RLR: Joint Reinforcement Learning and Attraction Reward for Mobile Charger in Wireless Rechargeable Sensor Networks

Cuijuan Shang, Chih-Yung Chang, *Member, IEEE*, Wen-Hwa Liao, *Member, IEEE*, and Diptendu Sinha Roy

Abstract—Advances in wireless charging technology give great new opportunities for extending the lifetime of a wireless sensor network (WSN) which is an important infrastructure of IoT. However, the existing greedy algorithms lacked learning from the experiences of energy dissipation trends. Unlike the existing studies, this article proposes a reinforcement learning approach, called reinforcement learning recharging (RLR), for mobile charger to learn the trends of WSNs, including the energy consumption of the sensors, the recharging cost as well as the coverage benefit, aiming to maximize the coverage contribution of the recharged WSN. The proposed RLR mainly consists of three modules, including sensor energy management (SEM), charger location update (CLP), and charger reinforcement learning (CRL) modules. In the SEM module, each sensor manages its energy and calculates its threshold for the recharging request in a distributed manner. The CLP module adopts the quorum system to ensure effective communication between sensors and the mobile charger. Meanwhile, the CRL module employs attraction rewards to reflect the coverage benefit and penalties of waiting time raised due to charger movement and recharging other sensors. As a result, the charger accumulates the learning experiences from the Q-Table such that it is able to execute the appropriate actions of charging or moving in a manner of state management. Performance results show that the proposed RLR outperforms the existing recharging mechanisms in terms of charging waiting time of sensors, the energy usage efficiency of the mobile charger, as well as the coverage contribution of the given sensor network.

Index Terms—Mobile charger, recharging mechanism, reinforcement learning, wireless rechargeable sensor network (WRSN).

Cuijuan Shang is with the School of Computer and Information Engineering, Chuzhou University, Chuzhou 239000, China (e-mail: shangcuijuan@chzu.edu.cn).

Chih-Yung Chang is with the Department of Computer Science and Information Engineering, Tamkang University, Taipei 25137, Taiwan (e-mail: cychang@mail.tku.edu.tw).

Wen-Hwa Liao is with the Institute of Information and Decision Sciences, National Taipei University of Business, Taipei 25137, Taiwan (e-mail: whliao@ntub.edu.tw).

Diptendu Sinha Roy is with the Department of Computer Science and Engineering, National Institute of Technology Meghalaya, Shillong 793003, India (e-mail: diptendu.sr@nitm.ac.in).

Digital Object Identifier 10.1109/JIOT.2023.3267242

I. INTRODUCTION

WIRELESS sensor network (WSN) is an important and basic infrastructure of the Internet of Things. It has been applied in many applications, such as environment monitoring [1], localizations [2], smart healthcare [3], intelligent manufacturing [4], and wisdom agriculture [5]. The major task of each sensor in WSNs is to perform the sensing and communication operations. However, each sensor is powered by battery which is difficult to be replaced in some harsh conditions. The WSNs will be inefficient if too many sensors run out of energy. Therefore, replenishing sensor's battery has attracted much attention of researchers and plenty of algorithms have been proposed to prolong the lifetime of WSNs [6], [7], [8], [9].

The WSNs that can be energy replenished are called wireless rechargeable sensor networks (WRSNs) [10]. In the literature, plenty of energy replenishing algorithms have been proposed. These algorithms can be classified into two types, including environmental energy harvesting [7], [11], [12] and mobile charger charging [13], [14], [16], [17], [18], [19], [20], [21]. There are many environmental energy resources, such as solar and wind, which can be harvested to renew sensor's battery. However, an inevitable drawback of environmental energy harvesting is the inherent dynamics of energy sources. When energy sources are not available, sensor nodes may stop working, leading to long data latency or even data loss in the network. With the development of the wireless charging technologies, energy recharging of sensors based on the mobile charger have become another hot topic in the WSN.

The recharging algorithms for the mobile charger can be mainly divided into two classes, including offline and online. In the offline class, the mobile charger moved according to a recharging schedule which was a decision made according to a sequence of recharging requests from low-energy sensors. Those algorithms aimed to plan the optimal path for the mobile charger and were suitable for the traffic balancing scenarios. However, the decision will not be changed with the new receiving requests until the recharging task has been completed. Different from the offline recharging algorithms, the online ones dynamically adjusted the recharging schedule based on the received recharging requests from sensors in real time. Since the mobile charger can receive new requests from sensors, the schedule which was made according to the previous requests was usually difficult to meet the requirement

of the new requests. As a result, it is a big challenge to develop a greedy algorithm which satisfies the on-demand requests received from sensors in the networks. The online algorithms have the advantage that they can greedily update the charging schedule according to the new charging requests. However, they lacked learning from the experiences of energy dissipation trends of the WSNs.

Different from the existing work, this article proposes a reinforcement learning approach, called reinforcement learning recharging (RLR), for the mobile charger to learn the trends of WSNs, including the energy consumption of the sensors, the recharging and moving costs, as well as the coverage benefit, aiming to maximize the coverage contribution of the given WRSN. The proposed RLR employs rewards to reflect the coverage benefit and the penalties of sensor's waiting time which are raised due to the charger's movement and battery recharging of those sensors scheduled in a prior order. This helps the mobile charger accumulate the learning experiences from the Q-Table such that the mobile charger learns to execute appropriate actions based on the state management. The contributions of this article are itemized as follows.

1) *Adaptive Recharging Request:* Different from the existing work [15], [16], [17], this article considers the scenario that sensors might have different energy consumption rates due to the different forwarding loads. Therefore, each sensor will adaptively calculate its own recharging threshold to send the recharging request at the right time in a distributed manner.

2) *Learning the Trend of Network Performance:* To the best of our knowledge, this article first employs reinforcement learning to learn the trends of WRSNs, including the energy consumption of the sensors, the recharging cost as well as the coverage benefit. It takes into account the learned trend when scheduling the recharging order, aiming to improve the recharging efficiency and maximize the coverage of the given WRSN. Performance results show that the proposed RLR mechanism outperforms its competitors in terms of the recharging efficiency and the network coverage.

3) *Design of Attraction Reward to Reflect the Coverage Benefit and Penalty of Waiting Time:* The proposed RLR designed the attraction rewards which reflect the coverage benefit and the penalty of waiting time. This helps the mobile charger accumulate the learning experiences from Q-Table such that it executes appropriate actions based on state management. The mobile charger performs an action in each state based on the attraction reward of multiple requested sensors. The learning experiences, stored in Q-Table, help the mobile charger learn to gain the maximal coverage benefit when the mobile charger takes an action in each state.

4) *Consideration of Sensor Coverage Contribution:* Most existing studies [19], [20], [21] considered the remaining energies of sensors and the distances between the requested sensors and the mobile charger, aiming to maximize the number of recharged sensors. Unlike the existing work, this article takes into consideration the coverage of each requested sensor in the design of

the attraction reward, aiming to maximize the surveillance quality of the monitoring region. Performance results demonstrate that the proposed RLR achieves better performance than its competitors in terms of coverage ratio.

The remainder of this article is organized as follows. Section II reviews the existing work related to the recharging issues in WRSNs. Section III presents the network environment and problem statement of the investigated issue. Section IV describes the proposed RLR in detail. Section V gives the simulation experiments and their results. Finally, a conclusion and future work of this article is drawn in Section VI.

II. RELATED WORK

In recent years, the recharging problem based on the mobile charger in WRSNs has received much attention. Plenty of recharging algorithms have been proposed to maintain the lifetime of WRSNs. These studies can be classified into two classes, including offline and online. The following reviews these studies and compares them with this work.

A. Offline Recharging Class

For the offline recharging algorithms, the mobile charger moved along a predefined path to charge sensors periodically, aiming to achieve the perpetual lifetime of the network. Study [13] investigated the paradigm that a mobile charger recharged multiple sensors simultaneously under the constraint of the charger's energy capacity. The charging scheme, named Alg01, constructed a closed charging path, aiming to maximize the accumulative charging utility gain. Meanwhile, the minimization problem of charging path length was discussed if all requested sensors must be charged, assuming that the mobile charger has sufficient energy to support all requested sensors. However, the proposed mechanism aimed to maximize the charging utility without considering the monitoring quality. The mechanism also lacks the support of on-demand service in terms of energy recharging since it did not consider the newly received recharging requests.

Study [14] employed a mobile charger to replenish sensors in the WRSNs. To estimate the energy efficiency of the mobile charger, it proposed a new metric named waste rate, which was defined as a function of the charging channel quality. Then, the energy efficiency optimization was modeled as the problem of minimizing the waste rate. A set of optimal sensor nodes was selected according to the distances between the mobile charger and sensor nodes. Then, a traveling path was constructed for the mobile charger by using the Hamiltonian circle. However, the coverage contribution of each sensor was not considered when selecting the optimal sensor node. Furthermore, it belongs to the offline class which did not dynamically consider the newly received requests.

Study [15] proposed a multinode temporal spatial partial-charging algorithm (MTSPC), aiming to jointly optimize the number of dead sensors and the energy usage effectiveness (EUE). The MTSPC employed the partial charging mechanism, which introduced the concept that the mobile charger can

fully charge a sensor by multiple times. At the beginning of a charging circle, a charging path was constructed for the mobile charger based on the remaining energies of sensors. Then, the mobile charger moved along the charging path to visit a series of selected sojourn locations. If some sensors were unable to be recharged in time, the mobile charger would apply the partial charging policy to rescue them. However, the partial charging mechanism leads to a long charging path, resulting in a low charging utility. In addition, the proposed mechanism belonged to the offline class, which did not consider the newly received recharging requests.

Study [16] proposed a utility-based collaborative charging (UBCC) strategy to maximize the charging utility of mobile chargers in large-scale WRSNs. Different mobile chargers played different roles and served diverse power replenishment areas. A path merging scheme was proposed to save path length and reduce the number of mobile chargers. An idle-difference alleviating scheme was devised to improve the utilization rate of mobile chargers. In addition, an energy-waste averting scheme was designed to maximize the energy utilization of mobile chargers. However, all mobile chargers were allocated a part of the same trajectory, aiming to exploit the opportunities of charging collaboration and reduce path length. The fixed path indicates that the proposed UBCC belongs to the offline class.

B. Online Recharging Class

The online recharging mechanisms timely adjusted the recharging schedule for the mobile charger according to the new received recharging requests from sensors. Therefore, they were able to handle the emergent requests. Furthermore, these algorithms can take advantage of path reduction when the locations of newly requested sensors were closed to the mobile charger.

Study [17] proposed a named data networking (NDN)-based real-time wireless recharging protocol for dynamic wireless recharging in sensor networks. To efficiently deliver sensor energy status information to vehicles in real time, it leveraged concepts and mechanisms from NDN and designed energy monitoring and reporting protocols. The charging schedule was designed according to the minimum weighted sum of traveling time and residual lifetime of sensors. Furthermore, given a set of to-be-charged sensors with different residual lifetimes, an adaptive algorithm was designed to charge a proportion of the sensors before their energy expirations [18]. However, studies [17], [18] aimed to improve the utility of recharging. They did not take into account the coverage of the sensor when determining the recharge schedule. Therefore, it was difficult to guarantee the surveillance quality.

Study [19] proposed a Primary and Passer-by Scheduling (P2S) algorithm for online charging architecture in large-scale WRSNs. In P2S, several primary sensors were selected based on the remaining energy and the distance to the mobile charger. Then, a Hamiltonian path was constructed for the mobile charger to visit the primary sensors. After that, a local searching algorithm was exploited to find nearby requested nodes as the passer-by nodes. Such a strategy not only made full use of the available remaining time of a charging deadline but also solved the complex scheduling problem with spatial and temporal task interdependency. However, similar to studies [17], [18], the surveillance quality was not guaranteed since the coverage contribution of each sensor was not considered when constructing the charging path.

Study [20] developed a temporal–spatial charging scheduling algorithm, namely, TSCA, aiming to minimize the number of dead nodes while maximizing energy efficiency to prolong the network lifetime. The mobile charger determined a feasible movement schedule after collecting a certain amount of charging requests. A basic path planning algorithm was then introduced to adjust the charging order for better efficiency. Then, a node deletion algorithm was developed to remove low efficient charging nodes. Finally, a node insertion algorithm was executed to avoid the death of abandoned nodes. However, the TSCA was unable to learn from the past charging experiences.

Study [21] proposed a charging scheduling mechanism based on the fussy logic, named FLCS, which was effective to deal with the uncertainties in the network. In study [21], different network parameters, such as residual energy, distance to the mobile charger, and critical node density, were blended to make decisions. In addition, it took into account the spatial, temporal, and energy constraints together. However, the coverage contribution of each sensor was not considered.

Study [22] proposed an importance-different charging scheduling (IDCS) strategy for improving charging utility as well as reducing the data loss. The proposed IDCS distinguished sensors in terms of different importance of data delivery. The sensor with greater importance and shorter deadline has a higher priority of being included in the early charging tasks. In addition, the charging sequence in the trajectory of the mobile charger could be adjusted to maximize the charging utility and reduce the path length. However, similar to study [21], the coverage contribution of each sensor was not considered.

Although the studies in the second category constructed the recharging path in an online manner, none of them considered the coverage contribution of the sensors. Hence, the surveillance quality of the WSNs is unpredictable. In addition, the proposed algorithms greedily made decisions according to the considered parameters at the moment, including the remaining energies of the requested sensors and the cost of path length. These lacks learning the trend of energy dissipation from long-time network operations. Different from the previous studies, this article proposes a reinforcement learning approach for the mobile charger to learn the trends of WSNs, including the energy consumption of the sensors, the recharging cost as well as the coverage benefit, aiming to maximize the coverage contribution of the recharged WRSN.

III. NETWORK ENVIRONMENT AND PROBLEM STATEMENT

This section initially introduces the network environment and assumptions of the given WSNs. Then, the problem statements and constraints are described.
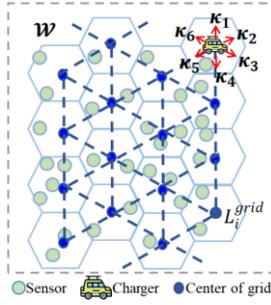
Fig. 1. Network environment.

Fig. 2. Independent coverage.

A. Network Environment and Assumptions

This article considers the following scenario. A given monitoring region $\mathcal{W}$ has been randomly deployed with a set of sensors $B = \{b_1, b_2, \ldots, b_{|B|}\}$. Each sensor b_i is aware of its physical location, denoted by $L_i^{\text{sensor}}(x, y)$ or L_i^{sensor} in short for $1 \leq i \leq |B|$. Each sensor is stationary and powered by a rechargeable battery with the capacity E. It is assumed that the sink node is placed at the center of region $\mathcal{W}$, aiming to collect data from all sensors. A mobile charger M is initially located at the center of the region. It will move to charge sensors in the network once it receives recharging requests from sensors. Let $L_t^{\text{charger}}(x, y)$ or L_t^{charger} in short, denote the location of M at time t. The energy of charger M is assumed to be unlimited. This assumption can be achieved if there is another backup mobile charger is on standby in a fix location. The charging radius is denoted by r^{ch}. The charger M charges sensors at a fixed charging rate α^{ch} and moves at a fixed speed v.

To simplify the actions of charger M, the monitoring region $\mathcal{W}$ is divided into N hexagonal grids $\mathbb{G} = \{g_i, 1 \leq i \leq N\}$. Let ζ denote the side length of the hexagon grid and $L_i^{\text{grid}}(x, y)$ or L_i^{grid} in short denote the center location of each grid g_i. If the charger M arrives at L_i^{grid}, it is said to stay in state s_i. We assume that charger M executes the charging task only if it stays at the grid center and all sensors in that grid can be charged at the same time. Meanwhile, the charger M only moves along the path formed by the centers of grids. Let κ_i denote the ith direction starting clockwise from the north. When a mobile charger locates at the center of a grid, it can move to its neighboring grid in six directions from κ_1 to κ_6, as shown in Fig. 1, or simply performs the recharging operation for the sensors in that grid.

B. Problem Statements

This article aims to develop a recharging mechanism for a mobile charger M based on reinforcement learning, aiming to maximize the coverage contribution of the given network. The following presents the problem formulation of this work.

Let o_i denote the coverage area of sensor b_i, which is a region enclosed by a circle with dashed lines as shown in Fig. 2. Let N_i denote the set of neighboring nodes of b_i. The area only covered by sensor b_i is defined as the independent coverage area of b_i. Let c_i denote the independent coverage
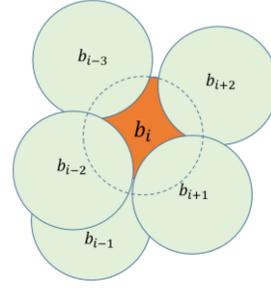
area for sensor b_i, as shown in Fig. 2. c_i can be calculated by

$$c_i = o_i \setminus \bigcup_{b_j \in N_i} o_j. \tag{1}$$

Each sensor b_i has three working states, including *Strong Working*, *Weak Working* and *Sleeping* states. The rechargeable battery of each sensor initially has the fixed capacity E. Since the forwarding loads of different sensors might be different, each sensor b_i has its own energy consumption rate, denoted by α_i^{disch}. Let $e_i^{\text{rem},t}$ denote the remaining energy of sensor b_i at time t. If sensor b_i has rich remaining energy, it is said to stay in Strong Working state. In this state, the remaining energy $e_i^{\text{rem},t}$ is larger than a recharging threshold $e_{\text{th},i}^{\text{ch}}$. If $e_i^{\text{rem},t}$ is reduced to the recharging threshold $e_{\text{th},i}^{\text{ch}}$, sensor b_i should send the recharging request to the mobile charger and switch to the *Weak Working* state. Furthermore, if $e_i^{\text{rem},t}$ is smaller than the sleep threshold $e_{\text{th}}^{\text{sleep}}$, sensor b_i should switch to *Sleeping* state. In the *Sleeping* state, sensor b_i stop its sensing operation for energy conservation. This also leads to the loss of its own independent coverage. Assume that sensor b_i switches to *Sleeping* state at time t_1. Then, the coverage loss of sensor b_i at current time t, denoted by c_i^{loss}, can be calculated by

$$c_i^{\text{loss}} = (t - t_1) * c_i. \tag{2}$$

Let $\lambda_i^{\text{work},t}$ denote a Boolean variable, representing whether or not sensor b_i works at time t. That is

$$\lambda_i^{\text{work},t} = \begin{cases} 1, & e_i^{\text{rem},t} \geq e_{\text{th}}^{\text{sleep}} \\ 0, & e_i^{\text{rem},t} < e_{\text{th}}^{\text{sleep}}. \end{cases} \tag{3}$$

Let τ denote a unit of time slot. Considering a given time period T which can be divided into many slots, denoted by $T = \{\tau_1, \ldots, \tau_x\}$, where $x = (T/\tau)$. Let C_j^{loss} denote the total coverage loss of all sensors during time slot τ_j. C_j^{loss} can be calculated by

$$C_j^{\text{loss}} = \sum_{b_i \in B} \left(1 - \lambda_i^{\text{work},\tau_j}\right) * c_i. \tag{4}$$

Recall that $\mathcal{W}$ denotes the considered monitoring region. Expression (5) presents the objective function of this article.

Objective: This article aims to maximize the coverage contribution of all sensors in the given network during a given time period T. That is

$$\max\left(\sum_{\tau_j \in T} \left(\mathcal{W} - C_j^{\text{loss}}\right)\right). \tag{5}$$

C. Constraints

The size of the grid can impact the charging and movement efficiencies of the mobile charger. Since the mobile charger M is assumed to be able to charge all sensors in the grid when it locates at the grid center, the side length ζ of the grid should not be larger than the charging radius r^{ch} of the mobile charger. On the contrary, the grid size should have a lower bound. This occurs because a smaller grid can cause that the mobile charger should maintain a large number of states. In case that there is no sensor in a small grid, it is impossible to execute the recharging action. This will result in the inefficiency management of the state.

The following derives the lower bound of the grid size. Recall that notations $\mathcal{W}$ and $|B|$ denote the monitoring region and number of sensors, respectively. The area size of a grid can be calculated by the expression ($[3\sqrt{3}\zeta^2]/2$). Let $|\mathcal{W}|$ denote the area size of region $\mathcal{W}$. The number of grids can be computed by the expression ($2|\mathcal{W}|/[3\sqrt{3}\zeta^2]$). Since it is expected that each grid has at least one sensor, we have ($2\mathcal{W}/[3\sqrt{3}\zeta^2]$) $\leq |B|$. The following grid constraint reflects the upper and lower bounds of each grid size.

1) Grid Constraint: The grid size should satisfy the following expression.

$$\sqrt{\frac{2\mathcal{W}}{3\sqrt{3}|B|}} \leq \zeta \leq r^{\text{ch}}. \tag{6}$$

Recall that each sensor b_i has three states which are related to the remaining energy $e_i^{\text{rem},t}$. Let $\lambda_i^{\text{request},t}$ denote a Boolean variable, representing whether or not sensor b_i has sent a recharging request to a mobile charger at time slot t. That is

$$\lambda_i^{\text{request, } t} = \begin{cases} 1, & e_i^{\text{rem},t} \leq e_{\text{th},i}^{\text{ch}} \\ 0, & e_i^{\text{rem},t} > e_{\text{th},i}^{\text{ch}}. \end{cases} \tag{7}$$

Let φ_i^{SW}, φ_i^{WW} and φ_i^{Sleep} denote three Boolean variables representing whether or not sensor b_i is in the *Strong Working* state, *Weak Working* state, and *Sleeping* state, respectively. Recall that notation $\lambda_i^{\text{work},t}$ denotes a Boolean variable which represents whether or not sensor b_i works at time t. We have

$$\varphi_i^{\text{SW}} = \begin{cases} 1, & \lambda_i^{\text{work},t} = 1, \lambda_i^{\text{request, } t} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\varphi_i^{\text{WW}} = \begin{cases} 1, & \lambda_i^{\text{work},t} = 1, \lambda_i^{\text{request, } t} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$\varphi_i^{S} = \begin{cases} 1, & \lambda_i^{\text{work},t} = 0, \lambda_i^{\text{request, } t} = 1 \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

The following presents the *sensor state constraint*.

2) Sensor State Constraint: Any sensor b_i can only stay in one of three states at any time slot. This constraint is given as shown in (11).

$$\varphi_i^{\text{SW}} + \varphi_i^{\text{WW}} + \varphi_i^{S} = 1, \text{ for } \forall b_i \in B. \tag{11}$$

In addition to the sensor state constraint, another constraint restricts that each sensor should send the recharging request to the mobile charger before it entering the sleeping state. The following presents the recharging request constraint.

3) Recharging Request Constraint: Any sensor b_i should send the recharging request to the mobile charger before it entering the sleeping state. This constraint is given as shown in (12).

$$\lambda_i^{\text{request},t} \geq \lambda_i^{\text{work},t}, \text{ for } \forall b_i \in B. \tag{12}$$

The goal, as shown in (5), of this article is to maximize the coverage contribution of all sensors in the given network during a given time period T while satisfying the Grid Constraint, Sensor State Constraint as well as the Recharging Request Constraint.

IV. PROPOSED RECHARGING MECHANISM

This section presents the proposed recharging mechanism that aims to achieve the goal given in (5) while satisfying constraints given in (6), (11), and (12). Each sensor, say b_i, should send a recharging request to the mobile charger M as soon as its remaining energy is lower than the recharging threshold $e_{\text{th},i}^{\text{ch}}$. Since the charger M keeps moving, sensors that need to be charged should acquire the location of charger M before sending the recharging request. The charger M should learn a recharging schedule by applying the proposed reinforcement learning. The proposed recharging mechanism mainly consists of three modules, including sensor energy management (SEM), charger location update (CLP) and charger reinforcement learning (CRL). In the SEM, the major task of each sensor is to manage its energy, including managing its remaining energy, determining its own request threshold as well as sending the recharging request to the mobile charger. The CLP mainly updates the location of the mobile charger using the Quorum system, aiming to help sensors acquire the location of charger M before sending the recharging requests. In the CRL, the mobile charger M aims to learn a better recharging schedule by adopting the proposed reinforcement learning algorithm.

A. Sensor Energy Management

Each sensor will manage its operations according to the proposed state diagram. As shown in Fig. 3, each sensor stays in one of the three possible states, including *Strong Working, Weak Working,* and *Sleeping states*. The change of states depends on the remaining energy of the sensor. Each sensor initially has a full energy and stays in the Strong Working state. In this state, each sensor contributes its coverage and sends its readings to the static sink in a multihop manner.

The following uses sensor b_i to present the operations of each sensor. As long as the remaining energy $e_i^{\text{rem},t}$ is smaller than the threshold $e_{\text{th},i}^{\text{ch}}$, sensor b_i switches to the Weak Working state and sends the recharging request to the mobile charger M to ask for recharging. Since the mobile charger M keeps moving, sensor b_i should acquire its location before sending the recharging request. To reduce the communication costs, the CLP procedure is proposed, which will be detailed in the next section. After obtaining the location of charger M, sensor b_i sends a *recharging request (CReq)* packet to charger M using the greedy perimeter stateless routing (GPSR) [2]. The content of *CReq* packet can be expressed
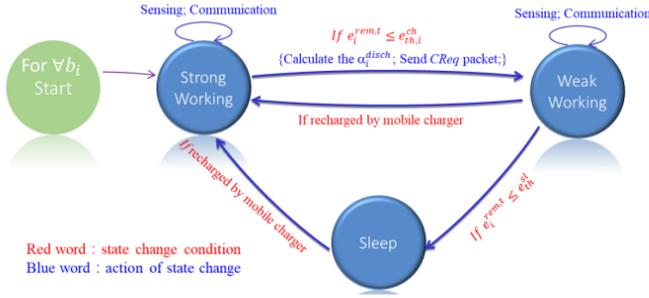
Fig. 3. Sensor state diagram.

Fig. 4. Construction of a quorum system.

as $[b_i, L_i^{\text{sensor}}, e_i^{\text{rem},t}, \alpha_i^{\text{disch}}, L_{t_{\text{now}}}^{\text{charger}}, t_{\text{now}}]$, where α_i^{disch} is the energy consumption rate of sensor b_i at current time t_{now}. Since the communication overloads of different sensors are different, the values of α_i^{disch} can be different. The following presents the calculation of α_i^{disch}.

Let $\mho$ denote a fixed time period used to compute α_i^{disch}. Recall that $e_i^{\text{rem},t}$ denotes the remaining energy of sensor b_i at time t. Similarly, $e_i^{\text{rem},t-\mho}$ denotes the remaining energy of sensor b_i at time $t - \mho$. Therefore, the recent energy consumption rate during time period $[t, t - \mho]$ can be expressed by $([(e_i^{\text{rem},t-\mho} - e_i^{\text{rem},t})]/\mho)$. Let $\alpha_i^{\text{disch}}[t]$ denotes energy consumption rate from the initial time t_0 to time t. Then, the value of $\alpha_i^{\text{disch}}[t]$ can be evaluated by applying

$$\alpha_i^{\text{disch}}[t] = (1 - \rho)\alpha_i^{\text{disch}}[t - \mho] + \frac{\rho * \left(e_i^{\text{rem},t-\mho} - e_i^{\text{rem},t}\right)}{\mho} \tag{13}$$

where ρ is a coefficient to adjust the weights of historical energy consumption rate $\alpha_i^{\text{disch}}[t - \mho]$ and recent energy consumption rate. When the value of ρ is larger, α_i^{disch} is more dependent on the recent energy consumption rate.

Since the α_i^{disch} values of different sensors might be different, the thresholds $e_{\text{th},i}^{\text{ch}}$ for sending the $CReqs$ could be different. According to its energy consumption rate, each sensor b_i should determine the threshold $e_{\text{th},i}^{\text{ch}}$ and the time of sending $CReq$ packet. Let t be the current time. Assume that sensor b_i has sent the $CReq$ packet k times and has been recharged k times. Let $t_{i,k}^{\text{request}}$ and $t_{i,k}^{\text{ch}}$ denote the kth time for sending the $CReq$ packet and the kth recharged time of sensor b_i, respectively. Then, the average waiting time for recharging of sensor b_i at current time t, denoted by $T_{i,t}^{\text{wait}}$, can be expressed by

$$T_{i,t}^{\text{wait}} = \frac{\sum_k \left(t_{i,k}^{\text{ch}} - t_{i,k}^{\text{request}}\right)}{k}. \tag{14}$$

Then, the value of $e_{\text{th},i}^{\text{ch}}$ can be calculated by

$$e_{\text{th},i}^{\text{ch}} = \alpha_i^{\text{disch}} * T_{i,t}^{\text{wait}}. \tag{15}$$

It implies that sensor b_i should send a $CReq$ packet to ask for recharging when its remaining energy is smaller than $e_{\text{th},i}^{\text{ch}}$.

Staying in the *Weak Working* state, sensor b_i still contributes its coverage and sends its readings to the sink node periodically. In case that it is actually recharged by charger M, it switches to the *Strong Working* state. Once the remaining
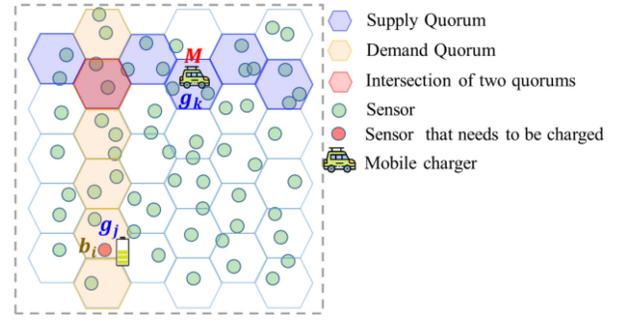
energy $e_i^{\text{rem},t}$ is smaller than the threshold $e_{\text{th}}^{\text{sleep}}$, sensor b_i switches to the *Sleeping* state. In the *Sleeping* state, sensor b_i stops to perform sensing operation. Once it is recharged by charger M, it switches to the *Strong Working* state.

B. Charger Location Update

To minimize the coverage loss, sensor b_i should send the recharging request to the mobile charger M and expect that M can visit it within its average waiting time $T_{i,t}^{\text{wait}}$. However, the location of M might be changed with time. To efficiently deliver the location updates of charger M, the mobile charger and each sensor will cooperatively perform the following location update procedure in a distributed manner. This can help each sensor maintain the current location of the mobile charger M with a low communication cost. The following presents the location update procedure.

The location update procedure is designed based on the quorum system. A quorum system is a set of subsets (called quorums) such that any two quorums have a nonempty intersection. Consider two quorums: 1) supply and 2) demand quorums. Let q_t^{supply} and $q_{i,t}^{\text{demmend}}$ denote the supply and demand quorums, which are constructed by mobile charger M and sensor b_i at time slot t, respectively. The following will introduce the construction of a quorum system by using an example as shown in Fig. 4.

Assume that mobile charger M arrives at the center of the grid g_a at the current time t_{curr}. Let $L_{t_{\text{curr}}}^{\text{charger}}(x, y)$ denote the location of M at time slot t_{curr}. The mobile charger M should actively construct a supply quorum $\mathcal{Q}_{t_{\text{curr}}}^{\text{supply}}$ which includes the grids marked with blue ink as show in Fig. 4 at the current time slot t_{curr}. That is

$$\mathcal{Q}_{t_{\text{curr}}}^{\text{supply}} = \left\{ g_k | L_k^{\text{grid}} \cdot x = L_{t_{\text{curr}}}^{\text{charger}} \cdot x \pm \frac{3\zeta}{2} l \right.$$
$$L_k^{\text{grid}} \cdot y = \begin{cases} L_{t_{\text{curr}}}^{\text{charger}} \cdot y & l\%2 = 0 \\ L_{t_{\text{curr}}}^{\text{charger}} \cdot y + \frac{\sqrt{3}\zeta}{2} & l\%2 = 1 \end{cases}$$
$$\left. \text{s.t. } 0 \leq L_k^{\text{grid}} \cdot x \leq \sqrt{\mathcal{W}}, \text{ where } l = \{0, 1, 2, \ldots\} \right\}.$$

To physically construct the supply quorum $\mathcal{Q}_{t_{\text{curr}}}^{\text{supply}}$, the following operations should be applied by the WRSNs. For each grid g_k, the sensor which locates in g_k and is nearest to the center of g_k will be treated as the header sensor of g_k, denoted by $g_k \cdot$header. The mobile charger M broadcasts quorum construction request (QCReq) packet to the headers of grids belonging

to the supply quorum $\mathcal{Q}_{t_{curr}}^{supply}$ along the horizontal direction. The QCReq packet can be expressed as $[M, L_{t_{curr}}^{charger}, t_{curr}]$, where M denotes that the packet is sent from the mobile charger and $L_{t_{curr}}^{charger}$ denotes the location of M at the current time slot t_{curr}. As a result, the supply quorum $\mathcal{Q}_{t_{curr}}^{supply}$ can be formed. It is noted that only the grid headers $g_k \cdot$header should cooperatively maintain the supply quorum $\mathcal{Q}_{t_{curr}}^{supply}$ until they receive a quorum cancelation request (QCCan) packet from the mobile charger M. The QCCan packet can be expressed as $[M, t_{curr}]$. This indicates that the mobile charger has arrived another new grid and intends to construct a new supply quorum $\mathcal{Q}_{t_{curr}}^{supply}$ at time slot t_{curr}.

Assume that sensor b_i is located at the center location $L_j^{grid}(x, y)$ of grid g_j. When sensor b_i aims to be recharged at time slot t_2, it should query the location of M. To achieve this, sensor b_i will construct a demand quorum $\mathcal{Q}_{i,t_2}^{demand}$. As shown in Fig. 4, the formed quorum $\mathcal{Q}_{i,t_2}^{demand}$ contains the grids marked with yellow ink. That is

$$\mathcal{Q}_{i,t_2}^{demand} = \left\{ g_k | L_k^{grid} \cdot x = L_j^{grid} \cdot x \right.$$
$$L_k^{grid} \cdot y = L_j^{grid} \cdot y \pm \sqrt{3}\zeta l$$
$$\text{s.t. } 0 \le L_k^{grid} \cdot y \le \sqrt{\mathcal{W}}$$
$$\left. \text{where } l = \{0, 1, 2, \ldots\} \right\}.$$

To construct the demand quorum $\mathcal{Q}_{i,t_2}^{demand}$, sensor b_i will broadcast a mobile charger location request (MLReq) packet to those grid headers belonging to its demand quorum $\mathcal{Q}_{i,t_1}^{demand}$ along the vertical direction. The MLReq packet can be expressed as $[b_i, L_i^{sensor}, t_{now}]$, where b_i denotes the sensor's ID, L_i^{sensor} denotes the location of sensor b_i and t_{now} is the current time.

Let g_{i,t_2}^{intsec} denote the intersection grid of $\mathcal{Q}_{t_2}^{supply}$ and $\mathcal{Q}_{i,t_2}^{demand}$ at current time slot t_2. We have $g_{i,t_2}^{intsec} = \mathcal{Q}_{t_2}^{supply} \cap \mathcal{Q}_{i,t_2}^{demand}$, which is marked with pink ink as shown in Fig. 4.

The header in g_{i,t_2}^{intsec} will be responsible to notify the sensor b_i about the receipt of MLReq packet by replying the mobile charger location response (MLRes) packet using GPSR [2]. The content of MLRes packet can be expressed as $[M, L_{t_{now}}^{charger}, t_{now}]$, where M denotes the mobile charger, $L_{t_{now}}^{charger}$ denotes M's location at time t_{now}.

According to the location update procedure, only a small number of headers maintain the location of mobile charger. Meanwhile, each sensor can acquire the current location of the mobile charger M with a low communication cost. Furthermore, each sensor can send a recharging request to M quickly when it needs to be recharged.

C. Charger Reinforcement Learning

This section presents the reinforcement learning algorithm for the mobile charger M to learn a recharging schedule, aiming to maximize the surveillance quality provided by the recharged sensors. Initially, charger M stays at the center of the monitoring area. Once M receives a CReq packet from any sensor b_i, it activates the reinforcement learning and moves to charge the requested sensor according to the Q-Table. Let

U denote the set of sensors which have already sent the CReq packets to M. Charger M learns its recharging schedule according to sensors in U.

Each sensor $b_i \in U$ will form an attraction force, aiming to pull charger M to recharge b_i itself. Furthermore, M learns its recharging schedule based on the attraction generated by all sensors $b_i \in U$. The following presents the reinforcement learning of M.

Definition: The reinforcement learning, denoted by **RL(S, A, AR, Q)**, for mobile charger M consists of four tuples, including **S**, **A**, **AR**, and **Q**, where:

1) $S = \{s_0, s_1, \ldots, s_N\}$ denotes the set of learning states for charger M, where N is the number of grids in the monitoring region;
2) $A = \{a_0, a_1, a_2, \ldots, a_6\}$ denotes the set of all possible actions that mobile charger M can take in a special learning state $s_k \in S^M$. In detail, a_0 denotes the charging action while a_1 to a_6 denote the moving actions;
3) AR is the attraction reward matrix which is used to store attraction generated by sensors in set U. Meanwhile, it is used to evaluate how Good/Bad an action is for a specific state;
4) Q is the Q-Learning function for evaluating $Q(s_k, a_l)$ which predicts the best action to be executed for a specific state.

The following presents the design of each element in **RL**.

Learning States S: The set S of states is the collection of all possible states s_k, $1 \le k \le N$. That is, $S = \{s_k | 1 \le k \le N\}$. The definition of the learning state is given as follows.

Definition of Learning State: The learning state is the location of the mobile charger. More specially, charger M is said to be in state $s_k \in S$, where $1 \le k \le N$, if it stays in grid g_k.

The mobile charger M is said to be in state $s_k \in S$, where $1 \le k \le N$, if it stays in grid g_k. When charger M stays in state s_k, it should check whether or not it should charge sensors in grid g_k. Let $B_k = \{b_1^k, b_2^k, \ldots, b_{|B_k|}^k\}$ denote the set of sensors located in g_k. If any sensor $b_j^k \in B_k$ has sent the CReq packet to M, charger M should execute the charging task in state s_k. It is also said that charger M should take the action a_0 in state s_k.

The following presents the charging criterion of the mobile charger staying in state s_k.

Charging Criterion: Charger M will execute the charging task in state s_k, only if the following criterion is satisfied:

$$\sum_{b_j^k \in B_k} \lambda_j^{request, \, t} \ge 1. \tag{16}$$

Actions A: There are seven actions in set A. That is, $A = \{a_l | 0 \le l \le 6\}$. Action a_0 indicates that mobile charger M executes the charging task while action $a_l \in A (l \ne 0)$ indicates that mobile charger M moves to the neighboring grid along direction κ_l.

Attraction Reward AR: The mobile charger M maintains an attraction reward matrix which stores the attraction generated by sensors $b_i \in U$. Charger M can evaluate the reward of an action in a specific state based on the attraction reward matrix. The following presents the calculation of the attraction reward.
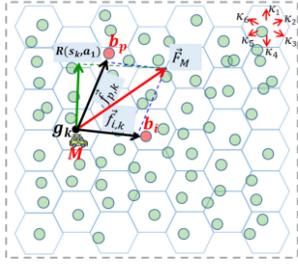
Fig. 5. Example of attraction.

Fig. 6. Attraction reward matrix of M_j.

As shown in Fig. 5, assume that the charger M locates at the center of grid g_k at time t.

Consider a sensor $b_i \in U$. Let $\overrightarrow{f}_{i,k}$ denote the attraction force generated by sensor $b_i \in U$ on charger M located in grid g_k. The direction of $\overrightarrow{f}_{i,k}$ is from g_k to b_i. The magnitude of $\overrightarrow{f}_{i,k}$ can be calculated by

$$|\overrightarrow{f}_{i,k}| = \begin{cases} \frac{\beta * c_i}{d(b_i, g_k)}, & L_{j,t}^{\text{charger}} = L_k^{\text{grid}} \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where β is a coefficient to adjust the weights of independent coverage c_i and the distance between sensor b_i and grid g_k.

Considering that each sensor $b_i \in U$ generates its attraction force. The total attraction on charger M in grid g_k, denoted by $\overrightarrow{F}_k$, can be calculated by

$$\overrightarrow{F}_k = \sum_{b_i \in U} \overrightarrow{f}_{i,k}. \quad (18)$$

Fig. 5 gives an example of the calculation of total attraction $\overrightarrow{F}_k$ when set U has two sensors b_i and b_p. Assume that mobile charger M is in grid g_k. This also indicates that M is in state $s_k \in S$. The movement of M will be affected by $\overrightarrow{F}_k$. Since M can move along one of six directions, the attraction of one direction will be treated as the reward for M moving along that direction. Let $R(s_k, a_l)$ denote the attraction reward that M takes action $a_l \in A$ in state $s_k \in S$. The value of $R(s_k, a_l)$ can be calculated by

$$R(s_k, a_l) = |\overrightarrow{F}_k| * \cos < \overrightarrow{F}_k, a_l > \quad (19)$$

where $\cos < \overrightarrow{F}_k, a_l >$ represents the cosine of the angle between attraction $\overrightarrow{F}_k$ and the direction κ_l of action a_l. As shown in Fig. 6, M can obtain reward $R(s_k, a_1)$ if it moves along direction κ_1. Based on (19), M can obtain its attraction reward matrix or AR in short, as shown in Fig. 6. The attraction generated by sensors $b_i \in U$ will fade over the time. Let ϱ denote the attenuation rate of the attraction. It indicates that each $R(s_k, a_l)$ in AR will reduce ϱ per time slot until the $R(s_k, a_l)$ is reduced to 0. Besides, AR will be updated when M

receives a new *CReq* packet. According to its *AR*, M obtains the reward $R(s_k, a_l)$ when it takes action a_l in state s_k.

The design of attraction reward indicates that sensors which have larger coverage contribution and short distance from the mobile charger will have higher priority to be charged. It is conducive to achieve the objective represented by (5).

Q-Learning Function Q: Assume M stays in state s_p after executing action a_l in state s_k. Then, the Q-Learning function $Q(s_k, a_l)$ is given in

$$Q(s_k, a_l) = Q(s_k, a_l) + \alpha * \left[R(s_k, a_l) + \gamma * \max_{a_q \in A} Q(s_p, a_q) - Q(s_k, a_l) \right] \quad (20)$$

where α is the learning rate and γ is the discount factor of Q-Learning. In addition, $\max_{a_q \in A} Q(s_p, a_q)$ indicates the max value of Q in the next state s_p. The expression $[R(s_k, a_l) + \gamma * \max_{a_q \in A} Q(s_p, a_q)]$ represents the current reward obtained by executing action a_l in state s_k, including the immediate and future parts. Furthermore, the expression $[R(s_k, a_l) + \gamma * \max_{a_q \in A} Q(s_p, a_q) - Q(s_k, a_l)]$ calculates the difference between the past experience and the current reward. In summary, expression (20) calculates the new experience after executing the action a_l in state s_k.

Besides, the ϵ-greedy movement policy is employed by mobile charger M to explore the probability $\theta(a_l | s_k)$ of taking action a_l when M is in state s_k. The value of $\theta(a_l | s_k)$ can be calculated by

$$\theta(a_l | s_k) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A|}, & a_l = \arg\max_{a_q \in A} Q(s_k, a_q) \\ \frac{\epsilon}{|A|}, & a_l \neq \arg\max_{a_q \in A} Q(s_k, a_q) \end{cases} \quad (21)$$

where ϵ is a random action exploration probability of trade-off between exploiting the current knowledge and exploring the environment to select an action, and $|A|$ is the number of actions. Let $\sigma_{j,k}$ denote a Boolean variable representing whether or not M has ever taken action a_j in state s_k. That is

$$\sigma_{j,k} = \begin{cases} 1, & a_j \text{ has been taken when } M \text{ is in } s_k \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

Then, the value of ϵ can be calculated by

$$\epsilon = 1 - \frac{\sum_{j=1}^{|A|} \sum_{k=1}^{|S|} \sigma_{j,k}}{|S| * |A|}. \quad (23)$$

Herein, charger M can estimate the probability $\theta(a_l | s_k)$ ($0 \leq l \leq |A|, 1 \leq k \leq |S|$) based on (21). Assume that action a_l has the largest probability when M stays in state s_k. That is

$$a_l = \arg\max_{a_i \in A} \theta(a_i | s_k). \quad (24)$$

It means that charger M learns to take action a_l in state s_k. Furthermore, it indicates that M moves to the neighboring grid, say g_u, of g_k along direction κ_l. If any sensor in U is located in the new grid g_u, it will charge all sensors in grid g_u. Otherwise, it learns to move to the next grid based on (21) and (24). After learning for a while, M will learn a recharging schedule for recharging sensors according to the received *CReq* packets and finally can charge sensors according to the schedule.

Here, the proposed RLR mechanism has been described. In the CRL module, Q-Learning algorithm is applied to help the charger learn a recharging schedule. Therefore, the proposed RLR mechanism can be denoted as *RLR-QL* specifically. The state-action-reward-state-action (SARSA) is a more conservative reinforcement learning algorithm compared with Q-Learning. SARSA can play the role of QL in the CRL module, which induces the proposed RLR-SARSA mechanism. The effect of different reinforcement learning algorithms on RLR will be discussed in Section V.

V. SIMULATION

This section investigates the performance improvement of the proposed RLR by comparing with the existing Alg01 [13] and FLCS [21]. The Alg01 aims to maximize the accumulative charging utility gain by greedily reducing the charging path length. The FLCS aims to effectively deal with the requested sensors by applying fuzzy logic. Both of them did not consider the coverage contribution of each sensor when making the decision of charging schedule. The following presents the simulation environment, performance metrics, and the simulation results.

A. Simulation Setup

The simulation is executed on the MATLAB R2021a platform. A WRSN consists of 200 sensors which are randomly deployed in the 200 m * 200 m region. The sensing radius of each sensor is set at 10 m. The battery capacity E of each sensor is set at 10.8 kJ [18]. The residual energy of each sensor is randomly set at a value between $0.1E$ to $1E$ before the simulation starts. The initial recharging threshold $e_{\text{th},i}^{\text{ch}}$ of each sensor is set at $0.2E$. The energy consumption rate α_i^{disch} of each sensor varies ranging from 0.1 to 0.2 J/s. The monitoring region $\mathcal{W}$ is divided into multiple hexagonal grids. The length of the hexagonal grid varies ranging from 10 and 20 m while the charging radius of the mobile charger is set at a value between 10 and 20 m. This guarantees the satisfaction of the proposed *Grid Constraint* as shown in (6). The charging rate of the mobile charger is set at 5 J/s [21]. The moving speed and the moving cost of the mobile charger are set at 1 m/s and 50 J/m, respectively. The learning rate α is set at 1 and the value of discount factor γ varies ranging between 0.1 and 1. The initial values of elements in the Q-Table are zero. The simulation parameters are listed in Table I.

B. Performance Metrics

The performance metrics mainly include the average waiting time, the EUE, the number of sleeping sensors and the accumulated coverage contribution of all sensors.

1) Average Waiting Time: Recall that notation $T_{i,t}^{\text{wait}}$ denotes the average waiting time of sensor b_i at the current time t and it can be calculated based on (14). Then, the average waiting time of all sensors deployed in the network at current time t, denoted by T_t^{wait}, can be calculated by

$$T_t^{\text{wait}} = \frac{\sum_{i=1}^{|B|} T_{i,t}^{\text{wait}}}{|B|}. \qquad (25)$$

TABLE I
SIMULATION PARAMETERS

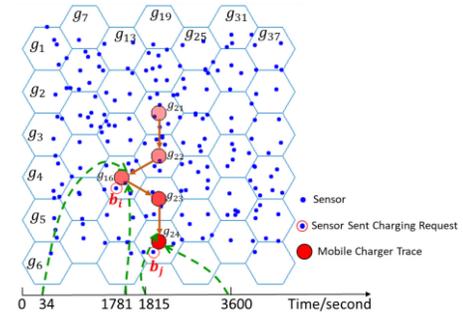| Parameters	Values
Monitoring region $\mathcal{W}$	200 m * 200 m
Sensor Number $|\mathbf{B}|$	200
Sensing Radius $r^{sensing}$	20 m
Sensor Battery Capacity E	10.8 KJ
Energy Consumption Rate α_i^{disch}	0.1~0.2J/s
Grid Length ζ	10~20 m
Charging Radius r^{ch}	10~20m
Charging Rate α^{ch}	5 W
Moving Speed v	1 m/s
Learning Rate α	0.9~1
Discount Factor γ	0.1~1
Simulation Time	200h

Fig. 7. Network snapshot after it operates 3600 s.

2) EUE: The EUE of the mobile charger is an evaluation indicator of the charging algorithm. Let E^{ch} and E^{tr} denote the energies of mobile charger consumed for recharging or moving, respectively. That is

$$\text{EUE} = \frac{E^{\text{ch}}}{E^{\text{ch}} + E^{\text{tr}}}. \qquad (26)$$

The EUE reflects the ratio of the energy of the mobile charger consumed for recharging sensors. An inefficient schedule of moving along a long path will result in a small value of EUE. On the contrary, a large EUE value indicates that the recharging schedule achieves a large ratio of energy consumption for recharging the sensors, instead of movement.

3) Accumulated Coverage Ratio: Recall the objective function given in (5) which aims to maximize the coverage contribution of all sensors. The accumulated coverage ratio, denoted by *ACR*, is further defined to evaluate the network coverage achieved by an algorithm. The *ACR* can be expressed by

$$\text{ACR} = \frac{\sum_{\tau_j \in T} (\mathcal{W} - C_j^{\text{loss}})}{\sum_{\tau_j \in T} (\mathcal{W})}. \qquad (27)$$

C. Simulation Results

Figs. 7 and 8 depict the network snapshot and the contents of Q-Table, respectively, after the network operates for 3600 s. The grid length ζ is set to 20 m and the number of sensors is set to 200. The mobile charger is initially located at the center of the grid g_{21} (i.e., s_{21}). It periodically constructs a supply quorum to update its location to the sensors. As shown in Fig. 7, the initial energies of sensors b_i and b_j are set

Action Grid/State		North a_1	North East a_2	South East a_3	South a_4	South West a_5	North West a_6
g_{16}	s_{16}	0	0	21.41	0	0	0
g_{17}	s_{17}	0	0	0	0	0	0
g_{18}	s_{18}	0	0	0	0	0	0
g_{19}	s_{19}	0	0	0	0	0	0
g_{20}	s_{20}	0	0	0	0	0	0
g_{21}	s_{21}	0	0	0	36.25	0	0
g_{22}	s_{22}	0	0	0	0	43.79	0
g_{23}	s_{23}	0	0	0	28.11	0	0

Fig. 8.   Snapshot of the Q-Table.
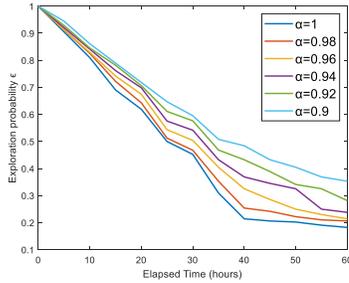
Fig. 9.   Effect of learning rate on exploration probability.

Fig. 10.   Comparison of average waiting time for different γ.

Fig. 11.   EUE comparison of mobile charger for different γ.

below 0.2E. they should first query the location of the mobile charger based on the CLP procedure and then send the recharging requests to the mobile charger. On receiving the requests, the mobile charger applies the reinforcement learning $RL(S, A, AR, Q)$ and selects the action a_4 in state s_{21} based on (24). Then, the mobile charger moves from grid g_{21} to grid g_{22}. Therefore, it switches from state s_{21} to state s_{22} accordingly and updates the value of $Q(s_{21}, a_4)$ to 36.25, as shown in Fig. 8. As long as the mobile charger decides to move to a new grid, the Q-Table is also updated. Let $\pi^{op}=[t_1, t_2]$ denote that the mobile charger executes the operation op during the period from t_1 to t_2. As shown in Fig. 7, a time line is given at the bottom to show the duration of each action taken by the mobile charger M. The mobile charger starts moving and finally arrives grid g_{22} during $\pi^{move} = [0, 34]$. According to the *Charging Criterion*, the mobile charger charges all sensors located in the grid g_{16} during time period $\pi^{charge}=[34, 1781]$. Afterward, it moves to grid g_{24} passing through g_{23} during $\pi^{move}=[1781, 1815]$. Finally, it charges all sensors located in grid g_{24} during $\pi^{charge}= [1815,3600]$.

The learning rate α affects the learning speed of the proposed RLR mechanism. The proposed RLR aims to learn a charging schedule by exploring state–action space in the network. During the experiment, it is found that when the learning rate α is set at a small value, the charger will be easily trapped in the grids that have been charged due to the experiences preserved in the Q-Table. It means that it could not to exploring more states and actions. Therefore, to make the charger get rid of the influence of local optimal solution, the learning rate is set at a larger value. Recall (23), the exploration probability ϵ is determined by the number of the explored states and actions. Furthermore, a smaller value of ϵ means that the charger has explored a larger number of states and actions. Fig. 9 shows the effect of learning rate α on exploration probability ϵ. Considering the boundary effect, ϵ is approximately equal to 0.18 when the mobile charger has
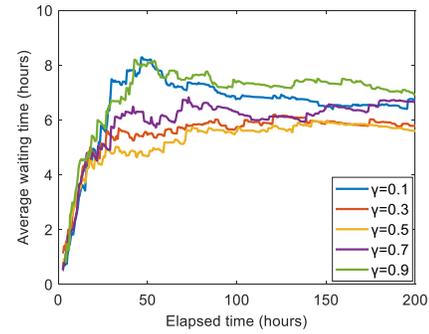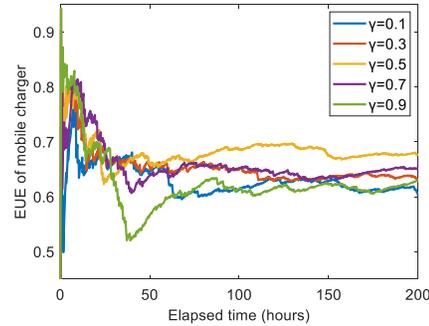
explored all available states and actions. As shown in Fig. 9, the larger the value of α, the faster the value of ϵ decreases, which means the faster the chargers can explore. In addition, when the charger does not fully explore all available states and actions, the learning is not sufficient, which will lead to the underfitting problem. Therefore, to avoid the underfitting problem, the algorithm should at least learn in ϵ approximation 0.18. To better discuss the other parameters, the parameter α will be set at 1 in the following experiments.

The parameter γ reflects the tradeoff between the immediate attraction reward and the accumulated experiences, which highly impact the waiting time, the energy usage efficiency as well as the network coverage. Figs. 10–12 aim to find an appropriate setting value of γ for the later experimental studies. Fig. 10 investigates the average waiting time of all sensors during the elapsed time under different discount factor γ. As shown in Fig. 10, there is a common trend that the average waiting time increases first, then decreases and finally almost levels off with the elapsed time. This occurs because a small number of sensors needs to be recharged at the beginning of the simulation and the mobile charger can quickly learn to charge those requested sensors. However, as time goes by, more and more sensors need to be charged and the mobile charger needs more time to accumulate the experiences of energy dissipation trend. This results in an increase in average waiting time. Furthermore, when the mobile charger learns enough experiences, it can make a better charging schedule, leading to a decrease in average waiting time. Fig. 10 shows that reinforcement learning can reduce the average waiting time. Finally, the mobile charger achieves a stable curve of learning experiences maintained in Q-Table. In comparison,
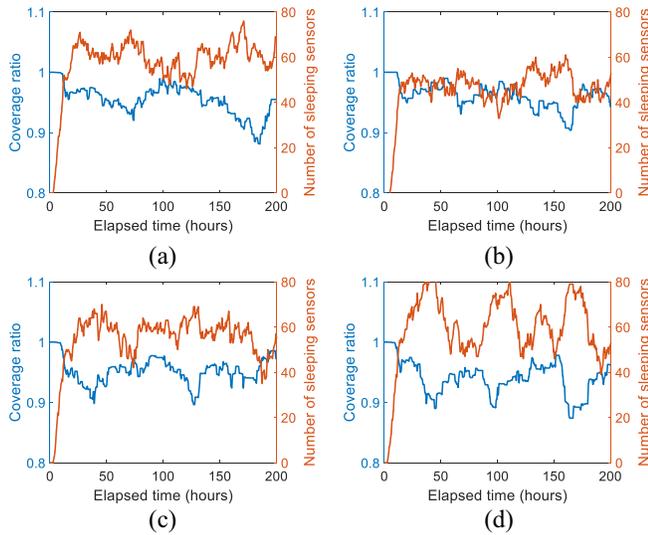
Fig. 12. Coverage ratio and the number of sleeping sensors under different γ. (a) $\gamma = 0.3$. (b) $\gamma = 0.5$. (c) $\gamma = 0.7$. (d) $\gamma = 0.9$.

Fig. 13. Comparison of average waiting time at different grid lengths.

Fig. 14. EUE comparison of mobile charger at different grid lengths.

when γ is set at 0.5, the convergence of the average waiting time is fastest and the values of average waiting time tend to be smaller. This occurs because of several reasons. First, if γ is set at a small value, the scheduling decision depends more on the immediate reward when deciding an action. This causes that the mobile charger has a smaller weight on the accumulated experience and hence decreases the learning speed. On the contrary, if γ is set at a large value, the movement of the mobile charger is severely influenced by experience, which indicates that the mobile charger is more likely to charge sensors that have been charged before, leading to a longer waiting time of those sensors that have not been charged.

Fig. 11 further presents the relation of EUE and the elapsed time by varying the discount factor γ under the same parameter settings of Fig. 10. As shown in Fig. 11, there is a common trend that the EUE rises sharply, then falls sharply, before leveling off. In the initial time period, the mobile charger is located at the center location of the region. The average distance between the mobile charger and the requested sensors is short. Therefore, the ratio of energy consumption for movement is small. However, when the mobile charger moves to the boundary region, the average distance between the requested sensors and the mobile charger is relatively large, as compared with the initial time period. As a result, the EUE drops sharply. Finally, the mobile charger learns the experience of the energy dissipation of the WSNs from the Q-Table and makes better schedules which discard the worst situations including long path but small gains in terms of the coverage. Therefore, the EUE is maintained with a stable curve. In comparison, the value of EUE is higher when γ is set at 0.5. It takes advantage of considering both the reward matrix and the Q-Table. The reward matrix reflects the current benefit while the Q-Table reflects benefits learned from the past experiences.

Fig. 12 compares the coverage ratios and the numbers of sleeping sensors over time, for different γ values. Fig. 12(a)–(d) depicts the results of γ values 0.3, 0.5, 0.7, and 0.9, respectively. The coverage ratio is defined as the ratio of
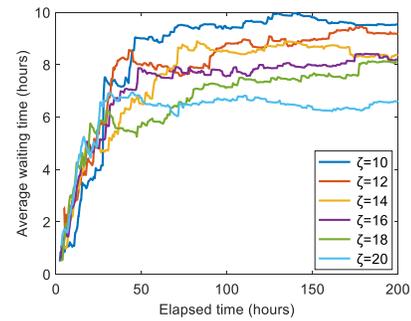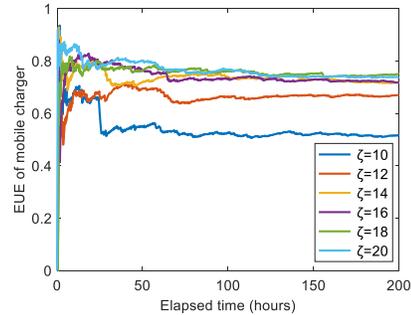
the size of the coverage area contributed by all working sensors to the size of the monitoring area. In comparison, the setting of $\gamma = 0.5$ outperforms the other settings in terms of coverage ratio and the number of sleeping sensors. More specifically, the coverage ratio with $\gamma = 0.5$ is generally higher than those with other settings while the number of sleeping sensors with $\gamma = 0.5$ is generally smaller than those with other settings. It occurs because that the mobile charger takes the advantage of both the immediate attraction reward and the accumulated experiences. Based on the results of experimental studies as shown in Figs. 10–12, the parameter γ will be set at 0.5 in the following experiments.

Since the size of the monitoring area is fixed, the value of the grid length ζ determines the number of the grids as well as the number of states. The number of states highly impacts the traveling path of the mobile charger and the refinement of accumulated experiences, which further impacts the waiting time, energy usage efficiency as well as network coverage. Figs. 13–15 aim to find an appropriate setting value of ζ for the later experimental studies.

Fig. 13 compares the trend of the average waiting time for different grid lengths. The number of deployed sensors is set at 250. The charging radius of the mobile charger is set to the value as long as the grid length, in order to guarantee the satisfaction of the proposed *Grid Constraint* as shown in (6). Similar to Fig. 10, there is a common trend that the average waiting time increases first, then decreases and finally almost levels off with the elapsed time. Finally, the mobile charger converges with a stable curve due to the learning experiences maintained in Q-Table. In comparison, the values of the average waiting time with $\zeta = 20$ are lower than those with other settings after the trends level off. It occurs because that the
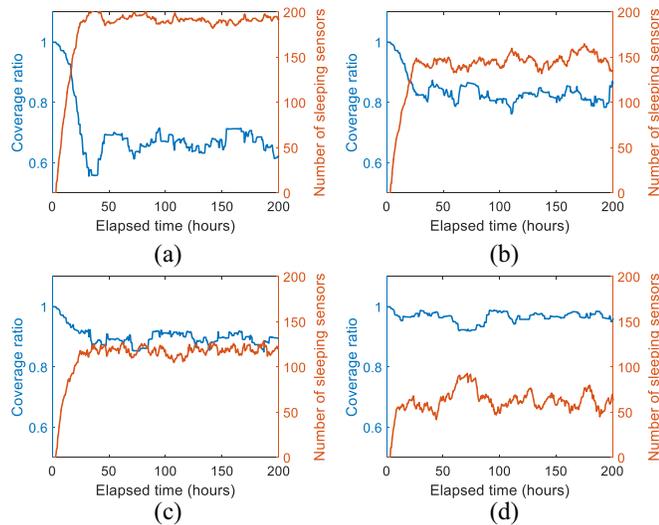
Fig. 15. Coverage ratio and the number of sleeping sensors under different grid lengths. (a) $\zeta = 10$. (b) $\zeta = 14$. (c) $\zeta = 16$. (d) $\zeta = 20$.

Fig. 16. EUE under a different number of deployed sensors.

Fig. 17. Comparisons of the four mechanisms in terms of the average waiting time.

larger the value of the grid length, the larger the number of sensors in a grid. This indicates that the mobile charger can charge a larger number of sensors simultaneously, reducing the average waiting time.

Fig. 14 compares the trends of the EUE for different grid lengths. The parameter settings are identical to those of Fig. 13. As shown in Fig. 14, the settings of $\zeta = 20$ and $\zeta = 18$ outperform the other settings in terms of the EUE. It occurs because that the mobile charger can simultaneously charge a larger number of sensors when the grid length has a big value. It further indicates that the gains of sensor energy recharged from the mobile charger are increased, when ζ is set at a large value. This leads to the increase of the EUE.

Fig. 15 shows the impact of the grid length ζ on the coverage ratio and the number of sleeping sensors under the same parameter settings of Figs. 13 and 14. In general, the setting of $\zeta = 20$ outperforms the other settings in terms of the coverage ratio and the number of sleeping sensors. It occurs because of several reasons. One major reason is the increased number of recharge sensors. The other reason is that the mobile charger can quickly learn the trend of energy consumption of the sensors when the grid size is large. Since the average waiting time of all sensors is generally reduced, the number of sleeping sensors drops accordingly. To sum up, the grid length is suggested to set at a large value on the premise of the *Grid Constraint* and will be set at 20 in the following experiments.

Fig. 16 observes the effect of the sensor density on EUE. The number of deployed sensors varies from 200 to 400. In general, the higher the sensor density, the higher the EUE. It occurs because the charger can simultaneously charge more sensors in one grid with higher density. However, it can be observed that the EUE increases with the elapsed time when the number of deployed sensors is set at 350 or 400, as compared with other settings. This is because that a higher density will lead to the havey network load. Then, the energy consumption rates of the sensors increase quickly. Consequently, after learning for a period, the charger will work around a
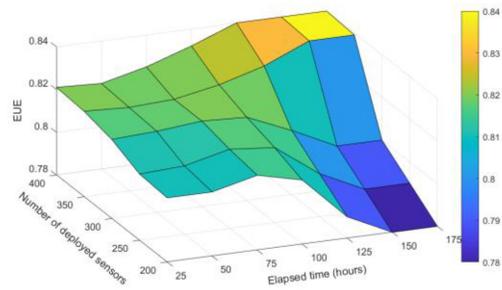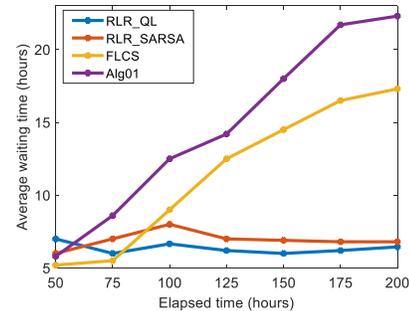
smaller region. It indicates that the charger can charge more sensors with less movement.

Fig. 17 further compares the proposed RLR_QL and RLR_SARSA against the existing Alg01 [13] and FLCS [21] in terms of the waiting time. As shown in Fig. 17, both the average waiting times of proposed RLR-*QL* and RLR-SARSA increase slightly and then decrease before leveling off. However, the average waiting times of FLCS and Alg01 always increase with the elapsed time. At the beginning of the simulation, the proposed RLR-QL and RLR-SARSA need to learn the experience of the energy dissipation trend, leading to the increase of the average waiting time. After learning for a while, the mobile charger has accumulated enough experiences, which likely guides it moving along the direction where some sensors have already sent the recharging requests and most of them are with large independent coverage, as compared with the requested sensors located in the other direction. The mobile charger learns experiences from the *Q*-Table and, thus, achieves a shorter average waiting time, as compared with the existing FLCS and Alg01. In addition, the RLR-QL has a faster downward trend than RLR-SARSA. It occurs because that the SARSA is more conservative than *Q*-Learning, leading to a lower rate of convergence. In comparison, the existing FLCS and A lg01 determine the charging schedule only according to the recently received charging requests, without learning from experiences. As a result, the number of sensors needed to be charged is increased with the elapsed time, and hence the average waiting time is increased.

Fig. 18 further compares the four mechanisms in terms of the EUE of the mobile charger. As shown in Fig. 18, there is a common trend that the EUE of the mobile charger decreases
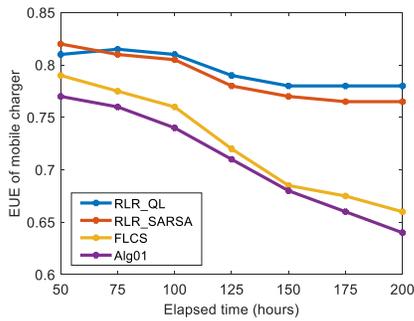
Fig. 18. Comparisons of the four mechanisms in terms of EUE.

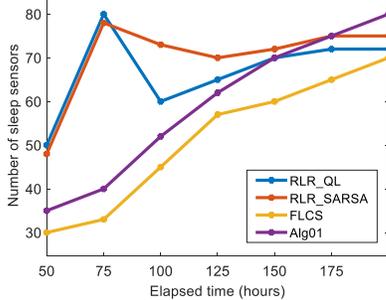Fig. 20. Comparisons of the four mechanisms in terms of ACR.

Fig. 19. Comparisons of the four mechanisms in terms of the number of sleeping sensors.

along with the elapsed time. It occurs because that the average path length for recharging the sensors is increased with the elapsed time, resulting in the decrease of the EUE. In comparison, the proposed RLR-QL and RLR-SARSA outperform the existing FLCM and Alg01. This occurs because that the proposed mechanisms learn the energy dissipation trend of the network. This indicates that the mobile charger learns to move to the areas where a lot of sensors are needed to be charged. It further indicates that the mobile charger can charge more sensors by traveling a shorter distance, resulting in a slow descent and then a stable tendency for the EUE.

Fig. 19 compares the four mechanisms in terms of the number of sleeping sensors. The numbers of sleeping sensors of the proposed RLR-*QL* and RLR-SARSA grow rapidly and then decrease before leveling off along with the elapsed time. It occurs because the proposed RLR-*QL* and RLR-SARSA still need to learn the appropriate states and the corresponding actions, without a rich experience, at the beginning of the simulation, leading to a large number of sleeping sensors. Then, the learning of the mobile charger converges and, thus, the number of sleeping sensors levels off. In general, the numbers of sleeping sensors of the existing FLCS and Alg01 increase along with the elapsed time. In comparison, the numbers of sleeping sensors of the proposed RLR-*QL* and RLR-SARSA are larger than those of the existing FLCS and Alg01 during the simulation time. It occurs because the proposed mechanism further considers the independent coverage of sensors when making the decision of the charging schedule. The sensors which have small or even no independent coverage might be ignored by the mobile charger. In contrast, the existing FLCS and Alg01 will try to charge all sensors which have
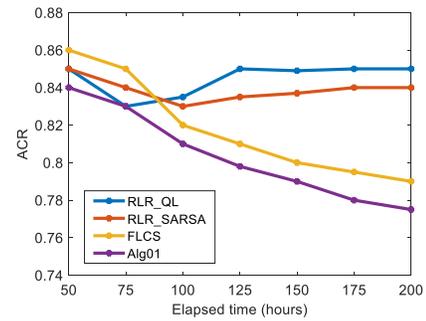
sent the charging requests. However, the proposed RLR-QL and RLR-SARSA outperform the existing works in terms of coverage, which will be depicted in Fig. 20 in advance.

Fig. 20 further compares the four mechanisms in terms of *ACR* along with the elapsed time. As shown in Fig. 20, the ACRs of the proposed RLR-QL and RLR-SARSA decrease slightly, then increase before leveling off while the ACRs of the existing FLCS and Alg01 decrease rapidly with the elapsed time. Recall the results obtained in Fig. 19. Though the proposed RLR-QL and RLR-SARSA have a larger number of sleeping sensors, they achieve higher coverage contribution, as compared with the existing FLCS and Alg01. It demonstrates that the proposed charging mechanisms achieve better performance in terms of coverage ratio during a given time period T.

VI. CONCLUSION

This article proposes a novel reinforcement learning approach for the mobile charger, named RLR, aiming to maximize the coverage contribution of the deployed sensors in a given WRSN. To achieve this, the mobile charger employs reinforcement learning to learn the energy dissipation trend of the network, the recharging cost as well as the coverage benefit. The attraction reward is designed by considering both the independent coverages of requested sensors and the distances between the mobile charger and requested sensors. As a result, the mobile charger tends to move to and recharge the requested sensors which have a short distance from the charger and can contribute large independent coverages, as compared with the other requested sensors. Extensive experiments show that the proposed RLR outperforms the compared algorithms in terms of charging waiting time of sensors, the energy usage efficiency of the mobile charger, as well as the coverage of the given sensor network. Our future work will further discuss the cooperation issue of multiple mobile chargers in a large-scale network.

REFERENCES

[1] M. T. Lazarescu and P. Poolad, "Asynchronous resilient wireless sensor network for train integrity monitoring," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3939–3954, Mar. 2021.

[2] B. Zhou, H. C. So, and S. Mumtaz, "Effect of signal propagation model calibration on localization performance limits for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3254–3268, May 2021.

[3] J. Feng, H. Chen, X. Deng, L. T. Yang, and F. Tan, "Confident information coverage hole prediction and repairing for healthcare big data collection in large-scale hybrid wireless sensor networks," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16801–16813, Dec. 2021.

[4] F. Ademaj, M. Rzymowski, H.-P. Bernhard, K. Nyka, and L. Kulas, "Relay-aided wireless sensor network discovery algorithm for dense industrial IoT utilizing ESPAR antennas," *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16653–16665, Nov. 2021.

[5] W. Lu et al., "Energy efficiency optimization in SWIPT enabled WSNs for smart agriculture," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4335–4344, Jun. 2021.

[6] G. Han, H. Guan, J. Wu, S. Chan, L. Shu, and W. Zhang, "An uneven cluster-based mobile charging algorithm for wireless rechargeable sensor networks," *IEEE Syst. J.*, vol. 13, no. 4, pp. 3747–3758, Dec. 2019.

[7] H. Dai et al., "CHASE: Charging and scheduling scheme for stochastic event capture in wireless rechargeable sensor networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 44–59, Jan. 2020.

[8] C. Wang, J. Li, Y. Yang, and F. Ye, "Combining solar energy harvesting with wireless charging for hybrid wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 560–576, Mar. 2018.

[9] C.-M. Yu, M.-L. Ku, and L.-C. Wang, "BMRHTA: Balanced multipath routing and hybrid transmission approach for lifecycle maximization in WSNs," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 728–742, Jan. 2022.

[10] M. Zhao, J. Li, and Y. Yang, "A framework of joint mobile energy replenishment and data gathering in wireless rechargeable sensor networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 12, pp. 2689–2705, Dec. 2014.

[11] R. Zhang, J. Yu, Y. Guan, and J. Liu, "A dominating set-based sleep scheduling in energy harvesting WBANs," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11923–11934, Nov. 2021.

[12] K. Ali and D. J. Rogers, "An orientation-independent multi-input energy harvesting wireless sensor node," *IEEE Trans. Ind. Electron.*, vol. 68, no. 2, pp. 1665–1674, Feb. 2021.

[13] Y. Ma, W. Liang, and W. Xu, "Charging utility maximization in wireless rechargeable sensor networks by charging multiple sensors simultaneously," *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1591–1604, Aug. 2018.

[14] M. Tian, W. Jiao, M. Liu, and S. Ma, "A charging algorithm for the wireless rechargeable sensor network with imperfect charging channel and finite energy storage," *Sensors*, vol. 19, p. 3887, Sep. 2019.

[15] T. Liu, B. Wu, S. Zhang, J. Peng, and W. Xu, "An effective multi-node charging scheme for wireless rechargeable sensor networks," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 2026–2035.

[16] W. Ouyang et al., "Utility-aware charging scheduling for multiple mobile chargers in large-scale wireless rechargeable sensor networks," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 4, pp. 679–690, Oct.–Dec. 2021.

[17] C. Wang, J. Li, F. Ye, and Y. Yang, "NETWRAP: An NDN based real-time wireless recharging framework for wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 6, pp. 1283–1297, Jun. 2014.

[18] C. Wang, J. Li, F. Ye, and Y. Yang, "A mobile data gathering framework for wireless rechargeable sensor networks with vehicle movement costs and capacity constraints," *IEEE Trans. Comput.*, vol. 65, no. 8, pp. 2411–2427, Aug. 2016.

[19] C. Lin, D. Han, J. Deng, and G. Wu, "P^2S: A primary and passer-by scheduling algorithm for on-demand charging architecture in wireless rechargeable sensor networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8047–8058, Sep. 2017.

[20] C. Lin, J. Zhou, C. Guo, H. Song, G. Wu, and M. S. Obaidat, "TSCA: A temporal-spatial real-time charging scheduling algorithm for on-demand architecture in wireless rechargeable sensor networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 1, pp. 211–224, Jan. 2018.

[21] A. Tomar, L. Muduli, and P. K. Jana, "An efficient scheduling scheme for on-demand mobile charging in wireless rechargeable sensor networks," *Pervasive Mobile Comput.*, vol. 59, pp. 1–17, Oct. 2019.

[22] W. Ouyang, M. S. Obaidat, X. Liu, X. Long, W. Xu, and T. Liu, "Importance-different charging scheduling based on matroid theory for wireless rechargeable sensor networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3284–3294, May 2021.

Cuijuan Shang received the B.S. degree in biomedical engineering from Zhengzhou University, Zhengzhou, China, in 2008, the M.S. degree in biomedical engineering from Shandong University, Jinan, China, in 2011, and the Ph.D. degree in computer science and information engineering from Tamkang University, New Taipei, Taiwan, in 2021.

She is currently an Associate Professor with the School of Computer and Information Engineering, Chuzhou University, Chuzhou, Anhui, China. Her current research interests include wireless sensor network and healthcare.

Chih-Yung Chang (Member, IEEE) received the Ph.D. degree in computer science and information engineering from the National Central University, Taoyuan, Taiwan, in 1995.

He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan. His current research interests include Internet of Things, wireless sensor networks, artificial intelligence, and deep learning.

Dr. Chang has served as an Associate Guest Editor for several SCI-indexed journals, including the *International Journal of Ad Hoc and Ubiquitous Computing* from 2011 to 2019, the *International Journal of Distributed Sensor Networks* from 2012 to 2014, *IET Communications* in 2011, *Telecommunication Systems* in 2010, the *Journal of Information Science and Engineering* in 2008, and the *Journal of Internet Technology* from 2004 to 2008.

Wen-Hwa Liao (Member, IEEE) received the Ph.D. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 2002.

He is currently a Full Professor with the Institute of Information and Decision Sciences, National Taipei University of Business, Taipei, Taiwan. His current research interests include Internet of Things, wireless sensor networks, and artificial intelligence.

Dr. Liao was an Associate Editor for the *International Journal of Distributed Sensor Networks* and *International Journal of Vehicle Information and Communication Systems*. He was also as an Associate Guest Editor for SCI-indexed *International Journal of Ad Hoc and Ubiquitous Computing*.

Diptendu Sinha Roy received the Ph.D. degree in engineering from Birla Institute of Technology, Mesra, India, in 2010.

He is currently an Associate Professor with the Department of Computer Science and Engineering, National Institute of Technology Meghalaya, Shillong, India. His current research interests include Internet of Things, big data, machine learning, and soft and evolutionary computing.