AoI-minimal Power Adjustment in RF-EH-powered Industrial IoT Networks: A Soft Actor-Critic-Based Method

Yiyang Ge, Ke Xiong, *Member, IEEE,* Qiong Wang, Qiang Ni *Senior Member, IEEE,* Pingyi Fan *Senior Member, IEEE,* and Khaled Ben Letaief, *Fellow, IEEE*

Abstract—This paper investigates the radio-frequency-energy-harvesting-powered (RF-EH-powered) wireless Industrial Internet of Things (IIoT) networks, where multiple sensor nodes (SNs) are first powered by a wireless power station (WPS), and then collect status updates from the industrial environment and finally transmit the collected data to the monitor with their harvested energy. To enhance the timeliness of data, age of information (AoI) is used as a metric to optimize the system. Particularly, an expected sum AoI (ESA) minimization problem is formulated by optimizing the power adjustment policy for the SNs under multiple practical constraints, including the EH, the minimal signal-to-noise-plus-interference ratio (SINR) and the battery capacity constraints. To solve the non-convex problem with no explicit AoI expression, we transform it into a Markov decision problem (MDP) with continuous state space and action space. Then, inspired by the Soft Actor-Critic (SAC) framework in deep reinforcement learning, a SAC-based age-aware power adjustment (SAPA) method is proposed by modeling the power adjustment as a stochastic strategy. Furthermore, to reduce the communication overhead of SAPA, a multi-agent version of SAPA, i.e., MSAPA, is proposed, with which each SN is able to adjust its transmit power based on its local observations. The communication overhead of SAPA and MSAPA is also analyzed theoretically. Simulation results show that the proposed SAPA and MSAPA converge well with different numbers of SNs. It is also shown that the ESA achieved by the proposed SAPA and MSAPA is lower than that achieved by the baseline methods.

Index Terms—Energy Harvesting (EH), power adjustment, Age of Information (AoI), Soft actor critic (SAC), multiple agent.

1 Introduction

1.1 Background

With its strong capacity in terms of sensing and transmission, the industrial Internet of Things (IIoT) is regarded as the key technology to facilitate Industry 4.0 [1], [2], [3], where numerous sensors, controllers, execution units and monitors are interconnected to establish smart systems for monitoring, collecting, exchanging and analyzing data for intelligent control. Conventionally, network latency and throughput were adopted as important performance indices to design IIoT systems for applications. Recently, information timeliness has become a novel and very critical metric for the newly emerged status update applications, including hazard monitoring, autonomous driving, and AR services, because outdated information may result in serious impacts, such as production accidents and economic loss or even casualties. In order to enhance the information timeliness for IIoT, age of information (AoI), which is defined as the time interval from the time a data packet is generated to the current time, has been presented and widely studied in various systems network scenarios [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15].

Meanwhile, in many IIoT systems, to effectively monitor environmental indicators, including temperature, humidity, and hazardous substance content, massive sensor nodes (SNs) equipped with small batteries are deployed. Since powering the SNs by wire or batteries often requires a relatively high deployment and maintenance cost, wireless energy harvesting (EH), which enables SNs to harvest energy from radio frequency (RF) signals, has been widely regarded as a promising solution to power low-power SNs wirelessly, due to its advantages in controllability [16], [17], [18], [19].

1.2 Related work

In order to simultaneously release the energy supply issue and also meet the information timeless demands, AoI-

- *Yiyang Ge and Ke Xiong are with the Engineering Research Center of Network Management Technology for High-Speed Railway of Ministry of Education, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, with the Collaborative Innovation Center of Railway Traffic Safety, Beijing Jiaotong University, Beijing 100044, China and also with the National Engineering Research Center of Advanced Network Technologies, Beijing Jiaotong University, Beijing 100044, China. E-mail: yiyangge{kxiong}@bjtu.edu.cn*
- *Qiong Wang is with the State Grid Beijing Electric Power Company, No. 41, Qianmen West Street, 100031, Beijing, China. E-mail: wangqiong@bj.sgcc.com.cn.*
- *Qiang Ni is with the School of Computing and Communications and Data Science Institute, Lancaster University, Lancaster LA1 4WA, U.K. E-mail: q.ni@lancaster.ac.uk.*
- *Pingyi Fan is with the Beijing National Research Center for Information Science and Technology, and with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: fpy@tsinghua.edu.cn.*
- *Khaled Ben Letaief is with the Department of Electrical and Computer Engineering, Hong Kong University of Science and Technology (HKUST), Hong Kong, and also with the Pengcheng Laboratory, Shenzhen 518055, Guangdong, China. E-mail: eekhaled@ust.hk.*

Manuscript received 26 December 2022; revised 30 September 2022; accepted 4 January 2024.
This work was supported in part by the National Natural Science Foundation of China under Grant 62071033 and also in part by the Changping Innovation Joint Fund of Beijing Natural Science Foundation under Grant L234084. (Corresponding author: Ke Xiong.)

oriented RF-EH-powered wireless network design has attracted increasing attention [20], [21], [22], [23], [24], [25], [26], [27], where RF-EH is employed to wirelessly charge the low-power wireless sensors and the AoI is adopted as a performance metric for deigning the IIoT. In [20], the average AoI of the RF-EH network was analyzed and minimized by optimizing the capacitor's size of SNs. In [21], the age-energy region and age-energy function were explored, where the average AoI of the single-SN RF-EH network was minimized. In [22], the urgency-aware AoI was minimized by optimizing transmission policy under the constraint of energy causality. In [23], the long-term average AoI was minimized by optimizing the online sampling policy. In [24], the average AoI and the peak AoI (PAoI) of the RF-EH-powered network were analyzed, where the relay and SN transmit in a cooperative way. In [25], the AoI-energy utility of the hybrid access point IoT device pair was maximized in RF-EH networks. In [26], the bounds of the average AoI of a multi-user cognitive radio RF-EH-powered network were analyzed. In [27], the average AoI was minimized in the unmanned aerial vehicle (UAV)-assisted multi-SN wireless powered IoT system.

It is noticed that all aforementioned works optimized the system's AoI performance by using traditional mathematical methods such as convex optimization theory, game theory, and heuristic algorithms. Since in most network optimization problems, multiple variables are mathematically coupled together, making the problems non-convex and difficult to solve, traditional convex optimization methods thus cannot be used directly to get good optimized solutions. Another challenge is that the explicit expressions of AoI are usually hard to model, which further enhances the difficulty of solving the optimization problem.

Therefore, recently, reinforcement learning (RL), as sparking instances of artificial intelligence, has been introduced to optimally design AoI-oriented RF-EH-power IoT networks, see e.g., [28], [29], [30], [31], [32]. In [28], the long-term on-demand AoI on SNs in a cache-enabled RF-EH-powered network was minimized by optimizing the caching and transmitting policy based on a Q-learning method. In [29], the average AoI of SNs in a UAV-assisted RF-EH-powered network was minimized by optimizing the trajectory of the UAV and the scheduling of SNs based on a deep Q-network (DQN) method. In [30], the average AoI of SNs and energy consumption of the UAV-assisted RF-EH-powered network were jointly minimized by optimizing the trajectory of multiple UAVs with a multi-agent DQN-based method. In [31], the long-term average weighted sum of AoI in a multiple-SN RF-EH-powered network was minimized by optimizing the scheduling policy with a DQN-based method. In [32], the average AoI in a multiple-SN RF-EH-powered network was minimized via optimizing the SNs' selection by a distributed Q-Learning method without knowing the battery and channel state of SNs.

1.3 Motivations and Contributions

In this paper, we study the AoI-aware RF-EH-powered networks with multiple SNs to simultaneously release the transmission and timeliness issues for IIoT, where multiple SNs are first powered by a WPS, then collect status updates from the industrial environment and finally transmit the collected data to the monitor with their harvested energy from the RF signals. Different from previous works [28], [29], [30], [31], [32], which mainly studied the optimization of the scheduling of SNs, this paper focuses on improving the AoI performance of the system via power adjustment of SNs because in practical industrial monitoring scenarios, scheduling SNs in the time domain requires high synchronization among SNs, which is hard to realize [33]. Therefore, in this paper, all SNs are allowed to share the same spectrum resource and transmit their collected status updates simultaneously rather than in a TDMA manner. By doing so, strict synchronization requirement among SNs is avoided, but the interference among SNs cannot be neglected. Thus, in order to make the simultaneous transmission of SNs more efficient, we optimize the power adjustment of SNs to coordinate the interference among SNs and enable them to utilize their harvested energy more efficiently. As a result, the system AoI performance is further improved. The main contributions of this paper are summarized as follows.

1) In order to improve the freshness of information in an RF-EH-powered network with multiple SNs, the expected sum AoI (ESA) of the system is analyzed and modeled, and then an optimization problem is formulated to minimize the ESA via optimizing the power adjustment at SNs, where the EH, the minimal signal-to-noise-plus-interference ratio (SINR) and the battery constraints are jointly taken into account. Since all SNs share the same spectrum resource to transmit status updates simultaneously, the interference among SNs is also taken into account.
2) Due to no explicit expression of the ESA and the formulated problem being non-convex, we transform it into a Markov Decision Process (MDP) problem. Moreover, as the formulated power adjustment problem is with continuous action space, we develop a Soft Actor-Critic (SAC) based Age-aware power adjustment (SAPA) method to optimize the power adjustment policy.
3) To reduce the communication overhead of the network, we further design a multi-agent version of SAPA (MSAPA), where SNs adjust the transmitting power according to the local observation. The communication overheads of SAPA and MSAPA are also analyzed theoretically. It shows that MSAPA is with no interaction overhead when deployed.
4) Simulation results demonstrate the convergence of the proposed SAPA and MSAPA. It also shows that the ESA achieved by SAPA and MSAPA is lower than that achieved by conventional DRL methods. Additionally, the effects of the sensing probability of SNs and the SINR threshold on the system ESA are also discussed, which indicates that MSAPA is more suitable for the network with a higher SINR threshold to achieve a better AoI performance.

The rest of the paper is organized as follows. Section II describes the network model and the ESA minimization problem is formulated. In Section III, the problem is transformed into an MDP. In Section IV, SAPA is proposed. In

Section V, the multi-agent version of SAPA, i.e., MSAPA, is designed. Section VI provides the simulation results and Section VII summarizes the paper.

2 SYSTEM MODEL AND PROBLEM FORMULATION

2.1 Network Model

We consider an RF-EH-powered monitoring system for IIoT, as shown in Fig. 1, which consists of a WPS, a monitor, and a set of SNs denoted as $\mathcal{M} = \{1, 2, \ldots, M\}$. In order to reduce the cost of manual battery replacement in industrial scenarios, the WPS is deployed to charge the SNs wirelessly by transmitting RF signals. The SNs are equipped with EH circuit modules, so they are capable of harvesting energy from the received RF signals and then use the harvested energy to upload the time-sensitive data collected from the industrial environment to the monitor. Without loss of generality, the time is discretized into small blocks with a small duration of τ. The block fading channel model is considered, so the channel gain is regarded unchanged in each τ, but it may change from one block to the next. Let $h_i(t)$ and $g_i(t)$ denote the downlink channel gain of the link from the WPS to SN i and the uplink channel gain of the link from SN i to the monitor in block t, respectively. As the links are spatially separated, their channel gains are regarded to be independent and identically distributed (i.i.d.).

To realize intelligent industrial monitoring, the monitor is integrated with a computing server, so it not only monitors the industrial environment but is also able to enhance the monitoring through its intelligent computing unit. Each SN is equipped with a battery with a capacity of $B_{\max}$ joules and it has an energy-receiving antenna and an information-transmitting antenna. In order to avoid interference between the power signals and information signals, the energy transfer and the information transfer are performed over orthogonal frequency bands.Moreover, all SNs are allowed to transmit status update packets to the monitor over the same frequency band at the same time. Thus, the inter-user interference cannot be neglected. To enhance the transmission of data, power adjustment among SNs should be employed. For clarity, the time frame structure is illustrated in Fig. 2, where each SN is able to harvest energy from WPS and transmit the sensed data to the monitor via the uplink in the same block. Each time interval is divided into two parts. The first part takes up a relatively very short time and is used for all SNs to send the feedback on their local state information in sequence for determining the power adjustment. The second part takes up a relatively much longer time and is used for SNs to transmit sensed data. More specifically, in the first part, time is divided into many tiny fixed time slots based on the number of SNs, with only one SN transmitting its state information per time slot. Due to the orthogonal channels, state information can be transmitted correctly without interference. In the second part, the power adjustment methods proposed in this paper are adopted to enhance the SN's transmission of the sensed data[1].

1. Since the first part of time used for local states feedback by SNs is much smaller than the second part of time used for sensed data transmission, the second part is roughly approximated by τ.

Similar to [34], packets arrive at SNs according to the Bernoulli distribution with probability p, based on the monitored industrial environment. More specifically, for each SN, a state update packet is arrived with a probability p at the beginning of each block. The packets are buffered, so the old one in the buffer will be replaced with the new one once a new packet is arrived.

To charge the SNs, the WPS broadcasts RF signals to them continuously. To be practical, the piecewise non-linear EH model in [25] is adopted to characterize the EH operation. Denoting the transmit power of the WPS as P_{W}, the energy harvested by SN i in block t is given by

$$E_i^{\mathrm{EH}}(t) = \min\{\eta\tau P_{\mathrm{W}} h_i(t), E_{\max}^{\mathrm{EH}}\}, \tag{1}$$

where $\eta \in (0, 1)$ is the EH efficiency coefficient and $E_{\max}^{\mathrm{EH}}$ is the maximum energy harvested in one block based on the EH circuit of SNs. $h_i(t)$ is given by $h_i(t) = |\varphi_i(t)|^2 L_i^{\mathrm{DL}}$, where $\varphi_i(t)$ is the small-scale fading gain and $L_i^{\mathrm{DL}}{=}{d_{\mathrm{W},i}}^{-\alpha}$ is the path-loss coefficient between the WPS and SN i with $d_{\mathrm{W},i}$ being the distance between the WPS and SN i and α being the pass-loss factor.

Denoting x_i as the signal transmitted by SN i with x_i being an i.i.d. random variable with zero mean and unit variance, the signal received by the monitor in block t is expressed as $y = \sum_{i=1}^{M} \sqrt{g_i(t)P_i(t)}x_i + n$, where $P_i(t)$ is the real transmission power at SN i in block t and n is the additive white Gaussian noise (AWGN) at the monitor with variance σ^2. $g_i(t)$ is given by $g_i(t) = |\upsilon_i(t)|^2 L_i^{\mathrm{UL}}$, where $\upsilon_i(t)$ is the small-scale fading gain and $L_i^{\mathrm{UL}} = {d_{i,\mathrm{M}}}^{-\alpha}$ is the path-loss coefficient between SN i and the monitor. Thus, the received SINR at the monitor is

$$\gamma_i = \frac{g_i(t)P_i(t)}{\sum_{j\neq i} g_j(t)P_j(t) + \sigma^2}. \tag{2}$$

To successfully decode x_i at the monitor, it should satisfy that

$$\gamma_i \geq \gamma_{\mathrm{th}}, \tag{3}$$

where γ_{th} is the minimal required received SINR for successfully decoding the data at the monitor.

Denote the battery level of SN i in block t as $B_i(t)$, and SN's maximum permitted transmit power as $P_{\max}$. The real average transmit power of SN i in block t, i.e., $P_i(t)$, satisfies that

$$P_i(t) \leq \min\{B_i(t)/\tau, P_{\max}\}. \tag{4}$$

Denoting $E_i^{\mathrm{C}}(t)$ as the energy comsumed in block t, we have that $E_i^{\mathrm{C}}(t) = \tau P_i(t)$. Since the SN i is equipped with a battery of limited capacity $B_{\max}$, after the EH operation and the data transmission in block t, the remaining energy in the battery at SN i at the beginning of block $(t+1)$ is

$$B_i(t+1) = \min\{B_i(t) + E_i^{\mathrm{EH}}(t) - E_i^{\mathrm{C}}(t), B_{\max}\}. \tag{5}$$

2.2 AoI Model

In industrial applications, once the system AoI reaches the maximum tolerant value, it means that no new data has been received for a long time and the measurement associated with the monitoring becomes stale, so further collecting and counting make no difference [28], [29], [31]. Therefore, denoting the maximum tolerant value of the AoI with $A_{\max}$, the AoI at the SNs and that at the monitor is discussed as follows.

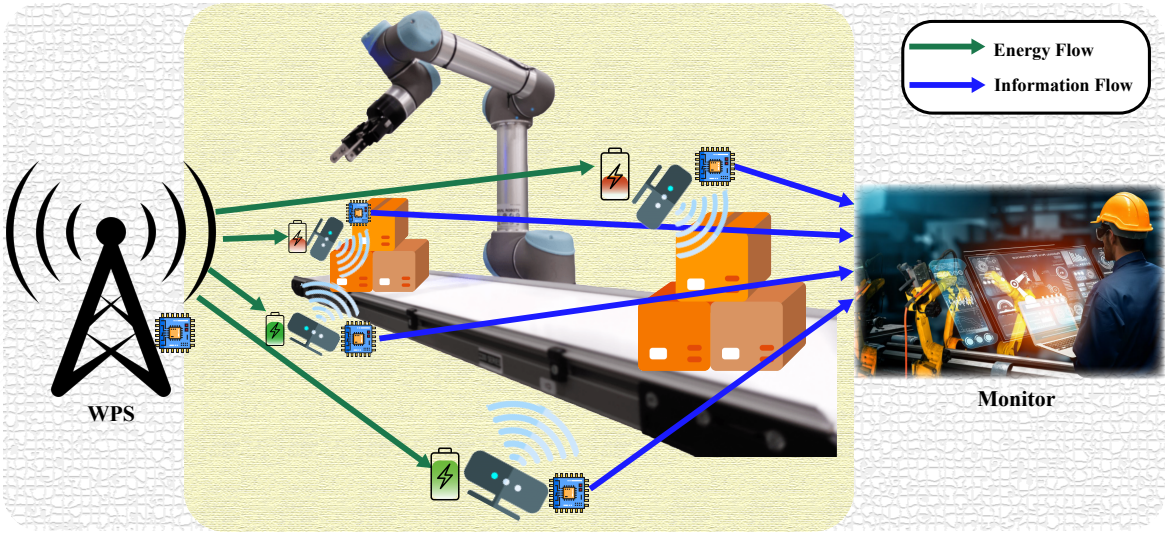


Fig. 1: The network model of the RF-EH-powered IIoT system.

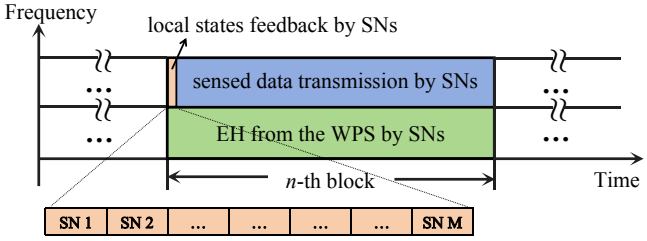


Fig. 2: The illustration of the time frame structure of the system.

2.2.1 AoI at SNs

In block t, the AoI of SN i will be reset to 1 if a new data packet arrives at SN; otherwise, it will be increased by one unit. Denoting $A_i^S(t)$ as the AoI of SN i in block t, one have that

$$A_i^S(t) = \begin{cases} 1, \text{ if } r_i(t) = 1, \\ \min(A_i^S(t-1)+1, A_{\max}), \text{ otherwise}, \end{cases} \quad (6)$$

with $r_i(t) \in \{0,1\}$ being the data arriving indicator, where $r_i(t) = 1$ means that a new data packet arrives at SN i and $r_i(t) = 0$ means that it doesn't.

2.2.2 AoI at the monitor

In block t, if SN i successfully transmits data to the monitor, the AoI of the data associated with SN i at the monitor will be decreased to be the value of the AoI of the newly received one. Otherwise, the AoI of the data associated with SN i at the monitor will be increased by one unit. At the end of block t, if the monitor successfully receives data from SNs, it will instantly broadcast an acknowledge (ACK) feedback message to all SNs. Denoting $A_i^D(t)$ as the AoI of SN i at the monitor in block t, SN i can easily acquire its $A_i^D(t)$ from the feedback message by a simple counting. Therefore, one have that

$$A_i^D(t) = \begin{cases} A_i^S(t), \text{ if } ack_i(t) = 1, \\ \min(A_i^D(t-1)+1, A_{\max}), \text{ otherwise}, \end{cases} \quad (7)$$

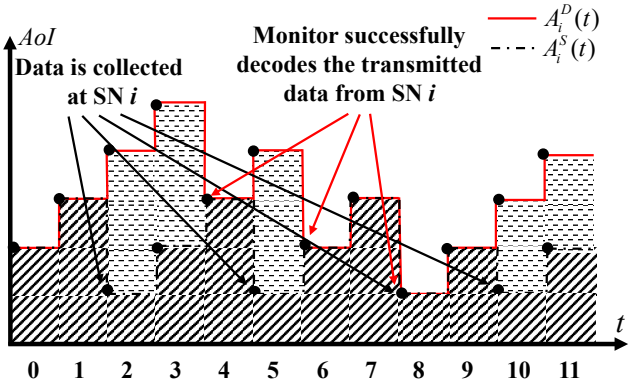


Fig. 3: AoI evolution of SN i.

with $ack_i(t) \in \{0,1\}$ being the feedback ACK indicator, where $ack_i(t) = 1$ means that the SN i successfully transmits the data packet to the monitor and constraint (3) is satisfied and $ack_i(t) = 0$ means that it doesn't.

For clarity, take the AoI evolution of SN i as an example, which is illustrated in Fig. 3. For the infinite horizon observation, denoting $\bar{\Delta}$ as the long-term ESA of the system, one have that

$$\bar{\Delta} = \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{M} \omega_i A_i^D(t) \,\middle|\, \mathbf{A}^D(0)\right], \quad (8)$$

where and $\mathbb{E}[\cdot]$ denotes the symbol that takes the expectation, $\mathbf{A}^D(0) = (A_0^D(0), A_1^D(0), \ldots, A_M^D(0))$ is the vector representing the initial AoI at the monitor, and ω_i denotes the weight factor of SN i. The larger the value of ω_i, the more important the timeliness of SN i in the network.

2.3 Problem Formulation

One can observe in Fig. 3 that the AoI is decreased when new data is transmitted successfully. That is, the ESA is able to be reduced by efficient data transmission. However, due to the interference among SNs and the uncertainty of battery level, it is difficult to successfully transmit all data from SNs in one block. Therefore, to make the data transmission more

efficient and further enhance data timeliness, an efficient power adjustment policy is required. As a result, we formulate the corresponding optimization problem $\mathcal{OP}$, which is expressed by

$$\begin{aligned}\mathcal{OP}: &\min_{\boldsymbol{\pi}} \bar{\Delta} \\ &\text{s.t. } (1), (3), (4), (5), (6) \text{ and } (7), \\ &\boldsymbol{\pi} = (\mathbf{P}(1), \mathbf{P}(2), \ldots, \mathbf{P}(T)),\end{aligned} \tag{9}$$

where $\boldsymbol{\pi}$ is the power adjustment policy and $\mathbf{P}(t) = (P_0(t), P_1(t), \ldots, P_M(t))$ is the power adjustment vector of SNs in block t.

Nevertheless, it is difficult to directly solve problem $\mathcal{OP}$: first, the objective function is neither convex nor concave w.r.t. $\boldsymbol{\pi}$; second, there is no explicit expression of the ESA w.r.t. $\boldsymbol{\pi}$; third, the monitor and SNs do not know the prior information about the channel quality and cannot know the future channel state, so it is hard to obtain the smallest ESA through the single-block power adjustment. Therefore, benefiting from DRL's powerful ability on solving complex non-convex problems, we develop a DRL-based method to solve $\mathcal{OP}$ efficiently.

3 MDP FORMULATION

Before solving problem $\mathcal{OP}$ with DRL, we first reformulate $\mathcal{OP}$ as an MDP with the goal of minimizing the long-term average cost, which is regarded as the system's ESA. Furthermore, the MDP is composed of five components and is defined as follows.

State space: To effectively characterize the RF-EH-powered monitoring system, we take the set containing all SN states as the system state. We denote $\mathbb{S}$, $\mathbf{s}(t)$ and $\mathbf{s}_i(t)$ as the system state space, the system state and the state of SN i in block t. Therefore, we have $\mathbf{s}(t) = (\mathbf{s}_1(t), \mathbf{s}_2(t), \ldots, \mathbf{s}_M(t))$. For the real-time monitoring, we use the AoI of SN i at the monitor, the battery level, and the channel gain between SN i and the monitor as the state of SN i. That is, the state of SN i in block t can be expressed as $\mathbf{s}_i(t) = (A_i^D(t), B_i(t), g_i(t))$. It should be noted that the state space $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2 \times \ldots \times \mathbb{S}_M$ is infinite because $B_i(t)$ and $g_i(t)$ are continuous variables.

Action space: To solve $\mathcal{OP}$, let the SNs' transmit power be the action of SN i. Therefore, let $\mathbf{a}(t)$ denote the joint action, which consists of the actions of all SNs, i.e., $\mathbf{a}(t) = (P_1(t), P_2(t), \ldots, P_M(t))$. Let $\mathbb{A}_i$ be the action space of SN i, so the joint action space is given by $\mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2 \times \ldots \times \mathbb{A}_M$. It should also be noticed that the action space is infinite since the transmit power of SN is continuous. Besides, in order to satisfy the constraint (4), the transmit power of SNs should satisfy that $P_i(t) = \min(B_i(t)/\tau, a_i(t)), \forall i$. For implementation, SN i will not send data packets to the monitor in block t if $P_i(t) = 0$. Otherwise, SN i will send status update with power $P_i(t)$.

Transition Probability: The system transition probability is defined as the probability of transition from state $\mathbf{s}(t)$ to $\mathbf{s}(t+1)$ after taking joint action $\mathbf{a}(t)$, which is $\mathbb{P}(\mathbf{s}(t+1) \,|\mathbf{s}(t), \mathbf{a}(t))$.

Cost function: To minimized the long-term $\bar{\Delta}$, we use the sum AoI of SNs in block t obtained by taking action $\mathbf{a}(t)$ from state $\mathbf{s}(t)$ to state $\mathbf{s}(t+1)$ as the cost function, which is expressed by

$$C(t) = C(\mathbf{s}(t+1), \mathbf{s}(t), \mathbf{a}(t)) = \sum\nolimits_{i=1}^{M} \omega_i A_i^D(t+1). \tag{10}$$

Discount factor: Let's denote $\gamma \in (0, 1)$ as the discount factor and $\mathbf{s}(0)$ as the initial state, the long-term discounted cost can be expressed as

$$V_{\boldsymbol{\pi}}(\mathbf{s}(0)) = \mathbb{E}_{\boldsymbol{\pi}}[\sum\nolimits_t \gamma^{t-1} C(\mathbf{s}(t+1), \mathbf{s}(t), \mathbf{a}(t)) \,|\mathbf{s}(0)\,], \tag{11}$$

where $\mathbb{E}_{\boldsymbol{\pi}}[\cdot]$ is the expectation of the state under the power adjustment policy $\boldsymbol{\pi}$.

The optimal power adjustment policy can be expressed as

$$\boldsymbol{\pi}^* = \arg\min_{\boldsymbol{\pi}} V_{\boldsymbol{\pi}}(\mathbf{s}(0)). \tag{12}$$

4 THE PROPOSED SAC-BASED AGE-AWARE POWER ADJUSTMENT METHOD

DRL can be used to solve problems with multi-dimensional continuous state space. However, for the problems with continuous high-dimensional action space, value-based DRL methods such as DQN are intractable. Although some conventional policy-based DRL methods such as DDPG can generate multidimensional continuous actions, they lack action exploration, resulting in insufficient policy performance. In view of the fact that SAC [35], [36] is capable of solving the problem with high-dimensional continuous action space by modeling and optimizing a stochastic policy and can enhance exploration by maximizing the entropy of the policy, SAPA is designed to solve problem $\mathcal{OP}$ in this section.

4.1 The framework of SAC

The goal of SAC is to minimize the entropy-regularized cumulative cost, which is expressed as

$$\begin{aligned}\boldsymbol{\pi}^* = \arg\min_{\boldsymbol{\pi}} [\sum\nolimits_t \gamma^{t-1} (&C(\mathbf{s}(t), \mathbf{a}(t)) \\ &- \alpha \mathcal{H}(\boldsymbol{\pi}(\cdot|\mathbf{s}(t))) \,|\mathbf{s}(0))\,],\end{aligned} \tag{13}$$

where α is the regularization coefficient and $\mathcal{H}(\boldsymbol{\pi}(\cdot|\mathbf{s}(t)))$ is the entropy of the action at $\mathbf{s}(t)$ under policy $\boldsymbol{\pi}$. The soft Q-function and the soft value function of SAC are given by

$$\begin{cases} Q_{\text{soft}}(\mathbf{s}(t), \mathbf{a}(t)) = \mathbb{E}_{\boldsymbol{\pi}}[\sum\nolimits_t (V^t \\ \qquad\qquad - \alpha \gamma^{t-1} \mathcal{H}(\boldsymbol{\pi}(\cdot \mid \mathbf{s}(t))))], \\ V_{\text{soft}}(\mathbf{s}(t)) = \mathbb{E}_{\mathbf{a}(t) \sim \boldsymbol{\pi}(\cdot|\mathbf{s}(t))}[Q(\mathbf{s}(t), \mathbf{a}(t)) \\ \qquad\qquad + \alpha \log \boldsymbol{\pi}(\mathbf{a}(t) \mid \mathbf{s}(t))], \end{cases} \tag{14}$$

where $V^t = \sum_k \gamma^t (C(t))$ is the accumulated cost and $Q(\mathbf{s}(t), \mathbf{a}(t)) = C(t) + \gamma \mathbb{E}\left[V(\mathbf{s}(t+1))\right]$ is the Q-function.

Then, the policy is improved by minimizing the Kullback-Leibler (KL) divergence [35], i.e.,

$$\boldsymbol{\pi}_{\text{new}} = \arg\min_{\boldsymbol{\pi}' \in \boldsymbol{\pi}} D_{\text{KL}} \left(\boldsymbol{\pi}'(\cdot \mid \mathbf{s}(t)) \| \frac{\exp\left(\frac{1}{\alpha} Q(\mathbf{s}(t)), \cdot)\right)}{Z(\mathbf{s}(t)))} \right), \tag{15}$$

where $\boldsymbol{\pi}_{\text{new}}$ is the improved policy, $D_{\text{KL}}(\cdot \| \cdot)$ is the KL divergence, π' is the old policy, and $Z(\mathbf{s}(t))) = \sum_{\mathbf{a}} \exp(Q(\mathbf{s}(t)), \cdot))$ is the normalization variable.

4.2 The presented SAPA

Based on the framework of SAC, SAPA has 5 networks, including a stochastic policy network $\boldsymbol{\pi_\theta}$, 2 soft Q-networks $\{Q_{\phi 1}, Q_{\phi 2}\}$ and their corresponding target networks $\{Q_{\phi 1'}, Q_{\phi 2'}\}$, constructed by artificial neural networks (ANN) with the parameters of $\boldsymbol{\theta}$, $\{\phi 1, \phi 2\}$, and $\{\phi 1', \phi 2'\}$, respectively. The double Q-learning mechanism and target network mechanism are adopted to improve the stability of training. In order to increase the training efficiency, the experience reply mechanism is adopted, where the memory replay buffer $\mathbb{D}$ is built to store historical experience tuples $< \mathbf{s}(t), \mathbf{a}(t), C(t), \mathbf{s}(t+1) >$. Moreover, in order to make the soft Q-value networks estimate accurately, the loss function $J_Q(\boldsymbol{\phi})$ of the soft Bellman residual should be minimized, i.e.,

$$\begin{aligned} J_Q(\boldsymbol{\phi}) =& \mathbb{E}_{k\in\mathbb{I}}[\frac{1}{2}(Q_{\boldsymbol{\phi}}(\mathbf{s}(k),\mathbf{a}(k)) - C(k) \\ & -\gamma V_{\boldsymbol{\phi}'}(\mathbf{s}(k+1)))^2], \end{aligned} \tag{16}$$

where $\mathbb{I}$ is the mini-batch $\mathbb{I}$ sampled from $\mathbb{D}$, $Q_{\boldsymbol{\phi}}(\mathbf{s}(k),\mathbf{a}(k)) = \max(Q_{\phi 1}((\mathbf{s}(k),\mathbf{a}(k)), Q_{\phi 2}((\mathbf{s}(k),\mathbf{a}(k)))$ and $V_{\boldsymbol{\phi}'}(\mathbf{s}(k+1)) = \max_{j\in\{1,2\}} \mathbb{E}_{\mathbf{a}(k+1)\sim\boldsymbol{\pi_\theta}}[Q_{\phi_j'}(\mathbf{s}(k+1), \mathbf{a}(k+1) - \alpha \log \boldsymbol{\pi_\theta}(\mathbf{a}(k+1) \mid \mathbf{s}(k+1))]$. The policy network is improved according to (15), which is

$$\begin{aligned} J_{\boldsymbol{\pi}}(\boldsymbol{\theta}) =& \mathbb{E}_{k\in\mathbb{I}}[\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi_\theta}}[\alpha \log \boldsymbol{\pi_\theta}(\mathbf{a} \mid \mathbf{s}(k)) \\ & +Q_{\boldsymbol{\phi}}(\mathbf{s}(k),\mathbf{a})]]. \end{aligned} \tag{17}$$

Thus, the policy network and soft Q-network are updated by

$$\begin{cases} \boldsymbol{\phi}_i \leftarrow \boldsymbol{\phi}_i - \lambda_C \nabla_{\boldsymbol{\phi}_i} J(\boldsymbol{\phi}_i),\ i \in \{1,2\}, & \text{(18a)} \\ \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda_A \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), & \text{(18b)} \end{cases}$$

where λ_C and λ_A are the update step size of the online soft Q-networks and policy network. Additionally, SAPA also updates the regularization coefficient α by minimizing the loss function, which is given by

$$J(\alpha) = \mathbb{E}_{k\in\mathbb{I}}[\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi_\theta}}[-\alpha \log \boldsymbol{\pi_\theta}(\mathbf{a} \mid \mathbf{s}(k)) - \alpha\mathcal{K}]]. \tag{19}$$

where $\mathcal{K}$ is the target desired entropy.

The update of the target soft Q-network $Q_{\boldsymbol{\phi}_i'}$ adopt the soft update method, i.e.,

$$\boldsymbol{\phi}_i' \leftarrow \tau_C \boldsymbol{\phi}_i + (1-\tau_C)\boldsymbol{\phi}_i',\ i \in \{1,2\}, \tag{20}$$

where $0 < \tau_A \ll 1$ and $0 < \tau_C \ll 1$ are the mixing weights of the online network and the target network, respectively.

For clarity, the proposed SAPA is summarized in Algorithm 1, whose framework is illustrated in Fig. 4. SAPA's training is deployed in the monitor, where it aggregates the local states from SNs, and then broadcasts the action to each SN. Denoting L_S bits as the data size of local state of each SN, L_A as the data size of transmitting power, $T_{\text{total}} = \text{MAX\_EPISODE} * \text{MAX\_STEP}$ as the total training time step, the total communication overhead of SAPA in the training stage is $T_{\text{total}}(ML_S + L_A)$ bits. In the implementation stage, the communication overhead of SAPA remains unchanged.

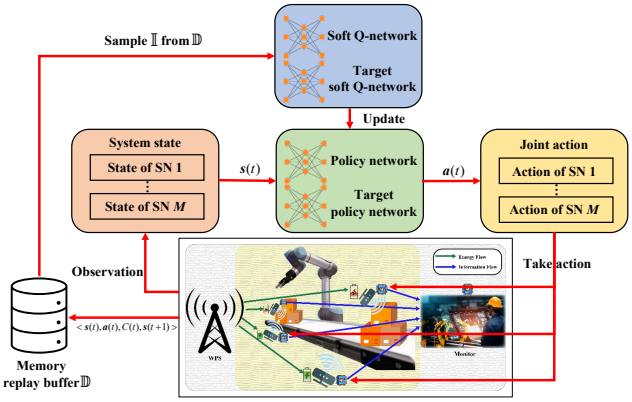


Fig. 4: The framework of the proposed SAPA.

Algorithm 1: The Pseudocode of SAPA

```
Initialize policy network and 2 soft Q-networks.
Initialize target soft Q-networks.
Initialize the replay memory 𝔻, the update step size
  λ_C and λ_A, the mixing weights τ_A and τ_C, the
  maximum episodes MAX_EPISODE, the
  maximum steps in each episode MAX_STEP, and
  the target network update interval T_Interval.
for episode = 1 to MAX_EPISODE do
    for step t = 1 to MAX_STEP do
        The monitor selects action
          a(t)=μ(s(t)|θ) + N_t according to the joint
          state s(t) and exploration noise and
          broadcasts it to SNs. Each SN adopts
          transmit power according to a(t) and
          P_i(t) = min(B_i(t)/τ, a_i(t)), ∀i. Each SN
          obtains the C(t) and next state s(t+1)
          through the interaction of SNs and the
          monitor. Store tuple
          < s(t), a(t), C(t), s(t+1) > in the replay
          buffer 𝔻.
        Randomly sample a mini-batch 𝕀 from the
          replay buffer 𝔻.
        Calculate the MSE loss of the soft Q-network
          by (16) and update its weights by (18a).
        Calculate the policy gradient by (17) and
          update its weight parameters by (18b) to
          update.
        Update the target networks every T_Interval
          steps according to (20).
    end
end
```

5 MULTI-AGENT VERSION OF SAPA

The presented SAPA is able to enhance the system's AoI performance if the RL model is well-trained. Nevertheless, SAPA always requires the monitor to collect all SNs' states, which causes communication overhead in IIoT. To reduce the communication overhead, we design a multi-agent version of SAPA (MSAPA) by treating each SN as an agent, where each SN is able to adjust the power based on its local state without information interaction required after training.

Similar to SAPA, MSAPA also includes 5 networks, i.e., a

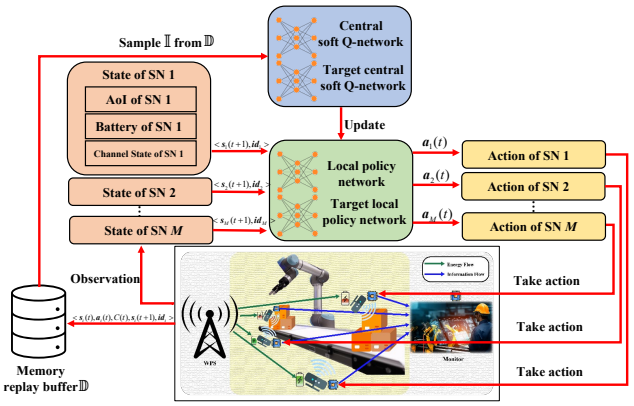


Fig. 5: The framework of the proposed MSAPA.

local stochastic policy network $\boldsymbol{\pi_\theta}$, 2 central soft Q-networks $\{Q_{\phi 1}, Q_{\phi 2}\}$ for double soft Q-learning and their corresponding target networks $\{Q_{\phi 1'}, Q_{\phi 2'}\}$, constructed by the parameters of $\boldsymbol{\theta}$, $\{\phi 1, \phi 2\}$, and $\{\phi 1', \phi 2'\}$, respectively. Different from SAPA, the local policy network only outputs the action for one SN and all SNs share the same policy network but are distinguished by their identity vectors, i.e., the one-hot vectors determined by their id. Specifically, the local policy generates the local action $a_i(t)$ of SN i by observing the local state $\mathbf{s}_i(t)$ and the identity vector $\mathbf{id}_i$, where $a_i(t) = \boldsymbol{\pi_\theta}(\mathbf{s}_i(t), \mathbf{id}_i)$. The central soft Q network evaluates the global action-state pair $(\mathbf{s}(t), \mathbf{a}(t))$. For the central soft Q-network, the loss function is

$$\begin{aligned} J_Q(\boldsymbol{\phi}) =& \mathbb{E}_{k\in\mathbb{I}}[\frac{1}{2}(Q_{\boldsymbol{\phi}}(\mathbf{s}(k), \mathbf{a}(k)) - C(k) \\ & - \gamma * \min_{j\in\{1,2\}} Q^{\text{targ}}_{\boldsymbol{\phi}'_j})^2], \end{aligned} \tag{21}$$

where

$$\begin{aligned} Q^{\text{targ}}_{\boldsymbol{\phi}'_j} =& \mathbb{E}_{\mathbf{a}_i\sim\boldsymbol{\pi_\theta}}[Q_{\boldsymbol{\phi}'_j}(\mathbf{s}(k+1), \mathbf{a})|_{\mathbf{a}=\{\mathbf{a}_i\}} \\ & - \alpha \sum\nolimits_{i\in\mathcal{M}} \log \boldsymbol{\pi_\theta}(\mathbf{a}_i \mid (\mathbf{s}_i(k+1), \mathbf{id}_i))]. \end{aligned} \tag{22}$$

For the local stochastic policy network, it can be improved by

$$\begin{aligned} J(\boldsymbol{\theta}) =& \mathbb{E}_{k\in\mathbb{I}}[\alpha \sum\nolimits_{i\in\mathcal{M}} \mathbb{E}_{\mathbf{a}_i\sim\boldsymbol{\pi_\theta}}[\log \boldsymbol{\pi}(\mathbf{a}_i(t) \mid \mathbf{s}_i(k), \mathbf{id}_i) \\ & + \max_{j\in\{1,2\}} Q_{\boldsymbol{\phi}'_j}(\mathbf{s}(k), \mathbf{a})]]. \end{aligned} \tag{23}$$

Similarly to (19), the regularization coefficient α of MSAPA is also updated by minimizing the loss function, which is

$$\begin{aligned} J(\alpha) =& \mathbb{E}_{k\in\mathbb{I}}[-\alpha \sum\nolimits_{i\in\mathcal{M}} (\mathbb{E}_{\mathbf{a_i}\sim\boldsymbol{\pi_\theta}}[\log \boldsymbol{\pi_\theta}(\mathbf{a_i} \mid (\mathbf{s_i}(k)), \mathbf{id}_i))) \\ & - \alpha\mathcal{K}]]. \end{aligned} \tag{24}$$

By taking one step update, the local stochastic policy and the soft Q network are updated in a similar way to (18a) and (18b) and the target networks are updated similarly to (20).

For clarity, the proposed SAPA is summarized in Algorithm 2, whose framework is illustrated in Fig. 5. Since SNs have no stable energy supply, all the agents' training is deployed in a monitor similar to SAPA. Therefore, the total communication overhead of MSAPA in the training stage is also $T_{\text{total}}(ML_S + L_A)$ bits. In the implement stage, the monitor broadcasts the local network to all SNs and then each SNs is able to adjust its power locally. Therefore, the communication overhead of MSAPA is zero. For comparison, the communication overhead of SAPA and MSAPA is summarized in TABLE 1. It is seen that both SAPA and MSAPA have the same communication overhead in the training stage, but MSAPA has no interaction overhead when deployed.

Algorithm 2: The Pseudocode of MSAPA

Initialize local policy network and 2 central soft-Q networks.
Initialize target soft Q-networks.
Initialize the replay memory $\mathbb{D}$, the update step size λ_C and λ_A, the mixing weights τ_A and τ_C, the maximum episodes MAX_EPISODE, the maximum steps in each episode MAX_STEP, and the target network update interval T_{Interval}.
for *eposide* $= 1$ *to* MAX_EPISODE **do**
  for *step* $t = 1$ *to* MAX_STEP **do**
    for *each* $\forall i \in M$ **do**
      Select action $\mathbf{a}_i(t) = \mu_i(\mathbf{s}_i(t), \mathbf{id}_i|\boldsymbol{\theta}_i) + \mathcal{N}_i$ according to the local state $\mathbf{s}_i(t)$ and exploration noise $\mathcal{N}_i$. SN i takes the transmit power according to $\mathbf{a}_i(t)$ with $P_i(t) = \min(B_i(t)/\tau, a_i(t))$, and obtains $C(t)$ and next state $\mathbf{s}_i(t+1)$. Store tuple $< \mathbf{s}_i(t), \mathbf{a}_i(t), C(t), \mathbf{s}_i(t+1), \mathbf{id}_i >$ in the replay buffer $\mathbb{D}$.
    end
    Randomly sample a mini-batch $\mathbb{I}$ from the replay buffer $\mathbb{D}$.
    Calculate the loss of the central soft Q-network by (21) and update the weights of the central soft Q-network.
    Calculate the policy gradient by (23), and update the weight parameters of the local policy network.
    Update the target networks every T_{Interval} steps.
  end
end

TABLE 1: The communication overhead of SAPA and MSAPA.

	Training stage	Implement stage
SAPA	$T_{\text{total}}(ML_S + L_A)$	$T_{\text{total}}(ML_S + L_A)$
MSAPA	$T_{\text{total}}(ML_S + L_A)$	0

6 PERFORMANCE EVALUATION

The simulations in this section refer to a smart industrial production scenario, where the monitored area is a $20\text{m} \times 20\text{m}$ industrial production workshop, with a WPS and a monitor being located at the coordinates of (0,10) and (20,10), respectively. In order to achieve comprehensive monitoring, multiple SNs are deployed in the workshop, where the locations of SNs follow the uniform distribution. A topology map is shown in Fig. 6. The system is powered

by RF-EH technology, where the network configuration is set based on [19] and [31]. Unless otherwise specified, the parameters of the network configuration are as follows. Specifically, the number of SNs is set to 4. The small-scale fading is set to Rayleigh fading and the path-loss coefficient α is set to 2. For the EH, the energy transmit power of WPS P_t is set to 4W, and the EH coefficient of the SNs η is set to 0.8. The battery capacity of the SN $B_{\max}$ is set to 0.4 mjoules. The sensing probability of SNs p and the SINR threshold γ_{th} are set to 0.2 and 2 dB, respectively. For the uplink data transmission, the received noise is -41 dBm. The weight factor of each SN is set to 1 without loss of generality.

To test the proposed methods and the benchmarks, we build an interactive environment based on Python 3.7.13 and use the PyTorch framework to implement the proposed methods and some learning-based benchmarks, with the version of 1.11 and Cuda 11.3. The related hardware platform is a desktop with an AMD 3600X CPU and an NVIDIA 2070 GPU. For clarity, more related parameters in the simulations are summarized in TABLE 2.

To evaluate the ESA performance of the proposed SAPA and MSAPA, we simulate and compare the two methods with several DRL-based benchmark methods, i.e., SDQN-PA and SDDPG-PA. SDQN-PA is based on the DQN algorithm, which is policy-based with a Q network, so it is necessary to discretize the action space to solve the formulated problem described in Section 2.3. SDDPG-PA is based on the DDPG algorithm, which is also actor-critic (AC)-based with policy networks and critic networks. In order to show the effectiveness of the power adjustment, we also simulated RAPA and the FET. The descriptions of the simulated benchmark methods are as follows:

- **The single-agent-based DQN Power Adjustment Method (SDQN-PA):** SDQN-PA is a policy-based method, where the monitor determines discrete transmit power to the SNs via the DQN-based method with (10) as the cost function. For each SN, the transmit power is set to $P_i(t) \in \{0, \frac{P_{\max}}{D}, \frac{2P_{\max}}{D}, \ldots, P_{\max}\}$, where it is discretized into D levels.
- **The single-agent-based DDPG Power Adjustment (SDDPG-PA) method:** SDDPG-PA is an AC-based method that can generate continuous action, where the monitor determines transmit power to the SNs via the DDPG-based method with (10) as the cost function. At each output step of SDDPG-PA, decode the action with the maximum Q value and map it to the actual power.
- **The Random Power Adjustment (RAPA) method:** Each SN randomly adopts its transmit power.
- **The Full energy transmit (FET) method:** Each SN transmits status updates only if its battery is fully charged; otherwise, the transmit power is set to 0.

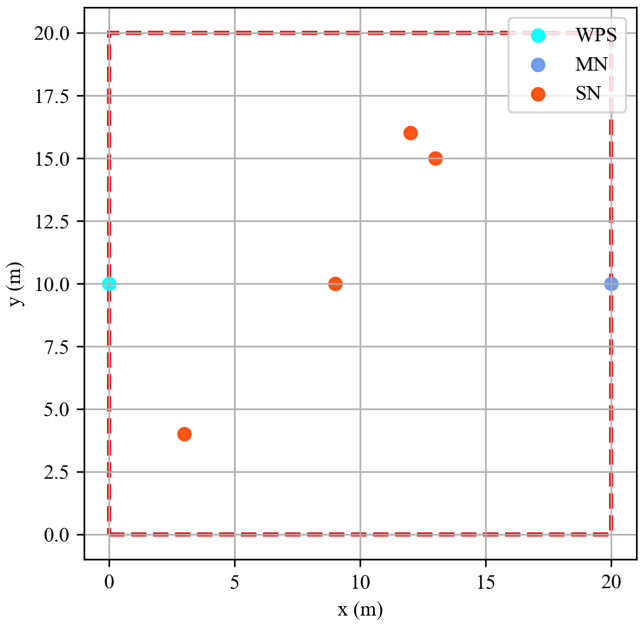


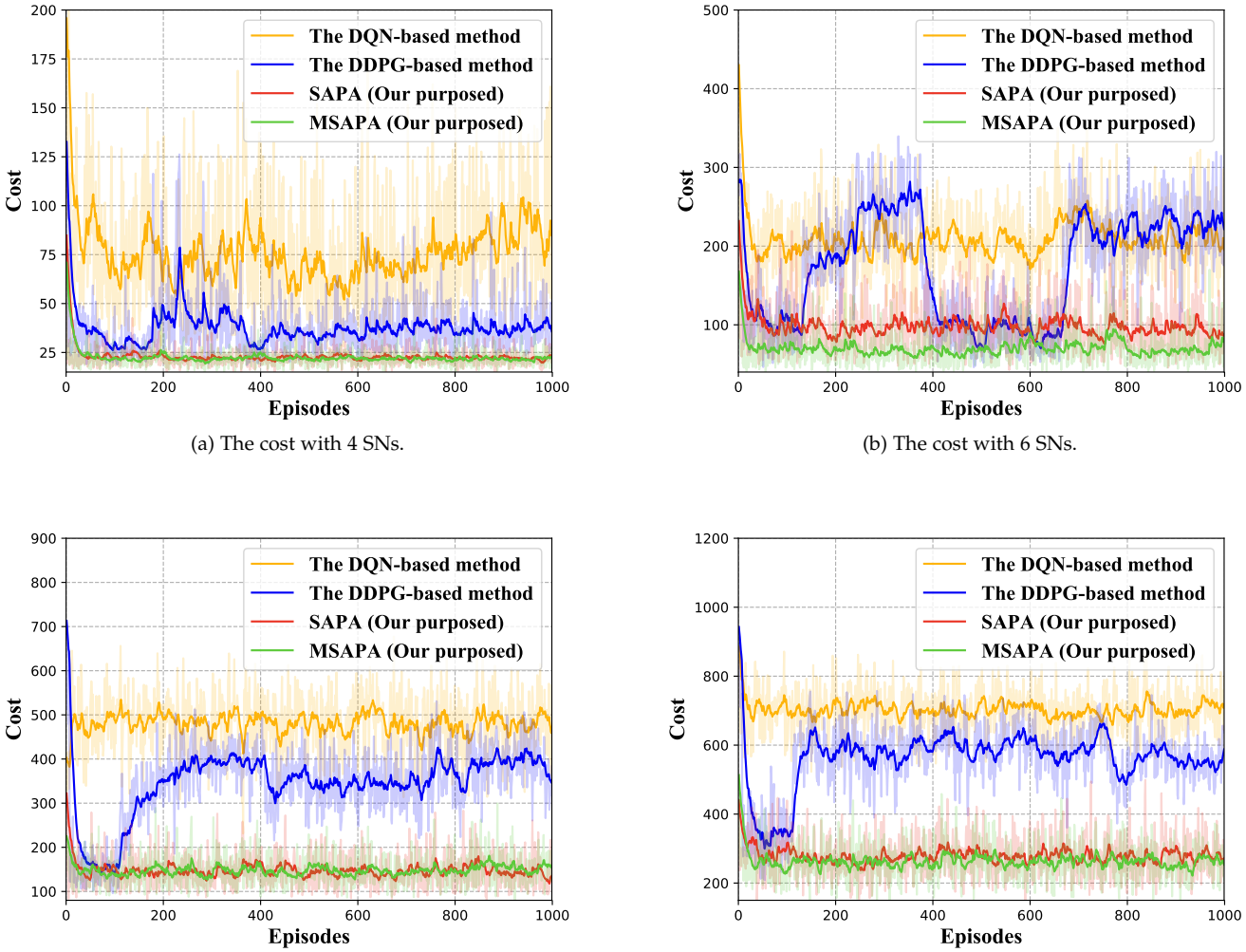
Fig. 6: An example topology map of the RF-EH-powered network.

TABLE 2: Parameters of the simulated RL-based methods.

Parameter	Value
The learning rate of the policy network	1e-3
The learning rate of the Soft Q-network	2e-3
Batch size	256
Discount factor	0.995
Memory capacity	1000
The update factor of target network	0.005
The policy network scale of SAPA	{64,32}
The soft Q-network scale of SAPA	{128,32}
The policy network scale of MSAPA	{64,32}
The soft Q-network scale of MSAPA	{128,32}
The Q-network scale of SDQN-PA	{200,200}
The power discrete gears D of the SDQN-PA	5
The policy network scale of SDDPG-PA	{64,32}
The critic network scale of SDDPG-PA	{128,32}
The learning rate of the policy network of SDDPG-PA	1e-3
The learning rate of the critic network of SDDPG-PA	2e-3

Fig. 7 shows the convergence of behavior of SAPA, MSAPA, SDQN-PA and SDDPG-PA. It is seen that the SDDPG-PA, SAPA, and MSAPA achieve lower system cost than SDQN-PA method because these methods are more suitable for problems with continuous action space. Moreover, the proposed SAPA and MSAPA converge well with different numbers of SNs, while SDDPG-PA method oscillates in networks with 6, 8 and 10 SNs, because the Q-value in the SDDPG-PA approximated by the critic network may gradually be overestimated in training, which leads to repeated selections of a deterministic action, periodically causing the policy of SDDPG-PA to fall into local optimal and thus causing oscillations. The proposed SAPA and MSAPA are based on the SAC algorithm, learning stochastic policy, which will not assign a very high probability to any action so that SAPA and MSAPA maintain stable performance. Besides, the proposed SAPA and MSAPA achieve the lowest cost among the four methods, and their performance is close.

Fig. 8 depicts the cost with different learning rates. We conducted multiple sets of experiments at different learning rates and obtained similar conclusions, so we selected a representative set for example. As can be seen from the figure, the learning rate has a small impact on the performance of our presented two methods. The cost associated with



(a) The cost with 4 SNs.

(b) The cost with 6 SNs.

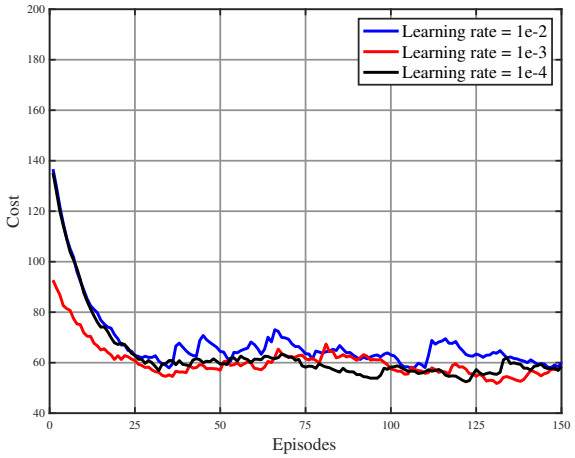
(c) The cost with 8 SNs.

(d) The cost with 10 SNs.

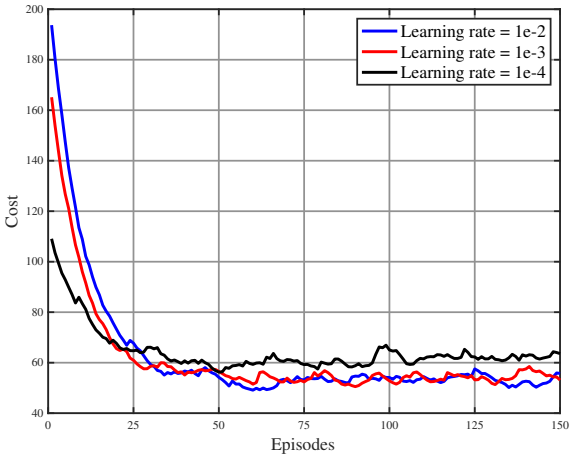
Fig. 7: Comparison of the cost of the four RL-based methods with different number of SNs.

the three different learning rates is roughly the same. For SAPA, compared to the learning rate being 1e-3 and 1e-4, the fluctuation is more obvious when the learning rate is 1e-2. Moreover, the learning rate has different effects on the two methods in terms of convergence speed. Specifically, for SAPA, the convergence is the fastest when the learning rate is 1e-3; while for MSAPA, although the convergence is faster when the learning rate is 1e-4, the arrival performance is relatively poor. Therefore, comprehensively considering cost, volatility, and convergence speed, we choose 1e-3 as the learning rate for SAPA and MSAPA, as shown in TABLE 2.

Fig. 9(a) depicts the ESA with different numbers of SNs. One can observe that All benchmark methods are able to achieve low ESA with a smaller number of SNs such as 2 and 4. The ESA increases as the number of SNs increases, because the interference among SNs increases as the number of SNs increases, and thus the transmission from SNs to the monitor is more likely to fail. When the number of that SN exceeds 4, the ESA of FET, SDQN-PA, RAPA is high. The reasons are different for the three methods. Particularly, for FET, the reason is that all power of each SN is used to transmit and the interference between SNs is very large. For RAPA, as the number of users increases, random power adjustment makes it difficult to ensure effective transmission of status updates. For SDQN-PA, the reason is the discretization of the action that it is difficult to learn a better policy to achieve lower ESA. In contrast, SDDPG-PA, SAPA, and MSAPA are able to maintain the effectiveness of power adjustment with different numbers of SNs, which shows that the AC-based method has better performance in solving MDPs with continuous action spaces. Among them, the ESA of SDDPG-PA is higher than that of SAPA and MSAPA, because of the insufficient exploration caused by the use of deterministic action strategies in DDPG. Moreover, SAPA and MSAPA achieve the lowest ESA, and when the number of SNs is greater than 8, MSAPA is slightly better than SAPA. This may be due to the fact that SAPA treats the system composed of all SNs as an agent, and the action space dimension is larger. MSAPA considers each SN as an agent, where the action space of each agent is relatively smaller, and it is easier to learn better strategies.

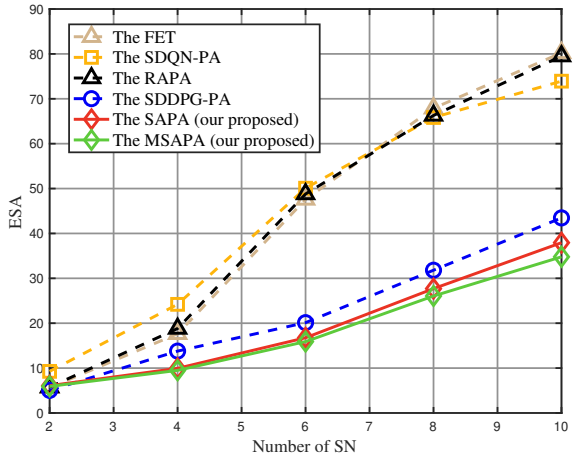


(a) SAPA's cost with different learning rates.

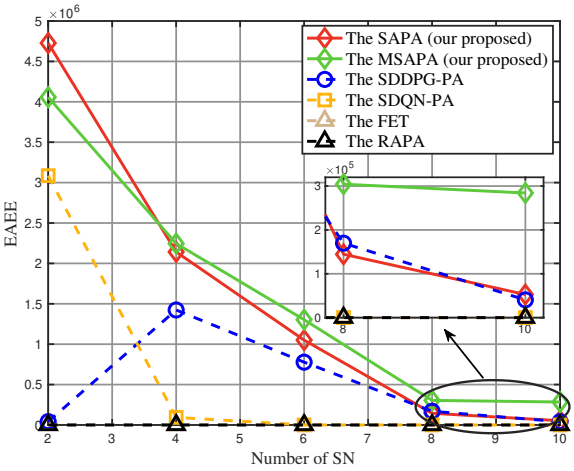


(b) MSAPA's cost with different learning rates.

Fig. 8: The cost of the proposed methods with different learning rates.



(a) ESA versus different numbers of SNs.



(b) EAEE versus different numbers of SNs.

Fig. 9: ESA and EAEE achieved by the simulated six methods versus different numbers of SNs.

Since energy efficiency is also an important performance index for RF-EH-powered networks, we also take the age-energy efficiency (AEE) newly presented in [37] as a metric to evaluate the proposed methods. Actually, AEE provides an insightful measure of the achievable AoI improvement per unit of energy consumption, which is given by

$$\Phi(t) = \frac{\sum_{i=1}^{M} A_{\max} - A_i^D(t+1)}{\sum_{i=1}^{M} \tau P_i(t)}. \tag{25}$$

Similar to ESA, denote EAEE as the average AEE in time, which is given by

$$EAEE = \frac{\sum_{t=1}^{T} \Phi(t)}{T}. \tag{26}$$

Fig. 9(b) depicts the EAEE versus the number of SNs. Without loss of generality, we normalize the energy to facilitate comparison by mapping the energy consumed to be between 0 and 1. As can be seen from the figure, the AEE achieved by the non-reinforcement learning methods is relatively low. Although the ESA achieved by the non-reinforcement learning methods is close to the lowest ESA when the number of SNs is small, this means that the non-reinforcement learning methods have relatively high energy consumption. For SDQN-PA, when the number of SNs is 2, it has a higher AEE, which means that the policy learned by SDQN-PA is very effective when the number of SNs is relatively small. However, when the number of SNs exceeds 2, the AEE achieved by SDQN-PA drops sharply, illustrating that SDQN-PA cannot cope with scenarios with more than 2 SNs. Interestingly, when the SN number is 2, the AEE achieved by SDDPG-PA is relatively low. However, as the number of SNs increases, the AEE achieved by SDDPG-PA experiences an increase, which means that the policy learned by SDDPG-PA makes power concessions to reduce ESA. Moreover, as the number of SNs increases, the decreases in ESA cause decreases in AEE accordingly.

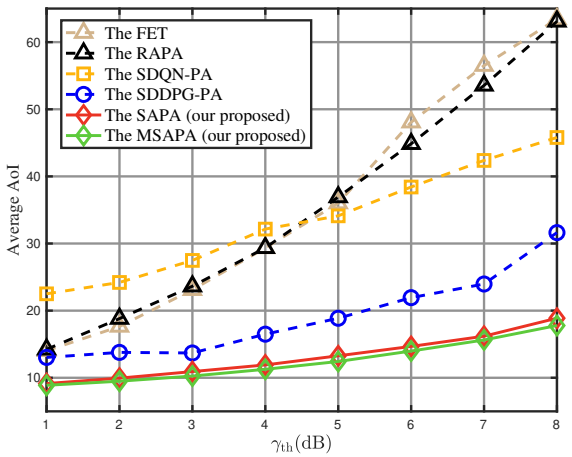


Fig. 10: ESA versus $\gamma_{\rm th}$.

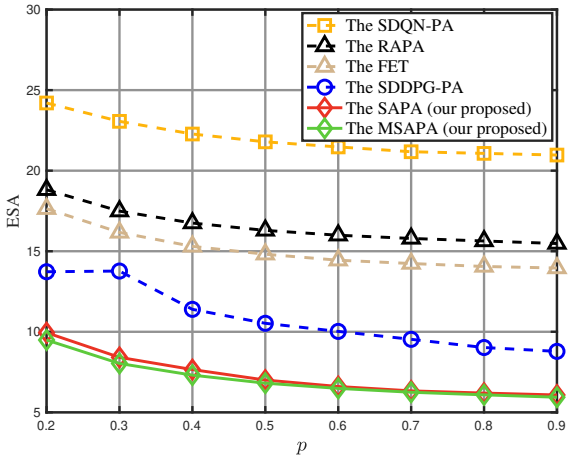


Fig. 11: ESA versus p.

The proposed SAPA and MSAPA achieve the highest AEE among the benchmark methods, indicating that the two methods can not only effectively reduce ESA, but also learn the concessions between SNs to save energy consumption, as well as improve AEE. In addition, when the number of SNs is greater than 4, the AEE achieved by MSAPA is higher than SAPA, which means that MSAPA is more suitable for scenarios with more SNs.

Fig. 11 depicts the ESA of SNs with different packet arrival probabilities, i.e., p. It is observed that the more frequently the packets arrive with larger p, the lower the ESA is. The reason is that when the packet arrival probability is higher, and the transmitted status updates from SNs are fresher, and the ESA becomes lower. It should be noticed that when the packet arrival probability is higher than 0.6, its impact on the ESA becomes small because not all newly arrived state updates are successfully transmitted in each block. Besides, with the same p, the proposed SAPA and MSAPA achieve the lowest ESA, which demonstrates an effective power adjustment is able to reduce the ESA greatly.

Fig. 10 plots the ESA of SNs with different $\gamma_{\rm th}$. One can observe that the ESA increases versus $\gamma_{\rm th}$, because the higher SINR threshold means higher data requirements. As a result, with high data requirements, the status update transmission of SNs is also likely to fail, so the ESA increases. The proposed SAPA and MSAPA also achieve the lowest ESA among the benchmark methods. It is noticed that MSAPA achieves lower ESA than SAPA at higher SNR thresholds, which inspires us that MSAPA is more suitable for networks with high data requirements.

Fig. 12 depicts the results of the scalability testing of the proposed MSAPA, where the red line represents the ESA performance achieved by MSAPA with the unchanged positions of SNs compared to the trained network topology and the blue line represents the performance achieved by MSAPA with changed positions compared to the trained network topology. As can be seen from the figure, MSAPA and MSAPA (with changed positions) have similar performance, but as the number of users increases, the performance gap between the two gradually widens. This shows that MSAPA is scalable w.r.t. the location of SNs. Besides, in networks with more SNs, when the position of SNs changes,

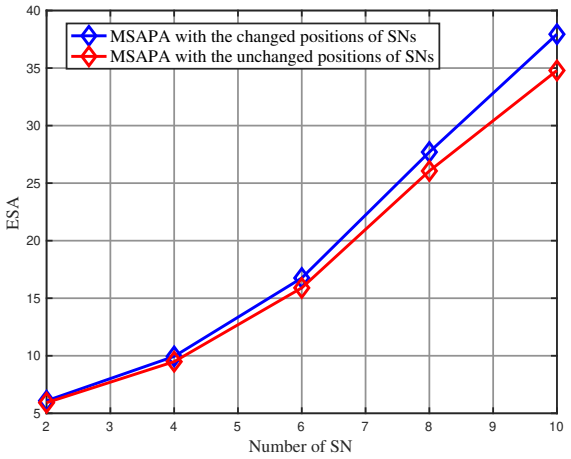


Fig. 12: ESA achieved by MSAPA with the unchanged/changed positions of SNs compared to the trained network topology under different numbers of SNs.

indicating that in this case the network should be retrained as much as possible.

7 Conclusion

This paper studied an RF-EH-powered wireless system and an optimization problem was formulated to minimize the ESA by optimizing the power adjustment of SNs under multiple practical constraints. An MDP problem was first established and SAPA was proposed to solve the problem. To further reduce the communication overhead in IIoT, a multi-agent version of SAPA, i.e., MSAPA, was also designed. Simulation results demonstrated the convergence of the proposed SAPA and MSAPA and also showed that the ESA of the system could be greatly reduced compared to several benchmark methods.

References

[1] L. Ma, X. Wang, X. Wang, L. Wang, Y. Shi, and M. Huang, "Tcda: Truthful combinatorial double auctions for mobile edge computing in industrial internet of things," *IEEE Trans. on Mobile Comput.*, vol. 21, no. 11, pp. 4125–4138, Nov. 2022.

[2] L. Liu, G. Han, Z. Xu, L. Shu, M. Martínez-García, and B. Peng, "Predictive boundary tracking based on motion behavior learning for continuous objects in industrial wireless sensor networks," *IEEE Trans. on Mobile Comput.*, vol. 21, no. 9, pp. 3239–3249, Sep. 2022.
[3] M. Rea and D. Giustiniano, "Location-aware wireless resource allocation in industrial-like environment," *IEEE Trans. on Mobile Comput.*, 2021.
[4] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, May 2012, pp. 2731–2735.
[5] M. Akbari, M. R. Abedi *et al.*, "Age of information aware vnf scheduling in industrial iot using deep reinforcement learning," *IEEE J. Sel. Areas in Commun.*, vol. 39, no. 8, pp. 2487–2500, Aug. 2021.
[6] T.-T. Chan, H. Pan, and J. Liang, "Age of information with joint packet coding in industrial iot," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2499–2503, Nov. 2021.
[7] Y.-H. Chiang, H. Lin, and Y. Ji, "Information cofreshness-aware grant assignment and transmission scheduling for internet of things," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14 435–14 446, Oct. 2021.
[8] J. Li, J. Tang, and Z. Liu, "On the data freshness for industrial internet of things with mobile edge computing," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13 542–13 554, Aug. 2022.
[9] J. Wang, X. Cao *et al.*, "Sleep-wake sensor scheduling for minimizing aoi-penalty in industrial internet of things," *IEEE Internet Things J.*, vol. 9, no. 9, pp. 6404–6417, May 2022.
[10] J. Zhao, Y. Wang *et al.*, "Timely device status updates in industrial wireless monitoring systems under resource constraints," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18 791–18 805, Oct. 2022.
[11] M. Li *et al.*, "Age-of-information aware scheduling for edge-assisted industrial wireless networks," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5562–5571, Aug. 2021.
[12] I. Kadota and E. Modiano, "Minimizing the age of information in wireless networks with stochastic arrivals," *IEEE Trans. on Mobile Comput.*, vol. 20, no. 3, pp. 1173–1185, Mar. 2021.
[13] Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Trans. on Mobile Comput.*, vol. 19, no. 12, pp. 2903–2915, Dec. 2020.
[14] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Optimizing information freshness in wireless networks: A stochastic geometry approach," *IEEE Trans. on Mobile Comput.*, vol. 20, no. 6, pp. 2269–2280, Jun, 2021.
[15] X. Wang, Z. Ning, S. Guo, M. Wen, and H. V. Poor, "Minimizing the age-of-critical-information: An imitation learning-based scheduling approach under partial observations," *IEEE Trans. on Mobile Comput.*, vol. 21, no. 9, pp. 3225–3238, Sep. 2022.
[16] S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey and implications," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 443–461, Jul. 2011.
[17] R. Zhang, K. Xiong *et al.*, "Q-learning-based adaptive power control in wireless rf energy harvesting heterogeneous networks," *IEEE Syst. J.*, vol. 15, no. 2, pp. 1861–1872, Jun. 2021.
[18] K. Xiong, P. Fan, C. Zhang, and K. B. Letaief, "Wireless information and energy transfer for two-hop non-regenerative mimo-ofdm relay networks," *IEEE J. Sel. Areas in Commun.*, vol. 33, no. 8, pp. 1595–1611, Aug. 2015.
[19] H. Li, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. Letaief, "Distributed design of wireless powered fog computing networks with binary computation offloading," *IEEE Trans. on Mobile Comput.*, 2021.
[20] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 628–631, Apr. 2019.
[21] H. Zheng, K. Xiong *et al.*, "Age-energy region in wireless powered communication networks," in *Proc. IEEE INFOCOM Workshops*, Aug. 2020, pp. 334–339.
[22] Y. Lu, K. Xiong *et al.*, "Online transmission policy in wireless powered networks with urgency-aware age of information," in *Proc. IWCMC*, Jul. 2019, pp. 1096–1101.
[23] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "Online age-minimal sampling policy for rf-powered iot networks," in *Proc. IEEE GLOBECOM*, Feb. 2019, pp. 1–6.
[24] Y. Khorsandmanesh, M. J. Emadi, and I. Krikidis, "Average peak age of information analysis for wireless powered cooperative networks," *IEEE Trans. Cog. Commun. and Netw.*, vol. 7, no. 4, pp. 1291–1303, Dec. 2021.
[25] H. Zheng, K. Xiong *et al.*, "Age of information-based wireless powered communication networks with selfish charging nodes," *IEEE J. Sel. Areas in Commun.*, vol. 39, no. 5, pp. 1393–1411, May 2021.
[26] S. Leng and A. Yener, "Age of information minimization for an energy harvesting cognitive radio," *IEEE Trans. on Cog. Commun. and Netw.*, vol. 5, no. 2, pp. 427–439, 2019.
[27] H. Hu, K. Xiong *et al.*, "Aoi-minimal trajectory planning and data collection in uav-assisted wireless powered iot networks," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1211–1223, Jan. 2021.
[28] M. Hatami, M. Leinonen, and M. Codreanu, "Aoi minimization in status update control with energy harvesting sensors," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8335–8351, Dec. 2021.
[29] L. Liu, K. Xiong *et al.*, "Average aoi minimization in uav-assisted data collection with rf wireless power transfer: A deep reinforcement learning scheme," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5216–5228, Apr. 2022.
[30] O. S. Oubbati, M. Atiquzzaman *et al.*, "Multi-uav-enabled aoi-aware wpcn: A multi-agent reinforcement learning strategy," in *Proc. IEEE INFOCOM Workshops*, May 2021, pp. 1–6.
[31] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "A reinforcement learning framework for optimizing age of information in rf-powered communication systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4747–4760, Aug. 2020.
[32] L. Zhang and K.-W. Chin, "A distributed device selection method to minimize aoi in rf-charging networks," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3733–3737, Nov. 2021.
[33] Y. Guo, K. Xiong *et al.*, "Slipt-enabled multi-led mu-miso vlc networks: Joint beamforming and dc bias optimization," *IEEE Trans. Green Commun. Netw.*, 2022.
[34] Z. Zhu, S. Wan *et al.*, "Federated multi-agent actor-critic learning for age sensitive mobile edge computing," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1053–1067, Jan. 2022.
[35] T. Haarnoja, A. Zhou *et al.*, "Soft actor-critic algorithms and applications," *arXiv:1812.05905*, 2018.
[36] Y. Pu *et al.*, "Decomposed soft actor-critic method for cooperative multi-agent reinforcement learning," *arXiv:2104.06655*, 2021.
[37] H. Zheng, K. Xiong, M. Sun, H. Wu, Z. Zhong, and X. Shen, "Maximizing age-energy efficiency in wireless powered industrial ioe networks: A dual-layer dqn-based approach," *IEEE Trans. on Wireless Commun.*, pp. 1–1, 2023.

Yiyang Ge received the B.E. degree from the School of Computer and Information Technology, Beijing Jiaotong University (BJTU), Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, BJTU, Beijing, China. His current research interests include wireless powered networks, mobile edge computing, age of information and machine learning technologies for wireless communications.

Ke Xiong (Member, IEEE) received the B.S. and Ph.D. degrees from Beijing Jiaotong University (BJTU), Beijing, China, in 2004 and 2010, respectively.

From April 2010 to February 2013, he was a Post-Doctoral Research Fellow with the Department of Electronics Engineering, Tsinghua University, Beijing. Since March 2013, he has been a Lecturer, an Associate Professor of BJTU, where he is currently a Full Professor and the Vice Dean of the School of Computer and Information Technology. From September 2015 to September 2016, he was a Visiting Scholar with the University of Maryland, College Park, MD, USA. He has published more than 100 academic papers in referred journals and conferences. His current research interests include wireless cooperative networks, wireless powered networks, and network information theory.

Dr. Xiong is a member of the China Computer Federation (CCF) and a Senior Member of the Chinese Institute of Electronics (CIE) and the Chinese Association for Artificial Intelligence (CAAI). He also serves as a reviewer for more than 15 international journals, including IEEE TRANS ACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS LETTERS, IEEE SIGNAL PROCESSING LETTERS, and IEEE WIRELESS COMMUNICATION LETTERS. He received the Best Student Paper Awards from HWMC'2014, the 25th and 26th Annual Conference of Information Theory of CIT (CIT-IT), the IEEE ICC'2020, and the TAOS Technical Committee at IEEE ICC'2020. He served as the Session Chair for IEEE GLOBECOM'2012, IET ICWMMN'2013, IEEE ICC'2013, and ACM MOMM'2014, the Publicity and the Publication Chair for IEEE HMWC'2014, and the TPC Co-Chair for IET ICWMMN'2017 and IET ICWMMN'2019. He serves as the Associate Editor-in Chief for the *Chinese Journal New Industrialization Strategy* and an Editor for International Journal of *Computer Engineering and Software Technology*. In 2017, he serves as a Lead Editor for the Special Issue "Recent Advances in Wireless Powered Communication Networks" for *EURASIP Journal on Wireless Communications and Networking* and the Guest Editor for the Special Issue "Recent Advances in Cloud-Aware Mobile Fog Computing" for *Wireless Communications and Mobile Computing*.

Qiong Wang is currently pursuing her Ph.D. degree in Electronic Engineering from Tsinghua University. She has been working at the headquarters of the State Grid Corporation of China since 2008 and has served as the Deputy General Manager of the Second Level Department of State Grid Beijing Company since 2019. She was responsible for the preparation of local plans for the 13th Five-Year Plan period at the Planning Department of the Energy Bureau from 2010 to 2011. During her work at State Grid, she mainly engaged in the development of software and hardware for charging stations and power-side infrastructure, alongside the integration of advanced security initiatives in vehicle networking platforms. She chaired and participated in more than 20 scientific research and engineering projects related to IoT communication, network security, and vehicle network interaction. She won the first prize of the China Electric Power Innovation Award in 2018. In 2022, she won the 26th "China Youth May Fourth Medal" and the first prize of the Science and Technology Award form the China Electric Power Promotion Council.

Qiang Ni (Senior Member, IEEE) is currently a Professor and the Head of the Communication Systems Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. His current research interests include future-generation communications and networking, including green communications and networking, millimeter-wave wireless communications, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, energy harvesting, wireless information and power transfer, the IoT, cyber-physical systems, AI and machine learning, big data analytics, and vehicular networks. He has authored or coauthored more than 300 articles in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to various IEEE wireless standards.

Pingyi Fan (Senior Member, IEEE) received the B.S. degree from the Department of Mathematics, Hebei University, Baoding, China, in 1985, and the M.S. degree from the Department of Mathematics, Nankai University, Tianjin, China, in 1990, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1994. He is currently a Professor with the Department of EE, Tsinghua University. From August 1997 to March 1998, he visited Hong Kong University of Science and Technology, Hong Kong, as Research Associate. From May 1998 to October 1999, he visited the University of Delaware, Newark, DE, USA, as Research Fellow. In March 2005, he visited NICT, Tokyo, Japan, as a Visiting Professor. From June 2005 to May 2014, he visited Hong Kong University of Science and Technology for many times and from July 2011 to September 2011, he is a Visiting Professor with the Institute of Network Coding, Chinese University of Hong Kong, Hong Kong.

Dr. Fan is a senior member of IEEE and an oversea member of IEICE. He has attended to organize many international conferences including as General co-Chair of EAI Chinacom2020, and IEEE VTS HMWC2014, TPC co-Chair of IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS 2010) and TPC member of IEEE ICC, Globecom, WCNC, VTC, Inforcom etc. He has served as an editor of IEEE Transactions on Wireless Communications, Inderscience International Journal of Ad Hoc and Ubiquitous Computing, Wiley Journal of Wireless Communication and Mobile Computing, MDPI Electronics, and Open Journal of Mathematical Sciences etc. He is also a reviewer of more than 30 international Journals including 20 IEEE Journals and 8 EURASIP Journals.

He has received some academic awards, including the IEEE WCNC'08 Best Paper Award, ACM IWCMC'10 Best Paper Award, IEEE Globecom'14 Best Paper Award, IEEE ICC'20 Best Paper Award, IEEE TAOS Technical Committee'20 Best Paper Award, and the CIEIT Best Paper Awards in 2018 and in 2019. Also, he has received IEEE ComSoc Excellent Editor Award for IEEE Transactions on Wireless Communications in 2009. His main research interests include B5G technology in wireless communications such as MIMO, OFDMA, Network coding, Network information theory, Machine learning and Big data analysis.

Khaled Ben Letaief (Fellow, IEEE) received the B.S. (Hons.), M.S., and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in December 1984, August 1986, and May 1990, respectively.

From 1990 to 1993, he was a Faculty Member with The University of Melbourne, Australia. Since 1993, he has been with The Hong Kong University of Science and Technology (HKUST), where he is currently the New Bright Professor of engineering. While at HKUST, he has held many administrative positions, including an Acting Provost, the Dean of the Engineering, the Head of the Electronic and Computer Engineering Department, and the Director of the Hong Kong Telecom Institute of Information Technology. He is an internationally recognized leader in wireless communications and networks. His research interests include artificial intelligence, mobile cloud and edge computing, tactile internet, and 6G systems. In these areas, he has over 720 articles along with 15 patents, including 11 U.S. inventions

Dr. Letaief is a member of the United States National Academy of Engineering, a fellow of Hong Kong Institution of Engineers, and a member of Hong Kong Academy of Engineering Sciences. He was a recipient of many distinguished awards and honors, including the 2019 Distinguished Research Excellence Award by HKUST School of Engineering (Highest Research Award and only one recipient/three years is honored for his/her contributions), the 2019 IEEE Communications Society and Information Theory Society Joint Paper Award, the 2018 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 IEEE Cognitive Networks Technical Committee Publication Award, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, the 2011 IEEE Wireless Communications Technical Committee Recognition Award, the 2011 IEEE Communications Society Harold Sobol Award, the 2010 Purdue University Outstanding Electrical and Computer Engineer Award, the 2009 IEEE Marconi Prize Award in Wireless Communications, the 2007 IEEE Communications Society Joseph LoCicero Publications Exemplary Award, and more than 16 IEEE best paper awards. He is well recognized for his dedicated service to professional societies and IEEE, where he has served in many leadership positions, including the Treasurer of the IEEE Communications Society, the IEEE Communications Society Vice- President for Conferences, the Chair of the IEEE Committee on Wireless Communications, an Elected Member of the IEEE Product Services and Publications Board, and the IEEE Communications Society Vice-President for Technical Activities. He is the Founding Editor-in-Chief of the prestigious IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and has served on the Editorial Board of other premier journals, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS: Wireless Communications Series (as the Editor-in-Chief). He has also been involved in organizing many flagship international conferences. He also served as the President of the IEEE Communications Society from 2018 to 2019, the world's leading organization for communications professionals with headquarter in New York City and members in 162 countries. He is also recognized by Thomson Reuters as an ISI Highly Cited Researcher and was listed among the 2020 top 30 of AI 2000 Internet of Things Most Influential Scholars.