Index Policies for Optimal Mean-Variance Trade-Off of Inter-delivery Times in Real-Time Sensor Networks

Rahul Singh*, Xueying Guo† and P.R. Kumar*
*Department of Electrical and Computer Engineering, Texas A&M University. Email: {rsingh1,prk}@tamu.edu
†Department of Electronic Engineering, Tsinghua University. Email: guo-xy11@mails.tsinghua.edu.cn

***Abstract*—A problem of much current practical interest is the replacement of the wiring infrastructure connecting approximately 200 sensor and actuator nodes in automobiles by an access point. This is motivated by the considerable savings in automobile weight, simplification of manufacturability, and future upgradability.**

A key issue is how to schedule the nodes on the shared access point so as to provide regular packet delivery. In this and other similar applications, the mean of the inter-delivery times of packets, i.e., throughput, is not sufficient to guarantee service-regularity. The time-averaged variance of the inter-delivery times of packets is also an important metric.

So motivated, we consider a wireless network where an Access Point schedules real-time generated packets to nodes over a fading wireless channel. We are interested in designing simple policies which achieve optimal mean-variance tradeoff in interdelivery times of packets by minimizing the sum of time-averaged means and variances over all clients. Our goal is to explore the full range of the Pareto frontier of all weighted linear combinations of mean and variance so that one can fully exploit the design possibilities.

We transform this problem into a Markov decision process and show that the problem of choosing which node's packet to transmit in each slot can be formulated as a bandit problem. We establish that this problem is indexable and explicitly derive the Whittle indices. The resulting Index policy is optimal in certain cases. We also provide upper and lower bounds on the cost for any policy. Extensive simulations show that Index policies perform better than previously proposed policies.

I. Introduction

Traditionally, throughput and delay have been used as performance metrics to judge quality of service (QoS) [1]–[7]. The steady-state variance of inter-delivery times of packets is considered as a measure of service regularity in [8]. Motivated by cyber-physical systems applications serving sensors, we address the problem of achieving an optimal "mean-variance trade-off" in the inter-delivery times of packets of N clients sharing K channels.

We consider an access point with K channels shared by N clients. The clients desire a high throughput with high service regularity. We can associate a reward function $\frac{\theta_i}{\bar{D}_i} - var(D_i)$ with client i, where θ_i is the parameter that client i uses to tune its trade-off between its throughput $\frac{1}{\bar{D}_i}$ (where $\bar{D}_i$ is the mean inter-delivery time between packets of client i) and the service regularity $var(D_i)$, the variance of the inter-delivery times for client i. By varying θ_i one can explore the full range of design freedom along the Pareto frontier of all mean-variance tradeoffs. In summary, the net function which captures the trade-off is,

$$\sum_{i=1}^{N} R_i \left(\frac{\theta_i}{\bar{D}_i} - var(D_i) \right),$$

where $R_i > 0$ is the weight attached to client i, and θ_i is a tunable parameter permitting full exploration of the Pareto frontier.

Our contributions can be summarized as follows. We show how one may obtain tractable decoupled solutions for the problem of scheduling the clients by addressing it as a Restless Multi-Armed Bandit Problem [9]. In particular we obtain the Whittle indices in a closed form, which yields a very elegant solution based merely on comparing the indices of the clients. We also derive upper bounds on the achievable performance of any policy. Simulation results show that the performance of the obtained Index policy is very close to optimal.

II. Related Works

The steady-state variance of the inter-delivery times of packets of clients as a measure of service regularity has been considered in [8]. References [8] and [10] consider the scenario where multiple queues are sharing a server and deal with the problem of stabilizing the queues while ensuring an optimal delay and service regularity. [11], [12] perform an analysis of the pathwise starvations in service for the case of a single-hop multi-user wireless network.

A detailed intoduction to Restless Multi-Armed Bandit Problems (RMBP) can be found in [13]. RMBP and its relaxation were first introduced in [9]. The RMBP model

This paper is partially based on work supported by NSF under Contract Nos. CNS-1302182 and CCF-0939370, and AFOSR under Contract No. FA-9550-13-1-0008.

has been used earlier in works such as [14], which considered the problem of choosing an appropriate channel for up and downlink transmissions in multichannel access. Reference [15] is another notable work which uses the RMBP model and derives index policies for optimizing convex holding costs in a multiclass queue.

We also note that optimality of Index policies has been established in certain cases as the population of arms goes to infinity [16] and extensive simulations have shown that Index policies have "good" performance even in the finite population regime [15], [17]. References [18]–[21] consider minimization of variance as an objective in Markov Decision Process.

III. System Model

We consider the situation where time has been discretized into slots, and the duration of a slot corresponds to the time taken to attempt a packet transmission. Each client is assumed to have one packet at the beginning of each slot. In each slot, a scheduler chooses K out of the N clients, and attempts to deliver their packets. Channel unreliability is modeled by supposing that if client i is served in slot t, then the packet is delivered with probability p_i, independent of the past attempts. Moreover the service times are independent across clients. The scheduler has to choose the K clients transmitted in each slot so as to maximize the reward function,

$$\sum_{i=1}^{N} R_i \left(\frac{\theta_i}{\bar{D}_i} - var(D_i) \right), \tag{1}$$

where $\bar{D}_i$ and $var(D_i)$ are the mean and variance of the inter-delivery times of packets for client i in the steady state distribution.

IV. Markov Decision Process Formulation

The system state at time t is given by the vector $\mathbf{s}(t) := (s_1(t), \ldots, s_N(t))$, with $s_i(t)$ denoting the time slots elapsed between the latest delivery of a packet of client i, and t. Because time is discretized, the state vector $\mathbf{s}(t)$ is updated only at the beginning of slot t, and remains unchanged within the slot. The state thus evolves as,

$$s_i(t+1) = \begin{cases} s_i(t) + 1 \text{ if no packet of client } i \text{ is} \\ \text{delivered in slot } t, \\ 0 \text{ if a packet of client } i \text{ is delivered in} \\ \text{slot } t. \end{cases}$$

The Access Point (AP) takes a decision at the beginning of the slot t to grant channel access to K clients by choosing a control $\mathbf{u}(t) \in \{0,1\}^K, \sum_i^N u_i(t) = K$, where $u_i(t) = 1$ implies that client i will be granted channel access in slot t. The decision can be based on the entire past history of the system up to time t.

The "reward earned" at time t when the system is in state $\mathbf{s}$ is given by $\sum_{i=1}^{N} R_i \left(\theta_i \mathbb{1}\left(s_i = 0\right) - s_i\right)$, and thus is solely a function of the system state $\mathbf{s}$. With this set-up, the process $\mathbf{s}(t)$ becomes a controlled Markov process.

For a positive discount factor $\beta < 1$, the β-discounted optimization problem is to design control policy $\mathbf{u}(t)$ so as to maximize the expected infinite horizon discounted reward,

$$\liminf_{T\to\infty} \mathbb{E} \sum_{t=0}^{T} \beta^t \left(\sum_{i=1}^{N} R_i \left(\theta_i \mathbb{1}\left(s_i = 0\right) - s_i\right) \right). \tag{2}$$

Similarly the average reward problem is to maximize the expected infinite horizon time-average reward,

$$\liminf_{T\to\infty} \mathbb{E} \frac{1}{T} \sum_{t=0}^{T} \left(\sum_{i=1}^{N} R_i \left(\theta_i \mathbb{1}\left(s_i = 0\right) - s_i\right) \right). \tag{3}$$

It is easily verified that the above reward function reduces to,

$$\sum_{i=1}^{N} R_i \left(\frac{\theta_i}{\mathbb{E}\left(D_i\right)} - \mathbb{E}\left(\frac{D_i\left(D_i+1\right)}{2} \right) \right), \tag{4}$$

and thus differs slightly from the original reward function (1).

V. Whittle Index

We will pose the MDP of the previous section as a Restless Multiarmed Bandit Problem (RMBP). First we briefly describe the RMBP. A detailed discussion can be found in [9], [13], [22].

Consider a bandit which has N arms modeled as Markov processes. At each time a player can choose to play any $K < N$ arms and collect a reward from each arm, where the reward is a function of the current state of the arm that is played. The time evolution of each arm depends on whether it was chosen to play or not; thus the bandits (arms) are "restless" and evolve even if they are not played. The player has to choose the K arms to play at each time, so as to maximize the expected reward.

A "Whittle" policy, or "Index-based" policy, for the RMBP, calibrates each of the N arms by deriving N positive functions (called "index functions") $W_i(\cdot), i = 1, \ldots, N$, which are defined for each possible value that the state of arm i can assume. At time t the policy simply chooses to play the K arms having the K largest values of $W_i(s_i(t))$. After a re-labeling so that $W_1(s_1(t)) \geq W_2(s_2(t)) \geq W_N(s_N(t))$, the choices at time t are

$$u_i(t) = \begin{cases} 1 \text{ for } i = 1, 2, \ldots, K, \\ 0 \text{ otherwise.} \end{cases}$$

The derivation of the functions $W_i(\cdot)$ follows the following procedure. Each arm is considered in isolation from the rest of the arms, and the reward function is now modified so that the player receives, in addition to the original reward of the arm, a "subsidy" each time that he chooses not to play the arm (chooses "passive action"), and the goal once again is to maximize the

average reward. After having solved this problem, let us denote by $\Pi(w)$ the set of states that an optimal policy chooses to not play arm (stay passive). Then the arm is said to be *indexable* if for any two values of subsidies w_1, w_2, we have $w_1 > w_2 \implies \Pi(w_2) \subseteq \Pi(w_1)$, and the original MDP is said to be indexable if all the N arms are indexable. In case the MDP is indexable, the *Whittle Index* as a function of the state value s is defined as the smallest value of subsidy that makes an optimal policy choose the passive action when the client is in state n, i.e.,

$$W(n) = \inf\{w : n \in \Pi(w)\}. \tag{5}$$

Thus, the Whittle index measures, in a sense, the "value" of an arm as a function of the present state, and the Whittle or Index policy chooses those K arms which have the highest value amongst the N arms.

VI. The Client Scheduling Problem is Indexable

We will consider the β-discounted MDP, show that it is indexable and derive the corresponding Whittle index. The results for the average reward MDP will be obtained by letting $\beta \to 1$. We begin with a brief description of the single-arm β discounted reward problem.

Consider the following single client β discounted bandit problem parametrized by w and β. The subscripts are suppressed for convenience since the discussion below applies to each of the N clients. Thus $s(t), p$ are used in place of $s_i(t), p_i$.

There is a single client, whose state at time t, $s(t)$, is the time-elapsed-since-last-packet-delivery. At each time-slot, we can choose from the following two control actions: either attempt the transmission of a packet for it (active), or stay idle (passive). The reward earned at time t is $= -Rs(t) + w + R\theta\mathbb{1}\{s(t) = 0\}$ if the client chooses the passive action of not transmitting, while a reward of $-Rs(t) + R\theta\mathbb{1}\{s(t) = 0\}$ is earned if client chooses the active action of transmitting. If the action at time t is active, then $s(t+1)$, the state at time $t+1$, becomes 0 with probability p, and $s(t)+1$ with probability $1-p$. If the action at time t is passive, then $s(t+1) = s(t)+1$. The costs are additive over time after discounting by a factor β^t. A policy whether to be active or remain passive at time t when the system state at time t is $s(t) = s$.

We will prove that there is an optimal policy which is of **threshold type**, i.e. there is a threshold "elapsed time since last delivery" T (which depends on β, w, p), such that the policy which keeps the client passive in slot t if $s(t) < T$, and active if $s(t) \geq T$, is optimal.

By $c_i(T)$ we will denote the β-discounted reward earned by a policy when the system starts with an initial state value of i at time 0, and the policy with threshold at T is used. Let τ_i be the first time that state i is hit, i.e. $\tau_i = \min\{t \geq 1 : s(t) = i\}$. By "reward earned in the cycle $i \to j \to 0 \to i$" we will mean the reward earned by the system starting in state i in the time slots $0, \ldots, \tau_{i-1}$, while operating under the policy with threshold at j. Expressions involving reward-functions belonging to a single value of threshold are at times not mentioned as a function of threshold. X_p is a random variable that is geometrically distributed with parameter p. Also, we define $X := \mathbb{E}\beta^{X_p}$ and $Y := \mathbb{E}X_p\beta^{X_p}$.

Lemma 1: Consider the single client β discounted MDP.

1) $c_i(i+1) - c_i(i)$ is a linear increasing function of the subsidy w for all $i \geq 0$ It is strictly negative when $w = 0$.
2) For each $n \geq 0$, there exists a unique value of the subsidy, denoted $W(n)$, such that $c_n(n+1) = c_n(n)$.
3) $W(n) \geq W(n-1)$; thus $W(n)$ form an increasing sequence.
4) For all values of thresholds T, if $j > i \geq T$, then $c_i(T) > c_j(T)$.

Proof: For $T \geq 0$, the infinite horizon discounted reward earned starting in state i and following a policy with threshold $T+i$ is,

$$\begin{aligned} c_i(i+T) = {} & w\sum_{j=0}^{T-1}\beta^j - \sum_{j=0}^{T-1} R\,(i+j)\,\beta^j \\ & + R\beta^T\left[\mathbb{E}\left[-\sum_{j=0}^{X_p-1}(i+T+j)\,\beta^j\right]\right] \\ & + \beta^T\left(\mathbb{E}\beta^{X_p}\right)\left[R\theta + \sum_{j=0}^{i}(w - Rj)\,\beta^j\right] \\ & + \beta^{T+i}\left(\mathbb{E}\beta^{X_p}\right)c_i(i+T). \end{aligned}$$

Thus $c_i(i+T)$ depends on w as,

$$\begin{aligned} & \left[w\sum_{j=0}^{T-1}\beta^j + w\beta^T\left(\mathbb{E}\beta^{X_p}\right)\sum_{j=0}^{i-1}\beta^j\right] \Big/ \left[1 - \beta^{T+i}\left(\mathbb{E}\beta^{X_p}\right)\right] \\ = & w\left[\frac{1-\beta^T}{1-\beta} + \beta^T\frac{p\beta}{p\beta+1-\beta}\cdot\frac{1-\beta^i}{1-\beta}\right] \Big/ \left(1 - \beta^{T+i}\frac{p\beta}{p\beta+1-\beta}\right) \\ = & \frac{w\left[1-\beta+p\beta-\beta^T(1-\beta+p\beta^{i+1})\right]}{(1-\beta)\left(1-\beta+p\beta-\beta^{T+i+1}p\right)}. \end{aligned} \tag{6}$$

Thus $c_i(i+1) - c_i(i)$ depends on w as, $\frac{w(1-\beta)(1-\beta+p\beta)}{(1-\beta+p\beta-p\beta^{i+1})(1-\beta+p\beta-p\beta^{i+2})}$, which is linear and increasing in w.

Now we consider the case when $w = 0$. If C_1 is the cost of cycle $i \to i \to 0 \to i$, then it follows via a simple coupling argument that the cost of cycle $i \to i+1 \to 0 \to i+1$, denoted C_2, is given by,

$$C_2 = -Ri + \beta C_1 - R\beta\mathbb{E}\sum_{j=0}^{X_p-1}\beta^j,$$

and thus to prove the second result of the first statement, we only have to show that

$$\frac{C_1}{1-\beta^i X} - \frac{-Ri + \beta C_1 - R\beta\mathbb{E}\sum_{j=0}^{X_p-1}\beta^j}{1-\beta^i\beta X} > 0.$$

This is equivalent to showing that,

$$C_1 > -Ri \cdot \frac{1-\beta^i X}{1-\beta} - R\beta \left(\mathbb{E} \sum_{j=0}^{X_p-1} \beta^j \right) \cdot \frac{1-\beta^i X}{1-\beta}$$

We observe that $-Ri \cdot \frac{1-\beta^i X}{1-\beta} - R\beta \left(\mathbb{E} \sum_{j=0}^{X_p-1} \beta^j \right) \cdot \frac{1-\beta^i X}{1-\beta}$ is the reward earned over the cycle $i \to i \to 0 \to i$ if one were to modify the original cost function and instead charge a penalty of $-Ri$ for value of states $s(t) \leq i$ and a penalty of $-Rs(t)$ if $s(t) > i$. However since the original reward function is $= -Rs(t) + R\theta \mathbf{1}\{s(t) = 0\}$ (note that $w = 0$), a simple coupling argument shows that the reward earned is lower with the modified function. This completes the proof of first statement.

Note that from the first statement it follows that $c_n(n+1) - c_n(n)$ is a linear increasing function of w which is less than 0 at $w = 0$. Hence there exists a value of w such that the function $c_n(n+1) - c_n(n)$ vanishes, and moreover vanishes at a unique point since the slope of this function is strictly positive. This value of w, where the function $c_n(n+1) - c_n(n)$ vanishes, is $W(n)$.

Let C_1, C_2 be the costs of cycles $n \to n \to 0 \to n$ and $n \to n+1 \to 0 \to n$. It is seen that,

$$c_n(n) = \frac{C_1}{1-\beta^n X}, \quad c_n(n+1) = \frac{C_2}{1-\beta^n \beta X}. \tag{7}$$

Using a coupling argument we obtain,

$$C_2 = (W(n) - Rn) + \beta C_1 - R\beta \mathbb{E} \sum_{j=0}^{X_p-1} \beta^j. \tag{8}$$

Combining (7),(8) and the fact that for $w = W(n)$ we have $c_n(n) = c_n(n+1)$,

$$\frac{C_1}{1-\beta^n X} = \frac{(W(n) - Rn) + \beta C_1 - R\beta \mathbb{E} \sum_{j=0}^{X_p-1} \beta^j}{1-\beta^{n+1} X}, \text{or ,}$$

$$C_1 (1-\beta) = \left(W(n) - Rn - R\beta \mathbb{E} \sum_{j=0}^{X_p-1} \beta^j \right) (1-\beta^n X). \tag{9}$$

Now let us check if under the value of subsidy set to $W(n)$, we have $c_{n-1}(n) > c_{n-1}(n-1)$. If this is the case, then from the first statement of this lemma, we will deduce that $W(n-1) < W(n)$. Now, $c_{n-1}(n) > c_{n-1}(n-1)$ is equivalent to showing

$$\frac{W(n) - R(n-1) + \beta C_1 - \beta^n X (W(n) - R(n-1))}{1-\beta^n X} > \frac{C_1 + R\mathbb{E} \sum_{j=0}^{X_p-1} \beta^j - \beta^{n-1} X (W(n) - R(n-1))}{1-\beta^{n-1} X}.$$

After some algebraic manipulations and using (9) it can be shown that proving the above inequality is equivalent to proving $X > 0$, which indeed is true. This completes the proof of third statement.

For the fourth statement, using a coupling argument, we obtain, $c_j(T) = c_i(T) - R(j-i) \sum_{j=0}^{X_p-1} \beta^j$, and hence $c_j(T) < c_i(T)$. ■

Lemma 2: Let the subsidy be $w = W(n)$. Then for the single client β discounted MDP,

1) $c_i(n) = c_i(n+1), \forall i \geq 0$.
2) $c_{i-1}(n) \geq c_i(n), \forall i \geq 1$.

Proof: Firstly recall that for subsidy $= W(n)$, $c_n(n) - c_n(n+1) = 0$. Thus for $i = 0, 1, \ldots, n-1$,

$$c_i(n) - c_i(n+1) = \beta^{n-i} (c_n(n) - c_n(n+1)) = 0. \tag{10}$$

For $i \geq n+1$,

$$c_i(n+1) - c_i(n) = \beta X (c_0(n+1) - c_0(n)) = 0,$$

where the last equality follows from (10). This proves the first statement.

To prove the second result, consider the following cases:

i) For $i > n$, Lemma 1 implies that the inequality is true.
ii) For $2 \leq i \leq n$, denote d_i as the cost incurred in the cycle $n \to 0 \to i-1$. Then both $c_i(n)$, and $c_i(n+1)$ can be derived in terms of d_i. When subsidy is equal to $W(n)$, we have $c_i(n) = c_i(n+1)$, i.e.,

$$-\beta^{n-i}(1-\beta) d_i = \tag{11}$$
$$W(n) \left(\beta^n X - \beta^{n-i} \right) + R\beta^{n-i} n - \beta^n X i \tag{12}$$
$$+ R \frac{\beta^{n-i}}{1-\beta} \left(\beta(1-X) - \beta^{i+1} X + \beta^{n+1} X^2 \right). \tag{13}$$

where the first equality follows from statement 1. Similarly, $c_{i-1}(n) - c_i(n) \geq 0$ is equivalent to

$$\sum_{j=0}^{n-i-1} (W(n) - Ri - Rj) \beta^j + \beta^{n-i} d_i \geq \sum_{j=0}^{n-i} (W(n) - Ri + R - Rj) \beta^j + \beta^{n-i+1} d_i - \beta^n X (W(n) - Ri + R),$$

i.e.,

$$-\beta^{n-i}(1-\beta) d_i + R \frac{1-\beta^{n-i}}{1-\beta} + (W(n) - nR + R) \beta^{n-i} - \beta^n X (W(n) - Ri + R) \geq 0 \text{ or}$$

$$\frac{(1-\beta^n X)(\beta^i - \beta^{n+1} X)}{\beta^i (1-\beta)} \geq 0,$$

where the second-last equivalence follows from (11). We note that the last inequality holds trivially for all $\beta \in (0,1)$ and hence the statement 2 holds for $i = 2, \ldots, n$.
iii) $i = 1$. We compare the cost incurred by the system starting in state 0 over the cycle $0 \to n \to 0$ (say C_0) with the cost incurred over the cycle $j \to n \to 0 \to j$ when starting in state j (denoted C_j) via coupling the processes associated with the two systems constructed on the same probability space. Clearly $C_0 > C_j$. Thus $c_0(T) > c_j(T)$ for any value of threshold T.

■

Lemma 3: The function $w + p\beta\left(c_i(T) - c_0(T)\right)$ (which depends on w, i, T) is linear, increasing in w. Also,

$$W(n) + p\beta\left(c_{n+1}(n) - c_0(n)\right) = 0 \text{ for } n = 0, 1, \ldots. \quad (14)$$

Proof: We consider the following cases:

i) For $i \leq T$, it follows from (6) that the function $w + p\beta\left(c_i(T) - c_0(T)\right)$ depends on w as

$$\frac{1 - \beta - p\beta + p\beta^{T-i+1}}{1 - \beta + p\beta - p\beta^{T+1}} w. \quad (15)$$

We have $1 - \beta + p\beta - p\beta^{T+1} > 0, \forall \beta < 1$. Also, $1 - \beta - p\beta + p\beta^{T-i+1} \geq 1 - 2\beta + \beta^{T-i+1} > 0$ since the function

$$1 - 2\beta + \beta^k \geq 0, \forall k > 1, \beta \in (0, 1).$$

Thus, in the expression (15) the coefficient of w is positive.

ii) For $i \geq T + 1$, we have,

$$c_i(T) = \mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j (-i - j) + X c_0(T).$$

The dependence of $c_0(T)$ on w can be obtained from (6). Combining, $w + p\beta\left(c_i(T) - c_0(T)\right)$ depends on w as,

$$\frac{1 - \beta}{1 - \beta + p\beta - p\beta^{T+1}} w,$$

which has a positive slope with respect to w.

This completes the proof of first statement. Note that for $w = W(n)$, we have

$$c_n(n+1) = c_n(n).$$

This implies

$$\begin{aligned} &- Rn + W(n) + \beta c_{n+1}(n+1) = -Rn \\ &\quad + \beta\left(p c_0(n) + (1-p) c_{n+1}(n)\right) \text{ i.e.} \\ &W(n) + \beta c_{n+1} = \beta\left(p(c_0 + (1-p) c_{n+1}\right) \text{ and so} \\ &W(n) + p\beta\left(c_{n+1} - c_0\right) = 0. \end{aligned}$$

Above, in the second implication, we have used the first statement of Lemma 2 to remove the dependence of $c_i(\cdot)$ on the threshold values. ∎

Theorem 1: For the β-discounted MDP with subsidy $w \in [W(n), W(n+1))$, the policy with threshold at n is optimal. Thus the MDP is indexable and $W(n)$ is the Whittle index when the state is n.

Proof: Fix a $w \in [W(n), W(n+1))$. If the policy is indeed optimal, then the Dynamic Programming optimality equation would be satisfied. Hence we only need to verify the inequality

$$\begin{aligned} &- Ri + w + \beta c_{i+1} \geq -Ri + \beta\left[(1-p)\, c_{i+1} + p c_0\right], \\ &\quad \text{for } i = 0, 1, \ldots, n, \\ &\text{or, equivalently,} \quad w + \beta p\left(c_{i+1} - c_0\right) \geq 0, \end{aligned} \quad (16)$$

with strict inequality holding if $w \in (W(n), W(n+1))$, and equality holding for $i = n, w = W(n)$. Similarly for $i = n+1, n+2, \ldots$ we have to verify the inequality

$$w + \beta p\left(c_{i+1} - c_0\right) \leq 0. \quad (17)$$

We will first prove (16). We use superscripts to distinguish between costs c_i calculated under different values of subsidy. We have,

$$\begin{aligned} w + \beta p\left(c_{i+1}^w - c_0^w\right) &\geq W(n) + \beta p\left(c_{i+1}^{W(n)} - c_0^{W(n)}\right) \\ &= p\beta\left(c_0^{W(n)} - c_{n+1}^{W(n)}\right) + p\beta\left(c_{i+1}^{W(n)} - c_0^{W(n)}\right) \\ &= p\beta\left(c_{i+1}^{W(n)} - c_{n+1}^{W(n)}\right) \\ &\geq 0, \end{aligned}$$

where the first inequality and equality follow from Lemma 3, and the last inequality follows from Lemma 2.

To prove (17) we have,

$$\begin{aligned} w + \beta p\left(c_{i+1}^w - c_0^w\right) &\leq W(n+1) + \beta p\left(c_{i+1}^{W(n+1)} - c_0^{W(n+1)}\right) \\ &= p\beta\left(c_0^{W(n+1)} - c_{n+2}^{W(n+1)}\right) \\ &\quad + p\beta\left(c_{i+1}^{W(n+1)} - c_0^{W(n+1)}\right) \\ &= p\beta\left(c_{i+1}^{W(n+1)} - c_{n+2}^{W(n+1)}\right) \\ &\leq 0, \end{aligned}$$

where first two steps follow from Lemma 3, and the last inequality follows from Lemma 2. This completes the optimality of the policy with threshold at $W(n)$. Following 5, the Whittle index for the state n is thus given by

$$\inf\{w : n \in \Pi(w)\} = \inf\{w : w \geq W(n)\} = W(n),$$

where the first equality follows from the first statement of Theorem. ∎

We now proceed to explicitly derive the values of the indices $W(n)$.

Theorem 2:

$$\begin{aligned} W(n) &= \frac{p\beta(f_1 - f_2 - f_3 + f_4)}{f_5}, \text{ where ,} \\ f_1 &= \frac{1 - \beta^n}{(1-\beta)^2} \cdot \left((1 - X)\left[n(1-\beta) + \beta\right] - Y(1-\beta)\right), \\ f_2 &= \frac{\beta(1 - \beta^n) - \beta^n n(1-\beta)}{(1-\beta)^2} \cdot (1 - X), \\ f_3 &= \frac{1 - X}{1 - \beta}(1 - \beta^n X), \\ f_4 &= \theta\left(1 - X\right), \\ f_5 &= 1 - \beta^n X - p\beta\left(\frac{1 - \beta^n}{1 - \beta}\right)(1 - X) \\ &= \frac{1 - \beta}{1 - \beta + p\beta}. \end{aligned}$$

Proof: From (14) we have,

$$W(n) = p\beta(c_0 - c_{n+1})$$

$$= p\beta(c_0 - c_n - \mathbb{E}\sum_{j=0}^{X_p}\beta^j). \tag{18}$$

Now,

$$c_0 - c_n = \frac{C_0 - C_n}{1 - \beta^n \mathbb{E}\beta^{X_p}}, \tag{19}$$

where C_0, C_n are the costs over the cycles $0 \to n \to 0$ and $n \to n \to 0 \to n$. We can compute $C_0 - C_n$ as,

$$C_0 - C_n = \left(\mathbb{E}\sum_{j=0}^{X_p-1}(n+j)\beta^j\right)(1-\beta^n) \tag{20}$$

$$+ \left(\sum_{j=0}^{n-1}(W(n)-j)\beta^j\right)(1-\mathbb{E}\beta^{X_p}) + \theta\left(1-\beta^{X_p}\right). \tag{21}$$

Combining (18,19,20) and setting $\Delta = \mathbb{E}\sum_{j=0}^{X_p-1}(n+j)\beta^j$, we have,

$$W(n) = p\beta\left(\frac{\Delta(1-\beta^n) + \left(\sum_{j=0}^{n-1}(W(n)-j)\,\beta^j\right)\left(1-\mathbb{E}\beta^{X_p}\right)}{1-\beta^n\mathbb{E}\beta^{X_p}} - \mathbb{E}\sum_{j=0}^{X_p-1}\beta^j + \frac{\theta\left(1-\mathbb{E}\beta^{X_p}\right)}{1-\beta^n\mathbb{E}\beta^{X_p}}\right),$$

or,

$$W(n)\left[1 - p\beta\cdot\frac{\left(\sum_{j=0}^{n-1}\beta^j\right)\left(1-\mathbb{E}\beta^{X_p}\right)}{1-\beta^n\mathbb{E}\beta^{X_p}}\right] = p\beta\left(\frac{\Delta(1-\beta^n) + \left(\sum_{j=0}^{n-1}-j\beta^j\right)\left(1-\mathbb{E}\beta^{X_p}\right)}{1-\beta^n\mathbb{E}\beta^{X_p}} - \mathbb{E}\sum_{j=0}^{X_p-1}\beta^j + \frac{\theta\left(1-\mathbb{E}\beta^{X_p}\right)}{1-\beta^n\mathbb{E}\beta^{X_p}}\right),$$

which simplifies to,

$$W(n)\cdot f_5 = p\beta(f_1 - f_2 - f_3 + f_4). \tag{22}$$

■

Theorem 3: The Whittle indices for the average cost MDP are given by,

$$W^{\text{Avg}}(n) = \lim_{\beta\to 1} W^\beta(n) = nRp\cdot\left(\frac{n}{2} + \frac{1-p}{1+p} + \frac{1}{2}\right) + Rp\theta. \tag{23}$$

Proof: The expression (23) is easily derived from (22). It remains to show that the quantities $W^{\text{Avg}}(n)$ are indeed Whittle indices for the average-cost problem. Fix the subsidy to be w, and without loss of generality let $w \in \left(W^{\text{Avg}}(n), W^{\text{Avg}}(n+1)\right)$. Below we use superscripts to exhibit the dependence of the cost on β. Now,

$$c_0^\beta(n) = \frac{1}{1-\beta^n X}\cdot\left(w\frac{1-\beta^n}{1-\beta} + \frac{\beta(1-\beta^n) - n\beta^{n+1}(1-\beta)}{(1-\beta)^2} - \beta^n\sum_{j=1}^{X_p-1}(n+j)\beta^j + \frac{R\theta}{1-\beta^n X}\right), \text{ and so}$$

$$\lim_{\beta\uparrow 1}(1-\beta)c_0^\beta(n) = \lim_{\beta\uparrow 1}\left(w\frac{1-\beta^n}{1-\beta^n X} + \frac{\beta(1-\beta^n) - n\beta^{n+1}(1-\beta)}{(1-\beta)(1-\beta^n X)} - (1-\beta)\beta^n\sum_{j=1}^{X_p-1}(n+j)\beta^j + \frac{R\theta\,(1-\beta)}{1-\beta^n X}\right)$$

$$= w\frac{np}{np+1} + \frac{Rp(n^2+n)}{2(np+1)} + \frac{Rp\theta}{np+1} \tag{24}$$

$$< \infty.$$

Since for each m, $W^\beta(m) \to W^{\text{Avg}}(m)$, it follows from Theorem 2 that there exists a $\beta^\star(w)$ such that the policy with the threshold at n is optimal for the single client β-discounted MDP for all $\beta \in (\beta^\star(w), 1)$. However since $\lim_{\beta\uparrow 1}(1-\beta)c_0^\beta(n)$ exists, the policy with threshold at n is also optimal for the average cost problem. However since w can assume any value in the interval $\left(W^{\text{Avg}}(n), W^{\text{Avg}}(n+1)\right)$, the policy with threshold at n is optimal for the average cost MDP for each value of subsidy $w \in \left(W^{\text{Avg}}(n), W^{\text{Avg}}(n+1)\right)$. Thus,

$$\inf\{w : \text{optimal policy chooses active at } n\} \le W^{\text{Avg}}(n). \tag{25}$$

Similarly, picking subsidy $w < W^{\text{Avg}}(n)$ shows that the active action is not optimal for any value of subsidy $w < W^{\text{Avg}}(n)$. Hence,

$$\inf\{w : \text{optimal policy chooses active at } n\} = W^{\text{Avg}}(n), \tag{26}$$

and we obtain that $W^{\text{Avg}}(n)$ are indeed the Whittle indices for the average cost problem. ■

We note that the expression (24) is the average reward earned under the subsidy w and threshold at n. We will denote this quantity as $C^{Avg}(W, n)$.

VII. Bounds on Optimal Reward.

Lemma 4: For the average cost MDP, the reward obtained under any policy is upper-bounded by the value of the following optimization problem:

$$\max \sum_{i=1}^{N} R_i\left[\bar{D}_i^2 + \theta_i\frac{1}{\bar{D}_i}\right]$$

$$\text{such that } \sum_{i=1}^{N}\frac{1}{\bar{D}_i p_i} \le 1, \bar{D}_i \ge 0, i = 1,\ldots,N. \tag{27}$$

Proof: The random reward earned in time steps $1, 2, \ldots, t$ is given by,

$$C(t) := \sum_{i=1}^{N}\frac{R_i}{t}\left[-\sum_{l=1}^{N_i(t)} D_i(l)^2 + \theta_i N_i(t)\right],$$

where $N_i(t)$ is the number of packets of client i delivered by time t and $D_i(l)$ is the interdelivery time of l-th packet of client i. Let us assume that the average interdelivery-time for client i under a policy is equal to $\bar{D}_i$. Thus,

$$\begin{aligned}\liminf_{t\to\infty}\mathbb{E}C(t) &\le \limsup_{t\to\infty}\mathbb{E}C(t)\\ &\le \mathbb{E}\limsup_{t\to\infty}C(t)\\ &= \mathbb{E}\limsup_{t\to\infty}\sum_{i=1}^{N}R_i\left[\frac{\sum_{l=1}^{N_i(t)}D_i(l)^2}{t}+\frac{\theta_i N_i(t)}{t}\right]\\ &\le \sum_{i=1}^{N}R_i\left[\bar{D}_i^2+\theta_i\frac{1}{\bar{D}_i}\right],\end{aligned}$$

where the second inequality follows from Fatou's lemma and the last is Jensen's inequality. Thus solving the optimization problem (27) gives a lower bound on the performance of any policy. We note that the constraint $\sum_{i=1}^{N}\frac{1}{\bar{D}_i p_i}\le 1, \bar{D}_i \ge 0$ is simply the capacity of the wireless channel. ■

Next we consider the Lagrangian relaxation of the RMBP [9]. For this, we relax the constraint of choosing K arms at each time, to the constraint that one plays K arms on average, i.e., $\lim_{t\to\infty}\frac{\text{Total numbers of arms played by time } t}{t} = K$. Clearly the maximum possible reward in the relaxed problem is greater than or equal to the reward earned by any policy for the original RMBP. Also since the Index policy is the optimal solution to this relaxed problem ([9]), its value function serves as an upper-bound for the value function of the RMBP.

Lemma 5: Let $C^{Avg,i}$ be the average reward earned by the policy maximizing the single-client average reward under the subsidy W (24). Then the reward for the average cost MDP obtained by any policy is less than or equal to,

$$\begin{aligned}&\inf_{W>0}\sum_{i=1}^{N}C^{Avg,i}(W)-W(N-K)\\ &=\inf_{W>0}\left(\sum_{i=1}^{N}W\frac{n_i p_i}{n_i p_i+1}+\frac{R_i p_i(n_i^2+n_i)}{2(n_i p_i+1)}\right.\\ &\left.+\frac{R_i p_i\theta_i}{n_i p_i+1}-W(N-K)\right),\\ &=\inf_{W>0}\left[W\left(\sum_{i=1}^{N}\frac{n_i p_i}{n_i p_i+1}+K-N\right)+\frac{R_i p_i(n_i^2+n_i)}{2(n_i p_i+1)}\right.\\ &\left.+\frac{R_i p_i\theta_i}{n_i p_i+1}\right],\end{aligned}$$

where n_i is such that $W\in(W(n_i),W(n_i+1))$.

VIII. Optimality of Index Policy

Now we consider several special cases of interest.

Theorem 4: Consider the average cost problem for the case where all the clients are identical, i.e., $R_i\equiv 1$ and $p_i\equiv p$ for all the clients. The index policy is optimal in this case.

Proof: Firstly we note that in this symmetric case, the Index policy serves the client with the largest value of the state, i.e. the policy is, "largest time-since-last-service-first". We will prove the result only for the case of two clients, each having channel reliability p. The case where there are multiple such clients follows in a straighforward manner.

Consider the time-horizon at t. If (s_1,s_2) is the initial value of the state vector, and $R_t(\mathbf{s})$ is the maximum reward that can be earned when there are t time-slots to go, then the Dynamic Programming optimality equation becomes,

$$\begin{aligned}R_t[(s_1,s_2)] = &-(s_1+s_2)+(1-p)R_{t-1}[(s_1+1,s_2+1)]\\ &+p\max\{R_{t-1}[(0,s_2+1)],R_{t-1}[(s_1+1,0)]\},\end{aligned}$$

where the optimal action corresponds to the one maximizing the expression on the right hand side. Let us assume without loss of generality that $s_1<s_2$. Then $R_{t-1}[(0,s_2+1)]\le R_{t-1}[(s_1+1,0)]$, which implies that the optimal action is to serve client 2. ■

IX. Simulations

We have carried out simulations to compare the performance of the optimal policy which was obtained via the Policy Iteration tool-box in Matlab vs. the Index policy which was obtained in Theorem 3. We present three plots in Figures 1-3. In all the cases considered 2 clients share a single channel. To obtain Figure 1, we fix client 1's parameter as $p_1=.8,\theta_1=3,R_1=1$, while for client 2 we fix $\theta_2=3,R_2=1$ and vary p_2 from 0 to 1. For Figure 2, we fix Client 1 parameters to be $p_1=.8,\theta_1=3,R_1=1$ while for Client 2 we fix $p_2=.6,R_2=1$ and vary the value of θ_2 from 1 to 10. To obtain Figure 3, we fix Client 1's parameters as $p_1=.8,\theta_1=5,R_1=5$, and for Client 2 we fix the parameters $p_2=.6,\theta_2=5$ while varying the value of R_2.

We observe that Index policy gives near-optimal performance in all the cases.

X. Concluding Remarks

We have proposed an analytical framework for exploring the full range of mean vs. variance tradeoffs in inter-delivery times in wireless sensor nerworks, i.e. Throughput vs. Service Regularity trade-off. The problem can be formulated as Restless Multiarmed Bandit Problem and indices can be obtained in closed form. Simulations indicate near-optimal performance of the resulting Index policy.

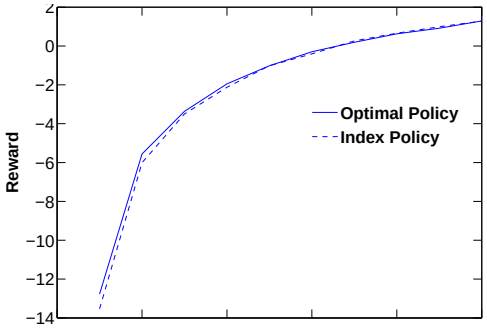

Fig. 1: Reward Optimal Policy vs. Index Policy for $p_1 = .8, \theta_1 = 3, R_1 = 1, \theta_2 = 3, R_2 = 1$, p_2 varying from .1 to 1.

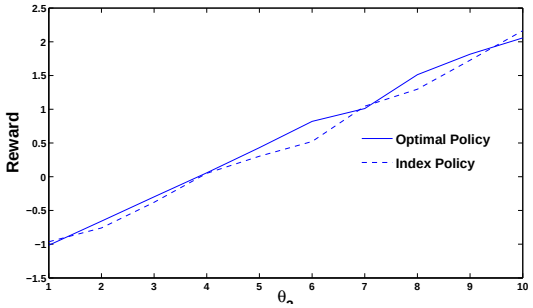

Fig. 2: Reward Optimal Policy vs.Index Policy for $p_1 = .8, \theta_1 = 3, R_1 = 1, p_2 = .6, R_2 = 1$ while θ_2 varies from 1 to 10.

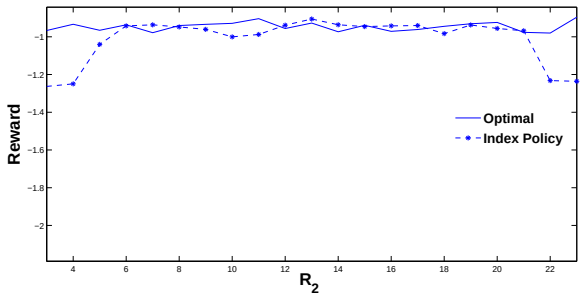

Fig. 3: Reward Optimal Policy vs. Index Policy for $p_1 = .8, \theta_1 = 5, R_1 = 5, p_2 = .6, \theta_2 = 5$ while R_2 is varied.

REFERENCES

[1] L. Bui, R. Srikant, and A. Stolyar, "Novel architectures and algorithms for delay reduction in back-pressure scheduling and routing," in *INFOCOM 2009, IEEE*, April 2009, pp. 2936–2940.

[2] Haozhi Xiong, Ruogu Li, A. Eryilmaz and E. Ekici, "Delay-aware cross-layer design for network utility maximization in multi-hop networks," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 5, pp. 951–959, May 2011.

[3] A. Eryilmaz and R. Srikant, "Asymptotically tight steady-state queue length bounds implied by drift conditions," *Queueing Syst. Theory Appl.*, vol. 72, no. 3-4, pp. 311–359, Dec. 2012.

[4] Xiaojun Lin and Ness B. Shroff, "Joint rate control and scheduling in multihop wireless networks," in *in Proceedings of IEEE Conference on Decision and Control*, 2004, pp. 1484–1489.

[5] A. L. Stolyar, "Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *The Annals of Applied Probability*.

[6] M.J. Neely, E. Modiano and Chih-ping Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 2, pp. 396–409, April 2008.

[7] Xueying Guo, Sheng Zhou, Zhisheng Niu and P.R. Kumar, "Optimal wake-up mechanism for single base station with sleep mode," in *International Teletraffic Congress (ITC)*, Sept 2013.

[8] 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.

[9] P. Whittle, "Multi-armed bandits and the Gittins index," *J. R. Statist. Soc. B*, vol. **42**, pp. 143–149, 1980.

[10] 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. New York, NY, USA: ACM, 2013, pp. 79–88.

[11] Rahul Singh, I-Hong Hou and P.R. Kumar, "Fluctuation analysis of debt based policies for wireless networks with hard delay constraints," in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 2400–2408.

[12] ——, "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.

[13] K. G. J.C. Gittins and R. Weber, *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

[14] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, Nov 2010.

[15] P. S. Ansell, K. D. Glazebrook, J. Nio-Mora, and M. O'Keeffe, "Whittle's index policy for a multi-class queueing system with convex holding costs," *Mathematical Methods of Operations Research*, vol. 57, no. 1, pp. 21–39, 2003. [Online]. Available: http://dx.doi.org/10.1007/s001860200257

[16] R.R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, vol. 27, Sep 1990.

[17] K.D. Glazebrook, D. Ruiz-Hernandez and C. Kirkbride, "Some indexable families of restless bandit problems," *Advances in Applied Probability*, vol. 38, no. 3, pp. 643–672, 2006.

[18] J.A. Filar, L.C.M. Kallenberg and H.M. Lee, "Variance-penalized markov decision processes," *Math. Oper. Res.*, vol. 14, no. 1, pp. 147–161, Mar. 1989.

[19] Hajime Kawai, "A variance minimization problem for a markov decision process," *European Journal of Operational Research*, vol. 31, no. 1, pp. 140–145, 1987.

[20] Masami Kurano, "Markov decision processes with a minimum-variance criterion," *Journal of Mathematical Analysis and Applications*, vol. 123, no. 2, pp. 572–583, 1987.

[21] Eitan Altman and Adam Shwartz, "Markov decision problems and state-action frequencies," *SIAM J. CONTROL AND OPTIMIZATION*, vol. 29, no. 4, pp. 786–809.

[22] P. Whittle, "Restless bandits: activity allocation in a changing world," pp. 287–298, 1988.