EURASIP Journal on Wireless Communications and Networking

RESEARCH

Open Access

Learning nodes: machine learning-based energy and data management strategy

Yunmin Kim and Tae-Jin Lee*

*Correspondence:
tjlee@skku.edu
College of Information and Communication Engineering, Sungkyunkwan University, 2066 Seobu-Ro, Jangan-Gu, Suwon, Gyeonggi-Do 16419, Republic of Korea

Abstract

The efficient use of resources in wireless communications has always been a major issue. In the Internet of Things (IoT), the energy resource becomes more critical. The transmission policy with the aid of a coordinator is not a viable solution in an IoT network, since a node should report its state to the coordinator for scheduling and it causes serious signaling overhead. Machine learning algorithms can provide the optimal distributed transmission mechanism with little overhead. A node can learn by itself by utilizing the machine learning algorithm and make the optimal transmission decision on its own. In this paper, we propose a novel learning Medium Access Control (MAC) protocol with learning nodes. Nodes learn the optimal transmission policy, i.e., minimizing the data and energy queue levels, using the Q-learning algorithm. The performance evaluation shows that the proposed scheme enhances the queue states and throughput.

Keywords: Energy-harvesting, Transmission policy, Q-learning, IoT

1 Introduction

With the advent of the Internet of Things (IoT), wireless communication function is employed not only in the electronic devices but also in every 'Things' [1]. These devices are expected to have low-cost and low-power consumption characteristics to operate in IoT networks [2]. Also, since tons of devices are expected to be deployed in IoT networks, energy should be provided in a sustainable way to maintain long-lasting networks [3]. In this sense, energy will play an important role to provide seamless services with limited resource.

One of the viable solutions is to produce energy by devices for themselves or provide energy to devices wirelessly, i.e., energy-harvesting, which enables devices to obtain energy from various physical phenomena such as wind, sun-light, and Radio Frequency (RF) signal [4]. In a network with energy-harvesting devices, the status of data and energy in the devices may vary [5, 6]. Devices may have different amount of traffic to transmit. Some devices frequently report the status or send information to the network, while other devices have relatively sparse traffic. The energy sustainability of devices also fluctuates. If a device is located near the power beacon, it receives much energy with minimal loss. However, the nodes far from the power beacon obtain small amount of energy. Without an appropriate data transmission strategy, the devices may suffer from shortage of energy to transmit or unnecessary charging, and accumulated data queue and packet loss. Therefore, depending on the status of devices, a desired transmission strategy is required.

Transmission policies for energy-harvesting devices have been researched. In [7, 8], the optimal packet scheduling policy for a single energy harvesting node is studied. The transmission power of a node related with the data rate is optimized to minimize the total transmission time of a node. The authors in [9, 10] present a transmission policy for point-to-point transmission in the fading channel. By controlling the time sequence, throughput is maximized and the total transmission time is minimized. In [11], a decentralized random access policy is studied to maximize the long-term network utility. Using the game theory, nodes decide the policy to transmit, remain idle, or discard packets. The optimal new solution is found and the heuristic algorithm is provided. The authors in [12] studied the power management policies for the dual energy harvesting links, where transmitter and receiver are both energy harvesting-capable nodes. Considering the battery size and the retransmission index, the packet drop probability (PDP) is modeled. The battery size highly is shown to impact on the PDP performance as it helps to overcome the randomness of energy availability. Also, the optimal retransmission policy to minimize PDP is designed. In [13], the selective sampling, which decodes the packet with a certain length, is proposed to reduce energy consumption. The selective sampling information is further utilized by piggybacking for more efficient energy use at the receiver. Also, the retransmission strategy and the power allocation scheme to ensure lower PDP are introduced using Markov Decision Process (MDP).

Recently, machine learning has drawn much interest as a powerful tool to solve complex problems, e.g., Google's AlphaGo [14]. This adaptive learning capability can be applied to tackle complex problems. The transmission strategy for energy harvesting nodes by machine learning is an attractive research issue. In [15], adaptation of duty cycle for energy harvesting sensor nodes is studied. To achieve the balance between the energy supply and the Quality of Service (QoS) requirement, a modified MDP using reinforcement learning is introduced.

Reinforcement learning-based energy management policies for single [16] and multiple [17] nodes are studied. The energy harvesting node is modeled as continuously to create data and to gain energy from the energy source. Data can be transmitted using a certain amount of energy defined in a conversion function. For a single node, the authors in [16] utilize Q-learning to find the optimal policy for a general conversion function. An extra energy source node providing energy to multiple nodes is considered [17]. To minimize the average delay of transmitting nodes, an efficient energy sharing method is presented using the Q-learning algorithm. In [18], Q-learning-based Medium Access Control (MAC) protocol for underwater sensor networks is studied. Without extra message exchange, a node learns to optimize back-off slots to reduce collision through trial-and-error. By intelligently selects a back-off slot through Q-learning, low-signaling overhead and low complexity can be obtained. Also, the authors design the reward function updates from messages, especially to consider the level of collision from Negative Acknowledgment (NACK). The authors in [19] proposed a machine learning-enabled MAC framework for IoT nodes coexisting with WiFi users. During the rendezvous phase, an intelligent gateway learns the type and expected amount of devices, i.e., WiFi and IoT nodes by monitoring the three-way handshake. Then, the gateway schedules frequency channels to IoT and WiFi devices based on the learning result. In the transmission phase, IoT devices and WiFi users contend for data transmission. The gateway can dynamically adjust the superframe length to achieve enhanced throughput.

In this paper, we propose a new learning MAC protocol to learn a differentiated transmission strategy method of a node in a network. We focus on the imbalance nature between the energy and the data in a node, which stems from the randomness of arrival rates of energy and data. We propose a MAC protocol with learning mechanism to mitigate the imbalance problem. The contributions of our work are:

- The 'imbalance' problem of energy and data management for energy-harvesting nodes in IoT networks is revealed.
- Based on the nature of nodes, we classify them into energy-dominant and data-dominant nodes, and, for each type of node, the multi-slot and high-rate transmission strategies are proposed to mitigate the imbalance problem.
- We utilize Q-learning to automatically determine better choice when using multi-slot and high-rate schemes. Nodes learn their best actions given the energy and data availability.
- Performance evaluation shows the learning behavior for stable queue states, and overall improved throughput.

We consider a more realistic environment in which the nodes have various data and energy profiles. Also, different transmission strategies, i.e., multi-slot and high-rate transmission, are proposed. Each node learns and selects a different transmission mechanism based on their evolutions of data and energy queue states. We utilize a Q-learning algorithm for individual nodes to learn the optimal parameters of the proposed learning MAC. As time evolves, a node learns the optimal transmission strategy, which can minimize the data and energy queue levels, by itself so that the nodes in a network harmoniously transmit data while boosting energy efficiency.

2 Proposed learning MAC protocol

We consider a network of devices, in which devices report the collected information to the sink node. The nodes are capable of producing electricity by energy-harvesting. Harvested energy can be stored in the battery of a node and used to transmit data to the sink node. If the energy is not sufficient to transmit a packet of data, it is not transmitted and remains in the data queue. If there is no data to transmit, the harvested energy is stored in the battery until the next data transmission. Dynamic data traffic and energy states of nodes may create unbalanced use of energy and data. Thus, an optimal and balanced transmission strategy of nodes is required to minimize the data and energy queue levels. To react to the status of a node considering both energy and data, we define E-node and D-node for which different transmission strategies are employed.

2.1 Energy dominant and data dominant nodes

In an IoT network, different types of nodes coexist depending on their own jobs. Nodes can have different tasks to do and different performance of energy-harvesting. So, the energy and data packet arrival rates can vary to each of the nodes.

The arrival rates of energy can be higher than those of data in the nodes. They tend to have a small number of jobs compared to the amount of energy. We refer this kind of nodes as Energy dominant nodes (E-node). The nodes may locate near the power beacon or sparsely transmit data. E-nodes are likely to have sufficient amount of energy. However, some of the stored energy will not be used properly and wasted. To mitigate the energy waste, E-node is required to have a proper transmission scheme.

On the other hand, some nodes may have heavier data arrival rates than the energy generation rates. The nodes may be placed far from the wireless power source or shaded by obstacles and hardly gets the sufficient energy. We refer this kind of nodes as Data dominant nodes (D-node). These nodes suffer from the shortage of energy when they try to transmit data. Then, the nodes tend to wait until sufficient energy arrives and the length of data queue may be increased. To resolve the energy shortage, an appropriate transmission policy to reduce energy consumption needs to be applied.

2.2 Multi-slot method for energy dominant nodes

We define E-node as the energy dominant node which has a larger arrival rate of energy than that of data traffic. The E-nodes are likely to have excessive energy and they tend to wait for the arrival of data traffic. With the conventional MAC, e.g., Frame Slotted ALOHA (FSA), the unused energy may keep accumulating in the energy queue. So, in order to increase the energy utilization, instead of storing excessive energy, the node may need to increase the energy consumption by transmitting more data.

To use surplus energy efficiently, we propose to use multi-slot transmission. In the multi-slot contention mechanism, the node can select multiple slots for transmitting data. Then, a node attempts to transmit data in selected time slots until the successful data transmission. If data transmission succeeds in the middle of the selected time slots, the node quits the procedure. Figure 1 shows an example of the multi-slot transmission of a certain E-node. At first, an E-node selects 4 time slots according to the multi-slot transmission scheme. In the first and the second selected slots the E-node fails to send data due to collisions with other nodes. Data is sent at the third selected time slot and the E-node ends the data transmission and does not operate in the fourth selected time slot. Then, the node consumes 3 energy units to transmit one data packet. Since the E-node consumes more energy for the transmitted data unit, the imbalance difference between energy units and data units at the beginning of the E-node is mitigated from 3 to 2 (Fig. 1). However, the number of multiple slots to be selected should be carefully determined since the degree of balance between data and energy among the nodes is affected by that and a congestion problem may arise by the heavy multi-slot transmission mechanism.

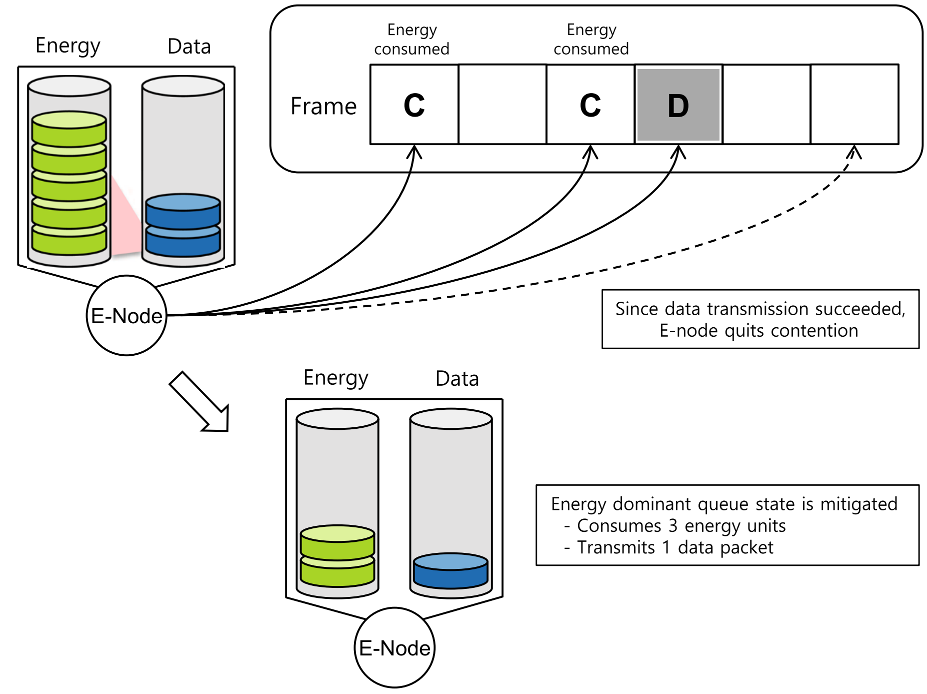
Energy
Data
Frame
Energy consumed
Energy consumed
C
C
D
E-Node
Since data transmission succeeded, E-node quits contention
Energy
Data
Energy dominant queue state is mitigated
- Consumes 3 energy units
- Transmits 1 data packet
E-Node

Fig. 1 Multi-slot operation and the effect of the E-node

2.3 High-rate method for data dominant nodes

The D-node is defined as the data dominant node which generates more data traffic than energy traffic. For the D-node, data packets tend to accumulate in the data queue due to insufficient energy to use. So, D-nodes are likely to suffer from the energy shortage and arriving packets can wait long in the queue. In this case, it is desirable to transmit data packets with less energy.

To decrease the energy consumption, we propose high-rate transmission. High-rate transmission can be done by shortening the operation time to reduce the energy consumption. The amount of energy consumed in the active state is known to be significant. In high-rate transmission, the node boosts the data rate by changing the modulation scheme and sacrifices the reliability of transmission. Then, the node can pump out more data packets in a slot and the transmission time of a single data packet can be reduced. If a node succeeds in the random access, it attempts to transmit as many data packets as possible in the data queue. Figure 2 shows an example of the high-rate transmission of a certain D-node. In the first frame, a D-node succeeds and the 3 data packets can be transmitted in a time slot. Then, the D-node unburden 3 data packets using one energy unit. The imbalance nature of the D-node is resolved using the high-rate transmission. However, in the second frame, the D-node fails to transmit data due to a collision. As collision happens, the D-node quits the transmission in a frame. Still, the D-node only consumes one third of the energy unit as it uses 3 times higher data rate. However, the data rate of a node should be carefully selected since the reliability of packet transmission decreases as the rate increases. Thus, appropriate selection of rate is required.

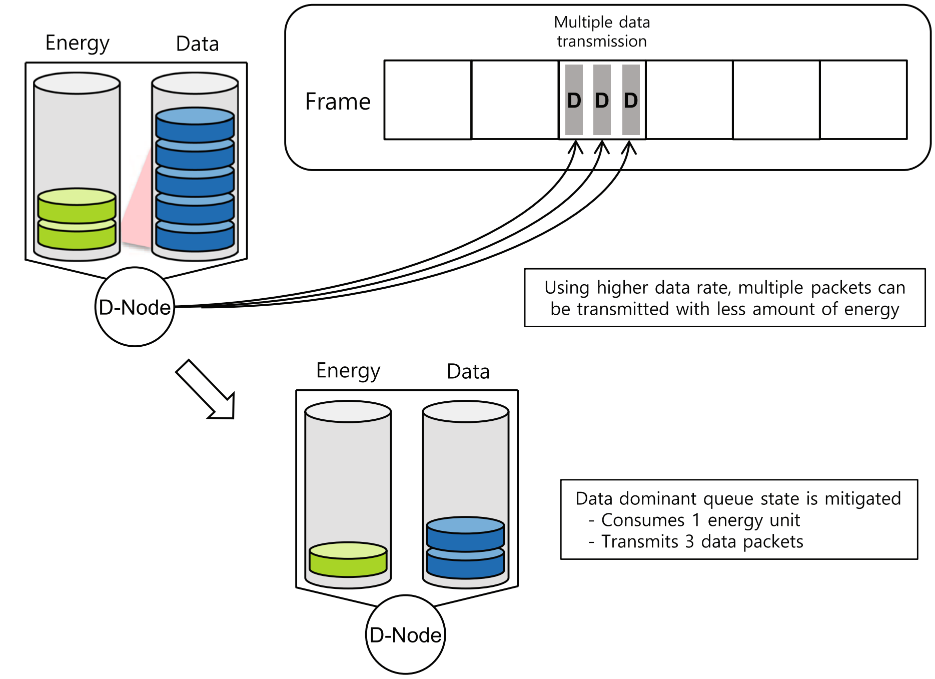
Energy
Data
Multiple data transmission
Frame
D D D
D-Node
Using higher data rate, multiple packets can be transmitted with less amount of energy
Energy
Data
Data dominant queue state is mitigated
- Consumes 1 energy unit
- Transmits 3 data packets
D-Node

Fig. 2 High-rate operation and the effect of the D-node

2.4 Channel model

An IoT node utilizes Phase Shift Keying (PSK) for transmitting data to the sink node. We consider the Additive White Gaussian Noise (AWGN) channel during the data transmission. The noise is modeled as a white Gaussian random process with zero mean and Power Spectral Density (PSD) $N_0/2$. Due to channel error, transmitted data from a node might be lost at the sink node. We assume that the data transmission fails when IoT nodes collide in the same time slot due to high level of interference. So, only the noise can be reasonably considered to distinguish the successful transmission in the PHY layer. The Signal-to-Noise Ratio (SNR) can be defined

$$\mathrm{SNR} = \frac{P_\mathrm{r}}{N_0 B}, \tag{1}$$

where P_r and B are the received power and the bandwidth. Total noise power within the bandwidth 2B is $N = N_0/2 \cdot 2B = N_0 B$. In terms of energy per bit (E_b) or energy per symbol (E_s), the SNR can be

$$\mathrm{SNR} = \frac{E_\mathrm{b}}{N_0 B T_\mathrm{b}} = \frac{E_\mathrm{s}}{N_0 B T_\mathrm{s}}, \tag{2}$$

where T_b and T are the bit time and symbol time.

We use a symbol error rate to determine the successful transmission of an IoT node. For an M-ary PSK, the symbol error rate can be modeled as [20]

$$p_e \approx 2Q\left(\sqrt{(2E_\mathrm{s}/N_0)\sin(\pi/2M)}\right), \tag{3}$$

where $Q(\cdot)$ is the Q-function. If a node wins the MAC layer contention and the transmission is successful, the node gains the reward and it is reflected to the corresponding Q-value matrix.

2.5 Proposed learning MAC

The proposed learning MAC utilizes both the multi-slot and high-rate schemes. In the learning MAC, nodes operate based on the frame broadcast by the sink node. Nodes select one of the method by comparing the energy and data queues. If the amount of energy is larger than the number of data packets in the queue, a node chooses multi-slot scheme. Otherwise, a node selects the high-rate scheme for data transmission. Then, nodes are required to determine the parameters used in the multi-slot and the high-rate scheme. In the multi-slot scheme, nodes need to decide the number of slots to be selected. On the other hand, in the high-rate scheme, the factor for boosting data rate is needed. The detailed process of parameter settings is described in Sect. 4. Using the methods and the parameters, nodes perform contention in the current frame. At the end of the frame, the energy and data queue states will be updated by the contention. In the next frame, the node operates based on the updated queue states.

Figure 1 shows an example of the proposed learning MAC protocol with multi-slot and high-rate transmissions. The nodes 1, 2, and 5 are the E-nodes while the nodes 3, 4, and 6 are the D-nodes. The E-nodes need to select multiple slots to utilize the surplus energy, while the D-nodes need to apply the high-rate strategy to reduce the transmission time of a node. The node 1 selects slots 1 and 7. Since the node 1 succeeds in slot 1, it does not transmit in slot 7. Collision between node 2 and node 5 occurs in slot 5. Node 5 recovers from the collision in the second transmission attempt in slot 6. Nodes 3, 4, and 6 terminate the successful high-rate data transfer to transmit more bits (three times) in slots 3, 4, and 7 to save the energy. As mentioned in the strategy, a node should select the transmission policy in a wise way. Considering dynamic arrival rates of energy and data of a node, it is desirable to track the optimal MAC parameters, i.e., number of multiple slots and transmission rate in a slot.

3 Learning MAC with Q-learning

The proposed learning MAC is affected by the parameters selected by the node. To do this, a central coordinator may collect the nodes' status and make decision. However, the process of collecting and making decision may cause the signaling overhead. Also, the status of nodes varies over contention process instantaneously. So, the centralized coordination scheme may not be a viable solution. As a solution, we apply Q-learning to each node to find its best strategy. Q-learning is a method to estimate the available actions by scoring based on the result caused by the actions. Based on its own previous choices, nodes find the optimal action. Using Q-learning, nodes can learn its current best parameters with the minimal interaction.

3.1 Q-learning

Q-learning is one of the reinforcement learning techniques that can be used to find the optimal action using the reward by learning. The agent, the learner, utilizes Q-learning to learn the optimal policy by interacting with its environment. Let S be the possible states and A be the possible actions of the agent. The learner senses its state $s_t \in S$ and chooses an action $a_t \in A$ based on its state at time t. After the action taken, the agent moves to the new states s_{t+1} with the probability of $P_{s_t,s_{t+1}}$. Then, the learner receives its reward $r(s, a)$. The objective of the learner is to find the optimal policy $\pi^*(s)$, which maximizes the cumulative reward $r_t = r(s_\mathrm{r}, a)$ over time. In the considered network and problem setup, the optimal criterion is to minimize the data and energy queue levels of an IoT node.

The total discounted return over an infinite time is

$$V^\pi(s) = E\left\{\sum_t^\infty \gamma^t r(s_t, \pi(s_t))|s_0 = s\right\}, \tag{4}$$

where γ is the discount factor from 0 to 1. The value function $V^\pi(s)$ can be further expressed as [21]

$$V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} P_{s,s'}(\pi(s)) V^\pi(s'), \tag{5}$$

where $R(s, \pi(s))$ is the expectation of $r(s, a)$ and $P_{s,s'}$ is the transition probability from state s to s'. Applying the Bellman optimality criterion [22], which shows the existence of at least one optimal strategy, the value function is

$$V^*(s) = V^{\pi^*}(s) = \max_{a \in A}\left\{R(s, a) + \gamma \sum_{s' \in S} P_{s,s'}(a) V^*(s')\right\}. \tag{6}$$

For a policy π, action a is taken at state s. Then, the expected return value, Q-value, is

$$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{s,s'}(a) V^\pi(s'). \tag{7}$$

When the optimal policy π^* is applied, the Q-value can be defined as

$$Q^*(s, a) = Q^{\pi^*}(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{s,s'}(a) V^{\pi^*}(s'). \tag{8}$$

Plugging Eq. (5) into Eq. (3), we get

$$V^*(s) = \max_{a \in A}[Q^*(s, a)]. \tag{9}$$

Therefore, the optimal value function can be obtained from the maximized $Q^*(s, a)$. Using the result of Eq. (6), the Q-value can be expressed as

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s' \in S}\left\{P_{s,s'}(a)[\max_{a' \in A} Q^*(s', a')]\right\}. \tag{10}$$

Let the learner performs action a_t in state s_t at time t. Then, the state changes to s_{t+1} and the learner returns the immediate reward r_{t+1}. In the Q-learning process, the optimal action can be found iteratively. So, the learner updates the Q-values as follows.

$$Q(s_t, a_t) \longleftarrow Q(s_t, a_t) + \alpha \left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right) \tag{11}$$

where α and r_{t+1} are the learning rate and the reward observed after performing a_t in s_t. The learning rate affects the update rate while the discount factor controls the importance of the future values. It is known that the Q-value $Q(s_t, a_t)$ converges to the optimal value $Q^*(s, a)$ as the pairs of state-action are performed.

3.2 Network model for Q-learning

We consider a networks of N nodes. An energy-harvesting node is equipped with the energy queue and the data queue. Each queue stores the harvested energy and the data. The energy is assumed to be quantized to be buffered in the energy queue. For a basic data rate, a node requires unit energy to transmit a data packet.

3.2.1 State

Let the state be the difference between the state of energy and that of data in the queues. High value of the difference denotes the severe unbalance between energy and data. At the beginning of each frame, nodes set their states as follows.

$$x_{E,t} = [E_{q,t} - D_{q,t}], \tag{12}$$

$$x_{D,t} = [D_{q,t} - E_{q,t}], \tag{13}$$

where $E_{q,t}$, $D_{q,t}$, and $[\cdot]$ refer to the energy queue, data queue states at the frame time t, and the nearest integer function. The state ranges from 0 to M, where M is the maximum capacity of each of the queues.

3.2.2 Action

Depending on the state, a node can select an appropriate action. Since the E-node operates by multi-slot transmission, the number of slots to be selected becomes the available actions. The available actions of an E-node can be written as

$$a_{E,t} = l, \quad l \in \{0, 1, 2, ..., l_{\max}\} \tag{14}$$

where l and $l_{\max}$ are the number of slots and the maximum number of slots to be selected. The D-node uses high-rate transmission. Then, the available actions of the D-node can be defined as the number of bits in a symbol to be used for transmission. The available actions of a D-node is defined as

$$a_{D,t} = g, \quad g \in \{0, 1, 2, ..., g_{\max}\} \tag{15}$$

where g and $g_{\max}$ are the number of bits in a symbol and the maximum number of bits in a symbol to be used in the high-rate transmission. The transmission rate is g multiples of the unit rate.

3.2.3 Reward

We define the reward as the change of difference after a contention. If the difference decreases after a contention, it can be concluded that the node is moving toward an appropriate direction. The reward induced from action a at the state x_t is

$$\begin{aligned} r_{E,t+1} &= r_1\left(x_{E,t} - x_{E,t+1}\right) - r_2(D_{q,t+1} + E_{q,t+1}), \\ r_{D,t+1} &= r_1\left(x_{D,t} - x_{D,t+1}\right) - r_2(D_{q,t+1} + E_{q,t+1}), \end{aligned} \tag{16}$$

where r_1 and r_2 indicate the weights of the first and the second term. The first term accounts for the change of the difference between energy and data. Decreasing the unbalance between energy and data contributes to the positive reward. The second term is for the overall queue states of a node. As the queues build up, the reward decreases.

3.2.4 Next state

After the contention in a frame, the queue states of nodes may change. If a node succeeds in contention, both energy and data queue states decrease. In the multi-slot policy, the amount of energy consumption depends on the number of transmission attempts. For high-rate nodes, the energy consumption varies according to the selected number of bits in a symbol.

3.3 Proposed Q-learning mechanism

The nodes construct the state-action matrix to manage the Q-values. The rows of the matrix indicate the current states and the columns of the matrix indicate the actions. If a node is an E-node, it uses a multi-slot Q-value matrix Q_E and if a node is a D-node, it uses a high-rate Q-value matrix Q_D. Using the state-action matrix, the nodes retrieve the Q-values.

$$Q_E = \begin{pmatrix} u_{11} & \dots & u_{1l_{\max}} \\ \vdots & \ddots & \vdots \\ u_{M1} & \dots & u_{Ml_{\max}} \end{pmatrix}, \; Q_D = \begin{pmatrix} v_{11} & \dots & v_{1g_{\max}} \\ \vdots & \ddots & \vdots \\ v_{M1} & \dots & v_{Mg_{\max}} \end{pmatrix}, \tag{17}$$

where u_{ij} and v_{ij} denote the expected Q-value when the action is taken to change state from i to j. For example, when a node tries to access the channel, it first checks the difference between the energy level and the data queue length. If a node has more energy than data, it become an E-node and utilizes Q_E matrix. Otherwise, a node is treated as a D-node and it uses Q_D matrix. Then, the optimal action (MAC parameter assignment) in a frame is determined by

$$a_{E,t} = \min\left(\underset{l \in \{0,\dots,l_{\max}\}}{\operatorname{argmax}} \{Q_E(x_{E,t}, l)\}, \lfloor E_{q,t} \rfloor \right), \tag{18}$$

$$a_{D,t} = \begin{cases} \underset{g \in \{0,\dots,g_{\max}\}}{\operatorname{argmax}} \{Q_D(x_{D,t}, g)\}, & E_{q,t} \geqq 1, \\ 0, & E_{q,t} < 1. \end{cases} \tag{19}$$

where $\lfloor \cdot \rfloor$ is the floor function. Since a node can transmit within the current energy state ($E_{q,t}$), the number of slots to be selected is limited by the current energy level.

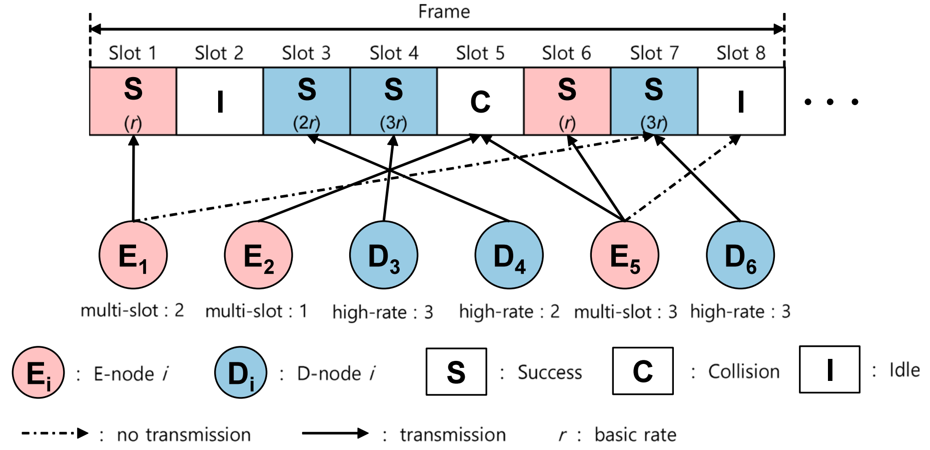

Fig. 3 Operation of the proposed learning MAC protocol

Table 1 Simulation parameters

Parameter	Value
Number of nodes (N)	20–80
Duration of a time slot	1 ms
Maximum queue capacity (M)	50
Maximum number of multi-slots ($l_{\max}$)	5 slots
Maximum number of bits in a symbol ($g_{\max}$)	5
Data arrival rate	0.5–3 packets/frame
Energy arrival rate	0.5–3 units/frame
Learning rate (α)	0.7
Discount factor (γ)	0.1
Weights (r_1, r_2)	10, 1
Signal-to-noise ratio	20 dB

With the achieved MAC parameter, the nodes perform contention with the selected mode. After each frame, the energy and data queue states of the nodes change by the learning actions. Based on the changed queue states, the nodes update the state-action matrix as follows.

$$\begin{aligned} Q_E(x_{E,t+1},k) = Q_E(x_{E,t},k) + \alpha \Big(r_{E,t+1} \\ +\gamma \max_{l} Q_E(x_{E,t+1},l) - Q_E(x_{E,t},k) \Big), \end{aligned} \tag{20}$$

$$\begin{aligned} Q_D(x_{D,t+1},k) = Q_D(x_{D,t},k) + \alpha \Big(r_{D,t+1} \\ +\gamma \max_{g} Q_D(x_{D,t+1},g) - Q_D(x_{D,t},k) \Big). \end{aligned} \tag{21}$$

The nodes then learn the optimal action at each state by updating the state-action matrix. After a certain learning time, nodes can choose the best action and the data and energy queue states are expected to be balanced (Fig. 3).

4 Results and discussion

We conduct simulations to verify the performance enhancement of our transmission strategy for learning nodes. In simulations, nodes have random arrival rates of data and energy in a predefined range (see Table 1). At the beginning of a frame, a node determines the action based on the Q-matrix and performs channel access. During the contention in a frame, newly arrived data and energy are put into the queues. Then, the Q-matrix is updated according to the queue status of the current and the previous frame. The conventional FSA protocol is chosen as a comparative scheme. In FSA, nodes perform channel access if at least one energy and data unit exist. Otherwise, they operate in the sleep mode. Simulations are conducted for 10,000 frames. The parameters used in the simulation are shown in Table 1.

Figure 4 shows the change of the difference between the energy queue and the data queue for Q-learning iterations (frames). Figure 4a indicates the node with similar energy arrival rate and data arrival rate. As the Q-learning mechanism evolves, the difference between the queues fluctuates over time. Since nodes perform contentions,

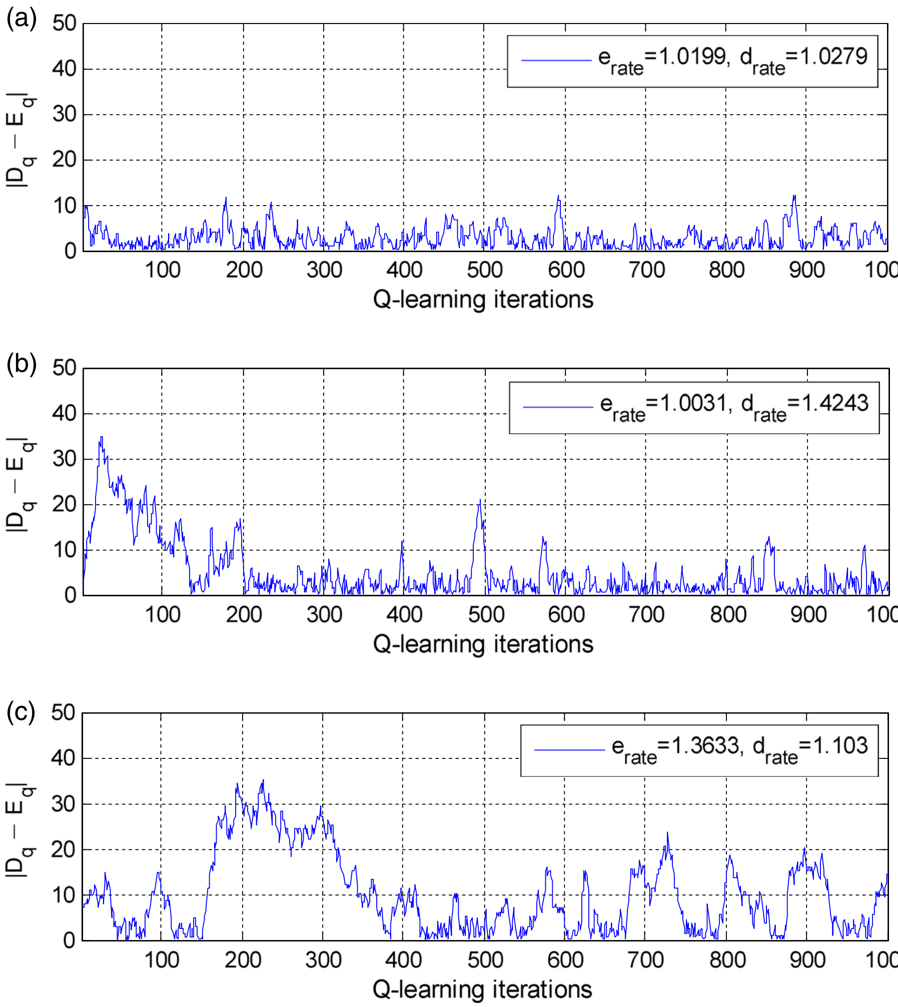

Fig. 4 Change of the difference between the energy and data of queues for evolving Q-learning iterations

the difference may suddenly rise up sometimes due to the randomness. However, the node recovers below a certain level by utilizing the proposed method. Figure 4b refers to the D-node, in which the arrival rate of energy is bigger than the arrival rate of data. The difference is alleviated after 200 Q-learning iterations. After that, the difference of the queue states, becomes under control. By transmitting more data with reduced amount of energy, the imbalance problem is mitigated. Finally, Fig. 4c stands for the E-node, with larger data arrival rate than the energy arrival rate. At first, the difference starts with the small amount of imbalance. When the difference becomes larger, the degree of imbalance is mitigated after 250 Q-learning iterations (from 150 to 400). By inducing the consumption of excess energy to transmit data, multi-slot schemes resolve the imbalance problem. The fluctuation of difference swings more than that of the case (b). As the multi-slot method largely depends on the contention result, the effect of resolving imbalance is weaker than that in the high-rate method.

Figure 5 shows the energy and data queue states of the individual nodes. The data and energy arrival rates are randomly chosen from 1 to 1.5. In the FSA, the frame size is the same as the number of nodes ($N = 20$) while the frame size is set to 4 times to the number of nodes in the proposed scheme due to the multi-slot operation. For the proposed learning MAC, the queue states of the nodes are distributed along the diagonal line. With the learning capability, the nodes take actions to balance the energy and data. Depending on the arrival rates of energy and data, the balancing point is appropriately determined. The proposed scheme mitigates the imbalance and the queue build-up problem. However, in the FSA scheme, queue states are shown to be deployed in the upper right corner area. Since the nodes suffer from the imbalance problem, the data packets and energy are unnecessarily built-up in the queues. At first, the data queue states are higher than the energy queue states. Energy can be consumed by successful transmissions, but that is not the case due to collisions and energy starts to accumulate. Then both energy and data queues become almost full.

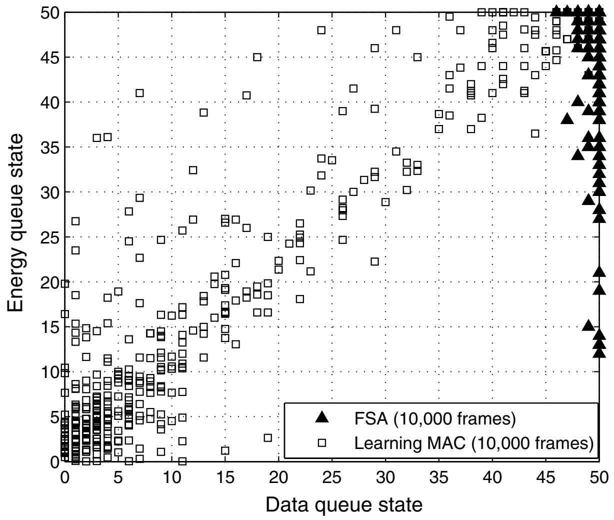

Fig. 5 Energy and data of queue states of the individual nodes ($N = 20$)

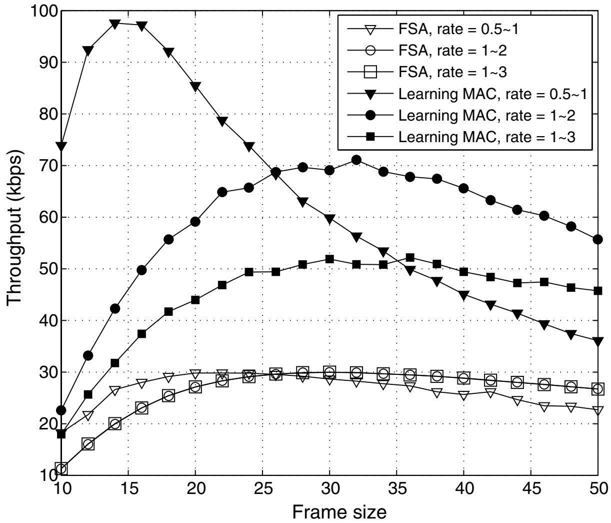

Fig. 6 The performance of FSA and the proposed learning MAC ($N = 30$) for different arrival rates

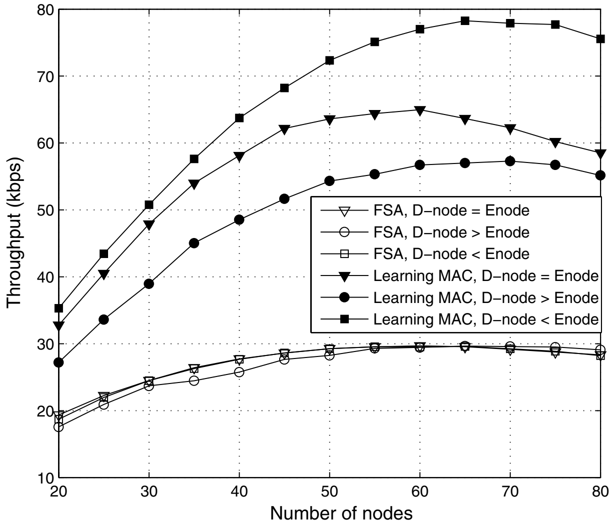

Fig. 7 The performance of FSA and the proposed learning MAC ($N = 30$) for D-node/E-node dominant and balanced cases

Figure 6 shows the saturation throughput for varying frame size. It is shown that the throughput of proposed learning MAC outperforms than that of the FSA. Since the multi-slot transmission improves the success probability in the contention, the number of packets transmitted in a certain time increases. Also, the high-rate transmission contributes to the throughput improvement by saving the transmission time of the nodes. In the proposed learning MAC, the optimal frame size changes for different data and energy rate conditions. As the data arrival rate increases, the use of multi-slot is needed and the optimal frame size increases.

Figure 7 indicates the throughput performance for varying numbers of nodes. In this simulation the frame size is set to 60. To implement different percentage of D-nodes and

E-nodes, we generate different energy and data arrival rates. For example, in the D-node dominant case we set the higher data arrival rate. As the number of nodes increases, the throughputs improve since the proposed MAC manages various imbalances. Also, the E-node dominant case shows the better performance than the other cases. Since nodes are expected to use multi-slot scheme, the transmitted data increases. When D-nodes are dominant, nodes are likely to utilize high-rate method. Then, the channel error rate might be increased since the transmission is performed with the reduced amount of bit energy.

5 Conclusion

We have proposed a new learning MAC protocol for energy-harvesting nodes to resolve the imbalance between energy and data. For E-nodes, multi-slot policy is used to enhance the success probability and energy efficiency. High-rate policy is utilized for D-nodes to decrease the energy consumption. The optimal MAC parameters depending on the data and energy queue states are automatically learned by the nodes using the Q-learning mechanism. Thus nodes learn the optimal actions for every queue states. The performance evaluation shows that our new learning MAC protocol with learning nodes outperforms in terms of the queue sizes and the network throughput. Nodes are shown to appropriately flush out data and energy in the queues and to achieve better throughput and lower packet drop rate.

Abbreviations

IoT: Internet of Things; RF: radio frequency; PDP: packet drop probability; MDP: Markov decision process; QoS: quality of service; MAC: medium access control; NACK: negative acknowledgment; E-node: energy dominant node; D-node: data dominant node; FSA: framed slotted ALOHA; PSK: phase shift keying; AWGN: additive white Gaussian noise; PSD: power spectral density; SNR: signal-to-noise ratio.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) through the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF- 2021R1A2B5B01002661.

Authors' contributions

YK's contribution is to write the paper and conduct performance analysis and simulations. T-JL's contribution is to write and revise the paper, and to guide the direction and organization of the paper. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 27 March 2018 Accepted: 28 August 2021
Published online: 15 September 2021

References

1. L. Atzori, A. Iera, G. Moralbito, Internet of Things: a survey. Comput. Netw. **54**(1), 2787–2805 (2012)
2. J. Gubbi, R. Buyya, S. Marusic, M. Palaniswamia, Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
3. P. Kamalinejad, C. Mahapatra, Z. Sheng, S. Mirabbasi, V.C.M. Leung, Y.L. Guan, Wireless energy harvesting for the Internet of Things. IEEE Commun. Mag. **53**(6), 102–108 (2015)
4. S. Sudevalayam, P. Kulkarni, Energy harvesting sensor nodes: survey and implications. IEEE Commun. Surv. Tutor. **13**(3), 443–461 (2011)
5. A. Biason, M. Zorzi, Transmission policies for an energy harvesting device with a data queue. In *Proceedings of International Conference on Computing, Networking and Communications (ICNC)* (2015), pp. 189–195

6. D. Liu, J. Lin, J. Wang, X. Chen, Y. Chen, Dynamic power allocation for a hybrid energy harvesting transmitter with multiuser in fading channels. In *Proceedings of IEEE Vehicular Technology Conference (VTC)* (2016), pp. 1–5
7. J. Yang, S. Ulukus, Optimal packet scheduling in an energy harvesting communication system. IEEE Trans. Commun. **60**(1), 220–230 (2012)
8. D.D. Testa, N. Michelusi, M. Zorzi, Optimal transmission policies for two-user energy harvesting device networks with limited state-of-charge knowledge. IEEE Trans. Wirel. Commun. **15**(2), 1393–1405 (2016)
9. O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, A. Yener, Transmission with energy harvesting nodes in fading wireless channels: optimal policies. IEEE J. Sel. Areas Commun. **29**(8), 1732–1743 (2011)
10. Q. Bai, J.A. Nossek, Joint optimization of transmission and reception policies for energy harvesting nodes. In *Proceedings of International Symposium of Wireless Communication Systems* (2015)
11. N. Michelus, M. Zorzi, Optimal adaptive random multiaccess in energy harvesting wireless sensor networks. IEEE Trans. Commun. **63**(4), 1355–1372 (2015)
12. M.K. Sharma, C.R. Murthy, On the design of dual energy harvesting communication links with retransmission. IEEE Trans. Wirel. Commun. **16**(6), 4079–4093 (2017)
13. A. Yadav, M. Goonewardena, W. Ajib, O.A. Dobre, H. Elbiaze, Energy management for energy harvesting wireless sensors with adaptive retransmission. IEEE Trans. Commun. **65**(12), 5487–5498 (2017)
14. D. Silver et al., Mastering the game of go with deep neural networks and tree search. Nature **529**, 484–489 (2016)
15. W.H.R. Chan, P. Zhang, I. Nevat, S.G. Nagarajan, A.C. Valera, H.-X. Tan, N. Gautam, Adaptive duty cycling in sensor networks with energy harvesting using continuous-time Markov Chain and fluid models. IEEE J. Sel. Areas Commun. **33**(12), 2687–2700 (2015)
16. K.J. Prabuchandran, S.K. Meena, S. Bhatnagar, Q-learning based energy management policies for a single sensor node with finite buffer. IEEE Wirel. Commun. Lett. **2**, 1 (2013)
17. S. Padakandla, K.J. Prabuchandran, S. Bhatnagar, Energy sharing for multiple sensor nodes with finite buffers. IEEE Trans. Commun. **63**, 5 (2015)
18. F. Ahmed, H.-S. Cho, A time-slotted data gathering medium access control protocol using Q-learning for underwater acoustic sensor networks. IEEE Access **9**, 48742–48752 (2021)
19. B. Yang, X. Cao, Z. Han, L. Qian, A machine learning enabled MAC framework for heterogeneous Internet-of-Things networks. IEEE Trans. Wirel. Commun. **18**, 7 (2019)
20. A. Goldsmith, *Wireless Communication* (Cambridge University, Cambridge, 2005)
21. C.J. Watkins, Learning from delayed rewards. Ph.D. dissertation, Cambridge University, Cambridge, UK (1989)
22. C.J. Watkins, P. Dayan, Q-learning. Mach. Learn. **8**(3–4), 279–292 (1992)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.