of the primary channels when they are free. Furthermore, the curves show a significant improvement in the data aggregation performance when the primary channels are used in the network compared to the ISM-channel-only scheme.

The throughput of the schemes was plotted as ρ_u changes from 0.4 to 0.8 in Figure 8 to explore the impact the successful packet reception ratio of the ISM channels on network performance. The GW can receive more packets sent by the devices when the ISM channels have lower interference for all schemes. Because the GW can have a higher probability of receiving the packets transmitted by the devices on the ISM channels with a high value of ρ_m when the assigned primary channels are sensed to be "active." The performance of proposed scheme was shown to be superior to other schemes

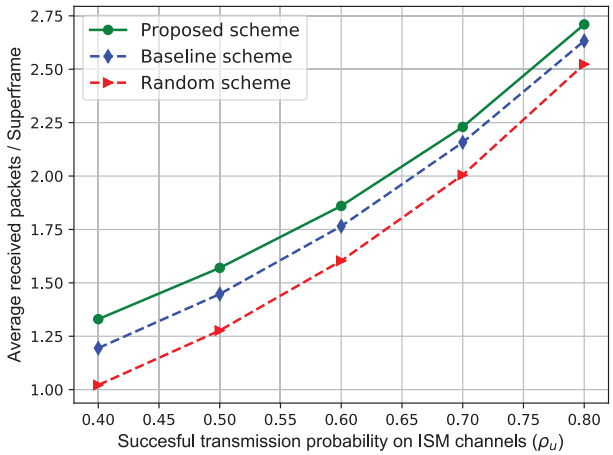

FIGURE 8. Received packets versus the successful packet reception ratio on the ISM channels.

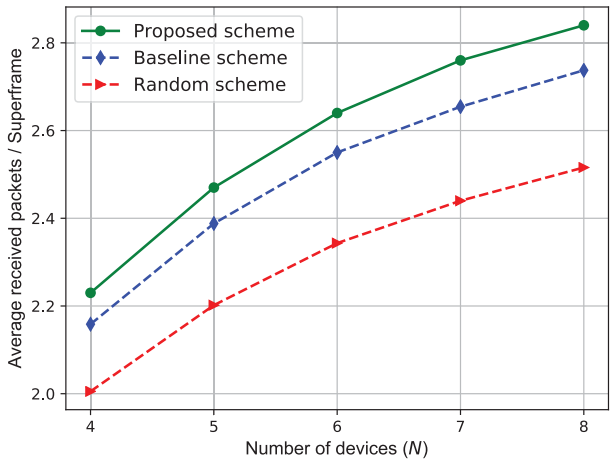

FIGURE 9. Received packets according to the number of devices.

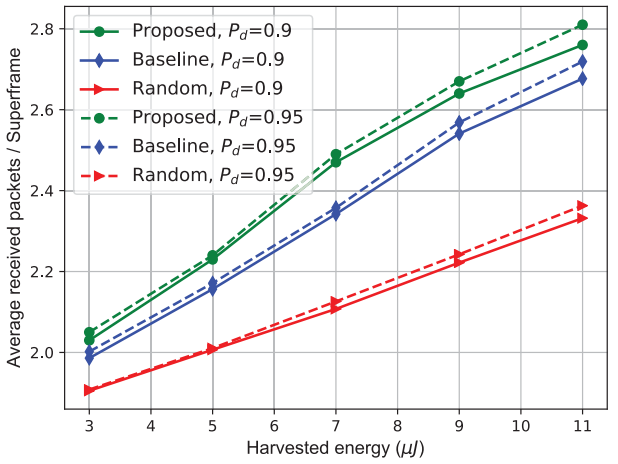

FIGURE 10. Received packets versus the mean value of harvested energy.

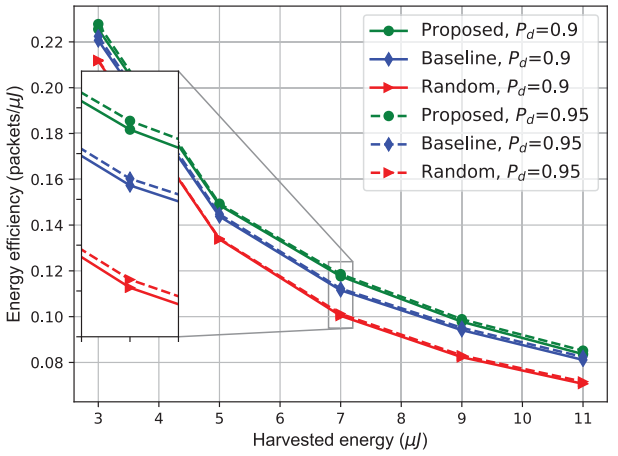

FIGURE 11. Energy efficiency versus the mean value of harvested energy.

when ρ_m was low, but the improvement using the proposed scheme became smaller as ρ_m was high. It is because the low interference on the ISM channels affects the data aggregation performance slightly. The system tends to frequently utilize ISM channels with a higher successful transmission probability than the primary channels with a lower belief. Thus, using primary channels when the reliability of the ISM channels is high will not greatly improve the throughput. Figure 9 shows the network performance of the schemes versus the number of devices employed in the network. The curves showed that the total number of packets generated in field devices became larger when the network size was increased, leading to the higher received packets at the GW.

In Figure 10, we plot the received packets according to the mean value of harvested energy of the CAPs along with different values of detection probability, $P_d = 0.9$ and $P_d = 0.95$. The number of packets received at the CAP can be enhanced gradually as the harvested energy increases. This is because the CAPs have more chance to sense the primary channels. Accordingly, the devices can frequently share the primary channels with primary users as they are sensed to be free. Moreover, correctly detecting the actual state "active" of the primary channels (i.e., with the high value of P_d) can help the system lower number of transmission collisions between the field devices and primary users. This results in a throughput improvement when detection probability increases. Therefore, the proposed scheme always offers higher packets received at the GW than the others.

Figure 11 plots to verify the energy efficiency of the proposed scheme according to the various harvested energy and the detection probability. In this article, the energy efficiency is defined as the average received packets over the average amount of utilized energy of CAPs. It is observed that when the value of P_d is large, the harvested energy can be utilized better by all schemes. For the various scenarios of $E^{h,mean}$ and P_d, the proposed scheme could efficiently utilize the harvested energy, compared to the others. The reason is that the proposed scheme not only considers the long-term reward with energy consumption at CAPs, but also updates the belief of primary channels using P_d to select the optimal scheduling in each superframe.

Generally, in spectrum sensing, when the sensing parameters are fixed, an increase in P_d will result in an increase

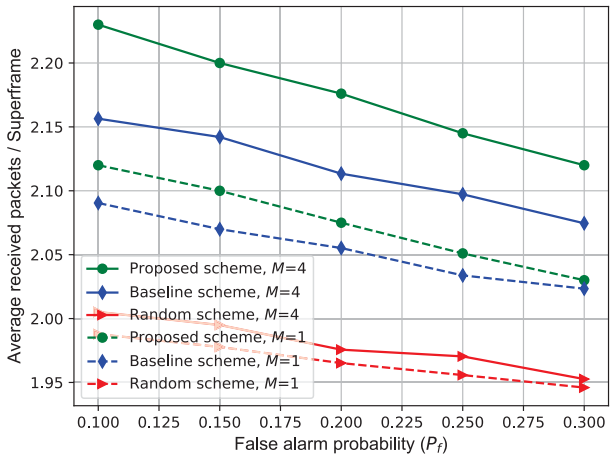

FIGURE 12. Received packets according to the false alarm probability.

in P_f. Therefore, the average packet delivery was examined according to the different probability of the false alarm and the various number of primary channels to evaluate the performance of the proposed scheme compared to the others, as shown in Figure 12. As can be seen from the figure, the false alarm can significantly deteriorate the received packets at the GW as P_f increases, while increasing the number of primary channels can help improve the throughput of the system. The reason for the degradation due to the false alarm is that the devices might miss many opportunities for their transmissions on the primary channels as P_f is large, which leads to poor transmission performance. Therefore, the sensing error is one of the key factors to consider when designing a transmission scheme for joint ISM/primary channel allocation in the network. On the other hand, the devices can have more opportunities of using multiple primary channels when M increases. From the presented simulation, the proposed scheme can outperform the traditional schemes under various network parameters since it not only consider the current reward, but also the future reward to obtain maximum long-term throughput. Moreover, the proposed scheme can be verified to work efficiently in industrial CR networks with spectrum sensing errors by considering spectrum factors such as, detection and false alarm probabilities.

We close this section by providing a concrete scenario that the proposed scheme could be effectively applied. Let us consider a big warehouse of a factory where the inside temperature and air quality should be monitored and controlled precisely. Since the placement of sensor nodes may be changed frequently based on the re-arrangement of the obstacles (e.g. goods and furniture) inside, it is quite challenging to deploy the conventional wired sensor networks such as the SCADA system. For this circumstance, the wireless monitoring and controlled sensor networks could be an effective option and the communication protocol could be the WirelessHART, in which data transmissions are on the ISM bands. Unfortunately, the increasing use of microelectronics devices and the attraction of unlicensed use have been leading to an overload in ISM bands recently. Since the proposed scheme allows the sensor nodes or field devices to switch between ISM channels and licensed channels for their transmissions, we can apply the proposed network settings to attain the higher reliability of data transmissions between sensor nodes. Furthermore, the wireless deployment flexibility of the field devices and solar-powered CAPs in our network setting help lower the long-term operating cost, especially when the arrangement of machines or products is frequently changed in the manufacturing facilities.

VII. CONCLUSION

In this article, we proposed deep reinforcement learning-based transmission scheduling for joint ISM/primary channel allocation to devices to maximize the throughput of the linear convergecast network under constraints of the required number of ISM channels and delay. By considering the long-term reward, the system can select the optimal scheduling policy for field device transmissions through each superframe under the awareness of limited energy in CAPs and the dynamics of the primary channels. As a result, the maximum number of packets received at the GW was obtained with a minimum delay and number of ISM channels through a trial-and-error action of the GW after training. The proposed method was assessed by comparing the system performance of the proposed scheme with those of other traditional schemes where the context of long-term reward maximization was not considered. The numerical simulation results were presented to verify the effectiveness of the proposed scheme under the various network parameters. From the simulation, the agent in the proposed approach can adapt its policy to network variations in terms of the harvested energy, successful packet reception ratio on the ISM channels, number of devices, probability of detection, and false alarm. Thus, a greater reward was obtained compared to the others. As a result, the optimal long-term throughput of the WirelessHART system can be guaranteed with energy-efficient utilization.

REFERENCES

[1] M. C. Lucas-Estan, B. Coll-Perales, and J. Gozalvez, "Redundancy and diversity in wireless networks to support mobile industrial applications in industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 311–320, Jan. 2021.

[2] S. Zoppi, A. V. Bemten, H. M. Gürsu, M. Vilgelm, J. Guck, and W. Kellerer, "Achieving hybrid wired/wireless industrial networks with WDetServ: Reliability-based scheduling for delay guarantees," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 2307–2319, May 2018.

[3] T.-Y. Huang, C.-J. Chang, C.-W. Lin, S. Roy, and T.-Y. Ho, "Delay-bounded intravehicle network routing algorithm for minimization of wiring weight and wireless transmit power," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 4, pp. 551–561, Apr. 2017.

[4] Y. Nakayama, K. Maruta, T. Tsutsumi, and K. Sezaki, "Wired and wireless network cooperation for wide-area quick disaster recovery," *IEEE Access*, vol. 6, pp. 2410–2424, 2018.

[5] (2007). *WirelessHART Specifications*. [Online]. Available: http://www.hartcomm2.org

[6] D. Gunatilaka and C. Lu, "Conservative channel reuse in real-time industrial wireless sensor-actuator networks," in *Proc. ICDCS*, 2018, pp. 344–353.

[7] M. K. Ehsan, "Performance analysis of the probabilistic models of ISM data traffic in cognitive radio enabled radio environments," *IEEE Access*, vol. 8, pp. 140–150, 2020.

[8] Spectrum Policy Task Force, "Federal communications commission," Tech. Rep. ET Docket 02-135, Nov. 2002.

[9] A. Mustafa, M. N. U. Islam, and S. Ahmed, "Dynamic spectrum sensing under crash and byzantine failure environments for distributed convergence in cognitive radio networks," *IEEE Access*, vol. 9, pp. 23153–23167, 2021.

[10] G. I. Tsiropoulos, O. A. Dobre, M. H. Ahmed, and K. E. Baddour, "Radio resource allocation techniques for efficient spectrum access in cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 824–847, 1st Quart., 2014.

[11] T. M. Chiwewe, C. F. Mbuya, and G. P. Hancke, "Using cognitive radio for interference-resistant industrial wireless sensor networks: An overview," *IEEE Trans. Ind. Informat.*, vol. 11, no. 6, pp. 1466–1481, Dec. 2015.

[12] S. Demirci and D. Gozupek, "Switching cost-aware joint frequency assignment and scheduling for industrial cognitive radio networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4365–4377, Jul. 2020.

[13] M. Liu, L. Liu, H. Song, Y. Hu, Y. Yi, and F. Gong, "Signal estimation in underlay cognitive networks for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5478–5488, Aug. 2020.

[14] G. Kakkavas, K. Tsitseklis, V. Karyotis, and S. Papavassiliou, "A software defined radio cross-layer resource allocation approach for cognitive radio networks: From theory to practice," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 740–755, Jun. 2020.

[15] Y.-H. You and J.-H. Paik, "Suboptimal maximum likelihood detection of integer carrier frequency offset for digital terrestrial television broadcasting system," *IEEE Trans. Broadcast.*, vol. 66, no. 1, pp. 195–202, Mar. 2020.

[16] W. U. Mondal, A. A. Sardar, N. Biswas, and G. Das, "Nash bargaining-based economic analysis of opportunistic cognitive cellular networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 242–255, Mar. 2020.

[17] C. C. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. J. Shellhammer, and W. Caldwell, "IEEE 802.22: The first cognitive radio wireless regional area network standard," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 130–138, Jan. 2009, doi: 10.1109/MCOM.2009.4752688.

[18] H. Zhang, P. Soldati, and M. Johansson, "Optimal link scheduling and channel assignment for convergecast in linear WirelessHART networks," in *Proc. 7th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw.*, Jun. 2009, pp. 1–8.

[19] H. Zhang, P. Soldati, and M. Johansson, "Performance bounds and latency-optimal scheduling for convergecast in WirelessHART networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2688–2696, Jun. 2013.

[20] K. Dang, J. Shen, and L. Dong, "A graph route-based superframe scheduling scheme in WirelessHART mesh networks for high robustness," *Wireless Pers. Commun.*, vol. 71, pp. 2431–2444, Apr. 2013.

[21] R. Tavakoli, M. Nabi, T. Basten, and K. Goossens, "Topology management and TSCH scheduling for low-latency convergecast in in-vehicle WSNs," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1082–1093, Jul. 2019, doi: 10.1109/TII.2018.2853986.

[22] S. Wang, S. M. Kim, L. Kong, and T. He, "Concurrent transmission aware routing in wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6275–6286, Dec. 2018.

[23] L. Yunhuan, C. Cailian, X. Honghua, Y. Baoqing, and G. Xinping, "A cognitive radio based reliability optimization for industrial wireless DSSS/CH transmission links," in *Proc. 31st Chin. Control Conf.*, Hefei, China, 2012, pp. 5542–5547.

[24] L. Lyu, C. Chen, Y. Li, F. Lin, L. Liu, and X. Guan, "Cognitive radio enabled transmission for state estimation in industrial cyber-physical systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[25] P. Yu, M. Yang, A. Xiong, Y. Ding, W. Li, X. Qiu, L. Meng, M. Kadoch, and M. Cheriet, "Intelligent-driven green resource allocation for industrial Internet of Things in 5G heterogeneous networks," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 520–530, Jan. 2022.

[26] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1823–1834, May 2010.

[27] W. Lu, P. Si, G. Huang, H. Han, L. Qian, N. Zhao, and Y. Gong, "SWIPT cooperative spectrum sharing for 6G-enabled cognitive IoT network," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15070–15080, Oct. 2021, doi: 10.1109/JIOT.2020.3026730.

[28] M. Hamza, M. U. Rehman, A. Riaz, Z. Maqsood, and W. T. Khan, "Hybrid dual band radio frequency and solar energy harvesting system for making battery-less sensing nodes," in *Proc. IEEE Radio Wireless Symp. (RWS)*, Jan. 2021, pp. 116–118.

[29] H. T. H. Giang, T. N. K. Hoan, and I. Koo, "Uplink NOMA-based long-term throughput maximization scheme for cognitive radio networks: An actor–critic reinforcement learning approach," *Wireless Netw.*, vol. 27, no. 2, pp. 1319–1334, Feb. 2021.

[30] D. Ohmann, A. Awada, and I. Viering, "SINR model with best server association for high availability studies of wireless networks," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 60–63, Oct. 2016.

[31] J.-Y. Choi, J. Park, S.-H. Lim, and Y.-B. Ko, "A RSSI-based mesh routing protocol based IEEE 802.11p/WAVE for smart pole networks," in *Proc. 23rd Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2021, pp. 104–108.

[32] G. Chen, R. Ma, M. Lei, and X. Cao, "Channel list selection based on quality prediction in WirelessHART networks," *J. Commun. Inf. Netw.*, vol. 3, no. 3, pp. 49–56, Sep. 2018.

[33] M. Raza, N. Aslam, H. Le-Minh, S. Hussain, Y. Cao, and N. M. Khan, "A critical analysis of research potential, challenges, and future directives in industrial wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 39–95, 1st Quart., 2018.

[34] D. Chen, M. Nixon, and A. W. Mok, *Real-Time Mesh Network for Industrial Automation*. Cham, Switzerland: Springer, Apr. 2010, pp. 1–6.

[35] H. Sawada, A. Nakajima, K. Yamamoto, and S. Suzuki, "Signaling protocols using mobile application part for call control in the digital mobile network," in *Proc. IEEE Global Telecommun. Conf. Exhib.*, 1990, pp. 1569–1573, doi: 10.1109/GLOCOM.1990.116754.

[36] N. Umeda and S. Onoe, "Design and performance evaluation of novel common control channels for digital mobile radio," in *Proc. 41st IEEE Veh. Technol. Conf.*, Dec. 1991, pp. 646–651, doi: 10.1109/VETEC.1991.140573.

[37] H. Chen, X. Li, and F. Zhao, "A reinforcement learning-based sleep scheduling algorithm for desired area coverage in solar-powered wireless sensor networks," *IEEE Sensors J.*, vol. 16, no. 8, pp. 2763–2774, Sep. 2016.

[38] I. Mathews, P. J. King, F. Stafford, and R. Frizzell, "Performance of III–V solar cells as indoor light energy harvesters," *IEEE J. Photovolt.*, vol. 6, no. 1, pp. 230–235, Jan. 2016.

[39] M. Miantezila, B. Guo, C. Zhang, and X. Bai, "Primary user channel state prediction based on channel allocation and DBHMM," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2020, pp. 335–338.

[40] S. P. Sheng, M. Liu, and R. Saigal, "Data-driven channel modeling using spectrum measurement," *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1794–1805, Sep. 2015.

[41] D. Gozupek, S. Buhari, and F. Alagoz, "A spectrum switching delay-aware scheduling algorithm for centralized cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1270–1280, Jul. 2013.

[42] S. Q. Jalil, S. Chalup, and M. H. Rehmani, "Cognitive radio spectrum sensing and prediction using deep reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[43] A. A. Olawole, F. Takawira, and O. O. Oyerinde, "Cooperative spectrum sensing in multichannel cognitive radio networks with energy harvesting," *IEEE Access*, vol. 7, pp. 84784–84802, 2019.

[44] P. Lee, Z. A. Eu, M. Han, and H.-P. Tan, "Empirical modeling of a solar-powered energy harvesting wireless sensor node for time-slotted operation," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2011, pp. 179–184.

[45] J. Rodway and P. Musilek, "Harvesting-aware energy management for environmental monitoring WSN," in *Proc. IEEE 16th Int. Conf. Environ. Electr. Eng. (EEEIC)*, Jun. 2016, pp. 1–6.

[46] H. O. Kpojime and G. A. Safdar, "Interference mitigation in cognitive-radio-based femtocells," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1511–1534, 3rd Quart., 2015, doi: 10.1109/COMST.2014.2361687.

[47] P. D. Thanh, T. N. K. Hoan, and I. Koo, "Joint resource allocation and transmission mode selection using a POMDP-based hybrid half-duplex/full-duplex scheme for secrecy rate maximization in multi-channel cognitive radio networks," *IEEE Sensors J.*, vol. 20, no. 7, pp. 3930–3945, Apr. 2020.

[48] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, 2016, pp. 1–21.

[49] J.-T. Chien and P.-Y. Hung, "Multiple target prediction for deep reinforcement learning," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2020, pp. 1611–1616.

PHAM DUY THANH received the B.E. degree in electronics and telecommunications engineering from Ton Duc Thang University, Vietnam, in 2013, the M.S. degree from the Graduate Institute of Digital Mechatronic Technology, College of Engineering, Chinese Culture University, Taiwan, in 2015, and the Ph.D. degree in electrical, electronic and computer engineering from the University of Ulsan, South Korea, in 2021. His current research interests include reinforcement learning, cognitive radio networks, wireless security, and next generation wireless communication systems.

TRAN NHUT KHAI HOAN received the B.E. degree in electronics engineering from Can Tho University, Can Tho, Vietnam, in 2002, the M.E. degree in electronics engineering from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2008, and the Ph.D. degree in electrical engineering from the University of Ulsan (UOU), South Korea, in 2018. His research interests include cognitive radio and next generation wireless communication networks.

HOANG THI HUONG GIANG received the B.E. degree in electronics and telecommunications engineering from Ton Duc Thang University, Vietnam, in 2013, the M.S. degree from the Graduate Institute of Digital Mechatronic Technology, College of Engineering, Chinese Culture University, Taiwan, in 2015, and the Ph.D. degree in electrical, electronic and computer engineering from the University of Ulsan, South Korea, in 2021. Her current research interests include NOMA, RIS, reinforcement learning, and deep learning in wireless communications.

INSOO KOO received the B.E. degree from Konkuk University, Seoul, South Korea, in 1996, and the M.S. and Ph.D. degrees from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 1998 and 2002, respectively. From 2002 to 2004, he was with the Ultrafast Fiber-Optic Networks Research Center, GIST, as a Research Professor. In 2003, he was a Visiting Scholar with the Royal Institute of Science and Technology, Stockholm, Sweden. In 2005, he joined the University of Ulsan, South Korea, where he is currently a Full Professor. His current research interests include next generation wireless communication systems and wireless sensor networks.