*International Journal of Advanced Smart Convergence Vol.7 No.4 19-26 (2018)*
*http://dx.doi.org/10.7236/IJASC.2018.7.4.19*

IJASC 18-4-3

# Reinforcement Learning-based Duty Cycle Interval Control in Wireless Sensor Networks

Shathee Akter, Seokhoon Yoon*

*Department of Electrical and Computer Engineering, University of Ulsan, Korea*
*erittrashathee@gmail.com, seokhoonyoon@ulsan.ac.kr**

***Abstract***

*One of the distinct features of Wireless Sensor Networks (WSNs) is duty cycling mechanism, which is used to conserve energy and extend the network lifetime. Large duty cycle interval introduces lower energy consumption, meanwhile longer end-to-end (E2E) delay. In this paper, we introduce an energy consumption minimization problem for duty-cycled WSNs. We have applied Q-learning algorithm to obtain the maximum duty cycle interval which supports various delay requirements and given Delay Success ratio (DSR) i.e. the required probability of packets arriving at the sink before given delay bound. Our approach only requires sink to compute Q-leaning which makes it practical to implement. Nodes in the different group have the different duty cycle interval in our proposed method and nodes don't need to know the information of the neighboring node. Performance metrics show that our proposed scheme outperforms existing algorithms in terms of energy efficiency while assuring the required delay bound and DSR.*

***Keywords:*** *Duty cycle, Wireless Sensor Networks, Q-learning, Delay bound*

## 1. Introduction

In Wireless Sensor Networks (WSNs), network lifetime maximization of battery-constrained sensors has received a great amount of attention as sensors are expected to operate autonomously for a long period. WSNs are generally deployed in a different environment such as remote and hostile regions which makes it costly to replace exhausted batteries or even impossible sometimes [1]. Hence, there are numerous proposed methods and ongoing studies related to optimizing power usage in battery-constrained sensors. One of the effective ways to reduce energy consumption is duty cycling mechanism. Energy consumption of the node is low when the duty cycle interval is large, however, a large duty cycle interval leads to the higher delay which can be crucial for delay constrained applications namely fire detection and accident signaling, etc. WSNs have various applications, for example; healthcare surveillance, environmental monitoring, and military surveillance, *etc*. [2]. Some applications have some specific delay requirements (*e.g*., packets should be delivered within a specific delay bound in public safety application) yet most of the existing studies only consider increasing energy efficiency while ignoring the delay requirements of the applications.

Some existing studies have considered delay requirements of the applications while obtaining optimal

Manuscript Received: Sep. 2, 2018 / Revised: Sep. 7, 2018 / Accepted: Sep. 13, 2018
Corresponding Author: seokhoonyoon@ulsan.ac.kr
Tel: +82-52-259-1403, Fax: +82-52-259-1687
Department of Electrical and Computer Engineering, University of Ulsan, Korea

energy efficiency. For example, [3] applied Reinforcement Learning (RL) scheme in a star topology to learn to adapt the duty cycle according to traffic conditions. Furthermore, MAC protocol was specified in their method. Vu *et al*. [4] and Dao *et al*. [5] has estimated the maximum duty cycle interval based on end-to-end delay distribution which is a sum of one-hop delay distribution. However [5] is only applicable on the random deployment of WSNs in a circular area. In terms of energy efficiency, our proposed method gets better performance compared to [4].

Considering these issues, we propose a duty cycle interval approximation scheme to maximize network lifetime based on Q-learning. The objective of our model is to expand the network lifetime, *i.e.* minimizing the total energy consumption of the network while satisfying the requirement, for instance, the delay bound and DSR. DSR is the ratio of the packets arrive at the sink before given delay bound.

The rest of this paper is organized as follows: Section 2 presents the network model, Section 3 provides a brief background on Reinforcement Learning and the optimization problem formulation for duty cycle interval using the RL framework. In section 4, parameters of the model and comparison of DCI-RL (Duty cycle interval control using reinforcement learning) with two other methods have shown. Finally, Section 5 concludes the paper.

## 2. Network Model

We consider a multi-hop WSN with $n$ static sensor nodes denoted by $n_k$ ($k = 1,2,3 \ldots . n$) and a single sink node [4] where transmission range is $R$. Different deployment strategies are applicable (such as Rectangular, Random, manual and circular, etc.) depending on the demand. According to the distance from the sink to the node, nodes are divided into multiple groups and indexes of the groups are fixed. Groups index is denoted by $G_j$ where $j = 1,2,3 \ldots . L$. $L$ is the total number of groups in the network. In this paper, we denote $n_0$ and $n_i^j$ as the sink node and $i^{th}$ node of group $j$ where $i = 1,2,3 \ldots N_j$ respectively. $N_j$ is the total number of nodes in group $j$.

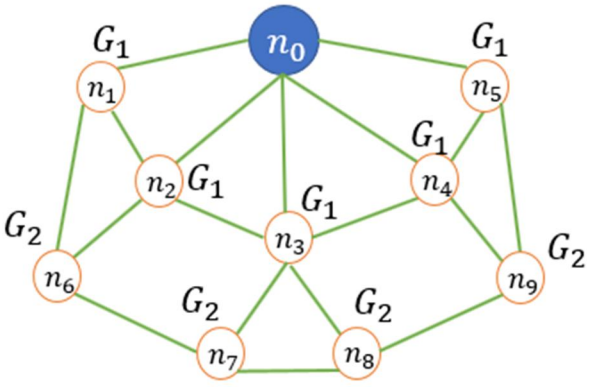


**Figure 1. An example of network model**

Sleep/wake-up scheme has been taken into account in this model to prolong the network lifetime. In each duty cycle interval, nodes stay awake for certain periods and sleep for the rest of the time whereas the sink is always active. Nodes follow asynchronous duty cycling mechanism, i.e., they don't maintain the time schedule of other nodes and wake up randomly, independently. In his paper, duty cycle interval and active period are denoted by $T$ and $a^t$ respectively. We assume that nodes wake up only once in each duty cycle interval whilst the active period is long enough to transmit all the packets.

## 3. Duty Cycle Interval Controller Based on Q-Learning

In this section, we present an optimization problem formulated as a Markov decision process (MDP) to estimate the duty cycle interval based on Reinforcement learning (RL) theory which will reduce the total energy consumption of the network. RL has been applied in a centralized manner in our proposed model where sink knows all the necessary network parameters to calculate the duty cycle interval (for instance number of potential forwarding candidate (PFC), groups, *etc.*).

At first, we are going to introduce background on RL to understand the duty cycle interval control scheme and then approximate the duty cycle interval using Q-learning.

### 3.1 Background on Reinforcement Learning

In Reinforcement Learning Framework, Agents interact with its environment directly and learn by taking action repeatedly [6] where the environment is fully observable MDP. An MDP is a tuple $(\boldsymbol{S}, \boldsymbol{A}, \tau, \boldsymbol{r}, \boldsymbol{\gamma})$, where $\boldsymbol{S}$ is the state space, $\boldsymbol{A}$ is a finite set of actions, $\tau$ is the state transition probability function;

$$\tau(s, a, s') = \mathbb{P}[S_{t+1} = s' \mid S_{t=}s, A_t = a] \tag{1}$$

$\boldsymbol{r}$ is the reward function which is a function of $s, a$ and $s'$, and defined by;

$$r(s, a, s') = \mathbb{E}[r_{t+1} \mid S_{t=}s, A_t = a] \tag{2}$$

$\gamma$ is the discount factor where $\gamma \in [0, 1]$. At each time step $t$ agent in state $s$ takes an action $a$ according to the policy $\pi$ and land into $s'$ with probability $\tau$ where transition probability only depends on $s$, $a$ and $s'$ [7]. Policy is a sequence of decision rules, each of which tells the agent what action to take depending on the current state, not the history. Policies can be deterministic and stochastic where stochastic policy allows the agent to explore the environment. In RL, the goal of the agent is to find the best policy while minimizing the penalty from its experience which leads to the perplexing situation between exploration and exploitation. To learn the optimal policy, agent evaluates the given policy by estimating Value function of $\pi$. Value functions are two types; state value function $V^{\pi}(s)$ and the state-action value function $Q^{\pi}(s, a)$. State value function evaluates "how good the state is" and calculated as;

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s, \pi\right] \tag{3}$$

Similarly, the state-action value function assesses actions of the state and defined by;

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s, a, \pi\right] \tag{4}$$

The optimal state value function and state-action value function is the maximum return over all policies and formally described as;

$$V^{*}(s) = \max_{\pi} V^{\pi}(s) \tag{5}$$

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{6}$$

### 3.2 Q-learning based duty cycle interval Control

Reinforcement-learning learns the optimal policy while maximizing the expected cumulative reward to reach the goal following policy $\pi$. The aim of this work is to find the optimal state with maximum return, i.e., the duty cycle interval of all groups.

In this subsection, we will describe the duty cycle interval control algorithm using RL that satisfies the delay requirement; i.e., given delay success ratio (DSR) $\Gamma$ and delay bound $\theta$, detailed below.

State $s \in S$ is defined by duty cycle interval $T$ of all groups where $S = v^L$. Thus, the state $s = [T_1, T_2, \ldots . T_L]$ where $T_{min} \leq T \leq T_{max}. T_{min}$, $T_{max}$ and $v$ are the threshold, maximum duty cycle interval and the possible value of duty cycle interval $T$ a node can take respectively. $v$ is derived as $v = \frac{T_{max}}{I}$ where $I$ is the interval between each possible value of the duty cycle interval. The agent will get negative rewards if it reaches below the threshold.

An action $a$ is defined by increasing, decreasing duty cycle interval of one group at each time step and staying in the same state; Therefore, the action space $A$ is $(L \times 2 + 1)$.

In this work, we have used Q-learning algorithm to find the maximum duty cycle interval which is a model free RL [8]. Q-learning requires to visit each state infinitely often to converge [9, 10, 11], therefore large state space leads to the large number of iterations. For this reason, we have divided our proposed model into two phases based on the interval $I$ where $I = I_1, I_2$. We first design 1st phase as the baseline. Since the value of the baseline can be suboptimal, therefore the result of the baseline has been used in 2nd phase to get the optimal value. Action space $A$ has been used in both phases.

In 1st phase, model has large interval $I = I_1$ between possible value of the duty cycle interval. Thus, the state space is small. Reward function $r$ for this phase calculated as follows;

$$r(s,a,s') = \begin{cases} \frac{1}{R_a}, & if(z'(s') \geq \Gamma) \\ -c, & otherwise \end{cases} \tag{7}$$

where $c$ is a constant and $R_a$ is the ratio of the active period in each duty cycle interval and calculated as;

$$R_a = \sum_{j=1}^{L-1} \frac{a^t}{T_j} \tag{8}$$

$z'$ is the probability that packets arrive to the sink before delay bound and calculated as $z' = p\,(d_e < \theta)$ where $d_e$ is the end-to-end delay distribution [4]. E2E delay distribution $d_e$ is the sum of one-hop delay distribution $d_h$ and can be expressed as;

$$d_e = \sum_{j=2}^{L} d_h^j \tag{9}$$

In the 2nd phase, the model has the large state space since the interval between the possible value of $T$ is small $(I = I_2)$. Thus, the result of the baseline has been used as the initial state which achieves superior

results in a relatively low computation cost. Furthermore, we have used higher negative reward in this phase for faster learning. The reward function of 2nd phase is given below;

$$r(s,a,s') = \begin{cases} \frac{1}{R_a}, & if\ \left[(z'(s') \geq \Gamma)\ \&\ \varphi \geq \psi\right]\ \|\ \left[z'(s) < \Gamma\ \&\ z'(s') \geq \Gamma\ \&\ \varphi < \psi\right] \\ -c, & otherwise \end{cases} \quad (10)$$

where $\varphi$ and $\psi$ are the sum of duty cycle interval T of all groups in state s and $s'$. Hence, we can write;

$$\varphi = \sum_{k=1}^{L} T(s') \quad (11)$$

$$\psi = \sum_{k=1}^{L} T(s) \quad (12)$$

Policy $\varepsilon$-greedy has been used in this work to ensure that we explore the environment enough and exploit the known information. In $\varepsilon$-greedy policy; Agent takes random action with $\varepsilon$ probability and acts greedily with 1- $\varepsilon$ probability.

Q-learning is a kind of Temporal difference (TD) learning [12, 13] where an agent takes action and evaluate the action according to immediate rewards or penalty it receives from the state it has taken action [14]. Since the transition probability is unknown in our work, we can sample each transition $(s, a, s', r)$ as follows;

$$Q(s,\ a) = r(s,\ a,\ s') + \gamma \max_{a'} Q(s',\ a') \quad (13)$$

The Q-learning algorithm updates the state-action value function (for discounted return) using equation (14) given learning rate alpha and defined as follows;

$$Q\big(s,\ a\big) \leftarrow\ (1-\alpha)Q(s,a) + \alpha\big[r(s,a,s') + \gamma \max_{a'} Q\big(s',\ a'\big)\big] \quad (14)$$

where $\alpha$ = [0,1] which specified how much new state-action information will override the old information.

## 4. Result and Analysis

In this section, the result of the Duty Cycle Interval Control using Reinforcement Learning in Wireless Sensor Networks (DCI-RL) and the parameters have been presented. In our analytical model, we have considered the following parameters; soda network with 78 nodes in a rectangular area where the sink is in the center; total group number is 7; transmission range is 10 m with 2Mbps bandwidth; data packet is 46 bytes where the event rate is 0.5 packet/s. The default value of the DSR and delay bound is 95% and 30 s respectively. Data transmission period is 50 s. Two different interval value ($I_1 = 4, I_2 = 0.95$) has been used in 1st and 2nd phase respectively. In Q-learning, $\alpha = 0.1$ and $\gamma = 0.9$ is used. Threshold $T_{min}$ has been obtained from [4]. To validate that our proposed model can adjust with various requirements (i.e. finding the optimal duty cycle interval with different delay bound and DSR), we used three different delay bound [20,30,40] and DSR [90,95,97].

We have compared our result with two methods PMS and ESW of Duty Cycle scheduling considering Delay Time Constraints in Wireless Sensor Networks [4] under the effect of different delay bound and DSR to validate the result. Network performance has been proved with two performance metrics; total energy consumption and maximum energy consumption of a node where our proposed model DCI-RL satisfy the given DSR requirement for all different delay bound and DSR.

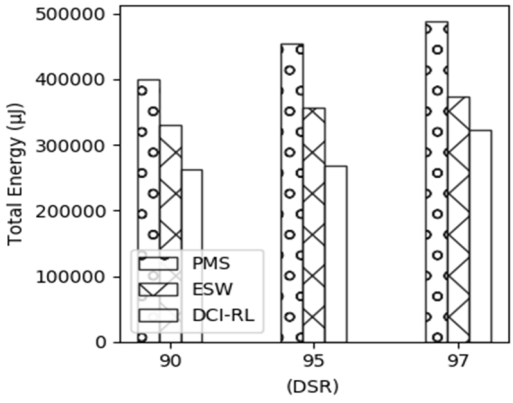


**Figure 2. (a): Total energy consumption**

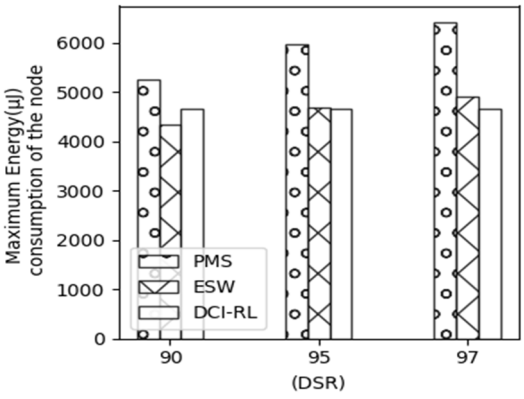


**Figure 2. (b): Maximum energy consumption of the node in Network**

Figure 2 shows the effect of different DSR on two performance metrics. Fig 2 (a) display the effects of the different DSR on total energy consumption of the network. We can see that in case of all three DSR; DCI-RL outperforms the other two Methods. Total energy consumption of DCI-RL for DSR 90% and 95% is less than 268,000µJ where ESW is between 330,000 µJ – 360,000 µJ and PMS is around 400,000µJ – 450,000 µJ respectively. When DSR is high, duty cycle interval decrease, thus the energy consumption of node increase. Hence, ESW and DCI-RL consume energy more than 300,000µJ at DSR 97 % while PMS consume around 500,000 µJ.

Fig 2(b) shows that ESW has the node that consumes the least energy for 90% DSR which is slightly more than 4000µJ while DCI-RL is less than 5000µJ. PMS consume higher energy than other two methods for 90% DSR. The node of DCI-RL consumes less energy than the other two methods in terms of 95% and 97% DSR whereas PMS method is maintaining an upward trend.

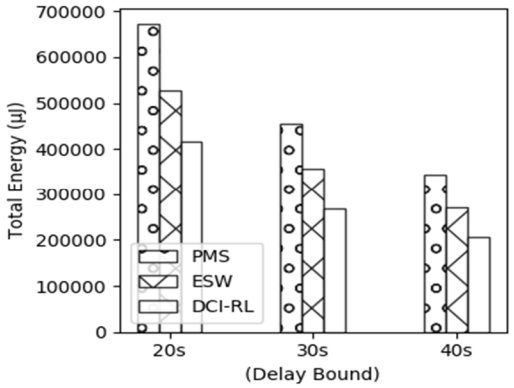


**Figure 3. (a): Total energy consumption**

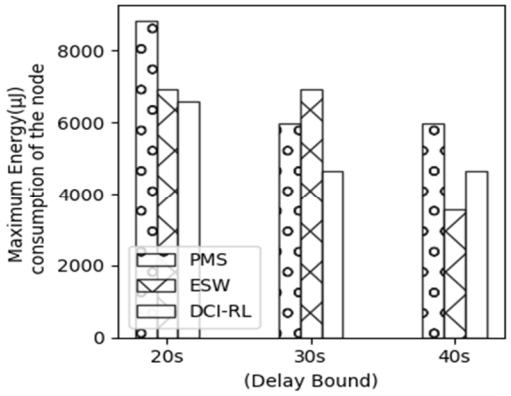


**Figure 3. (b): Maximum energy Consumption of the node in network.**

In Fig 3: performance metrics have been shown with different delay bounds. In Fig 3(a); DCI-RL has consumed the less energy around 200,000μJ at 40 s delay bound while PMS suppress the 400,000μJ in 30 s delay bound. PMS consumed around 700,000μJ energy when delay bound is 20 s. Energy consumption is higher at 20 s because the frequency of nodes waking up increases at lower delay bound.

In fig 3(b): when delay bound is 40 s, node in DCI-RL consume higher energy than the node in ESW while PMS consume higher energy compared to DCI-RL. In 30 s and 20 s delay bound node of ESW method and PMS method consumes higher energy than the DCI-RL.

## 5. Conclusion

In this paper, we formulated duty cycle interval learning problem for minimizing power consumption of Wireless Sensor Networks, which ensures the required probability of packets delivering to the sink and application delay requirements. We developed a noble Q-learning based duty cycle control for the multi-hop network, named DCI-RL. Our approach doesn't require sensors to implement the Q-learning algorithm which makes it suitable for resource constraint system such as WSNs. In order to meet the different demand of various applications, the proposed scheme enables the operator to adjust the duty cycle interval according to demand. It was shown using performance metrics that DCI-RL is more energy efficient compared to existing approaches and assures a longer network lifetime.

## Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF2016R1D1A3B03934617).

## References

[1] H. Wang, N. Agoulmine, M. Ma and Y. Jin, "Network lifetime optimization in wireless sensor networks", *IEEE Journal on Selected Areas in Communications*, Vol. 28, No. 7, pp. 1127 – 1137, 2010.
DOI: http://dx.doi.org/10.1109/JSAC.2010.100917

[2] T. Rault, A. Bouabdallah and Y. Challal, "Energy efficiency in wireless sensor networks: A top-down survey", *Computer Networks*, Vol.67, pp.104–122, 2014.
DOI: https://doi.org/10.1016/j.comnet.2014.03.027

[3] R. Alberola and D. psch," Duty Cycle Learning Algorithm (DCLA) for IEE 802.15.4 Beacon-enabled Wireless-Sensor Networks*", Journal of Ad hoc networks*, Vol.10, no-4, pp. 664-679, 2012.
DOI: https://doi.org/10.1016/j.adhoc.2011.06.006

[4] V. D. Son and S. Yoon, "Duty Cycle Scheduling considering Delay Time Constraints in Wireless Sensor Networks", *The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)*, Vol. 18, No. 2, pp. 169-176, Apr. 30, 2018.
DOI: https://doi.org/10.3390/electronics7110306

[5] T. N. Dao, S. Yoon, and J. Kim, "A deadline-aware scheduling and forwarding scheme in wireless sensor networks", *Sensors*, vol. 16, no. 1, 2016.
DOI: https://doi.org/10.3390/s16010059

[6] R. Sutton and A. Barto., "*Reinforcement Learning*", MIT Press., Cambridge, MA., 1998.

[7] D. White, "Real applications of Markov decision processes", *Interfaces*, Vol. 15, no. 6, pp. 73–83, 1985.
DOI: https://doi.org/10.1287/inte.15.6.73

[8] Watldns, C.J.C.H., *Learning from delayed rewards*, PhD Thesis, University of Cambridge, England, 1989.

[9] D. Bertsekas and J. Tsitsiklis., "*Neuro-Dynamic Programming*", Athena Scientific, Belmont, MA, 1996.

[10] J. Tsitsiklis., "Asynchronous stochastic approximation and Q-learning", *Machine Learning*, Vol. 16, pp. 185-202, 1994.
DOI: https://doi.org/10.1023/A:102268912504
[11] T. Jaakkola, M. Jordan, and S. Singh., "On the convergence of stochastic iterative dynamic programming algorithms", *Neural Computation*, Vol. 6, pp. 1185 – 1201, 1994.
DOI: https://doi.org/10.1162/neco.1994.6.6.1185
[12] C. Watkins and P. Dyan., "Q-learning", *Machine Learning*, Vol. 8, pp. 279–292, 1992.
DOI: https://doi.org/10.1007/BF00992698
[13] Sutton, R.S., *Temporal credit assignment in reinforcement learning,* PhD Thesis, University of Massachusetts, Amherst, MA, 1984
[14] R. Sutton, "Learning to predict by the methods of temporal difference", *Machine Learning*, Vol.3, pp. 9-44, 1988.
DOI: https://doi.org/10.1007/BF00115009