# Optimality of Myopic Scheduling and Whittle Indexability for Energy Harvesting Sensors

[1,2]Fabio Iannello, [2]Osvaldo Simeone, and [1]Umberto Spagnolini
[1]Dipartimento di Elettronica e Informazione, Politecnico di Milano, 20133, Milan, Italy
[2]CWCSPR, New Jersey Institute of Technology, 07102 Newark, New Jersey, USA
Email: {iannello, spagnoli}@elet.polimi.it, osvaldo.simeone@njit.edu

***Abstract*—Consider a single-hop wireless sensor network, where a central node (or fusion center, FC) collects data from a set of $M$ energy harvesting (EH)-capable sensors (or nodes). In each time-slot only a subset of $K \leq M$ nodes can be scheduled by the FC for transmission over $K$ orthogonal communication resources (e.g., frequencies). The scheduling problem is tackled by assuming that the FC has no direct access to the instantaneous states of the nodes' batteries, but it only knows the outcomes of previous transmissions attempts and the statistical properties of the energy harvesting/discharging processes. Based on a simple Markovian modeling of the EH and battery leakage processes, the FC's scheduling problem is formulated as partially observable Markov decision processes (POMDPs) and then cast into a restless multi-armed bandit (RMAB) framework. It is shown that in some special cases, a myopic (or greedy) scheduling policy is optimal, and that such a policy coincides with the so called Whittle index policy.**

## I. Introduction and System Model

Energy harvesting (EH) technologies not only provide a means to mitigate the energy footprint of wireless communications, but also enable new applications, such as the deployment of wireless sensor networks (WSNs), in environments where maintenance is impractical or too costly [1]. EH-capable wireless sensors can collect the energy needed to operate from the surrounding environment (e.g., from solar power). However, unlike battery-powered sensors, EH-devices generally depend on unreliable energy sources, which call for the design of robust and adaptive energy management strategies [2].

In this paper, we consider a single-hop WSN, where a central node, referred to as fusion center (FC), collects data from a set of $n$ EH-capable sensors (or nodes), labeled as $U_1, ..., U_n$, and deployed in its surrounding as shown in Fig. 1. Time is slotted with slots indexed as $t = 1, 2, ....$ In each slot $t$, the FC schedules a subset $\mathcal{U}(t) \subseteq \{U_1, ..., U_n\}$ of $|\mathcal{U}(t)| = K$ nodes for transmission, where each of the $K$ scheduled nodes is allocated an orthogonal transmission resource, e.g., a frequency. Each node has always data to transmit (i.e., it is backlogged) and, when it is scheduled by the FC, it can transmit a packet within the allocated resource only if it has enough energy for transmission as detailed in Sec. I-A. We assume that communications between the scheduled nodes and the FC are free of errors.

The scheduling policies' design is tackled by assuming that the FC has no direct access to the nodes' instantaneous energy availability. Instead, the FC can take scheduling decisions only based on the outcomes of previous transmissions attempts, and on the known statistical properties of the EH and battery leakage processes at each node (see Sec. I-A). The design goal is the maximization of the average number of packets collected in a given time of interest, i.e., the *throughput*.

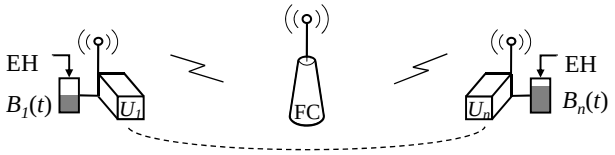


Figure 1. A WSN where a fusion center (FC) collects data from energy-harvesting (EH) sensors.

### A. Markov Formulation

An elementary model that describes the evolution of a node's battery over slots and that accounts for EH and leakage processes, is depicted in Fig. 2. The battery of node $U_i$ is discrete and of capacity one ($C = 1$): at time-slot $t$ it contains an energy $B_i(t) \in \{0, \varepsilon\}$, where $\varepsilon$ represents the unit of energy consumed by a node for transmitting a packet in a slot, which is normalized to one for simplicity, i.e., $\varepsilon = 1$.

At each slot, node $U_i$ can be either scheduled ($U_i \in \mathcal{U}(t)$) or not ($U_i \notin \mathcal{U}(t)$). If it is not scheduled ($U_i \notin \mathcal{U}(t)$), and the battery is full (i.e., $B_i(t) = 1$), the battery gets discharged in the next slot with probability $p_{10}^{(0)}$, while it remains full with probability $p_{11}^{(0)} = 1 - p_{10}^{(0)}$ (see Fig. 2-a)). If node $U_i$ is scheduled ($U_i \in \mathcal{U}(t)$) instead, and $B_i(t) = 1$, it transmits successfully to the FC and its battery in the next slot is either empty or full with probability $p_{10}^{(1)}$ and $p_{11}^{(1)} = 1 - p_{10}^{(1)}$ respectively (see Fig. 2-b)). If $B_i(t) = 0$ the probabilities of harvesting one energy unit when $U_i$ is not scheduled and scheduled are $p_{01}^{(0)}$ and $p_{01}^{(1)}$ respectively, while the probabilities of remaining empty are $p_{00}^{(0)} = 1 - p_{01}^{(0)}$ and $p_{00}^{(1)} = 1 - p_{01}^{(1)}$. The FC knows $p_{xy}^{(u)}$, for $x, y, u \in \{0, 1\}$. We assume that the FC has no direct access to the batteries' states at each slot $t$, i.e., $B_1(t), ..., B_n(t)$. The FC's scheduling decision problem can thus be formalized as a partially observable Markov decision processes (POMDP) [3] and, more specifically, as a restless multiarmed bandit problem (RMAB) [4].

### B. Related Work and Contributions

In this work, we address the FC's scheduling problem, by assuming that: *i*) nodes are symmetric and evolve according

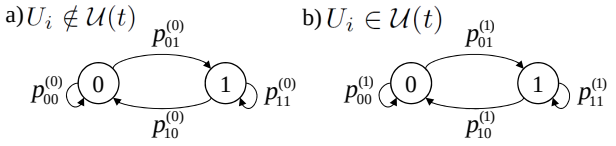


Figure 2. The considered Markov model for the evolution of the battery $B_i(t)$, of capacity $C = 1$, when the node $U_i$: a) is not scheduled in slot $t$ (i.e., $U_i \notin \mathcal{U}(t)$); b) is scheduled in slot $t$ (i.e., $U_i \in \mathcal{U}(t)$).

to the Markov models of Fig. 2, independently of one another given the scheduling command $\mathcal{U}(\cdot)$; *ii*) the transition probabilities for the Markov chains in Fig. 2 satisfy the inequalities

$$p_{11}^{(1)} \leq p_{01}^{(1)} \leq p_{01}^{(0)} \leq p_{11}^{(0)}; \tag{1}$$

and *iii*) the number of nodes in the system is proportional to the transmission resources available, i.e.,

$$n = Km, \text{ for } m \text{ integer.} \tag{2}$$

The conditions in (1) are inspired by physical considerations as follows. The inequality $p_{01}^{(0)} \leq p_{11}^{(0)}$, or equivalently $p_{01}^{(0)} \leq 1 - p_{10}^{(0)}$, implies the assumption that the probability $p_{10}^{(0)}$ that the battery loses one energy unit due to self-discharge only is smaller than the probability $p_{01}^{(0)}$ of harvesting one energy unit. The second inequality $p_{01}^{(1)} \leq p_{01}^{(0)}$ indicates that the probability $p_{01}^{(1)}$ of harvesting one energy unit when the battery is empty and the node is scheduled is no larger than when it is not scheduled (i.e., $p_{01}^{(0)}$). In the relevant setting in which the energy harvesting process is independent of the scheduling decision we have $p_{01}^{(1)} = p_{01}^{(0)}$. The inequality $p_{11}^{(1)} \leq p_{01}^{(1)}$ follows immediately since the battery is of capacity one.

Our main contributions are as follows. For the scheduling problem at hand, we show the optimality of a myopic (or greedy) policy (MP). We then show that for the special case in which $p_{01}^{(0)} = p_{01}^{(1)}$ and $p_{10}^{(0)} = p_{11}^{(1)} = 0$, this optimal MP coincides with the Whittle index policy, which is a generally suboptimal policy for RMAB problems [4]. We remark that the characterizations of optimal policies are usually hard to obtain for general RMABs problems and even the numerical computation is highly complex [5]. Therefore, this paper represents a rare example [4] in which a RMAB is indexable and the generally suboptimal Whittle index policy is optimal.

Our derivations are related to, and inspired by, the works [6], [7], [8], in which a RMAB problem similar to the one considered in this paper, was motivated by cognitive radio applications. In the RMAB in [6], [7], [8], however, the scheduling decision was assumed not to affect the transition probabilities in the Markov chains in Fig. 2 (so that $p_{01}^{(0)} = p_{01}^{(1)}$ and $p_{11}^{(0)} = p_{11}^{(1)}$). However, accounting for the effects of the scheduling decisions on battery's evolution is of critical importance in EH applications, thus motivating our work.

## II. Problem Formulation

We now formalize the scheduling problem at the FC when the goal is the maximization of the average throughput in a finite number of slots $T$ (i.e., in a finite horizon scenario). Extension to infinite horizon scenarios will be briefly discussed in Sec. IV. Let $\mathbf{B}(t) = [B_1(t), ..., B_n(t)]$ be the vector collecting the states of the batteries at slot $t$. At slot $t = 1$, the FC is only aware of the initial probability distribution $\boldsymbol{\omega}(1) = [\omega_1(1), ..., \omega_n(1)]$ of $\mathbf{B}(1)$, whose $i$th entry is $\omega_i(1) = \Pr[B_i(1) = 1]$. The scheduling decision (or action) $\mathcal{U}(1)$ is thus a function of the initial distribution $\boldsymbol{\omega}(1)$ only.

Note that, for any set $\mathcal{U}(t)$ of scheduled nodes at slot $t$, a set of *observations* are made available to the FC at the end of the slot. Specifically, if $B_i(t) = 1$ and $U_i \in \mathcal{U}(t)$, the packet of the scheduled node $U_i$ is received correctly by the FC within slot $t$, and thus the FC learns that $B_i(t) = 1$. Conversely, if $B_i(t) = 0$ and $U_i \in \mathcal{U}(t)$, the packet is not transmitted and the FC realizes that $B_i(t) = 0$. No observations are available for non-scheduled nodes $U_i \notin \mathcal{U}(t)$. Accordingly, we define the observations available for the FC's decision at slot $t + 1$ as $\mathcal{O}(t) = \{B_i(t) : U_i \in \mathcal{U}(t)\}$. At time $t$, the FC thus knows the history of all actions and previous observations along with the initial distribution $\boldsymbol{\omega}(1)$, namely $\mathcal{H}(t) = \{\mathcal{U}(1), ..., \mathcal{U}(t-1), \mathcal{O}(1), ..., \mathcal{O}(t-1), \boldsymbol{\omega}(1)\}$, with $\mathcal{H}(1) \triangleq \{\boldsymbol{\omega}(1)\}$. The scheduling decision $\mathcal{U}(t)$ for $t = 1, 2, ...$ is a function of the history $\mathcal{H}(t)$.

A policy $\pi = [\mathcal{U}^\pi(1), ..., \mathcal{U}^\pi(T)]$ is defined as a collection of functions $\mathcal{U}^\pi(t)$, $t = 1, ..., T$, where each $\mathcal{U}^\pi(t)$ maps the history $\mathcal{H}(t)$ to a subset $\mathcal{U}(t)$ of $K$ nodes. Note that we will refer to $\mathcal{U}^\pi(t)$ as the subset of scheduled nodes, even though, strictly speaking, it is the mapping function defined above.

Let us define the *immediate reward* $R(\mathbf{B},\mathcal{U})$ as the number of packets correctly received by the FC in a slot where the battery states are $\mathbf{B}$ and the scheduled set is $\mathcal{U}$:

$$R(\mathbf{B},\mathcal{U}) = \sum_{i=1}^{n} 1(B_i = 1) 1(U_i \in \mathcal{U}), \tag{3}$$

where $1(A)$ is the indicator function of event $A$, with $1(A) = 1$ if event $A$ is true and zero otherwise. The performance of a policy $\pi$ is measured by the (possibly discounted) throughput $V_1^\pi(\omega(1))$ over the horizon $t = 1, ..., T$, which is given by

$$V_1^\pi(\omega(1)) = \sum_{t=1}^{T} \beta^{t-1} \mathrm{E}^\pi \left[ R(\mathbf{B}(t), \mathcal{U}^\pi(t)) | \boldsymbol{\omega}(1) \right], \tag{4}$$

where $0 \leq \beta \leq 1$ is a discount factor, and the expected value $\mathrm{E}^\pi[\cdot|\boldsymbol{\omega}(1)]$ is with respect to the probability distribution of the Markov process $\mathbf{B}(t)$ as determined by the Markov chains in Fig. 2 and by policy $\pi$, when the initial distribution is $\boldsymbol{\omega}(1)$.

The optimization goal is to find a policy $\pi^* = [\mathcal{U}^*(1), ..., \mathcal{U}^*(T)]$ that maximizes the throughput (4) so that

$$\pi^* = \arg\max_{\pi} V_1^\pi(\boldsymbol{\omega}(1)), \tag{5}$$

$$\text{and } V_1^*(\boldsymbol{\omega}(1)) = V_1^{\pi^*}(\boldsymbol{\omega}(1)) = \max_{\pi} V_1^\pi(\boldsymbol{\omega}(1)) \tag{6}$$

### A. Formulation as Belief MDP and RMAB

Problem (5)-(6) is a POMDP since the controller (i.e., the FC) has only partial information about the current state of the system $\mathbf{B}(t)$ through the observations $\mathcal{O}(t)$. We now reformulate (5)-(6), without loss of optimality, in an equivalent

MDP with full state knowledge, referred to as *belief MDP*. To this end, we start by noting that, while decision $\mathcal{U}^\pi(t)$ at time $t$ depends in general on the entire past history $\mathcal{H}(t)$, it is well-known that a sufficient statistics for the optimization problem (5)-(6) is given by the conditional probability distribution $\boldsymbol{\omega}(t)$ of $\mathbf{B}(t)$ conditioned on the history $\mathcal{H}(t)$ [9]. This conditional probability is referred to as *belief* and it is given by vector $\boldsymbol{\omega}(t)$ with $i$th entry $\omega_i(t) = \Pr[B_i(t) = 1|\mathcal{H}(t)]$, for $i = 1, ..., n$. An optimal decision $\mathcal{U}^*(t)$ at each $t$th slot can thus be found as function of the belief $\boldsymbol{\omega}(t)$ only. Note that, the belief vector $\boldsymbol{\omega}(t)$ is known by the FC. Therefore, a policy $\pi$ can be equivalently defined by a collection of functions $\mathcal{U}^\pi(t)$ that map the current state $\boldsymbol{\omega}(t)$ (instead of the whole history $\mathcal{H}(t)$) into the set of the $K$ scheduled nodes.

For the belief MDP at hand we need to find the transition probabilities over the belief state $\boldsymbol{\omega}(t)$ and the expression of the throughput (4) as a function of $\boldsymbol{\omega}(t)$ instead of $\mathbf{B}(t)$.

**Transition probabilities:** Since node batteries evolve independently given the scheduling decision, also the beliefs $\omega_i(t+1)$ do. The probability that the next slot's belief (or state) is $\boldsymbol{\omega}(t+1) = \boldsymbol{\omega}' = [\omega_1', ..., \omega_n']$, given decision $\mathcal{U}(t) = \mathcal{U}$ and belief $\boldsymbol{\omega}(t) = \boldsymbol{\omega} = [\omega_1, ..., \omega_n]$ is

$$\begin{aligned} p^{(\mathcal{U})}_{\boldsymbol{\omega}\boldsymbol{\omega}'} &= \Pr[\boldsymbol{\omega}(t+1) = \boldsymbol{\omega}'|\boldsymbol{\omega}(t) = \boldsymbol{\omega}, \mathcal{U}(t) = \mathcal{U}] \\ &= \prod_{i=1}^{n} \Pr[\omega_i(t+1) = \omega_i'|\omega_i(t) = \omega_i, \mathcal{U}(t) = \mathcal{U}] \end{aligned} \quad (7)$$

while the transition probabilities $\Pr[\omega_i(t+1) = \omega_i'|\omega_i(t) = \omega_i, \mathcal{U}(t) = \mathcal{U}]$ of the belief $\omega_i(t)$ of node $U_i$ are given by

$$\omega_i(t+1) = \begin{cases} p_{11}^{(1)} & \text{w.p. } \omega_i(t) & \text{if } U_i \in \mathcal{U}(t) \\ p_{01}^{(1)} & \text{w.p. } (1-\omega_i(t)) & \text{if } U_i \in \mathcal{U}(t) \\ \tau_0^{(1)}(\omega_i(t)) & \text{w.p. } 1 & \text{if } U_i \notin \mathcal{U}(t) \end{cases}. \quad (8)$$

In (8), the first two lines reflect the fact that, when node $U_i$ is scheduled ($U_i \in \mathcal{U}(t)$) it has enough energy to transmit with probability (w.p.) $\omega_i(t)$, and thus, from Fig. 2-b), the probability that its battery is full in the next slot, i.e., the belief $\omega_i(t+1)$, is $p_{11}^{(1)}$; similarly, with probability $(1-\omega_i(t))$ the scheduled node $U_i$ does not have energy and hence, from Fig. 2-a), the new belief is $p_{01}^{(1)}$. The last line in (8) states that, if node $U_i$ is not scheduled (i.e., $U_i \notin \mathcal{U}(t)$), then its belief in the next slot can be calculated through a function

$$\begin{aligned} \tau_0^{(1)}(\omega) &= \Pr[B_i(t+1) = 1|\omega_i(t) = \omega, U_i \notin \mathcal{U}(t)] \\ &= \omega p_{11}^{(0)} + (1-\omega)p_{01}^{(0)} = \omega\delta_0 + p_{01}^{(0)}, \end{aligned} \quad (9)$$

where $\delta_0 = p_{11}^{(0)} - p_{01}^{(0)} \geq 0$ due to inequalities (1). Eq. (9) follows from Fig. 2-a), since the next slot's belief is $p_{11}^{(0)}$ if $B_i(t) = 1$ (w.p. $\omega$) or $p_{01}^{(0)}$ if $B_i(t) = 0$ (w.p. $(1-\omega)$). For convenience of notation, we also define the vector

$$\boldsymbol{\tau}_0^{(1)}(\omega_1, ..., \omega_K) = [\tau_0^{(1)}(\omega_1), ..., \tau_0^{(1)}(\omega_K)]. \quad (10)$$

If conditions (1) hold, function (9) has the following properties

$$p_{11}^{(1)} \leq p_{01}^{(1)} \leq \tau_0^{(1)}(\omega), \textit{ for all } \omega \in [0,1]; \quad (11)$$

$$\tau_0^{(1)}(\omega) \leq \tau_0^{(1)}(\omega'), \textit{ for all } \omega \leq \omega' \textit{ with } \omega, \omega' \in [0,1]. \quad (12)$$

Inequalities (11) guarantee that the belief of a non-scheduled node is always larger than that of a scheduled node. Inequality (12) says that the belief ordering of two non-scheduled nodes is maintained across a slot. Inequalities (11)-(12) play a crucial role in the analysis below.

**Throughput**: We now observe that the throughput (4) can be written in terms of an immediate reward

$$R(\boldsymbol{\omega}, \mathcal{U}) = \sum_{i=1}^{n} \omega_i 1(U_i \in \mathcal{U}), \quad (13)$$

which depends only on the belief $\boldsymbol{\omega}$ and the scheduling decision $\mathcal{U}$. A scheduled node $U_i$ thus provides the FC with an immediate reward equal to its belief $\omega_i$. From (13) and (3), the throughput (4) can be written as

$$V_1^\pi(\boldsymbol{\omega}(1)) = \sum_{t=1}^{T} \beta^{t-1} \mathrm{E}^\pi\left[R\left(\boldsymbol{\omega}(t), \mathcal{U}^\pi(t)\right)|\boldsymbol{\omega}(1)\right]. \quad (14)$$

In (14), expectation $\mathrm{E}^\pi[\cdot|\boldsymbol{\omega}(1)]$ is with respect to the distribution of the Markov process $\boldsymbol{\omega}(t)$ determined by probabilities (8) and by policy $\pi$, given the initial belief $\boldsymbol{\omega}(1)$. The optimal policy and the optimal throughput are defined as in (5)-(6).

### *B. Optimality Equations*

In this section, we introduce the standard dynamic programming (DP) optimality conditions that characterize an optimal policy $\pi^*$ in (5). To start with, let the probability that the $K$ scheduled nodes have energies $b_1, ..., b_K \in \{0, 1\}$ be

$$q(b_1, ..., b_K, \omega_1, .., \omega_K) = \prod_{i=1}^{K} \omega_i^{b_i}(1-\omega_i)^{1-b_i}. \quad (15)$$

Let $V_t^\pi(\boldsymbol{\omega})$ be the throughput in the horizon $\{t, ..., T\}$, then by exploiting (7), (8), (10), (13) and (15), the throughput (14) can be written through the following recursion (see e.g., [10])

$$V_T^\pi(\boldsymbol{\omega}) = R(\boldsymbol{\omega}, \mathcal{U}^\pi(T)) = \sum_{i \in \mathcal{U}^\pi(T)} \omega_i \quad (16)$$

$$V_t^\pi(\boldsymbol{\omega}) = R(\boldsymbol{\omega}, \mathcal{U}^\pi(t)) + \beta \sum_{\boldsymbol{\omega}'} V_{t+1}^\pi(\boldsymbol{\omega}') p^{(\mathcal{U}^\pi(t))}_{\boldsymbol{\omega}\boldsymbol{\omega}'} \quad (17)$$

$$\sum_{i \in \mathcal{U}^\pi(t)} \omega_i + \beta \sum_{b_1, ..., b_K \in \{0,1\}} q(b_1, ..., b_K, \boldsymbol{\omega}_{\mathcal{U}^\pi(t)}) \cdot$$

$$V_{t+1}^\pi\left(\gamma(b_1), ..., \gamma(b_K), \boldsymbol{\tau}_0^{(1)}(\boldsymbol{\omega}_{(\mathcal{U}^\pi(t))^c})\right), \quad \text{for } t \in \{1, ..., T\},$$

where the notation $\boldsymbol{\omega}_{\mathcal{S}}$ indicates a vector that contains the belief of the nodes in set $\mathcal{S}$, $(\mathcal{U}^\pi(t))^c = \{U_1, ..., U_n\} \setminus \mathcal{U}^\pi(t)$ and $\gamma(b) = p_{01}^{(1)}(1-b) + p_{11}^{(1)}b$, with $\gamma(b) = p_{01}^{(1)}$ if $b = 0$ and $\gamma(b) = p_{11}^{(1)}$ if $b = 1$. The optimality conditions can then be expressed through functions (16)-(17) as (see [10])

$$V_t^*(\boldsymbol{\omega}) = \max_{\mathcal{U}^\pi(t)} \{V_t^\pi(\boldsymbol{\omega})\}, \text{ for } t \in \{1, ..., T\}, \quad (18)$$

while an optimal policy $\pi^* = [\mathcal{U}^*(1), ..., \mathcal{U}^*(T)]$ (5) is such that $\mathcal{U}^*(t)$ attains the maximum in (18) for all $t \in \{1, ..., T\}$.

Note that (16)-(17) are the conventional recursive DP equations for a policy $\pi$, in which one averages over the distribution $p^{(\mathcal{U})}_{\boldsymbol{\omega}\boldsymbol{\omega}'}$ (7) of the next-slot's belief given the current belief

and scheduling decision $\mathcal{U}$. Lastly, recall that due to (8), in (17) the beliefs of all the nodes not in $\mathcal{U}$, i.e., in $\mathcal{U}^c$, evolve deterministically as $\tau_0^{(1)}(\boldsymbol{\omega}_{\mathcal{U}^c})$. Instead, the beliefs of scheduled nodes in $\mathcal{U}$ can be either equal to $p_{11}^{(1)}$ or $p_{01}^{(1)}$ with probability $\omega_i$ and $(1-\omega_i)$ respectively.

## III. Myopic Scheduling Policy

Here, we first define the myopic policy (MP) and show that, under conditions (1)-(2) it is a round-robin (RR) strategy. Then we prove its optimality for problem (5).

The MP $\pi^{MP} = \{\mathcal{U}^{MP}(1), ..., \mathcal{U}^{MP}(T)\}$ is a greedy policy that in each $t$th slot schedules the $K$ nodes with the largest beliefs so as to maximize the immediate reward (13) as

$$\mathcal{U}^{MP}(t) = \arg\max_{\mathcal{U}} R(\boldsymbol{\omega}(t), \mathcal{U}) = \arg\max_{\mathcal{U}} \sum_{i\in\mathcal{U}} \omega_i(t). \quad (19)$$

**Proposition 1.** If conditions (1)-(2) hold, the MP $\pi^{MP}$ (19), given initial belief $\boldsymbol{\omega}'(1)$, is a RR policy that operates as follows: **1)** sort vector $\boldsymbol{\omega}'(1)$ in a decreasing order to obtain $\boldsymbol{\omega}(1) = [\omega_1(1), ..., \omega_n(1)]$ such that $\omega_1(1) \geq ... \geq \omega_n(1)$. Renumber the nodes so that $U_i$ has belief $\omega_i(1)$; **2)** divide the nodes into $m$ groups of $K$ nodes each, so that the $g$th group $\mathcal{G}_g$, $g = 1, ..., m$, contains all nodes $U_i$ such that $g = \left\lfloor \frac{i-1}{K} \right\rfloor + 1$, i.e., $\mathcal{G}_1 = \{U_1, ..., U_K\}$, $\mathcal{G}_2 = \{U_{K+1}, ..., U_{2K}\}$, and so on; **3)** schedule the groups in a RR (periodic) fashion with period $m$ slots, so that groups $\mathcal{G}_1, ..., \mathcal{G}_m, \mathcal{G}_1, ...$ are respectively scheduled at slot $t = 1, ..., m, m+1, ...$ and so on.

*Proof:* According to (19), the first scheduled set of nodes is $\mathcal{U}^{\pi^{MP}}(1) = \mathcal{G}_1 = \{U_1, U_2, ..., U_K\}$. Nodes' beliefs are then updated through (8). Recalling (11), scheduled nodes in $\mathcal{G}_1$ have their belief updated to either $p_{11}^{(1)}$ or $p_{01}^{(1)}$, which are both smaller than the belief of any non-scheduled node in $\{U_1, ..., U_n\} \setminus \mathcal{G}_1$. Moreover, the ordering of non-scheduled nodes' beliefs is preserved due to (12). Consequently, the second scheduled group is $\mathcal{U}^{\pi^{MP}}(2) = \mathcal{G}_2$, the third is $\mathcal{U}^{\pi^{MP}}(3) = \mathcal{G}_3$, and so on. This proves that the MP, upon an initial ordering of the beliefs, is RR. ∎

The throughput (16)-(17) of a RR policy $\pi^{RR}$ that operates according to steps **2)** and **3)** of Proposition 1, can be expressed recursively through functions $V_t(\boldsymbol{\omega})$ as

$$\tilde{V}_T(\boldsymbol{\omega}) = \sum_{i=1}^{K} \omega_i \quad (20)$$

$$\tilde{V}_t(\boldsymbol{\omega}) = \sum_{i=1}^{K} \omega_i + \beta \sum_{b_1,...,b_K \in \{0,1\}} q(b_1, ..., b_K, \omega_1, ..., \omega_K) \cdot \tilde{V}_{t+1}\left(\boldsymbol{\tau}_0^{(1)}(\boldsymbol{\omega}_{(\mathcal{U}^\pi(t))^c}),\ p_{01}^{(1)}\mathbf{1}_{K-\sum_{i=1}^K b_i},\ p_{11}^{(1)}\mathbf{1}_{K-\sum_{i=1}^K b_i}\right), \quad (21)$$

$$\text{for } t \in \{1, ..., T-1\},$$

in which, the policy $\pi^{RR}$ in each slot: *i*) schedules the $K$ nodes whose beliefs are in the first $K$ positions of the argument $\boldsymbol{\omega}$ of $\tilde{V}_T(\boldsymbol{\omega})$; *ii*) the argument $\boldsymbol{\omega}'$ for the next slot is updated (through (8)) so that the beliefs of the scheduled nodes are (i.e., in the set $\mathcal{U}^\pi(t)$) decreasingly ordered and put at the $K$ rightmost positions of $\boldsymbol{\omega}'$, while the ordering the other beliefs (i.e., in the set $(\mathcal{U}^\pi(t))^c$) are preserved so that $\boldsymbol{\omega}' = [\boldsymbol{\tau}_0^{(1)}(\boldsymbol{\omega}_{(\mathcal{U}^\pi(t))^c}),\ p_{01}^{(1)}\mathbf{1}_{K-\sum_{i=1}^K b_i},\ p_{11}^{(1)}\mathbf{1}_{K-\sum_{i=1}^K b_i}]$. Note that, when the initial belief $\boldsymbol{\omega}$ is ordered so that $\omega_1 \geq ... \geq \omega_n$, then $\tilde{V}_T(\boldsymbol{\omega}) = V_t^{MP}(\boldsymbol{\omega})$.

### A. Optimality of the Myopic Policy

Here, we prove the optimality of the MP $\pi^{MP}$ in (19).

**Theorem 2.** If conditions (1) and (2) hold, then the MP $\pi^{MP}$ is optimal for problem (6) in the sense that $\pi^{MP} = \pi^*$ (with $\pi^*$ in (5)) and $V_1^{MP}(\boldsymbol{\omega}) = V_1^*(\boldsymbol{\omega})$.

*Sketch of the Proof*: We show that the MP policy $\pi^{MP}$ satisfies the DP optimality conditions (18) by backward induction, similar to the approach in [6]. Specifically, the basis of the induction is the last slot $T$. In fact, from (19), $\mathcal{U}^{MP}(T)$ clearly attains the maximum in (18). Now, suppose that the MP is optimal at slot $t+1, ..., T$ (i.e., it satisfies (18)). Then, to prove that the MP is optimal at all slots $t \in \{1, ..., T\}$, it is sufficient to show that $\tilde{V}_t(\boldsymbol{\omega}_{\mathcal{S}}, \boldsymbol{\omega}_{\mathcal{S}^c}) \leq V_t^{MP}(\boldsymbol{\omega}_{\mathcal{S}}, \boldsymbol{\omega}_{\mathcal{S}^c}) = \tilde{V}_t(\omega_1, \omega_2, ..., \omega_n)$, for all $\omega_1 \geq \omega_2 \geq ... \geq \omega_n$ and all sets $\mathcal{S} \subseteq \{1, ..., n\}$ of $K$ elements, with the $n-K$ elements in $\boldsymbol{\omega}_{\mathcal{S}^c}$ decreasingly ordered. In fact, since the MP is optimal from $t+1$ on, it is sufficient to show that scheduling $K$ nodes with arbitrary beliefs at slot $t$ and then following the MP from slot $t+1$ onward, is no better than following the MP immediately at slot $t$. The performance of the former policy is given by $\tilde{V}_t(\boldsymbol{\omega}_{\mathcal{S}}, \boldsymbol{\omega}_{\mathcal{S}^c})$, since for any set $\mathcal{S}$, it represents the throughput of a policy that schedules the $K$ nodes with beliefs $\boldsymbol{\omega}_{\mathcal{S}}$ at slot $t$, and then operates as the MP from $t+1$ onward, since beliefs $\boldsymbol{\omega}_{\mathcal{S}^c}$ are decreasingly ordered (see (20)-(21)). The MP's performance is instead given by the $V_t^{MP}(\boldsymbol{\omega}_{\mathcal{S}}, \boldsymbol{\omega}_{\mathcal{S}^c}) = \tilde{V}_t(\omega_1, \omega_2, ..., \omega_n)$. Inequality $\tilde{V}_t(\boldsymbol{\omega}_{\mathcal{S}}, \boldsymbol{\omega}_{\mathcal{S}^c}) \leq V_t^{MP}(\boldsymbol{\omega}_{\mathcal{S}}, \boldsymbol{\omega}_{\mathcal{S}^c})$ can be shown to hold under conditions (1)-(2), thanks to the RR structure of the MP policy. The derivations are omitted and can be found in [11].

## IV. Infinite Horizon Scenario and Optimality of the Whittle Index Policy

We now discuss the extension of problem (14) to the infinite-horizon case and present the Whittle index policy [4].

**Infinite Horizon Scenario**. The throughput in the infinite-horizon case under policy $\pi$ and discount factor $0 \leq \beta < 1$, and its optimal value, are given by

$$V^\pi(\boldsymbol{\omega}(1)) = \sum_{t=1}^{\infty} \beta^{t-1} \mathrm{E}^\pi \left[ R\left(\boldsymbol{\omega}(t), \mathcal{U}^\pi(t)\right) | \boldsymbol{\omega}(1) \right], \text{ and } \quad (22)$$

$$V^*(\boldsymbol{\omega}(1)) = \max_\pi V^\pi(\boldsymbol{\omega}(1)), \quad (23)$$

where the optimal policy is $\pi^* = \arg\max_\pi V^\pi(\boldsymbol{\omega}(1))$. From standard DP theory, the optimal policy $\pi^*$ is *stationary*, so that $\pi^*$ is such that the optimal scheduling decision $\mathcal{U}^{\pi^*}(t)$ is a function of the current state $\omega(t)$ only independently of slot $t$ [10]. Following the same reasoning as in [6, Theorem 3],

it is easy to show that the optimality of MP for the finite-horizon setting implies the optimality also for the infinite-horizon scenario. Moreover, it can be proved that $V^*(\boldsymbol{\omega}(1)) = \lim_{T\to\infty} V_1^*(\boldsymbol{\omega}(1))$, where $V_1^*(\boldsymbol{\omega}(1))$ is (18).

**Whittle Index Policy**. The Whittle index policy assigns a numerical value $W(\omega_i)$ (the *Whittle index*) to each state $\omega_i$ of node $U_i$ to measure how rewarding it is to schedule $U_i$ in the current slot. The Whittle index is calculated independently for each node and the $K$ nodes with the largest index are scheduled in each slot. The Whittle index policy is thus not generally optimal for RMAB problems. However, the following results hold for the RMAB at hand.

**Proposition 3.** The Whittle index policy is optimal for problem (23) under the special case of conditions (1) given by

$$0 = p_{11}^{(1)} \leq p_{01}^{(1)} = p_{01}^{(0)} = p_{01} \leq p_{11}^{(0)} = 1. \tag{24}$$

We emphasize that, our results provide a rare example [4] in which, as in [8], not only indexability is established, but also the Whittle index is obtained in closed form and the Whittle policy proved to be optimal. Below, we provide a sketch of the proof of Proposition 3 by studying the corresponding restless single-armed bandit (RSAB) model [4]. We show that the latter is indexable and that the corresponding Whittle index is increasing in $\omega$. Therefore, since the Whittle index policy selects the $K$ arms with the largest index at each slot, the Whittle policy coincides with the MP, and thus it is optimal for the RMAB at hand.

### A. Proof of Proposition 3

To prove Proposition 3, we start by introducing the RSAB model at hand and then study its indexability [4].

The Whittle index is based on the concept of *subsidy for passivity*, whereby the FC is given a subsidy $m \in \mathbb{R}$ when the arm is not scheduled. At each slot $t$, the FC, based on the state $\omega(t)$ of the arm, can decide to activate (or schedule) it, i.e., to set $u(t) = 1$, obtaining an immediate reward $R_m(\omega(t), 1) = \omega(t)$. If, instead, the arm is kept passive, i.e., $u(t) = 0$, a reward $R_m(\omega(t), 0) = m$ equal to the subsidy is accrued. The state $\omega(t)$ evolves through (8), which under (24) and adapted to the simplified notation used here becomes

$$\omega(t+1) = \begin{cases} 0 & \text{w.p. } \omega(t) & \text{if } u(t) = 1 \\ p_{01} & \text{w.p. } (1-\omega(t)) & \text{if } u(t) = 1 \\ \tau_0^{(1)}(\omega(t)) & \text{w.p. } 1 & \text{if } u(t) = 0 \end{cases}. \tag{25}$$

The throughput, given policy $\pi = \{u^\pi(1), u^\pi(2), ...\}$ and initial belief $\omega(1)$, is

$$V_m^\pi(\omega(1)) = \sum_{t=1}^{\infty} \beta^{t-1} \mathrm{E}^\pi \left[ R_m(\omega(t), u^\pi(t)) | \omega(1) \right]. \tag{26}$$

The optimal throughput is $V_m^*(\omega(1)) = \max_\pi V_m^\pi(\omega(1))$, while the optimal policy $\pi^* = \arg\max_\pi V_m^\pi(\omega(1))$ is stationary in the sense that the optimal decisions $u_m^*(\omega) \in \{0, 1\}$ are functions of the belief $\omega$ only, independently of slot $t$ [8]. Removing the slot index from the initial belief, the optimal throughput $V_m^*(\omega)$ and the optimal decision $u_m^*(\omega)$ satisfy the following DP optimality equations for the infinite-horizon scenario (see [8])

$$V_m^*(\omega) = \max_{u\in\{0,1\}} \{V_m(\omega|u)\}, \tag{27}$$

$$\text{and } u_m^*(\omega) = \arg\max_{u\in\{0,1\}} \{V_m(\omega|u)\}. \tag{28}$$

In (27)-(28) we defined $V_m(\omega|u)$, $u \in \{0, 1\}$, as the throughput (26) of a policy that takes action $u$ at the current slot and then uses the optimal policy $u_m^*(\omega)$ onward, we have

$$V_m(\omega|0) = m + \beta V_m^*(\tau_0^{(1)}(\omega)), \text{ and} \tag{29}$$

$$V_m(\omega|1) = \omega + \beta \left[\omega V_m^*(0) + (1-\omega) V_m^*(p_{01})\right]. \tag{30}$$

*1) Indexability and Whittle Index:* We use the notation of [8] to define indexability and Whittle index for the RSAB at hand. We first define the so called *passive set*

$$\mathcal{P}(m) = \{\omega\colon 0 \leq \omega \leq 1 \text{ and } u_m^*(\omega) = 0\}, \tag{31}$$

as the set that contains all the beliefs $\omega$ for which the passive action is optimal (i.e., all $0 \leq \omega \leq 1$ such that $V_m(\omega|0) \geq V_m(\omega|1)$, see (29)-(30)) under the given subsidy for passivity $m \in \mathbb{R}$. The RMAB at hand is said to be *indexable* if the passive set $\mathcal{P}(m)$, for the associated RSAB problem, is monotonically increasing as $m$ increases within the interval $(-\infty, +\infty)$, in the sense that $\mathcal{P}(m') \subseteq \mathcal{P}(m)$ if $m' \leq m$ and $\mathcal{P}(-\infty) = \emptyset$ and $\mathcal{P}(+\infty) = [0, 1]$.

If the RMAB is indexable, the Whittle index $W(\omega)$ for each arm with state $\omega$ is the infimum subsidy $m$ such that it is optimal to make the arm passive. Equivalently, the Whittle index $W(\omega)$ is the infimum subsidy $m$ that makes passive and active actions equally rewarding, i.e.,

$$W(\omega) = \inf\{m\colon V_m(\omega|0) = V_m(\omega|1)\}. \tag{32}$$

*2) Optimality of the Threshold Policy for the RSAB:* We now show that the RSAB's optimal policy $u_m^*(\omega)$ is a threshold policy over the belief $\omega$. This is a crucial step in our proof of indexability given in Sec. IV-A3.

**Proposition 4.** The optimal policy $u_m^*(\omega)$ in (28) is given by

$$u_m^*(\omega) = \begin{cases} 1, & \text{if } \omega > \omega^*(m) \\ 0, & \text{if } \omega \leq \omega^*(m) \end{cases}, \tag{33}$$

where $\omega^*(m) \in \mathbb{R}$ is the optimal threshold for a given subsidy $m \in \mathbb{R}$. Clearly $u_m^*(\omega) = 1$ if $m < 0$ and $u_m^*(\omega) = 0$ if $m \geq 1$. The optimal threshold $\omega^*(m)$ is $0 \leq \omega^*(m) \leq 1$ if $0 \leq m < 1$.

*Sketch of the proof*: The proof is based on the following properties: *i*) function $V_m(\omega|1)$ in (30) is linear over the belief $\omega$; *ii*) function $V_m(\omega|0) = m + \beta V_m^*(\tau_0^{(1)}(\omega))$ in (29) is convex over $\omega$. The convexity of $V_m^*(\omega)$ is a general property of POMDPs (see [8], [9]). Moreover, a set of inequalities among functions $V_m(0|1)$, $V_m(1|1)$, $V_m(0|0)$ and $V_m(1|0)$, can be derived for different values of the subsidy $m$ as graphically shown in Fig. 3. From Fig. 3 and the properties *i*) and *ii*) above the optimality of the threshold policy in (33) can be easily inferred. The full derivations can be found in [11].

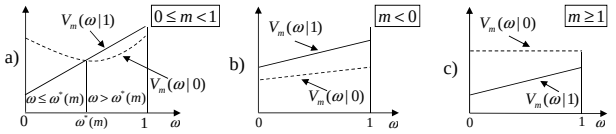


Figure 3. Illustration of the optimality of a threshold policy for different values of the subsidy for passivity $m$: a) $0 \le m < 1$; b) $m < 0$; c) $m \ge 1$.

*3) Proof of Indexability and Whittle Index:* Similarly to [8], a crucial step in our proof of indexability of the RMAB at hand is the derivation of the closed-form expression of the throughput $V_m^*(\omega)$ in (27) (see Appendix A for a sketch of the derivation and [11] for more details). This enables to prove that the passive set $\mathcal{P}(m)$ (see (31)) is monotonically increasing with the subsidy $m$ as discussed in Sec. IV-A1.

**Theorem 5.** The RMAB at hand is indexable.

*Sketch of the proof*: . Following the discussion in Sec. IV-A1, to prove indexability it is sufficient to show that the threshold $\omega^*(m)$ is monotonically increasing with the subsidy $m$, for $0 \le m < 1$. In fact, from Proposition 4 the passive set (31) for $m < 0$ is $\mathcal{P}(m) = \emptyset$, while for $m \ge 1$ is $\mathcal{P}(m) = [0, 1]$. We thus need to prove the monotonicity of $\omega^*(m)$ for $0 \le m < 1$, which has been shown to hold in [8, Lemma 9] if $\frac{dV_m(\omega|1)}{dm}\Big|_{\omega=\omega^*(m)} < \frac{dV_m(\omega|0)}{dm}\Big|_{\omega=\omega^*(m)}$. From (29)-(30) and by exploiting the closed-form expression of the throughput $V_m^*(\omega)$ (see Appendix A), it can be shown that the latter inequality holds (see the full derivation in [11]).

Finally, the Whittle index (32) can be derived in closed-form (see [11] for details), and also shown being an increasing function of $\omega$, thus concluding the proof of Proposition 3.

## V. Conclusions

In this paper, we have considered a scheduling problem with applications to energy harvesting (EH) networks, where a fusion center (FC) schedules a set of wireless sensors to acquire their measurements. By modeling the EH and battery leakage processes through simple Markov models, the FC's scheduling problem is formulated as a partially observable Markov decision process (POMDP), and cast into a restless multi-armed bandit (RMAB) problem. Under the assumption that node batteries are of capacity one, a myopic (or greedy) policy (MP) that operates in the space of the a posteriori probabilities (beliefs) of the battery levels is proved to be optimal for both finite horizon and infinite-horizon throughput criteria. Finally, we have established that the RMAB problem at hand is indexable and proved that the Whittle index policy is equivalent to the MP and thus is optimal.

## References


[1] J. A. Paradiso, T. Starner, "Energy scavenging for mobile and wireless electronics," *IEEE Perv. Comput. Mag.*, vol. 4, no. 1, pp. 18-27, Jan.-Mar. 2005.

[2] C. K. Ho, P. D. Khoa, P. C. Ming, "Markovian models for harvested energy in wireless communications," in *Proc. IEEE ICCS*, Singapore, pp. 311-315, Nov. 2010.

[3] G. E. Monahan, "A survey of partially observable Markov decision processes: Theory, models, and algorithm," *Manag. Sci.*, vol. 28, no. 1, pp. 1-16, 1982.

[4] J. Gittins, K, Glazerbrook, R. Weber, *Multi-armed bandit allocation indices*. West Sussex, UK, Wiley, 2011.

[5] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Math. Oper. Res.*, vol. 24, pp. 293-305, May 1999.

[6] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, No. 9, pp. 4040-4050, Sept. 2009.

[7] S. H. A. Ahmad, L. Mingyan, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Proc. 47th Ann. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, pp. 1361-1368, Sept. 2009.

[8] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547-5567, Nov. 2010.

[9] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artif. Intell.*, vol. 101, pp. 99-134, May 1998.

[10] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, Wiley, 2005.

[11] F. Iannello, O. Simeone and U. Spagnolini, "On the optimal scheduling of independent, symmetric and time-sensitive tasks," submitted, available at http://arxiv.org/pdf/1112.1229v1.


## Appendix A
## Closed-Form Expression of the Throughput $V_m^*(\omega)$

Since the threshold policy (33) is optimal, the expression of the optimal throughput $V_m^*(\omega)$ in (27) can be found in closed-form following an approach similar to [8]. To start with, let $\tau_0^{(k)}(\omega)$ be a function that gives the belief of a node that is not scheduled for $k$ consecutive slots when its initial belief is $\omega$. Function $\tau_0^{(k)}(\omega)$ can be obtained by applying (9) recursively to itself as $\tau_0^{(k)}(\omega) = \tau_0^{(1)}(\tau_0^{(k-1)}(\omega))$, for all $k \ge 1$, with $\tau_0^{(0)}(\omega) = \omega$. Under conditions (24) we have $\tau_0^{(k)}(\omega) = 1 - (1 - p_{01})^k(1 - \omega)$, which is a monotonically increasing function of $k$, so that $\tau_0^{(k)}(\omega) \ge \tau_0^{(i)}(\omega)$ for any $k \ge i$. Based on such monotonicity, we can define the average number $L(\omega, \omega')$ of slots it takes for the belief to become larger than $\omega'$ when starting from $\omega$ while the arm is kept passive, as $L(\omega, \omega') = \min\left\{k\text{: } \tau_0^{(k)}(\omega) > \omega'\right\}$. According to Proposition 4, the optimal policy $u_m^*(\omega)$ keeps the arm passive for $L(\omega, \omega^*(m))$ slots (i.e., as long as the current belief is $\omega \le \omega^*(m)$) during which a reward $R_m(\omega, 0) = m$ is accrued in each slot. This leads to a total reward within the passivity time given by the following geometric series $\sum_{k=0}^{L(\omega,\omega^*(m))-1} \beta^k m = \frac{1-\beta^{L(\omega,\omega^*(m))}}{1-\beta} m$. After $L(\omega, \omega^*(m))$ slots of passivity, the belief becomes larger than the threshold $\omega^*(m)$ and the arm is activated and the contribution becomes $\beta^{L(\omega,\omega^*(m))} V_m(\tau_0^{(L(\omega,\omega^*(m)))}(\omega)|1)$ since $V^*(\omega) = V(\omega|1)$ when $\omega > \omega^*(m)$. Therefore, the throughput can be written as $V_m^*(\omega) = \frac{1-\beta^{L(\omega,\omega^*(m))}}{1-\beta} m + \beta^{L(\omega,\omega^*(m))} V_m(\tau_0^{(L(\omega,\omega^*(m)))}(\omega)|1)$. The last step to obtain $V_m^*(\omega)$, is to explicitly calculate $V_m(\omega|1)$. However, from (30), evaluating $V_m(\omega|1)$ only requires $V_m^*(0)$ and $V_m^*(p_{01})$, which can be promptly calculated by plugging (30) into $V_m^*(\omega)$ given above, and evaluating $V_m^*(\omega)$ for $\omega = 0$ and $\omega = p_{01}$. We thus get a linear system of two equations in the two unknowns $V_m^*(0)$ and $V_m^*(p_{01})$, which can be easily solved (see [11] ).