Vacation policy optimization with application to IEEE 802.16e power saving mechanism

Amar Prakash Azad, Sara Alouf, Eitan Altman, Vivek Borkar, Georgios
Paschos

▶ **To cite this version:**

HAL Id: hal-00640988
https://hal.inria.fr/hal-00640988

Submitted on 12 Jul 2019

Vacation Policy Optimization with Application to IEEE 802.16e Power Saving Mechanism[*]

Amar P. Azad,[1] Sara Alouf,[1] Eitan Altman,[1] Vivek Borkar,[2] and Georgios Paschos[3]
[1] INRIA, Sophia Antipolis, France
[2] STCS, TIFR, Mumbai, India
[3] CERTH, Greece

Abstract

Much research has been devoted to optimizing the power saving mechanism in wireless mobile devices. Recent advances in wireless radio technology facilitate the implementation of various possible sleep policies. One basic question that arises is: which policy performs best under a certain condition? Furthermore, what are the optimal parameters for a given policy? To answer these questions, we formulate an optimization problem, which entails cost minimization for a given parameterized policy and selection of the best policy among a class. We propose a cost function which captures the inherent tradeoff of delay and energy saving. This takes into account the cost of response time due to the extra sleep, the energy saving during the sleep, and the cost for periodic waking up (for listening). As an application, we consider IEEE 802.16e's power saving mechanism. We study various practical policies and check their performance. We show that the constant duration policy is optimal for Poisson inactivity periods, but not for hyper-exponentially distributed inactivity periods. In the policy where vacations are i.i.d. exponential random variables, we derive analytically the optimal control as a function of the expected inactivity period. This result holds for general inactivity periods. Our framework allows us to compare the performance of several optimal and suboptimal practical policies with that of the IEEE 802.16e standard.

1 Introduction

A wireless device using contemporary radio technology can obtain great energy benefits by shutting off the transceiver in the absence of communications—a state that is called "sleep mode." Assuming that the device is connected to Internet through a gateway (e.g., a base station), the attention of the mobile may be required by an incoming activity. Since the mobile transceiver is shut off, a response delay will be incurred. One then has to be very careful as how to schedule sleep periods in order to minimize energy consumption and reduce delays.

Since the initial announcement of IEEE 802.16e Standards for mobility [1], there has been an important volume of performance studies on the subject. The first approach chronologically is found in [11]. In an effort to relax some assumptions, [14, 12] study the impact of outgoing traffic, [8, 7] study the effect of setup time while [9, 3] deal with queueing implications in the analysis.

The above models assume a Poisson process for the packet arrivals. In [13], the authors are using hyper-Erlang distribution for the packet interarrival period. In [6, 2], hyper-exponential arrivals are proposed. In any of the above cases, an exogenous arrival process that does not depend on the energy management scheme is considered. Moreover, the delay metric taken is the average packet delay in the system.

Rather than assuming an exogenous independent arrival process, we have in mind elastic arrival processes in which (i) the duration of the activity period does not depend on the response delay—defined as the duration between the instant a request is issued and the instant at which the service actually begin—and (ii) the inactivity period begins when the activity of the mobile ends. Both assumptions are appropriate to interactive applications such as web browsing. As a result, the measure for delay is taken to be the mobile's response delay to the oldest activity taking place while idle.

The distribution of the duration of the inactivity period is often modeled as an exponential random variable. We shall also consider the case when the parameter of the exponential distribution is unknown but we have a known prior distribution on that parameter. This is equivalent to using a hyper-exponential distribution for the inactivity period.

In the literature, there are other works that provide evidence of heavy-tailed inactivity period distributions on the Internet and on the World Wide Web (e.g., a Pareto type distribution). In [5] the operator's idle periods are found to be heavy-tailed. As heavy-tailed distributed random variables can be well approximated by hyper-exponential distributions [10, 6], we face yet another motivation to study inactivity periods with hyper-exponential distributions.

Our contributions are:

i. We formulate an optimization framework for a large class of practical policies;

ii. We analytically derive the optimal behavior of two simple policies (constant or i.i.d. exponential vacations);

iii. We show that, with a Poisson arrival process, the optimal policy is the constant duration policy;

iv. We provide numerical results for cost minimization for a set of parameterized sleep policies;

v. We compare the optimal performance to the performance of the policy used in the IEEE 802.16e Standards (called the "Standard" policy).

The rest of the paper is structured as follows: Section 2 outlines our system model, introduces the cost function and states the optimization problem that is considered in the rest of the paper. Section 3 investigates policies with identically distributed vacations while those with non-identically distributed ones are tackled in Section 4. Numerical results and a comparative study of the different optimal policies and of the IEEE 802.16e Standard are reported in Section 5. We last conclude the paper in Section 6.

Due to space limitations, the proofs are omitted from the paper, but can be found in [4].

2 System Model

Consider a mobile wireless device connected to the Internet through a gateway. When the device is idle, keeping the transceiver working requires a power P_L; this is called *listen* state. Instead, while being in *sleep* state (the transceiver is turned off), the device requires a power P_S, with $P_S < P_L$. The device is then eager to go to sleep state in order to save energy and extend battery lifetime. If, however, the device is in sleep state, an incoming activity (packet or flow) will be stalled at the gateway. The device should frequently turn to listen state in order to check for such incoming activity.

In each idle period, the device goes through a sequence of sleep and listen states until an incoming activity is detected, see Fig. 1. In particular, in the beginning of each sleep period, the device chooses the sleep window while the listen window is considered fixed and negligible. At the end of the sleep window, the device switches to listen state. In case of no awaiting activity at the gateway, a fixed energy cost is incurred for checking the system state. In case at least one activity has arrived, the idle period is finished and a delay cost is incurred depending on the waiting time of the first activity, which itself depends on the device's idle period. In the spirit of achieving a Quality of Service (QoS) tradeoff, we are interested in finding the optimal policy that minimizes the total cost.

An equivalent system model is one that considers a server that goes on repeated vacations. The incoming traffic load is replaced by customers waiting to be served. In the following, we use the notation of a server with vacations as in [3]. The vacation length is then equivalent to the sleep window.

Let X denote the number of vacations in an idle period, where X is a discrete random variable taking values in $\mathbb{N}^*$. Given that our objective is to study the sleep mode, the value $X = 0$ is not allowed. Let τ denote the time between the start of the first
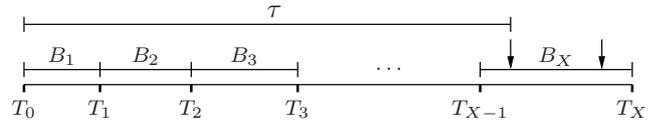

Figure 1: An idle period T_X. At time T_i, the mobile decides on a random vacation B_{i+1} and returns to sleep. At time $T_{i+1} = T_i + B_{i+1}$, it wakes up to check for activity. The idle period ends when an activity is detected.

vacation and the arrival of a customer; this time is nothing but the *inactivity* period. τ is a random variable whose probability density function is $f_\tau(t), t \geq 0$.

The duration of the kth vacation is a random variable denoted by B_k, for $k \in \mathbb{N}^*$. The time until the end of the kth vacation is a random variable denoted by T_k, for $k \in \mathbb{N}^*$. We denote T_0 as the time at the beginning of the first vacation; by convention $T_0 = 0$. Therefore $T_k = T_{k-1} + B_k = \sum_{i=1}^k B_i$. Observe that a generic idle period lasts for T_X units of time. All above definitions impose that $T_X \geq \tau$. We will use the following notation $\mathcal{Y}^*(s) := \mathbb{E}[\exp(-sY)]$ to denote the Laplace-Stieltjes transform of a generic random variable Y evaluated at s. Hence, we can readily write $\mathcal{T}_k^*(s) = \prod_{i=1}^k \mathcal{B}_i^*(s)$.

In this paper, we will be considering a certain number of different policies that constitute a set $\mathcal{R}$. We will focus on policies in which vacations are mutually independent random variables. Each element of $\mathcal{R}$ is a generic policy denoted $\mathbf{B} := \{B_k\}_{k \in \mathbb{N}^*}$; each entry of the random vector $\mathbf{B}$ has a predetermined distribution whose parameters are not specified. The parameters relative to policy $\mathbf{B}$ form a vector $\mathbf{b}$.

2.1 Cost Function

We now define the total cost of the sleep mode for an arbitrary idle period. This cost is simply the weighted sum of the energy consumed during the power save mode and the extra delay incurred on the traffic by a sleeping mobile. Given that the inactivity period is τ and the mobile's idle period is T_X, the service of the first customer to arrive during T_X is delayed for $T_X - \tau$ units of time.

The energy consumed by a mobile while listening to the channel and checking for customers is a constant denoted by E_L. This is actually a penalty paid at the end of each vacation. The energy consumed by a mobile during vacation B_k is therefore equal to $E_L + P_S B_k$ and that consumed during a generic idle period is equal to $E_L X + P_S T_X$.

The cost is then expressed as follows

$$V(\mathbf{b}) := \bar{\epsilon}\,\mathbb{E}[T_X - \tau] + \epsilon\left(E_L \mathbb{E}[X] + P_S \mathbb{E}[T_X]\right) \quad (1)$$

where ϵ is a normalized weight that takes value between 0 and 1, and $\bar{\epsilon} = 1 - \epsilon$. Note that $\mathbf{b}$ is hidden in the expectations appearing in (1). We would like to stress that this definition of the cost is meaningful only if the device is in sleep mode.

The following can be directly derived,

$$\mathbb{E}[T_X - \tau] = \sum_{k=1}^{\infty} \mathbb{E}[(T_k - \tau)\mathbb{1}\{T_{k-1} < \tau \leq T_k\}];$$

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k P(X = k); \qquad \mathbb{E}[T_X] = \sum_{k=1}^{\infty} \mathbb{E}[T_k] P(X = k);$$

where, for $k \in \mathbb{N}^*$, $P(X = k) = P(T_{k-1} < \tau \leq T_k)$. The cost can then be rewritten as

$$
\begin{aligned}
V(\mathbf{b}) &= \sum_{k=1}^{\infty} \Big\{ \bar{\epsilon}\, \mathbb{E}\big[(T_k - \tau)\mathbb{1}\{T_{k-1} < \tau \leq T_k\}\big] \\
&\quad + \epsilon P(X = k)\big(E_L k + P_S \mathbb{E}[T_k]\big) \Big\}. \quad (2)
\end{aligned}
$$

2.2 Problem Formulation

We aim to obtain the sleep policy and its optimal parameter which minimize the cost function $V(\mathbf{b})$ over the set of sleep policies $\mathcal{R}$. The cost for a given policy $r \in \mathcal{R}$ is denoted $V_r(\mathbf{b})$ and its minimum is denoted V_r^*. The policy r is optimized at $\mathbf{b}^*$ such that $\mathbf{b}^* := \arg\min_{\mathbf{b}} V_r(\mathbf{b})$ and $V_r^* := V_r(\mathbf{b}^*)$. Considering now the set of policies $\mathcal{R}$, the minimal cost W^* and the best policy $\mathbf{B}^*$ in $\mathcal{R}$ are given by $W^* := \min_{r \in \mathcal{R}} V_r^*$ and $\mathbf{B}^* := \arg W^*$.

2.3 Hyper-Exponential Inactivity Period

The inactivity period τ plays an important role in the above problem. Unless otherwise specified, we assume in the following that τ is hyper-exponentially distributed with n phases and parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$. In other words, we have

$$f_\tau(t) = \sum_{i=1}^{n} q_i \lambda_i \exp(-\lambda_i t), \qquad \sum_{i=1}^{n} q_i = 1. \quad (3)$$

Recall that τ represents the duration from the beginning of the save mode (i.e., beginning of the first vacation) until the arrival of the first packet (i.e., customer). Therefore, τ is the conditional residual inter-arrival time. Observe that when $n = 1$, τ is exponentially distributed with rate λ_1. This is equivalent to having a Poisson arrival process with rate λ_1, thanks to the memoryless property of the exponential distribution.

We will now compute the elements of (2) when τ is hyper-exponentially distributed. We obtain

$$
\begin{aligned}
P(X = k) &= P(\tau > T_{k-1}) - P(\tau > T_k) \\
&= \sum_{i=1}^{n} q_i \mathcal{T}_{k-1}^*(\lambda_i)\big(1 - \mathcal{B}_k^*(\lambda_i)\big); \\
\mathbb{E}[(T_k - \tau)&\mathbb{1}\{T_{k-1} < \tau \leq T_k\}] \\
&= \mathbb{E}[(T_{k-1} + B_k - \tau)\mathbb{1}\{T_{k-1} < \tau \leq T_{k-1} + B_k\}] \\
&= \sum_{i=1}^{n} q_i \mathcal{T}_{k-1}^*(\lambda_i)\left(\mathbb{E}[B_k] - \frac{1 - \mathcal{B}_k^*(\lambda_i)}{\lambda_i}\right).
\end{aligned}
$$

After some calculus, the cost simplifies to (denote $\eta = \bar{\epsilon} + \epsilon P_S$)

$$V(\mathbf{b}) = -\bar{\epsilon}\, \mathbb{E}[\tau] + \sum_{k=0}^{\infty} \sum_{i=1}^{n} q_i \mathcal{T}_k^*(\lambda_i)\big(\epsilon E_L + \eta \mathbb{E}[B_{k+1}]\big), \quad (4)$$

where $\mathbb{E}[\tau] = \sum_{i=1}^{n} q_i/\lambda_i$ is the expectation of τ. Equation (4) is very interesting in that the inactivity period τ appears only through its expectation. Given that the cost depends only on $\mathbb{E}[\tau]$, so will the minimal cost and the optimal control. This is true for any distribution of the vacations.

In the following sections, we will study several policies and derive explicitly V_r^* whenever possible.

3 Identically Distributed Vacations

We assume in this section that all vacations are identically distributed, in other words, the control is static. Let B be a generic random variable having the same distribution as any of the vacations. Thence, (4) can be rewritten as

$$V(\mathbf{b}) = -\bar{\epsilon}\, \mathbb{E}[\tau] + (\epsilon E_L + \eta \mathbb{E}[B]) \sum_{i=1}^{n} \frac{q_i}{1 - \mathcal{B}^*(\lambda_i)}. \quad (5)$$

We now propose different policies and derive the optimal control in each case. The policies that are considered are:

"Exponential" policy B is exponentially distributed; one can control b, the expectation of B;

"Constant" policy B is deterministic; one can control the constant vacation size b;

"Scaled" policy B is a scaled version of a known random variable S; one can control the scale α;

"General discrete" policy B has a discrete distribution with known possible values; one can control the distribution $\mathbf{p}$.

3.1 The Exponential Policy

In this policy, vacations are i.i.d. exponential random variables with mean $\mathbb{E}[B] = b$. The variable T_k is then Erlang distributed with shape k and rate $1/b$; $\mathbb{E}[T_k] = kb$. The inactivity period τ can have any distribution (we do not need τ to be hyper-exponentially distributed). With this policy, the cost, denoted $V_e(b)$, depends only on $\mathbb{E}[\tau]$, as detailed hereafter. Conditioning on a given inactivity period τ, the number of vacations decremented by one is a Poisson variable with rate τ/b. It is straightforward to write

$$\mathbb{E}[X] = \mathbb{E}[\tau]/b + 1; \qquad \mathbb{E}[T_X] = b\mathbb{E}[X] = \mathbb{E}[\tau] + b.$$

Equation (1) can be rewritten (recall that $\eta = \bar{\epsilon} + \epsilon P_S$)

$$V_e(b) = \epsilon\big(P_S + E_L/b\big)\mathbb{E}[\tau] + (\epsilon E_L + \eta b). \quad (6)$$

3

Remark 3.1 *Equation* (6) *stands for any distribution of* τ. *We naturally obtain the same expression if we substitute* $\mathcal{B}^*(s)$ *for* $1/(1 + bs)$ *in* (5).

Proposition 3.1 *The cost* $V_e(b)$ *is a convex function having a minimum at*

$$b_e^* = \sqrt{\frac{\epsilon E_L \mathbb{E}[\tau]}{\eta}} = \sqrt{\frac{\epsilon E_L \mathbb{E}[\tau]}{\bar{\epsilon} + \epsilon P_S}}. \tag{7}$$

The minimal cost is

$$V_e^* = V_e(b_e^*) = \epsilon(P_S \mathbb{E}[\tau] + E_L) + 2\sqrt{\epsilon \eta E_L \mathbb{E}[\tau]} \tag{8}$$

Proposition 3.1 is really interesting in that it says that with i.i.d. exponential vacations, only the expected inactivity period defines the optimal control. The inactivity period τ can be generally distributed. Therefore, Proposition 3.1 stands valid for any user application.

3.2 The Constant Policy

In this policy, all vacations are equally sized. In other words, $B = b$. The performance is optimized by controlling the size of b. Substituting $\mathcal{B}^*(s)$ for $\exp(-sb)$ in (5) yields the following simplified expression for the cost (the subscript stands for "constant")

$$V_c(b) = -\bar{\epsilon}\mathbb{E}[\tau] + (\epsilon E_L + \eta b)\sum_{i=1}^{n}\frac{q_i}{1 - \exp(-\lambda_i b)}. \tag{9}$$

Proposition 3.2 *When* $n = 1$, *the cost* $V_c(b)$ *is a convex function having a minimum at*

$$b_c^* = -\frac{1}{\lambda_1}\left(\zeta_1 + W_{-1}(-e^{-\zeta_1})\right) \text{ with } \zeta_1 := \frac{\lambda_1 \epsilon E_L}{\eta} + 1, \tag{10}$$

where W_{-1} *denotes the branch of the Lambert W function*[1] *that is real-valued on the interval* $[-\exp(-1), 0]$ *and always below* -1. *The minimal cost is*

$$V_c^* = V_c(b_c^*) = -\frac{1}{\lambda_1}\left(\bar{\epsilon} + \eta W_{-1}(-e^{-\zeta_1})\right). \tag{11}$$

Proposition 3.3 *The cost* $V_c(b)$ *is a convex function having a minimum in* $]0, \infty[$.

Proposition 3.3 proves the existence of a global minimum. Unfortunately, we are not able to derive the optimal b_c^* analytically and use numerical methods to find b_c^*. The dimensionality of the problem can be showcased by the following result.

Proposition 3.4 *When* $n > 1$, *no optimal policy can be independent of* $\mathbf{q} = (q_1, \ldots, q_n)$.

[1] The Lambert W function, satisfies $W(x)\exp(W(x)) = x$. As the equation $y\exp(y) = x$ has an infinite number of solutions y for each (non-zero) value of x, the function $W(x)$ has an infinite number of branches.

3.3 The Scaled Policy

In this third policy, we consider the random vacation B to be a factor α of a random variable S with a general distribution, i.e. $B = \alpha S$. For a given distribution of S, the scaling factor α is controlled to optimize the performance. The cost $V_s(\alpha)$ (the subscript stands for "scaled") follows readily from (5) using $\mathcal{B}^*(s) = \mathcal{S}^*(\alpha s)$ and $\mathbb{E}[B] = \alpha \mathbb{E}[S]$.

We consider now that S is a discrete random variable taking values in a finite set $\{a_j\}_{j=1,\ldots,J}$ with a probability distribution $\mathbf{p} = (p_1, \ldots, p_J)$, i.e., $P(S = a_j) = p_j$ and $\sum_{j=1}^{J} p_j = 1$. Hence, $\mathcal{S}^*(s) = \sum_{j=1}^{J} p_j \exp(-sa_j)$, and $\mathbb{E}[S] = \sum_{j=1}^{J} p_j a_j$.

This policy advocates to have each vacation follow a discrete general distribution, taking values in $\{\alpha a_j\}_{j=1,\ldots,J}$. The probability distribution $\mathbf{p}$ is assumed fixed whereas the set of possible values can be scaled for minimal cost.

The optimization problem can be stated as

$$V_s^* = \min_{\alpha > 0} V_s(\alpha); \quad \alpha^* = \arg V_s^*. \tag{12}$$

It is intractable to solve analytically (12), we will therefore resort to a numerical resolution (cf. Section 5).

3.4 The General Discrete Policy

The fourth policy resembles the third one in that it equally considers a discrete general vacation for the variable B. However, the set of possible values is now fixed (i.e., $\alpha = 1$) whereas the probability distribution $\mathbf{p}$ can be optimized for minimal cost. We denote the cost as $V_g(\mathbf{p})$, where the subscript stands for "general," and write

$$V_g(\mathbf{p}) = -\bar{\epsilon}\mathbb{E}[\tau] + \sum_{i=1}^{n}\frac{q_i\left(\epsilon E_L + \eta \sum_{j=1}^{J} p_j a_j\right)}{1 - \sum_{j=1}^{J} p_j \exp(-\lambda_i a_j)}.$$

Our objective is to find $\mathbf{p}^* = \arg\min_{\mathbf{p}} V_g(\mathbf{p})$ such that $0 \leq p_j \leq 1$ for $j = 1, \ldots, J$ and $\sum_{j=1}^{J} p_j = 1$. This optimization problem can only be solved numerically.

4 Non-Identically Distributed Vacations

If we relax the constraint of identically distributed vacations, the mobile is free to choose any vacation distribution at each waking up instant, a fact that complexes the problem immensely. We will narrow the problem by considering only deterministic vacations.

The kth vacation is now of fixed size b_k, the instants $\{T_k\}_{k \in \mathbb{N}}$ are now deterministic, and we let $t_k = T_k$ for any k to reflect this. We have $t_0 = 0$ and $t_k = \sum_{j=1}^{k} b_j$.

The policies that are considered in this section are:

"Semi-Constant" policy Most vacations are equal;

"Multiplicative" policy Vacations increase over time;

"General Deterministic" policy Vacations can last for any positive time.

4.1 The Semi-Constant Policy

As the name indicates, only a few vacations are allowed to be of different size. This is expected to bring some improvement with respect to the Constant policy. For the sake of illustration, let the first vacation be of size b_0 while all the subsequent ones are of size b. This particular policy will be referred to as "one-stage" policy. The cost is

$$V_{sc}(b_0, b) = -\bar{\epsilon}\mathbb{E}[\tau] + (\epsilon E_L + \eta b) \sum_{i=1}^{n} \frac{q_i \exp(-\lambda_i b_0)}{1 - \exp(-\lambda_i b)}. \quad (13)$$

Proposition 4.1 *For $n = 1$, all vacations are equal at optimum, i.e., $b_0^* = b^* = b_c^*$ (recall (10)).*

We conjecture here that, for $n = 1$, the Constant policy is optimal among all policies, as hinted by Proposition 4.1. One can use dynamic programming to prove this conjecture, which is beyond the scope of this paper.

Notice that the m-stage Semi-Constant policy approaches the absolute optimal policy when $m \to \infty$. However, it adds immense computational complexity as m grows. Hence, for practical purposes, one can optimize until a few stages.

4.2 The Multiplicative Policy

This policy is inspired by the power save mode of the IEEE 802.16e [1], and more precisely, by type I power saving classes. There, the size of a sleep window (i.e., a vacation) is doubled over time until a maximum permissible sleep window, denoted $b_{\max}$, is reached. The size of the kth vacation is then

$$b_k = b_1 2^{\min\{k-1,l\}}, \quad k \in \mathbb{N}^*$$

where $l := \log_2(b_{\max}/b_1)$. We also have

$$t_k = b_1 \left(2^{\min\{k,l\}} - 1 + 2^l (k - l) \mathbb{1}\{k > l\} \right), \quad k \in \mathbb{N}^*.$$

The cost of the power save mode of the IEEE 802.16e Standard can be derived from (4), yielding

$$V_{\text{Std}} = -\bar{\epsilon} \, \mathbb{E}[\tau] + \sum_{k=0}^{\infty} \sum_{i=1}^{n} q_i e^{-\lambda_i t_k} \left[\epsilon E_L + \eta b_1 2^{\min\{k,l\}} \right]. \quad (14)$$

Instead of doubling the vacations over time, the multiplicative policy increases the vacations by a multiplicative factor f (in the Standard policy, $f = 2$). The performance is then optimized by controlling the factor f. In this policy, we have

$$b_k = b_1 f^{\min\{k-1,l\}}, \quad k \in \mathbb{N}^*$$

$$t_k = b_1 \left(\frac{f^{\min\{k,l\}} - 1}{f - 1} + f^l (k - l) \mathbb{1}\{k > l\} \right), \quad k \in \mathbb{N}^*$$

$$V_m(f) = -\bar{\epsilon}\mathbb{E}[\tau] + \sum_{k=0}^{\infty} \sum_{i=1}^{n} q_i e^{-\lambda_i t_k} [\epsilon E_L + \eta b_{k+1}] \quad (15)$$

$$f^* = \arg \min_{f > 1} V_m(f). \quad (16)$$

The optimal f^* and the minimal cost $V_m^* = V_m(f^*)$ ("m" stands for "multiplicative") will be computed numerically.

4.3 The General Deterministic Policy

In this section, no particular pattern is imposed on the vacations. This policy is a generalization of the Semi-Constant policy as $m \to \infty$. We denote the cost as $V_d(\boldsymbol{\mu})$ where the subscript stands for "deterministic" and $\boldsymbol{\mu} := (b_1, b_2, \dots)$ is the deterministic policy. The cost has the same expression as (15). A necessary condition for the existence of an optimal control sequence $\boldsymbol{\mu}^* = (b_1^*, b_2^*, \dots)$ is that $\mathbf{grad}\, V_d(\boldsymbol{\mu}^*) = 0$.

Proposition 4.2 *When $n > 1$, no optimal policy can be independent of $\mathbf{q} = (q_1, \dots, q_n)$.*

4.4 Recapitulation

In Sections 3 and 4, we have characterized various sleep policies. The optimal control for the Exponential and Constant policies has been derived analytically. For the rest of the policies, the optimal control will be derived numerically as reported in the following section.

Comparing the Exponential and Constant policies for $n = 1$ (Poisson arrival process), it comes that the Constant policy achieves a better minimal cost than the Exponential policy (cf. (8) and (11)). This observation derives from the fact that $W_{-1}(-e^{-x}) \geq -\exp(2\sqrt{x-1})$ for $x \geq 1$; thus $V_c^* \leq V_e^*$. Moreover, and in light of Proposition 4.1, we believe that the Constant policy is optimal among all policies, for $n = 1$.

5 Numerical Investigation

In this section, we show some numerical results of our modeling. We study the performance of the IEEE 802.16e's power saving scheme, which allows devices to go on sleep mode when there are no packets to serve.

In practice, the mobile needs to check for any packet arrived while it was off. So, at the end of each sleep duration, it will wake up and listen to the channel for any notification message from the base station. The mobile needs to switch on the radio every time it wakes up and then listen to the channel for a (small) fixed duration listen period. Hence, each waking up costs E_L amount of energy consumption. The rate of energy consumption during sleep, P_S, is considerably less than that during listening.

The cost V, defined in (1), captures the main performance measures: energy consumed during the sleep duration and extra delay incurred due to the sleep mode. The cost V is a weighted sum of both metrics. From (1), it comes that V is more sensitive to the energy consumption when ϵ is large, whereas a small ϵ gives more weight to the delay.

The various policies discussed in previous sections are now evaluated numerically. Policies are compared through: (i) the optimal expected sleep duration, (ii) the minimal cost achieved, and (iii) the relative improvement with respect to the IEEE 802.16e

protocol. The improvement ratio, denoted I, is defined as follows:

$$I := \frac{V_{\text{Std}} - V_r^*}{V_{\text{Std}}}. \qquad (17)$$

The physical parameters are set to the following values: $E_L = 10$, and $P_S = 1$. The parameters of the Standard policy are $b_1 = 2$ and $l = 10$.

5.1 Exponential Inactivity Period

In this section, we consider arrivals to form a Poisson process with rate $\lambda = 1/\mathbb{E}[\tau]$. Three sleep policies (Exponential, Constant and Standard) are evaluated and compared. The performance of each policy depends on the arrival rate λ and on the normalized weight ϵ. In the following evaluation, we will alternatively vary one of the parameters and fix the other.

We first vary λ and fix ϵ to 0.1 and 0.9. The weight ϵ equal to 0.1 mimics the situation when energy consumption is given lower priority over delay, while ϵ equal to 0.9 mimics the opposite situation. Looking at Fig. 2, one can observe the impact of the arrival rate λ on (i) the optimal expected sleep duration (cf. Fig. 2(a)), (ii) the minimal cost (cf. Fig. 2(b)), and (iii) the cost improvement (cf. Fig. 2(c)). We naturally find that the expected sleep duration decreases as λ increases, as forseen in (7) and (10). The physical explanation is that, a large arrival rate forces the server to be available after shorter breaks, otherwise the cost is too high.

Of more interest are the curves reported in Fig. 2(b), where the optimal cost achieved by the Constant policy always outperforms the costs of the two other policies. This is in agreement with the discussion in Section 4.4, namely, that the constant policy should be the optimal among all possible policies. The Exponential policy outperforms the Standard policy for a large range of values of λ as seen more clearly in Fig. 2(c) where $\epsilon = 0.1$.

Observe in Fig. 2(b) how the cost decreases asymptotically to ϵE_L (1 for $\epsilon = 0.1$ and 9 for $\epsilon = 0.9$) as the rate λ increases. The same trend is observed for the cost of the Standard policy. As λ decreases, the increase in V_c^* and V_e^* is due to the increase of the optimal expected sleep duration, while for V_{Std} the increase is due to the extra (useless and costly) listening.

We next vary ϵ and fix λ to 0.1 (low traffic) and 5 (high traffic). The results are depicted in Fig. 3. As ϵ gets smaller, the extra delay gets more penalizing, enforcing then smaller optimal sleep durations; cf. Fig. 3(a). As mentioned earlier, smaller optimal sleep durations yield smaller optimal costs. Thus, the optimal costs increase as ϵ increase, as can be observed in Fig. 3(b). For $\epsilon < 0.1$, the Standard policy's cost is fairly insensitive to ϵ. This is because the Standard intrinsically favors delay over energy: the first sleep duration is small ($b_1 = 2$) and it takes a while before the sleep duration becomes penalizing in terms of delay. This is confirmed by the sudden cost increase as $\epsilon \geq 1$ (notice the logarithmic scale): when energy consumption costs start to have more weight, the Standard policy's performance degrades.

Looking at Fig. 3(c), we find again that *the Constant policy is the best* and that *the Exponential policy outperforms the Standard*

policy in most cases as it yields a substantial improvement over a large range of values of λ and ϵ.

5.2 Hyper-Exponential Inactivity Period

In this section, we consider the situation in which the inactivity period follows an n-phase hyper-exponential distribution. We start with the comparison between the Exponential, Constant, Scaled (cf. Section 3.3), Semi-Constant (cf. Section 4.1) and Standard policies. Analytical expression are available only for the Exponential policy, for the rest of the policies we resort to using numerical solutions.

For this study, we consider for τ two distinct distributions. We let $n = 3$ and $\boldsymbol{\lambda} = [0.2, 3, 10]$ but consider two possible values for $\mathbf{q}$, namely, $\mathbf{q}^1 = [0.1, 0.3, 0.6]$ and $\mathbf{q}^2 = [0.6, 0.3, 0.1]$, yielding an expected inactivity period $\mathbb{E}[\tau]$ equal to 0.66 (high incoming traffic) and 3.11 (moderate incoming traffic), respectively. These values of $\mathbf{q}^1, \mathbf{q}^2$ have been intentionally chosen so as to show different behavior of the policies. The parameters of the Scaled policy are (distribution of the variable S) $\{a_1, a_2, a_3\} = \{0.2, 1, 3\}$ and $\mathbf{p} = [0.6, 0.3, 0.1]$. The optimal expected sleep duration is then $0.72\alpha^*$. As for the Semi-Constant policy, we report the results of the two-stage one (the first two vacations are allowed to have a different size from the rest of the vacations).

We vary the weight ϵ between 0.001 and 1. The impact of ϵ on the expected vacation size, the cost and the cost improvement can be observed in Fig. 4. We observe the same trends for the optimal expected vacation size and the optimal cost as with Poisson arrivals (cf. Fig. 3). Unlike the case in Fig. 3(b), the optimal cost achieved by the Constant policy (i.e., V_c^*) is not the smallest among all costs, at least at high arrival rate ($\mathbb{E}[\tau] = 0.66$ in Fig. 4(b)). The best performance at this arrival rate is achieved by the Exponential and Scaled policies for most values of ϵ. Notice the poor performance of the Constant and Semi-Constant policies, which, interestingly enough, exhibit the same trend as the Standard policy.

The performance of the policies at moderate rate can be seen in 4(c). For $\mathbb{E}[\tau] = 3.11$, the Exponential policy is the best whatever the weight ϵ, performing at least as good as the Standard policy if not better.

The last policy that we evaluate is the Multiplicative policy (cf. Section 4.2). We want to compute the optimal multiplicative factor for a variety of distributions of τ. To this end, the rates of the $n = 3$ phases is taken to be $C_\lambda \boldsymbol{\lambda} = [0.2C_\lambda, 3C_\lambda, 10C_\lambda]$ and the probabilities of the phases are $\mathbf{q}^2 = [0.6, 0.3, 0.1]$. The expected inactivity period is then $\mathbb{E}[\tau] = 3.11/C_\lambda$. We vary the scaling factor C_λ from 0.001 (extremely low traffic) to 1 (moderate traffic).

Results are depicted in Fig. 5. On can deduce from Fig. 5(a) that the value $f = 2$ used in the Standard policy is actually optimal (considering the Multiplicative policy) when there is almost no traffic ($\mathbb{E}[\tau] = 3110$). Even though the values of f^* for different ϵ are very close to each other at $C_\lambda = 0.01$, we observe a large impact on the cost improvement (cf. Fig. 5(b)). We can conclude that the optimal cost is highly sensitive to the Multiplicative fac-
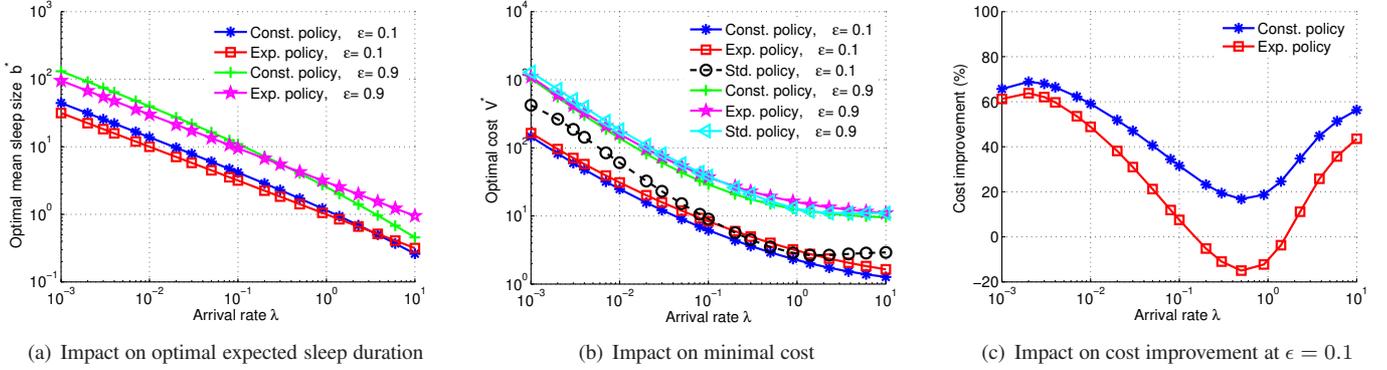
(a) Impact on optimal expected sleep duration

(b) Impact on minimal cost

(c) Impact on cost improvement at $\epsilon = 0.1$

Figure 2: Exponential τ: Impact of λ for various sleep policies.

(a) Impact on optimal expected sleep duration

(b) Impact on minimal cost

(c) Impact on cost improvement at $\lambda = 0.8$

Figure 3: Exponential τ: Impact of ϵ for various sleep policies.

(a) Impact on optimal expected sleep duration

(b) Impact on minimal cost

(c) Impact on cost improvement at $\boldsymbol{\lambda} = (0.2, 3, 10)$
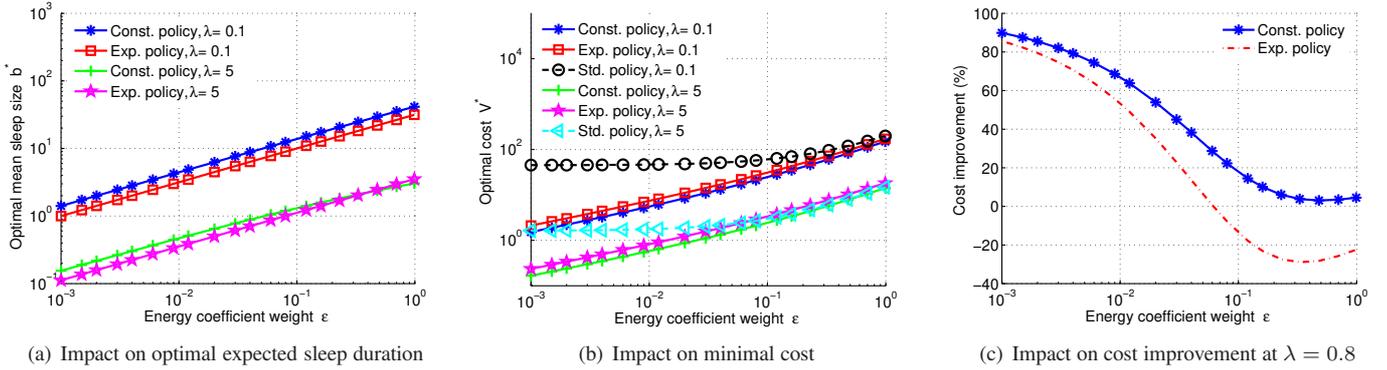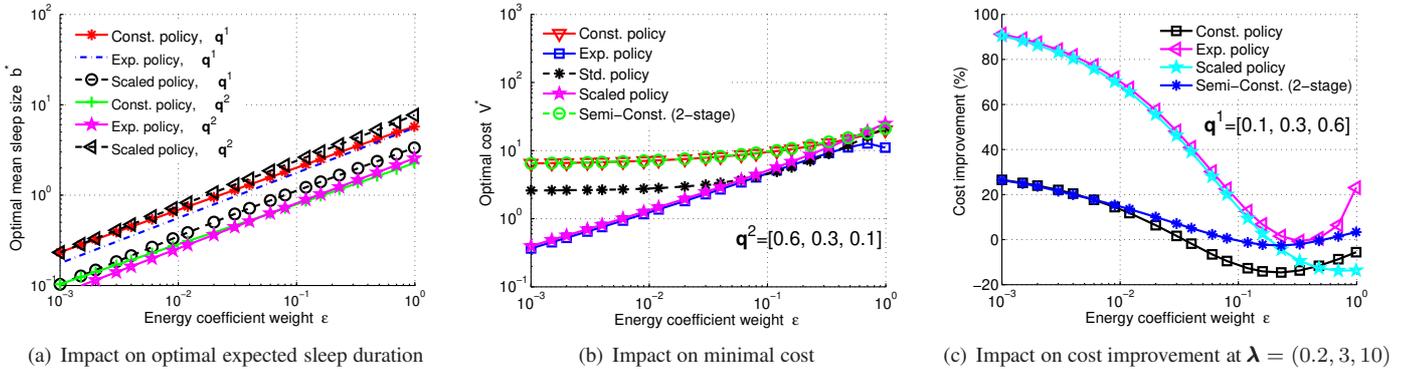
Figure 4: Hyper-exponential τ: Impact of ϵ for various sleep policies.

(a) Impact on optimal factor (b) Impact on cost improvement
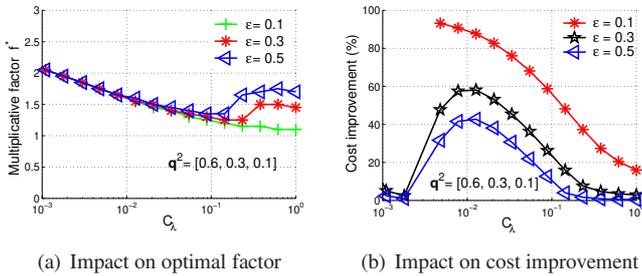
Figure 5: Hyper-exponential τ: Impact of C_λ on the multiplicative policy.

tor at very low traffic. This is not surprising as vacations increase exponentially in the Multiplicative policy, and this is much likely to happen when traffic is very low.

6   Conclusion

We introduced an optimization framework for controlling the vacation length as well as selecting the best vacation policy. The approach can be directly applied to centrally controlled wireless devices for optimizing individual energy saving while taking into account the delays. Among the policies proposed in the paper, the Constant policy achieves the best performance if arrivals are Poisson. When the inactivity period is hyper-exponentially distributed, we showed that the Standard can be improved substantially if the multiplicative factor is optimized. Also, if one gives little weight to the mobile's response delay and favors the minimization of energy use, then both the Exponential and Scaled policies are candidate to substitute the Standard. The optimal control for the Exponential policy is found in closed form for a general inactivity period as well as that for the Constant policy for a Poisson arrival process.

Using dynamic programming, one can show that in case of Poisson inactivity periods, the optimal sleep policy should be the Constant policy. Also, it may be possible to identify the optimal policy when the inactivity period is hyper-exponentially distributed, or even generally distributed.

In this paper, we have focused on policies that keep no memory of the past. Alternatively, one can study adaptive policies which can exploit the history of the inactivity period.

References

[1] IEEE Std 802.16e 2005. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Amendment: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2005.

[2] J. Almhana, Z. Liu, C. Li, and R. McGorman. Traffic estimation and power saving mechanism optimization of IEEE 802.16e networks. In *Proc. of IEEE ICC 2008, Beijing, China*, pages 322–326, May 2008.

[3] S. Alouf, E. Altman, and A. P. Azad. Analysis of an M/G/1 queue with repeated inhomogeneous vacations with application to IEEE 802.16e power saving mechanism. In *Proc. of QEST 2008*, pages 27–36, September 2008.

[4] A. P. Azad, S. Alouf, E. Altman, V. Borkar, and G. Paschos. Vacation policy optimization with application to IEEE 802.16e power saving mechanism. Research Report RR-7017, INRIA, Aug. 2009.

[5] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic-evidence and possible causes. In *Proc. of ACM Sigmetrics, Philadelphia, PE*, pages 160–169, 1996.

[6] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(8):963–976, August 1998.

[7] K. Han and S. Choi. Performance analysis of sleep mode operation in IEEE 802.16e mobile broadband wireless access systems. In *Proc. of IEEE VTC 2006-Spring, Melbourne, Australia*, May 2006.

[8] N-H. Lee and S. Bahk. MAC sleep mode control considering downlink traffic pattern and mobility. In *Proc. of IEEE VTC 2005-Spring, Stockholm, Sweden*, volume 3, pages 3102–3106, May 2005.

[9] Y. Park and G. U. Hwang. Performance modelling and analysis of the sleep mode in IEEE 802.16e WMAN. In *Proc. of IEEE VTC 2007-Spring, Melbourne, Australia*, pages 2801–2806, April 2007.

[10] A. Riska, V. Diev, and E. Smirni. Efficient fitting of long-tailed data sets into hyperexponential distributions. In *Proc. of IEEE GLOBECOM*, volume 3, pages 2513–2517, November 2002.

[11] Y. Xiao. Energy saving mechanism in the IEEE 802.16e wireless MAN. *IEEE Communications Letters*, 9:595–597, July 2005.

[12] Y. Xiao. Performance analysis of an energy saving mechanism in the IEEE 802.16e wireless MAN. In *Proc. of IEEE CCNC 2006*, pages 406–410, January 2006.

[13] Y. Zhang. Performance modeling of energy management mechanism in IEEE 802.16e mobile WiMAX. In *Proc. of IEEE WCNC 2007*, pages 3205–3209, March 2007.

[14] Y. Zhang and M. Fujise. Energy management in the 802.16e MAC. *IEEE Communications Letters*, 10:311–313, April 2006.