CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Journal Paper

Reinforcement Learning for Scheduling Wireless Powered Sensor Communications

Kai Li
Wei Ni
Mehran Abolhasan
Eduardo Tovar

CISTER-TR-181116

Reinforcement Learning for Scheduling Wireless Powered Sensor Communications

Kai Li, Wei Ni, Mehran Abolhasan, Eduardo Tovar

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP-IPP)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail:

http://www.cister.isep.ipp.pt

Abstract

In a wireless powered sensor network, a base station transfers power to sensors by using Wireless Power Transfer (WPT). Inadequately scheduling WPT and data transmission causes fast battery drainage and data queue overflow of some sensors who could have potentially gained high data reception. In this paper, scheduling WPT and data transmission is formulated as a Markov decision process (MDP) by jointly considering sensors 19 energy consumption and data queue. In practical scenarios, the prior knowledge about battery level and data queue length in MDP is not available at the base station. We study reinforcement learning at the sensors to find a transmission scheduling strategy, minimizing data packet loss. An optimal scheduling strategy with full-state information is also investigated, assuming that the complete battery level and data queue information are well known by the base station. This presents the lower bound of the data packet loss in wireless powered sensor networks. Numerical results demonstrate that the proposed reinforcement learning scheduling algorithm significantly reduces network packet loss rate by 60%, and increases network goodput by 67%, compared to existing non-MDP greedy approaches. Moreover, comparing the optimal solutions, the performance loss due to the lack of sensors 19 full-state information is less than 4.6%.

Reinforcement Learning for Scheduling Wireless Powered Sensor Communications

Kai Li, *Member, IEEE,* Wei Ni, *Senior Member, IEEE,* Mehran Abolhasan, *Senior Member, IEEE,* and Eduardo Tovar

***Abstract*—In a wireless powered sensor network, a base station transfers power to sensors by using Wireless Power Transfer (WPT). Inadequately scheduling WPT and data transmission causes fast battery drainage and data queue overflow of some sensors who could have potentially gained high data reception. In this paper, scheduling WPT and data transmission is formulated as a Markov decision process (MDP) by jointly considering sensors' energy consumption and data queue. In practical scenarios, the prior knowledge about battery level and data queue length in MDP is not available at the base station. We study reinforcement learning at the sensors to find a transmission scheduling strategy, minimizing data packet loss. An optimal scheduling strategy with full-state information is also investigated, assuming that the complete battery level and data queue information are well known by the base station. This presents the lower bound of the data packet loss in wireless powered sensor networks. Numerical results demonstrate that the proposed reinforcement learning scheduling algorithm significantly reduces network packet loss rate by 60%, and increases network goodput by 67%, compared to existing non-MDP greedy approaches. Moreover, comparing the optimal solutions, the performance loss due to the lack of sensors' full-state information is less than 4.6%.**

***Index Terms*—wireless sensor network, wireless power transfer, markov decision process, reinforcement learning, optimization.**

I. Introduction

In this section, we introduce research background on Wireless Powered Sensor Network (WPSN), and motivation of scheduling Wireless Power Transfer (WPT) and data transmission in the WPSN.

A. Wireless Powered Sensor Network

Inexpensive sensors capable of computation and wireless communications are becoming increasingly available. However, sensor nodes are severely restrained by battery power, limiting the network lifetime and quality of service. Wireless Powered Sensor Network (WPSN) has been extensively studied, where energy is harvested by Wireless Power Transfer (WPT) to recharge battery of a sensor node [1]–[3]. Figure 1 depicts a scalable WPSN that is composed of multiple clusters, where each cluster is covered by one base station (BS). A number of fixed sensor nodes serve as data sources with sensing ability, for example, monitoring the environment and detecting anomaly. The nodes, equipped with a data communication antenna and a wireless power receiver, generate data packets at an application-specific sampling rate, and put them into a data queue for future transmission. A BS is deployed to transfer power to sensor nodes via WPT for charging their batteries, and collect sensory data from the nodes [4]–[7]. Beamforming can be used at the BS, either electronically or mechanically. The use of beamforming allows for the concentrated transfer of energy towards the intended nodes, avoiding the dispersion and waste of energy.

During each time slot (or epoch), one of the nodes is scheduled to transmit data to and harvest energy from the BS. To this end, the BS only generates one beam per time slot for both energy transfer and data transmission, thereby reducing the overhead of beamforming. Data transmission and WPT can be carried out in multiple clusters, simultaneously, where frequency division multiple access (FDMA) or code division multiplex access (CDMA) approach is applied to overcome inter-cluster interference. Within each cluster, we consider that WPT and data transmission work in the same radio frequency band, but different time slots, so that the nodes only need a single RF chain with reduced hardware cost. It is critical to schedule the WPT and data transmission to minimize packet loss and extend network lifetime since inadequately scheduling WPT and data transmission could cause some of the nodes to drain their battery and have their data queue overflow, while the other nodes waste their harvested energy.

Learning-based algorithms have been designed for energy harvesting systems [8]–[10]. In [8], the energy harvesting point-to-point communication is modeled by MDP. A reinforcement learning algorithm is applied at the power transmitter to find a power allocation policy that can increase network throughput. A learning-based algorithm is presented in [9] to schedule the data transmission under an energy availability constraint and an individual deadline constraint for each packet. It is assumed that the amount of harvested energy, the channel coefficients, and the transmit power at each time slot are taken from a finite discrete set, whose performance suffers from the "curse of dimensionality". In [10], reinforcement learning algorithms are studied for energy management at a single sensor node with a finite data queue. The algorithms learn randomness of the generated sensory data and harvested energy. However, the learning-based algorithms in the literature focus on

K. Li, and E. Tovar are with Real-Time and Embedded Computing Systems Research Centre (CISTER), 4249–015 Porto, Portugal (E-mail:{kaili,emt}@isep.ipp.pt).

W. Ni is with the Digital Productivity and Services Flagship, Commonwealth Scientific and Industrial Research Organization (CSIRO), Sydney, Australia (E-mail: wei.ni@csiro.au).

M. Abolhasan is with The University of Technology Sydney. (E-mail: mehran.abolhasan@uts.edu.au).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

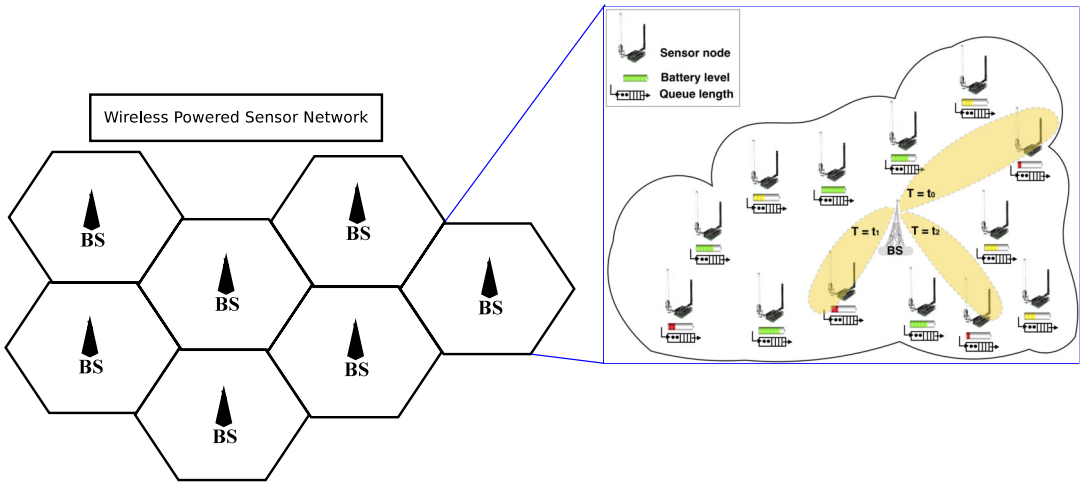


Fig. 1: Data transmission and WPT in the WPSN that is composed of multiple clusters, where each cluster is covered by one BS.

point-to-point communications, which is different from data transmission scheduling with multiple nodes.

Simultaneous energy harvesting and data transmission is formulated as MDP for addressing the scheduling problems in [11]–[13]. A cognitive radio communication system is considered in [11], where the user is powered by an energy harvester. Energy is consumed, during spectrum sensing and subsequent signal processing, and when the secondary user decides to carry out data transmission. Cognitive sensing and access policies are developed for the energy management of the secondary user with an energy buffer. Li *et al.* present scheduling strategies for cooperative communications, where a node uses a relay for its transmissions [12]. The scheduling policies choose different transmission mode for the sensors, depending on their available energy, as well as the states of their energy harvesting and data generating. In [13], energy management policies are developed to reduce packet queueing delay for a sensor node with an energy harvesting source. Conditions for energy neutral operation, i.e., the sensor node has sufficient energy and stable data queue, are also identified.

Several studies have integrated WPT technologies into wireless networks [14]–[18]. Chu *et al.* [14] study a WPSN for Internet-of-Things applications, where a multiantenna power charging station is employed to transfer power wirelessly to multiple sensor nodes. The power station and the sensor network can belong to different service providers. Incentives are designed for the sensors to purchase the energy from the power station. A hierarchical energy trading framework is designed to improve the net income of the sensor network. In [15], a data sender harvests energy from a power transmitter via WPT in the downlink before transmitting information to a data receiver in the uplink. It is found that data reception performance degrades when the time for energy harvesting and data transmission is unbalanced. Zhou *et al.* present a resource allocation problem for WPT and data transmission in downlink multiuser OFDM systems, where users harvest energy and decode data using the same signals received from the BS [16]. A tradeoff between the weighted sum-rate of all users and transferred energy is obtained, given the lower bound of harvested energy on each user and the upper bound of total transmission power. An energy harvesting resource allocation scheme considering imperfect channel state information is studied in [17]. It is found that channel training process and estimation error can affect energy harvesting efficiency of the network. However, the work in the literature only focuses on improving energy efficiency of WPT. The data loss caused by buffer overflows is not considered. A scheduling strategy is studied to reduce packet loss, considering battery level and data queue statues [18]. Based on specific packet collision probability distributions, a contention-based semi-decentralized algorithm is employed to allow the nodes to spontaneously self-nominate for data transmission and energy harvesting. However, transmission probability of all the sensor nodes has to be known, namely, channel statistical information needs to be predetermined before the scheduling. In contrast to [18], we study a more realistic scenario in which the statistical information about the underlying MDP is not available to the sensor nodes, and that, all the data arrivals, WPT, and the channel states are known only causally.

B. Research Motivation

To address the challenge of fast battery drainage and data queue overflow in WPSNs, in this paper, the scheduling problem is formulated as a finite-state Markov Decision Process (MDP) by jointly considering the sensor nodes' energy consumption and data queue state information. Considering practical scenarios, the prior knowledge about battery level and data queue length in the formulated MDP model is not available at the BS. We propose a new reinforcement learning strategy based on partial outdated statistical information of the MDP, aiming to minimize the packet loss of entire network and extend the battery lifetime of all the nodes. Moreover, to analyze lower bound of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

packet loss in WPSNs, we further investigate a MDP-based scheduling optimization, where the sensor nodes report their packet arrivals and battery levels in prior to every time slot. In this case, the complete up-to-date statistical information of the battery level and the data queue length at each sensor node are known a-priori by the BS. Numerical evaluations show that the proposed reinforcement learning approach is indistinguishably close to the full-state optimal scheduling solution in terms of packet loss rate and network goodput given different network size, learning iteration, discount factor, and length of the time slot. Particularly, the proposed reinforcement learning scheduling algorithm reduces the packet loss by 60% and raises the network goodput (defines the amount of data packets collected by the BS from all the sensor nodes) by 67%, compared to existing non-MDP greedy algorithms. In addition, while this paper is structured around WSNs, our framework can be applied to any wireless networks with energy harvesting capabilities.

The rest of this paper is organized as follows. Section II studies system model of the WPSN. In Section III, scheduling WPT and data transmission is modeled as MDP, and a reinforcement learning algorithm is proposed. The MDP-based scheduling optimization with full statistical knowledge is investigated in Section IV. Numerical results and evaluation are presented in Section V. Section VI concludes the paper.

II. System and Protocol Design

In this section, we introduce system model of WPSN and communication protocol. Notations used in this paper are listed in Table I.

TABLE I: The list of fundamental variables defined in system model

Notation	Definition
N	total number of sensor nodes
i	sensor node ID
$\mathbf{h}$	channel gain between the BS and the node
P^E	power that is transferred to the node
P_e	output power of the BS
L	number of bits of the sensory data packet
ϵ	required BER of the node
e_i	battery level of node i
E	battery capacity of the node
K	the highest battery level of the node
v_i	queue length of node i
V	maximum queue length of the node
ρ	modulation scheme of the node
M	the highest modulation order
λ	arrival probability of the data packet
T_λ	time when a new incoming data packet is put into the queue
$\widehat{T}$	length of the scheduling time slot
$\mathcal{A}$	action set of MDP
ϱ	learning rate
ω	discount factor for future states

A. System Model

We focus on scheduling WPT and data transmission in one of the clusters in the WPSN, as shown in Fig. 1. The network under consideration consists of a BS and N geographically distributed wireless powered nodes. The BS, connected to persistent power supplies, is responsible for remotely charging the nodes using WPT. Equipped with N_c antennas ($N_c \gg 1$), the BS can exploit transmit beamforming techniques to produce a narrow beam to each node. As a result, energy is transferred with improved focus and transfer efficiency. The BS is also responsible for collecting sensory data. Additionally, receive beamforming techniques enable the BS to enhance the received signal strength and reduce bit error rate (BER). Other advanced multi-user beamforming techniques, e.g., zero-forcing beamforming, are not considered in this work. Although they achieve spatial multiplexing or diversity, they would require real-time feedback on channel state information in most cases [19].

Each of the wireless powered nodes, e.g., node i ($i = 1, \cdots, N$), harvests energy from the BS to power its operations, e.g., sensing and communication. The rechargeable battery of the node is finite with the capacity of E_i Joules, and the battery overflows if overcharged. The complex coefficient of the reciprocal wireless channel between the BS and the node is $\mathbf{h}_i$, which can be known by channel reciprocity. Considering the non-persistent power supply and subsequently limited signal processing capability of the sensor nodes, we assume that each node is equipped with a single antenna, and we have $\mathbf{h}_i \in \mathcal{C}^{N_c \times 1}$. We consider the maximal ratio transmission (MRT) and maximal ratio combing (MRC) at the BS for the downlink energy transfer and uplink information detection, respectively. This is due to the fact that MRT and MRC maximize the receive signal-to-noise ratio (SNR) by exploiting the spatial diversity of wireless channels [20]. Furthermore, it has been shown that WPT efficiency is jointly affected by distance between WPT transmitter and receiver, and their antenna orientation [6]. Therefore, the power transferred to node i via WPT can be given by

$$P_i^E = \delta(d, \theta) P_e \|\mathbf{h}_i\|^2, \tag{1}$$

where $\delta(d, \theta) \in (0, 1]$ is a constant indicating WPT efficiency factor given the distance d and the antenna alignment θ between node i and BS. P_e is the constant output power of the BS, and $\|\cdot\|$ stands for norm. Besides, we consider that the WPT and data transmission work in the same radio frequency band, but different time slots, so that the nodes only need a single RF chain with reduced hardware cost.

Node i also keeps sensing its ambient environment, packetizes and queues (in a first-in-first-out (FIFO) fashion) the sensory data in packets of L_i bits, and transmits the packets to the BS through the wireless channel. The arrival/queueing process of sensory data at node i is modeled as a random process, where a new packet is put into the FIFO queue at a probability λ_i within every T_λ seconds. The node has a finite data queue to accommodate the maximum queue length of V_i packets or $V_i \cdot L_i$ bits. The data queue starts overflowing, once the maximum queue length is reached while the transmission of the sensor is withheld due to insufficient energy harvested. In other words, the arrival rate of sensory data exceeds the departure rate when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

the battery level of the node is insufficient to transmit data. To this end, the scheduling algorithm needs to be appropriately designed to assign the node sufficient channel and energy resources.

The modulation scheme that node i uses to transmit packets is denoted by ρ_i. $\rho_i \in \{1, 2, \cdots, M\}$, where $\rho_i = 1, 2$, and 3 indicates binary phase-shift keying (BPSK), quadrature-phase shift keying (QPSK), and 8 phase-shift keying (8PSK), respectively, and $\rho_i \geq 4$ corresponds to 2^{ρ_i} quadrature amplitude modulation (QAM). M is the highest modulation order.

Suppose that the BER requirement of node i is ϵ_i. The required transmit power of node i depends on ρ_i, ϵ_i, and $\mathbf{h}_i$, and can be given by [21]

$$P_i^D(\rho_i) \approx \frac{\kappa_2^{-1} \ln \frac{\kappa_1}{\epsilon_i}}{\|\mathbf{h}_i\|^2}(2^{\rho_i} - 1), \tag{2}$$

where κ_1 and κ_2 are channel related constants.

For illustration convenience, we consider a homogeneous network, where all the nodes have the same battery size, queue length, packet length, packet arrival probability, and the BER requirement. Their wireless channels are independent and identically distributed. The subscript "$_i$" is suppressed in E_i, V_i, L_i, λ_i, and ϵ_i. However, the proposed scheduling protocols can be extended to a heterogeneous network, where the complexity of the scheduling problem can grow. Moreover, the channels are assumed to be block fading. In other words, the channels remain unchanged during each time slot, and can change from slot to slot. The assumption of block fading channel is reasonable, because the duration of a time slot is typically milliseconds [22] while the channels between a pair of static transmitter and receiver are stable for tens to hundreds of milliseconds [23].

B. Communication Protocol

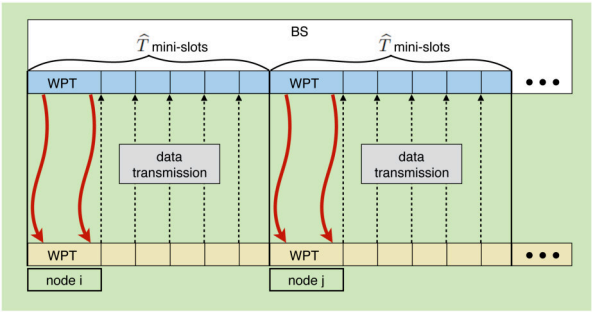


Fig. 2: The schedule of WPT and data transmission in the WPSN.

Figure 2 depicts the schedule of WPT and data transmission, where i and j ($i \neq j$ and $i, j \in [1, N]$) denote the node ID. Specifically, one time slot, also known as a scheduling interval, lasts $\widehat{T}$ mini-slots, during which the BS wirelessly transfers power to the node [24], followed by the data transmission from the node to the BS. The BS takes the action of choosing a node and deciding the associated modulation order at each time slot. Given such coordination, the nodes are activated in a TDMA fashion to transmit sensory data and harvest energy. We note that there is no collision between the data transmissions of the nodes, therefore, the receive and transmit beams can also be produced ahead of the actual data transmission or energy transfer.

We also note that the BS's actions depend on the current network state, i.e, the battery level e_i and queue length v_i of every node i loss. The actions also account for the potential influence on the future evolutions of the network. Particularly, the current action that the BS takes can affect the future battery level and queue length of every node, and in turn, influences the future actions to be taken. Such action taking is a discrete-time stochastic control process which is partly random (due to the random and independent arrival/queueing process of sensory data at every node) and partly under the control of the decision-making BS. The action of selecting node and modulation can be optimized in a sense that the optimality in regards of a specific metric, e.g., packet loss, is achieved in long term over the entire stochastic control process (rather than myopically at an individual time slot).

III. A Reinforcement Learning Approach

In this section, scheduling WPT and data transmission is formulated as a finite-state MDP. A new scheduling algorithm using Q-learning is proposed to minimize packet loss from a network of sensor nodes based on partial statistical information available at the BS. Furthermore, the modulation scheme of the node is optimized to maximize the energy harvested during a time slot.

A. MDP Formulation

We consider optimizing the actions to minimize the overall packet loss of the entire WPSN. The packet loss can result from queue overflows at the node where its data transmission is withheld due to insufficient energy harvested. The packet loss can also result from unsuccessful data transmissions over wireless fading channels. It is known that battery readings are continuous variables with variances difficult to be traced in real time. Therefore, to improve the mathematical tractability of the problem and illustration convenience, the continuous battery is discretized into K levels, as $0 < \mathcal{E} < 2\mathcal{E} < \cdots < K\mathcal{E} = E$. $e_i \in \{0, \mathcal{E}, 2\mathcal{E}, \cdots, K\mathcal{E}\}$ [25]. In other words, the battery level of a node is rounded downwards to the closest discrete level. Initially, the nodes may have inconsistent battery levels, i.e., $e_i \neq e_j$, where $i \neq j \in [1, N]$, and queue length $v_i \in \{0, 1, \cdots, V\}$. In other words, the premier MDP states of the nodes could be different from each other. Furthermore, the applied quantization of the continuous battery readings and discrete queue length facilitate generating MDP states. Although the accuracy of the quantization can be improved by reducing the quantization interval, it can result in an increased number of MDP states, and hence an increasing complexity of solving the MDP problem.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

The node can use a small control message to update the BS with its battery level and queue length at the beginning of every time slot. For example, consider a network of 40 nodes, battery level of 10 and queue length of 100 packets, the overhead of one node takes 11 bits, and total overhead is 440 bits, which is much smaller than the size of a data packet. Therefore, we assume that both the transmission time and the energy consumption of the overhead are negligible.

As noted earlier, the optimization of the action, i.e., the selection of node and modulation, at every time slot needs to be conducted over the entire scheduling process. The correlation between actions taken at different time slots needs to be captured, and to validate the long-term optimality of the actions. For this reason, we consider MDP for which the actions are chosen in each state to minimize a certain long-term objective. The MDP formulation requires a description of states, actions, and costs. The scheduling problem of interest can be formulated as a discrete-time MDP with $\widehat{T}$ mini-slots, where each state, denoted by $\mathcal{S}_\alpha$, collects the battery levels and queue lengths of all the nodes in the network, i.e., $\{(e_{\alpha,n}, v_{\alpha,n}), n = 1, \cdots, N\}$. The size of the state space is $\big(K(V+1)\big)^N$. The action to be taken, denoted by $\mathcal{A}$, is to select the node to be activated and specify its modulation. $\mathcal{A} \in \Big\{(i, \rho_i) : i = 1, \cdots, N, \rho_i \in \{1, \cdots, M\}\Big\}$. The size of the action set is NM.

Here, $\mathcal{A}$ can be reduced to only consist of the selected node, i.e., $n \in \mathcal{A}$. The maximum energy that can be harvested into the battery of the selected node n can be given by

$$\Delta\mathcal{E}_n = \Bigg\lfloor (\widehat{T} - \frac{L}{\rho_n^\star W})\frac{P_e\|\mathbf{h}_n\|^2}{\mathcal{E}} - \frac{L\kappa_2^{-1}\ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_n\|^2\rho_n^\star W\mathcal{E}}\big(2^{\rho_n^\star}-1\big)\Bigg\rfloor \mathcal{E}, \tag{3}$$

where $\rho_n^\star$ denotes the optimal modulation scheme of node n, and can be obtained by the optimization of ρ_i (see Section III-C).

B. Q-Learning Algorithm for Scheduling Optimization

Given a practical scenario where the BS has no prior knowledge on the transition probabilities and the immediate costs, we propose a new scheduling algorithm that utilizes Q-learning, one of the reinforcement learning techniques, to minimize the packet loss of the network and extend the battery lifetime of all the nodes. Specifically, we define that a policy π is a mapping from states to actions, and the set of all policies is defined as Π. Particularly, the MDP is said to be stationary if the corresponding policy is independent of the current state. The goal of the algorithm is to minimize the expected total packet loss which is denoted as $v(S_\alpha)$,

$$v(S_\alpha) = \min_{\pi\in\Pi}\ \mathbb{E}^\pi_{\mathcal{S}_\alpha}\left\{\sum_{t=1}^{\infty}\omega^{t-1}C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}\right\}, \tag{4}$$

where $\omega \in [0,1]$ is a discount factor for future states. $C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ is the cost from state $\mathcal{S}_\alpha$ to $\mathcal{S}_\beta$ when action k is carried out. $\mathbb{E}^\pi_{\mathcal{S}_\alpha}\{\cdot\}$ denotes the expectation with respect to policy π and state $\mathcal{S}_\alpha$. To find the action k to calculate $v(S_\alpha)$, we evaluate for each action $k \in \mathcal{A}$ and select the actions that achieve (4). Thus, we have

$$\begin{aligned} v(S_\alpha) = \ & \min_{k\in\mathcal{A}}\{C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\} \\ & + \textstyle\sum_{\mathcal{S}_\beta\in\mathcal{S}}\omega\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}v(\mathcal{S}_\beta)\}, \end{aligned} \tag{5}$$

where $\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ is transition probability of the MDP, from state $\mathcal{S}_\alpha$ to $\mathcal{S}_\beta$, given the action k. $v(\mathcal{S}_\beta)$ is the expected total packet loss in the posterior state $\mathcal{S}_\beta$.

The optimal action, namely k^*, which satisfies (5), can be given by

$$\begin{aligned} k^* = \ & \arg\min_{k\in\mathcal{A}}\{C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\} \\ & + \textstyle\sum_{\mathcal{S}_\beta\in\mathcal{S}}\omega\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}v(\mathcal{S}_\beta)\}. \end{aligned} \tag{6}$$

Note that k^* exists due to a finite $\mathcal{A}$ and may not be unique. k^* leads to the minimum expected packet loss.

In the absence of prior knowledge on the values of $C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ or $\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$, the BS can learn an action-value function $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ corresponding to the expected accumulated cost when action $k \in \mathcal{A}$ is taken in the state $\mathcal{S}_\alpha$ under the decision policy π. In particular, the action-value function following action k is defined as

$$\begin{aligned} Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\} = & (1-\varrho)Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\} + \\ & \varrho\Big[C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\} + \omega Q\Big\{\mathcal{S}_{\beta'}\Big|\mathcal{S}_\beta,k'\Big\}\Big], \end{aligned} \tag{7}$$

where $\varrho \in (0,1]$ is a small positive fraction which influences the learning rate. When action $k' \in \mathcal{A}$ is taken, $\mathcal{S}_{\beta'}$ denotes the next state of $\mathcal{S}_\beta$. From (7), $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ is estimated considering the transitions from a state-action pair $(\mathcal{S}_\alpha,k)$ to another state-action pair $(\mathcal{S}_\beta,k')$ while obtaining cost $C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$. In other words, when the nodes are in state $\mathcal{S}_\alpha$, the BS schedules node k for WPT and data transmission. Afterwards, the BS obtains a cost $C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ and moves to state $\mathcal{S}_\beta$. According to the current values of $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$, Q-learning algorithm can decide the next action k'. At this point, $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ is updated using the gained experience and the current values of $Q\Big\{\mathcal{S}_{\beta'}\Big|\mathcal{S}_\beta,k'\Big\}$. By performing the optimal action k^*, the optimal action-value function based on (7) can be expressed as a combination of the expected cost and the minimum value of $Q^*\Big\{\mathcal{S}_{\beta'}\Big|\mathcal{S}_\beta,k'\Big\}$, where

$$\begin{aligned} Q^*\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k^*\Big\} = & (1-\varrho)Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k^*\Big\} + \\ & \varrho\Big[C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k^*\Big\} + \omega\min_{k'\in\mathcal{A}}Q\Big\{\mathcal{S}_{\beta'}\Big|\mathcal{S}_\beta,k'\Big\}\Big]. \end{aligned} \tag{8}$$

In particular, the convergence rate of $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k\Big\}$ to $Q^*\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha,k^*\Big\}$ depends on the learning rate ϱ. The convergence rate also decreases with the number of actions, states, and the discount factor ω; and increases with the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

Algorithm 1 Q-learning Scheduling Algorithm (QSA)

1: **1. Initialize**:
2: $\mathcal{S}_\alpha \in \mathcal{S}$, $k \in \mathcal{A} \rightarrow Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$, ϱ, and w.
3: An action-value table consists of three vectors which includes current state $\mathcal{S}_\alpha$, next state $\mathcal{S}_\beta$, and Q value.
4: Learning time $\rightarrow t_{\text{learning}}$.
5: **2. Learning**:
6: **for** time t_{learning} **do**
7: **while** $\mathcal{S}_\alpha \in \mathcal{S}$ **do**
8: Perform $k \in \mathcal{A}$.
9: Calculate the corresponding expected cost $\rightarrow C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$.
10: Obtain the action-value function $\rightarrow$ (7).
11: **end while**
12: Select the optimal action $k^* \leftarrow$ (8).
13: The next state can be updated to $\mathcal{S}_\beta$ with regards to $Q^*\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k^*\Big\}$.
14: Create a record of $\mathcal{S}_\alpha$, $\mathcal{S}_\beta$, and $Q^*\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k^*\Big\}$ in the action-value table.
15: **end for**

learning time. More details of the convergence rate of the Q-learning algorithm can be found in [26].

Based on (8), the action in each $\widehat{T}$ is chosen by optimizing $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$ in a recursive manner. Algorithm 1 presents the proposed Q-learning Scheduling Algorithm (QSA), which solves (8), in the case that the transition probability $\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$ is unknown. Given the learning time of t_{learning}, the BS observes the next state $\mathcal{S}_\beta \in \mathcal{S}$ based on the learning outcome of $\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$, which estimates the battery levels and queue lengths of all the nodes in the network. Then, the BS performs an action $k \in \mathcal{A}$, namely, scheduling one node to transmit data and harvest energy. Accordingly, the immediate cost $C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$ and the next state $\mathcal{S}_\beta$ can be calculated. The action-value function $Q\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$ can be obtained due to (7). By applying (8), the BS is able to select the optimal action k^*. Given k^*, the next state $\mathcal{S}_\beta$ can be accordingly updated. The BS can create a table to store the results of the action-value function, and record each state-action pair, i.e., $\mathcal{S}_\alpha$, $\mathcal{S}_\beta$, and $Q^*\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k^*\Big\}$. Iteratively, QSA explores all the states during t_{learning}, and determines the optimal action at each state.

C. Selecting ρ_i

Given the goal of QSA to minimize the packet loss stemming from insufficient energy, ρ_i is to be chosen to maximize the energy harvested during a time slot, with a duration of $\widehat{T}$. The optimal modulation of node i, $\rho_i^\star$, is independent of the battery level and the queue length of the node i. This is because $\rho_i^\star$ is selected to maximize the increase of the battery level at node i, under the bit error rate requirement ϵ_i for the packet transmitted. As a result, ρ_i can be decoupled from $\mathcal{A}$, and optimized in priori by

$$\rho_i = \arg\max_{\rho=1,\cdots,M}\left\{(\widehat{T}-\frac{L}{\rho W})P_i^E - \frac{L}{\rho W}P_i^D(\rho)\right\}, \quad (9)$$

the right-hand side (RHS) of which, by substituting (1) and (2), can be rewritten as

$$\max_{\rho=1,\cdots,M}\left\{(\widehat{T}-\frac{L}{\rho W})P_e\|\mathbf{h}_i\|^2 - \frac{L\kappa_2^{-1}\ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i\|^2\rho W}\big(2^\rho-1\big)\right\}, \quad (10)$$

where W is the bandwidth of the uplink data transmission, $\frac{1}{W}$ is the duration of an uplink symbol, $\frac{L}{\rho W}$ is the duration of uplink data transmission, and $(\widehat{T}-\frac{L}{\rho W})$ is the rest of the time slots used for downlink WPT.

By using the first-order necessary condition of the optimal solution, we have

$$\frac{d}{d\rho}((\widehat{T}-\frac{L}{\rho W})P_e\|\mathbf{h}_i\|^2 - \frac{L\kappa_2^{-1}\ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i\|^2\rho W}\big(2^\rho-1\big)) = 0, \quad (11)$$

$$\begin{aligned}\rho^{-2}\frac{L}{W}P_e\|\mathbf{h}_i\|^2 - \frac{L\kappa_2^{-1}\ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i\|^2 W}(\rho^{-1}2^\rho\text{ln}2 - \rho^{-2}2^\rho) \\ - \frac{L\kappa_2^{-1}\ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i\|^2 W}\rho^{-2} = 0.\end{aligned} \quad (12)$$

The ρ values are then given as follows:

$$\rho 2^\rho\text{ln}2 - 2^\rho = \frac{L}{W}P_e\|\mathbf{h}_i\|^2\frac{\|\mathbf{h}_i\|^2 W}{L\kappa_2^{-1}\ln(\frac{\kappa_1}{\epsilon})} - 1. \quad (13)$$

Since the left-hand side (LHS) of (13) monotonically increases with ρ, the optimal value $\rho^\star$ can be obtained by applying a bisection method, and evaluating the two closest integers about the fixed point of the bisection method [27]. Specifically, $\rho_- = 1$ and $\rho_+ = M$ are initialized. Each iteration of the bisection method contains 4 steps applied over the range of $\rho = [1; M]$, as follows.

- The midpoint of modulation interval $[\rho_-; \rho_+]$ is calculated, which gives $\rho_{\text{mid}} = \frac{\rho_- + \rho_+}{2}$.
- Substitute ρ_{mid} into (13) to obtain the function value $f(\rho_{\text{mid}})$.
- If convergence is satisfactory (that is, the modulation interval or $|f(\rho_{\text{mid}})|$ can not be further reduced), return ρ_{mid} and stop iterating.
- Replace either $(\rho_-, f(\rho_-))$ or $(\rho_+, f(\rho_+))$ with $(\rho_{\text{mid}}, f(\rho_{\text{mid}}))$.

IV. Scheduling Optimization with Full-State Information

For comparison purpose, we process to consider the scheduling problem with full-state information of $\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$, where $\mathcal{S}_\alpha, \mathcal{S}_\beta \in \mathcal{S}$ and $k \in \mathcal{A}$. Actions of the finite-state MDP, i.e., the selection of sensor nodes, can be optimized by meticulously deriving the transition probabilities, and solved by using dynamic programming (DP) techniques.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

A. Transition Probability and Packet Loss

Given action $k \in \mathcal{A}$ and $1 \leq (\alpha, \beta) \leq K^N(V+1)^N$, the transition probability of the MDP, from state $\mathcal{S}_\alpha$ to $\mathcal{S}_\beta$, can be given by

$$\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\} = \Pr\Big\{(e_{\beta,k}, v_{\beta,k})\Big|(e_{\alpha,k}, v_{\alpha,k}), k \in \mathcal{A}\Big\} \times \prod_{n=1, n\neq k}^{N} \Pr\Big\{(e_{\beta,n}, v_{\beta,n})\Big|(e_{\alpha,n}, v_{\alpha,n}), n \neq k\Big\}, \tag{14}$$

where the two parts of the RHS are specified in (15). Specifically, (15a) corresponds to the selected node k with three different cases. The first case is that the queue of the node increases due to the failed transmission of a packet and the arrival of a new sensory packet. The second case is that the queue decreases due to a successful transmission and no new packet arrival. The third case is that the queue does not change, due to either (a) a successful transmission and a new packet arrival, or (b) a failed transmission and no new packet arrival. The battery level of the selected node increases by $\Delta\mathcal{E}_k$ in all the three cases, given the optimized modulation order ρ_k.

(15b) corresponds to the unselected nodes $n \neq k$ with two different cases. The first case is that the queue of the node increases due to a new packet arrival. The second case is that the queue does not change, without a new packet arrival. The battery level of the node does not change, since the node is not selected to harvest energy.

The packet loss, resulting from queue overflow during the transition, can be given by

$$C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\} = C\Big\{(e_{\beta,k}, v_{\beta,k})\Big|(e_{\alpha,k}, v_{\alpha,k}), k \in \mathcal{A}\Big\} + \sum_{n=1, n\neq k}^{N} C\Big\{(e_{\beta,n}, v_{\beta,n})\Big|(e_{\alpha,n}, v_{\alpha,n}), n \neq k\Big\}; \tag{16}$$

$$C\Big\{(e_{\beta,k}, v_{\beta,k})\Big|(e_{\alpha,k}, v_{\alpha,k}), k \in \mathcal{A}\Big\} = \begin{cases} \Big(1-(1-\epsilon)^L\Big)\lambda, & \text{if } v_{\alpha,k} = v_{\beta,k} = V; \\ 0, & \text{otherwise.} \end{cases} \tag{17a}$$

$$C\Big\{(e_{\beta,n}, v_{\beta,n})\Big|(e_{\alpha,n}, v_{\alpha,n}), n \neq k\Big\} = \begin{cases} \lambda, & \text{if } v_{\alpha,n} = v_{\beta,n} = V; \\ 0, & \text{otherwise.} \end{cases} \tag{17b}$$

The first case of (17a) is the probability that the data queue of the selected node k overflows. Specifically, the new generated packet at State $\mathcal{S}_\beta$ with the arrival probability λ has to be dropped since the data queue is fully occupied, and the node k fails to transmit the packet at State $\mathcal{S}_\alpha$. Similarly, the first case of (17b) gives the probability that the data queue of an unselected node $n \neq k$ overflows.

B. Dynamic Programming Algorithm

The actions of the MDP can be optimized by using DP techniques [28]. Generally, value iterations and policy iterations are two popular methods for computing optimal actions for MDP with expected total discounted costs. In this paper, the value iteration method [29], [30] is applied. Basically, the value iteration method computes the optimal cost functions by assuming first a one-stage finite horizon, then a two-stage finite horizon, so on and so forth. The cost functions are computed to converge in the limit to the optimal cost function. Therefore, the policy associated with the successive cost functions converges to the optimal policy in a finite number of iterations. In addition, policy iteration is also adoptable to obtain the optimal actions of the MDP problems.

Algorithm 2 Full-State Scheduling Algorithm (FSSA)

1: **1. Initialize**:
2: $v^0(S_\alpha) = 0$ for the state S_α, $t \in [0, \infty]$, specify $\varepsilon > 0$, and set $w = 0$.
3: **2. Iteration**:
4: **while** $S_\alpha \in \mathcal{S}$ **do**
5: $v^{w+1}(S_\alpha) = \min_{k\in\mathcal{A}}\{C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\} + \sum_{\mathcal{S}_\beta\in\mathcal{S}} \omega \Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\} v(\mathcal{S}_\beta)\}$
6: **if** $||v^{w+1} - v^w|| < \varepsilon(1-\omega)/(2\omega)$ **then**
7: **for** $S_\alpha \in \mathcal{S}$ **do**
8: $k = \arg\min_{k\in\mathcal{A}}\{C\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\} + \sum_{\mathcal{S}_\beta\in\mathcal{S}} \omega \Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\} v(S_\beta)\}$
9: **end for**
10: **end if**
11: **end while**

The proposed Full-State Scheduling Algorithm (FSSA) is summarized in Algorithm 2. Since the running time for each iteration is $O(MN^2)$, the presented FSSA is polynomial. Note that the scheduling optimization with full-state information of $\Pr\Big\{\mathcal{S}_\beta\Big|\mathcal{S}_\alpha, k\Big\}$ is ideal, and can only be conducted off-line. Under the assumption of the availability of up-to-date global information at the BS, the optimal actions of the MDP are inapplicable to practical environments with partial state information. Nevertheless, the MDP with full-state information (or its optimal actions) can provide the lower-bound benchmark of packet loss to any scheduling designs of WPSN.

V. Numerical Results and Discussions

The simulations prototyping our proposed scheduling are carried out in this section. We compare the performance between our proposed QSA and FSSA, and existing non-MDP scheduling strategies (used in [31], [32]) given different network sizes and learning iterations. Moreover, we also show the impact of discount factor and length of $\widehat{T}$ on the performance of QSA and FSSA.

A. Simulation Configuration

The sensor nodes ($N \in [10, 40]$) are deployed in the range of 50 meters, which are all one-hop away from the BS. The node has the maximum discretized battery level

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

$$\Pr\left\{(e_{\beta,k},v_{\beta,k})\middle|(e_{\alpha,k},v_{\alpha,k}),k\in\mathcal{A}\right\}=\begin{cases}\left(1-(1-\epsilon)^L\right)\lambda, & \text{if } e_{\beta,k}=e_{\alpha,k}+\Delta\mathcal{E}_k \text{ and } v_{\beta,n}=v_{\alpha,n}+1;\\ (1-\epsilon)^L(1-\lambda), & \text{if } e_{\beta,k}=e_{\alpha,k}+\Delta\mathcal{E}_k \text{ and } v_{\beta,n}=v_{\alpha,n}-1;\\ \left(1-(1-\epsilon)^L\right)(1-\lambda)+(1-\epsilon)^L\lambda, & \text{if } e_{\beta,k}=e_{\alpha,k}+\Delta\mathcal{E}_k \text{ and } v_{\beta,n}=v_{\alpha,n};\\ 0, & \text{otherwise.}\end{cases} \tag{15a}$$

$$\Pr\left\{(e_{\beta,n},v_{\beta,n})\middle|(e_{\alpha,n},v_{\alpha,n}),n\neq k\right\}=\begin{cases}\lambda, & \text{if } e_{\beta,n}=e_{\alpha,n} \text{ and } v_{\beta,n}=v_{\alpha,n}+1;\\ 1-\lambda, & \text{if } e_{\beta,n}=e_{\alpha,n} \text{ and } v_{\beta,n}=v_{\alpha,n};\\ 0, & \text{otherwise.}\end{cases} \tag{15b}$$

$E=5$ and queue length $V=6$, and the highest modulation M is 5. Each data packet has a payload of 32 bytes, i.e., $L=256$. The data transmission period is given as a sequence of time slots, where each node generates one data packet per time slot and put into the queue. The length of the time slot for scheduling and learning process, i.e., $\widehat{T}$, lasts 5 mini-slots, unless otherwise specified. For $P_i^D(\rho_i)$ in (2), the two constants, κ_1 and κ_2 are set to 0.2 and 3, respectively. We set the target $\epsilon=0.05\%$ for the numerical results, i.e., the transmitted bits have an error no more than 0.05, however, this value can be configured depending on the traffic type and quality-of-service (QoS) requirement of the sensory data. The transmit power of WPT transmitter at BS is 3 watts, and the power transfer efficiency, i.e., $\delta(d,\theta)$, is set to 0.5.

For comparison purpose, two non-MDP heuristic algorithms are simulated. The first algorithm, referred to as "Full Queue (FQ)", is a greedy algorithm, where the scheduling is based on the data queue length of the node. The node with full queue has a higher priority to transmit data and harvest power. The second algorithm is named as "Random Selection (RS)", where the BS randomly selects one node to transmit data and transfer power.

B. Performance Evaluation

1) Comparing to FSSA and non-MDP algorithms: Figure 3 shows the network packet loss rate with an increasing number of nodes, where the learning rate of QSA is set to 0.1 or 0.3. We can see that the packet loss rate of QSA is lower than FQ and RS by around 29% and 60%, respectively. The reason is that QSA learns the sensor nodes' energy consumption and data queue state so that the scheduling of WPT and data transmission can minimize the packet loss of the entire network. Furthermore, the packet loss rate achieved by QSA is much closer to those of the FSSA, which takes advantage of the complete up-to-date state information of MDP, as compared to the other scheduling algorithms. Note that in FSSA, the BS needs to have the complete knowledge on the packet arrivals and battery levels of all the nodes. This requires the nodes to report in prior to every time slot. However, QSA allows the BS to learn the schedule of WPT and data transmission based on the partial outdated statistical information, as well as scheduling the modulation level if the transmission is turned on.

In Figure 3, we also see that convergence speeds up with an increase of ϱ. In particular, QSA with $\varrho=0.3$ has 17% higher packet loss rate than the one with $\varrho=0.1$ when $N=40$. This is because a large ϱ results in a fast convergence of learning $Q\left\{\mathcal{S}_\beta\middle|\mathcal{S}_\alpha,k\right\}$ with regards to (7) given a limited t_{learning}. Furthermore, decreasing the learning rate of QSA (i.e., reducing the step size of the learning) reduces the network packet loss in WPSNs. However, the low learning rate slows down the convergence of QSA. In this sense, the setting of the learning rate is important, and needs to be fine-tuned in accordance with different performance requirement of the WPSN.

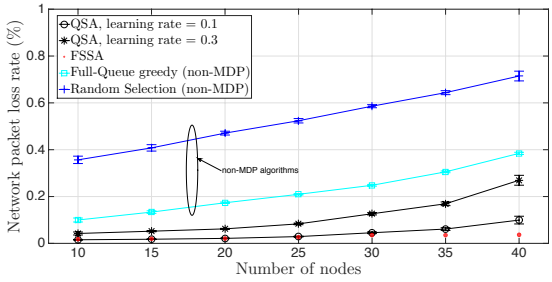


Fig. 3: A comparison of network packet loss rate by QSA, FSSA and the typical scheduling strategies, where the error bars show the standard deviation over 30 runs.

In general, the packet loss rate of QSA is lower than FQ and RS. This is also observed by Figure 4, which presents the network goodput. QSA generally achieves higher network goodput than FQ and RS. Specifically, QSA have similar goodput to FQ and RS, when there are 10 nodes in the network. However, from $N=15$ to 40, QSA outperforms the two non-MDP scheduling algorithms. In particular, QSA with $\varrho=0.1$ achieves 25%, and 67% higher goodput than FQ and RS, respectively, when $N=40$. Moreover, the performance of QSA with $\varrho=0.1$ is close to the one achieved by FSSA, which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

has the complete up-to-date state information of MDP. It is also observed that the goodput of QSA converges with an increase of ϱ since a large learning rate leads to a deterministic action-value function.

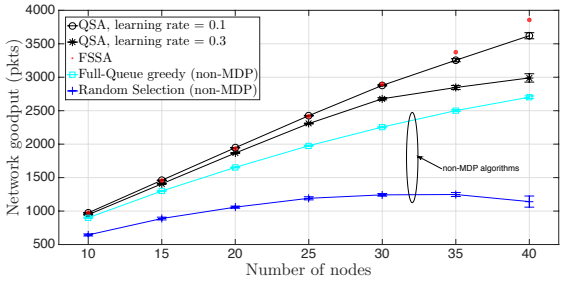


Fig. 4: A comparison of network goodput using different scheduling strategies, where the error bars show the standard deviation over 30 runs.

2) Performance of QSA given different t_{learning}: Figure 5 studies the network packet loss rate with regards to t_{learning}, where t_{learning} is given as the number of learning iterations, and N is 25. As expected, the packet loss of QSA drops when the amount of learning iterations increases, while FSSA, FQ, and RS maintain a fixed packet loss regardless of t_{learning}. This is reasonable because a large t_{learning} allows $Q^*\left\{\mathcal{S}_\beta \middle| \mathcal{S}_\alpha, k^*\right\}$ in QSA to be derived with more states and actions, which gets its performance closer to the full-state MDP optimization. Therefore, the packet loss rate of QSA approximates those in FSSA with the growth of t_{learning}. Since increasing the learning iterations only updates the action-value function, the performance of FSSA, FQ, and RS is unaffected by t_{learning}. Moreover, we observe that for $t_{\text{learning}} \geq 1900$ iterations QSA achieves lower packet loss rates than FQ, while QSA outperforms RS with substantial gains about 32% when $t_{\text{learning}} = 1900$, and the gains keep growing with t_{learning}.

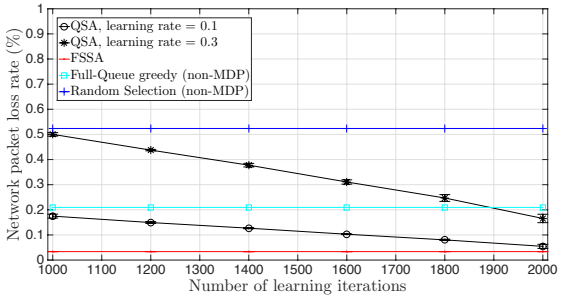


Fig. 5: The network packet loss rate with regards to the learning iterations, i.e., t_{learning}, where the error bars show the standard deviation over 30 runs.

In terms of network goodput, it is observed in Figure 6 that the performance of QSA generally grows with t_{learning}. The reason is that the nodes are scheduled to transmit data with the minimized data loss. Moreover, the nodes using QSA transmit 450 and 1150 more packets than those using FQ and RS, respectively, when t_{learning} increases to 2000 iterations. Additionally, the RS algorithm shows a performance well below the others.

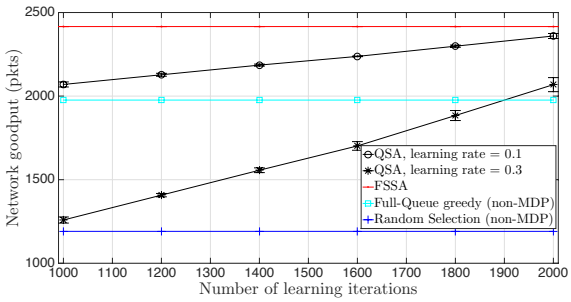


Fig. 6: The network goodput with regards to t_{learning}, where the error bars show the standard deviation over 30 runs.

Figures 5 and 6 also show that QSA provides a practical solution to WPSN, as the packet loss rate and network goodput of QSA approach those of FSSA. The reason is that as t_{learning} increases, the optimal action of QSA is learned increasingly correctly with a growing number of iterations of updating the action-value functions. In this case, the BS can learn increasingly accurate statistical information on the battery level and data queue length of each sensor node.

3) Impact of ω: Figures 7 and 8 depict the packet loss rate and network goodput of MDP-based scheduling algorithms, i.e., QSA and FSSA, with respect to the discount factor $\omega \in [0.9, 0.98]$. N, t_{learning} and ϱ are set to 25 nodes, 2000 iterations and 0.1, respectively, and the other settings are as configured in Section V-A. As observed, QSA achieves a similar packet loss rate and goodput to FSSA with a difference about 2.9%∼4.6% and 89∼124 packets, respectively. The performance of QSA confirms that the convergence rate of QSA decreases with the discount factor ω. In other words, a low discount factor accelerates reinforcement learning. The performance of FSSA also confirms the validity of QSA, which efficiently schedules the wireless powered sensor communications with partial outdated statistical information.

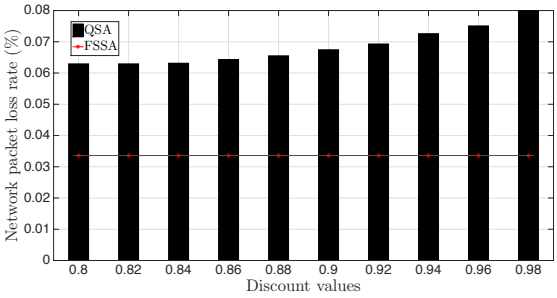


Fig. 7: The packet loss rate with different discount factors.

4) Impact of $\widehat{T}$: We study impact of the length of time slot for MDP and learning process that lasts $\widehat{T}$ mini-slots. Particularly, ω is set to 0.8. Figures 9 and 10 show the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

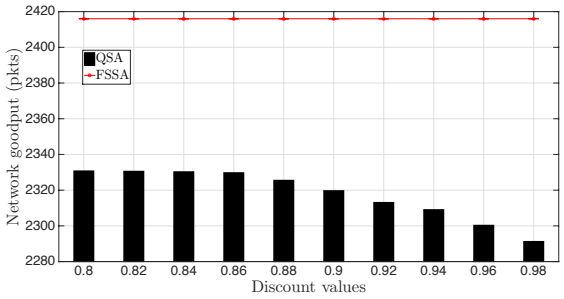


Fig. 8: The network goodput with different discount factors.

packet loss rate and network goodput of QSA and FSSA, where $\widehat{T} \in [10, 70]$. It is observed that the packet loss rates of QSA and FSSA grow with $\widehat{T}$ while their goodputs drop. This confirms that an increasing number of packets are generated and injected into the data queues of nodes with a long $\widehat{T}$. However, only one node is scheduled to transmit data and harvest energy during each $\widehat{T}$, though the node can harvest more energy with a longer $\widehat{T}$ due to (3). As a result, the nodes that are not scheduled suffer from severer packet losses caused by queue overflows, compared to those with shorter $\widehat{T}$. Moreover, we also observe that FSSA achieves lower packet loss and higher network goodput than QSA. This is reasonable since the former has full-state up-to-date information of each node and can accordingly achieve the optimal schedules.

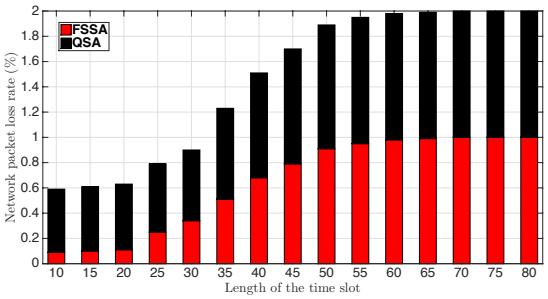


Fig. 9: The network packet loss with regards to the length of the time slot, i.e., $\widehat{T}$.

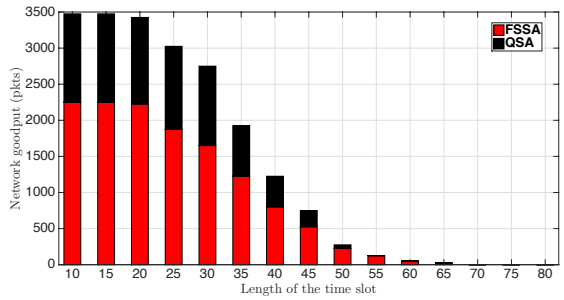


Fig. 10: The network goodput with regards to $\widehat{T}$.

VI. Conclusion

We have considered the scheduling problem of WPT and data transmission in WPSNs. Based on partial outdated statistical information of the sensor nodes' energy consumption and data queue, we propose a reinforcement learning algorithm, QSA, to minimize the packet loss caused by data queue overflows or unsuccessful data transmissions of the sensors. Moreover, the modulation of the sensor node is also optimized to maximize the energy harvested during a time slot. To further analyze the lower bound of the packet loss in WPSNs, we also investigate a MDP-based scheduling optimization, where the complete up-to-date statistical information of each sensor node is known a-priori. Numerical results have shown that QSA provides a near-optimal scheduling to the WPT and data transmission in WPSNs. QSA can also reduce the packet loss rate by 60%, and increase the network goodput by 67%, as compared to the existing algorithms.

Acknowledgements

This work was partially supported by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology), within the CISTER Research Unit (CEC/04234); also by the Operational Competitiveness Programme and Internationalization (COMPETE 2020) through the European Regional Development Fund (ERDF) and by national funds through the FCT, within project POCI-01-0145-FEDER-029074 (ARNET). The authors would like to thank the editors and the anonymous reviewers for their constructive comments on the article.

References

[1] K. Li, W. Ni, L. Duan, M. Abolhasan, and J. Niu, "SWPT: A joint-scheduling model for wireless powered sensor networks," in *Proceedings of IEEE Global Communications Conference (Globecom)*, 2017.

[2] F. Sangare, Y. Xiao, D. Niyato, and Z. Han, "Mobile charging in wireless-powered sensor networks: Optimal scheduling and experimental implementation," *IEEE Transactions on Vehicular Technology*, 2017.

[3] K. W. Choi, L. Ginting, P. A. Rosyady, A. A. Aziz, and D. I. Kim, "Wireless-powered sensor networks: How to realize," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 221–234, 2017.

[4] J. Huang, C.-C. Xing, and C. Wang, "Simultaneous wireless information and power transfer: technologies, applications, and research challenges," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 26–32, 2017.

[5] Y. Shu, H. Yousefi, P. Cheng, J. Chen, Y. J. Gu, T. He, and K. G. Shin, "Near-optimal velocity control for mobile charging in wireless rechargeable sensor networks," *IEEE Transactions on Mobile Computing*, vol. 15, no. 7, pp. 1699–1713, 2016.

[6] K. Li, C. Yuen, and S. Jha, "Fair scheduling for energy harvesting WSN in smart city," in *SenSys*. ACM, 2015, pp. 419–420.

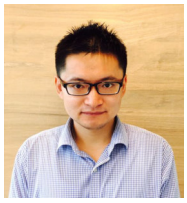
[7] S. Bi, C. K. Ho, and R. Zhang, "Wireless powered communication: Opportunities and challenges," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 117–125, 2015.

[8] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Reinforcement learning for energy harvesting point-to-point communications," in *International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.

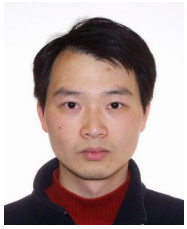
[9] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1872–1882, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2879023, IEEE Transactions on Green Communications and Networking

[10] K. Prabuchandran, S. K. Meena, and S. Bhatnagar, "Q-learning based energy management policies for a single sensor node with finite buffer," *IEEE Wireless Communications Letters*, vol. 2, no. 1, pp. 82–85, 2013.
[11] A. Sultan, "Sensing and transmit energy optimization for an energy harvesting cognitive radio," *IEEE wireless communications letters*, vol. 1, no. 5, pp. 500–503, 2012.
[12] H. Li, N. Jaggi, and B. Sikdar, "Relay scheduling for cooperative communications in sensor networks with energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 2918–2928, 2011.
[13] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, "Optimal energy management policies for energy harvesting sensor nodes," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, 2010.
[14] Z. Chu, F. Zhou, Z. Zhu, R. Q. Hu, and P. Xiao, "Wireless powered sensor networks for internet of things: Maximum throughput and optimal power allocation," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 310–321, 2018.
[15] F. Zhao, L. Wei, and H. Chen, "Optimal time allocation for wireless information and power transfer in wireless powered communication systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1830–1835, 2016.
[16] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer in multiuser ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2282–2294, 2014.
[17] K. Lee and J.-P. Hong, "Energy-efficient resource allocation for simultaneous information and energy transfer with imperfect channel estimation," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2775–2780, 2016.
[18] K. Li, W. Ni, L. Duan, M. Abolhasan, and J. Niu, "Wireless power transfer and data collection in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, 2017.
[19] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 102–111, 2011.
[20] X. Chen, X. Wang, and X. Chen, "Energy-efficient optimization for wireless information and power transfer in large-scale mimo systems employing energy beamforming," *IEEE Wireless Communications Letters*, vol. 2, no. 6, pp. 667–670, 2013.
[21] T. He, X. Wang, and W. Ni, "Optimal chunk-based resource allocation for OFDMA systems with multiple BER requirements," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 9, pp. 4292–4301, 2014.
[22] L. Shi and A. O. Fapojuwo, "TDMA scheduling with optimized energy efficiency and minimum delay in clustered wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 7, pp. 927–940, 2010.
[23] K. Li, C. Yuen, B. Kusy, R. Jurdak, A. Ignjatovic, S. S. Kanhere, and S. Jha, "Fair scheduling for data collection in mobile sensor networks with energy harvesting," *IEEE Transactions on Mobile Computing*, 2018.
[24] A. Kadrolkar, R. X. Gao, R. Yan, and W. Gong, "Variable-word-length coding for energy-aware signal transmission," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 4, pp. 850–864, 2012.
[25] K. H. Liu, "Selection cooperation using RF energy harvesting relays with finite energy buffer," in *Wireless Communications and Networking Conference (WCNC)*. IEEE, 2014, pp. 2156–2161.
[26] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
[27] D. Estep, "The bisection algorithm," *Practical Analysis in One Variable*, pp. 165–177, 2002.
[28] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
[29] E. F. Arruda, F. Ourique, and A. Almudevar, "Toward an optimized value iteration algorithm for average cost markov decision processes," in *IEEE Conference on Decision and Control (CDC)*, 2010, pp. 930–934.
[30] C. Sun, E. Stevens-Navarro, and V. W. Wong, "A constrained mdp-based vertical handoff decision algorithm for 4g wireless networks," in *IEEE International Conference on Communications (ICC)*. IEEE, 2008, pp. 2169–2174.
[31] A. Dimakis and J. Walrand, "Sufficient conditions for stability of longest-queue-first scheduling: Second-order properties using fluid limits," *Advances in Applied probability*, vol. 38, no. 2, pp. 505–521, 2006.
[32] P. Levis and D. Culler, "The firecracker protocol," in *SIGOPS European workshop*. ACM, 2004, p. 3.

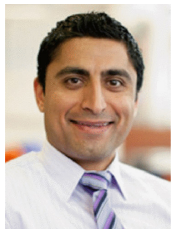


Kai Li (S'09–M'14) received the B.E. degree from Shandong University, China, in 2009, the M.S. degree from The Hong Kong University of Science and Technology, Hong Kong, in 2010, and the Ph.D. degree in Computer Science from The University of New South Wales, Sydney, Australia, in 2014. Currently he is a research scientist and project leader at Real-Time and Embedded Computing Systems Research Centre (CISTER), Portugal. Prior to this, Dr. Li was a postdoctoral research fellow at The SUTD-MIT International Design Centre, The Singapore University of Technology and Design, Singapore (2014-2016). He was a visiting research assistant at ICT Centre, CSIRO, Australia (2012-2013). From 2010 to 2011, he was a research assistant at Mobile Technologies Centre with The Chinese University of Hong Kong. His research interests include vehicular communications and security, resource allocation optimization, Cyber-Physical Systems, Internet of Things (IoT), human sensing systems, sensor networks and UAV networks. Dr. Li serves as the Associate Editor for IEEE Access Journal, the Demo Co-chair for ACM/IEEE IPSN 2018, the TPC member of IEEE Globecom'18, MASS'18, VTC-Spring'18, Globecom'17, VTC'17, and VTC'16.



Wei Ni (M'09-SM'15) received the B.E. and Ph.D. degrees in Electronic Engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. Currently he is a Team Leader at CSIRO, Sydney, Australia, and an adjunct professor at the University of Technology Sydney (UTS). He also holds adjunct positions at the University of New South Wales (UNSW) and Macquarie University (MQ). Prior to this, he was a postdoctoral research fellow at Shanghai Jiaotong University from 2005-2008; Deputy Project Manager at the Bell Labs R&I Center, Alcatel/Alcatel-Lucent from 2005-2008; and Senior Researcher at Devices R&D, Nokia from 2008-2009. His research interests include stochastic optimization, game theory, graph theory, as well as their applications to network and security.

Dr Ni has been serving as Vice Chair of IEEE NSW VTS Chapter and Editor of IEEE Transactions on Wireless Communications since 2018, secretary of IEEE NSW VTS Chapter from 2015 - 2018, Track Chair for VTC-Spring 2017, Track Co-chair for IEEE VTC-Spring 2016, and Publication Chair for BodyNet 2015. He also served as Student Travel Grant Chair for WPMC 2014, a Program Committee Member of CHINACOM 2014, a TPC member of IEEE ICC'14, ICCC'15, EICE'14, and WCNC'10.



Mehran Abolhasan (M'01–SM'11) received the B.E. degree in computer engineering and the Ph.D. degree in telecommunications from the University of Wollongong, Wollongong, Australia, in 1999 and 2003, respectively.He is currently an Associate Professor with and the Deputy Head of the School for Research, School of Computing and Communications, University of Technology Sydney, Ultimo, Australia. He has led several major R&D projects in various areas of wireless communications since 2003, authored more than 60 international publications, and won more than $1 million in research funding. His current research interests are in wireless mesh, fifth-generation cooperative networks, and body-area/sensor networks.

Eduardo Tovar received the Ph.D. degree in electrical and computer engineering from the University of Porto, Porto, Portugal, in 1999. Currently, he is Professor of Industrial Computer Engineering in the Computer Engineering Department, Polytechnic Institute of Porto (ISEP-IPP), where he is also engaged in research on real-time distributed systems, wireless sensor networks, multiprocessor systems, cyber-physical systems and industrial communication systems. He heads the CISTER Research Unit (UI 608), a top ranked ("Excellent") unit of the FCT Portuguese network of research units. Since 1991, he authored or coauthored more than 100 scientific and technical papers. He has been consistently participating in top-rated scientific events as member of the Program Committee, as Program Chair or as General Chair. He is team leader within the EU Seventh Framework ICT Network of Excellence on Cooperating Objects, www.cooperating-objects.eu.