An Energy-Efficient Strategy based on Q-Learning for Energy Harvesting-based Wireless Sensor Network

Jiayuan Wei, Xingyu Miao, and Yongqi Ge*
School of Information Engineering, Ningxia University, Yinchuan, 750021, China
jiayuanwei_nxu@outlook.com, miaoxy97@163.com, geyongqi@nxu.edu.cn

***Abstract*—In this paper, we consider a scenario in which the energy consumption transitions from sleep mode to the active mode for energy harvesting wireless sensor network (EHWSN). The sensor's energy consumption is primarily caused by the communication module. Due to the fact that the energy consumed during sensor state transitions is not the most essential component of the energy consumed, many experimental investigations exclude the energy used during state transitions when calculating the overall energy used by the sensors. However, when the frequency of sensor state transitions rises, the energy consumption associated with wireless sensor state transitions becomes unavoidable. As a result, we present an energy-efficient algorithm based on Q-learning that train the agent to minimize sensor state transitions and prevent further energy consumption in order to accomplish the goal of energy conservation. The experiment demonstrates that our method can save 18.07% of energy when compared to the Q-learning approach in ACES (ACES-QL), while also improving the residual energy of the energy storage device by 3.40%.**

***Index Terms*—reinforcement learning, wireless sensors network , energy harvesting**

I. Introduction

Over the last several years, a significant challenge has been how to reduce the energy consumption of sensors in the EHWSN area. Due to the limited energy capacity of battery-powered sensors, the battery life of sensor nodes is crucial for the stability of wireless sensor networks (WSN). When the energy stored in batteries runs out, the only option is to replace the battery. However, battery replacement costs are too expensive in a large-scale sensor deployment setting, particularly in remote places. Therefore, energy harvesting technology is used to the WSN [1]–[3], which harvests energy from the physical environment's resources. However, owing to the instability and unpredictability of energy harvesting, it becomes difficult to keep the sensor node's remaining energy in a stable state, and so several approaches are proposed to minimize sensor node energy consumption.

Many methods have been developed in recent years to meet the objective of reducing energy usage by adjusting the duty cycle of sensors. According to the energy consumption statistics for each sensor node module, the energy consumed by the sensor node in the wireless communication module (sending data, receiving data, and idle) accounts for a significant proportion, and thus reducing the energy consumption of sensor nodes can be considered by adjusting the sensing period (duty cycle) of sensor nodes. Q-Learning is introduced into the EHWSN to control the power management in various works [4], [5]. These methods achieve the sustainable operation of EHWSN by adjusting the duty cycle. However, they neglect the energy cost of sensor state transitions when computing the energy consumption of nodes. We expect that this portion of the energy usage is included in the overall energy consumption, which will make the experiment more representative of the real-world WSN environment. We observed that decreasing the node's total energy consumption is beneficial when combined with a reduction in transiton energy consumption.

The purpose of this paper is to examine an energy-saving technique based on Q-learning for adaptively adjusting the duty cycle of sensor nodes. The reinforcement learning (RL) agent determines the optimal duty cycle by interacting with the environment in response to the sensor node's energy status. Different duty cycle selections will result in the sensor transforming into various states. However, sensor node state transitions use energy. Specifically, when the data packet that a sensor node must broadcast is short, the energy used during the node state transition phase will be more than the energy used during data transmission. Frequent state changes of sensor nodes will result in node data collection and transmission delays, lowering node efficiency and using a significant amount of unnecessary energy. Thus, we propose a method for training the agent to maintain the current state of the sensor node rather than transitioning to a new state when the energy cost of state transition is high, which not only reduces the sensor node's energy consumption but also enables the sensor to operate sustainably. This is a departure and improvement from other classic duty cycle adjustment algorithms, and also makes our simulation experiments more convincing and more comprehensive. The following are the article's primary contributions.

- The EHWSN's sustainability is ensured by the use of a Q-learning dynamic power management system based on the TBE (break-even Time) strategy (TQL).
- Comprehensive testing findings demonstrate the TQL approach's benefit in terms of energy savings, the TQL approach can save 18.07 % used by nodes during operation.

The following article is divided into five sections. The related work is introduced in Section 2, and Section 3 illustrates the EHWSN sensor module. Section 4 describes the foundational information regarding the TBE approach, as

2767-6684/21/$31.00 ©2021 IEEE
DOI 10.1109/DSA52907.2021.00063

well as the concrete experimental technique and Q-learning algorithm. Section 5 presents the experimental findings and compares them to the ACES-QL method, and then give the conclusions in Section 6.

II. Related Work

Numerous research have been conducted in recent years on sensor power management. Kansal et al. [3] proposed the first adaptive duty cycle algorithm based on an energy prediction model, which adjusts the duty cycle based on predicted energy, and that article introduced the concept of energy neutrality, which is defined as when the energy consumed by a node in a wireless sensor network is less than or equal to the energy collected by the node. Additionally, Kansal et al. [6] subsequently presented a dynamic duty cycle adjustment (DDCA) approach for lowering the duty cycle when the collected energy is low and increasing it when the harvested energy is high. With the development of reinforcement learning applications in wireless sensor networks, Hsu et al. [5] proposed the RLDPM method, which employs RL agent to adjust the duty cycle. Their policy's objective is to ensure that the QoS trained through Q-learning is capable of satisfying the QoS required by the sensor node itself in the energy neutral state, which means that sensors are capable of performing data transformations in a timely manner in wireless sensor networks, and they build the reward function in such a manner that it is also associated with the degree of service provided.

Furthermore, Hsu et al. [7] proposed the RLTDPM method for meeting demand for throughput and sustaining perpetual operation by varying the duty cycle, and the throughput is separated into several grades. They simplified the designed reward function, the sigmoid and Mexican hat functions, to a specified reward function value in consideration of the low power consumption needs of sensor nodes. Additionally, Wu et al. [8] suggested a novel way for meeting the throughput requirement and compared it to the RLTDPM technique, which has a greater impact and also enhances energy consumption efficiency. The article contributes by detailing the energy consumption calculations for sensors used in data collecting, processing, transmission, and reception, and emphasizing that energy consumption is proportional to the distance of data transmission. Simultaneously, the residual energy of nodes with varying beginning energies will all converge to a stable state.

Many scholars have also recommended using a fuzzy inference system to manage energy dynamically in order to decrease the complexity of state input and reward in reinforcement learning. Hsu et al. [9] recommended that dynamic power management be accomplished via the use of the reinforcement learning technique RLFR with fuzzy rewards. The scenario with fuzzy inference system incorporates uncertainty and ambiguity, and the fuzzy reward function is linked to the fuzzy state and energy neutrality. Additionally, in [10], a strategy for dynamic energy management using fuzzy inference systems is described. Hsu et al. also offer fuzzy-RL, which divides the state vector into fuzzy subsets as input and obtains the weighted average duty cycle as output, by comparing the three-month remaining energy information with other approaches, the suggested technique has the lowest root mean square deviation (RMSD) of energy neutrality. Hsu et al. have provided a method for examining the energy neutrality of sensors using fuzzy Q learning (FQL) [11]. Under varying beginning energy conditions, the residual energy of nodes will converge to a state that would sustain perpetual operation.

Due to the harvested energy's significant fluctuation, it's difficult to forecast. In [12], Aoudia et al. introduced an energy harvesting management strategy focused at the harvested energy, which included a power manager (PM) for each energy harvesting node. To our knowledge, this is the first paper to address the challenge of developing PM for EH-nodes via the use of fuzzy energy harvesting. Shresthamali et al. [13] proposed configuring a power management for each sensor node, with the power management adjusting the duty cycle when the RL agent interacts with the environment. In order to lower system energy consumption to attain self-sustainability, Prauzek et al. in [14] reduced that the energy consumption of the hardware and software and the temperature sensors are capable sample constantly to monitor environmental characteristic. Fraternali et al. [15] proposed the ACES method, which uses the RL to adjust the sleeping time of sensor nodes to reduce energy consumption and sensor node death rates, which is the first deployed in the real world and collect measurements (light intensity and supercapacitor voltage level), using these data traces to establish a simulation environment, in a deployment experiment of 60 nodes. In a deployment trial with 60 nodes, the nodes only stopped operating 0.1% of the time over the course of two weeks.

III. Sensors module of the EHWSN

There are usually two types of energy harvesting models, one is the harvest-storage-use [16] model, and the other is the harvest-use (storage) [17] model. The harvest-storage-use model means that part of the harvested energy is used to power the nodes and a small part is stored in the energy buffer. Therefore, this method will cause energy to be wasted in the energy storage process. The basic idea of the harvest-use (storage) mechanism is that the harvested energy directly powers the node, and only the collected remaining energy is stored. If the collected energy is insufficient, the node Energy will be drawn from storage devices, the harvest-storage-use mechanism used in this article. Therefore, the energy harvesting embedded system (EHES) generally consists of three parts, including the energy harvesting unit, the energy storage unit, and the energy consumed unit. And we will present the relationships between them as shown in Figure1.

A. Energy harvesting unit

The energy harvesting unit, such as the solar panel is used to collect energy from the sun, collects energy from the ambient environment with an uncertain and fluctuant collection rate $R_p(t)$, which is a function of time, but it is gradually assumed to be constant value based on the [18]–[20]. The harvested energy during the $[t_1, t_2]$ time interval is defined $E_{harvest}(t_1, t_2) = \int_{t_1}^{t_2} R_p(t)\,dt$, and which is stored in the energy storage unit to provide energy for the sensor node.

B. Energy storage unit

When the energy output of the energy harvesting unit is insufficient, the energy storage unit releases energy to maintain task execution. The energy of the energy storage unit fluctuates between two thresholds E_{max} and E_{min}, where E_{max} denotes the maximum capacity that storage device and E_{min} denotes the minimum capacity of that storage device storage. And during the time $[t_1, t_2]$, the storage energy of storage device is denoted $E_{max} \geq E_{store}[t_1, t_2] \geq 0$. If the $E_{store}[t_1, t_2]$ is positive, which means the harvested energy is more than the consumed energy during the time $[t_1, t_2]$ interval, on the other hand, if the $E_{store}[t_1, t_2]$ is negative, which means the harvested energy is less than the consumed energy during that time interval.

C. Energy consume unit

The energy consumption unit refers to running real-time tasks, its energy comes from the energy harvesting unit and the energy storage unit in the embedded system. The energy management unit is responsible for ensuring the energy storage unit can meet the energy consumption requirements of the energy consume unit to avoid energy exhaustion.

According to the EHES energy model shown in Figure 1, we can observe the harvested energy in the energy harvesting unit, denoted as $E_{harvest}$, and we defined the stored energy in the energy storage unit as E_{store}, and the consumed energy of sensor node in energy consumed unit, denoted as E_{node}. The main energy management unit that is responsible for coordinating the balance between the various parts of the energy according to the received information, $E_{harvest}$, E_{store}, E_{node}, determines the duty cycle of the sensor nodes.

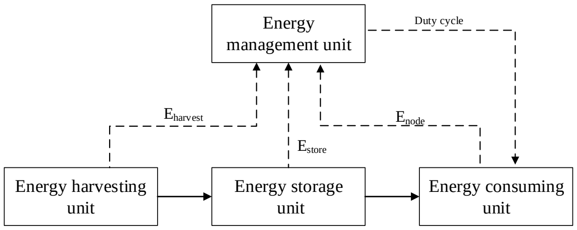

Fig. 1. Wireless sensor energy composition model of energy harvesting embedded system.

In our experiment, the data is real and derived from [15], the current is measured by using the National Instrument USB-6210 with MATLAB, and the lighting data acquired according to the actual lighting collection situation, we approximate the irregular real data to the data within a certain range to establish a simulation environment. Francesco et. al divided the lighting data into 10 different levels and verified the rationality of the lighting data according to the actual deployment. And we use the super-capacity (SC) as the energy storage unit because the super-capacitor can carry out multiple charges and discharges cycles compared to batteries.

IV. The basical knowledege and Q-learning algorithm

A. TBE Strategy

In this paper, Q-earning based on the TBE strategy for the energy harvesting wireless sensor network method is proposed. TBE, an energy monitoring management strategy, is composed of three parts, as shown in formulation (1), T_t is the time to transition from high-energy mode to low-energy mode, T_o is the operating time during the low-energy mode, and T_r is the transition from low-energy mode to high-energy mode time. The TBE strategy means that the energy consumption of the computing resource when it is in the high energy consumption mode during the period is the same as the energy consumption of the computing resource when the computing resource is converted to the low energy mode during the period and returns to the high energy mode.

$$TBE = T_t + T_o + T_r \tag{1}$$

Taking the StrongARM SA-1100 processor as an example, suppose the delay for the StrongARM SA-1100 to switch from operating mode to sleep mode is T_t, and the delay for switching from sleep mode to operating mode is T_r, and the energy of the conversion process The energy consumption is E_0, the energy consumption in running mode and sleep mode are respectively P_w and P_s, and the time from entering sleep mode to returning to running mode is T_{be}. If the StrongARM SA-1100 consumes equal energy in running mode and sleep mode, Which satisfies the formula (2)

$$P_w \times T_{be} = E_0 + P_s \times (T_{be} - T_t - T_r) \tag{2}$$

B. Sensor Duty Cycle

The state of the sensor node includes gradually the sensing state and the sleeping state, we define the duty cycle period of wireless sensor nodes i as T_i, which is composed of several continuous sensing slots and sleeping slots. And the duty-cycle period is given by the following equation.

$$Duty - cycle - period = \frac{1}{\tau}\sum\nolimits_{t=0}^{\tau} T_{sensing} + T_{sleeping} \tag{3}$$

where the $T_{sensing}$ represents the time of sensor nodes remain active and is set to a constant here due to that it relies on the sensing characteristics and the communication mechanism [21], therefore, And the duty cycle represents the ratio of sensing time to the duty cycle period, therefore, we can compute the duty cycle ratio by the following equation.

$$Duty - cycle - ratio = \frac{T_{sensing}}{Duty - cycle - period} \tag{4}$$

C. Reinforcement learning

We choose Q-learning as our algorithm among RL algorithms. The decision-maker, agent of reinforcement learning, chooses an action a by interacting with the environment in a random initialization state s according to ϵ-greedy policy, which means to choose the max Q-value in ϵ probability in Q-table, otherwise, choose the random Q-value in 1- ϵ probability, after the action is performed, the agent will receive a reward or punishment, which indicates how well the action a performed. If a reward is received, the agent will update by the following equation, its Q-value to a larger Q value to increase the opportunity that to be chosen once again, but if a punishment is received, the agent will update its Q-value to a smaller Q value to reduce the opportunity that to be chosen once again. Then the Q-value will also be updated to reach a new state $\tilde{s}$ by the following equation and choose a action again until the episode is final.

$$\begin{aligned} Q(s_t,a_t) = \alpha\{r_{t+1} + \gamma \max Q(s_{t+1},a_{t+1}) - Q(s_t,a_t) \\ + Q(s_t,a_t)\} \end{aligned} \quad (5)$$

Q(s_t,a_t) is updated by adding the old Q-value (the first on the right side of the equation Q(s_t,a_t)) to the part that needs to be adjusted, whereas Q$(s_t,a_t)$is the original Q-value of state-action quality, where the s_t represents the state in t time, then the agent will receive a reward r_{t+1} and a future discount reward, a_t represents the chosen action by the agent in t time, and γ is called discount factor $\in$ [0,1] that is set to 0.99, which is introduced to show the degree of dependence for the reward of future and avoid the total reward is infinite to be unable to converge, when the action a is performed, the agent will obtain a discount for future rewards (the next Q-value) besides the timely reward, and we call it q_{target}. The original value in the Q-table is the estimated value, we call it $q_{predict}$, the part that needs to be adjusted is the difference between q_{target} and $q_{predict}$, the α $\in$[0,1] is set to 0.1 here, a constant value representing the learning rate.

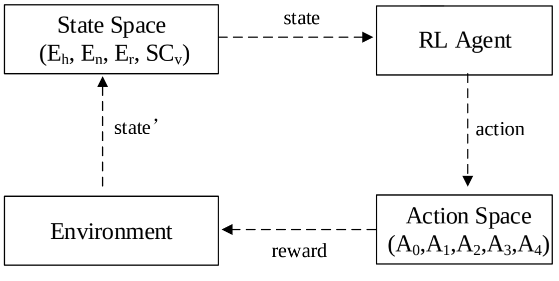

Fig. 2. The reinforcement learning operation scenario of energy harvesting sensor node.

The goal of reinforcement learning is to learn a strategy to maximize reward, that is, the agent is expected to perform a series of actions to obtain as many average returns as possible. To evaluate the expected return of a strategy, a value function needs to be defined. The value function is divided into a state-value function and aan action-value function. The state-value function represents the future obtained reward by the state s based on the strategy π (refers to the probability of performing action a in state s). And the action-value function refers to the expected reward return obtained after performing action a for state s based on the strategy π. Based on formula 6, the quality of the action a can be observed from formula 7. The equation 7 can also be written in the form of a compound reward (timely reward in state s and the probability of the next state multiplied by the state value of the next state), which can more intuitively show the relationship between equation 6 and equation 7.

$$V_\pi(s) = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s] \quad (6)$$

$$Q_\pi(s) = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a] \quad (7)$$

D. RL formulation

States: Based on the energy harvesting wireless sensor network model, we design that the state space consists of four state vectors. And in order to observe the energy consumption of sensor node, we introduce the residual energy as a state vetor criterion.

$$(S_h, SC_v, S_n, S_r) \in S.$$

$S_h(t)$: represents the harvested energy in t time solt according to the collected indoor light energy, and the light is divided into 10 levels due to the difference in the light intensity.

$$S_h(t) \in L(i) = \{Light_1, Light_2, \cdots, Light_{10}\}, 1 \leq i \leq 10$$

where $Light_i$ represents that the higher light intensity with the larger number of i.

$SC_v\ (t)$: represents the SC voltage level in t time solt, the SC voltage is also divided into 10 levels.

$$SC_v(t) \in V(i) = \{SC_{v1}, SC_{v2}, \cdots, SC_{v10}\}, 1 \leq i \leq 10$$

where SC_{v1} represents the min SC voltage level and is 2.1V, if the super-capacity voltage reaches $<$ 2.1V, the system will terminate all operations to replenish energy. SC_{v10} represents the max SC voltage level, and in this study the super-capacitor, providing maximum voltage 5V is applied.

$S_n\ (t)$: represents the consumed energy of nodes in t time solt, which also are used to design the reward function. And we also divided it into different levels to reduce the computational complexity.

$S_r\ (t)$: represents the residual energy of nodes in t time and the residual energy is expressed by the residual energy level. It is given by the following equation.

$$E_r(t+1) = E_r(t) + E_h(t) - E_n(t) - E_p(t) \quad (8)$$

where $E_r(t+1)$ is the residual energy in t+1 time solt, which is composited of the residual energy of sensor node in t time solt $E_r(t)$, a finite amount energy that the system harvests in t time solt $E_h(t)$, and the total consumed energy $E_c(t)$ including the energy consumed by the node $E_n(t)$ in t time solt and the energy consumed by sensor state transition $E_p(t)$ in t time solt.

Action: We assign the 15minutes as a timestep of the take action of node and 24 hours as an episode, in other words, the sensor nodes take another action every fifteen minutes, which is reasonable if the timestep is too small, the communication energy of the nodes will be wasted. The action $(A_0, A_1, A_2, A_3, A_4)$ is that the sleep time of the designed node is 900s, 300s, 60s, 15s, 0s.

TABLE I
NODE SLEEP TIME BASED ON ACTION INDEX

Action Index	Node Sleeping Time(s)
0	900
1	300
2	60
3	15
4	0

Reward function: Based on the main idea of saving energy of sensor node, we compared the energy consumption of sensors with different sleep times with those without sleep time. Since the switching energy consumption is calculated when the state transitions, it is likely that the energy consumption with sleep time is greater than the energy consumption continuously in the sensing state. In this case, we train the agent to maintain a continuous sensing state, not only ensure timeliness of data transmission but also save energy. And the setting of the reward function should consider ensuring the sustainable operation of the node. Therefore, the reward function is designed if the SC voltage reaches $\leq$ 3V, the agent will receive a -300 reward, otherwise, the agent will consider the difference between the energy consumed by the executed action and the sensing energy $E_s(t)$ without sleep time, we introduce the μ to balance the node sleeping time T_{node} and the energy consumption, alternative reward is receied by learning agent is follows.

$$reward = \begin{cases} \mu T_{node} + \frac{1-\mu}{E_c(t)}, & E_c(t) = E_s(t) \\ \mu T_{node} + \frac{1-\mu}{(E_s(t)-E_c(t))E_c(t)}, & otherwise \end{cases} \tag{9}$$

Algorithm 1 describes the application implementation of Q-learning on sensor nodes. Lines 1-3 initialize the Q table, and line 4 means to obtain the current state, supercapacitor voltage and current informations. We define the time step as 15 minutes, and the agent chooses an action a according to the ϵ-greedy strategy in this interval. The ϵ is set to the minimum value until the time interval T ends, and ϵ will be updated, the agent will obtain a reward and the next state to update the Q table. Then the sensor node updates the sleep time and sends the new state again after the next 15-minute interval. In the experiment, we set some Q-learning hyper-parameters, as detailed in the Table II.

Algorithm 1 Q-learning Algorithm for EHWSN

```
1: Initialize $q_{table}$ as an empty set
2: Initialize action a, state $s_{curr}, s_{next}$ time passed = 0
3: $\epsilon = \epsilon_{min}$
4: $s_{curr} \leftarrow state$
5: while $timepassed < episode\_duration$ do
6:     Choose a from s using policy derived from $\epsilon$-greedy
7:     wait for T time units, time passed += T
8:     receive reward and $s_{next} \leftarrow state$
9:     Update $q_{table}$ using below interation formula
10:    $Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma maxQ(s',a') - Q(s,a)\}$
11:    $\epsilon = \epsilon + \triangle$
12:    $s_{curr} \leftarrow s_{next}$
13: end while
```

TABLE II
Q-LEARNING HYPER-PARAMETERS USED FOR SIMULATIONS.

Hyper-Parameter	Value
Learning rate	0.1
Reward-decay (γ)	0.99
Epsilon max (ε_{max})	1
Epsilon min (ε_{min})	0.1
Epsilon increment ($\triangle$)	0.0004
Episode Duration	24 hours
Wait Time T	15 mins

Since we changed the sleeping time of the sensor in the 15-minute interval, the working mode of the sensor will also be different. We simulated the working mode of the sensor in the 15-minute interval as shown in Figure 3, the sensing mode is shown in Figure 3(a) when the sleeping time is 300, the sensor will perform three times sensing, the sensing mode is shown in Figure 3(b) when the sleeping time is 900, the sensor will perform one time sensing, and the sensing mode is shown in Figure 3(c) and Figure 3(d) when the sleeping time is 60 and 15 respectively, the sensor will perform several times sensing. We can observe that the working modes of task sets τ_1 and task sets τ_2 from Figure 3(e) and Figure 3(f), when

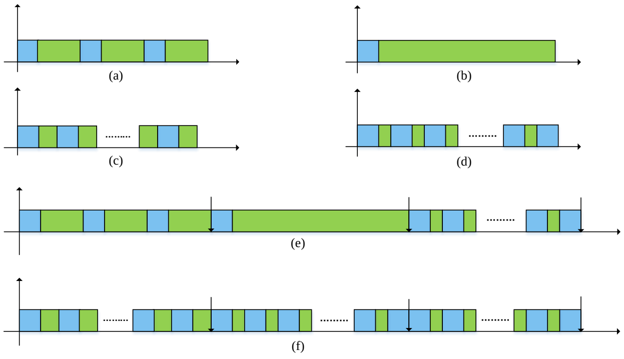

Fig. 3. Wireless sensor working modes of energy harvesting embedded system (Here blue slot means senseing, green means sleeping).

the action with a sleep time of 15 is executed, compared with the action with a sleep time of 900, the number of sensor sensing will significantly increase. And we compare that the energy consumption of the first hundred episodes as shown in Figure 4, we can also observe that sometimes that the energy consumed by using ACES-QL method is more than the energy consumed by using TQL from Figure 4. This is due to the increase in the number of times the sensor switches between the sensing state and the sleep state, resulting in an increase in the total energy consumption. This provides us with a good trend to show that the method is implementable and useful.

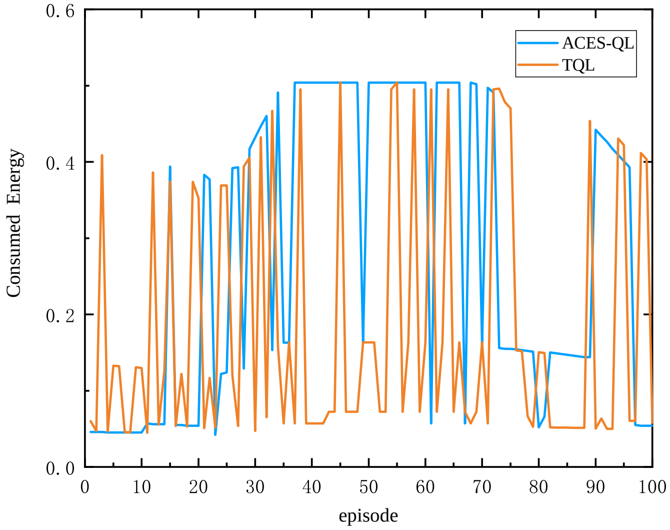

Fig. 4. Energy consumption of the wireless sensor node in different working modes.

V. Simulation Results

Our simulation results are based on the initial energy setting of 52.8% of the maximum capacity of the super capacitor, and simulate indoor lighting environment. The experiment was carried out after modifying some data of [8]. To further understand how well the energy saving of the proposed method TQL compared with the ACES-QL mmethod, in this study, we focus on the three-month energy consumption comparison from the vernal equinox to the summer solstice as shown in Figure 5. It can be intuitively seen that the ACES-QL method consumes more energy than the TQL method from Figure 5, and the convergence rate of the proposed method is more quickly than the ACES-QL method, this is because the more times of state transition are preferred to consume the more energy in ACES-QL method, and the agent of the TQL method avoids to perform the different state transition rather than executing sensing in a continuous-time solt. which save the energy consumption of sensor node and satisfy the requirement of the throughput that is the actual application object for EHWSN system [22]. We can observe that the TQL saves about 18.07 % of the energy consumption than the ACES-QL method, the saved energy consumption contributes to extend the life of the wireless sensors. The TQL algorithm satisfies the low power consumption requirements of sensor nodes without degrading the performance of sensor nodes.

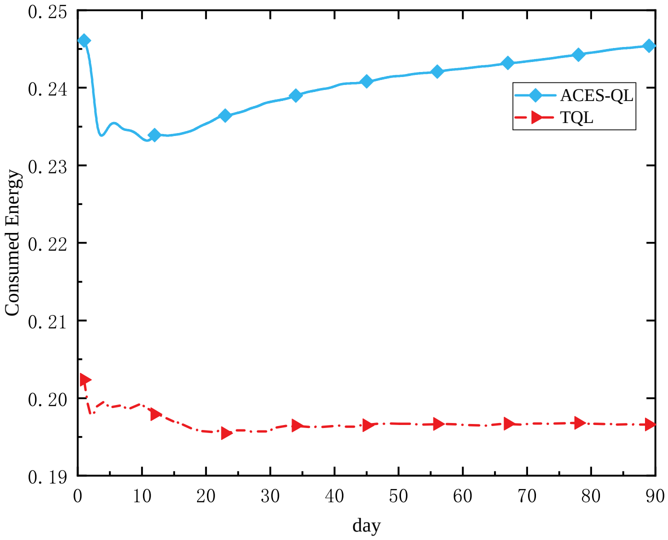

Fig. 5. Average consumed energy result comparisons of long term simulation showing day 0 to 90.

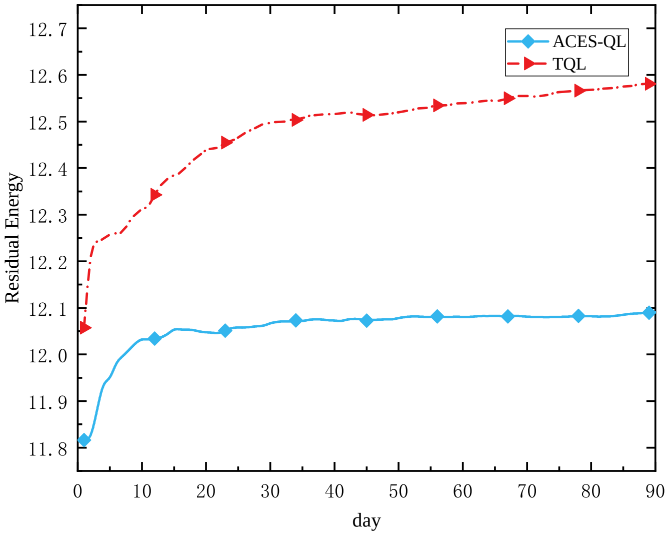

Fig. 6. Average residual energy result comparisons of long term simulation showing day 0 to 90.

We compared the remaining energy of the sensor node similarly as shown in Figure 6. The amount of remaining energy (RBE) also indicates the degree of energy saving of the sensor node. Both TQL and ACES-QL can achieve the goal that the residual energy of the sensor node tends to a stable state, and reduce the energy wasted by the agent during the exploration period. However, due to the consumed energy of the sensor node by ACES-QL is more the consumed energy of the sensor node by the TQL, the remaining energy by the ACES-QL is also expected. We can observe that the proposed method TQL has a significant improvement over the remaining energy of the node for ACES-QL method, and increases the remaining energy by about 3.40% compared with ACES-QL method, which also proves that the proposed method TQL can reduce efficiently the energy consumption of the sensor node, extend the life of the sensor node, and ensure the sustainable operation of the sensor node. Regarding the impact of energy saving on sensor performance, about the sampling performance of the sensor, it is verified that under different sampling frequencies, the sampling performance by the sensor in the simulation and the sampling performance by the actual sensor almost coincide, which also indicates the real-time and accuracy of the sensor sampling in the simulated environment.

VI. Conclusion

In this study, we propose a method for controlling the energy consumption of sensor nodes based on reinforcement learning. This method considers the energy consumption of sensor state transition and tries to find a suitable transition time to reduce the number of sensor transitions. The experimental results demonstrated that the energy consumption of the sensor nodes of the proposed method is in a lower energy consumption state, and the average remaining energy is maintained in a greater energy state among other methods. And we analyzed the influence of energy saving on the sampling rate of the sensor. Of course, the metric of the sensor performance not only includes the sampling rate, we hope to study more about the performance analysis in the actual deployment in the future. And the consideration of experiments under different conditions, we hope that the exploration is not limited to simulation experiments, but also on actual deployment, the indicators for evaluating the performance of the algorithm will also be improved, and more detailed experimental discussions will be explored to support the credibility of the experimental results.

Acknowledgment

This study was supported in part by the Young Scholar in Western China of Chinese Academy of Sciences under grant no. XAB2018AW12, in part by the Ningxia Key Research and Development Projects under grant no. 2018BEB04020, in part by the Natural Science Foundation of Ningxia under grant no. 2021AAC03118, and in part by the Natural Science Foundation of Ningxia under grant no. 2021AAC03041.

References

[1] C. Moser, L. Thiele, D. Brunelli, and L. Benini, "Adaptive power management in energy harvesting systems," in *2007 Design, Automation & Test in Europe Conference & Exhibition*, pp. 1–6, IEEE, 2007.

[2] C. Moser, J.-J. Chen, and L. Thiele, "Reward maximization for embedded systems with renewable energies," in *2008 14th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, pp. 247–256, IEEE, 2008.

[3] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, "Power management in energy harvesting sensor networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 6, no. 4, pp. 32–es, 2007.

[4] R. C. Hsu, C.-T. Liu, and W.-M. Lee, "Reinforcement learning-based dynamic power management for energy harvesting wireless sensor network," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 399–408, Springer, 2009.

[5] R. C. Hsu, C.-T. Liu, K.-C. Wang, and W.-M. Lee, "Qos-aware power management for energy harvesting wireless sensor network utilizing reinforcement learning," in *2009 International Conference on Computational Science and Engineering*, vol. 2, pp. 537–542, IEEE, 2009.

[6] A. Kansal, J. Hsu, M. Srivastava, and V. Raghunathan, "Harvesting aware power management for sensor networks," in *Proceedings of the 43rd annual design automation conference*, pp. 651–656, 2006.

[7] 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 Transactions on Emerging Topics in Computing*, vol. 2, no. 2, pp. 181–191, 2014.

[8] Y. Wu and K. Yang, "Cooperative reinforcement learning based throughput optimization in energy harvesting wireless sensor networks," in *International Journal of Innovative Computing, Information and Control(IJICIC)*, pp. 1–18, IEEE, 2018.

[9] C.-T. Liu and R. C. Hsu, "Dynamic power management utilizing reinforcement learning with fuzzy reward for energy harvesting wireless sensor nodes," in *IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society*, pp. 2365–2369, IEEE, 2011.

[10] R. C. Hsu, T.-H. Lin, S.-M. Chen, and C.-T. Liu, "Dynamic energy management of energy harvesting wireless sensor nodes using fuzzy inference system with reinforcement learning," in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, pp. 116–120, IEEE, 2015.

[11] R. C. Hsu and T.-H. Lin, "A fuzzy q-learning based power management for energy harvest wireless sensor node," in *2018 International Conference on High Performance Computing & Simulation (HPCS)*, pp. 957–961, IEEE, 2018.

[12] F. A. Aoudia, M. Gautier, and O. Berder, "Fuzzy power management for energy harvesting wireless sensor nodes," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2016.

[13] S. Shresthamali, M. Kondo, and H. Nakamura, "Adaptive power management in solar energy harvesting sensor node using reinforcement learning," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 16, no. 5s, pp. 1–21, 2017.

[14] M. Prauzek, N. R. Mourcet, J. Hlavica, and P. Musilek, "Q-learning algorithm for energy management in solar powered embedded monitoring systems," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–7, IEEE, 2018.

[15] F. Fraternali, B. Balaji, Y. Agarwal, and R. K. Gupta, "Aces: Automatic configuration of energy harvesting sensors with reinforcement learning," *ACM Transactions on Sensor Networks (TOSN)*, vol. 16, no. 4, pp. 1–31, 2020.

[16] S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey and implications," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 443–461, 2010.

[17] S. Peng and C. P. Low, "Throughput optimal energy neutral management for energy harvesting wireless sensor networks," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 2347–2351, IEEE, 2012.

[18] C. Moser, D. Brunelli, L. Thiele, and L. Benini, "Lazy scheduling for energy harvesting sensor nodes," in *IFIP Working Conference on Distributed and Parallel Embedded Systems*, pp. 125–134, Springer, 2006.

[19] Y. Abdeddaïm, Y. Chandarli, and D. Masson, "The optimality of pfpasap algorithm for fixed-priority energy-harvesting real-time systems," in *2013 25th Euromicro Conference on Real-Time Systems*, pp. 47–56, IEEE, 2013.

[20] R. Jayaseelan, T. Mitra, and X. Li, "Estimating the worst-case energy consumption of embedded software," in *12th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS'06)*, pp. 81–90, IEEE, 2006.

[21] S. M. Finnigan, A. K. Clear, G. Farr-Wharton, K. Ladha, and R. Comber, "Augmenting audits: Exploring the role of sensor toolkits in sustainable buildings management," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–19, 2017.

[22] A. Murad, F. A. Kraemer, K. Bach, and G. Taylor, "Autonomous management of energy-harvesting iot nodes using deep reinforcement learning," in *2019 IEEE 13th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, pp. 43–51, IEEE, 2019.