International Journal of Innovative
Computing, Information and Control
Volume **14**, Number **6**, December **2018**

 ISSN 1349-4198
pp. **1993–2010**

COOPERATIVE REINFORCEMENT LEARNING BASED THROUGHPUT OPTIMIZATION IN ENERGY HARVESTING-WIRELESS SENSOR NETWORK

Yin Wu[1], Wenbo Liu[2] and Yanyi Liu[1]

[1]Department of Information Science and Technology
Nanjing Forestry University
No. 159, Longpan Road, Nanjing 210037, P. R. China
{ wuyin; yyliu }@njfu.edu.cn

[2]College of Automation
Nanjing University of Aeronautics and Astronautics
No. 29, Jiangjun Avenue, Nanjing 211106, P. R. China
wenboliu@nuaa.edu.cn

Received January 2018; revised May 2018

Abstract. *Energy Harvesting-Wireless Sensor Network (EH-WSN) has got increasing attention in recent years. During its actual deployment, we find that the energy that can be harvested from the environment is always continually changing and unpredictable. This paper aims to investigate the energy management approach of EH-WSN under such circumstance and propose a corresponding dynamic scheme to optimize the network throughput. Here we adopt a Cooperative Reinforcement Learning (CRL) method for analysis. Firstly, we model the external environment status, and then the CRL algorithm based on Q-learning starts regulating the EH-node's duty cycle according to the external energy's variation; meanwhile, the feedback reward takes responsibility for the evaluation of CRL's regulation. Different from traditional reinforcement learning, CRL facilitates EH-nodes to share their local knowledge with others periodically. With this information, EH-node chooses which action to take for the current time slot: (i) idling, (ii) sensing, (iii) calculating, and (iv) transmitting. Experimental results show that the proposed scheme can make EH-node work energy-balanceable, and satisfy the network throughput requirement effectively, and it also improves the energy utilization efficiency obviously in contrast with existing strategies.*

Keywords: Energy harvesting, Wireless sensor network, Energy management, Cooperative reinforcement learning, Energy neutral, Throughput

1. **Introduction.** The limited available lifetime is a key bottleneck for most battery-powered Wireless Sensor Networks (WSNs). Therefore, harvesting energy from the environment has been widely investigated to ensure the sustainability of the network. As for this Energy Harvesting-WSN (EH-WSN), many studies have been carried out [1-5]. Ongaro and Saggini propose a power management architecture that utilizes both supercapacitor and lithium battery as energy storages for a solar-powered WSN [1]. Lee et al. develop a cross-layer optimization-based scheduling scheme called binding optimization of duty cycling and networking through energy tracking (BUCKET) to maximize the utilization of solar energy [2]. In [3], the authors propose an energy prediction algorithm that uses the light intensity of fluorescent lamps in an indoor environment, and then an optimal transmission interval is calculated using the amount of predicted harvested energy and residual energy. The authors in [4] just propose a stochastic Markov

DOI: 10.24507/ijicic.14.06.1993

chain framework, which captures the degradation status of the battery to improve the lifetime of sensor while guaranteeing the minimum required Quality of Service (QoS). [5] considers the problem of communication coverage for sustainable data forwarding in EH-WSN, where an energy-aware deployment model of relay nodes is proposed. From these achievements, we can see that the main research issue lies in two aspects: how to maximize the harvested energy and how to maximize the energy utilization efficiency; plus one research target: keeping the EH-node "energy-neutral" [6]. Therefore, the energy management algorithm of EH-WSN is particularly attractive because it is just like the "brain" of whole system. This paper proposes a novel energy management strategy using the Cooperative Reinforcement Learning (CRL) method to regulate the EH-node's work/sleep duty cycle based on the incoming energy's changing status, with the purpose of maximizing the number of sampled data aggregated at the sink while keeping all the EH-nodes working under "energy-neutral" mode.

Recently, there are several works in which the authors used RL method to optimize WSN's performance. Pourpeighambar et al. considered a routing problem in the cognitive radio networks such that each cognitive user wants to select best route that minimizes its own end-to-end delay provided that the QoS requirements of the primary users are met [7], and they used a multi-agent Q-learning algorithm for solving the routing problem that can avoid information exchange between the competing cognitive users. In [8], Khan and Rinner proposed a method for scheduling the tasks using cooperative reinforcement learning where each node determines the next task based on the observed application behavior. By exchanging data among neighboring nodes, they could further improve the energy/performance tradeoff. Simulations showed that cooperative approaches are superior to non-cooperative approaches in a target tracking application. Chen et al. investigated a reinforcement learning based sleep scheduling for coverage algorithm in rechargeable time-slotted sensor networks [9]: it includes the precedence operator-based group formation algorithm and the Q learning-based active node selection algorithm. Experiments on a solar-powered wireless sensor network were presented, and the results showed that it could effectively adjust the working modes of nodes. In addition, it achieved the energy consumption balance between nodes while maintaining the desired coverage. Especially [10] introduced a cooperative reinforcement learning scheme, namely Cooperative Q, to let cognitive radios learn and adapt to the environment they are in, and share their information among themselves. Its proposal aimed to maximize energy efficiency while ensuring buffer occupancy kept below some predetermined level. However, none of the previous researches have investigated the adaptivity of their algorithms to changes in environmental energy and task allocation, much less on the optimization of precise duty cycle.

Our work intends to maximize the total number of sampled data under energy harvesting constraints. The distributed and stochastic nature of environment energy model lends itself to a learning approach. So we design an effective and reliable CRL method which considers the energy buffer, duty cycle, task schedule, and power consumption together to improve the energy utilization efficiency. To the best of our knowledge, this paper is the first to use CRL in network performance optimization of EH-WSN. The main jobs and innovations are as follows.

(1) A novel CRL formulation that can jointly optimize the duty cycle, the energy balance and the throughput has been introduced. The optimized result notably improves the information quality.

(2) The EH-WSN equipped with our novel algorithm can adapt to the energy changes in environment effectively and sensitively.

(3) We evaluate the performance of our CRL algorithm with a different non-cooperative method. Results show that CRL outperforms non-cooperative way in terms of collected data quantity and operation stability.

The rest of this paper is organized as follows. Section 2 explains our system model and describes the problem formulation. Section 3 presents the cooperative RL approach and our CRL based online duty-cycle regulation algorithm. Section 4 discusses simulation results for a solar powered EH-WSN application. In Section 5, we conclude this paper with a brief summary.

2. **System Model and Problem Statement.** We consider an EH-WSN that is formed with a set of cooperative EH-nodes which seek for data transmissions to the sink. Each EH-node mainly contains three parts: energy harvesting module, energy storage battery and wireless sensor node. CRL algorithm takes responsibility for the system parameters monitoring and optimal energy management. From [11] we know that the energy cost of EH-node is proportional to its working duty cycle; therefore, CRL should calculate the optimal duty ratio to make EH-node work in "energy-neutral" strategy. A structural diagram of EH-node is shown below in Figure 1.

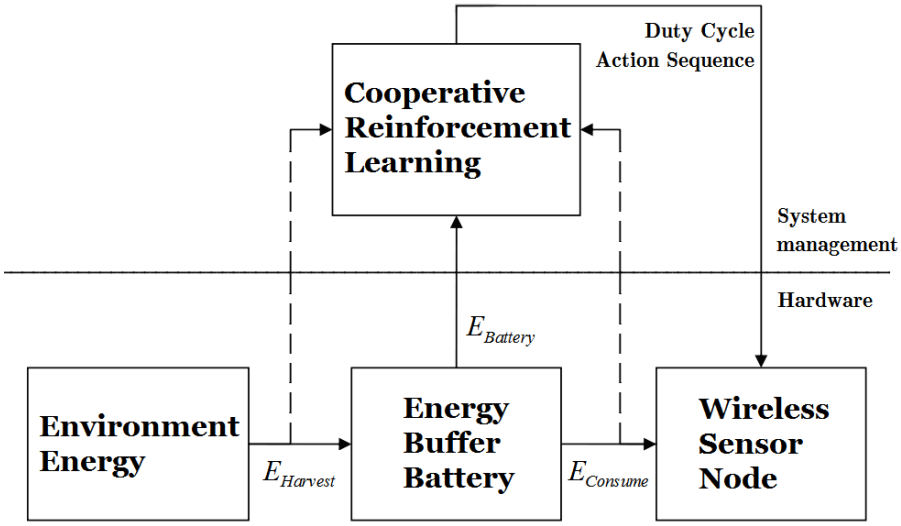


FIGURE 1. Electrical diagram of an EH-node with CRL

At the beginning of each time slot, every EH-node should start to harvest energy based on its battery buffer level (here it is assumed to have two kinds: insufficient and adequate). If the buffer shows insufficient, the node should turn down to low power state and harvest energy until the energy level changes to adequate; otherwise, the node would take jobs of three contents: data collection, data processing and data transmission. Note that these three operations also consume different energy costs. Hence, the node would run a task chosen from a certain combination of three actions under "energy-neutral" constraints, as shown in Figure 2.

The objective of each EH-node is to transmit the packets with maximum volume while not causing energy exhaustion. Therefore, when the environment energy available for harvesting has changed, the EH-node should automatically regulate its work/sleep duty cycle and decide on an action sequence to execute for this time round.

Meanwhile, the whole network works in a common operation status based on a pre-setting routing protocol, i.e., the relay node needs to receive former node's packets and transmit it with its own data to the next, and repeat the process until the sink. Hence,

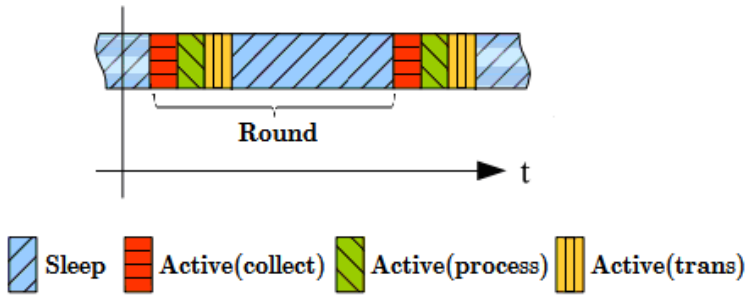


FIGURE 2. A typical duty cycle sequence (Assume that energy harvesting is always running)

every EH-node still needs to cooperate with its front and back nodes to choose the optimal next action.

2.1. **Energy buffer model.** We can deduce the energy model of each EH-node:

$$E_i(\tau) = E_i(\tau - 1) + E_{EH,i}(\tau) - I[a_i(j,k)] \cdot E_{Trans} - b_i(m) \cdot E_{Process} - c_i(n) \cdot E_{Collect} \quad (1)$$

$E_i(\tau)$ is the residual energy of node i at the end of time round τ, $E_{EH,i}(\tau)$ is the harvested energy of node i during round τ; $I[\cdot]$ is a binary indicator function and $a_i(j,k)$ is the event that node i receives data from node j and transmits packets to node k, $b_i(m)$ and $c_i(n)$ are the events that node i executes data process or data collection m and n times, respectively; E_{Trans}, $E_{Process}$ and $E_{Collect}$ just represent the energy consumption of data transmission, data processing and data collection.

2.2. **Data traffic model.** We can deduce the energy model of each EH-node. As an important part of energy management, data transmission cost model plays a crucial role. In the paper we refer to a simple transmission consumption model in [12], and to power an l bit of messages over distance d, the consumed energy is:

$$E_{Send}(l,d) = E_{elec} \cdot l + \xi_{amp} \cdot l \cdot d^2 \quad (2)$$

Moreover, to receive this message, the consumed energy is:

$$E_{Receive}(l) = E_{elec} \cdot l \quad (3)$$

where E_{elec} is the radio dissipation, and ξ_{amp} is the emission amplification factor.

2.3. **Actions and outcomes.** Each EH-node should decide to stay sleep or collect data or do something, etc., and acts based on the outcome of its choice. The cost (both time and energy) of switching from one operation mode to another is ignored in order for a convenient calculation: for example, CC2530 chip opens a radio connection during a setup phase (10.5 mS) and closes it during a teardown phase (2.8 mS), compared with the time length of one slot (1 S) in our experiment the mode switching cost is relatively negligible. Following we would present all possibilities resulting in various throughput and energy consumption.

(i) **Stay Sleep.** Independent of the data collecting state, the EH-node decides to stay sleep for this time slot due to internal factors such as out of power or external factors such as next hop node disabled. The energy consumption in this state is assumed to be zero, also as no packets are transmitted, throughput $\Phi = 0$ in this case.

(ii) **Turn on Collection.** When the battery shows adequate, EH-node should start to work. Data collection is a fundamental step to accomplish the design target of EH-WSN.

The energy consumption in this case can be formulated as:

$$E = c(n) \cdot E_{Collect} \tag{4}$$

Here $c(n)$ means that the node chooses to collect data in n continuous slots during one time round. Consequently, we think the node produces n byte data; however, these outputs cannot be transferred directly, which still need data processing in advance.

(iii) **Turn on Processing.** When the data collection action completes, it needs to conduct data processing operation. As for the EH-node nowadays always has a relatively strong computational capability, we regard that only one slot data processing is enough for the above continuous sampled data, and it yields n byte output. Note that data collection could be discontinuous in a time round due to many reasons, but it must be closely followed by a one-slot data processing procedure. The energy consumption of this case is:

$$E = b(m) \cdot E_{Process} \tag{5}$$

Here m means that data collections turn on m times intermittently in one time round, so the throughput $\Phi = m \times n$ bytes afterwards.

(iv) **Turn on Transmission.** EH-node wants to transfer its data to the sink based on a usable routing path. We think every node could send data freely to its forward node if there is enough energy; on contrast it could also receive the former node's message except in sleep state. In light of using the Direct Memory Access (DMA) technique, we can use just one slot to represent the whole communication procedure, and we consider that one slot is long enough to receive and send all the available data. The resulting energy consumption of this state is:

$$E = E_{Send}(l + \phi, d) + E_{Receive}(l) \tag{6}$$

Consequently the throughput $\Phi = l + \phi$ bits (ϕ is the node's own data, while l is the received data).

(v) **Continuous Collection, Processing and Transmission.** In this case, EH-node chooses to act according to a sequential order based on a sufficient battery level. For example: the node executes an n continuous slots data collection, a one-slot data processing and a one-slot data transmission. Therefore, the whole energy consumption is:

$$E = n \cdot E_{Collect} + E_{Process} + E_{Send}(l + 8n, d) + E_{Receive}(l) \tag{7}$$

and the throughput $\Phi = l + 8n$ bits.

(vi) **Intermittent Collection, Processing and Transmission.** If the energy stored in battery cannot afford a continuous activities employment, then the node could work in a discontinuous mode: it may first enter in collection state of $n1$ slots, followed by one slot processing; after that it turns down to sleep unless the energy buffer shows to be adequate; once it wakes up to collect data again, the above process might repeat until near the end of a round; and when coming to this last moment, the node should prepare to transmit all the data to its next hop node. Hence, the energy utilization during this case is:

$$\begin{aligned} E = {} & (n1 + n2 + \cdots + nm) \cdot E_{Collect} + m \cdot E_{Process} \\ & + E_{Send}[l + 8 \cdot (n1 + n2 + \cdots + nm), d] + E_{Receive}(l) \end{aligned} \tag{8}$$

and the whole throughput $\Phi = l + 8 \cdot (n1 + n2 + \cdots + nm)$ bits.

(vii) **Interrupted Transmission.** Once the EH-node completes its data collection and processing, it needs to transmit all the valid data to next hop node. However, if the next node is in extremely low battery condition, and cannot wake up to receive messages, this EH-node only has to drop the data and turn down to idle. Thus, we regard the case as a packet loss event that causes severe damage to the network performance. To avoid

the problem CRL must be adopted to coordinate the neighbor nodes' status. Energy cost during this case is considered to be:

$$E = E_{EH,i}(\tau) \tag{9}$$

Obviously the throughput equals zero.

2.4. **Problem formulation.** Let $b(T)$ denote the number of bits received by sink during time rounds $1 \sim T$, $\Phi_i(\tau)$ represents the EH-node i's corresponding throughput in round τ, X is the total number of time slots in one time round, and Z is the total number of EH-nodes in EH-WSN. Then we can formulate the throughput maximization problem as follows:

$$\textbf{Maximize} \quad b(T) = \sum_{i=1}^{Z} \sum_{\tau=1}^{T} \Phi_i(\tau) \tag{10}$$

$$\textbf{s.t.} \quad B \geq E_i(\tau) \geq 0,\ \forall i, \tau,\ 1 \leq i \leq Z,\ 1 \leq \tau \leq T \tag{11}$$

where B is the battery's maximum capacity. An EH-node can choose only one action per time slot from actions (i)-(iv), or carry on one working sequence from (v)-(vii) per time round. While EH-WSN tries to maximize its total throughput, as described by (10), it also needs to maintain its "energy-neutral" condition to satisfy the battery constraint as shown in restriction (11).

Instead of solving this problem, which requires knowledge of vast system parameters, we propose a learning based throughput computing scheme that is an online algorithm and approximates the above-defined solution.

3. **Duty Cycle Regulation Algorithm with Cooperative Reinforcement Learning.** Here in this section, we introduce cooperative Q-learning based algorithm dubbed as DR-CRL which makes an EH-node (referred to as agent) learn while taking actions and making observations in its environment. We first define the states of our system as well as actions and the corresponding rewards.

3.1. **States.** We represent the state of an EH-node as a tuple $s = [E_i(\tau), \Phi_i(\tau)]$. For simplicity, we quantize the energy buffer occupancy to β levels denoted by $\beta = \{0, 1, 2, \ldots, B\}$. Given $\psi = \{0, 1, 2, \ldots, (X-2) \cdot 8\}$ is the set of data throughput, then the state space of our system is $\beta \times \psi$ which consists of $(B+1) \cdot 8 \cdot (X-2)$ states.

3.2. **Actions.** An EH-node can choose one of four actions: stay sleep, turn on collection, turn on processing, and turn on transmission in every slot during one round, and the three practical sets of action sequences A in a whole time round: Continuous Collection, Processing and Transmission; Intermittent Collection, Processing and Transmission; and Interrupted Transmission. Therefore, there are possibly $X + (X^2/2)$ actions an EH-node can take.

3.3. **Rewards.** After observing the outcomes of its actions, each EH-node gets a reward. The reward function $r_x(s, a) : S \times A \rightarrow R$ defines the desirability of an action a performed on a state s. Reward function takes different values for each possible outcome defined in Section 2.3. We calculate $r_x(s, a)$ as follows.

(i) **Stay Sleep.** In this case, we think the EH-node is out of energy mainly due to an extremely low energy-harvesting rate. So the node has no alternative but to sleep until the energy buffer meets requirement. In addition, we have nothing to adjust unless the node wakes up at the present circumstances.

(ii) **Continuous Collection, Processing and Transmission.** In this case, reward function may get a positive value that depends on the residual battery level, energy harvesting rate and action sequence. It is defined as:

$$r_x(s,a) = \left[1 - 2\frac{E_i(\tau-1)}{B}\right] \cdot \left[1 - \frac{E}{E_{EH,i}(\tau)}\right] \cdot \frac{\Phi_i(\tau)}{X} \tag{12}$$

Here E, $\Phi_i(\tau)$ are the energy consumption and the corresponding throughput in current time round which is defined in (7), $E_{EH,i}(\tau)$ is the energy harvested in the present time round, and $E_i(\tau-1)$ is the energy state of node i according to last time round $\tau-1$ as shown in (1). This reward function aims to enhance the working duty ratio when energy harvesting rate is relatively high, and vice versa.

(iii) **Intermittent Collection, Processing and Transmission.** Similarly, E, $\Phi_i(\tau)$, m are the parameters defined in (8). Due to its inherent low energy buffer level, EH-node should try hard to improve the residual battery while maximizing the data throughput.

$$r_x(s,a) = \left[1 - \frac{E}{E_{EH,i}(\tau)}\right] \cdot \frac{m}{X} \cdot \Phi_i(\tau) \tag{13}$$

(iv) **Interrupted Transmission.** In this case, the EH-node should get negative reinforcement to evade further interruptions. $\alpha_{Interrupt}$ is the penalty coefficient for interruption and is problem and algorithm specific, ν is the bit rate that EH-node would transmit, and T_{slot} is duration of a time slot.

$$r_x(s,a) = -\alpha_{Interrupt} \cdot \nu \cdot \frac{X \cdot T_{slot}}{E_{EH,i}(\tau)} \tag{14}$$

3.4. **Algorithm.** After finishing the aforementioned work, we can study the DR-CRL algorithm. Generally speaking, reinforcement learning is a heuristic unsupervised learning method which tries to search the appropriate policies from interaction with the environment; Q-learning is frequently used to calculate the accumulative rewards and decide the best policy [13-17]. Based on this above basis, thus we just propose a novel cooperative reinforcement learning algorithm that uses neighbor nodes' interactions: every EH-node must check out the battery level of its next hop node after every round. If the battery level of next hop is adequate, then the current node should choose action sequence 5 or 6 based on its local battery level, and it will adjust the parameters n or $n1, n2, \ldots, nm$ based on the reward functions to maximize the whole throughput. Otherwise, if the next hop node is in an insufficient state (certainly, it should work in the intermittent mode in the first place), the current EH-node should only work in the intermittent collection, processing and transmission mode also. The most important task in this case is to keep the next hop node alive and improve its energy storage level until reaching the adequate status. The detailed regulation method deals with parameters $n1, n2, \ldots, nm$ based on RL as well. Finally, the special case is that the next hop node has no residual energy. We consider it is caused by inappropriate management or harsh external condition so the current node and all the nodes previous along the routing table lost their data during these time rounds unless it wakes up. Therefore, the current node also has nothing to do except being idle, but note that it has to wake up to regulate the battery storage level to prevent overflow. An energy states analytical diagram is shown in Figure 3: the left side icons represent the energy status of transmitting node, and the right side means the receiving node's energy state. Actually if the right side's energy shows to be sufficient, then the left current node has two options, i.e., continuous mode or intermittent mode, based on its energy condition; else if the right side turns to be insufficient, then the right

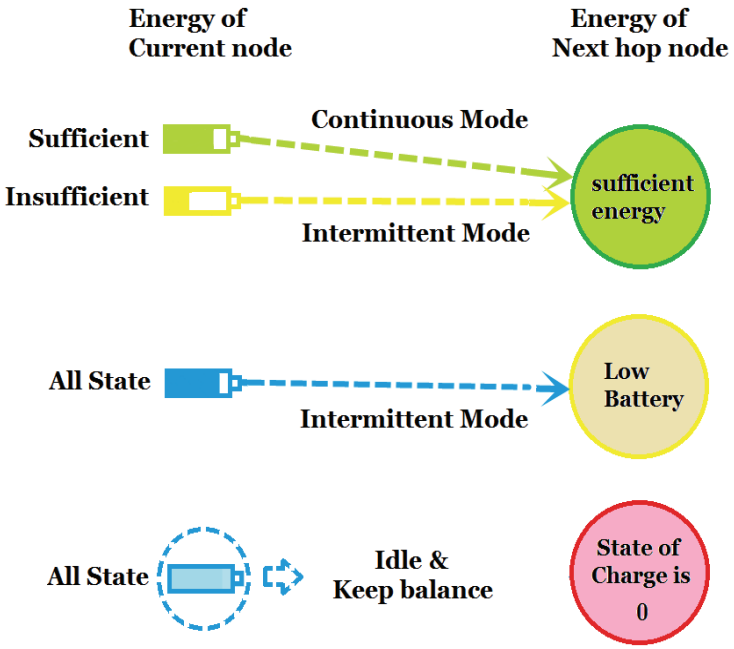


FIGURE 3. Energy state diagram of an EH-node placed in the network

Algorithm 1. **D**uty cycle **R**egulation algorithm based on **C**ooperative **R**einforcement **L**earning

Initialize $Q(s,a)=0$ and $e(s,a)=0$
While energy harvesting rate is not equal to zero **do**
  Determine current state s by application variable
  Select an action a, using Exploration-Exploitation Policy ($a_t = \arg\max_{a\in A} Q_t(s,a)$)
  Execute the selected action a
  Calculate reward for the executed action (Equations (12), (13), (14))
  Update the learning rate α (Equation (15))
  Calculate the temporal difference error $\delta_t = r_{t+1} + \gamma \cdot f^i \cdot Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$
  Update the eligibility traces $e_t(s,a)$:

$$\begin{cases} e_t(s,a) = \gamma \cdot \lambda \cdot e_{t-1}(s,a) + 1 & \text{if } s = s_t \text{ and } a = a_t \\ e_t(s,a) = \gamma \cdot \lambda \cdot e_{t-1}(s,a) & \text{otherwise} \end{cases}$$

  Update the Q value: $Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha \cdot \delta_t \cdot e_t(s,a)$
End While

receiving node should work under intermittent mode to save energy above all; hence the left node only has to work in intermittent mode also to cooperate with it.

As in the above described Algorithm 1, α is the learning rate which controls how much a learning step will impact the Q-value, and it is calculated as:

$$\alpha = \frac{\zeta}{visited(s,a)} \tag{15}$$

where ζ is a positive constant, and $visited(s,a)$ represents the visited state-action pairs so far [18]. It is proven that the above Q-learning will converge to the optimal policy that maximizes rewards, i.e., the optimal action sequences that maximize the throughput.

With regard to the detailed information in the learning procedure of Algorithm 1, we take the circulation optimization in intermittent mode for example.

Algorithm 2. Learning loop of intermittent collection, processing and transmission mode

Learning:
Loop
    Observe current state $s(n1, n2, \ldots, nm)$ on time round $\tau - 1$
    Generate a uniform random number $R \in (0, 1)$
    If $R < \varepsilon$ **then**
        Select action $a_t \in A$ randomly
    Else
        Select action $a_t = \arg\max_{a \in A} Q_t(s, a)$
    If $a_t = [n1' = (n1 + 1), n2, \ldots, nm]$ **then**
        Switch from duty cycle $n1, n2, \ldots, nm$ to $n1', n2, \ldots, nm$
        Sense the energy consumption and data throughput in time round τ
        Get reward $r_t(s_t, a_t)$ according to Equation (13)
        Update the Q value
    Else if $a_t = [n1'' = (n1 + 2), n2, \ldots, nm]$ **then**
        $\cdots\cdots$

In Algorithm 2, ε is the exploration ratio and notice that the numerical value of working slots is $n1 + n2 + \cdots + nm + m + 1$, which must be less than X.

At each time step, the above Algorithms 1 and 2 use the exploration-exploitation mechanism to learn the states iteratively, and choose the action with the highest reward based on the Q-value table.

Finally, if the interrupted mode appears, the current EH-node should keep idle in the whole time round and monitor the battery status. In case of a high-energy harvesting rate, the node must take some energy-extensive consumption tasks to prevent energy overflow. For it is not the main concern of this paper, we just use a general duty cycle ratio for regulation which is based on Equation (14).

4. Experimental Results and Evaluation.

We present the empirical studies in this section to evaluate the performance of our proposed algorithm DR-CRL. Simulations are carried out on MATLAB platform. We compare the performance of our algorithm against existing algorithms from five aspects respectively that are EH-node's residual energy, the work/sleep duty cycle ratio, the action sequence versus the energy-harvesting rate, the individual EH-node's throughput and whole EH-WSN's throughput.

4.1. Simulation setup.

The parameters used in the simulations are shown in Table 1.

For the communication parameters, we use the value obtained from [9]. We consider that each round lasts for 20 minutes. The demarcation point of battery shortage is set to 40% to strengthen its robustness. For the Q-learning parameters, as with most RL problems, these values were determined empirically rather than through mathematical methods. We evaluated the system with different values of parameters and chose the combination that performed the best. The system is somewhat sensitive to the values of α and ε. Using high values for α (learning rate) and ε causes large oscillations in Q-values during training. Therefore, we chose smaller values but compensated with a larger number of iterations during learning.

We design an EH-WSN with the topology shown as Figure 4, where nodes have been half-arbitrary placed around the sink. Each node's transmission distance is listed as follows: nodes 1, 2, 3, 4 are linked directed with the sink and $d1 = d2 = d3 = d4 = 10$ m; node 5 and node 6 also transmit to sink straightly and $d5 = 10\sqrt{2}$ m, $d6 = 5\sqrt{2}$ m

TABLE 1. Simulation parameters

Parameter	Values
Number of time rounds (T)	10
Number of time slots in one time round (X)	20
Time slot duration (T_{slot})	1 s
Number of EH-nodes (Z)	10
Radio dissipation (E_{elec})	50 nJ/bit
Emission amplification factor (ξ_{amp})	100 pJ/bit/m^2
Energy consumption of data collection ($E_{Collect}$)	120 mW$\cdot T_{slot}$
Energy consumption of data processing ($E_{Process}$)	80 mW$\cdot T_{slot}$
Battery energy level of insufficient	40%
Battery's maximum capacity (B)	7000 mAH
Bit rate that agents transmit (ν)	3.5 Mbps
Penalty coefficient for interruption ($\alpha_{Interrupt}$)	1.2
Q-Learning Parameters	
Exploration probability ε	0.03
Discount rate γ	0.9
Learning parameter α, λ	0.5

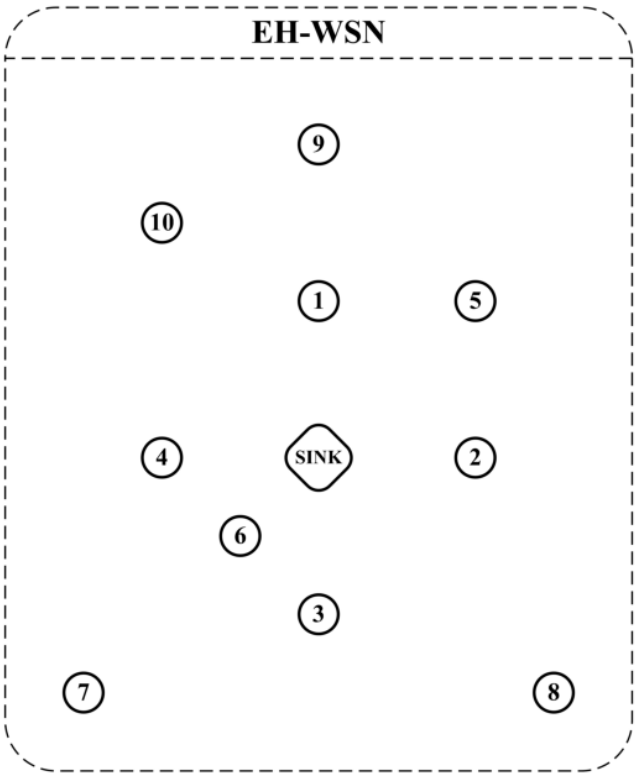


FIGURE 4. Topology of the proposed EH-WSN

respectively; however, node 7 needs to transfer data to node 6 first and then be relayed to the sink, so $d7 = 10\sqrt{2}$ m; similarly node 3 is responsible for relaying node 8's data and $d8 = 5\sqrt{10}$ m; as well node 9 uses node 1 as the relay station and $d9 = 10$ m; at last the node 1 is also in charge of relaying node 10's data additionally, and $d10 = 5\sqrt{5}$ m.

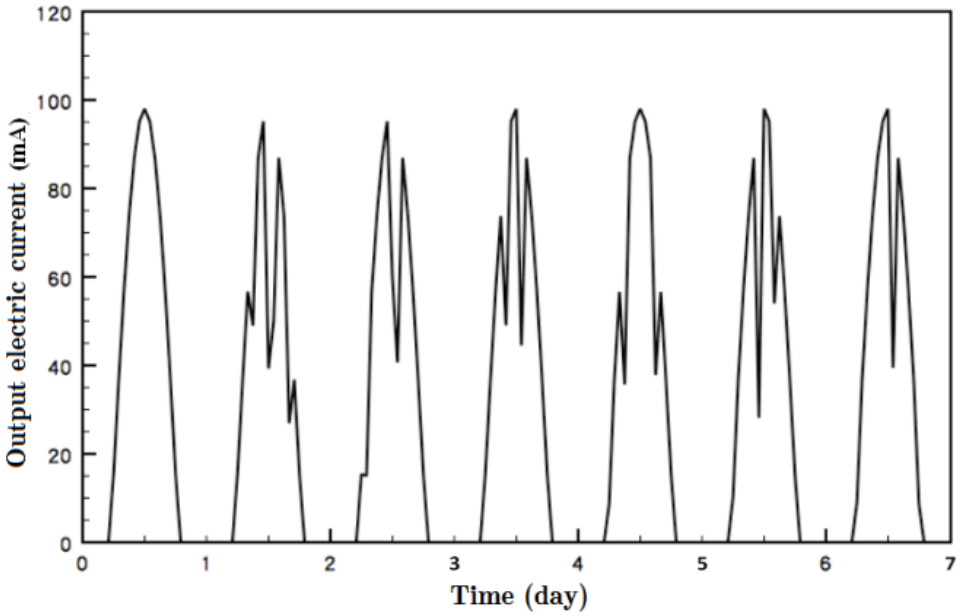


FIGURE 5. Operation diagram of selected energy harvesting module

Moreover, we adopt an energy-harvesting module (solar panel) with typical electrical values from [19], and its output energy during one week is recorded in Figure 5. Because in the database the collection of solar radiation occurs at the rate of about a quarter of an hour, when the present round is within a certain quarter, the solar radiation is viewed as a constant in this round.

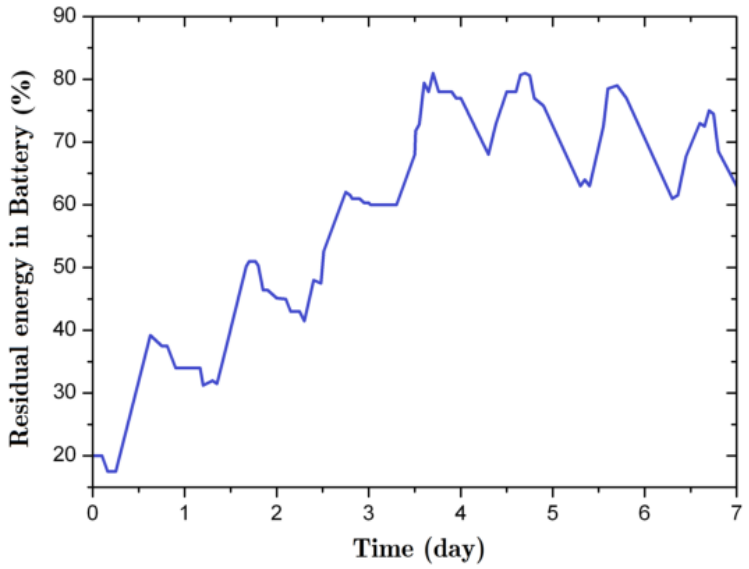
So based on the above basis, we could have the following experiments aiming at testing and evaluating the individual node's energy sustainability, task scheduling capability and the overall network throughput performance. We choose to neglect the energy and time cost caused by the proposed DR-CRL algorithm for the convenience of calculation.

4.2. **Energy-neutral testing.** First of all, we do some research on the performance of EH-nodes' energy variation trend: we assume that all nodes endure the same energy harvesting condition of Figure 5 in this exam, and the DR-CRL mechanism has been loaded on all, then we set the nodes with different initial energies in the battery, and introduce a statistic analysis on the duty cycle ratio (for a direct perception, we classify the ratio into 5 sorts: 0%, 25%, 50%, 75% and 100%, P.s. data collection, processing and transmission are all seemed as working state), and the record of some node's residual energy variation and duty cycle changes are shown in Figure 6.

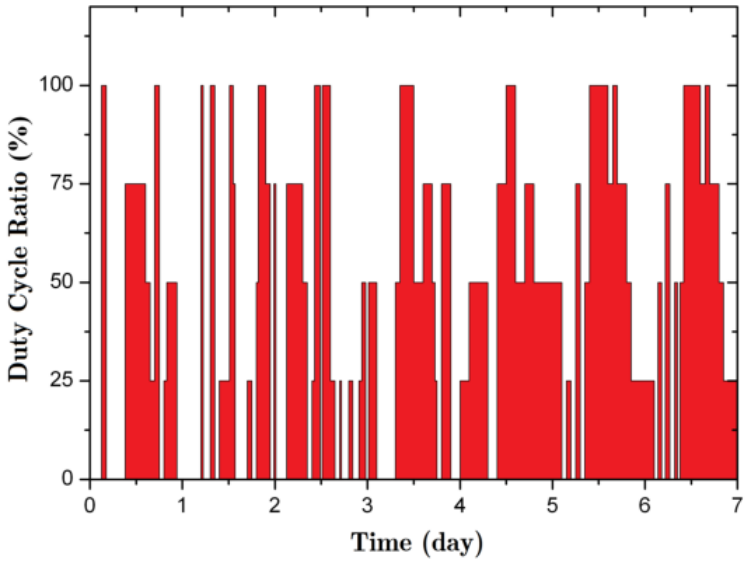
In Figure 6(a), we can see that the residual energy of node 2 rises to approximately 75% from day 1 to day 4, and maintains around 65%-70% in days 5, 6, 7. Figure 6(b) shows the detailed work/sleep ratio variations from day 1 to day 7, apparently the ratio rises to 100% when coming up to the noon hour and drop to 0% when is in the dark night, and the average ratio is lower than 55% in the first four days in order to increase the residual energy, and improves to near 60% in the last three days to maximize the throughput. From these data, we can infer that the proposed energy management algorithm can effectively charge the EH-node that is given a lower initial energy, and make it work reliably.

Next, we check the node 6's performance that has been set with a medium initial energy, and the result is shown in Figure 7.

We can see that the residual energy rises up quickly in the first day, and stays between 52%-66% during day 2 to day 6, and then gradually increases to 65% in day 7. While



(a)



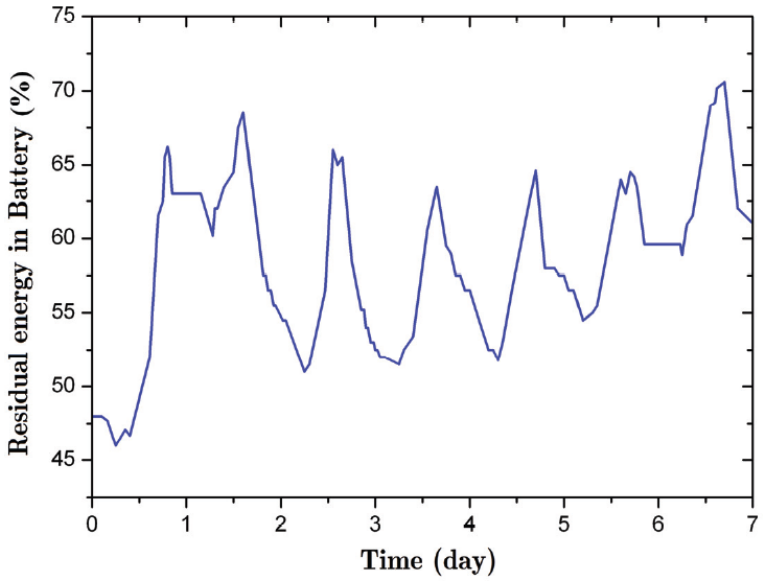
(b)

FIGURE 6. (a) Changes of residual energy (node 2 with 20% initial electricity) and (b) changes of duty cycle ratio (node 2 with 20% initial electricity)

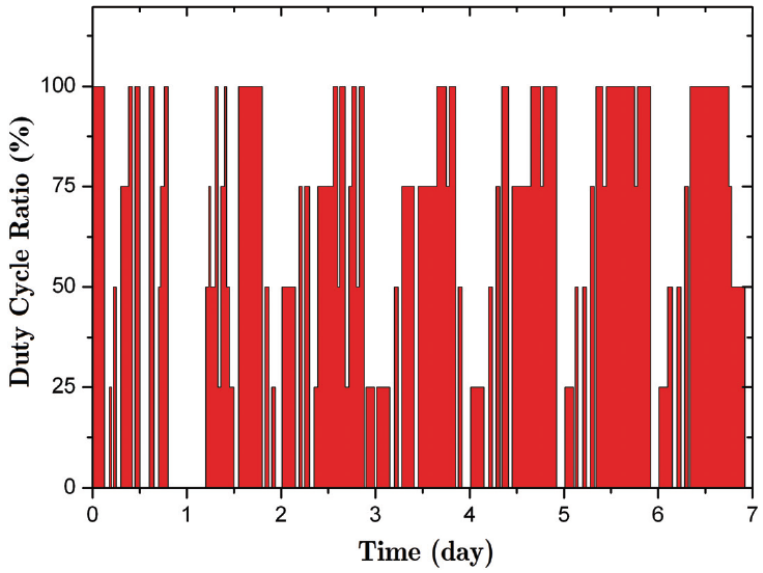
the average work/sleep ratio in day 1 is lower than 40%, keeps up between 50%-60% at day 2 to day 5, and finally achieves stabilization in days 6 and 7. This result proves that DR-CRL could maintain the energy-neutral status of EH-node with a medium level starting energy.

Thirdly, we measure the node 1's energy profile and its duty cycle, as shown in Figure 8.

Node 1 is equipped with a higher initial energy as for it has to relay two front nodes' data, from Figure 8(a) we discover that the battery level rises to 90% rapidly after day 1, and then it declines slowly to about 60% after day 3. The average duty-cycle ratio is higher than 65% in the first three days with more energy consumption that aims at



(a)



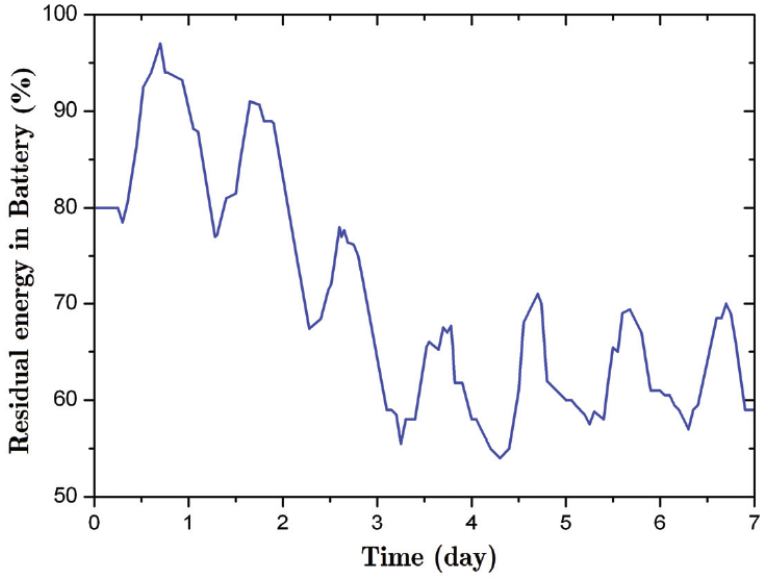
(b)

FIGURE 7. (a) Changes of residual energy (node 6 with 48% initial electricity) and (b) changes of duty cycle ratio (node 6 with 48% initial electricity)

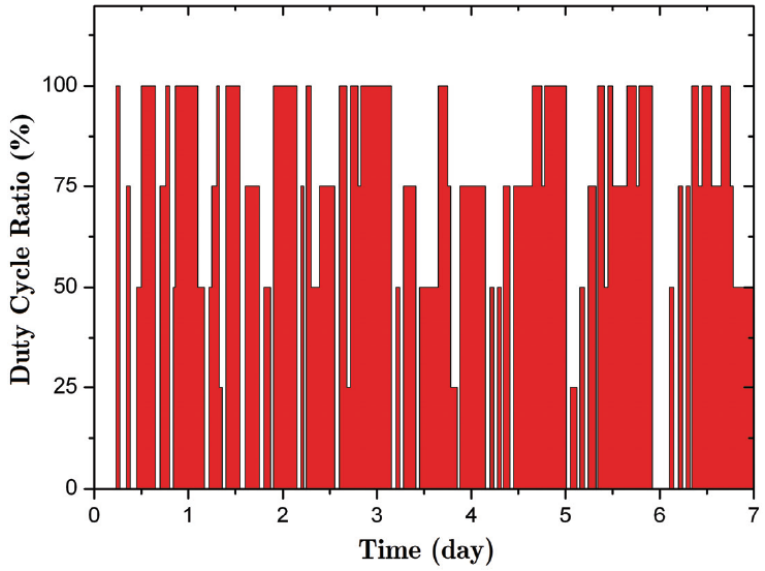
reducing the residual energy, and then it stays around 50%-60% in the last four days to keep energy balance.

From what has been mentioned above, we may find that the proposed DR-CRL algorithm can make EH-nodes with different initial energies work in the "energy-neutral" state effectively and efficiently, which satisfies the principal rule of EH-WSN.

4.3. **Action sequence vs. energy harvesting rate.** In this part, we study the action sequences of EH-nodes when the energy harvesting rate changes. In theory, the EH-nodes would automatically regulate their working time ratio based on the current harvested energy under DR-CRL mechanism. For individual node the action carried out in every slot needs to follow two principles: one is the "energy-neutral" rule of itself, and the other



(a)

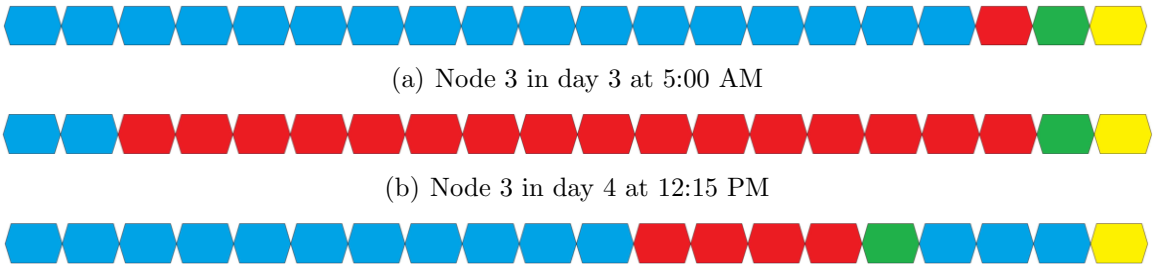


(b)

FIGURE 8. (a) Changes of residual energy (node 1 with 80% initial electricity) and (b) changes of duty cycle ratio (node 1 with 80% initial electricity)

is the cooperation with the rest nodes on the routing table. Here we choose node 3 as an illustration, some operation sequences of certain moment are shown in Figure 9, action time order from left to right.

As can be seen from the diagram, node 3 first needs to charge the battery to an adequate level, so it has to sleep most of the time; next when approaching the noon hour, it has gathered enough energy to collect, process and transmit data as much as possible in a majority of time; then if coming to the nightfall, node has to turn down to sleep discontinuously because of the insufficient residual energy and declined harvesting rate.

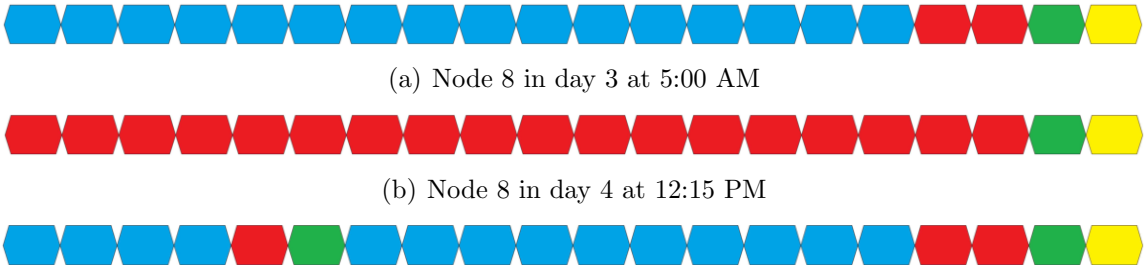


(a) Node 3 in day 3 at 5:00 AM

(b) Node 3 in day 4 at 12:15 PM

(c) Node 3 in day 5 at 6:30 PM

FIGURE 9. (color online) Typical action sequences of node 3 with 20% initial capacity (Here blue slot means sleep, red means data collection, and green means data process, yellow means data transmission)



(a) Node 8 in day 3 at 5:00 AM

(b) Node 8 in day 4 at 12:15 PM

(c) Node 8 in day 5 at 6:30 PM

FIGURE 10. (color online) Typical action sequences of node 8 with 50% initial capacity (Here blue slot means sleep, red means data collection, and green means data process, yellow means data transmission)

FIGURE 11. (color online) Action sequences of testing node with a fixed duty cycle of 50%

Moreover, we still need to check the previous node's status, i.e., node 8; therefore, we synchronously record its action sequences at the exact time with node 3 and display in Figure 10. It can be figured out that node 8 has a more active sequence due to its comparatively adequate initial energy, besides it is not required to relay a former node's data. However, note that it uses an intermittent mode in Figure 10(c), as its next hop node (node 3) is in an insufficient state regardless of current node's state.

4.4. **Data throughput.** In this part, we evaluate the performance of our proposed method with a comparison to the following algorithms: one is for single node-level throughput comparison, we use a fixed duty cycle solution (50%, as shown in Figure 11) for the task arrangement in every time round, simply because it is in close proximity to the optimized numerical model described above, and then we statistically analyze the detailed data throughput of two algorithms during the same testing period; the other is for whole network-level contrast, where optimization mechanism RLTDPM [20] is used to satisfy the throughput on demand requirements, we also calculate and comparatively discuss the variance.

The first comparison results are shown in Figures 12 and 13, in which node 4's throughputs and node 7's throughputs during day 4 are recorded: here node 4 transfers directly to the sink and node 7 uses node 6 as a relay. It can be seen in Figure 12 that node 4 under our proposed algorithm wakes up much earlier and falls into deep sleep later than the case with fixed duty cycle, mainly due to its automatically self-regulation based on the external energy harvesting rate, and obviously the total throughput in DR-CRL is greater than the comparative object in fixed duty cycle. The other contrasting result is represented in Figure 13, where nodes 6 and 7 both experience with the fixed duty cycle sequence. We can find that node 7's throughput in DR-CRL is less than node 4, for it has a longer transmission distance and an unstable changing relay. However, it is greater

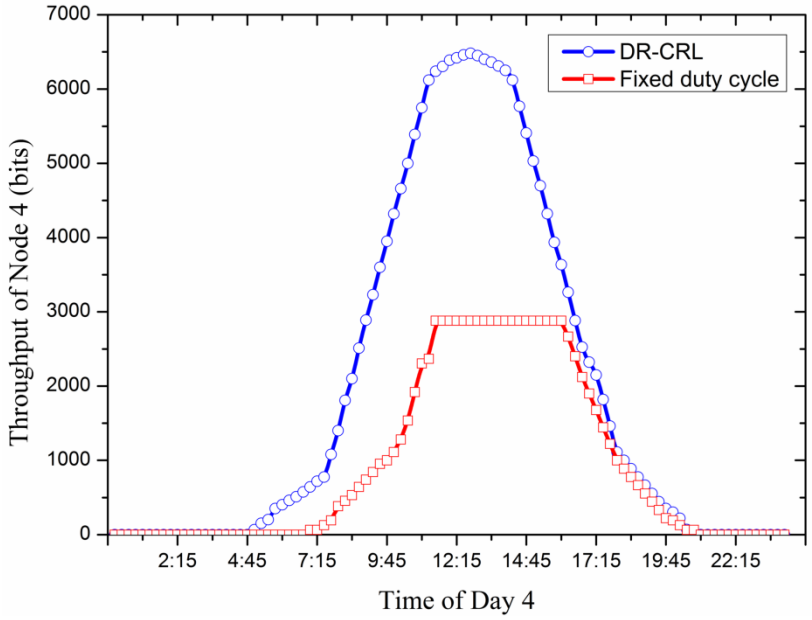


FIGURE 12. Throughput comparison of node 4 with 36% initial energy

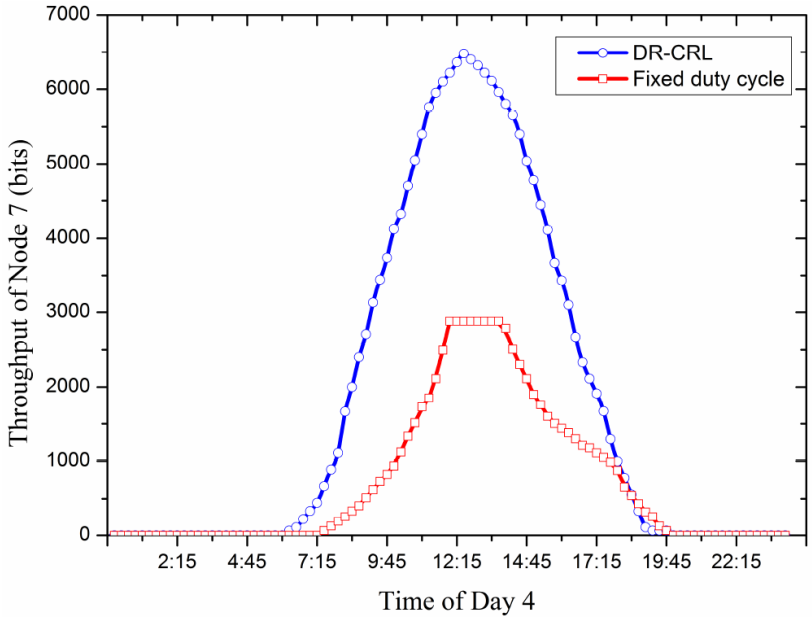


FIGURE 13. Throughput comparison of node 7 with 45% initial energy

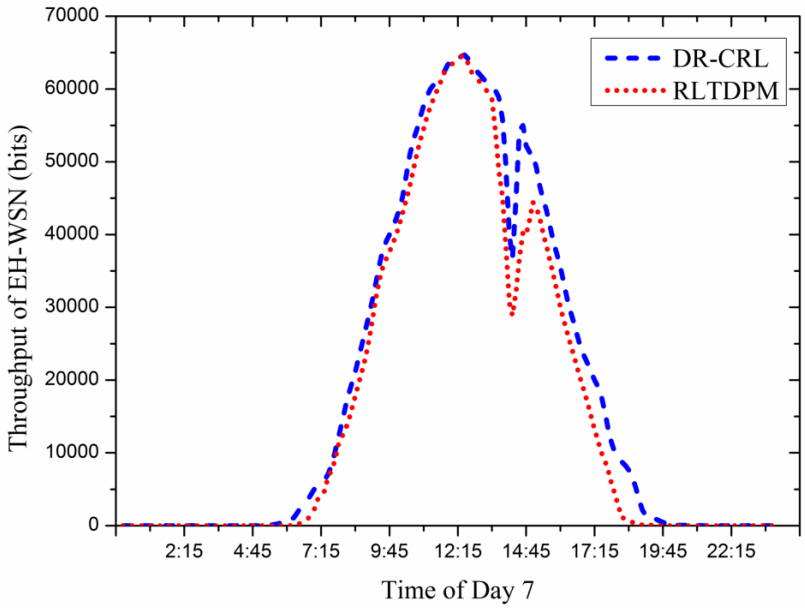


FIGURE 14. Throughput comparison of whole EH-WSN

than the fixed duty cycle version, mostly because of the intermittent transmission mode caused by the relay.

Then we conduct the whole throughput evaluation of EH-WSN, we run the two algorithms separately and calculate the number of whole data received by the sink in day 7, and results are shown in Figure 14. From the figure it can be deduced that both two algorithms keep up with the changing environmental energy sensitively, the total numbers of output data are almost identical until a sudden sinking of external energy, after that the DR-CRL resumes the total data transmission quickly and dependably, on the contrary RLTDPM mechanism could not handle the accident timely and result in a poorer performance during the subsequent period, on account of lacking the task schedule method in each time round.

In conclusion, these simulations indicate that the proposed DR-CRL algorithm could effectively keep the EH-nodes working in "energy-neutral" status, and efficiently regulate the task schedule to maximize the whole throughput. We also see that our definitions of the state result in a highly adaptive behavior. Our proposed method is able to adapt to changes in initial battery electricity, weather, and device parameters, which make the EH-node robust in its operation. In addition, our state definition and general reward formulation scheme allows for general application of our power management method such as wind energy harvesting, radio frequency energy harvesting and biological energy harvesting.

5. **Conclusions.** In this paper, a novel algorithm for sustaining perpetual operation of EH-WSN as well as maximizing its corresponding throughput by utilizing cooperative reinforcement learning method, named DR-CRL, was proposed. Numerical simulations are evaluated for a solar-power WSN to analyze the performance of the proposed algorithm. Parameters such as the EH-node's residual energy, the work/sleep duty cycle ratio, the action sequence versus the energy-harvesting rate, the individual EH-node's throughput and whole EH-WSN's throughput have been analyzed and reviewed. Results show the effectiveness and efficiency of the proposed DR-CRL algorithm. Future work will focus on the proposal of a Medium Access Control (MAC) protocol for EH-WSN and consider cluster-based topology structures.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 31700478, No. 31670554) and the Natural Science Foundation of Jiangsu Province, China (BK20150880, BK20161527). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] F. Ongaro and S. Saggini, Li-ion battery-supercapacitor hybrid storage system for a long lifetime, photovoltaic based wireless sensor network, *IEEE Trans. Power Electronics*, vol.27, no.9, pp.3944-3952, 2012.

[2] S. Lee, B. Kwon, S. Lee and A. C. Bovik, BUCKET: Scheduling of solar-powered sensor networks via cross-layer optimization, *IEEE Sensors Journal*, vol.15, no.3, pp.1489-1503, 2015.

[3] M. Shin and I. Joe, Energy management algorithm for solar-powered energy harvesting wireless sensor node for Internet of Things, *IET Communications*, vol.10, no.12, pp.1508-1521, 2016.

[4] N. Michelusi, L. Badia, R. Carli, L. Corradini and M. Zorzi, Energy management policies for harvesting-based wireless sensor devices with battery degradation, *IEEE Trans. Communications*, vol.61, no.12, pp.4934-4947, 2013.

[5] D. Djenouri and M. Bagaa, Energy-aware constrained relay node deployment for sustainable wireless sensor networks, *IEEE Trans. Sustainable Computing*, vol.2, no.1, pp.30-42, 2017.

[6] R. Rana, W. Hu and C. T. Chou, Optimal sampling strategy enabling energy-neutral operations at rechargeable wireless sensor networks, *IEEE Sensors Journal*, vol.15, no.1, pp.201-208, 2015.

[7] B. Pourpeighambar, M. Dehghan and M. Sabaei, Non-cooperative reinforcement learning based routing in cognitive radio networks, *Computer Communications*, vol.106, pp.11-23, 2017.

[8] M. I. Khan and B. Rinner, Energy-aware task scheduling in wireless sensor networks based on cooperative reinforcement learning, *IEEE International Conference on Communications Workshops (ICC)*, Sydney, Australia, pp.871-877, 2014.

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

[10] M. Emre, G. Gür, S. Bayhan and F. Alagöz, CooperativeQ: Energy-efficient channel access based on cooperative reinforcement learning, *IEEE International Conference on Communication Workshop (ICCW)*, London, United Kingdom, pp.2799-2805, 2015.

[11] S. Wu, J. Niu, W. Chou and M. Guizani, Delay-aware energy optimization for flooding in duty-cycled wireless sensor networks, *IEEE Trans. Wireless Communications*, vol.15, no.12, pp.8449-8462, 2016.

[12] A. Goldsmith, *Wireless Communications*, Cambridge Univ. Press, Cambridge, UK, 2005.

[13] M. Wiering and M. van Otterlo, *Reinforcement Learning: State-of-the-Art, ser. Adaptation, Learning, and Optimization*, Springer, 2012.

[14] K. J. Prabuchandran, S. K. Meena and S. Bhatnagar, Q-learning based energy management policies for a single sensor node with finite buffer, *IEEE Wireless Communication Letter*, vol.2, no.1, pp.82-85, 2013.

[15] R. C. Hsu, C.-T. Liu and W.-M. Lee, Reinforcement learning-based dynamic power management for energy harvesting wireless sensor network, *Next-Generation Applied Intelligence*, pp.399-408, 2009.

[16] C. Szepesvari, *Algorithms for Reinforcement Learning*, Morgan & Claypool Publishers, California, 2010.

[17] W. Liu, G. Qin, Y. He and F. Jiang, Distributed cooperative reinforcement learning-based traffic signal control that integrates V2X networks' dynamic clustering, *IEEE Trans. Vehicular Technology*, vol.66, no.10, pp.8667-8681, 2017.

[18] U. A. Khan and B. Rinner, Online learning of timeout policies for dynamic power management, *ACM Trans. Embedded Computing Systems*, vol.13, no.4, pp.1-25, 2014.

[19] I. Reda and A. Andreas, Solar position algorithm for solar radiation applications, *Solar Energy*, vol.76, no.5, pp.577-589, 2004.

[20] R. C. Hsu, C.-T. Liu and H.-L. Wang, A reinforcement learning-based ToD provisioning dynamic power management for sustainable operation of energy harvesting wireless sensor node, *IEEE Trans. Emerging Topics in Computing*, vol.2, no.2, pp.181-191, 2014.