Optimal Energy-Efficient Regular Delivery of Packets in Cyber-Physical Systems

Xueying Guo*, Rahul Singh†, P. R. Kumar† and Zhisheng Niu*
*Department of Electronic Engineering, Tsinghua Uviersity, R. R. China
Email: guo-xy11@mails.tsinghua.edu.cn, niuzhs@tsinghua.edu.cn
†Department of Electrical and Computer Engineering, Texas A&M University, USA
Email:{rsing1, prk}@tamu.edu

***Abstract*—In cyber-physical systems such as in-vehicle wireless sensor networks, a large number of sensor nodes continually generate measurements that should be received by other nodes such as actuators in a regular fashion. Meanwhile, energy-efficiency is also important in wireless sensor networks. Motivated by these, we develop scheduling policies which are energy efficient and simultaneously maintain "regular" deliveries of packets. A tradeoff parameter is introduced to balance these two conflicting objectives. We employ a Markov Decision Process (MDP) model where the state of each client is the time-since-last-delivery of its packet, and reduce it into an equivalent finite-state MDP problem. Although this equivalent problem can be solved by standard dynamic programming techniques, it suffers from a high-computational complexity. Thus we further pose the problem as a restless multi-armed bandit problem and employ the low-complexity Whittle Index policy. It is shown that this problem is indexable and the Whittle indexes are derived. Also, we prove the Whittle Index policy is asymptotically optimal and validate its optimality via extensive simulations.**

I. Introduction

Cyber-physical systems typically employ wireless sensors for keeping track of physical processes such as temperature and pressure. These nodes then transmit data packets containing these measurements back to the access point/base station. Moreover, these packets should be delivered in a "regular" way. So, time between successive deliveries of packets, i.e. inter-delivery time, is an important performance metric [1], [2]. Furthermore, many wireless sensors are battery powered. Thus, energy-efficiency is also important.

We address the problem of satisfying these dual conflicting objectives: inter-delivery time requirement and energy-efficiency. We design wireless scheduling policies that support the inter-delivery requirements of such wireless clients in an energy-efficient way. In [3], [4], the authors analyzed the growth-rate of service irregularities that occur for the case of multiple clients sharing a wireless network and when the system is in heavy traffic regime. The inter-delivery performance of the Max Weight discipline under the heavy traffic regime was studied in [5]. To the authors' best knowledge, the inter-delivery time was first considered in [1], [2] as a performance metric for queueing systems, where a sub-optimal policy is proposed to trade off the stablization of the queues and service regularity. However, this is different from our problem, where the arrival process does not need to be featured. In our previous work [6], throughput is traded off for better performance with respect to variations in inter-delivery times. However, tunable and heterogeneous inter-delivery requirements have not been considered.

In this paper, we formulate the problem as a Markov Decision Process (MDP) with a system cost consisting of the summation of the penalty for exceeding the inter-delivery threshold and a weighted transmission energy consumption. An energy-efficiency weight parameter η is introduced to balance these two aspects. To solve this infinite-state MDP problem, we reduce it to an equivalent MDP comprising of only a finite number of states. This equivalent finite-state finite-action MDP can be solved using standard dynamic programming (DP) techniques.

The significant challenge of this MDP approach is the computational complexity, since the state-space of the equivalent MDP increases exponentially in the number of clients. To address this, we further formulate this equivalent MDP as a restless multi-armed bandit problem (RMBP), with the goal of exploiting a low-complexity index policy.

In this RMBP, we first derive an upper bound on the achievable system reward by exploring the structure of a relaxed-constraint problem. Then, we determine the Whittle index for our multi-armed restless bandit problem, and prove that the problem is indexable. In addition, we show the resulting index policy is optimal in certain cases, and validate the optimality by a detailed simulation study. The impact of the energy-efficiency parameter η is also studied in the simulation results.

II. System Model

Consider a cyber-physical system in which there are N wireless sensors and one access point (AP). We will assume that time is discrete. At most L sensors can simultaneously transmit in a time slot. In each time-slot, a control message is broadcasted at the beginning by the AP to inform which set of L sensors can transmit in the current time-slot. Each of the assigned sensors then makes a sensor measurement and transmits its packet. The length of a time slot is the time required for the AP to send the control message plus the time required for the L assigned clients to prepare and transmit a package.

The wireless channel connecting the sensor and the AP is unreliable. When client n is selected to transmit, it succeeds in delivering a packet with a probability $p_n \in (0,1)$. Furthermore, each attempt to transmit a packet of client n consumes E_n units of energy.

The QoS requirement of client n is specified through an integer, the *packet inter-delivery time threshold* τ_n. The cost incurred by the system during the time interval $\{0,1,\ldots,T\}$ is given by,

$$\mathrm{E}\Big[\sum_{n=1}^{N}\Big(\sum_{i=1}^{M_T^{(n)}}(D_i^{(n)}-\tau_n)^+ + (T-t_{D^{(n)}_{M_T^{(n)}}}-\tau_n)^+ \\ + \eta \hat{M}_T^{(n)} E_n\Big)\Big], \qquad (1)$$

where $D_i^{(n)}$ is the time between the deliveries of the i-th and $(i+1)$-th packets for client n, $M_T^{(n)}$ is the number of packets delivered for the n-th client by the time T, $t_{D_i^{(n)}}$ is the time slot in which the i-th package for client n is delivered, $\hat{M}_T^{(n)}$ is the total number of slots in $\{0,1,\cdots,T{-}1\}$ in which the n-th client is selected to transmit, and $(a)^+ := \max\{a,0\}$. The second term is included since, otherwise, no transmission at all will result in the least cost. The last term weights the total energy consumption in T time-steps by a non-negative *energy-efficiency parameter* η, which tunes the weightage given to energy conservation. The access point's goal is to select at most L clients to transmit in each time-slot from among the N clients, so as to minimize the above cost.

III. Reduction to Finite State Problem

In the following, vectors will be denoted bold font, i.e., $\mathbf{a} := (a_1,\ldots,a_N)$. Define $\mathbf{a}\wedge\mathbf{b} := (a_1\wedge b_1,\ldots,a_N\wedge b_N)$. Random processes will be denoted by capitals.

We formulate our system as a Markov Decision Process, as follows. The system state at time-slot t is denoted by a vector $X(t) := (X_1(t),\cdots,X_N(t))$, where $X_n(t)$ is the time elapsed since the latest delivery of client n's packet. Denote the action at time t as $U(t) := (U_1(t),\cdots,U_N(t))$, with $\sum_{n=1}^N U_n(t) \le L$ for each t, where

$$U_n(t) = \begin{cases} 1 \text{ if client } n \text{ is selected to transmit in slot } t, \\ 0 \text{ otherwise.}\end{cases}$$

The system state evolves as,

$$X_n(t+1) = \begin{cases} 0 \ \text{ if a packet of client } n \text{ is delivered in } t, \\ X_n(t)+1 \text{ otherwise.}\end{cases}$$

Thus, the system forms a controlled Markov chain (denoted *MDP-1*), with the transition probabilities given by,

$$P^{\text{MDP-1}}_{\mathbf{x},\mathbf{y}}(\mathbf{u}) := \mathrm{P}\left[X(t+1)=\mathbf{y}\middle|X(t)=\mathbf{x},U(t)=\mathbf{u}\right] \\ = \prod_{n=1}^{N}\mathrm{P}\left[X_n(t+1)=y_n\middle|X_n(t)=x_n,U_n(t)=u_n\right],$$

with $\mathrm{P}\left[X_n(t+1)=y_n\middle|X_n(t)=x_n,U_n(t)=u_n\right]$

$$:= \begin{cases} p_n & \text{if } y_n = 0 \text{ and } u_n = 1, \\ 1-p_n & \text{if } y_n = x_n+1 \text{ and } u_n = 1, \\ 1 & \text{if } y_n = x_n+1 \text{ and } u_n = 0, \\ 0 & \text{otherwise.}\end{cases}$$

The T-horizon optimal cost-to-go from initial state $\mathbf{x}$ is given by,

$$V_T(\mathbf{x}) := \min_{\pi:\sum_n U_n(t)\le L} \mathrm{E}\Big\{\sum_{t=0}^{T-1}\sum_{n=1}^{N}\Big(\eta E_n U_n(t) \\ + (X_n(t){+}1{-}\tau_n)^+\,\mathbf{1}\{X_n(t{+}1)=0\}\Big)\Big|X(0)=\mathbf{x}\Big\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and $X(T) := \mathbf{0}$ (which leads to recovering the second term in the cost (1)), and the minimization is over the class of history dependent policies.

The Dynamic Programming (DP) (see [7]) recursion is,

$$V_T(\mathbf{x}) = \min_{\mathbf{u}:\sum_n u_n \le L} \mathrm{E}\Big\{\eta\sum_{n=1}^{N}E_n u_n + \sum_{\mathbf{y}} P^{\text{MDP-1}}_{\mathbf{x},\mathbf{y}}(\mathbf{u}) \\ \cdot\left[\sum_{n=1}^{N}(x_n+1-\tau_n)^+\,\mathbf{1}\{y_n=0\} + V_{T-1}(\mathbf{y})\right]\Big\}. \qquad (2)$$

The above problem, denoted as MDP-1, involves a countably infinite state space. The following results show that it can be replaced by an equivalent finite state MDP.

Lemma 1. For the MDP-1, we have, $\forall x_1,\cdots,x_N \ge 0$,

$$V_T(x_1,\cdots,\tau_i{+}x_i,\cdots,x_N) = x_i{+}V_T(x_1,\cdots,\tau_i,\cdots,x_N).$$

Moreover, the optimal actions for the states $(x_1,\cdots,\tau_i+x_i,\cdots,x_N)$ and $(x_1,\cdots,\tau_i,\cdots,x_N)$ are the same.

Proof. Let us consider the MDP-1 starting from two different initial states, $\mathbf{x} = (x_1,\cdots,\tau_i+x_i,\cdots,x_N)$ and $\tilde{\mathbf{x}} = (x_1,\cdots,\tau_i,\cdots,x_N)$, and compare their evolutions. Construct the processes associated with both the systems on a common probability space and couple stochastically the successful transmissions for the two systems. Let π be an arbitrary history-dependent policy that is applied to in the first system (starting in state $\mathbf{x}$). Corresponding to π, there is a policy $\tilde{\pi}$ in the second system, which takes the same actions as the policy π at each time slot. Then all the packet-inter-delivery times for both the processes are the same, except for the first inter-delivery time of the i-th client, which is larger for the former system as compared to the latter by x_i. In addition, Since the policy π is arbitrary, $V_T(\mathbf{x}) \ge x_i + V_T(\tilde{\mathbf{x}})$. The inequality in the other direction is proved similarly. The proof of the second statement follows by letting π be the optimal policy. □

Corollary 2. For any system state $\mathbf{x}$ such that $x_n \le \tau_n, \forall n$,

$$V_T(\mathbf{x}) = \min_{\mathbf{u}:\sum_n u_n \le L} \mathrm{E}\Big\{ \sum_n \left(\eta E_n u_n + \mathbf{1}\left\{x_n = \tau_n\right\}\right) + \sum_{\mathbf{y}} P^{\text{MDP-1}}_{\mathbf{x},\mathbf{y}} V_{T-1}(\mathbf{y} \wedge \boldsymbol{\tau})\Big\}. \tag{3}$$

Proof. Consider the equation (2) and the following two cases:

1) The initial state $\mathbf{x}$ is such that $x_n < \tau_n, \forall n$. Then $(x_n + 1 - \tau_n)^+ = 0$ and $\mathbf{1}\{x_n = \tau_n\} = 0$. In addition, for any action $\mathbf{u}$, if $\mathbf{y}$ is any state such that $P^{\text{MDP-1}}_{\mathbf{x},\mathbf{y}}(\mathbf{u}) > 0$, then $\mathbf{y}$ satisfies $y_n \le \tau_n, \forall n$, which shows, $\mathbf{y} = \mathbf{y} \wedge \boldsymbol{\tau}$.
2) There exists an i such that the initial state $\mathbf{x}$ satisfies $x_i = \tau_i$. Let us first assume there is only one client i satisfying $x_i = \tau_i$ and that $x_j < \tau_j, \forall j \ne i$. Then, for any action $\mathbf{u}$, if $\mathbf{y}$ is any state such that $P^{\text{MDP-1}}_{\mathbf{x},\mathbf{y}}(\mathbf{u}) > 0$, we have $y_j \le \tau_j, \forall j \ne i$, and also y_i is either 0 or $\tau_i + 1$. If $y_i = 0$ and $y_j \le \tau_j, \forall j \ne i$, then $(x_i + 1 - \tau_i)^+ \mathbf{1}(y_i = 0) = 1$ and $\mathbf{y} = \mathbf{y} \wedge \boldsymbol{\tau}$. If $y_i = \tau_i + 1$, and $y_j \le \tau_j, \forall j \ne i$, then from Lemma 1, $V_{T-1}(\mathbf{y}) = 1 + V_{T-1}(\mathbf{y} \wedge \boldsymbol{\tau})$. Thus, when there is only one client i satisfying $x_i = \tau_i$, the r.h.s (right-hand side) of (2) can be rewritten as,

   $$\min_{\mathbf{u}:\sum_n u_n \le L} \mathrm{E}\left\{ \eta \sum_n E_n u_n + 1 + \sum_{\mathbf{y}} P^{\text{MDP-1}}_{\mathbf{x},\mathbf{y}} V_{T-1}(\mathbf{y} \wedge \boldsymbol{\tau}) \right\}.$$

   The case where there are one or more clients $j \ne i$ satisfying $x_j = \tau_j$ is proved similarly.

□

The following lemma can be easily derived, the proof of which is omitted due to space constraints.

Lemma 3. $Y(t) := X(t) \wedge \boldsymbol{\tau}$ is a Markov Decision Process

$$\begin{aligned} \text{with}\quad & \mathrm{P}\left[Y(t+1) | Y(t), \cdots, Y(0), U(t), \cdots, U(0)\right] \\ =& \mathrm{P}\left[Y(t+1) | Y(t), U(t)\right]. \end{aligned}$$

Now we construct another MDP, denoted *MDP-2*, which is equivalent to the MDP-1 in an appropriate sense. We will slightly abuse notation and continue to use the symbols $Y(t)$ and $U(t)$ for states and controls.

For $Y_n(0) \in \{0, 1, \cdots, \tau_n\}$, let $Y_n(t)$ evolves as,

$$Y_n(t+1) = \begin{cases} 0 & \text{if a packet is delivered for client } n \text{ at } t, \\ (Y_n(t) + 1) \wedge \tau_n & \text{otherwise.} \end{cases}$$

Denote by $P^{\text{MDP-2}}_{\mathbf{x},\mathbf{y}}$ the transition probabilities of the resulting process $Y(t) := (Y_1(t), \cdots, Y_N(t))$ on the state space $\mathbb{Y} := \prod_{n=1}^{N} \{0, 1, \cdots, \tau_n\}$, where the transition probabilities,

$$\begin{aligned} &\mathrm{P}\left[Y_n(t+1) = y_n \middle| Y_n(t) = x_n, U_n(t) = u_n\right] \\ =& \begin{cases} p_n & \text{if } y_n = 0 \text{ and } u_n = 1, \\ 1 - p_n & \text{if } y_n = (x_n + 1) \wedge \tau_n \text{ and } u_n = 1, \\ 1 & \text{if } y_n = (x_n + 1) \wedge \tau_n \text{ and } u_n = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

The optimal cost-to-go function for MDP-2 is,

$$V_T(\mathbf{x}) := \min_{\pi:\sum_n U_n(t) \le L} \mathrm{E}\Big\{ \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mathbf{1}\{Y_n(t) = \tau_n\} + \eta E_n U_n(t) \Big| Y(0) = \mathbf{x} \Big\}, \forall \mathbf{x} \in \mathbb{Y}. \tag{5}$$

Theorem 4. MDP-2 is equivalent to the MDP-1 in that:

1) MDP-2 has the same transition probabilities as the accompanying process of MDP-1, i.e., the process $X(t) \wedge \boldsymbol{\tau}$;
2) Both MDPs satisfy the recursive relationship in (3); thus, their optimal cost-to-go functions are equal for each starting state $\mathbf{x}$ with $x_n \le \tau_n, \forall n$;
3) Any optimal control for MDP-1 in state $\mathbf{x}$ is also optimal for MDP-2 in state $\mathbf{x} \wedge \boldsymbol{\tau}$.

Proof. Statement 1) directly follows Lemma 3. The DP recursion for the optimal cost in MDP-2 is

$$V_T(\mathbf{x}) = \min_{\mathbf{u}:\sum_n u_n \le L} \mathrm{E}\Big\{ \sum_n \left(\eta E_n u_n + \mathbf{1}\left\{x_n = \tau_n\right\}\right) + \sum_{\mathbf{y}} P^{\text{MDP-2}}_{\mathbf{x},\mathbf{y}} V_{T-1}(\mathbf{y})\Big\}. \tag{6}$$

Thus, statement 2) is obtained from (6) and Corollary 2. In addition, statement 3) follows Lemma 1 and statement 1). □

As a result, we focus on MDP-2 in the sequel.

IV. Optimal Index Policy for the Relaxed Problem

A. Formulation of Restless Multi-armed Bandit Problem

MDP-2, with a finite state space, can be solved in a finite number of steps by standard DP techniques (see [7]). However, even for a finite time-horizon, it suffers from high computational complexity, since the cardinality of the state space increases exponentially in the number N of clients.

To overcome this, we formulate MDP-2 as an infinite-horizon restless multi-armed bandit problem ([8], [9]), and obtain an Index policy which has low complexity.

We begin with some notations: Denote by α the *maximum fraction* of clients that can simultaneously transmit in a time slot, i.e., $\alpha = L/N$. The process $Y_n(t)$ associated with client n is denoted as *project* n in conformity with the bandit nomenclature. If $U_n(t) = 1$, the project n is said to be *active* in slot t; while if $U_n(t) = 0$, it is said to be *passive* in slot t.

The infinite-horizon problem is to solve, with $Y(0) = \mathbf{x} \in \mathbb{Y}$,

$$\max_{\pi} \ \liminf_{T \to +\infty} \frac{1}{T} \mathrm{E}\Big[\sum_{t=0}^{T-1} \sum_{n=1}^{N} -\mathbf{1}\{Y_n(t) = \tau_n\} - \eta E_n U_n(t) \Big] \tag{7}$$

$$\text{s.t.} \qquad \sum_{n=1}^{N} (1 - U_n(t)) \ge (1 - \alpha) N, \ \forall t. \tag{8}$$

Note that the system reward is considered instead of the system cost.

B. Relaxations

We consider an associated relaxation of the problem (7)-(8) which puts a constraint only on the *time average* number of active projects allowed:

$$\max_{\pi} \ \liminf_{T\to+\infty} \frac{1}{T}\mathrm{E}\Bigg[\sum_{t=0}^{T-1}\sum_{n=1}^{N} -\mathbf{1}\{Y_n(t)=\tau_n\}-\eta E_n U_n(t)\Bigg] \tag{9}$$

$$\text{s.t.} \ \liminf_{T\to+\infty} \frac{1}{T}\mathrm{E}\left[\sum_{t=0}^{T-1}\sum_{n=1}^{N}(1-U_n(t))\right] \geq (1-\alpha)N. \tag{10}$$

Since constraint (10) relaxes the stringent requirement in (8), it provides an upper bound on the achievable reward in the original problem.

Let us consider the Lagrangian associated with the problem (9)-(10), with $Y(0)=\mathbf{x}\in\mathbb{Y}$,

$$\begin{aligned} l(\pi,\omega) := &\liminf_{T\to+\infty}\frac{1}{T}\mathrm{E}_\pi\Bigg[\sum_{t=0}^{T-1}\sum_{n=1}^{N}-\mathbf{1}\{Y_n(t)=\tau_n\}-\eta E_n U_n(t)\Bigg] \\ &+\omega\liminf_{T\to+\infty}\frac{1}{T}\mathrm{E}_\pi\left[\sum_{t=0}^{T-1}\sum_{n=1}^{N}(1-U_n(t))\right]-\omega(1-\alpha)N, \end{aligned}$$

where π is any history-dependent scheduling policy, while $\omega \geq 0$ is the Lagrangian multiplier. The Lagrangian dual function is $d(\omega) := \max_\pi l(\pi,\omega)$:

$$\begin{aligned} d(\omega) \leq &\max_{\pi} \ \liminf_{T\to+\infty}\frac{1}{T}\mathrm{E}\Bigg[\sum_{t=0}^{T-1}\sum_{n=1}^{N}-\mathbf{1}\{Y_n(t)=\tau_n\} \\ &-\eta E_n U_n(t)+\omega\left(1-U_n(t)\right)\Bigg|Y(0)=\mathbf{x}\Bigg]-\omega(1-\alpha)N \\ \leq &\max_{\pi} \ \limsup_{T\to+\infty}\frac{1}{T}\mathrm{E}\Bigg[\sum_{t=0}^{T-1}\sum_{n=1}^{N}-\mathbf{1}\{Y_n(t)=\tau_n\} \\ &-\eta E_n U_n(t)+\omega\left(1-U_n(t)\right)\Bigg|Y(0)=\mathbf{x}\Bigg]-\omega(1-\alpha)N \\ \leq &\max_{\pi} \ \sum_{n=0}^{N}\limsup_{T\to+\infty}\frac{1}{T}\mathrm{E}\Bigg[\sum_{t=0}^{T-1}-\mathbf{1}\{Y_n(t)=\tau_n\} \\ &-\eta E_n U_n(t)+\omega\left(1-U_n(t)\right)\Bigg|Y(0)=\mathbf{x}\Bigg]-\omega(1-\alpha)N, \end{aligned} \tag{11}$$

where the first and the third inequalities hold because of the super/sub-additivities of the limit inf/sub (respectively).

Now, consider the unconstrained problem in the last two lines of (11). It can be viewed as a composition of N independent ω-subsidy problems interpreted as follows: For each client n, besides the original reward $-\mathbf{1}\{Y_n(t)=\tau_n\}-\eta E_n U_n(t)$, when $U_n(t)=0$, it receives a subsidy ω for being passive.

Thus, the *ω-subsidy problem* associated with client n is defined as,

$$\begin{aligned} R_n(\omega) = &\max_{\pi_n} \ \limsup_{T\to+\infty}\frac{1}{T}\mathrm{E}\Bigg[\sum_{t=0}^{T-1}-\mathbf{1}\{Y_n(t)=\tau_n\} \\ &-\eta E_n U_n(t)+\omega\left(1-U_n(t)\right)\Bigg|Y_n(0)=x_n\Bigg], \end{aligned} \tag{12}$$

where π_n is a history dependent policy which decides the action $U_n(t)$ for client n in each time-slot.

In the following, we first solve this ω-subsidy problem, and then explore its properties to show that strong duality holds for the relaxed problem (9)-(10), and thereby determine the optimal value for the relaxed problem.

For $\theta \in \{0,1,\cdots,\tau_n\}$ and $\rho\in[0,1]$, we define $\sigma_n(\theta,\rho)$ to be a *threshold policy* for project n, as follows: The policy $\sigma_n(\theta,\rho)$ keeps the project passive at time t if $Y_n(t)<\theta$. However when $Y_n(t)>\theta$, the project is activated, i.e., $U_n(t)=1$. If $Y_n(t)=\theta$, then at time t, the project stays passive with probability ρ, and is activated with probability $1-\rho$.

For each project n, associate a function

$$W_n(\theta) := p_n(\theta+1)(1-p_n)^{\tau_n-(\theta+1)}-\eta E_n, \tag{13}$$

where $\theta = 0,1,\cdots,\tau_n-1$. (We elaborate on the physical meaning of $W_n(\cdot)$ later in Section V).

Lemma 5. Consider the ω-subsidy problem (12) for project n. Then,

1) $\sigma_n(0,0)$ is optimal iff the subsidy $\omega \leq W_n(0)$.
2) For $\theta \in \{1,\cdots,\tau_n-1\}$, $\sigma_n(\theta,0)$ is optimal iff the subsidy ω satisfies $W_n(\theta-1)\leq\omega\leq W_n(\theta)$.
3) $\sigma_n(\tau_n,0)$ is optimal iff $\omega = W_n(\tau-1)$.
4) $\sigma_n(\tau_n,1)$ is optimal iff $\omega \geq W_n(\tau-1)$.

In addition, for $\theta\in\{0,1,\ldots,\tau\}$, the policies $\{\sigma_n(\theta,\rho):\rho\in[0,1]\}$ are optimal when,

i) $0\leq\theta\leq\tau-1$ and $\omega=W_n(\theta)$,
ii) $\theta=\tau$ and $\omega=W_n(\tau-1)$.

Furthermore, for any $\theta\in\{0,\cdots,\tau\}$, under the $\sigma(\theta,0)$ policy, the average reward earned is,

$$\frac{p_n\theta\omega-\eta E_n-(1-p_n)^{\tau_n-\theta}}{1+\theta p_n}. \tag{14}$$

Meanwhile, under the $\sigma_n(\tau_n,1)$ policy, the reward is $\omega-1$.

Proof. For the ω-subsidy problem of project n, let us first analyze the $\sigma_n(\theta,0)$ policy. The subscript n is suppressed in the following. For each $\theta\in\{0,1,\cdots,\tau\}$, $\sigma(\theta,0)$ is a deterministic stationary policy. That is, for each $\sigma(\theta,0)$, there exists a function $g(\cdot)$ defined on the state space $\{0,1,\cdots,\tau\}$ of the project, such that $U_n(t)=g(Y_n(t))$. Further, there exist a real number R and a real function f on the state space with $f(0)=0$ such that,

$$\begin{aligned} R+f(i) = &-\mathbf{1}\{i=\tau\}-g(i)E\eta+\omega\left(1-g(i)\right) \\ &+pg(i)f(0)+(1-p)g(i)f\Big((i+1)\wedge\tau\Big) \\ &+\Big(1-g(i)\Big)f\Big((i+1)\wedge\tau\Big), \forall i=0,1,\cdots,\tau. \end{aligned}$$

The value of R and $f(i), i=1,\cdots,\tau$ can be obtained by solving the τ+1 equations above, and it can be shown that the R is the average expected system reward under this $\sigma(\theta,0)$ policy (see [7]). Then, by standard results in infinite-horizon dynamic programming, see [7], policy $\sigma(\theta,0)$ is optimal if and only if the following optimality equation is satisfied,

$$R+f(i)=\max_{u\in\{0,1\}}\Big\{-\mathbf{1}\{i=\tau\}-uE\eta+\omega(1-u)\\+puf(0)+(1-p)uf\Big((i+1)\wedge\tau\Big)\\+(1-u)f\Big((i+1)\wedge\tau\Big)\Big\},\forall i=0,\cdots,\tau. \quad (15)$$

Similar results hold for the policy $\sigma(\tau,1)$, under which the system is always passive. The conditions in 1)-4), and the average expected system reward under these policies are obtained.

To obtain the conditions i) and ii), note that $\sigma(\theta,0)=\sigma(\theta+1,1)$, and the policy $\sigma(\theta,\rho),\rho\in(0,1)$ can be regarded as a combination of $\sigma(\theta,0)$ and $\sigma(\theta,1)$. □

Theorem 6. For the relaxed problem (9)-(10) and its dual $d(\omega)$, the following results hold:

1) The dual function $d(\omega)$ satisfies,

$$d(\omega)=\sum_{n=0}^{N-1}R_n(\omega)-\omega(1-\alpha)N.$$

2) Strong duality holds, i.e., the optimal average reward for the relaxed problem, denoted R_{rel}, satisfies,

$$R_{\text{rel}}=\min_{\omega\geq 0}d(\omega).$$

3) In addition, $d(\omega)$ is a convex and piecewise linear function of ω. Thus, the value of R_{rel} can be easily obtained.

Proof. For 1), it follows from Lemma 5 that for the ω-subsidy problem associated with each project n, there is at least one stationary optimal policy, and under this policy, the optimality equation holds true. Thus, under the optimal policy, the limit of the time average reward exists (which is closely related to the optimality equation, see [7]). That is, the $\limsup_{T\to+\infty}$ in (12) can be replaced by $\lim_{T\to+\infty}$. As a result, all the "less than or equal to" in (11) can be replaced by equality signs. This proves the first statement.

For 2), the strong duality is proved by showing complementary slackness. The details are omitted due to space constraints.

For 3), it follows from equation (14) that each $R_n(\omega)$ is a piecewise linear function. To prove convexity of $R_n(\omega)$, note that the reward earned by any policy is a linear function of ω, and the supremum of linear functions is convex. Thus, by statement 1), $d(\omega)$ is also convex and piecewise linear. In addition, since each $R_n(\omega)$ can be easily derived from Lemma 5, the expression of $d(\omega)$ easily follows. Thus, R_{rel}, which is the minimum value of this known, convex, and piecewise linear function $d(\omega)$, can be easily obtained. □

V. The Large Client Population Asymptotic Optimality of the Index Policy

The *Whittle index* (see [8]) $W_n(i)$ of project n at state i is defined as the value of the subsidy that makes the passive and active actions equally attractive for the ω-subsidy problem associated with project n in state i. The n-th project is said to be *indexable* if the following is true: Let $B_n(\omega)$ be the set of states for which project n would be passive under an optimal policy for the corresponding ω-subsidy problem. Project n is *indexable* if, as ω increases from $-\infty$ to $+\infty$, the set $B_n(\omega)$ increases monotonically from $\emptyset$ to the whole state space of project n. The bandit problem is indexable if each of the constituent projects is indexable.

Lemma 7. The following are true:

1) The Whittle index $W_n(i)$ of project n at state i is,

$$W_n(i)=p_n(i+1)(1-p_n)^{\tau_n-(i+1)}-\eta E_n,$$

when $i=0,1,\cdots,\tau_n-1$; while $W_n(\tau_n)=W_n(\tau_n-1)$.

2) The stringent-constraint scheduling problem (7)-(8) is indexable.

3) For each project n, the transition rates of its states in the associated ω-subsidy problem form a unichain (there is a state $j\in\{0,1,\cdots,\tau_n\}$ such that there is a path from any state $i\in\{0,1,\cdots,\tau_n\}$ to state j), regardless of the policy employed.

Proof. Statements 1) and 2) directly follow from Lemma 5 and the definition of Whittle index, indexability. To prove statement 3), note that since $p_n<1$, there is a positive probability that there is no packet delivery for τ_n successive time slots, regardless of the policy employed. Thus, from any state $i\in\{0,1,\cdots,\tau_n\}$, there is a path to the state τ_n. □

As a result, the Whittle indices induce a well-defined order on the state values of each project. This gives the following heuristic policy.

Whittle Index Policy: At the beginning of each time slot t, client n is scheduled if its Whittle index $W_n(Y_n(t))$ is positive, and, moreover, is within the top αN index values of all clients in that slot. Ties are broken arbitrarily, with no more than αN clients simultaneously scheduled.

Now, we show the asymptotic optimality property of the Whittle Index Policy. Classify the N projects into K *classes* such that the projects in the same class have the same values of p_n, τ_n and E_n, while projects not in the same class differ in at least one of these parameters. For each class $k\in\{1,\cdots,K\}$, denote by γ_k the *proportion* of total projects that it contains; that is, there are $\gamma_k N$ projects in class k.

Assumption 1. Construct the *fluid model* of the restless bandit problem (7)-(8) as in [9] and [11], and denote the fluid process as $\mathbf{z}(t)$. We assume that, under the Whittle Index Policy, $\mathbf{z}(t)$ satisfies the global attractor property. That is, there exists $\mathbf{z}^\star$ such that from any initial point $\mathbf{z}(0)$, the process $\mathbf{z}(t)$ converges to the point $\mathbf{z}^\star$, under the Whittle Index Policy.

This assumption is not restrictive because of the following: First note that the MDP-2 itself also satisfies the unichain property. Then, under the Whittle Index policy, MDP-2 also forms a unichain. As a result, this N client bandit problem has a single recurrent class, and has a global attractor. Thus, it is not restrictive to assume that its fluid model also satisfies the global attractor property.

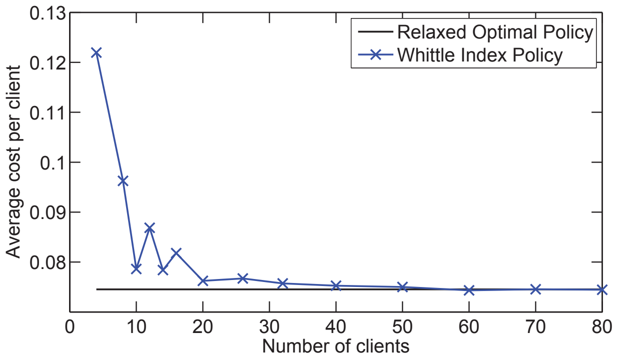

Fig. 1. The time average cost per client vs. the total number of clients for the optimal policy under the relaxed constraint and the Whittle Index policy are shown. (The parameters are $\alpha = 0.3$, $\eta = 0.1$, with $K = 2$ classes of projects, and $\gamma_1 = 0.5$, $\gamma_2 = 0.5$ proportion of projects in each class. For each client n in the first class, $p_n = 0.6$, $\tau_n = 10$, $E_n = 2$; while for each client n in the second class, $p_n = 0.8$, $\tau_n = 5$, $E_n = 3$.)

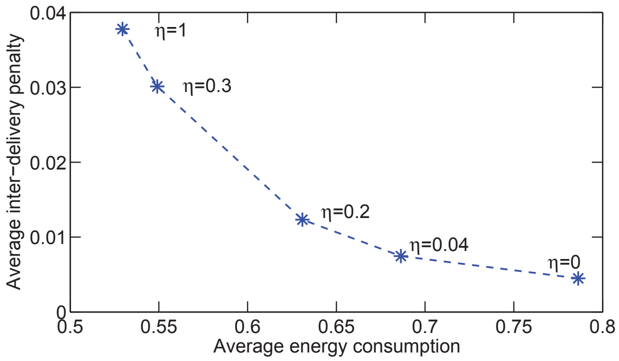

Fig. 2. The time average inter-delivery penalty per client vs. the time average energy consumption per client under the Whittle Index Policy for different values of energy-efficiency parameter η are shown. (The parameters are $N = 100$, $\alpha = 0.3$, with $K = 2$ classes of projects, and $\gamma_1 = 0.5$, $\gamma_2 = 0.5$ proportion of projects in each class. For each client n in the first class, $p_n = 0.6$, $\tau_n = 10$, $E_n = 2$; while for each client n in the second class, $p_n = 0.8$, $\tau_n = 5$, $E_n = 3$.)

Theorem 8. When Assumption 1 holds, as the number N of clients increases to infinity, $R_{\text{ind}}/N \to R_{\text{rel}}/N$, where R_{ind} and R_{rel} is the system reward under the Whittle Index policy and the optimal relaxed policy, respectively. (Here, the fraction of active bandit α and the proportion of each bandit class γ_k remain the same when N increases. In addition, the client number N is such that all $\gamma_k N$ are integers.) Thus, the Whittle Index policy is asymptotically optimal.

Proof. By Assumption 1 and Lemma 7, $R_{\text{ind}}/N \to R_{\text{rel}}/N$ directly from the result in [11]. Note that R_{rel} is an upper-bound for the stringent-constraint problem; thus, the asymptotic optimality holds. □

VI. Simulation Results

We now present the results of simulations of Whittle Index policy with respect to its average cost per client. The numerical results of the relaxed-constraint problem (9)-(10), which is derived by Theorem 6 and Lemma 5, are also employed to provide a bound on the stringent-constraint problem.

Fig. 1 illustrates the average cost per client under the relaxed optimal policy and the Whittle Index policy for different total numbers of clients. It can be seen that when the total number of clients increases, the gap between the relaxed optimal cost and the cost under the Whittle Index policy shrinks to zero. Since the optimal cost of the relaxed-constraint problem serves as a lower bound on the cost in the stringent-constraint problem, this means the Whittle Index policy approaches the optimal cost as the total number of clients increases, i.e., the Whittle Index policy is asymptotically optimal.

Fig. 2 illustrates the average inter-delivery penalty per client versus the average energy consumption per client under the Whittle Index policy for different values of the energy-efficiency parameter η. As η increases, the average energy consumption decreases, while the average inter-delivery penalty increases. Thus, there is a tradeoff between energy-efficiency and inter-delivery regularity. By changing η, we can balance these two important considerations.

Acknowledgment

This work is sponsored in part by the National Basic Research Program of China (2012CB316001), the Nature Science Foundation of China (61201191, 60925002, 61021001), NSF under Contract Nos. CNS-1302182 and CCF-0939370, AFOSR under Contract No. FA-9550-13-1-0008, and Hitachi Ltd.

References

[1] R. Li, A. Eryilmaz, and B. Li, “Throughput-optimal wireless scheduling with regulated inter-service times,” in *INFOCOM, 2013 Proceedings IEEE*, April 2013, pp. 2616–2624.

[2] B. Li, R. Li, and A. Eryilmaz, “Heavy-traffic-optimal scheduling with regular service guarantees in wireless networks,” in *Proceedings of the Fourteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc ’13. ACM, 2013, pp. 79–88.

[3] R. Singh, I.-H. Hou, and P. Kumar, “Pathwise performance of debt based policies for wireless networks with hard delay constraints,” in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, Dec 2013, pp. 7838–7843.

[4] ——, “Fluctuation analysis of debt based policies for wireless networks with hard delay constraints,” in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 2400–2408.

[5] R. Singh and A. Stolyar, in *Sigmetrics, title=MaxWeight Scheduling: Asymptotic Behavior of Unscaled Queue-Differentials in Heavy Traffic, year=2015,*.

[6] R. Singh, X. Guo, and P. R. Kumar, “Index policies for optimal mean-variance trade-off of inter-delivery times in real-time sensor networks,” in *INFOCOM*, 2015.

[7] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1994.

[8] P. Whittle, “Restless bandits: Activity allocation in a changing world,” *Journal of Applied Probability*, vol. 25, pp. pp. 287–298, 1988.

[9] R. R. Weber and G. Weiss, “On an index policy for restless bandits,” *Journal of Applied Probability*, vol. 27, no. 3, pp. pp. 637–648, 1990.

[10] D. P. Bertsekas, *nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[11] I. M. Verloop, “Asymptotic optimal control of multi-class restless bandits,” 2014. [Online]. Available: http://hal-univ-tlse2.archives-ouvertes.fr/docs/01/01/91/94/PDF/RBP.pdf